BIOMARKER DISCOVERY IN EXOME DATA

by

An-kwok Ian Wong

B.S.E., Biomedical Engineering &

Electrical and Computer Engineering, Duke University, 2007

M.S., Intelligent Systems Program / Biomedical Informatics,

University of Pittsburgh, 2009

M.D., Case Western Reserve University, 2015

Submitted to the Graduate Faculty of

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

fulfillment

of the requirements for the degree of

Doctor of Philosophy in Intelligent Systems

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

An-kwok Ian Wong

It was defended on

27 June 2016

and approved by

Shyam Visweswaran, Department of Biomedical Informatics

M. Michael Barmada, Department of Human Genetics

Gregory F. Cooper, Department of Biomedical Informatics

Milos Hauskrecht, Department of Computer Science

Dissertation Director: Shyam Visweswaran, Department of Biomedical Informatics

BIOMARKER DISCOVERY IN EXOME DATA

An-kwok Ian Wong, PhD

University of Pittsburgh, 2016

Current DNA sequencing technology enables inexpensive sequencing of the exome or the protein-coding regions of the genome. The primary goal of the analyses of exome data is to identify sequence variants, such as single nucleotide variations (SNVs), that will help elucidate the genetic causes of common polygenic diseases such as Alzheimer's disease and chronic pancreatitis. Exome data analysis presents several challenges. These challenges include the large number of SNVs compared to the relatively small sample size, the rarity of many of the SNVs, and potential interactions among SNVs on their effect on disease.

In this work, I develop, implement, and evaluate a new multivariate biomarker ranking algorithm called Bayesian averaged probabilistic rules (BAPR) that has several novel characteristics. It (1) learns probabilistic rule models from data, (2) performs Bayesian model averaging to rank biomarkers like SNVs, and (3) incorporates biological knowledge as structure priors of biomarkers. The BAPR algorithm was evaluated on several exome datasets with both synthetic outcomes and real outcomes, and using a range of variant deleteriousness scores as structure priors. The quality of SNV rankings was evaluated with biomarker recovery plots, area under the Receiver Operating Characteristic curves, and evidence of biological validity as supported by the literature.

The BAPR algorithm performed statistically significantly better in identifying previously known disease-associated SNVs and biologically meaningful SNVs when compared to chi-square and random forests. BAPR with uniform and expected number of predictors priors performed better than priors that were derived from variant deleteriousness scores. Also, combining several variant deleteriousness scores performed at least as well as the best per-

forming single deleteriousness score. The variant deleteriousness scores have sparse coverage and typically scores are available only for a small proportion of SNVs that are measured in an exome dataset. The encouraging results obtained with these scores suggests that as coverage of the scores increases the performance of algorithms like BAPR that incorporate them will also improve.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

PREFACE

I would like to thank:

Dr. M. Ilyas Kamboh, University of Pittsburgh, Department of Human Genetics, for his work on "Deep resequencing of candidate gene regions in late-onset Alzheimer's disease" (R01 AG041718) that provided whole exome sequencing data for LOAD critical to this dissertation.

Dr. David C. Whitcomb, University of Pittsburgh Medical Center, Department of Gastroenterology, for his work in the "Consortium for the study of pancreatitis: Pittsburgh Clinical Center" (U01 DK108306) that provided GWAS sequencing data for chronic pancreatitis and recurrent acute pancreatitis used in this dissertation.

Dr. Charles K. Smith, Case Western Reserve University, School of Medicine, for supporting me in my pursuit of a medical doctorate concurrently with this graduate work.

My parents and family for their love and support through a quarter-century of education.

1.0 INTRODUCTION

The human genome project sequenced the first human genome a decade and half ago for about a billion dollars. With the advances in genomics today, the human genome can now be sequenced for less than $10,000. This sequencing ability has generated vast amounts of sequence data for research. The goal of the analyses of most sequence data is to identify sequence variants that will help elucidate the genetic causes of rare (monogenic or Mendelian) diseases such cystic fibrosis and common (polygenic or complex) diseases such as Alzheimer's disease and chronic pancreatitis.

The majority of sequence variants that play a role in common diseases are yet to be determined. Discovering these variants will help us to understand the genetic basis for common diseases, and lead to the development of improved methods for disease prediction, prevention, diagnosis, and treatment.

In this dissertation, I develop and evaluate a new Bayesian machine learning algorithm called Bayesian averaged probabilistic rules to identify variants likely to be associated with common diseases in high-dimensional sequence data. This method has the ability to combine biological knowledge of variants with variant data in a coherent fashion using a probabilistic framework. In particular, I apply and evaluate the algorithm on variant data obtained from the exome (the protein-coding regions of the genome).

1.1 SEQUENCE VARIANT DATA

First-generation sequencing technologies measure millions of *single nucleotide polymorphisms* (SNPs) across the genome with high-throughput genotyping technologies like SNP microar-

ray chips. This led to *genome-wide association studies* (GWASs) that can identify SNPs that are highly associated with a disease of interest. The GWAS approach is based on the *common disease common variant* (CDCV) hypothesis, which posits that SNPs - mutations that occur in at least 5% of the population and also known as common variants - are responsible for the genetic variability in many common diseases.

Second-generation sequencing technologies sequence protein coding regions (exons) in the genome and generate whole-exome data. Compared to SNP microarray chips, exome sequencing captures all single nucleotide variants (SNVs), including both common and rare variants in the coding regions of the genome.

The eventual arrival of cost-effective third-generation sequencing technologies will lead to whole-genome sequencing of all 3 billion base pairs in the human DNA and the characterization of all common and rare variants across the entire genome.

Many current genomic studies are focused on generating whole exome data, and I focus on this type of sequence data in this dissertation. Until now, GWASs have been somewhat successful in identifying common SNVs associated with common diseases. The current belief is that many rare SNVs (unmeasured by first-generation sequencing technologies) are associated with common diseases, especially those SNVs that are located in the coding regions. Thus, analyses of exome variant data will likely lead to the discovery of rare as well as common SNVs that underlie common diseases.

Exome data analysis for identifying both rare and common disease-associated SNVs presents several challenges. These challenges include the large number of SNVs compared to the relatively small sample size, the rarity of many of the SNVs, and potential interactions among SNVs on their effect on disease. New machine learning methods that are computationally efficient are needed to identify SNVs (and, more generally, biomarkers other than SNVs) in exome data that are associated with disease.

1.2 OVERVIEW OF BAYESIAN AVERAGED PROBABILISTIC RULES ALGORITHM

The Bayesian averaged probabilistic rules (BAPR) algorithm is a Bayesian machine learning algorithm that learns a restricted class of Bayesian networks that represent probabilistic rules from data. This algorithm (1) learns probabilistic rules in a Bayesian fashion from exome data, (2) performs *Bayesian model averaging* over probabilistic rules to rank biomarkers like SNVs, and (3) combines biological knowledge as prior probabilities with data using a Bayesian framework. Given a dataset containing SNV measurements in a set of cases (e.g., Alzheimer's disease) and controls (e.g., healthy), BAPR outputs a list of SNVs that are ranked according to their ability in discriminating cases from controls. BAPR can be applied either to data alone or applied to data with the addition of biological knowledge related to SNVs.

1.3 HYPOTHESIS AND SPECIFIC AIMS

My main hypothesis is that the proposed Bayesian machine learning algorithm, BAPR, that learns **data-driven probabilistic rules** and incorporates **biological knowledge** will yield better biomarker discovery than existing methods. The specific hypothesis that I tested is that BAPR will have better performance than existing methods of biomarker discovery in exome data.

To evaluate the hypothesis, I developed and implemented the following specific aims:

Aim 1. Develop and implement a new Bayesian algorithm, BAPR, for ranking biomarkers, such as SNVs, in high-dimensional exome data.

Aim 2. Evaluate BAPR's ability to rank biomarkers using real exome data with synthetic outcomes and real exome data with actual disease outcomes. Evaluate BAPR with and without the addition of biological knowledge and compare its performance to existing methods used in exome data analysis.

1.4 CONTRIBUTIONS

The main goal of this work is to address the problem of ranking biomarkers associated with common diseases using exome data. Deterministic methods have been developed to identify relevant SNVs in exome data in Mendelian diseases that are based on filtering SNVs in stages based on biological knowledge. In Mendelian diseases, SNVs are highly associated with disease; consequently, deterministic methods based on SNV filtering have been successful. However, in common diseases, SNVs associated with disease are likely to be less strongly associated with disease; hence, probabilistic methods are better suited than simple rule-based methods.

The BAPR algorithm makes contributions both to machine learning and to genomic analysis. From a machine learning perspective, BAPR is a novel, multivariate feature ranking method that effectively combines biological knowledge with data using a Bayesian approach. From a genomic standpoint, BAPR can be applied effectively to genomic data to discover new variants that are associated with disease. Moreover, since BAPR is multivariate, it can be applied to discover not only variants with main effects, but also interacting variants. In this dissertation, the BAPR algorithm was extensively evaluated using real exome data with synthetic and real outcomes.

1.5 OVERVIEW OF DISSERTATION

Chapter 2 provides relevant background on biomarker discovery in genomic data and briefly describes related work in genomic analysis. Chapter 3 describes the BAPR algorithm in detail, including some variations. Chapter 4 describes the experimental methods, including the datasets used in the experiments and performance metrics used to evaluate the algorithms. Chapter 5 provides the results of the biomarker ranking experiments. Chapter 6 summarizes the results and conclusions drawn from this work.

2.0 BACKGROUND

This chapter provides relevant background on types of sequence variants, genetic foundations of common diseases, brief descriptions of the main sequencing technologies, SNVs and their relationship to diseases, and past work on discovery of biomarkers in high-dimensional sequence data.

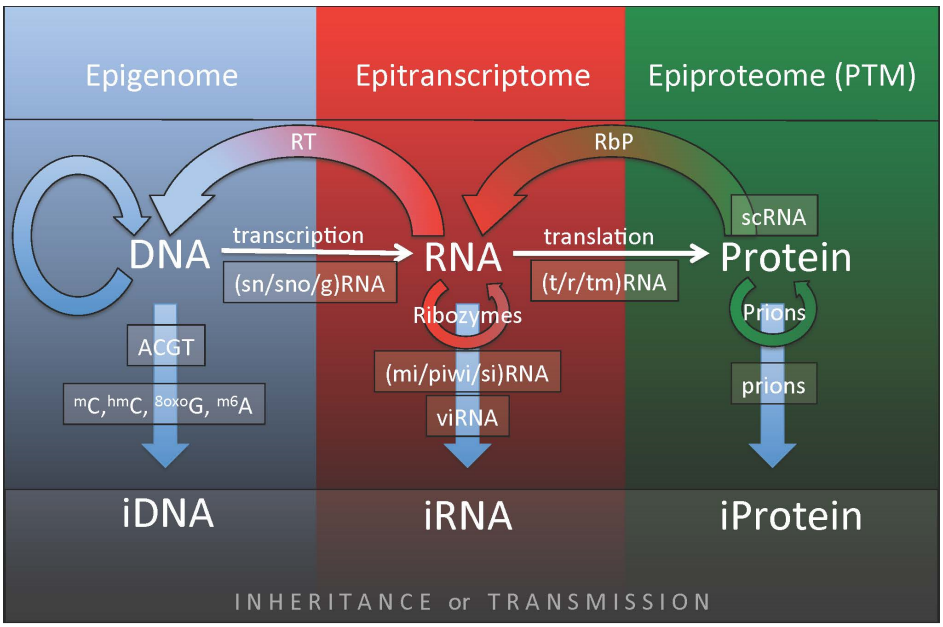


Figure 1: Central dogma of biology.

According to classical view of the central dogma of biology, genetic information in the genes in the DNA is transcribed into individual transportable sequences composed of messenger RNA (mRNA) and each mRNA is translated into a protein or a small number of proteins [80, 83]. The regions of the DNA that are transcribed into mRNA are known as *coding regions* or *exons*, and DNA regions that are not transcribed are known as *noncoding*

regions or *introns*. The entire DNA sequence is referred to as the *genome* and the entire set of coding regions is referred to as the *exome*.

2.1 SEQUENCE VARIANTS

The commonest sequence variation is the *single nucleotide variant* (SNV) which is a DNA sequence variation occurring 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.

A typical SNV has only two variant nucleotides that are called *alleles*: the *minor allele* is the less common nucleotide and the *major allele* is the more common nucleotide in the population. For example, rs13387042 is a SNV that is located on human chromosome 2 and has two alleles: A and G, where A is the minor allele. Humans are diploid organisms and chromosomes come in pairs. Hence, the *genotype* of an individual at a locus may be *homozygous* (having the same allele on both chromosomes) or *heterozygous* (having different alleles on the two chromosomes). For example, the genotype for rs13387042 can take the values AA, AG, and GG. In sequence data, the three genotypes at a SNV are often coded as 0, 1, and 2, which represent 0, 1, or 2 copies of the minor allele, respectively.

SNVs are often categorized as common SNVs and rare SNVs. For a given SNV, when its *minor allele frequency* (MAF) is 5% or higher in the population, it is considered to be a common SNV [106]. Common SNVs are also known as *single nucleotide polymorphisms* (SNPs). Otherwise, when the MAF is less than 5%, the SNV is considered to be a rare SNV.

Many of the known SNVs occur in the DNA between the coding regions (in the intronic, inter-genic regions) and do not directly impact either the protein function or the biological pathways. In contrast, when SNVs occur in exons, they may cause amino acid substitutions in the resultant protein. SNVs that do not change the resultant protein are called *synonymous* (s); SNVs that change the resultant protein are called *nonsynonymous* (ns). Many believe that over 50% of SNVs involved in inherited human diseases are nonsynonymous [50].

Single nucleotide replacements are not the only sequence variants; several other kinds of variants exist. Some of the more frequently encountered variants include indels and copy

number variations. *Indels* refer to a specific kind of mutation known as *insertions and deletions*. These require that the DNA be changed with the insertion or deletion of one or more base pairs. If these indels do not occur in sets of three, the result is a frameshift mutation, where all amino acid encodings following the mutation point are changed [94]. This tends to end in nonsense mutations, which result in RNA sequences that encode shorter, non-functional, or abnormal proteins. A specific subclass of indels known as single nucleotide indels are much more prone to causing these frameshift mutations [64, 98]. *Copy number variations* (CNVs) occur in areas of the genome that are highly repeated, with the number of copies varying per individual - Huntington's disease is a well-known example. These CNVs can be both positively correlated [90] and negatively correlated [54] to gene expression levels.

2.2 GENETIC BASIS OF COMPLEX DISEASES

Genetically, diseases are classified as either (1) rare, monogenic, or Mendelian diseases, or (2) common, polygenic, or complex diseases. Although this dichotomy is overly simplified, it is useful in distinguishing the genetic origins of monogenic disorders from polygenic diseases. Monogenic diseases, such as cystic fibrosis and Huntington's disease, are entirely or mostly caused by sequence variations in a single gene. This characteristic makes elucidating the genetic mechanisms simpler. These diseases are not very common in the population; for example, cystic fibrosis occurs 1 in 4,000 patients [49], and Kabuki syndrome occurs 1 in 32,000 births [62]. Due to their monogenicity, these diseases tend to be strongly Mendelian, with easily identifiable inheritance patterns.

In contrast to Mendelian diseases, complex diseases, such as Alzheimer's disease, are common in the population and have complex inheritance patterns [44]. These diseases are responsible for the majority of morbidity, mortality, and health-care costs [46]. If we better understand the causes of common diseases, we can better assess who is at risk, more accurately predict outcomes, design more effective therapies, and develop better prevention strategies. We have thus far been unable to identify all the genetic causes of most common diseases, like hypertension, coronary artery disease, and Alzheimer's disease.

Less is known about the genetic architecture of complex diseases compared to Mendelian diseases, though they are likely to have significant genetic components. Determining the genetic basis of complex diseases is more challenging than Mendelian diseases [3]. Mendelian diseases typically arise due to a defect in a single gene. They often manifest at birth or in childhood, and, since they follow the Mendelian laws of inheritance, it is easy to predict the risk in a family. Moreover, the causative genotype is highly predictive of the disease and often has a clear biological explanation.

Early attempts to elucidate this genetic structure in common diseases include a Wellcome Trust study using SNP chips that examined 17,000 individuals for seven major diseases [105]. This study found that relatively few SNPs contributed to the susceptibility of disease. Several reasons have been put forward for the discovery of fewer than expected genetic variants in common diseases. First, in common diseases several genes may be acting in concert, possibly in complex ways that defy traditional analysis techniques. Second, genetic heterogeneity is common, where different disease mechanisms lead to the same disease. Third, the environment likely plays a significant role in the development of disease.

Several models of heritability in common disease have been proposed in the literature. Four of the most prevalent models include the infinitesimal model, the rare variant model, broad-sense heritability, and the CDCV model [10, 33].

The *infinitesimal* model proposes that common variants are the major source for disease variability. Each change incrementally increases the risk for disease; many changes are needed to have a high risk for disease. This model was initially formulated as part of quantitative genetic theory and is described in [29]. According to this theory, every gene contributes to every trait in effect sizes so small that an intractable number of samples are required to detect them. For example, GWAS meta-analyses for height and body mass index involving several hundred thousand people indicate that more than a few hundred loci can be confirmed and these loci will not necessarily explain more than half of the genetic variance [53, 91]. In short, this model proposes that heritability is not missing, but hidden below the significance threshold [29, 33].

Broad-sense heritability proposes that there are several gene-gene or gene-environment interactions that result in disease. Additive contributions of common variants and large

effects of rare variants are insufficient to explain missing heritability. This model is strongly supported by gene-gene and gene-environment interactions detected in model organisms [59, 60]. Further research documenting parent-of-origin genetic contributions and inheritance of DNA methylation patterns seems to support this model [43, 48, 89]. Under this model, GWASs cannot tackle family-level heterogeneity because they measure average effects across thousands of people [33].

The CDCV model suggests that there are common alleles that, in combination, result in disease. Common variants seem to play a strong role according to the study described in [109], where approximately 40% of variability in height is explained by common SNPs. For example, authors of [39] note that there exist a large number of common SNPs found to be highly associated and probably causal for disease, such as $\Delta F508$ on CFTR or ApoE4 on ApoE. Some perceive that GWAS data analyses often are not consistent with rare variant explanations, suggesting instead that causal common variants may be likely [106]. In further support, despite differences in allele frequencies across different populations, the associations found in GWASs still hold across these different populations. If rare variants were dominant, CDCV proponents propose that these differences in allele distribution in these different populations should result in different associations [86, 103]. These inconsistencies with rare variants suggest that common variants might be more likely to underpin common disease [10, 39, 109, 15, 33].

Finally, the common disease, rare variant (CDRV) model proposes that rare variants underlie most common diseases. The authors of [109] note that it is impossible to tell how many discrepancies are caused by very low frequency variants with a $MAF < 0.001$ [10, 109, 33]. Theoretically, disease should be selected against, which should make disease-causing biomarkers rarer [13, 8, 75]. This theory is further supported by empirical studies that show that deleterious variants that cause amino acid substitutions are rare in the population [37, 16, 51, 114].

Of these models, the CDCV and CDRV are the two leading models. Gorlov et al (2008) graph of SNVs based on the SeattleSNPs dataset by MAF and disease causing potential seems to support a CDRV hypothesis. The graph shown in Figure 2 shows the proportion of functional or nonsynonymous SNVs plotted against MAF. It clearly demonstrates a higher

proportion of functionally deleterious SNVs at lower MAFs [35]. This result plays a strong role in our belief in emphasizing rare SNVs with prior knowledge in exome analyses.

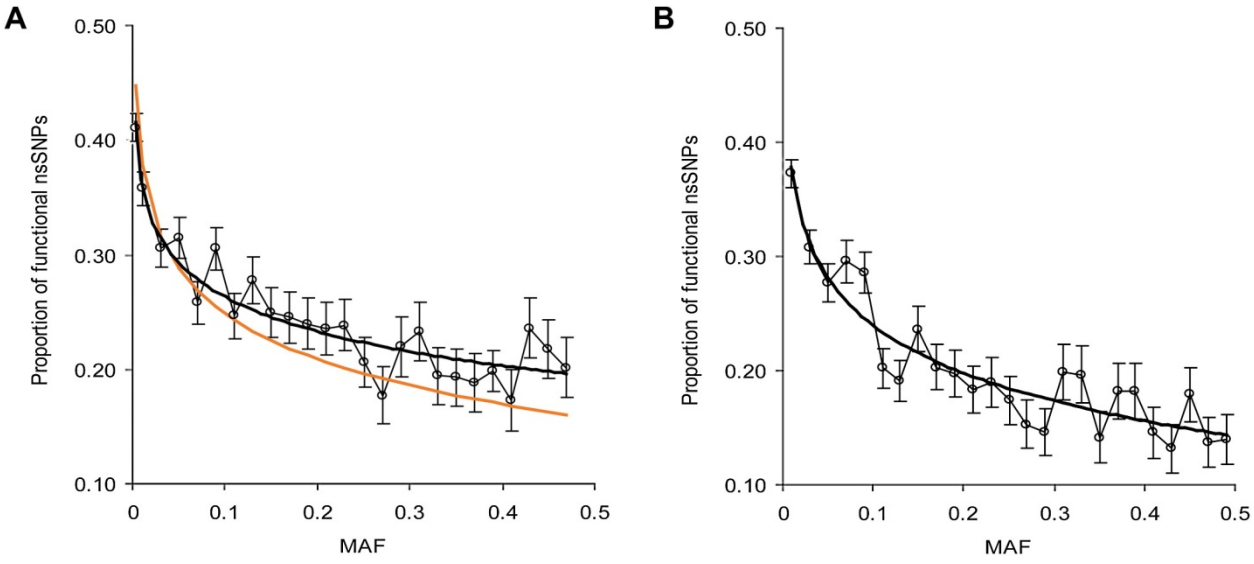


Figure 2: Graph A demonstrates the proportion of nonsynonymous SNVs predicted to be protein damaging plotted against MAF. The dark solid lines are the logarithmic regression curves. The orange line is the regression curve adjusted for PolyPhen's sensitivity and specificity. Graph B shows the proportion predicted by SIFT.

2.3 SEQUENCING TECHNOLOGIES

The human genome consists of about 3 billion nucleotide base-pairs that are packaged into 22 autosomal chromosomes and a pair of sex chromosomes. Only a small portion of the genome consists of genes, where a gene is defined as a sequence of DNA (the exon) that is converted into mRNA and, subsequently, into a protein. About 1.5% of the genome consists of genes and this 1.5% is called the exome or the expressed part of the genome. The remaining 98.5% of the genome does not code for proteins and comprises the inter-genic regions of the DNA. Sequence variants including rare and common SNVs are present in all parts of the genome

including the genic and the inter-genic regions.

The past two decades have seen the development of a succession of sequencing technologies characterized by increasing sophistication, ever increasing coverage of the human genome, and decreasing cost [63]. This section briefly describes measurement of common SNVs in GWASs, whole-exome sequencing and whole-genome sequencing.

The genotyping chip used in a GWAS measures only common SNVs because they rely on the CVCD hypothesis. Previously, the older chips typically measured several hundred thousand common SNVs across the genome, while the currently available chips measure about a million common SNVs. To date, GWASs have discovered over 4,500 SNPs that are associated with a variety of human traits and common diseases [93, 69]. However, the SNVs discovered by GWASs explain only a small amount of the anticipated genetic effect for most diseases. For example, more than 50 SNPs have been discovered to be associated with heart disease, but because of their small effects, they explain only a fraction of the genetic heritability in heart disease [28].

Sequencing of all exons is an intermediate step along the path from the genotyping chip to whole genome sequencing. Whole-exome sequencing is a new and powerful technique in which next-generation sequencing technology makes it feasible to sequence the entire protein-coding region of the genome [95]. With decreasing cost and increasing reliability of next-generation sequencing, exome studies are replacing GWASs quite rapidly. Since exome sequencing captures both common and rare SNVs, it provides richer data than does the GWAS approach. However, exome sequencing provides no data on the intergenic regions of the genome. The promise of exome sequencing is that it will uncover the genetic causes of common diseases that include both common and rare SNVs. Thus, both CDCV and CDRV hypotheses are supported by whole-exome sequencing.

Sarah Ng, along with her coworkers, demonstrated the power of exome sequencing in 2010 and 2011. With a sample of just four patients, her team identified DHODH as a causal gene in Miller syndrome, a recessive Mendelian disease [67]. A year later, her team identified a causal gene, MLL2, in Kabuki syndrome from a sample of ten patients [67]. Several other similar studies have shown that whole-exome sequencing can be successfully used to identify causative SNVs in Mendelian diseases.

However, current whole-exome datasets in common diseases that consist of a few hundred individuals and several hundred thousand SNVs pose several challenges. The small number of individuals compared to a large number of SNVs presents significant statistical and computational challenges. While univariate analyses may be adequate for identifying common SNVs associated with disease, they are unlikely to have the power to detect rare SNVs, which may be more important in terms of functional relevance. Beyond the number of SNVs and the rarity of some of them, the SNVs may interact in non-linear ways to influence the disease. It is therefore important to develop efficient methods to learn multivariate, SNV-disease relationships from whole-exome data.

The limitation of exome sequencing is that there are no measurements from the noncoding regions of the DNA which forms the bulk of the human genome. The recent ENCODE (Encyclopedia of DNA Elements) project is focused on examining the non-coding regions of the DNA for functional relevance. The ENCODE project is revealing that the majority of the non-coding regions which till recently were thought to be non-functional have actually a variety of functions. The ENCODE project has assigned some sort of function to roughly 80% of the genome, including more than 70,000 promoter regions - the sites just upstream of genes where proteins bind to control gene expression - and nearly 400,000 enhancer regions that regulate expression of distant genes [20, 21].

In the near future, we are likely to transition from whole-exome sequencing to whole genome sequencing as the method of choice for genetic research studies. The ongoing 1000 Genomes Project is aimed at whole-genome sequencing about 2,500 samples [72, 70]. Whole genome sequence data will allow identification of both rare and common SNVs from both the coding and non-coding regions of the genome. Analyses of whole-genome data will likely have the same challenges as the analyses of whole-exome data as described earlier including high dimensionality and relatively low sample sizes. And, methods that will be developed for exome data analyses will likely be useful for whole-genome analysis.

Table 1: Description of variant deleteriousness scores.

Score	Score range	Approach	Reference
SIFT	0 - 1	protein structure and function	[52]
PolyPhen-2	0 - 1	protein structure and function	[1]
PhyloP	0 - 1	conservation	[88]
GERP	-12.3 - 6.17	conservation	[23]
Phast	0 - 1	conservation	[41]
CADD	0 - ∞	combination of protein structure and function and conservation	[47]

2.4 VARIANT DELETERIOUSNESS SCORES

This section describes types of knowledge that is available at the level of the single nucleotide that can be used in sequence data analysis.

A range of methods have been developed to score single nucleotide loci in terms how likely a variant at a locus is to be associated with disease. Two main approaches are in use. In one approach, the score of a variant is based on predicting the effect of the nucleotide change on protein structure and function. Thus, variants that have a greater predicted effect on protein structure and function are considered to be more likely to cause disease. In another approach, the score of a variant is based on the degree to which a locus in the genome is evolutionarily conserved. Thus, variants that occur at more conserved loci are considered to be more likely to cause disease. The first approach can provide scores only for variants in the exome, while the second approach can provide scores for variants in the entire genome. Examples of scores that use the first approach include SIFT and Polyphen-2 while examples of the second approach include PhyloP. These scores are described below.

2.4.1 *S*orting *I*ntolerant *f*rom *T*olerant (SIFT)

When SNVs occur in exons, they may cause amino acid substitutions in the resultant protein. SNVs that do not change the resultant protein are called synonymous; SNVs that change the resultant protein are called nonsynonymous. Over 50% of SNVs involved in inherited human diseases are nonsynonymous [50]. The approximately 122,000 nsSNPs in dbSNP indicate that there is a strong need to characterize nonsynonymous variants with respect to their effects on protein function. The Sorting Intolerant From Tolerant (SIFT) algorithm is a multi-step, sequence homology-based approach for classifying amino acid substitutions. This technique focuses on evolutionary conservation of amino acids within protein families. Sequence conservation is inversely proportional to tolerance to mutations [66, 52].

Given a protein sequence, SIFT identifies related proteins with PSI-BLAST. Next, it builds an alignment from homologous sequences. At each amino acid position, SIFT analyzes the probabilities for each of the 20 amino acids. These probabilities are then normalized against the probability of the most frequent amino acid and stored in a matrix. The scaled probability also known as the SIFT score predicts protein function if it falls below a threshold [30].

SIFT also generates a confidence score based upon the diversity of reference proteins; the more diverse a reference set, the more confident SIFTs predictions are. SIFT does not use protein structure to assess amino acid substitution effects. However, evaluations of this additional information appear to not significantly improve prediction accuracy [84, 52].

2.4.2 *Poly*morphism *Phen*otyping (PolyPhen)

PolyPhen and PolyPhen-2 are methods to predict the functional effects of sequence variants. Unlike SIFT, which relies solely on sequence homology, PolyPhen attempts to understand the change in the actual amino acid structure by comparing annotated UniProt entries. For example, changes in functional sites or between hydrophilic and hydrophobic regions in a protein can alter the folding of the protein and its function. An amino acid substitution at a location may be tolerated if it has a similar charge, size, or hydrophilicity as the original amino acid. For example, glyceine can be substituted with alanine with minor effects on the

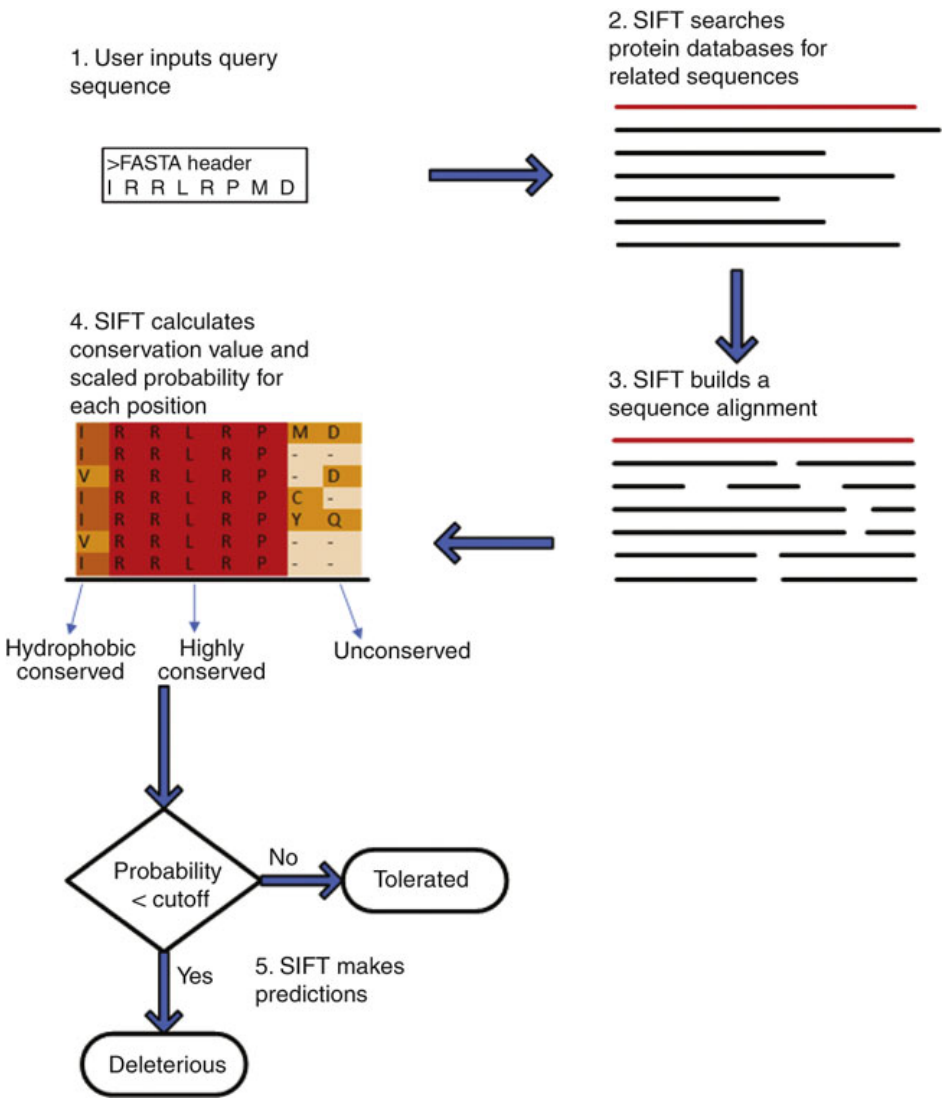


Figure 3: SIFT methology.

protein; however, when substituted with arginine- an amino acid that is quite different in terms of charge, size, and hydrophilicity- the effects on the protein are much larger [1, 79, 30].

2.4.3 *Phylo*genetic *P*-values (PhyloP)

Many geneticists believe that our current genetic structure has been strongly shaped by evolution into its current state. As a result, genetic regions that are critical to life are highly conserved. PhyloP therefore computes the conservation or acceleration p-values based on alignment and a model of neutral evolution. Higher PhyloP scores are more highly conserved [58, 24].

2.4.4 *G*enome *e*volutionary *r*ate *p*rofiling (GERP)

The GERP algorithm finds constrained elements by measuring substitution deficits under the premise that areas with fewer substitutions are functionally conserved. The number of rejected substitutions in an area measures the strength of the constraint [25].

2.4.5 *Ph*ylogenetic *a*nalysis with *S*pace-*T*ime Models (Phast)

In contrast to comparative evolution models that assume uniform selection pressures, Phast detects sequences under selection or drift on any lineage using a hidden Markov model. It does not require pre-assessed element boundaries and computes a p-value based on prior and posterior distributions on the number of substitutions in the evolution of predicted elements [87, 41].

2.4.6 *C*ombined *A*nnotation *D*ependent *D*epletion (CADD)

The CADD approach integrates diverse annotations of variants into a single metric called the C score. The authors annotated a dataset of about 15 million real variants with 63 types of annotations including conservation metrics, regulatory and transcript information, and protein-level scores, and created another dataset of 15 million simulated variants. They

trained a support vector machine on the 30 million variants to discriminate between the real and simulated variants [47].

The CADD scores come in two forms, namely "raw" and "scaled". Raw CADD are the output of the support vector machine model. The scaled scores are normalized raw scores such that reference genome single nucleotide variants at the 10th-% of CADD scores are assigned to CADD-10, the top 1% to CADD-20, the top 0.1% to CADD-30, etc.

2.5 BIOMARKER DISCOVERY FROM SEQUENCE DATA

This section also describes some of the methods used to analyze sequence data to identify variants that are associated with disease.

Typically, variant analysis of GWAS data or exome data uses a univariate test of association to test the association of each SNV with the disease or phenotype. A commonly used statistical test is chi-square. The genotypes of a SNV on a set of cases and controls can be summarized in a 2 x 3 contingency table of the genotype counts for each group. For example, for a SNV with alleles G and T, we tabulate the number of cases and controls with each genotype GG, GT, and TT. Pearson's chi-square test is used to assess departure from the null hypothesis that case and controls have the same the distribution of genotype counts. This test as presented uses no genetic information. Further chi-square tests can be used to test specific genetic hypotheses for example, that the SNV alleles increase disease risk under a dominant or a recessive model. Assuming T is a high-risk minor allele, these tests compare GG genotypes to GT + TT genotypes (dominant model), or GG + GT to TT genotypes (recessive model).

Chi-square analysis at the gene level can be applied in a similar way to identify genes associated with disease. Here, the SNV data for each gene is abstracted to represent a gene. One simple method that has been used is to create a binary genetic variable that takes the value 1 in a subject if there are one or more high-risk minor alleles present at any SNV in the gene; otherwise, the variable takes the value 0.

Biomarker discovery analyses can leverage known biological knowledge to improve the

analysis. Although an intuitively sensible notion, combining this prior information with the statistical evidence may be a difficult task. For example, should a SNV with a more extreme p-value in an intergenic region be ranked above or below a less extreme p-valued SNV near or in a candidate gene [96]?

The authors of [82] analyzed the GAW17 data with a waited-sum pooling method for a gene-level association test, showed that multiple-trait analysis was on par with single trait analysis. The authors of [112] used multiple strategies - including univariate collapsing, combined multivariate and collapsing, and single-marker methods - to analyze the same data and found it difficult to control for type 1 errors. The authors of [55] used a different perspective by combining a burden test with a non-burden sequence kernel association test to blend the prediction abilities of two orthogonal algorithms that emphasized different qualities on NHLBI Exome Sequencing Project data.

2.5.1 Filter methods

The objective of filter methods [42] is to filter out irrelevant information to reduce noise, integrate pertinent information to increase effect size, and choose tests that find promising biomarkers while still maintaining a certain false-discovery rate [42, 57, 93].

Filters generally are all-or-nothing, fully incorporating or excluding a variable based upon certain criteria. Likewise, integrators are also binary operators, unilaterally combining certain variables. An example of a flowchart showing how filters are applied to exome data analysis is shown in Figure 4.

2.5.2 Univariate methods

Single locus techniques, utilizing univariate methods, are often a first iteration analysis to identify the effects a single locus - or point mutation - can have on a final result.

The chi-square test is a commonly used univariate statistic that has a probabilistic interpretation that creates a statistic that can be compared to the chi-square distribution to determine the probability of association between the SNV's genotypes and the target variable.

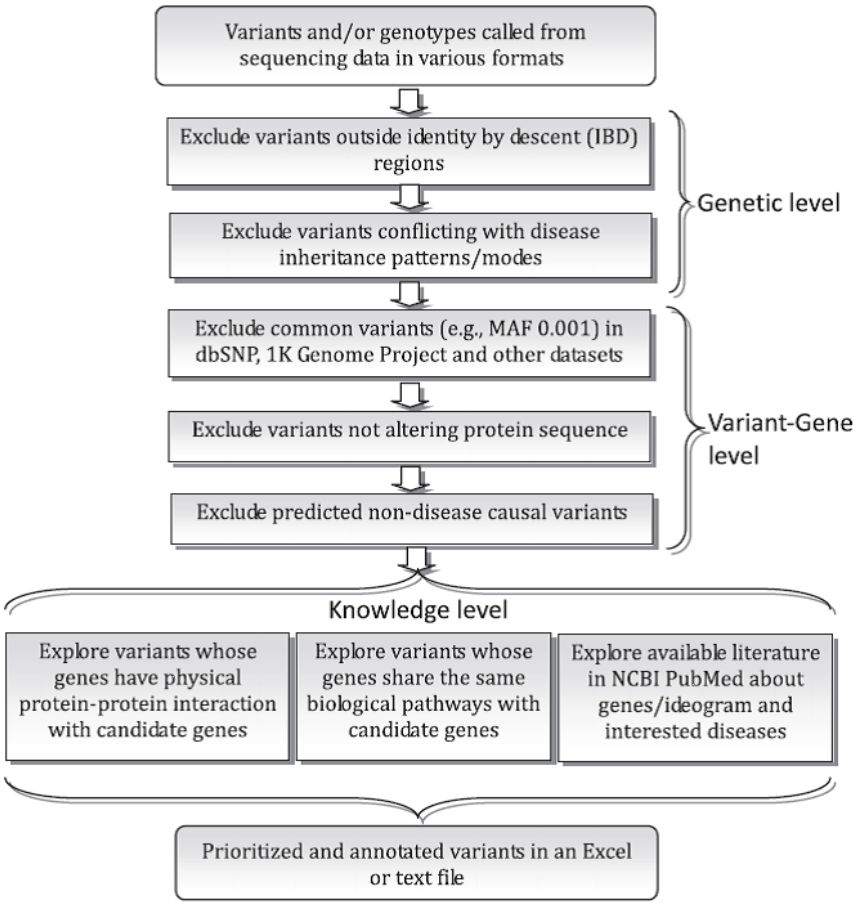


Figure 4: Filter method example.

Logistic regression measures a relationship between a categorical dependent variable and one or more independent predictor variables by estimating probabilities with a logistic function. More precisely, it can map a binary outcome variable to a continuous predictor variable that can take any real value.

2.5.3 Multivariate methods

Univariate techniques aren't able to characterize the interaction between loci, also known as epistasis. Epistasis affects how genes can be expressed. A classical example are two genes - one for hair color (e.g., black or brown hair) and another for baldness. Even if a person had a gene for brown hair, a gene for baldness would hide the value. Multivariate techniques are necessary to gain further insight into these interactions.

Multifactorial dimensionality reduction [36] partitions the data into multivariate cells in N-dimensional space. Cells deemed 'high risk' - identified by a predetermined threshold - are grouped together; the rest of the cells, deemed 'low risk', are also grouped together. Cross-validation is then used to validate the generated model. MDR requires a prespecified dimensionality of the final model.

Random forests (RF) are a classification and regression method that rely on an ensemble of decision trees that grew in prominence. The original RF method uses standard classification and regression trees (CART) derived from a random subset of the data and uses the decrease of Gini impurity to select variables on which to split the training data. This has been successfully applied to GWAS, such as in multiple sclerosis [11, 34].

2.5.3.1 Bayesian approaches Bayesian methods applied to high throughput data and biomarker discovery especially exome data are much less common. In the GAW17 workshop, authors of [74, 108] and [76] used Bayesian analytic techniques to detect rare variants as designed by the GAW17 semisynthetic exome committee.

Instead of relying on predictor variables as parents, the naive Bayes classifier uses predictor variables as children; in this arrangement, it can be learned with a smaller subset of data.

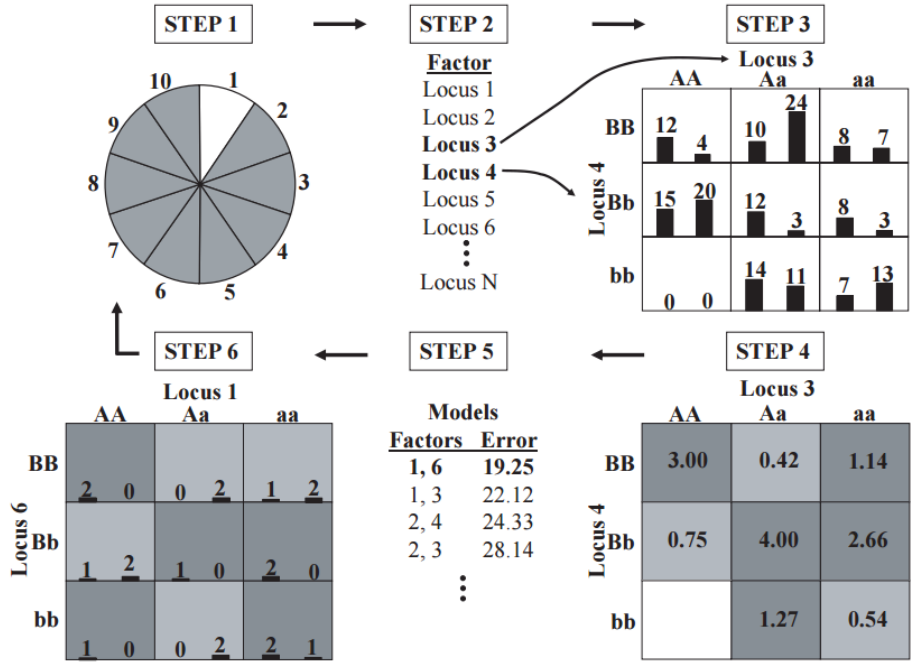


Figure 5: Multifactor dimensionality reduction (MDR) example.

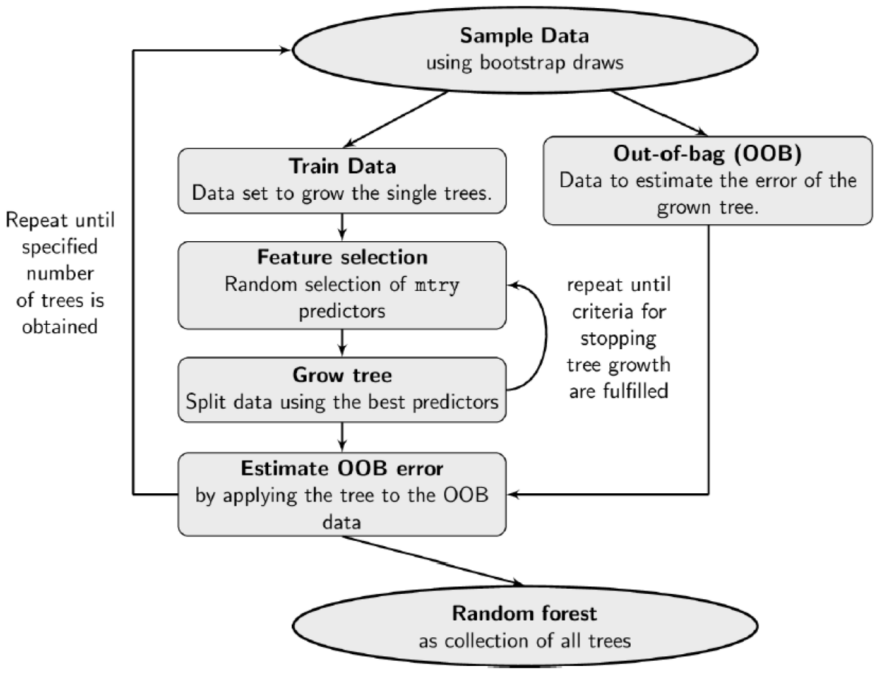


Figure 6: Random forest example.

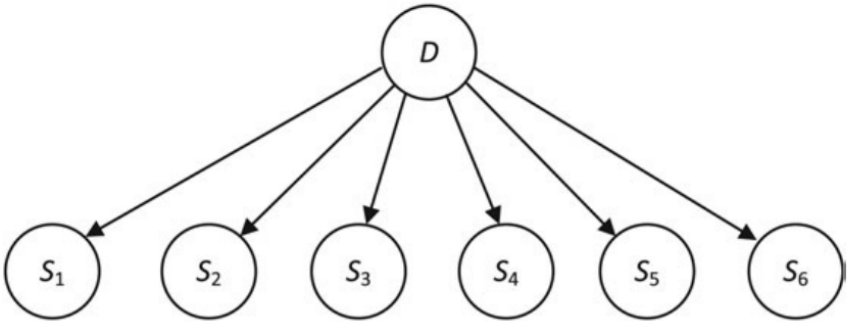


Figure 7: Naive Bayes example.

Bayesian model averaging (BMA) methods are rarer yet. The authors of [110] developed an iterative BMA technique that ranks genes based upon microarray data and found that their technique selects smaller numbers of relevant genes with high prediction accuracy. The authors of [4] apply their approach to diffuse large B-cell lymphoma microarray data, again with similar success. The authors of [45] apply a Bayesian averaging approach to combine SNV level scores to generate a gene score.

Fewer yet incorporate prior knowledge into their modeling for genomic data. For example, the authors of [78] utilized a Bayesian risk index that integrates external biological variant covariates and shows increased performance at pathway, region, and variant level inference for breast cancer (BRCA1) data.

2.5.4 Biomarker discovery in exome data in monogenic diseases

The goal of exome sequencing is to identify functional sequence variants including SNVs in the exome that underlie both Mendelian and complex diseases. Most of the success in exome sequencing has come in Mendelian diseases. Whole-exome sequencing has been successfully used to discover of disease causing variants in rare monogenic diseases. Ng et al. sequenced the exome of four unrelated individuals with Freeman Sheldon syndrome (a rare inherited disorder) and eight healthy HapMap individuals. They were able to correctly identify the gene previously known to cause the syndrome [68].

In a subsequent study, Ng et al. sequenced the exomes of four individuals with Miller syndrome (a rare malformation disorder) and identified a new casual mutation [67]. They used a stepwise filtering approach to screen the identified variants in order to select those likely to be implicated in the disorder. They then compared the four exomes to those of eight control individuals and to the dbSNP database [27] to exclude common variants. Finally, they used PolyPhen [79, 1] to exclude variants that are not predicted to be damaging. The four individuals with Miller syndrome were found to have six rare variants in the *DHODH* gene.

Choi et al. used exome sequencing to make an unanticipated genetic diagnosis of congenital chloride diarrhea in a patient referred with a suspected diagnosis of Bartter syndrome,

a renal salt-wasting disease [19]. The molecular diagnosis was based on the finding of a homozygous missense mutation and was confirmed by clinical follow-up.

In monogenic diseases, exome sequencing of a small number of affected individuals can greatly reduce the number of candidate genes and may even identify the responsible gene specifically [9, 67, 6]. In complex diseases, rare variants are predicted to have stronger effect sizes and straightforward functional significance than common variants. However, the analysis of the rare variants is more challenging than common variants. Even with large sample sizes, the power to detect an association with a single rare variant is low. Several strategies have been developed to address this challenge. For example, one strategy is to assess the collective effects of rare variants across a gene or across multiple genes. A second strategy is to incorporate prior evidence about variants (e.g., functional class of SNV), genes and pathways. Additional strategies include enhancing the statistical power of analysis using quantitative rather than dichotomized phenotypes [56, 7].

2.5.5 Biological knowledge

Incorporating external biological information offers new possibilities for analysis in conjunction with the typical data-driven approach [92].

As Ng et al noted, it was statistically difficult to identify candidate genes. The primary problem in pure data driven approaches is that the signal-to-noise ratio is extremely low. Some portray the central dogma as a layman's first encounter with a radio. On one level, a radio is a black box that takes an electromagnetic signal and, through some feat of magic, converts that signal into music and sound. Likewise, biology converts the genetic code in our cells into the phenotypes we see [68, 67, 42].

There are two main methods to tackling this complex signal. We know some things about complexity, which can become prior assumptions. For example, we can take the principles of modularity, hierarchical organization, evolution, and inheritance to guide how we present our data to our models. Second, many complementary layers exist genomes, transcriptomes, metabolomes, and more [42].

Increasing signal to noise ratios, especially in single -omics scenarios where we look at

just one source of information, like genomics, transcriptomics, or proteomics, has involved using complementary datasets and prior knowledge in two fashions: filters and integrators. Filters use prior information to reject some information as noise, such as rejecting some signals if they don't meet certain criteria. These methods reduce noise, thus reducing the false discovery rate. On the other hand, integrators increase the strength of signals by aggregating individual occurrences into larger units or integrating different types of information. For example, SNVs in a gene or genes in a pathway might be aggregated into similar clusters [42].

Furthermore, each effect is modular; for each problem, various filters and integrators can be combined in different ways to achieve tailored responses.

GWASs, in their attempt to identify SNVs that cause a phenotype of interest, have only been able to explain a small percentage of variation. One possible explanation, proposed by Wang and Yang [102, 109], is that there are many functionally similar loci that together have a large impact but individually don't reach significance in GWASs [35].

Segr of the Broad Institute [85] tested that hypothesis by analyzing gene variation in mitochondria in non-insulin dependent diabetes mellitus, also known as type II diabetes. Their method, named MAGENTA, performs a metaanalysis of various GWASes to achieve larger sample sizes and, thus, have greater statistical power. MAGENTA incorporates both filters and integrations. First, SNVs that are far from genes are filtered out. Subsequently, an integrator transforms SNVs to genes by assigning the highest p-value of a gene's SNVs to a gene. Further corrections integrate prior knowledge by correcting scores for confounding factors like gene size, number of SNVs per kilobase, and genetic linkage. A second integration combines scores of genes by pathway, giving a pathway-based p-value of association [85].

Evaluation of MAGENTA simulations indicates a strong potential to detect disease associations. In detecting a pathway with 10 out of 100 genes, MAGENTA had a five-fold increase in power from 10% to 50% over single SNV detection. Although there were not any mitochondrial pathways highlighted for diabetes, identified pathways for fatty acid metabolism in cholesterol-influencing genes.

Learning networks of transcription factors and understanding how regulation drives complex behaviors is a difficult task; researchers have applied probabilistic networks to try to

understand such interactions. Small sample sizes, however, lead to the inability to distinguish between many equivalently probable possibilities. To solve this problem, Zhu enriched gene expression profiles with known biological information on genotypes. Their work on analyzing the regulatory network of the common yeast *Sacchromyces cerevesiae* first combined mRNA expression profiles with genotypes, then analyzed expression quantitative trait loci (eQTL). eQTLs were used to prioritize genes with greater likelihoods of being associated [113, 96].

3.0 ALGORITHMIC METHODS

This chapter describes the Bayesian averaged probabilistic rules (BAPR) algorithm that learns probabilistic rule models from data. The probabilistic rule models are a class of restricted BNs. This algorithm: **(1)** learns probabilistic rules in a Bayesian fashion from exome data, **(2)** performs Bayesian model averaging over probabilistic rules to rank biomarkers like SNVs, and **(3)** combines biological knowledge with exome data using a Bayesian framework.

First, I provide background on BNs, scoring of BN structures, Bayesian model averaging and probabilistic rules. Then, I describe the BAPR algorithm in detail.

3.1 TERMINOLOGY AND NOTATION

This section introduces some notation and definitions in the context of probabilistic models. A probabilistic model is a family of probability distributions indexed by a set of parameters. More specifically, a *graphical probabilistic model* is a parametric family of probability distributions that satisfy independence relationships that are asserted by an independence graph [65].

Model selection reflects a method that utilizes data to select one model from a set of models under consideration. On the other hand, *model averaging* reflects the process of estimating a quantity under each considered model and then averaging the estimates [99].

Model selection and model averaging can be done with non-Bayesian or Bayesian approaches. Non-Bayesian model selection includes choosing models by maximum likelihood, maximum penalized likelihood, or cross validation. Non-Bayesian *model averaging* includes bagging and boosting, where resampled data allows for multiple model construction [65].

In contrast to non-Bayesian modeling methods, Bayesian model selection chooses the model with the highest posterior probability. In model averaging, the prediction is the weighted average of the predictions by the individual models, with the model posteriors comprising the weights. If optimizing for prediction accuracy, Bayesian model averaging is often preferred. However, this method incurs high computational costs and can impair understanding of a model. Single models are both computationally and conceptually simpler than their model averaged counterparts [65].

Capital letters (X,Z) denote random variables. The corresponding lower case letters (x, z) denote specific values assigned. Bold uppercase letters **(X,Z)** represent sets of variables or random vectors. Their instantiation is indicated by bold lowercase letters **(x,z)**. Thus, $\mathbf{X} = \mathbf{x}$ indicates that variables in $\mathbf{X}$ are assigned the values in $\mathbf{x}$. In addition, $\mathbf{Z}$ denotes the target (or class) variable being predicted, $\mathbf{X}$ denotes the set of features (or predictors), M denotes a model, D denotes the training dataset [65].

A glossary of symbols and brief descriptions is given in Table 2.

Table 2: Glossary of symbols and brief descriptions.

Symbol	**Brief description**
M	model with structure S and parameters Θ
S	structure of M
Θ	parameters of M
θ_{jk}	parameter for the k^{th} value and j^{th} instantiation of the parent nodes
j	instantiation of the parent nodes
k	value of the target node
Z	target node
D	data
$d_{u,v}$	data point for the u^{th} individual at the v^{th} indexed SNV
$\mathbf{U}$	set of all individuals u
$\mathbf{V}$	set of all variations v
$\mathbf{X}$	set of all variables in M

Continuation of Table 2	
Symbol	**Brief description**
$\mathbf{x}$	instantiation of $\mathbf{X}$
X_i	i^{th} variable in $\mathbf{X}$
x_i	instantiation of X_i
$\mathbf{L}$	set of leaves in the decision tree (local structure)
l	index of a leaf in $\mathbf{L}$
r_i	number of states of the random variable X_i
N_{ij}	count of data points in D fitting the j^{th} column for node i
N_{ijk}	count of data points in D fitting the j^{th} column for the k^{th} value for node i
N_{lk}	count of data points in D fitting the l^{th} leaf for the k^{th} value of Z
α_{ij}	prior for the j^{th} column for node i
α_{ijk}	prior for the j^{th} column for the k^{th} value for node i
α_{lk}	prior for the l^{th} leaf for the k^{th} value of Z
R_i	i^{th} ranked result in R
R	ranked results
tp	true positives
tn	true negatives
fp	false positives
fn	false negatives

3.2 BAYESIAN NETWORKS

A *Bayesian network* (BN) is a probabilistic model that combines a network structure (e.g. a graphical representation) with network parameters (e.g. quantitative information) to represent the joint probability distribution over a set of random variables [71, 65].

A BN M that represents the set of variables X consists of two parts: a structure S and parameters Θ. The structure S is a *directed acyclic graph* (DAG) that contains a node for

each variable in $\mathbf{X}$ and an arc between every pair of nodes that are directly probabilistically dependent. An absence of an arc between two nodes implies probabilistic independence between those two nodes and Θ represents the parameters for the graph [71, 65, 99].

The DAG defines the relationship between nodes. For a node X_i, its immediate predecessors in $\mathbf{X}$ are called *parents*. Parents are more remote predecessors area called ancestors. The immediate successors of X_i are called *children*. Children and more remote successors are called descendants. An *undirected path* in the network follows arcs while ignoring the direction of the arcs. An *undirected loop* is an undirected path that starts and ends at the same node while passing through at least one other node and does not cross itself [71, 65, 99].

Figure 8 shows an example graph for S of a BN. Here, *history of smoking* is the parent of *lung cancer* and *chronic bronchitis*, and fatigue is the child of both *lung cancer* and *chronic bronchitis*. Both *lung cancer* and *mass* seen on chest X-ray are descendants of *history of smoking*.

Each node has a local probability distribution given the state of its' parents. The set of all local probability distributions are parametrized by the set of parameters Θ. Table 3 shows a parameterization for the network in Figure 8.

Given the independences and dependencies provided by S, we can factor the complete joint probability distribution over $\mathbf{X}$. Given an extension of Bayes' rule, for $\mathbf{X} = \{X_1, X_2, ..., X_n\}$

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | parents(X_i)) \tag{3.1}$$

we can thus factor the joint probability distribution as follows [71, 65, 99]:

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1) P(X_2|X_1) P(X_3|X_1) P(X_4|X_2, X_3) P(X_5|X_3) \tag{3.2}$$

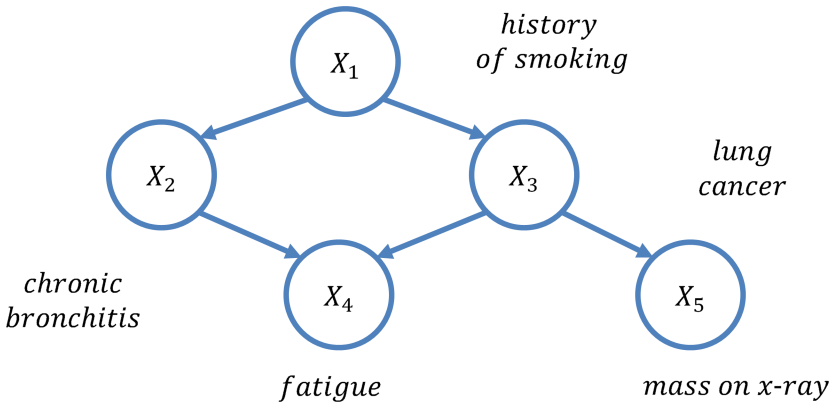


Figure 8: An example BN structure.

3.3 LOCAL STRUCTURE IN BAYESIAN NETWORKS

The DAG of a BN concisely summarizes statements of variable independence. Consider the following example. A variable X is independent of Y given Z if $P(x|y,z) = P(x|z)$ for all possible values x, y, and z for X, Y, Z. The graphical structure makes explicit independence relations in a BN of the form $X \perp Y|Z$, implying that $P(X|Y,Z) = P(X|Z)$ for all values of the variables X, Y, and Z. However, these are not the only independencies that may be present in a domain. For instance, value-specific independence relationships may exist that hold for only particular assignments of values to certain nodes. These relationships cannot be entirely represented by a BN graphical structure. Value-specific independence relationships can be elucidated in the form $X \perp Y|Z = z_1$, implying that $P(X|Y,Z = z_1) = P(X|Z = z_2)$ holds true for all values of X and Y only when Z takes the particular value z_1. For other values of Z, such as z_2, X and Y may not be conditionally independent, that is, $P(X|Y,Z = z_2) \neq P(X|Z = z_2)$. This type of independence is also known as context-specific independence and can be interpreted as X is independent of Y in the context of Z

Table 3: Probabilities associated with the example BN structure in Figure 8.

Node X_1	$P(X_1 = no) = 0.8$	$P(X_1 = yes) = 0.2$
Node X_2	$P(X_2 = no \mid X_1 = no) = 0.95$	$P(X_2 = yes \mid X_1 = no) = 0.05$
	$P(X_2 = no \mid X_1 = yes) = 0.75$	$P(X_2 = yes \mid X_2 = no) = 0.25$
Node X_3	$P(X_3 = no \mid X_1 = no) = 0.995$	$P(X_3 = yes \mid X_1 = no) = 0.005$
	$P(X_3 = no \mid X_1 = yes) = 0.997$	$P(X_3 = yes \mid X_1 = no) = 0.003$
Node X_4	$P(X_4 = no \mid X_1 = no \mid X_3 = no) = 0.95$	$P(X_4 = yes \mid X_1 = no \mid X_3 = no) = 0.05$
	$P(X_4 = no \mid X_1 = no \mid X_3 = yes) = 0.40$	$P(X_4 = yes \mid X_1 = no \mid X_3 = no) = 0.60$
	$P(X_4 = no \mid X_1 = yes \mid X_3 = no) = 0.90$	$P(X_4 = yes \mid X_1 = yes \mid X_3 = no) = 0.10$
	$P(X_4 = no \mid X_1 = no \mid X_3 = yes) = 0.40$	$P(X_4 = yes \mid X_1 = yes \mid X_3 = yes) = 0.60$
Node X_5	$P(X_5 = no \mid X_3 = no) = 0.98$	$P(X_5 = yes \mid X_3 = no) = 0.02$
	$P(X_5 = no \mid X_3 = yes) = 0.30$	$P(X_5 = yes \mid X_3 = no) = 0.70$

taking the value z_1, but not the value z_2. In general, these independent statements imply that in some contexts, defined by an assignment of specific values to the variables in the BN, specific independences hold [12, 99, 100].

I refer to BNs without explicitly represented context-specific structure as *BNs with global structure* and I refer to BNs with explicitly captured context-specific structure as *BNs with local structure*. Conditional probability tables are a typical representation used in BNs with global structure. Associated with a node in a BN is a set of conditional probability distributions (CPDs) that in domains with discrete random variables are typically represented by a table. In this representation, $P(X_i|Pa_i)$ is a table that contains an entry for each joint instantiation of X_i and Pa_i. Each column (or row) in the table represents a single conditional probability distribution, $P(X_i|Pa_i = pa_i)$, corresponding to a particular instantiation of the variables in Pa_i to a set of values given by pa_i. Tabular CPDs are called conditional probability tables (CPTs) and are commonly the representation used in discrete BNs.

There are several possible representations for local structure to capture context-specific independencies. Friedman and Goldszmidt describe a default table representation similar to a CPT, except that it provides both a default CPD for a subset of the parent states and a decision tree representation for the remaining parent states, where a *decision tree* is used to represent the local structure for a BN node X_i. A decision tree is a graph (not a BN graph) where the root node has no parents, and all other nodes have one parent. Nodes that have children and appear in the interior of the tree are called interior nodes, and terminal nodes are called leaf nodes. An example of a BN with a decision tree local structure is shown in Figure 11 [31, 99, 100]. An example of global structure for a target node Z is shown in Figure 10 and an example of a local decision tree structure for the node Z is shown in Figure 11.

Learning a tree structure requires several search operators, including:

(1) *Grow*: Randomly pick a terminal node and split it into two new ones by randomly assigning it a splitting rule.

(2) *Prune*: Randomly pick a parent of two terminal nodes and turn it into a terminal node by collapsing the nodes below it.

(3) *Change*: Randomly pick an internal node, and randomly reassign it a splitting rule.

(4) *Swap*: Randomly pick parent and child pairs that are both internal nodes. Swap their splitting rules.

Each of the above operators can be extended to accommodate more than two child nodes per parent node. In this dissertation, I use only the *grow* operator for simplicity.

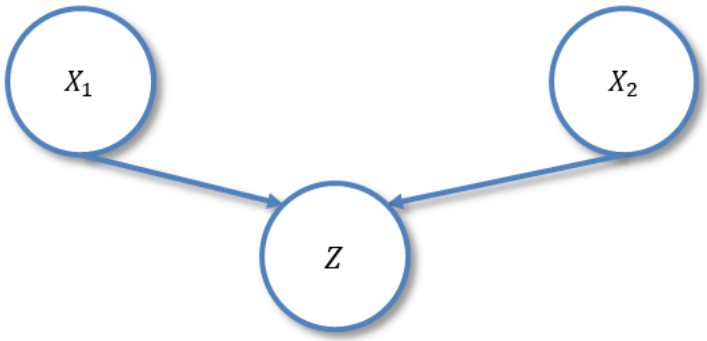
3.4 PROBABILISTIC RULES

Consider a BN that consists of a child node that has one or more parent nodes and there are no arcs among the parent nodes. Such a BN can be interpreted as a *probabilistic rule model*. A probabilistic rule model consists of a set of if-then rules, where the variables represented by the parent nodes form the antecedent and the variable represented by the child node forms the consequence in the rules. Note that the parameters associated with the rule model represents the parameters associated with the child node in the BN. The parameters associated with the child node may be represented by a global structure (by a CPT) or by a local structure (e.g. local decision tree). For example, Figure 11 shows parameters for the global structure of node Z and Figure 12 shows the corresponding parameters for the local decision tree structure of a child node Z. Compared to the global structure models, local structure models are likely to result in more succinct probabilistic rules.

3.5 SCORING BAYESIAN NETWORK STRUCTURES

A typical approach for learning BNs from data is to employ a score-and-search algorithm where a scoring metric (score for short) is adopted to evaluate candidate BN structures while a heuristic search strategy is used to find a structure with the best score. Heuristic search, typically greedy search, is employed because the model space for high dimensionality data is enormous and evaluation of every model is intractable.

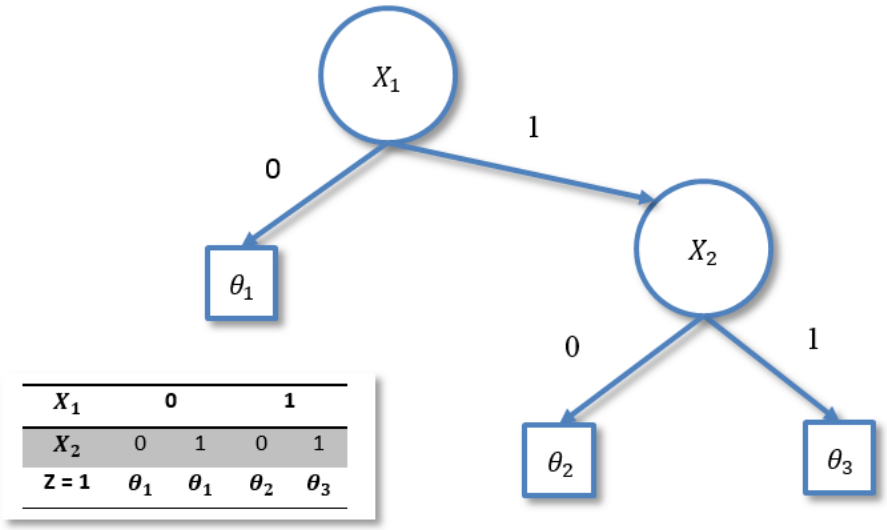
This section describes Bayesian scores for BN structures. I first describe the Bayesian Dirichlet score and then briefly describe several variations on that score.



X_1	0		1	
X_2	0	1	0	1
Z=1	θ_1	θ_2	θ_3	θ_4

$$X_1 = 0 \wedge X_2 = 0 \Longrightarrow Z = 1\ (\theta_1)$$
$$X_1 = 0 \wedge X_2 = 1 \Longrightarrow Z = 1\ (\theta_2)$$
$$X_1 = 1 \wedge X_2 = 0 \Longrightarrow Z = 1\ (\theta_3)$$
$$X_1 = 1 \wedge X_2 = 1 \Longrightarrow Z = 1\ (\theta_4)$$

Figure 9: BN model.



X_1	0		1	
X_2	0	1	0	1
Z = 1	θ_1	θ_1	θ_2	θ_3

$$X_1 = 0 \wedge X_2 = 0 \implies \mathrm{Z} = 1\,(\theta_1)$$
$$X_1 = 0 \wedge X_2 = 1 \implies Z = 1\,(\theta_1)$$
$$X_1 = 1 \wedge X_2 = 0 \implies Z = 1\,(\theta_2)$$
$$X_1 = 1 \wedge X_2 = 1 \implies Z = 1\,(\theta_3)$$

Figure 10: Local tree model.

3.5.1 Bayesian Dirichlet (BD) score

In the Bayesian approach, the model score is based on the posterior probability of the model structure given data. The Bayesian approach treats the structure and parameters as uncertain quantities and incorporates prior distributions for both. Given data D, the Bayesian score of S is a measure of how probable structure S is in light of the D over all possible parameterizations of S. The score of S is defined as a value that is proportional to the posterior probability $P(S|D)$ that is given by Bayes' theorem:

$$P\left(S|D\right) = \frac{P\left(D|S\right)P\left(S\right)}{P\left(D\right)} \tag{3.3}$$

For a fixed dataset D, $P\left(D\right)$ is constant and hence

$$P\left(S|D\right) \propto P\left(D|S\right)P\left(S\right) \tag{3.4}$$

We define the score of S as

$$score\left(S\right) = P\left(D|S\right)P\left(S\right) \tag{3.5}$$

To compute this score, we first derive $P(S, D)$ by integrating over the parameters Θ:

$$P\left(S,D\right) = \int_{\Theta} P\left(S,D,\Theta\right)d\Theta \tag{3.6}$$

$$= \int_{\Theta} P\left(D|S,\Theta\right)f\left(\Theta|S\right)P\left(S\right)d\Theta \tag{3.7}$$

$$= P\left(S\right)\int_{\Theta} P\left(D|S,\Theta\right)f\left(\Theta|S\right)d\Theta \tag{3.8}$$

In Equation 3.7, the second expression is obtained by applying the chain rule to the first expression, and in the third expression $P(S)$ is moved outside the integral since it is not dependent on Θ. The first term in the integral $P(D|S, \Theta)$ is the likelihood of the data D given structure S parameterization Θ, and the second term $f(\Theta|S)$ is a Dirichlet probability

density function that provides a parameter prior distribution over Θ. The marginal likelihood has a closed-form solution that is obtained from the derivation given in [22, 17]:

$$P(D|S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(N_{ijk})} \tag{3.9}$$

where Γ is the gamma function, r_i is the number of values of the variable denoted by node i, q_i is the number of parent states of node i, n is the number of nodes (variables), n_{ijk} is the number of samples in D that have values corresponding to parent state j for node i and node i taking the value k, $n_j = \sum_k n_{jk}$, α_{ijk} is a parameter prior that can be interpreted as belief equivalent to having previously (prior to obtaining D) seen α_{ijk} samples that have values corresponding to parent state j for node i and node i taking the value k, and $\alpha_{ij} = \sum_k \alpha_{ijk}$.

Heckerman et al proposed the BD score by making four assumptions on $P(S, D)$ [17].

The first assumption is that data D is a multinomial sample drawn from a BN structure S. The second assumption is that the parameter prior distributions follow Dirichlet distributions [17].

The Dirichlet distribution is a continuous multivariate probability distribution commonly used as a prior distribution in Bayesian statistics. It is a generalization of the beta distribution; the Dirichlet distribution is the conjugate prior of the multinomial distribution. Its probability density function returns the belief that the probabilities of K virtual events are x_i given that each event has been observed α_{i-1} times [17].

$$p(\theta_{ij}) = \frac{\Gamma(\prod_{k=1}^{r_i} \alpha_{ijk})}{\prod_{k=1}^{r_i} (\Gamma(\alpha_{ijk}))} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} \tag{3.10}$$

The third assumption is that the parameters are independent both globally and locally. Global parameter independence implies that the probability of the global graph parameters is equivalent to the product of the probability of the local graph parameters [17].

$$p(\Theta|S) = \prod_{i=1}^{n} p(\Theta_i|S) \tag{3.11}$$

Local parameter independence implies that the probability of the local graph parameters

is equivalent to the probability of the parameters [17].

$$p(\Theta_i|S) = \prod_{j=1}^{n} p(\Theta_{ij}|S) \forall j{=}1,...,n \tag{3.12}$$

The final assumption is that of parameter modularity such that if two DAGs have a node with the same parents, then the parameters for any local distribution are equivalent. [17]

$$p(\Theta_i|S) = p(\Theta_i|S') \tag{3.13}$$

Typically, the BD score is implemented computationally in log form to avoid numerical underflow arising from very small probability numbers. Equation 3.9 in log form is as follows: [17]

$$BD\left(S|D\right) = \log P(S) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{\Gamma(\alpha_{ij})}{\Gamma\left(n_{ij}+\alpha_{ij}\right)}\right)\sum_{k=1}^{r_i}\log\left(\frac{\Gamma\left(n_{ijk}+\alpha_{ijk}\right)}{\Gamma(\alpha_{ijk})}\right)\right) \tag{3.14}$$

Several specializations of the BD score include the K2, BDe, and the BDeu scores, and these are described next.

3.5.2 K2 score

The K2 score is a specialization of the BD score where the parameter prior is set to 1. Given that the counts are integers, and leveraging the fact that $\Gamma(x) = (x-1)!$, we can reduce the equation to the following form [22, 38].

$$K2\left(S|D\right) = \log P(S) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{(\alpha_{ij}-1)!}{(n_{ij}+\alpha_{ij}-1)!}\right)\sum_{k=1}^{r_i}\log\left(\frac{(n_{ijk}+\alpha_{ijk}-1)!}{(\alpha_{ijk}-1)!}\right)\right) \tag{3.15}$$

$$K2\left(S|D\right) = \log P(S) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{(r_i-1)!}{(n_{ij}+r_i-1)!}\right)\sum_{k=1}^{r_i}\log\left(n_{ijk}!\right)\right) \tag{3.16}$$

3.5.3 BDe score

Heckerman makes two additional assumptions of likelihood equivalence and structure possibility to derive the BDe score, where e stands for likelihood equivalence. Two DAGs are equivalent if they encode the same joint probability distributions. The assumption of likelihood equivalence implies that if two graphs are equivalent, then the parameters of the graphs are equivalent. For any complete DAG, $P(S) > 0$. With these six assumptions, the BDe score is given by:

$$BDe\left(S|D\right) = \log P(S) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{\Gamma\alpha_{ij}}{\Gamma\left(n_{ij}+\alpha_{ij}\right)}\right)\sum_{k=1}^{r_i}\log\left(\frac{\Gamma\left(n_{ijk}+\alpha_{ijk}\right)}{\Gamma\alpha_{ijk}}\right)\right) \tag{3.17}$$

3.5.4 BDeu score

Buntine (1991) proposed a specialization of the BDe score [14, 17]:

$$BDeu\left(S|D\right) = \log P(S) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{\Gamma\left(\frac{\alpha}{q_i}\right)}{\Gamma\left(n_{ij}+\frac{\alpha}{q_i}\right)}\right)\sum_{k=1}^{r_i}\log\left(\frac{\Gamma\left(n_{ijk}+\left(\frac{\alpha}{r_i q_i}\right)\right)}{\Gamma\left(\frac{\alpha}{r_i q_i}\right)}\right)\right) \tag{3.18}$$

which happens when

$$P\left(X_i = x_{ik}, Pa_i = pa_i|S\right) = \frac{\alpha}{r_i q_i} \tag{3.19}$$

This scores needs the specification of only one hyperparameter - the equivalent sample size α.

3.5.5 Conditional Bayesian score

I prove here that the marginal score of the child node is the same as the conditional score. The conditional likelihood of a structure S given data D is given by [101]:

$$CL\left(S|D\right) = \prod_{i=1}^{n} P\left(z^i|x^i\right) \tag{3.20}$$

where i indexes the instances in the data. This is equivalent to a conditional log-likelihood of a model S given data D:

$$CLL\left(S|D\right) = \sum_{i=1}^{n} \log P\left(z^i|x^i\right) \tag{3.21}$$

This is generally different from the marginal log-likelihood

$$LL\left(S|D\right) = \sum_{i=1}^{n} \log P\left(z^i, x^i\right) \tag{3.22}$$

The maximization of the conditional log likelihood and the marginal log likelihood are equivalent. From the definitions above, the following derivation applies.

$$LL\left(S|D\right) = \sum_{i=1}^{n} \log P\left(z^i, x^i\right) \tag{3.23}$$

$$= \sum_{i=1}^{n} \log \left(P\left(z^i, x_1^i, ...x_k^i\right) \times \prod_{j=1}^{k} P\left(x_j^i\right) \right)$$

$$= \sum_{i=1}^{n} \log P\left(z^i, x_1^i, ...x_k^i\right) \times \sum_{i=1}^{n} \log \prod_{j=1}^{k} P\left(x_j^i\right)$$

$$= CLL\left(S|D\right) + \sum_{i=1}^{n} \log \prod_{j=1}^{k} P\left(x_j^i\right)$$

where j indexes the variables in S. Both $CLL(S|D)$ and $\sum_{i=1}^{n} log \prod_{j=1}^{k} P(x_j^i)$ are negative. $P(z^i|x_1^i, ..., x_k^i)$ and $P(x_j^i, j = 1...k)$ are assumed to have independent parameterizations. A $P(z^i|x_1^i, ..., x_k^i)$ that maximizes the marginal log likelihood also maximizes the conditional log likelihood [101].

3.6 SCORING PROBABILISTIC RULES

A probabilistic rule model is essentially a BN with a child node and one or more parent nodes. Moreover, in a probabilistic rule model, we are interested in accurately predicting the child node from the parent nodes; thus, the model score of interest is the BN score associated

with the child node. Specifically, we want the score to represent the conditional marginal likelihood of the child node given the parent nodes. It can be shown that the conditional marginal likelihood is the same as the marginal likelihood of the child node that is typically computed in Bayesian methods that do BN structure learning.

For a probabilistic rule model, the K2 score is obtained from Equation 3.16 by noting that there is only one node to be scored:

$$K2\left(S|D\right) = \log P(S) + \sum_{j=1}^{q_i} \left(\log \left(\frac{(r-1)!}{(n_{ij}+r-1)!} \right) \sum_{k=1}^{r_i} \log \left(n_{jk}! \right) \right) \tag{3.24}$$

More generally, a probabilistic rule model that is represented by a BN with a local decision tree, the above K2 score is modified to give:

$$K2\left(S|D\right) = \log P(S) + \sum_{l=1}^{|L|} \left(\log \left(\frac{(r-1)!}{(n_{ij}+r-1)!} \right) \sum_{k=1}^{r_i} \log \left(n_{lk}! \right) \right) \tag{3.25}$$

where l indexes the leaves in the tree.

3.7 MODEL SELECTION AND BAYESIAN MODEL AVERAGING

Most methods in statistical machine learning learn a single model from data to apply for predicting a target variable Z; this is model selection. Model selection ignores the uncertainty in choosing a model based on limited data. In model selection, we must therefore use a training dataset D to search for a locally good model M that predicts well. Therefore, we learn a model for $P(Z|X, M)$. We assume that the model M must be correct. Ergo, we assume that $P(M|D) = 1$. In other words, model selection assumes:

$$P\left(Z|X, D\right) = P\left(Z|X, M\right) P\left(M|D\right) = P\left(Z|X, D\right) \tag{3.26}$$

However, we can consider intermediate prospects instead, where $P(M|D)$ ranges from 0 to 1, representing our confidence in each model [111]. Model averaging is a coherent approach to dealing with uncertainty in model selection improving predictive performance and provide more accurate estimates of prediction error. The authors of [40] describe several

examples of a significant prediction error decrease using Bayesian model averaging. Bayesian model averaging is most useful when no single model in the considered model space has a high posterior probability.

Complete Bayesian model averaging, where we average over the entire model space, is generally not computationally feasible. In these cases, selective Bayesian model averaging is generally used instead, where I average over a subset of models.

The traditional Bayesian model averaging algorithm to predict a target variable Z based on a training dataset D across k models M_k in our model space M can be elucidated as [111]:

$$P\left(Z=1|D\right)=\sum_{k\in M}P\left(Z=1|D,M_k\right)P\left(M_k|D\right) \tag{3.27}$$

As long as the sum of probabilities $\sum_M P(M|D) = 1$, the final prediction reflects the combination model efforts.

$$P\left(Z|X,D\right)=\sum_{M}P\left(Z|X,M\right)P\left(M|D\right) \tag{3.28}$$

An analysis of Bayesian model averaging by Madigan and Raftery indicates that model averaging is expected to perform better than model selection [61, 92, 111, 104].

Model averaging offers many advantages over model selection. It is possible to average over genetic models - for example, to include considerations for additive and non-additive models - by offering weights that reflect our prior belief in how likely certain models can be [92].

3.8 THE BAPR ALGORITHMS

This section describes two BAPR algorithms that I developed, implemented, and evaluated for biomarker discovery. The algorithms use probabilistic rule (PR) models that are evaluated with a Bayesian score and perform greedy search to identify high scoring models and average over them. The two BAPR algorithms differ in the representation used for the

model. One algorithm uses BNs with global structure which is called Bayesian Averaged Probabilistic Rules - Full Tree (BAPR-FT). Note that the global structure that is typically represented by a conditional probability table can be represented by a local full tree. The second algorithm uses BNs with local decision trees and is called Bayesian Averaged Probabilistic Rules (BAPR). Note that the model space of BAPR is a superset of the model space of BAPR-FT, since the space of all full trees is a subset of all possible trees.

Given a dataset containing SNV measurements in a set of cases and controls, a BAPR algorithm outputs a list SNVs that are ranked according to their ability in discriminating the cases from the controls.

3.8.1 Model score

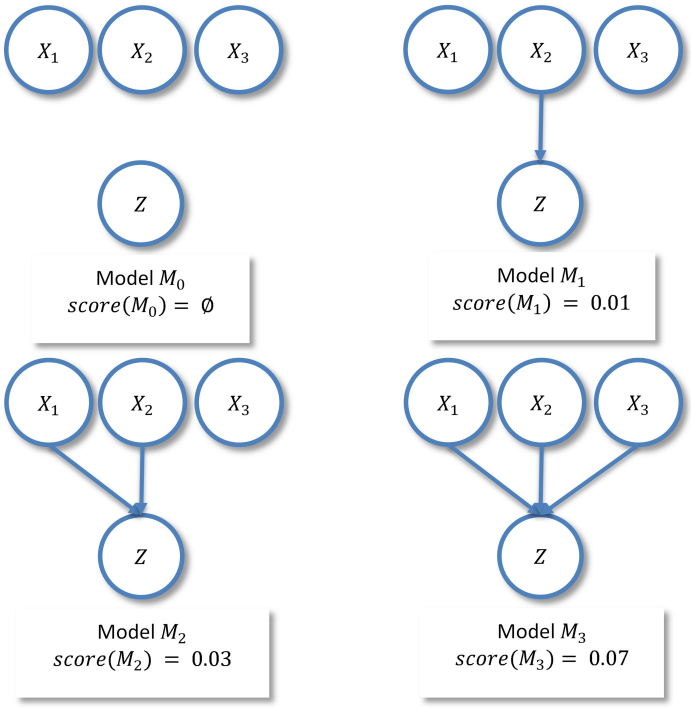
The score of a predictor variable or a biomarker (such as a SNV) is defined as the sum of the scores of PR models that include the predictor (i.e., the BN structure has an arc from the predictor node to the target node). The PR model score is the conditional marginal score of the child node given the parent nodes as described in Section 3.5.5.

$$score\,(\text{biomarker}) = \sum_{i \in \{S\}} score\,(S_i|D)\, indicator\,(\text{biomarker}) \tag{3.29}$$

where i indexes structures in S, $score(S_i|D)$ is given by the $K2$ score in Equation 3.24 for PR models represented by BN global structure, and by the $K2$ score in Equation 3.25 for the local tree structure. Additionally, $indicator\,(\text{biomarker}) = 1$ if biomarker is in structure S_i and 0 otherwise. Note that the $score(S|D)$ score includes a structure prior probability distribution $P(S)$. The next section describes several structure priors that can be incorporated into BAPR.

3.8.2 Structure priors

3.8.2.1 Uniform prior. The simplest prior is the uniform structure prior where all model structures are deemed to be equally likely. I set $P(S)$ to 1 when using the uniform prior: this implies that with the uniform prior, the model score is just the marginal likelihood.



$score(X_1)$	$= score(M_2) + score(M_3)$ $= 0.03 + 0.07$ $= 0.10$	$score(X_3)$	$= score(M_3)$ $= 0.07$
$score(X_2)$	$= score(M_1) + score(M_2) + score(M_3)$ $= 0.01 + 0.03 + 0.07$ $= 0.11$	$score(Z)$	$= score(M_1) + score(M_2) + score(M_3)$ $= 0.01 + 0.03 + 0.07$ $= 0.11$

Figure 11: Bayesian averaged score for probabilistic rules.

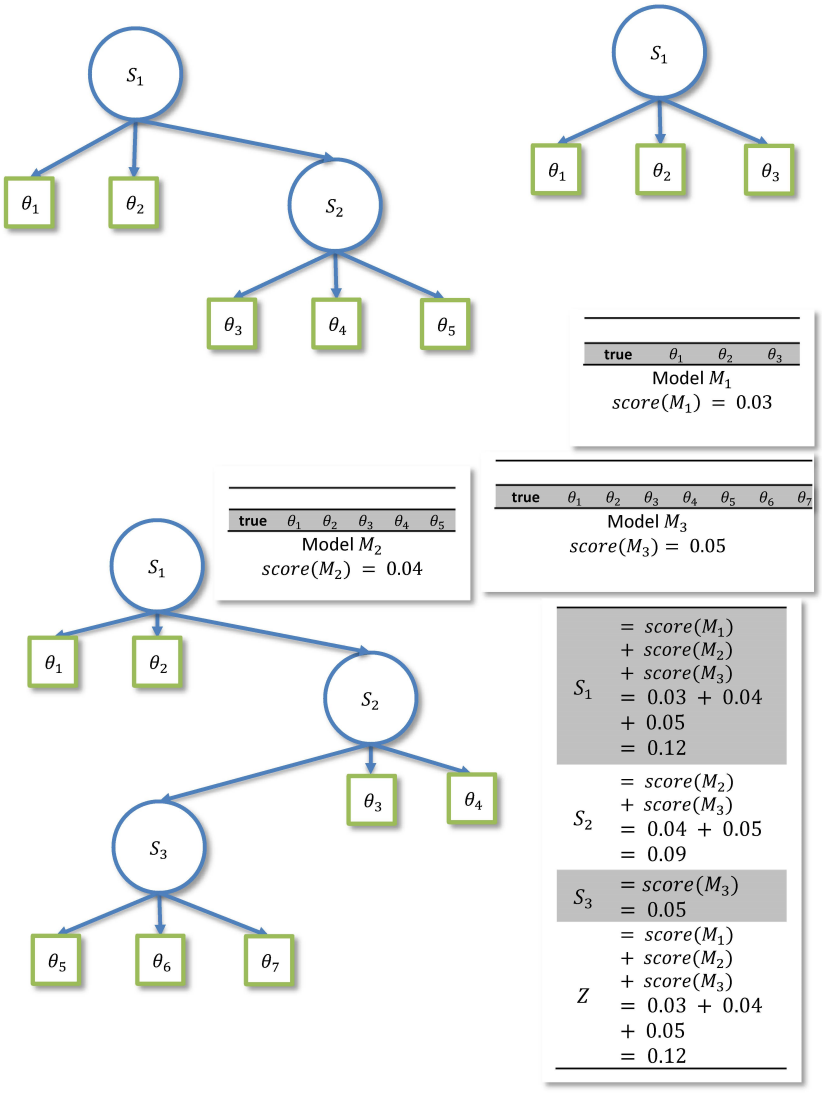


Figure 12: Model averaged probabilistic tree rules.

3.8.2.2 Binomial prior with enp.

A simple binomial structure prior incorporates a single parameter called the expected number of predictors (enp), and is given by:

$$P(S) = \prod_{i \in Q} p_i \prod_{i \notin Q} (1 - p_i) \tag{3.30}$$

where Q is the subset of variables in D that are in S. Let m be the **e**xpected **n**umber of **p**redictors (enp) in D that a domain expert expects to be predictive of Z, and let n be the total number of variables in D. Then, a simple binomial prior is given by $p_i = m/n$, when $i \in Q$, and otherwise $p_i = 1 - (m/n)$.

3.8.2.3 Binomial prior with SNV deleteriousness scores.

Given a deleteriousness score for SNVs (e.g., SIFT and PhyloP), I incorporate the score using a simple binomial structure prior as follows. If necessary, the score is scaled to the range 0 to 1 to provide a prior probability for a SNV such that 0 denotes no risk of disease and 1 denotes that disease is certain.

If the SNV is present in a model, I take the prior probability. If the SNV is absent, I take the complement of the prior probability. The structure prior is the product of the probabilities for the presence or absence of these biomarkers.

3.8.2.4 Combination priors.

Since several SNV scoring methods are available that can potentially provide independent information, I combine two or more scores using the heuristic of taking the maximum of the scores.

3.8.3 Search strategies

The BAPR algorithms use greedy search to explore a space of PR models. A PR model contains one or more SNVs as parents of the disease of interest, and the SNVs are constrained to be all on the same gene. In other words, each PR model contains only SNVs from a single gene. These models will capture interaction effects among SNVs located on a single gene but will miss interactions among SNVs on different genes. I adopted this approach to make the algorithm computationally tractable.

For each of these search strategies, we generate new probabilistic tree rules and probabilistic rules by the functions seen in Algorithm 1 and Algorithm 2. Each of these functions takes the previous model and creates a new model space by adding various variables to the model. After the search algorithm selects the successor model(s), it averages over them as shown in Algorithm 3.

Algorithm 1 Pseudocode for generating new probabilistic tree rule models. *PtrModel* represents *probabilistic tree rule model*.

```
1:  procedure GeneratePtrModel(PreviousPtrModel, Prior, TrainingData)
2:      ModelSpace ← ∅
3:      for all LeafNode ∈ PreviousPtrModel.LeafNodes do
4:          for all VariableToTest ∈ VariablesLeftToTest do
5:              NewPtrModel ← PreviousPtrModel.LeafNode.AddNode(VariableToTest)
6:              NewPtrModel.UpdateCounts(TrainingData)
7:              ComputePrior(NewPtrModel)
8:              ComputeScore(NewPtrModel)
9:              ModelSpace.Add(NewPtrModel)
10:     return ModelSpace
```

Algorithm 2 Pseudocode for generating new probabilistic rule models. *PrModel* represents *probabilistic rule model*.

```
1:  procedure GeneratePrModels(PreviousPrModel, Prior, TrainingData)
2:      ModelSpace ← ∅
3:      for all VariableToTest ∈ VariableRemainingToTest do
4:          NewPrModel ← AddNode(PrevousPrModel, LeafNode, VariableToTest)
5:          UpdateCounts(NewPrModel, TrainingData)
6:          ComputePrior(NewPrModel)
7:          ComputeScore(NewPrModel)
8:          ModelSpace.Add(NewPrModel)
9:      return ModelSpace
```

This implementation of model averaging combined with greedy search poses a caveat regarding the score summations. For each level of greedy search, we iterate throughout the

Algorithm 3 Model averaging pseudocode.

```
1: function ModelAverage(ListOfModels)
2:     VariableScores(Variable) ← ∅
3:     for all Model ∈ ListOfModels do
4:         for all Variable ∈ Model do
5:             VariableScores[Variable] = LogSum(VariableScores[Variable], Model.Score)
6:     return VariableScores                ▷ Score for each variable in the model space
```

space of possible extensions to our current model. A best first approach will then create models that have every biomarker in the best first model of the previous level with a new biomarker. Consequently, the number of models averaged for each variable is significantly higher for variables within the selected model compared to variables not in the selected model.

For example, for variables X_1 through X_n, ordered in descending significance, we can write the score of X_1 as the sum of the scores of 1-biomarker models containing X_1 added to the sum of the scores of all 2-biomarker models containing X_1, and so on.

$$score(X_1) = \sum_{|S|=1, X_1 \in S} score(S) + \sum_{|S|=2, X_1 \in S} score(S) + \sum_{|S|=3, X_1 \in S} score(S) + \dots \tag{3.31}$$

This then reduces to

$$score(X_1) = score(S = \{X_1\}) + \sum_{|S|=2, \{X_1\} \in S} score(S) + \sum_{|S|=3, \{X_1, X_2\} \in S} score(S) + \dots \tag{3.32}$$

and therefore explores

$$\text{number of models examined} = 1 + (n-1) + (n-2) + \dots \tag{3.33}$$

For X_2, however, X_2 begins to have an included score at the 2-biomarker model level

$$score(X_2) = score(S = \{X_2\}) + score(S = \{X_1, X_2\}) + \sum_{|S|=3, \{X_1, X_2\} \in S} score(S) + \dots \tag{3.34}$$

and therefore explores

$$\text{number of models examined} = 1 + 1 + (n-2) + \dots \tag{3.35}$$

This is more evident for X_3, which starts to have more coverage in the 3-biomarker model space.

$$+ score(X_3) = score(S = \{X_3\}) + score(S = \{X_1, X_3\}) + score(S = \{X_1, X_2, X_3\}) + \sum_{|S|=4, \{X_1, X_2, X_3\} \in S} score(S) + \cdots + \quad (3.36)$$

and therefore explores

$$\text{number of models examined} = 1 + 1 + 1 + (n - 3) + \ldots \quad (3.37)$$

By extension, for a search with variable threshold *depth*, variables $X_1, \ldots, X_{depth}$ will have significantly more models examined and, thus, suffer from scoring inflation. All other variables will incorporate scores from a fixed number of models, *depth*.

Due to this uneven weighting of scores for variables within the final model, for a study within a gene, the model averaged and greedy results should be identical from 1 to *depth*. Biomarkers not selected within the model will average across *depth* scores, and should be unbiased in the number of models to average over.

4.0 EXPERIMENTAL METHODS

This chapter describes the datasets, the characterization of SNV scores used for structure priors, and the experimental methods and evaluation of the algorithms. Section 4.1 describes the datasets used in the experiments, Section 4.2 describes variant deleteriousness scores that are used for structure priors, Section 4.3 describes the performance metrics used to quantify algorithmic performance, and Section 4.4 gives details of the comparison algorithm.

4.1 DATASETS

For the experiments, I used several exome datasets, including a semi-synthetic SNV dataset called GAW17, a small-sample Alzheimer's disease dataset, a large-sample Alzheimer's disease dataset, and a chronic pancreatitis dataset. All datasets have one binary target variable that denotes the case/control status of an individual and predictor variables that are SNVs. Each SNV has up to four states that denote the three genotypes and a fourth state for missing when the genotype measurement for an individual is not available. Details of the datasets are provided in the following sections.

4.1.1 Genome Analysis Workshop 17 (GAW17) semi-synthetic mini-exome

The GAW17 dataset is a mini-exome semi-synthetic dataset that was created for the Genetic Analysis Workshop 17. The exome data comes from 697 unrelated individuals (sequenced in the 1000 Genomes Project) and consists of 24,487 autosomal SNVs assigned to 3,205 genes [2, 32]. Four quantitative risk factors were simulated as normally distributed phenotypes;

Table 4: Brief description of datasets.

Dataset	Target	# genes	# SNVs	# individuals (# cases / # controls)
16-gene GAW17	Q2	16	112	6,970 (3457 / 3513)
full GAW17	Q2	3,205	24,487	6,970 (3457 / 3513)
TGen	LOAD	12,535	115,059	1,411 (861 / 550)
Kamboh-small	LOAD	19,444	352,693	22 (15 / 7)
Kamboh-large	LOAD	21,585	787,586	584 (299 / 285)
NAPS	Chronic pancreatitis	19	191	2,201 (980 / 1221)

however, values of only three of the risk factors are provided in the dataset. The genes associated with the risk factors were chosen from the cardiovascular disease (CVD) risk and inflammation pathways. In addition, a binary disease phenotype representing CVD was modeled as a function of the four quantitative risk factors. In the simulated phenotype data, the values of three of the risk factors (named Q1, Q2, and Q4) and the binary phenotype were provided for each individual; the values of the remaining risk factor was not provided to simulate a latent factor. Q1 was influenced by age and 39 SNVs in 9 genes and included a genotype-smoking interaction. Q2 was influenced by 72 SNVs in 13 genes and was not influenced by age, sex, or smoking status. Q4 was influenced by age, sex and smoking; while it had a genetic component, it was not influenced by any of the SNVs in the dataset. The latent factor was influenced by 51 SNPs in 15 genes. A total of 200 datasets were provided; each dataset contains 697 individuals with simulated disease/healthy status.

This dataset simulated a common disease using rare and common SNVs based on the current thinking that both rare and common SNVs contribute to the genetic basis of common diseases. Of the causative SNVs, 38.4% are private variants - with only one individual out of 697 having the variant - and 12.8% SNVs are common with MAF > 0.05.

I processed the GAW17 data to create two derivative datasets.

4.1.1.1 16-gene GAW17:

This dataset was created by pooling the first ten replicates and contains 16 genes and associated 112 SNVs, where 5 of the 16 genes were chosen from the 36 causal genes and the remaining 11 genes were non-causal.

4.1.1.2 Full GAW17:

This dataset was created by pooling the first ten replicates and contained 24,487 SNVs and 6,970 individuals. The target variable was the Q2 risk factor that was converted to a binary variable, since Q2's underlying model uses only genetic variables and no latent factors.

4.1.2 TGen Alzheimer's disease dataset

The Translational Genomics Research Institute's (TGen) is a late-onset Alzheimer's disease (LOAD) GWAS dataset that was collected by Reiman et al. [81].The genotype data were collected on 1,411 individuals, of which 861 had LOAD and 550 did not. Of the 1,411 individuals, 1,047 were brain donors in whom the status of LOAD or control was neuropathologically determined, and 364 were 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 outcome is the binary LOAD variable. In this dataset, 61% (861 of 1411) had LOAD. For each individual, the genotype data consists of 312,318 common SNVs that includes the two well-established ApoE SNVs associated with LOAD, namely, rs429358 and rs7412.

I processed this data with SNPnexus to annotate the the SNVs and used PLINK to extract only the exonic SNVs [77, 18, 26, 97]. The final dataset consisted of 115,059 common SNVs in 12,535 genes for 1,411 individuals.

4.1.3 Kamboh-small Alzheimer's disease dataset

This is a whole-exome dataset that contains 22 individuals of which 15 have LOAD and 7 are controls. This dataset comes from Dr. Kamboh at the University of Pittsburgh Alzheimer's Disease Research Center (Kamboh) and consists of 352,693 autosomal SNVs assigned to 19,444 genes.

I annotated this dataset with Ingenuity Variant Analysis to obtain a set of unique variants and additional information, such as relevant genes, MAFs, SIFT scores, and PhyloP scores.

4.1.4 Kamboh-large Alzheimer's disease dataset

This is a whole-exome dataset that contains 584 individuals, of which 299 had LOAD and 285 were controls. This dataset also comes from the Kamboh and Dr. M. Ilyas Kamboh and Dr. M. Barmada were critical in the development and acquisition of this dataset. Sequencing was performed with the Illumina HiSeq 2x100bp whole genome sequencer, with SureSelect All Exon v5 enrichment.

I processed this data as follows. PLINK was used to remove variants that were not SNVs (e.g., multiple base pair indels)[77]. SNPnexus was used to annotate the SNVs, and SNVs that could not be assigned to a gene or coding sequence (including long-intergenic non-protein coding RNAs and open reading frames) were removed [18, 26]. The final dataset consisted of 787,586 SNVs in 21,585 genes or coding sequences for 584 individuals.

4.1.5 NAPS2 chronic pancreatitis dataset

The chronic pancreatitis dataset consists of 2,201 individuals, of whom 980 were cases and have recurrent acute pancreatitis or chronic pancreatitis and 1,221 were controls. This cohort, which is one of the largest of its kind, was enrolled by the North American Pancreatitis Study (NAPS2) Group from pancreatic care centers across the US. For each individual, GWAS was filtered to provide 191 exonic SNVs and a case/control status.

4.2 INFORMATIVE STRUCTURE PRIORS

A major advantage of using a Bayesian approach is the ability to incorporate existing biological knowledge in the form of prior probabilities. The power to identify associations of rare variants, and common ones, is likely to be improved by including appropriate prior knowledge. Moreover, these priors play an important role in mitigating overfitting.

Table 5 gives the type of structure prior, source of information, and the formula for computing the prior probability of a SNV being associated with disease. Details of each of the variant deleteriousness scores is described in 2.4.

4.3 EVALUATION

Several methods are useful in evaluating algorithmic performance in biomarker discovery and ranking. The evaluation methods that I used included biomarker recovery plots, area under the receiver operating characteristic curve (AUC), and evidence of biological validity.

I refer to *algorithmically relevant* biomarkers as biomarkers that have been selected by the algorithm of interest and refer to *truly relevant* biomarkers as biomarkers that are known to be causative of disease. Also, I differentiate between *biomarker discovery* - identifying novel biomarkers - and *biomarker recovery* - identifying known biomarkers.

4.3.1 Biomarker recovery plots

For common diseases, partial knowledge of truly relevant biomarkers including SNVs is available from the literature or from online gene-disease association databases. Taking advantage of these sources, I obtained lists of truly relevant SNVs associated with LOAD and chronic pancreatitis. I evaluate an algorithm by the ranking of SNVs it produces by creating a biomarker recovery plot. In this plot, the x-axis denotes number (or percent) of top-ranked SNVs n and the y-axis denotes the number of truly relevant SNVs in the algorithmically relevant top-ranked n SNVs (biomarker recovery rate). To compare two algorithms, I com-

Table 5: Brief description of structure priors, source of information, and formulas.

Structure prior	**Source of information**	**Formula**
uniform	none	$P(S) = 1$
bin enp	# of expected SNVs associated with disease	$p_i = \frac{m}{n}$ where m is the expected number of disease-associated SNVs
bin MAF	MAF denotes rarity of SNV	$p_i \propto -0.04ln(MAF_i) + 0.17$
bin SIFT	effect on protein function	$p_i \propto 1 -$ SIFT score
bin Polyphen	effect on protein function	$p_i \propto$ Polyphen-2 score
bin PhyloP	degree of conservation	$p_i \propto$ PhyloP score
bin GERP	degree of conservation	$p_i \propto$ GERP score
bin Phast	degree of conservation	$p_i \propto$ Phast score
bin CADD	derived from several scores	$p_i \propto$ scaled CADD score
bin max 4	maximum of four scores	$p_i \propto max$(SIFT, Polyphen, Phast, CADD scores)

pare the proportions of truly relevant biomarkers to algorithmically relevant biomarkers at the 95^{th} percentile (equivalently the top 5%) using the chi-square test.

4.3.2 Evaluation of predictive performance

A relevant biomarker should be predictive of disease and a classifier developed from relevant biomarkers should perform well in discriminating between cases and controls. I evaluate predictive performance of the top-ranked SNVs by measuring the AUCs of a series of classification models using increasing numbers of top-ranked SNVs.

Given a set of top-ranked SNVs, I derive a k-Nearest Neighbor (kNN) classification model to predict the target (case/control status). I evaluate the performance of the classification model using five-fold cross-validation. The dataset is randomly partitioned into five approximately equal sets such that each set has a similar proportion of individuals who have the disease. I apply the algorithm on four sets taken together as the training data, and evaluated the top-ranked SNVs' predictions on the remaining test set. I repeat this process for each possible test set to obtain a prediction for each individual in the dataset. I use the predictions to compute the AUC which is a widely used measure of classification performance.

The kNN algorithm is a simple non-parametric classification algorithm that utilizes pairwise distances between a test individual and the individuals in the training sets. For SNV data, the pairwise distance simply counts the number of SNVs which have different values between the test and training individuals. The classification result for the test individual is then computed as the average target value among the k most similar training individuals. I used a setting of $k = 10$.

4.3.3 Evidence of biological validity

For exome datasets with real outcomes (e.g., LOAD, chronic pancreatitis), I examined the top-ranked SNVs and associated genes for biological significance and evidence of previously documented association with disease. I used a publically available database called DisGeNET v4.0 that integrates information on gene-disease and variant-disease associations from several

public data sources and the literature [73].

4.4 COMPARISON ALGORITHM

The chi-square statistic is typically used to rank SNVs in exome data analysis, and I use chi-square as the main univariate comparison algorithm. The test tabulates observations in a contingency table, which records co-occurrences of variable states and the target variable. The chi-square 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-square 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 i^{th} row and j^{th} column, and $E_{i,j}$ is the expected value in the same cell. The resulting statistic can be compared to the chi-square distribution to determine the probability of association between the SNV's genotypes and the target variable. I utilized the chi-square function in PLINK to calculate chi-square statistics and p-values for each locus.

Random forests (RFs) are commonly used as multivariate analysis methods. An ensemble of classification and regression trees (CARTs) are learned from a random subset of the variables in the dataset. I used GenABEL [5] and Ranger [107] to train RFs with default parameters.

5.0 EXPERIMENTAL RESULTS

This chapter describes the results of applying the BAPR algorithms described in Chapter 3 to the datasets described in Chapter 4. Section 5.4 describes the availability of the various variant deleteriousness scores across the datasets. Section 5.1 provides results of several versions of the BAPR algorithms on the 16-gene GAW 17 dataset. Sections 5.2 and 5.3 provide results obtained from an Alzheimer's disease GWAS dataset and an Alzheimer's disease exome dataset with a small sample size. Section 5.4 describes the availability of the various variant deleteriousness scores across the datasets. Section 5.5 validates BAPR with biologic priors and provides results from the full GAW17 dataset. Section 5.6 gives results from an Alzheimer's disease exome dataset with a large sample size and Section 5.7 implement BAPR on a chronic pancreatitis exome dataset.

For clarity, names that are italicized refer to the algorithms used (e.g. *bin SIFT*, *chi-square*) and names that are not italicized refer to a statistical test itself (e.g. chi-square test).

5.1 RESULTS FROM THE 16-GENE GAW17 DATASET

This section describes qualitative evaluation of BAPR on the 16-gene GAW 17 dataset that explores two model structures (global structure vs. local tree structure), two structure scores (K2 vs. BDeu), two search strategies (exhaustive vs. greedy), and with and without informative structure priors. The goal of this evaluation was to identify a good performing single BAPR algorithm (structure, score, and search) to use in later experiments on larger datasets. I followed this approach since BAPR is computationally expensive and my goal

Table 6: Model structures, model scores, and search strategies used by the BAPR algorithm for results shown in Figure 15.

Label	Algorithm	Search strategy	Model score
exhaustive K2	BAPR-FT	exhaustive	K2
greedy K2	BAPR-FT	greedy	K2
exhaustive BDeu	BAPR-FT	exhaustive	BDeu
greedy BDeu	BAPR-FT	greedy	BDeu

was to comprehensively evaluate one version of the BAPR algorithm.

5.1.1 Evaluation of model structures, model scores and search strategies

Figure 13 shows the biomarker recovery plots in retrieving causal SNVs for the BAPR-FT and *chi-square* algorithms. Four versions of the BAPR-FT algorithm included combinations of model score (K2 vs. BDeu with $\alpha = 1$) and search strategy (exhaustive vs. greedy). The biomarker recovery plots indicate that *BAPR-FT with K2* score using greedy search has the best performance and also performs better than *chi-square*.

Though exhaustive search outperforms greedy search, it does so by a small margin, and it seems reasonable to use greedy search for larger model spaces where exhaustive search would be intractable.

Figure 14 shows the biomarker recovery plots for BAPR-FT, BAPR, and *chi-square*. *BAPR-FT with K2* performs the best, followed by *BAPR with BDeu*. All four BAPR versions perform better than *chi-square*.

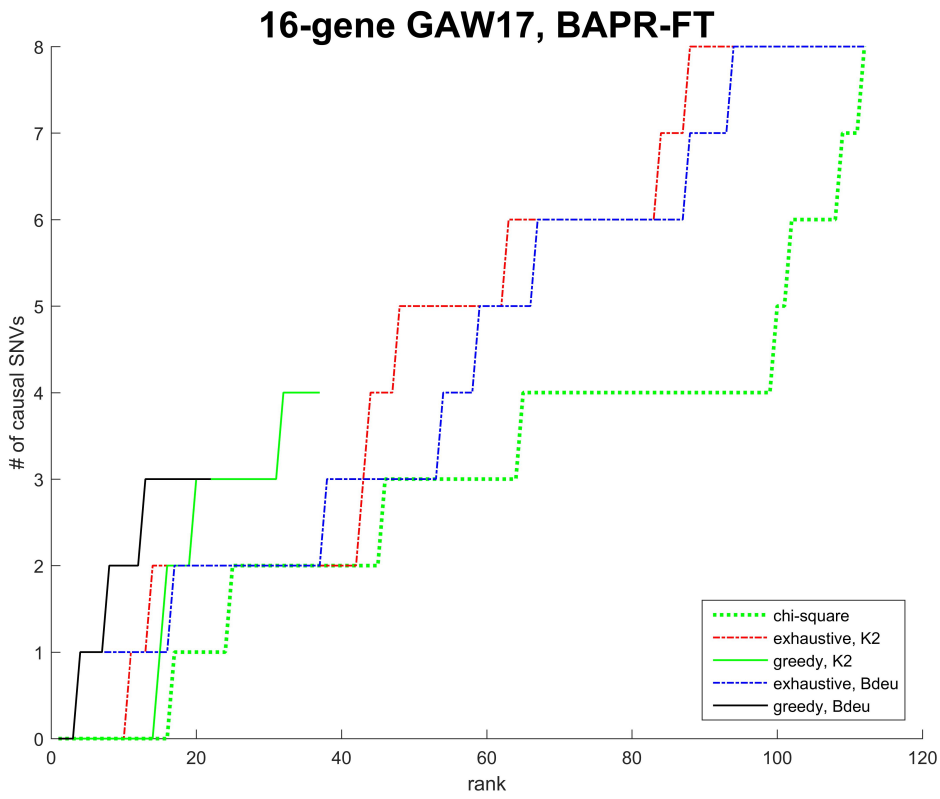


Figure 13: Biomarker recovery plots for the 16-gene GAW17 dataset using two search strategies and two model scores as shown in Table 6. The x-axis shows the number of top-ranked SNVs being considered, and the y-axis shows the number of true SNVs (recovery rate).

Table 7: Model structures, model scores, and search strategies used by the BAPR algorithm for results shown in Figure 15 on the 16-gene GAW17 dataset.

Label	Algorithm	Search strategy	Model score
global greedy K2	BAPR-FT	greedy	K2
global greedy BDeu	BAPR-FT	greedy	BDeu
tree greedy K2	BAPR	greedy	K2
tree greedy BDeu	BAPR	greedy	BDeu

5.1.2 Evaluation of structure priors

I evaluated the performance of BAPR using uniform and binomial priors with K2 and BDeu scores on the 16-gene GAW17 dataset (see Table 8). Figure 14 shows the biomarker recovery plots for various priors. The binomial prior with the K2 score performed the best. Overall, all of the BAPR methods outperformed *chi-square*.

Table 8: Model scores and structure priors used by the BAPR algorithm for results shown in Figure 15.

Label	Algorithm	Search strategy	Model score	Prior
uniform prior K2	BAPR	greedy	K2	uniform
binomial prior K2	BAPR	greedy	K2	binomial
uniform prior BDeu	BAPR	greedy	BDeu	uniform
binomial prior BDeu	BAPR	greedy	BDeu	binomial

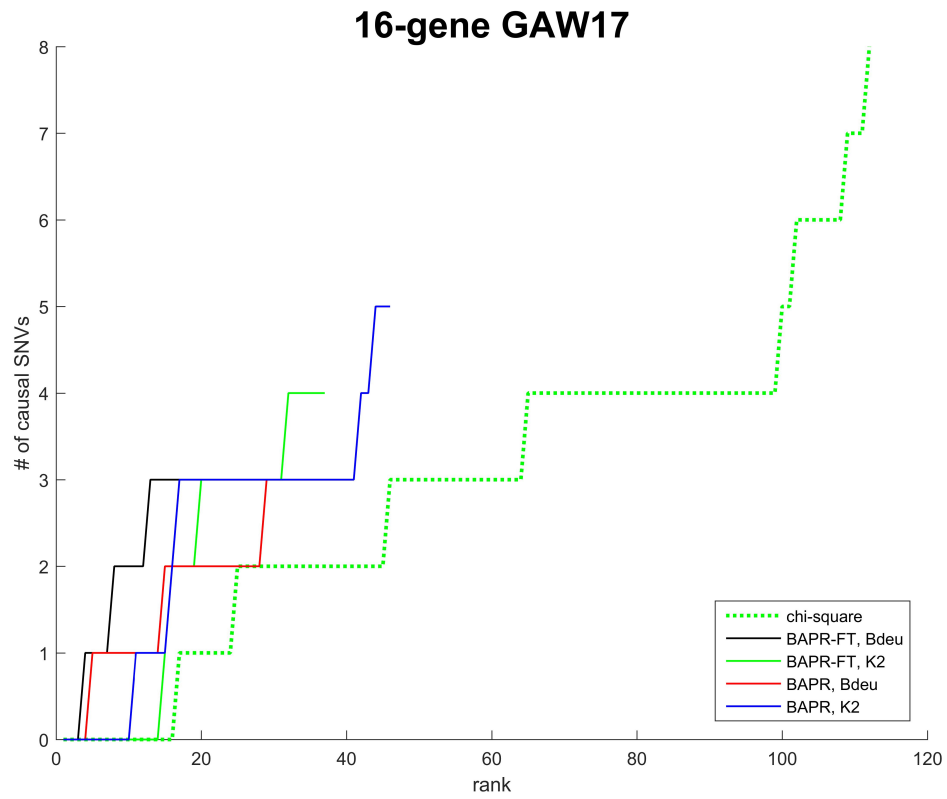


Figure 14: Biomarker recovery plots for the 16-gene GAW17 dataset using two model scores as shown in Table 8.

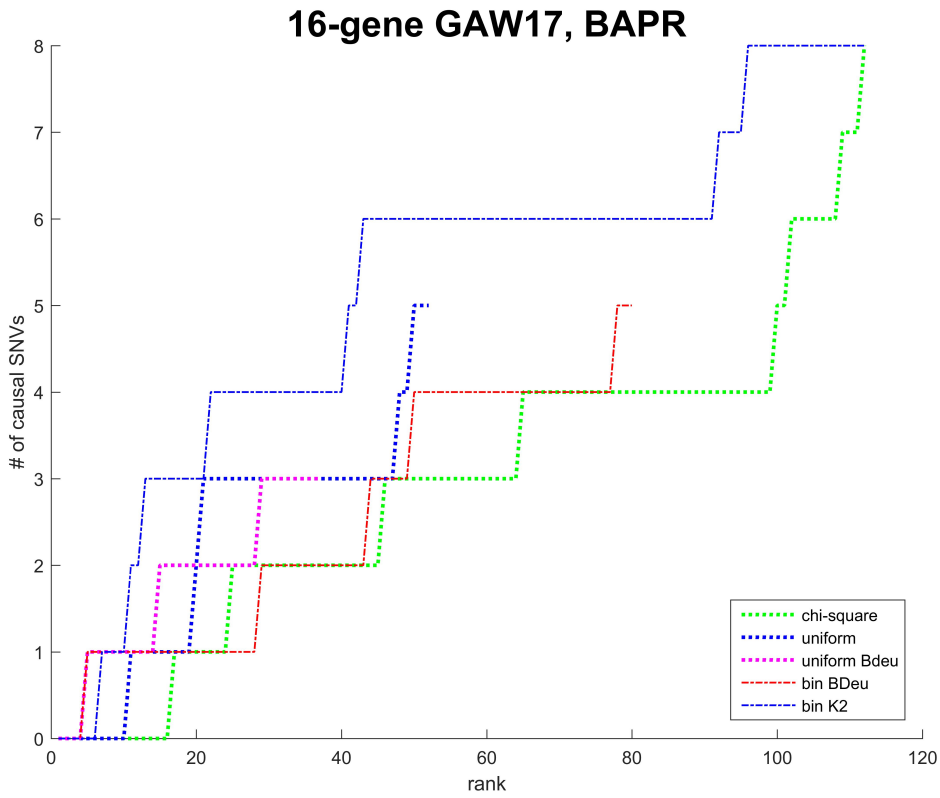


Figure 15: Biomarker recovery plots for the 16-gene GAW17 dataset for different model scores and structure priors.

Table 9: List of top ten SNVs and corresponding genes ranked by *uniform* and by *chi-square*. SNVs in bold indicate that the SNV is known to be associated with LOAD.

uniform	**chi-square**
rs7412	**rs7412**
rs429358	**rs4420638**
rs4420638	**rs429358**
rs9398855	rs16974268
rs428016	rs934745
rs270044	rs9453276
rs476366	rs1202774
rs6921729	rs4486000
rs17054975	rs207952

5.2 RESULTS FROM THE TGEN DATASET

The TGen dataset is a LOAD dataset that has only common SNVs from the exome; it consists of 115,059 common SNVs and 1,411 individuals. The performance of the BAPR algorithm (tree structure, K2 score, greedy search) was compared to that of *chi-square*.

Figure 16 shows the biomarker recovery plots for *uniform* and *chi-square*. The recovery rates at the highest ranks are similar between *uniform* and *chi-square*. Further down the rankings, *uniform* has a higher recovery rate than *chi-square*. As we progress further, however, BAPR outstrips the performance of *chi-square*. At the 50^{th} percentile, *uniform* recovers 187 SNVs while *chi-square* recovers 93 SNVs.

Table 9 shows the ten top ranked SNVs ranked by *uniform* and by *chi-square*. Both algorithms identified SNVs (*rs7412*, *rs429358*, and *rs4420638*) that are known to be strongly associated with LOAD. SNVs *rs7412* and *rs429358* together characterize the ApoE4 haplotype, which is a well-known genetic risk factor for developing LOAD.

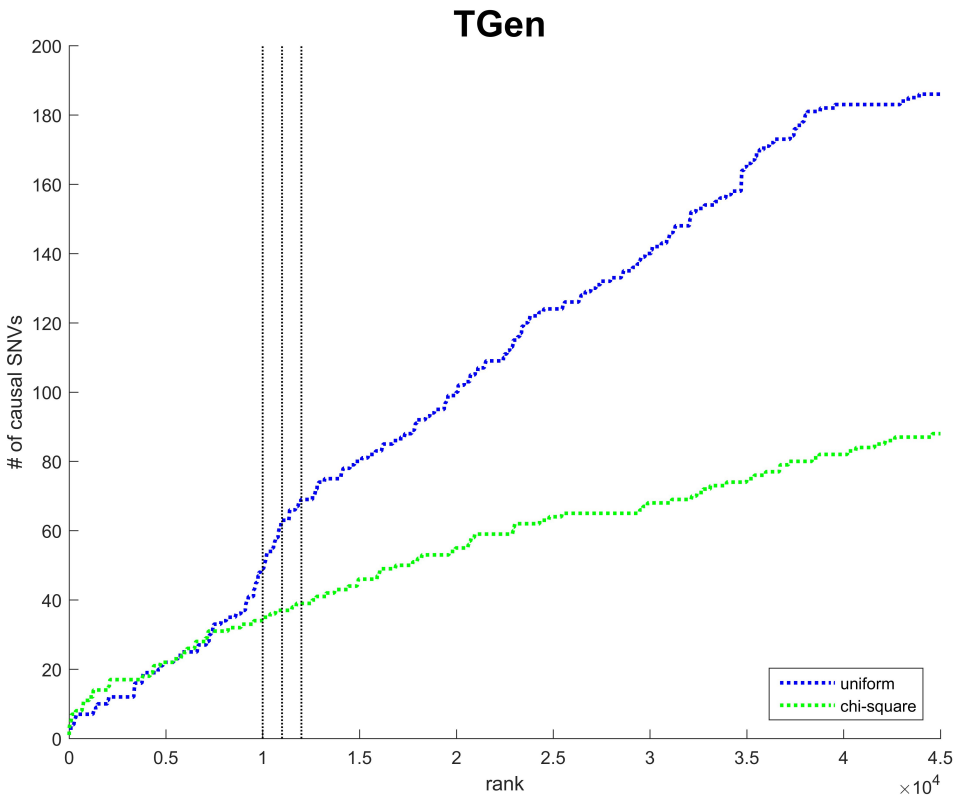


Figure 16: Biomarker recovery plot for the TGen dataset.

Table 10: P values from Fischer's exact test comparing the biomarker recovery rates of *uniform* with *chi-square* in the top-ranked 10,000, 11,000, and 12,000 SNVs.

top ranked n	**uniform**	**chi-square**	**Fisher's exact p-value**
10,000	49	35	0.078
11,000	63	37	**0.006**
12,000	69	39	**0.003**

5.3 RESULTS FROM THE KAMBOH-SMALL DATASET

The Kamboh-small dataset is an exome LOAD dataset that consists of 352,693 common and rare SNVs and 22 individuals. The performance of the BAPR algorithm (tree structure, K2 score, greedy search, *uniform*, *bin enp* - see Table 5 in Chapter 4 for a brief description of structure priors) was compared to that of *chi-square*.

Figure 17 shows the biomarker recovery plots for *uniform* and *chi-square*. The recovery rate of known LOAD-associated SNVs at the 90^{th} percentile is higher with *uniform* compared to *chi-square* (Fischer exact test, $p = 0.003$).

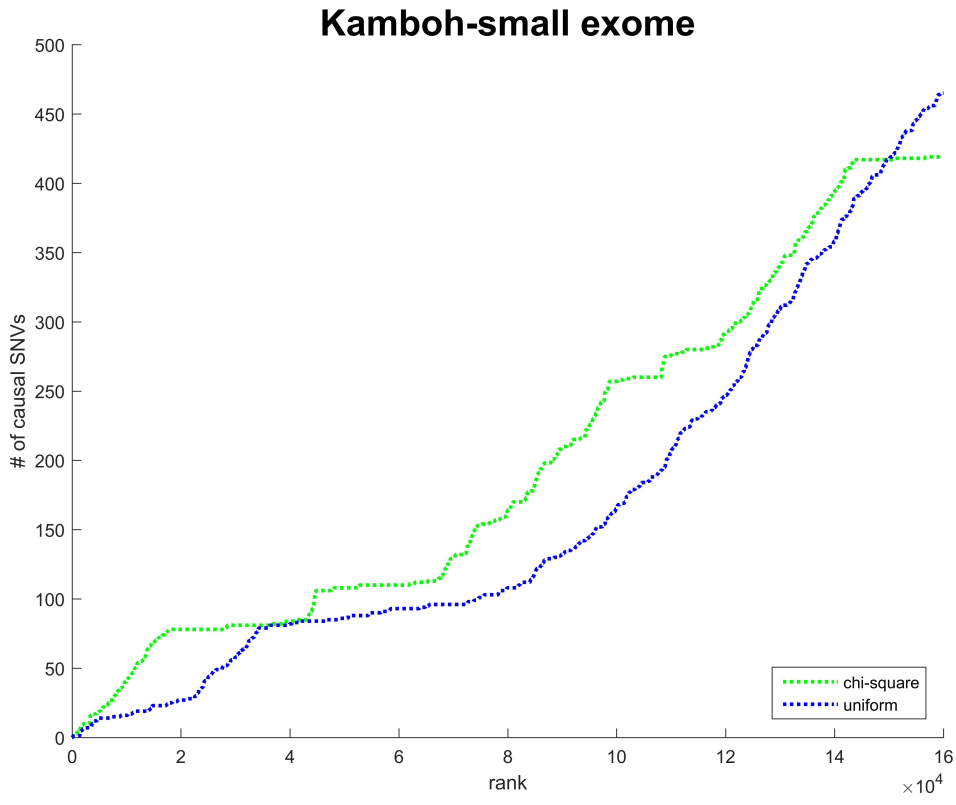


Figure 17: Biomarker recovery plots for the Kamboh-small dataset.

Table 11: Number of SNVs in each dataset and number of SNVs for which values were available for each variant deleteriousness score.

Dataset	# SNVs	CADD	GERP	Phast	PhyloP	Polyphen-2	SIFT
16-gene GAW17	112	25	8	0		0	0
full GAW17	24,487	5,789	1,905	0		505	152
TGen	115,059						
Kamboh-small	352,693						
Kamboh-large	787,586				505		1,905
NAPS	191	82	74	82		63	67

5.4 VARIANT DELETERIOUSNESS SCORE CHARACTERIZATION

For use as informative priors, I examined six different variant deleteriousness scores including two scores that are based on predicting changes in protein structure and function (SIFT and PolyPhen), three on estimating conservation (PhyloP, GERP and Phast), and one that combined multiple scores (CADD). The availability of these scores for the SNVs in the datasets is quite sparse, as shown in Table 11. Overall, the CADD score was available for more SNVs than any other score. I did not use every score for every dataset due to availability and capacity of annotation servers.

Given the sparsity of variant deleteriousness scores, I estimated a prior probability for SNVs that did not receive a score; this probability was set to the expected number of SNVs for the disease of interest divided by the number of SNVs in the dataset.

5.5 RESULTS FROM THE FULL GAW17 DATASET

The results from sections 5.1 to 5.3 support better performance of BAPR over BAPR-FT. Hence, I ran BAPR on the full GAW17. The full GAW17 dataset contains 24,487 SNVs and 6,970 individuals. The target variable was the Q2 risk factor that was converted to a binary variable. In the following sections, Q2 is modeled with 72 SNVs on 13 genes. I report the results from biomarker recovery rate plots followed by prediction performance with AUCs.

The BAPR algorithm was run with *uniform* and binomial priors (including *bin enp*, *bin Polyphen*, *bin SIFT*, *bin GERP*, *bin Phast*, *bin CADD*, and *bin max 4*). When a deleteriousness value was not available for a SNV, it was assigned a prior probability of 0.001. In addition, two comparison algorithms were run (*chi-square* and *CADD*).

5.5.1 Biomarker recovery

Figure 18 shows the biomarker recovery plots for all SNVs in the dataset, while Figure 19, shows the biomarker recovery plots for the top-ranked 25% SNVs. Table 11 gives the p values from chi-square testing comparing the performance of algorithms in a pairwise fashion at the 95^{th} percentile on the biomarker recovery plots. At the 95^{th} percentile (or top 5%) there are 1,250 SNVs and at this threshold 3.6 causal SNVs are likely to be discovered by chance.

CADD has poor performance, with no causal SNVs recovered until after the first 700 SNVs and only 3 causal SNVs recovered at the 95^{th} percentile. *Chi-square* performs better than *CADD*, recovering 20 SNVs, along with *bin CADD*, recovering 20 SNVs, and *bin SIFT*, recovering 22 SNVs. *Bin GERP* starts poorly, recovering only 4 SNVs by rank 750; however, by the 90^{th} percentile it performs equivalently to *chi-square*, *bin CADD*, and *bin SIFT*. *Uniform* performs better than *chi-square*, *bin CADD*, and *bin SIFT*, recovering 32 SNVs at the 95^{th} percentile. *Bin enp*, *bin Phast*, and *uniform* perform the best at the 95^{th} percentile. The performance of the combination method *bin max 4* is similar to that of *bin Phast*.

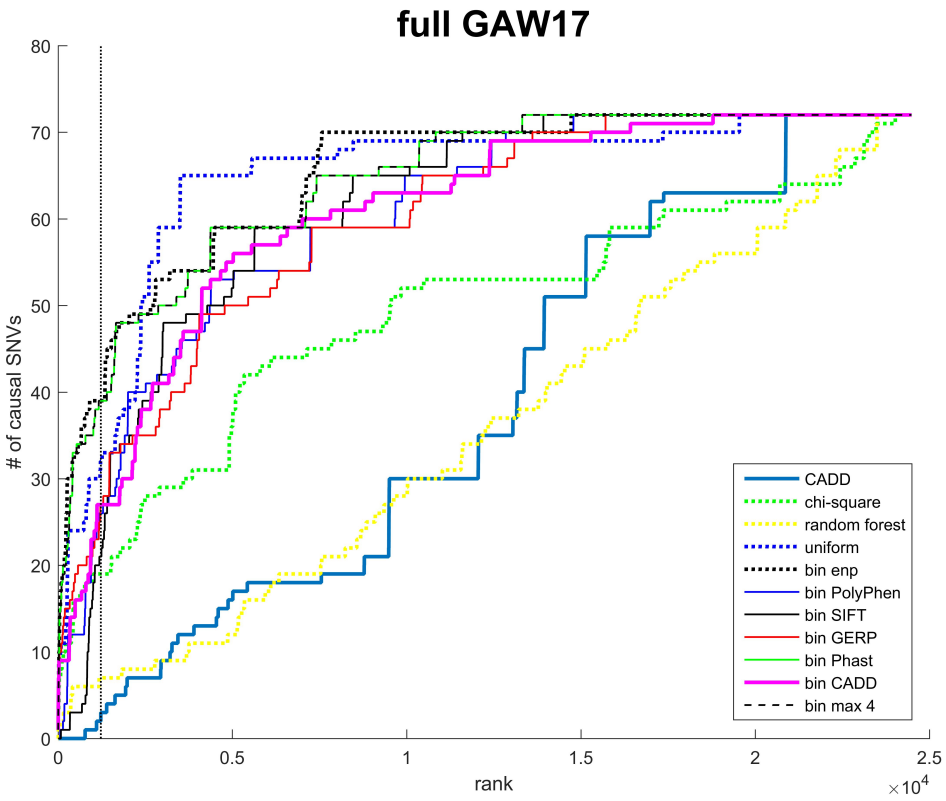


Figure 18: Biomarker recovery plots for the full GAW17 dataset.

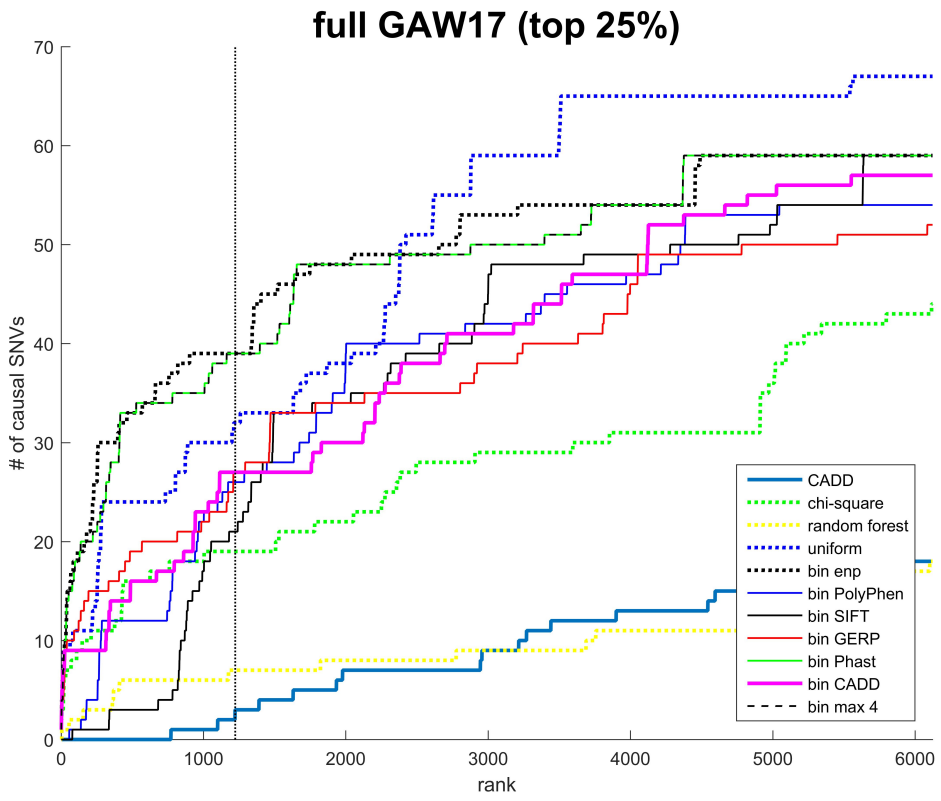


Figure 19: Biomarker recovery plots for the full GAW17 dataset for the top-ranked 25% SNVs.

Table 12: P-values from chi-square test comparing the performance of algorithms in a pair-wise fashion on the full GAW17 dataset at the 95^{th} percentile on the biomarker recovery plots, continued in Table 13.

	chi-square	**uniform**	**bin enp**	**bin PolyPhen**	**bin SIFT**
CADD	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
chi-square		**0.012**	**0.002**	0.592	0.755
uniform			0.589	0.113	**0.023**
bin enp				**0.042**	**0.007**
bin PolyPhen					0.422

Table 13: P-values from chi-square test comparing the performance of algorithms in a pair-wise fashion on the full GAW17 dataset at the 95^{th} percentile on the biomarker recovery plots, continued from Table 12.

	bin GERP	**bin Phast**	**bin CADD**	**bin max 4**
CADD	**<0.001**	**<0.001**	**<0.001**	**<0.001**
chi-square	0.655	**0.008**	0.964	**0.008**
uniform	0.097	0.914	**0.048**	0.914
bin enp	**0.035**	0.679	**0.016**	0.679
bin PolyPhen	0.933	**0.047**	0.642	**0.047**
bin SIFT	0.432	**0.002**	0.712	**0.002**
bin GERP		**0.037**	0.701	**0.037**
bin Phast			**0.038**	1.000
bin CADD				**0.010**

5.5.2 Prediction performance

The prediction performance of the kNN classifier as measured by AUCs is given in Table 14. With increasing number of features from 1 to 1,000, the AUCs increase slightly until 500 features and then decrease at 1,000 features. Overall, the mean AUCs are in the 0.50 range and all algorithms perform similarly.

Table 14: Mean AUCs with standard errors obtained from the kNN classifier on the Kamboh-large exome dataset.

	1	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
chi-square	0.502 ±0.015	0.502 ±0.015	0.506 ±0.015	0.506 ±0.015	0.502 ±0.015	0.500 ±0.015	0.492 ±0.015	0.510 ±0.015
uniform	0.490 ±0.015	0.516 ±0.015	0.512 ±0.015	0.493 ±0.015	0.488 ±0.015	0.504 ±0.015	0.505 ±0.015	0.504 ±0.015
bin enp	0.496 ±0.015	0.493 ±0.015	0.500 ±0.015	0.500 ±0.015	0.508 ±0.015	0.496 ±0.015	0.497 ±0.015	0.498 ±0.015
bin GERP	0.496 ±0.015	0.506 ±0.015	0.501 ±0.015	0.499 ±0.015	0.512 ±0.015	0.493 ±0.015	0.512 ±0.015	0.501 ±0.015
bin Phast	0.491 ±0.015	0.505 ±0.015	0.503 ±0.015	0.493 ±0.015	0.488 ±0.015	0.501 ±0.015	0.508 ±0.015	0.492 ±0.015
bin PolyPhen	0.499 ±0.015	0.500 ±0.015	0.509 ±0.015	0.484 ±0.015	0.476 ±0.015	0.506 ±0.015	0.505 ±0.015	0.489 ±0.015
bin SIFT	0.499 ±0.015	0.508 ±0.015	0.506 ±0.015	0.502 ±0.015	0.493 ±0.015	0.502 ±0.015	0.510 ±0.015	0.495 ±0.015
bin max 4	0.511 ±0.015	0.506 ±0.015	0.494 ±0.015	0.499 ±0.015	0.499 ±0.015	0.487 ±0.015	0.514 ±0.015	0.498 ±0.015

5.6 RESULTS FROM THE KAMBOH-LARGE EXOME DATASET

The Kamboh-large dataset is an exome dataset that consists of 787,586 SNVs and 584 individuals. The target variable is LOAD. In the following sections, I report the results of known disease-associated biomarker recovery with biomarker recovery rate plots followed by prediction performance with AUCs. I obtained a list of 2,524 known LOAD-associated SNVs from the AlzGene database, of which 776 SNVs were present in the Kamboh-large dataset. This list of 776 SNVs was used in the biomarker recovery plots.

The BAPR algorithm was run with *uniform* and binomial priors (including *bin enp*, *bin SIFT*, *bin PhyloP*, and *bin max 2*). When a deleteriousness value was not available for a SNV, it was assigned a prior probability of 0.001. In addition, two comparison algorithms were run (*chi-square* and *CADD*).

5.6.1 Biomarker recovery

Figure 20 shows the biomarker recovery plots for all SNVs in the dataset while Figure 21 shows the biomarker recovery plots for the top-ranked 25% SNVs. At the 95^{th} percentile (or top 5%), there are 39,392 SNVs. Table 15 gives the p values from chi-square tests comparing the performance of algorithms in a pairwise fashion at the 95^{th} percentile on the biomarker recovery plots.

Chi-square performs significantly worse than all other algorithms and at the 95^{th} percentile, it only recovers 28 SNVs. *Uniform* is one of the highest performing algorithms and ties in performance with *bin enp*.

For the two algorithms with prior knowledge using PhyloP and SIFT, *bin PhyloP* performs significantly worse than *bin SIFT*. *Bin SIFT* performs significantly better than *chi-square* and *bin PhyloP*, even performing better than *bin enp*. *Bin max 2*, which is a combination prior, has similar performance to *bin SIFT*. Overall, the best performing algorithms, not significantly different from one other, are *uniform*, *bin SIFT*, *bin enp*, and *bin max 2*.

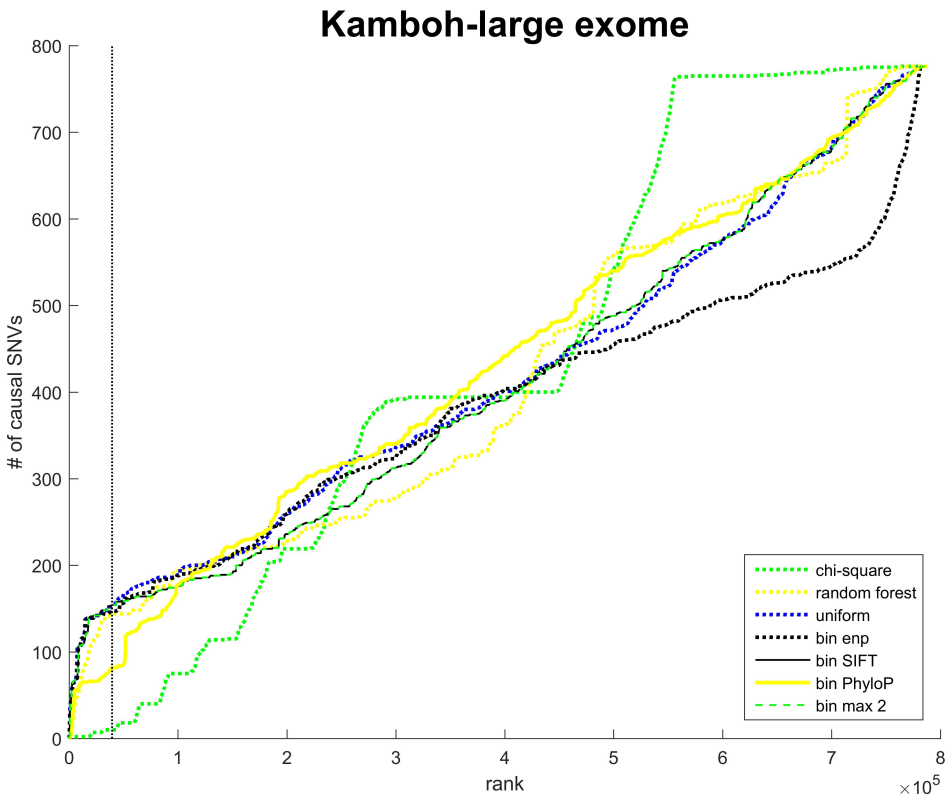


Figure 20: Biomarker recovery plots for the Kamboh-large dataset.

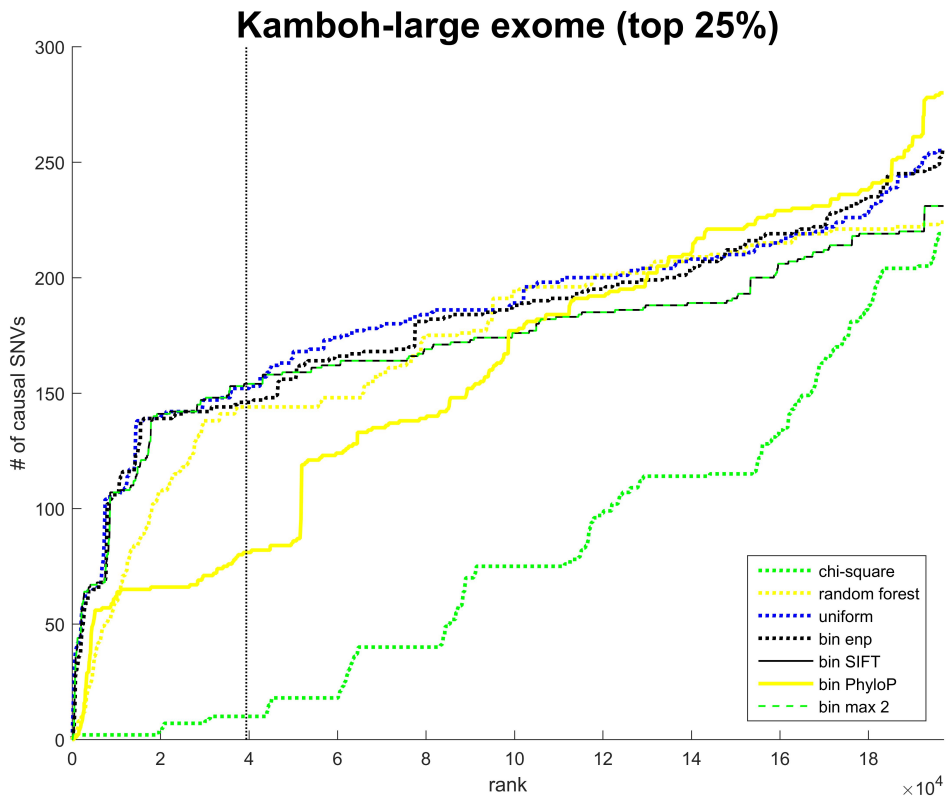


Figure 21: Biomarker recovery plots for the Kamboh-large dataset for the top-ranked 25% SNVs.

Table 15: P-values from chi-square test comparing the performance of algorithms in a pair-wise fashion on the Kamboh-large exome dataset at the 95^{th} percentile on the biomarker recovery plots.

	uniform	**bin enp**	**bin SIFT**	**bin PhyloP**	**bin max 2**
chi-square	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
uniform		0.324	0.710	**<0.001**	0.710
bin enp			0.522	**<0.001**	0.522
bin SIFT				**<0.001**	1.000
bin PhyloP					**<0.001**

5.6.2 Prediction performance

The prediction performance of the kNN classifier as measured by AUCs is given in Table 16. With increasing number of features from 1 to 1,000, the AUCs increase slightly until 500 features and then decrease at 1,000 features. The highest AUC that was achieved is 0.62. Overall, the mean AUCs are in the 0.53 range and most algorithms perform similarly.

5.6.3 Evidence of biological validity

The top 20 SNVs ranked by *bin enp* spanned 8 genes and had 8 rsIDs, as shown in Table 17. The 7^{th} ranked SNV, *rs429358*, is a known SNP that is highly associated with Alzheimer's disease.

Five of the eight recovered genes are associated with Alzheimer's disease. Two genes, APOE and SORL1, are known to be highly associated with Alzheimer's disease. Of the remaining six genes, three (GCN1L1, FAM163A, and ATP8A2) are associated with neurologic disease.

Table 16: Mean AUCs with standard errors obtained from the kNN classifier on the Kamboh-large exome dataset.

	1	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
chi square	0.638 ±0.044	0.564 ±0.046	0.555 ±0.046	0.518 ±0.046	0.580 ±0.045	0.549 ±0.046	0.479 ±0.046	0.623 ±0.044
uniform	0.509 ±0.046	0.482 ±0.046	0.507 ±0.046	0.505 ±0.046	0.563 ±0.046	0.591 ±0.045	0.510 ±0.046	0.584 ±0.045
bin enp	0.579 ±0.046	0.533 ±0.046	0.484 ±0.046	0.496 ±0.046	0.571 ±0.046	0.539 ±0.046	0.527 ±0.046	0.529 ±0.046
bin SIFT	0.499 ±0.046	0.521 ±0.046	0.512 ±0.046	0.529 ±0.046	0.519 ±0.046	0.509 ±0.046	0.539 ±0.046	0.528 ±0.046
bin PhyloP	0.490 ±0.046	0.535 ±0.046	0.481 ±0.046	0.467 ±0.046	0.555 ±0.046	0.547 ±0.046	0.507 ±0.046	0.515 ±0.046
bin max 2	0.520 ±0.046	0.440 ±0.046	0.517 ±0.046	0.558 ±0.046	0.563 ±0.046	0.496 ±0.046	0.500 ±0.046	0.523 ±0.046

Table 17: Top ranked 20 SNVs and rsIDs by *bin enp* on the Kamboh-large exome dataset. Genes in bold are known to be associated with Alzheimer's. Genes in italics are known to be associated with neurocognitive diseases.

rank	**rsID**	**gene**	**chr**	**position**
1	rs199901349	*GCN1L1*	12	120612953
2		*GCN1L1*	12	120632299
3		WIBG	12	56320843
4	rs202244671	*FAM163A*	1	179783095
5		WIBG	12	56321555
6	rs201822155	*ATP8A2*	13	26153950
7	rs9553698	*ATP8A2*	13	26586788
8	rs429358	**APOE**	19	45411941
9		**APOE**	19	45409590
10	rs61945823	*GCN1L1*	12	120613484
11		ZSCAN18	19	58600105
12		WIBG	12	56320827
13	rs72240167	**SORL1**	11	121457406
14		**SORL1**	11	121322793
15	rs367790148	ZSCAN18	19	58596112
16		PRC1	15	91537757
17		PRC1	15	91517745
18		*GCN1L1*	12	120572093
19		*GCN1L1*	12	120607971
20		*GCN1L1*	12	120613792

5.7 RESULTS FROM THE NAPS EXOME DATASET

The NAPS dataset is an exome dataset that consists of 191 SNVs and 2,201 individuals. The target variable is chronic pancreatitis. In the following sections, I report the results of known disease-associated biomarker recovery with biomarker recovery rate plots followed by prediction performance with AUCs. I obtained a list of 37 known pancreatitis-associated SNVs from the DisGeNET database, of which 12 SNVs were present in the dataset. This list of 12 SNVs was used in the biomarker recovery plots.

5.7.1 Biomarker recovery

Figure 22 shows the biomarker recovery plots for all SNVs in the dataset while Figure 23 shows the biomarker recovery plots for the top-ranked 25% SNVs.

All algorithms retrieved 1 SNV (8.3%) at the 95th percentile, slightly above random expectation of 0.62 SNVs. Using only *CADD* for ranking gave the worst performance, recovering 1 causal SNV at the 90^{th} percentile before staying slightly ahead of random expectation after the 85^{th} percentile with 2 causal SNVs, and finally recovering 8 of 12 causal SNVs by the 50^{th} percentile. *Chi-square* performed better than *CADD*, recovering 3 causal SNVs at the 85^{th} percentile, 5 causal SNVs at the 75^{th} percentile, and 10 causal SNVs at the 50^{th} percentile. All of the BAPR methods - *uniform*, *bin enp*, *bin PolyPhen*, *bin SIFT*, *bin GERP*, *bin Phast*, *bin CADD*, and *bin max 4* - perform equally well, and all outperform *chi-square* and recover 5 causal SNVs at the 85^{th} percentile, 7 causal SNVs at the 75^{th} percentile, and 10-11 causal SNVs by the 50^{th} percentile.

5.7.2 Prediction performance

The prediction performance of the kNN classifier as measured by AUCs is given in Table 18. With an increasing number of features from 1 to 200, the AUCs increase slightly until 50 features and then decrease at 200 features. The highest AUC that was achieved is 0.55. Overall, the mean ACUs are in the 0.50 range and most algorithms perform similarly.

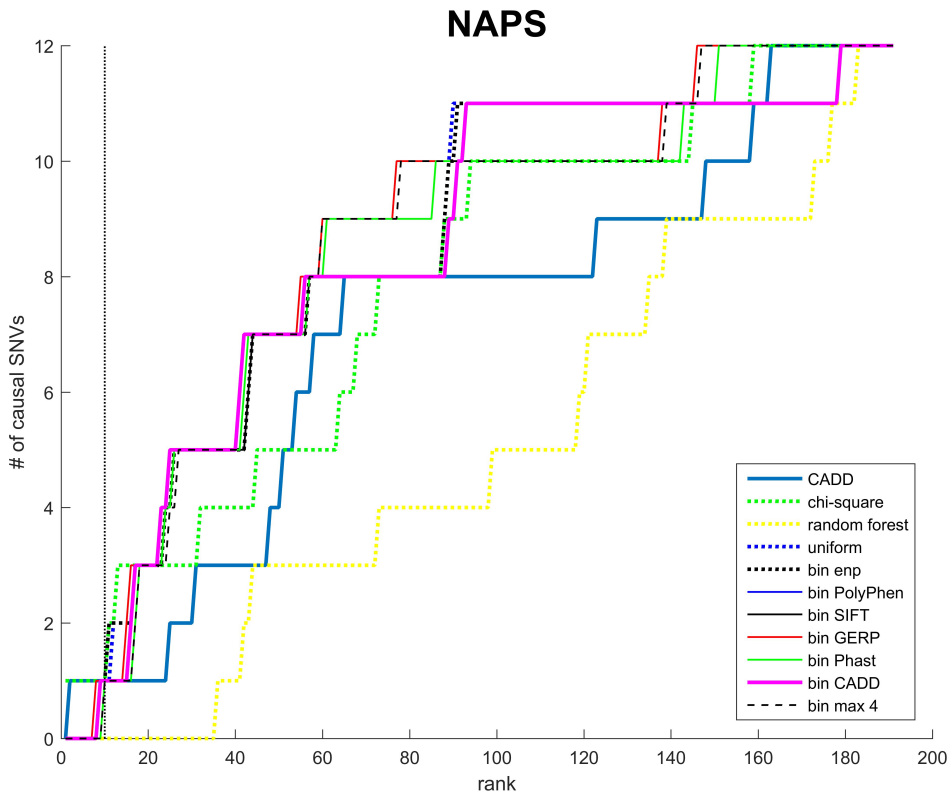


Figure 22: Biomarker recovery plots for the NAPS dataset.

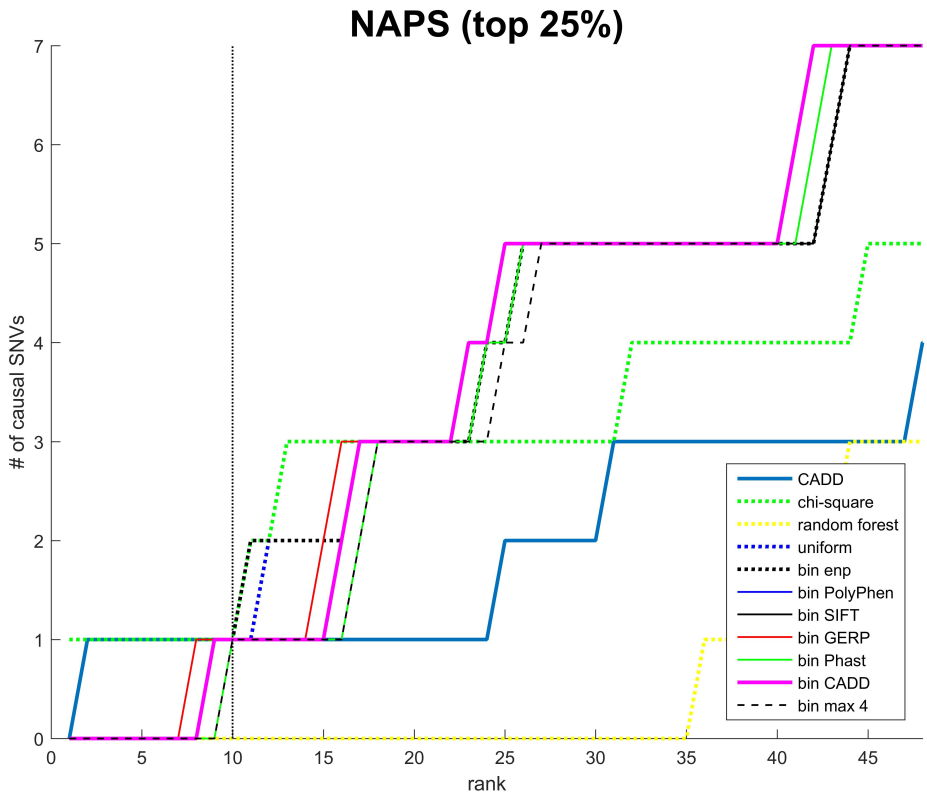


Figure 23: Biomarker recovery plots for the NAPS dataset for the top-ranked 25% SNVs.

Table 18: Mean AUCs with standard errors obtained from the kNN classifier on the NAPS exome dataset.

	1	**2**	**5**	**10**	**50**	**100**	**200**
chi square	0.523 ±0.029	0.541 ±0.029	0.505 ±0.029	0.509 ±0.029	0.536 ±0.029	0.501 ±0.029	0.491 ±0.029
uniform	0.520 ±0.029	0.525 ±0.029	0.474 ±0.029	0.524 ±0.029	0.495 ±0.029	0.447 ±0.029	0.425 ±0.029
bin enp	0.518 ±0.029	0.497 ±0.029	0.507 ±0.029	0.481 ±0.029	0.510 ±0.029	0.463 ±0.029	0.414 ±0.029
bin PolyPhen	0.519 ±0.029	0.503 ±0.029	0.526 ±0.029	0.528 ±0.029	0.492 ±0.029	0.510 ±0.029	0.447 ±0.029
bin SIFT	0.507 ±0.029	0.538 ±0.029	0.512 ±0.029	0.506 ±0.029	0.480 ±0.029	0.475 ±0.029	0.404 ±0.029
bin GERP	0.494 ±0.029	0.509 ±0.029	0.517 ±0.029	0.492 ±0.029	0.492 ±0.029	0.477 ±0.029	0.439 ±0.029
bin Phast	0.498 ±0.029	0.502 ±0.029	0.498 ±0.029	0.504 ±0.029	0.490 ±0.029	0.488 ±0.029	0.409 ±0.029
bin max 4	0.507 ±0.029	0.506 ±0.029	0.519 ±0.029	0.515 ±0.029	0.492 ±0.029	0.468 ±0.029	0.423 ±0.029

Table 19: Runtimes across datasets.

	genes	**SNVs**	**runtime per iteration**
NAPS	19	191	1.31 sec
full GAW17	3,205	24,487	2.17 min
Kamboh-large	21,585	787,586	108.12 min

5.8 RUNTIMES

These experiments were conducted on Amazon Web Services' Elastic Compute Cloud (EC2). These experiments were conducted on a c4.4xlarge instance (16 virtual CPUs, 30 GB RAM, 1.5TB SSD). Runtimes are noted in Table 19.

6.0 CONCLUSIONS AND FUTURE WORK

In this dissertation, I developed, implemented, and evaluated a new multivariate biomarker ranking algorithm called BAPR. The BAPR algorithm has a combination of several novel characteristics including (1) learning probabilistic rule models from data, (2) performing Bayesian model averaging to rank biomarkers like SNVs, and (3) incorporating biological knowledge as structure priors of biomarkers. I applied the BAPR algorithm with a variety of variant deleteriousness scores as priors to several exome datasets with both synthetic outcomes and real outcomes. The performance was evaluated with biomarker recovery plots, AUCs, and evidence of biological validity. The BAPR algorithm almost always performed better than chi-square, the comparison algorithm. Moreover, the use of prior knowledge also improved the performance of BAPR. A summary of 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 contributions and findings of the work presented in this dissertation.

6.1.1 BAPR model structure and search strategy

I first evaluated BAPR on semi-synthetic data, GWAS data, and a small exome dataset to determine whether BAPR-FT and BAPR performed better the chi-square. I also evaluated the efficiency of BAPR-FT and BAPR under various search strategies.

I needed to demonstrate algorithm efficacy first before extending it. Semi-synthetic data was a natural starting point for showing algorithmic efficacy, as the causal biomarkers are known. On the semi-synthetic data, BAPR performed significantly better at biomarker recovery when compared to chi-square.

Search strategies significantly influence the utility of an algorithm in practice. Exhaustive search is more expensive and often intractable in comparison to heuristic search, but is guaranteed to find the global optimum. On a small dataset, greedy search performed quite well compared to exhaustive search. I also explored the richer model space of tree structures compared to the smaller model space of global structures. The tree structure space provided better performance. Based on these results, the BAPR algorithm that was applied to high-dimensional exome datasets performed greedy search and used tree models.

6.1.2 BAPR structure priors

Informative structure priors that I investigated were of two types: one prior was based on the expected number of predictors associated with a disease of interest, and the other was a group of priors that were derived from variant deleteriousness scores. Variant deleteriousness scores assess the deleteriousness of SNVs based on predicting changes in structure or function of the relevant protein or the degree of nucleotide conservation at the locus.

The results obtained from experiments on the full GAW 17 and the Kamboh-large datasets were similar. BAPR performed better than CADD for all priors, and BAPR also performed better than chi-square for most priors at the 95th percentile on biomarker recovery plots. Moreover, BAPR with priors *uniform* and *enp* performed better than priors using variant deleteriousness scores.

One observation is that variant deleteriousness scores are available only for a small proportion of the SNVs that are measured in exome studies. Even with the sparse availability, the incorporation of scores improves biomarker recovery. Moreover, it seems that scores with more coverage like CADD often provide better performance than scores with less coverage. These encouraging result suggest that as the coverage of these scores increases the performance of biomarker discovery algorithms like BAPR that incorporate them will improve.

An additional observation is that *chi-square*, a method that is solely data driven, performs better than using just a variant deleteriousness score like CADD. The reason for this is likely due to the sparse coverage of CADD, since when a CADD score is not available for a casual SNV the SNV will not be identified.

6.1.3 Combining variant deleteriousness scores for priors

Variant deleteriousness scores have sparse coverage and often individual scores cover very different sets of SNVs. This observation led to the idea of combining all scores using the *max* function. The intuition is that this will increase coverage since a SNV will be assigned a value even if one of the scores provides a value. However, a disadvantage of this simplistic function is that it ignores the fact that the same value across several scores likely does not denote the same probability of deleteriousness, and sometimes scores may disagree strongly on the degree of deleteriousness at a locus.

Results obtained from the full GAW17 and Kamboh-large datasets indicate that the max prior performs at least as well as the best performing single deleteriousness score. However, it usually performs worse than *uniform* and *enp* priors.

6.2 FUTURE WORK

The experimental work presented in this research explored the application of one version of BAPR for biomarker discovery. Several extensions of the BAPR algorithm as directions for future work are possible.

6.2.1 Alternate genetic models

The current BAPR algorithm assumes a genotype model where each of three genotypes at a SNV is treated as an independent value. Examples of other common genetic models include recessive, dominant, and additive models. These models, in general, arise because each locus has two alleles, one on each of two homologous chromosomes. Consider the typical case

where the possible alleles at a locus are A (wildtype) and a (variant). A recessive genetic model assumes that having two copies of the a allele will result in disease. A dominant genetic model assumes that having one or more copies of the a allele will result in disease. An additive models implies that the effect on the risk of disease of having two copies of the a allele doubles when compared to having one copy of the a allele. The BAPR algorithm can be modified to handle any of these genetic models and even search among the various models. This is an interesting area of future investigation.

6.2.2 Alternate model averaging strategies

As mentioned in Section 3.8.3, the model averaging I used is limited by inflated scoring for variables included in the model. Future work can further investigate using techniques such as bagging and simulated annealing to reduce the effects of this bias.

6.2.3 Alternate approaches for combining scores for priors

The main combination prior that I investigated was the max function that assigned the highest deleteriousness score from four scoring methods. A method that combines *enp* with multiple deleteriousness scores is a promising direction of investigation. Also, using alternate ways of combining scores is another avenue for future work.

6.2.4 Using pathway information for search

For reasons of computational efficiency, BAPR partitions SNVs such that each set of SNVs is associated with a gene, and then searches for high-scoring PR models to average over in each set. An extension to this search involves a two stage approach. In the first stage, BAPR searches over gene-related SNVs as it currently does. In the proposed second stage, it obtains high-ranking SNVs from the first stage and partitions them such that each set of SNVs is associated with a pathway, and then searches for high-scoring PR models to average over in each set. Since biological pathways consist of a group of proteins (and hence corresponding genes) that perform a specific function, organizing gene-related SNVs into

pathways provides a natural grouping of SNVs. Moreover, pathway information is readily available from numerous online databases that organize genes into pathway-specific lists.

BIBLIOGRAPHY

[1] I. A. Adzhubei, S. Schmidt, L. Peshkin, et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7:248–249, 2010.

[2] L. Almasy, T. D. Dyer, J. M. Peralta, et al. Genetic analysis workshop 17 mini-exome simulation. In *BMC Proceedings*, volume 5, page 1. BioMed Central, 2011.

[3] D. Altshuler, M. J. Daly, and E. S. Lander. Genetic mapping in human disease. *Science*, 322:881–888, 7 Nov. 2008.

[4] A. Annest, R. Bumgarner, A. Raftery, and K. Y. Yeung. Iterative bayesian model averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics*, 10:72, 2009.

[5] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. Van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[6] M. J. Bamshad, S. B. Ng, A. W. Bigham, et al. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12:745–755, 2011.

[7] V. Bansal, O. Libiger, A. Torkamani, and N. J. Schork. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11:773–785, 2010.

[8] N. H. Barton and M. Turelli. Evolutionary quantitative genetics: how little do we know? *Annual Review of Genetics*, 23:337–370, 1989.

[9] L. G. Biesecker. Exome sequencing makes medical genomics a reality. *Nature Genetics*, 42:13–14, 2009.

[10] W. Bodmer and C. Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40:695–701, 2008.

[11] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

[12] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in bayesian networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 115–123, Portland, OR, 1996. Morgan Kaufmann Publishers Inc.

[13] M. G. Bulmer. The effect of selection on genetic variability. *American Naturalist*, 105:201–211, 1971.

[14] W. Buntine. Theory refinement on bayesian networks. In *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pages 52–60, Los Angeles, CA, 1991. Morgan Kaufmann Publishers Inc.

[15] W. S. Bush and J. H. Moore. Chapter 11: Genome-Wide association studies. *PLoS Computational Biology*, 2012.

[16] M. Cargill, D. Altshuler, J. Ireland, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22:231–238, 1999.

[17] A. M. Carvalho. Scoring functions for learning bayesian networks. *Inesc-id Tec. Rep*, 2009.

[18] C. Chelala, A. Khan, and N. R. Lemoine. Snpnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5):655–661, 2009.

[19] M. Choi, U. I. Scholl, W. Ji, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, 106:19096–19101, 2009.

[20] Consortium, Encode Project, E. Birney, J. A. Stamatoyannopoulos, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[21] Consortium, Encode Project, I. Dunham, A. Kundaje, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[22] G. F. Cooper and E. Herskovits. A bayesian method for constructing bayesian belief networks from databases. In *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pages 86–94, Los Angeles, CA, 1991. Morgan Kaufmann Publishers Inc.

[23] G. M. Cooper, D. L. Goode, S. B. Ng, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods*, 7(4):250–251, 2010.

[24] G. M. Cooper, E. A. Stone, G. Asimenos, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913, 2005.

[25] E. V. Davydov, D. L. Goode, M. Sirota, et al. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Computational Biology*, 6(12):e1001025, 2010.

[26] A. Z. Dayem Ullah, N. R. Lemoine, and C. Chelala. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Research*, 40:W65–W70, 1 July 2012.

[27] dbSNP. Short genetic variations. http://www.ncbi.nlm.nih.gov/projects/SNP/, 2012.

[28] EGAPP. Recommendations from the EGAPP working group: genomic profiling to assess cardiovascular risk to improve cardiovascular health. *Genetics in Medicine*, 12:839–843, 2010.

[29] R. A. Fisher. *The genetical theory of natural selection.* Clarendon Press, Oxford, England, 1930.

[30] S. E. Flanagan, A. M. Patch, and S. Ellard. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers*, 14:533–537, 2010.

[31] N. Friedman and M. Goldszmidt. Learning bayesian networks with local structure. In M. Jordan, editor, *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, pages 421–459. Springer Netherlands, 1998.

[32] S. Ghosh, H. Bickeböller, J. Bailey, et al. Identifying rare variants from exome scans: the gaw17 experience. In *BMC Proceedings*, volume 5, page S1. BioMed Central Ltd, 2011.

[33] G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13:135–145, 2012.

[34] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos. An application of random forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics*, 11(1):1, 2010.

[35] I. P. Gorlov, O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. Shifting paradigm of association studies: Value of rare Single-Nucleotide polymorphisms. *American Journal of Human Genetics*, 82:100–112, 2008.

[36] L. W. Hahn, M. D. Ritchie, and J. H. Moore. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, 19(3):376–382, 2003.

[37] D. L. Hartl and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[38] D. Heckerman, D. Geiger, and D. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1 Sept. 1995.

[39] K. Hemminki, A. Försti, and J. L. Bermejo. The common Disease-Common variant hypothesis and familial risks. *PLoS One*, 3:e2504, 2008.

[40] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.

[41] M. J. Hubisz, K. S. Pollard, and A. Siepel. Phast and rphast: phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, 12(1):41–51, 2011.

[42] T. Ideker, J. Dutkowski, and L. Hood. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, 144(6):860–863, 2011.

[43] E. Jablonka and G. Raz. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology*, 84:131–176, 2009.

[44] A. C. J. W. Janssens and C. M. van Duijn. Genome-based prediction of common diseases: advances and prospects. *Human Molecular Genetics*, 17:R166–R173, 15 Oct. 2008.

[45] J. Kang, W. Zheng, L. Li, et al. Use of bayesian networks to dissect the complexity of genetic disease: application to the genetic analysis workshop 17 simulated data. *BMC Proceedings*, 5:S37, 2011.

[46] R. A. King, R. I. Rotter, and A. G. Motulsky. *The Genetic Basis of Common Diseases*. Oxford University Press, 2nd edition, 2002.

[47] M. Kircher, D. M. Witten, P. Jain, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310, 2014.

[48] A. Kong, V. Steinthorsdottir, G. Masson, et al. Parental origin of sequence variants associated with complex diseases. *Nature*, 462:868–874, 2009.

[49] M. R. Kosorok, W.-H. Wei, and P. M. Farrell. The incidence of cystic fibrosis. *Statistics in Medicine*, 15:449–462, 1996.

[50] M. Krawczak, E. V. Ball, I. Fenton, et al. Human gene mutation database - a biomedical information and research resource. *Human Mutation*, 15:45–51, 2000.

[51] G. V. Kryukov, L. A. Pennacchio, and S. R. Sunyaev. Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *American Journal of Human Genetics*, 80:727–739, 2007.

[52] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4:1073–1081, 2009.

[53] H. Lango Allen, K. Estrada, G. Lettre, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832–838, 2010.

[54] J. A. Lee, R. E. Madrid, K. Sperle, et al. Spastic paraplegia type 2 associated with axonal neuropathy and apparent plp1 position effect. *Annals of Neurology*, 59(2):398–403, 2006.

[55] S. Lee, M. J. Emond, M. J. Bamshad, et al. Optimal unified approach for Rare-Variant association testing with application to Small-Sample Case-Control Whole-Exome sequencing studies. *American Journal of Human Genetics*, 91:224–237, 2012.

[56] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*, 83:311–321, 2008.

[57] M.-X. Li, H.-S. Gui, J. S. Kwan, S.-Y. Bao, and P. C. Sham. A comprehensive framework for prioritizing variants in exome sequencing studies of mendelian diseases. *Nucleic Acids Research*, 40(7):e53,gkr1257, 2012.

[58] C.-T. Liu, H. Lin, and H. Lin. Functional analysis of HapMap SNPs. *Gene*, 511:358–363, 2012.

[59] T. F. C. Mackay. The genetic architecture of quantitative traits. *Annual Review of Genetics*, 35:303–339, 2001.

[60] T. F. C. Mackay, E. A. Stone, and J. F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10:565–577, 2009.

[61] D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 1994.

[62] N. Matsumoto and N. Niikawa. Kabuki make-up syndrome: A review. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 117C:57–65, 2003.

[63] M. L. Metzker. Sequencing technologiesthe next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[64] R. E. Mills, C. T. Luttig, C. E. Larkins, et al. An initial map of insertion and deletion (indel) variation in the human genome. *Genome Research*, 16(9):1182–1190, 2006.

[65] R. E. Neapolitan. *Learning Bayesian networks*. Prentice Hall, Upper Saddle River [etc.], 2004.

[66] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Research*, 11:863–874, 1 May 2001.

[67] S. B. Ng, K. J. Buckingham, C. Lee, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42:30–35, 2009.

[68] S. B. Ng, E. H. Turner, P. D. Robertson, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461:272–276, 2009.

[69] NHGRI. A catalog of published Genome-Wide association studies. http://www.genome.gov/gwastudies/, 2012.

[70] K. Patterson. 1000 GENOMES: A world of variation. *Circulation Research*, 108:534–536, 4 Mar. 2011.

[71] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[72] E. Pennisi. 1000 genomes project gives new map of genetic diversity. *Science*, 330:574–575, 29 Oct. 2010.

[73] J. Piñero, N. Queralt-Rosinach, À. Bravo, et al. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, 2015.

[74] K. Pradhan, S. Yoon, T. Wang, and K. Ye. Identification of genes and variants associated with quantitative traits using bayesian factor screening. *BMC Proceedings*, 5:S4, 2011.

[75] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease–common variant or not? *Human Molecular Genetics*, 11:2417–2423, 1 Oct. 2002.

[76] V. Pungpapong, L. Wang, Y. Lin, D. Zhang, and M. Zhang. Genome-wide association analysis of GAW17 data using an empirical bayes variable selection. *BMC Proceedings*, 5:S5, 2011.

[77] S. Purcell, B. Neale, K. Todd-Brown, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[78] M. A. Quintana, F. R. Schumacher, G. Casey, et al. Incorporating prior biologic information for High-Dimensional rare variant association studies. *Human Heredity*, 74:184–195, 2012.

[79] V. Ramensky, P. Bork, and S. Sunyaev. Human nonsynonymous SNPs: server and survey. *Nucleic Acids Research*, 30:3894–3900, 1 Sept. 2002.

[80] R. Raynes. The central dogma of biology. http://www.rachelraynes.com/the-central-dogma-of-biology/.

[81] E. M. Reiman, J. A. Webster, A. J. Myers, et al. Gab2 alleles modify alzheimer's risk in apoe 4 carriers. *Neuron*, 54(5):713–720, 2007.

[82] M. Saad, A. Pierre, N. Bohossian, M. Mace, and M. Martinez. Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. *BMC Proceedings*, 5:S33, 2011.

[83] Y. Saletore, K. Meyer, J. Korlach, et al. The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biology*, 13:1–12, 4 Nov. 2012.

[84] C. T. Saunders and D. Baker. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of Molecular Biology*, 322:891–901, 2002.

[85] A. V. Segrè, Consortium, DIAGRAM, M. Investigators, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics*, 6(8):e1001058, 2010.

[86] D. Shriner, A. Adeyemo, N. P. Gerry, et al. Transferability and Fine-Mapping of Genome-Wide associated loci for adult height across human populations. *PLoS One*, 4:e8398, 2009.

[87] A. Siepel and D. Haussler. Phylogenetic hidden markov models. In *Statistical methods in molecular evolution*, pages 325–351. Springer, 2005.

[88] A. Siepel, K. Pollard, and D. Haussler. New methods for detecting Lineage-Specific selection. In A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, and M. Waterman, editors, *Research in Computational Molecular Biology*, volume 3909 of *Lecture Notes in Computer Science*, pages 190–205. Springer Berlin Heidelberg, 2006.

[89] K. S. Small, A. K. Hedman, E. Grundberg, et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics*, 43:561–564, 2011.

[90] M. J. Somerville, C. B. Mervis, E. J. Young, et al. Severe expressive-language delay related to duplication of the williams–beuren locus. *New England Journal of Medicine*, 353(16):1694–1701, 2005.

[91] E. K. Speliotes, C. J. Willer, S. I. Berndt, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42:937–948, 2010.

[92] M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10:681–690, 2009.

[93] N. O. Stitziel, A. Kiezun, and S. Sunyaev. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology*, 12:227, 2011.

[94] G. Streisinger, Y. Okada, J. Emrich, et al. Frameshift mutations and the genetic code. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 31, pages 77–84. Cold Spring Harbor Laboratory Press, 1966.

[95] J. K. Teer and J. C. Mullikin. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*, 19:R145–R151, 15 Oct. 2010.

[96] J. R. Thompson, M. Gogele, C. X. Weichenberger, et al. SNP prioritization using a bayesian probability of association. *Genetic Epidemiology*, 2012.

[97] A. Z. D. Ullah, N. R. Lemoine, and C. Chelala. A practical guide for the functional annotation of genetic variations using snpnexus. *Briefings in Bioinformatics*, 14(4):437–447, 2013.

[98] Ü. Väli, M. Brandström, M. Johansson, and H. Ellegren. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*, 9(1):1, 2008.

[99] S. Visweswaran. *Learning patient-specific models from clinical data*. ProQuest, 2007.

[100] S. Visweswaran, D. C. Angus, M. Hsieh, et al. Learning patient-specific predictive models from clinical data. *Journal of Biomedical Informatics*, 43:669–685, 2010.

[101] J. Vomlel. Noisy-or classifier. *International Journal of Intelligent Systems*, 21:381–398, 2006.

[102] K. Wang, M. Li, and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 2010.

[103] K. M. Waters, D. O. Stram, M. T. Hassanein, et al. Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genetics*, 6:e1001078, 2010.

[104] W. Wei, S. Visweswaran, and G. F. Cooper. The application of naive bayes model averaging to predict alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, 2011.

[105] Wellcome Trust Case Control, Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[106] N. R. Wray, S. M. Purcell, and P. M. Visscher. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biology*, 9:e1000579, 2011.

[107] M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[108] A. Yan, N. Laird, and C. Li. Identifying rare variants using a bayesian regression approach. *BMC Proceedings*, 5:S99, 2011.

[109] J. Yang, B. Benyamin, B. P. McEvoy, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010.

[110] K. Yeung, R. Bumgarner, and A. Raftery. Bayesian model averaging: Development of an improved Multi-Class, gene selection and classification tool for microarray data. *Bioinformatics*, 21:2394–2402, 2005.

[111] K. Y. Yeung. *A Practical Guide to Bioinformatics Analysis*. Bayesian Model Averaging for Biomarker Discovery from Genome-Wide Microarray Data. CreateSpace, 2010.

[112] J. Zhao and A. Thalamuthu. Gene-based multiple trait analysis for exome sequencing data. *BMC Proceedings*, 5:S75, 2011.

[113] J. Zhu, B. Zhang, E. N. Smith, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 2008.

[114] Q. Zhu, D. Ge, J. M. Maia, et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American Journal of Human Genetics*, 88:458–468, 2011.