In silico Prioritization of Genetic Risk Variants Using Functional Genomic Information

by

Sarah A Gagliano

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Institute of Medical Science
University of Toronto

In silico Prioritization of Genetic Risk Variants Using Functional Genomic Information

Sarah A Gagliano

Doctor of Philosophy

Institute of Medical Science
University of Toronto

2016

Abstract

Complex traits are the result of a contribution of both genetic risk variants throughout the genome, and environmental risk factors and their interactions. Genome-wide association studies (GWAS) have identified some of these associated variants, but there remain two fundamental issues to move forward in understanding the genetic etiology of complex traits: (1) The "missing heritability" for complex traits persists, possibly in part due to lack of statistical power as a result of insufficiently large sample sizes (2) The identity of the causal variant- a variant identified by GWAS could result in a functional consequence, or it could merely tag the causal variant. I hypothesize that integrating functional information, such as chemical modifications to DNA, along with statistical data from an association study can help prioritize variants for further analysis in both of these areas. I developed a method to prioritize genetic variants using hundreds of functional annotations (Gagliano et al., 2014a) using penalized logistic regression. I compared my prioritization method to two other methods that use data-trained classifiers to determine if there is an ideal algorithm or annotation set for prioritizing risk variants (Gagliano et al., 2015a).

In this work, I also investigated using different databases of disease-associated variants to define genetic risk variants. The models created all had some accuracy for detecting risk variants. I assessed the accuracy of these models using measures investigated in a review I undertook (Gagliano et al., 2015b). Finally, I investigated if allele-specific methylation (ASM) is a useful novel annotation to prioritize risk variants. I demonstrated that variants that exhibit ASM in brain tissue are enriched for functional annotations, and are also enriched in sub-genome-wide significant variants in a schizophrenia GWAS.

Acknowledgments

I would not have had the opportunity to conduct the research described in this thesis without my supervisors: Drs. Jo Knight and James L Kennedy from the Centre for Addiction and Mental Health (CAMH), and so my thanks and gratefulness goes out first and foremost to them.

I also thank the members of my Program Advisory Committee and my mentors for the CIHR Strategic Training in Genetic Epidemiology (STAGE) program, Drs. Andrew Paterson, Arturas Petronis and Cathy Barr, for all of their insight and assistance throughout my graduate studies.

I also acknowledge the funding I received throughout my graduate education. I would like to thank the donors of the Peterborough K.M. Hunter Graduate Studentship (whom I had the privilege of meeting) the Armstrong Family (via the CAMH Foundation), and the CIHR STAGE– CIHR Training Grant in Genetic Epidemiology and Statistical Genetics.

I thank those at CAMH with whom I worked, including postdoctoral fellows, professors, fellow students, and other staff, all of whom have helped make my experience memorable.

Contributions

This thesis consists of four original research studies. Chapter 3 is published in *PLoS ONE* (Gagliano et al., 2014a), Chapter 4 is published in *BMC Genomics* (Gagliano et al., 2015b), Chapter 5 is published in *Scientific Reports* (Gagliano et al., 2015a), and Chapter 6 has been submitted.

The author performed all experiments described in the thesis, except as noted below:

Chapter 3

David Kevans and Mira Ryten defined the phenotype-specific lists. For the R script used to create the Manhattan plot (Figure 3.3), the wrapper function was written by Michael Weale, and the internal "wgplot" function was written by Matt Settles.

Chapter 5

Reena Ravji helped prepare and run the Python code for the random forest and support vector machine algorithms.

Chapter 6

Carolyn Ptak performed the genotyping and all wet laboratory procedures. Denise Mak performed the piecewise linear regression to identify the single nucleotide polymorphisms (SNPs) exhibiting allele-specific methylation (ASM), and constructed the plots in Figure 6.3. Denise Mak also performed the principal components analysis (PCA) to determine ancestry in the samples and constructed the PCA plot (Figure 6.4).

I would also like to thank my PhD final oral examination defense examiners: Angelo Canty, Anna Goldenberg and Michael Hoffman. Finally, I would like to acknowledge the reviewers of my papers for their useful comments, suggestions, and ideas.

Table of Contents

List of Tables

List of Figures

Abbreviations

1KG- 1000 Genomes Project
Affy6- Affymetrix SNP 6.0 genotyping microarray
APOE- apolipoprotein E
ASM- Allele-specific methylation
AUC- Area Under the (receiver operating characteristic) Curve
BF_{annot}- Bayes Factor for Annotation
BF_{assoc}- Bayes Factor for Association
bp- base pair
BPD- bipolar disorder
CADD- Combined Annotation Dependent Depletion
ChiP-seq- Chromatin Immunoprecipitation, followed by sequencing
chr- chromosome
ChroMoS - Chromatin Modified SNPs
CMC- collapsing multivariate and collapsing
CNV- copy number variation
dbGAP- database of Genotypes and Phenotypes
DNA- deoxyribonucleic acid
DNAse I- DNase I hypersensitive sites
Duke- Duke University
EBI- European Bioinformatics Institute
EFO- Experimental Factor Ontology
ENCODE- Encyclopedia of DNA Elements
eQTLs- Expression Quantitative Trait Loci
FAIRE- Formaldehyde-Assisted Isolation of Regulatory Elements
FDR- false discovery rate
Gencode Txnstart- Transcription start sites as defined by Gencode
GERP- Genome Evolutionary Rate Profiling
GIANT- Genetic Investigation of Anthropomorphic Traits
GM12878- a lymphoblastoid cell line; a tier 1 cell line from the ENCODE Project
GNL3- guanine nucleotide binding protein-like 3
GTEx- Genotype-Tissue Expression
GWAS- Genome-wide association study
GWAVA- Genome Wide Annotation of VAriants
H1-hESC- embryonic stem cells; a tier 1 cell line from the ENCODE Project
H3K4Me1- monomethylation of the fourth lysine of histone protein H3
H3K4Me3- trimethylation of the fourth lysine of histone protein H3

H3K27Ac- acetylation of the twenty-seventh lysine of histone protein H3
hits- risk variants (variants associated with a complex trait)
hg19- human genome build 19
HGMD- Human Gene Mutation Database
HS- hypersensitive (i.e. as in DNase I HS)
ICBP- International Consortium for Blood Pressure
IQR- Interquartile range
K- 1000
K562- a leukemia cell line; a tier 1 cell line from the ENCODE Project
LD- Linkage DisequilibriummiRNA- micro-RNA
mRNA- messenger RNA
MSRE- Methylation Specific Restriction Enzymes
NCBI- National Center for Biotechnology Information
NHGRI- National Human Genome Research Institute
NIH- National Institutes of Health
NPV- Negative Predictive Value
PAINTOR- Probabilistic Annotation INTegratOR
PGC- Psychiatric Genomics Consortium
PhastCons- Evolutionary conservation measure
Phevor- Phenotype Driven Variant Ontological Re-ranking tool
PhyloP- Evolutionary conservation measure
PPV- Positive Predictive Value
PRKAR2B- Protein kinase cAMP-dependent regulatory type II beta gene
PWL- Piecewise Linear Regression
RNA- ribonucleic acid
ROC- Receiver Operating Characteristic
SBP- Systolic Blood Pressure
SCZ- schizophrenia
sFDR- stratified False Discovery Rate
SIFT- Sorting Intolerant From Tolerant
SilVA- Silent Variant Analyzer
SKAT- sequence kernel association test
SNP- Single Nucleotide Polymorphism
SVM- Support Vector Machine
TFBS- Transcription Factor Binding Sites
Txn- Transcription
UCSC- University of California Santa Cruz
UW- University of Washington
VCF- Variant Call Format
VEP- Variant Effect Predictor

VIM- Variable Importance Measure

Chapter 1
Why and How to Prioritize Genetic Risk Variants using Functional Information?

1.1 Lay Summary

Humans are largely identical in our DNA sequence, but about 5% of the genome, containing genetic differences or genetic variants, is a contributing factor as to why we look different, and these variants partially explain why some people develop an illness while others do not. A minority of these genetic variants falls into regions in our DNA sequence that encode proteins and other molecules important for cellular function (genes). Many of the genetic variants fall into known regulatory regions where they may work in controlling or regulating gene function. The genetic variants that are harmful (increase our risk of developing an illness) or are protective (reduce our risk of developing an illness) are called genetic risk variants.

Since it is difficult to differentiate risk variants from all variants based on current techniques, I developed a computer algorithm to do so based on their regulatory and other genomic information. My method (Gagliano et al., 2014a) was published around the same time as two other methods, but they use different computer algorithms and different regulatory information. I decided to determine the best combination of computer algorithm and regulatory information that most accurately predicts genetic risk variants (Gagliano et al., 2015a). I found that there are several combinations that offer some accuracy, but there is still a lot of room for improvement. In order to improve my method, I refined it to examine a subset of genetic risk variants: those specifically involved in mental health disorders. I also explored a new piece of regulatory information: chemical modifications to the DNA that differ between alleles at heterozygous sites. This new piece of information shows good potential for identifying novel risk variants because the variants that exhibit this quality fall into known regulatory regions significantly more than expected by chance. Identifying genetic risk factors helps in earlier diagnosis and better treatment options for a range of diseases.

1.2 The human genome

It has long been known that genetic material, deoxyribonucleic acid (DNA), plays a role in determining the phenotype or the manifestation of observed characteristics (Race et al., 1949).

DNA is a double helical structure with a sugar-phosphate backbone (Watson and Crick, 1953) composed of two complementary strands containing a sequence of four nucleotides (adenine, guanine, cytosine and thymine). Human DNA consists of about three billion base pairs, and is organized into chromosomes that are stored in the nucleus of each cell in the human body (Alberts et al., 2007). Some sections are transcribed into messenger RNA (mRNA) by the use of enzymes and regulatory factors (e.g. transcription factors). The mRNA is then transported to the cytosol of the cell where it is translated into a chain of amino acids to create a protein. Three mRNA bases (which make up a "codon") translate to one amino acid (and there is redundancy in this genetic code, meaning that there is more than one codon that translates to the same amino acid). Stop codons cause the translational machinery to stop translating. For further information about transcription and translation see this *Nature Education* review (Clancy and Brown, 2008). DNA also encodes for non-coding RNA molecules (i.e. DNA does not encode for only protein), such as micro-RNA (translation regulation), small nuclear RNAs (involved in splicing) and small nucleolar RNAs (involved in ribosomal RNA modification) (Eddy, 2001; Mattick and Makunin, 2006). The DNA is wrapped around proteins called histones (two proteins each of H2a, H2b, H3, H4), which have an effect on DNA conformation, and consequently the accessibility of the DNA sequence to regulatory factors and other proteins (McGhee and Felsenfeld, 1980). Regulatory factors will be discussed in detail later in **Sections 1.8 to 1.13**, inclusive.

1.3 Variation in the human genome

A change in nucleotide could alter the function of the stretch of DNA, and may contribute to an observed characteristic. The different variations possible at a position (locus) are called alleles. In our nuclear DNA, humans have two of each chromosome, one from the father and one from the mother (apart from sex chromosomes). The set of alleles carried across a set of loci on either the paternal or maternal chromosome are called a haplotype (Griffiths et al., 2008). If the allele at one locus is known, then the allele present at a nearby locus can often be inferred; this non-random association of alleles at different loci is called linkage disequilibrium (LD) (Reich et al., 2001). During the formation of gametes (egg or sperm) there is the crossing over of homologous chromosomes, resulting in the exchange of DNA segments between the two chromosomes (Griffiths et al., 2008). Different regions of the genome have different crossover frequencies. Areas of high crossover are called recombination hotspots (Petes, 2001). In stretches of DNA where there is a low crossover frequency the alleles at different alleles tend to segregate together through multiple generations and hence are in high LD. As a consequence there are a limited number of haplotypes within each region (Griffiths et al., 2008). Haplotypes and allele frequencies differ depending on the ancestral population.

The completion of the initial draft sequence of the human genome in 2001 (International Human Genome Sequencing Consortium et al., 2001) provided researchers with a map of the DNA sequence, but since then mapping human variation is still being refined. Many genotyping (e.g. HapMap Project (Altshuler et al., 2010)), and sequencing (e.g. 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010)) projects in various human populations have been possible as the price for such technologies decreases. These large-scale projects provide insight into DNA variation, the frequencies of these variants, and LD patterns throughout the genome in various world populations.

HapMap was conducted in three phases. Phase 1 investigated common variants (minor allele frequency >5%) in individuals from three populations genotyping at least one common SNP every 5 kilobases across the genome (The International HapMap Consortium, 2005). Phase 2 genotyped a small number of individuals (n=270) from only four human populations (Frazer et al., 2007). Phase 3 (HapMap3) provided the opportunity to look at low frequency (rare) variants (e.g. minor allele frequency <5%) in addition to common variants by genotyping over one thousand individuals. HapMap3 mapped 1.6 million variants in 1,184 reference individuals from 11 populations (Altshuler et al., 2010).

1000 Genomes too has three phases and has been able to identify common and rare variants throughout the human genome in diverse populations. Phase 1 came out in 2012, and there were several versions of this phase published to refine the genotypes. The data consisted of low-coverage whole-genome and high-coverage exome sequencing. This phase is comprised of 1,092 individuals from 14 human populations across the globe with a mean read depth of 5.1 times for over 37 million autosomal sites (1000 Genomes Project Consortium et al., 2012). Phase 2 was primarily for methods development, and there was no public release. Phase 3 came out in 2014, and it assessed 2,535 individuals from a total of 26 world populations (details from the 1000 Genomes Project website: http://www.1000genomes.org/faq/what-do-pilot-project-phase-1-phase-2-and-phase-3-mean). Although many more variants were called in Phase 3 than Phase 1, 2.3 million variants in Phase 1 were not in Phase 3 but these were either very low frequency or low quality calls so may have been false positives in Phase 1. (More details on the differences between these two phases are available from the 1000 Genomes Project website: http://www.1000genomes.org/category/frequently-asked-questions/phase-3.)

Through these projects, genotyping and quality control procedures were further refined for looking at variation.

The type of genetic variant that has been the focus in the HapMap and 1000 Genomes Projects has been the single nucleotide polymorphism (SNP): at a single base position in the DNA sequence there can be a different DNA nucleotide base that is present depending on the individual. The 1000 Genomes Project has also investigated indels: small insertions and deletions (Mullaney et al., 2010), microsatellites, CNVs, and structural variants (Sudmant et al., 2015; Zarrei et al., 2015).

In the coding regions of the genome, there are different types of changes that could occur depending on the location of the SNP in the sequence (see **Box 1**), which can explain why such a variant may alter the phenotype (Griffiths et al., 2008). For variation in non-coding regions in the genome, the biological explanation resulting in an altered phenotype could be due to the SNP falling within the DNA binding site for a protein or other regulatory signatures or functions such as splicing (more details in **Section 1.8**).

Box 1. **Types of alterations in the coding regions of the genome.**

Synonymous- the change that does not alter the amino acid sequence (due to the redundancy in the genetic code). However a proportion of synonymous changes could still have an effect on the protein. For instance, a synonymous mutation could disrupt a splice site, or it could alter mRNA folding.

Nonsynonymous- the change that does alter the amino acid sequence. There are a few types, and the phenotypic effect of the alteration depends on protein structure and function. A missense change occurs in a protein, and the effect on the protein depends on how similar (for instance, charge or hydrophobicity) the new amino acid is from the one it is in the wild type protein. A nonsense change creates a premature stop codon, and the effect depends at which point the premature stop codon is inserted. If earlier on in the amino acid chain, often the more devastating the alteration is to the protein's structure and thus function.

Splice- a change in a site in the DNA sequence involved in splicing out introns

Frameshift- an insertion or deletion that shifts the three base-pair reading frame, thus altering the string of amino acids translated

Another type of variation is copy number variants (CNVs), which involve a different number of a set of ordered bases in the sequence. CNVs arise either *de novo* (meaning not preset in either parent, but present in the progeny) or are inherited (Wain et al., 2009). However, the focus of this thesis will be on SNPs.

All humans contain genetic variants. For instance, the 1000 Genomes Projects identified around 38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000

larger deletions in their Phase 1 data (n=1,092) (1000 Genomes Project Consortium et al., 2012). This Consortium found that on average, an individual carries approximately 250 to 300 loss of function variants in genes and 50 to 100 variants that are previously implicated in inherited disorders (The 1000 Genomes Project Consortium, 2010).

1.4 The role of genetics in disease

The initial understanding of the genetic contributions to traits dates back to Gregor Mendel (Mendel, 1866). Mendel bred pea plants to obtain desired traits from a series of binary outcomes, such as smooth or wrinkled peas, long or short stems, and axial or terminal flowers (Weir, 1990). Of course, at the time of Mendel it was not known that variants in the DNA sequence were the causes leading to these particular traits. Furthermore, Mendel had only been experimenting with single-gene traits or disorders (which in the human context would include traits such as blood groups (Race et al., 1949)). Such disorders present in the simplest case as a variant within a gene that results in an alternate form of the protein, leading to a phenotype that deviates from the wild-type phenotype (Antonarakis and Beckmann, 2006). The vast majority of human traits do not follow such a simplistic mode of transmission.

Complex traits (e.g. height, blood pressure, schizophrenia, adverse drug response, for example) are the result of genetics (possibly many variants in multiple genes and in intergenic regions) (Lango Allen et al., 2010; Ehret et al., 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Ozeki et al., 2011) and environmental factors (Leask, 2004; Sinclair, 1989; Pickering, 1997; Vesell, 1991), as well as possibly their interactions. Heritability is the proportion of the variance that can be attributed to genetic variation; further details in the following review (Tenesa and Haley, 2013). Humans contain millions of genetic variants, but not all are genetic *risk* variants, or in other words are associated with a disease or trait. There are variants that increase one's risk of developing a trait (or disease), and there are also variants that are protective, meaning they decrease one's risk of developing a disease. The known

disease/trait-associated variants do not account for all of the heritability (Manolio et al., 2009).

Heritability does not pinpoint the genetic architecture of the disease, for instance the number and/or types of DNA variation involved, and the frequency of those variants. Heritability can be determined through twin-studies (Boomsma et al., 2002). It is important to keep in mind that there are assumptions and limitations of twin-study determined heritability. These studies assume that the environments are similar for both twins in a pair, which is not necessarily true, but more importantly within pair environment similarity is similar for monozygotic and dizygotic twins. Such studies also assume an additive model of inheritance at a locus, and thus do not take into account other models such as dominance (which need multi-generation family studies) or epistatic effects (interactions among multiple genes), for example (Neale, 1992).

One hypothesis describing the effect of variants on a complex trait is the common disease common variant hypothesis (Reich and Lander, 2001). One samples a large number of individuals, some of whom are affected with the disease of interest (cases) and others who are not (controls). Given the hypothesis, one can identify those variants that are common enough in the population to be detected as statistically significant: variants that have a genotype appearing more often in the cases compared to the controls or vice versa. Variants detected by this procedure tend to have low or moderate effect sizes.

Another hypothesis is the common disease rare variant hypothesis (Schork et al., 2009). The idea is that the disease results from rare variants (for instance, variants with the minor allele appearing in less than 1% of the sample). Such variants are thought to have high effect sizes (high penetrance), but they can also have more moderate effects.

Likely, the genetic component attributed to complex diseases is a result of both common and rare variants both with varying effect sizes, in addition to other factors such as gene-

environment interactions. Researchers have used a variety of techniques to find those variants.

1.5 Identification of genetic variants involved in disease

Identifying associated variants among all variants is important for advances in medical care (Manolio, 2013). Knowledge of the variants results in information about the role of genes, and pathways in disease, which can provide mechanistic insight. This information ultimately can help with diagnosis, and in personalizing treatment (for example, using genetic information to improve the selection of medication that is most likely to not have negative side effects and/or is most likely to be effective in treating symptoms).

There has been an evolution of methods employed to identify the genetic variants that modify (increase/decrease) one's risk of developing a complex trait as technologies and methodologies have developed.

Linkage studies were conducted using family data (for example, Lathrop et al., 1984). Alleles on one chromosome co-segregate together with another allele on another chromosome with 50% probability. Alleles on the same chromosome co-segregate at a rate related to the distance between them on the chromosome: the recombination fraction. Two loci are linked when the recombination fraction is less than one half. A trait was said to be linked to a locus if the recombination fraction was less than half (assessed through parametric studies) (Terwilliger and Ott, 1994). Non-parametric studies were developed for complex traits and include quantitative trait linkage studies which correlate sharing of chromosomal segments among relatives with their similarity for a given trait (Purcell et al., 2003). These studies identify broad regions making it difficult to pinpoint precise locations in the genome that are associated with the outcome (phenotype) of interest. Genome-wide scans were initially conducted using microsatellite markers and restriction fragment length polymorphisms (for example, (Rice et al., 2000)).

To refine the resolution of the detected associated loci, studies were then carried out comparing the frequencies of alleles or genotypes in a set of unrelated individuals with the trait/disease of interest (cases) and a set of individuals without that particular phenotype (controls) in particular genes. Alternatively, family-based association methods can also be used (Ott et al., 2011). Keeping in mind the costs associated with genotyping, rather than interrogating variants throughout the entire genome, variants in a subset of genes were assessed. These candidate gene association studies are hypothesis-driven association studies where genes with potential biological evidence, for instance for possible association with the phenotype, are selected. Variants in those regions are tested for association with the phenotype in a sample of individuals (Tabor et al., 2002). These studies look at correlations between genotype and a phenotype. There can be relatively simple biological rationale to implicate variants in genes as disease-causing. A variant that produces a different amino acid or stop gain or stop loss could affect the protein structure and thus function, and contribute to the observed characteristic. Unfortunately, few significant findings identified through candidate gene studies have replicated in larger samples, suggesting that most candidate gene study findings may be spurious (Hart et al., 2013). However, one of the few examples of a gene that came up as associated to a phenotype that has been replicated in many larger genome-scan studies was the association between the epsilon4 haplotype of the apolipoprotein E (*APOE*) gene and Alzheimer's disease (Combarros et al., 2002; Lambert et al., 2013).

As microarray technology developed, genome-wide association studies (GWAS) became increasingly popular (and less expensive) in the mid/late 2000s and the genome was able to be interrogated by genotyping individuals at variants present on genotyping arrays. GWAS have been successful in identifying risk variants for complex diseases and traits (for example, (Jostins et al., 2012; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014)), but much of the heritability is still unaccounted for. GWAS are association studies where variants throughout the genome are tested (using regression for instance) one-by-one for an association with the trait of interest. Since

GWAS do not only assess associations between SNPs in candidate genes and the phenotype of interest, associations between never-before implicated loci and disease can be detected. For instance, through GWAS, for Crohn's disease, loci in genes involved in autophagy have been discovered, and it is because of these studies that it is now understood that autophagy plays a role in Crohn's (Xavier et al., 2008).

The Wellcome Trust Case Control Consortium set the standards for sample size and analysis pipelines. They identified risk variants for seven common diseases using over 14,000 cases and a set of shared controls (Wellcome Trust Case Control Consortium, 2007). Quality control procedures for both SNPs (e.g. genotyping rate, Hardy-Weinberg equilibrium) and individuals (e.g. population stratification) are important (Anderson et al., 2010). This latter point relates to the importance of ensuring that a homogenous population is used in GWAS because the association of a SNP with the trait of interest may be confounded by that SNP being associated with ancestral differences between the cases versus the controls (Anderson et al., 2010). These procedures and the association analysis can be conducted in tools such as PLINK (Purcell et al., 2007).

GWAS interrogate variants on a genotyping array platform, and are useful for identifying common variants. Such array platforms are offered by several companies including Affymetrix and Illumina. Through projects such as HapMap and through technological advances, the arrays have been updated. For example, more variants have been added to new arrays over the years. The variants on the arrays have been selected largely because of their LD correlation with many other variants, and thus able to cover a vast amount of the genome; these variants are not necessarily chosen because they are likely to have functional consequences (Edwards et al., 2013). Additionally, there are specialized arrays for investigating a subset of traits (for example: Barrans and Liew, 2006; Cortes and Brown, 2011; Voight et al., 2012). These specialized arrays contain customized content informative for the trait of interest, such as SNPs in or close to genes that are likely candidates for the disease.

Procedures and software for imputing the genotypes at other variants have been developed. Such software take advantage of LD patterns in the genome in reference samples, and examples include Impute2 (Howie et al., 2009) and Minimac2 (Fuchsberger et al., 2014). In imputation, missing genotypes are estimated based on haplotypes from a cosmopolitan population. Imputation is useful for combining samples that were genotyped on different arrays as well as for fine-mapping signals at an associated locus (Verbeek et al., 2012). Moreover, imputation can also be used to investigate low-frequency and rare variants at a genome-wide level; for instance see Surakka et al. (2015) where they imputed in over 62,000 samples to identify novel loci involved in lipid levels.

As mentioned, GWAS arrays mainly contain common variants. More recently, sequencing has become cheaper and faster (through technological advances). Whole-genome (or whole-exome) sequencing interrogates the genome (or the exons of genes: the exome) more thoroughly than genotyping arrays, including the less frequent (rare) variants. There can be low power due to small sample size to detect associations with less frequent variants. In order to address these issues, in addition to testing single-variants for association with the phenotype, several gene-based (or region-based) tests have been developed such as the combined multivariate and collapsing (CMC) method (Li and Leal, 2008), C-alpha (Neale et al., 2011), and sequence kernel association test (SKAT) (Wu et al., 2011). CMC is a burden test, whereas C-alpha and SKAT are non-burden. Burden tests collapse rare variants in a defined region into a single burden variable (Lee et al., 2012), whereas non-burden tests do not. Burden tests work best when the variants themselves are responsible for disease risk (i.e. not just tagging the variant resulting in the effect because they are in high LD with each other) and all influence risk in the same direction, whereas non-burden tests are more flexible, having the power to detect the effects of variants whether increasing risk or protective. There is evidence supporting the impact of rare variants in many complex diseases and traits ranging from neurodevelopmental disorders such as autism (Krumm et al., 2015) to lipid levels (Surakka et al., 2015). Similar to GWAS, for sequencing too, there have been some novel

associated variants identified (Cirulli et al., 2015; Sanders et al., 2012), but a large proportion of risk variants still remain undiscovered due to small sample sizes, variants with small effects, or a focus on the coding sequence, for instance.

1.6 Characteristics of disease-associated variants

As more and more variants that predispose individuals to disease have been identified, efforts have been made to share this knowledge with the scientific community.

In the literature there is no consistent term used to describe risk variants; different terms all have some nuances. MacArthur et al. (2014) differentiates between pathogenic variants (those that contribute mechanistically to the disease that may not be alone sufficient to cause the disease) from damaging (those that result in altered levels or function of a gene or gene product, but may not have a pathogenic effect), for example. Regardless of more specific categorization, these variants may be able to be used to partially predict risk of disease in the individuals that carry them.

There are several databases that report genetic risk variants. One example of a database includes the National Human Genome Research Institute (NHGRI)-European Bioinformatics Institute (EBI) Genome-wide association study (GWAS) Catalogue (Hindorff et al., 2010)), which catalogues genetic variants from a GWAS. Another example of a database is the Human Gene Mutation Database (HGMD) (Stenson et al., 2009). It reports variants for all known genetic mutations responsible for causing classes of human inherited diseases from the peer-reviewed literature. ClinVar (Landrum et al., 2014), another database, reports relationships between medically important variants (variants that result in a health-related phenotype) and phenotypes. HGMD and ClinVar largely contain SNPs, but they are not restricted to this type of variation; for instance they contain insertions, deletions and repeat variations as well. (See **Box 2** for more details on these databases.)

Box 2. Databases of Genetic Risk Variants.

GWAS Catalogue This Catalogue started in 2010 as a manually curated collection from the literature of variants associated with complex diseases or traits that looked at a minimum of 100,000 SNPs in the initial stage. The Catalogue moved to a new website through the European Bioinformatics Institute (EBI) in March 2015: http://www.ebi.ac.uk/gwas/. It contains variants from GWAS studies with a combined p-value $<1.0x10^{-5}$ (discovery plus replication populations), and studies are excluded if they were restricted to just candidate genes, not published in the English language, if samples were to assess somatic mutations (e.g. tumor samples), or if the study does not include any new GWAS data. Information is extracted from PubMed searches using terms "genome-wide" OR "genome AND identification" OR "genome AND association", with limits on the current year and human status.

HGMD Available at http://www.biobase-international.com/product/hgmd, there is a public (free) and professional (paid) version. The public version is less up to date and provides less information on the variants (for instance, neither chromosome number and base position nor rsID). The database was first made publically available in 1996. It was first established to catalogue variants in human genes that cause inherited disease, but has since been expanded to germ-line disease-related functional variants (Stenson et al., 2009). It reports mutations for all known gene lesions responsible for causing human inherited disease from the peer-reviewed literature.

ClinVar The database (at http://www.ncbi.nlm.nih.gov/clinvar/) does not include unreviewed data from GWAS studies, but accepts variants identified through clinical testing and literature curation.

There are design differences among the databases. For example, variants in the GWAS Catalogue are explicitly not necessarily the disease-causing variants. Furthermore, the Catalogue includes variants that are not just associated with diseases per se (also with complex traits: for example height and platelet count among many others). With regard to HGMD, the variants in the database have been included based on multiple (and vastly different) lines of evidence. For instance, some have evidence of direct functional relevance, while others are predicted to alter the length of a resulting gene-product but there is no reported disease association (Stenson et al., 2009). What is more, there is not necessarily 100% penetrance of the variants, and there is an inherent bias to variants found in genes (because originally the database was created to study mutational mechanisms in human genes). As for ClinVar, variants are correlated with the trait in a clinical sample, but there is not necessarily 100% penetrance. Different clinical labs often have different opinions on the clinical significance of the same genetic variant (Rehm et al., 2015). Variants can be inputted into the database if evidence of causality is seen in a sample of one, such as from a clinical testing lab (Landrum et al., 2014).

The difference in design leads to fundamental differences between the variants in the GWAS Catalogue and HGMD (and ClinVar), such as minor allele frequency. HGMD variants have significantly lower minor allele frequencies compared to the GWAS Catalogue variants (**Figure 1.1**).

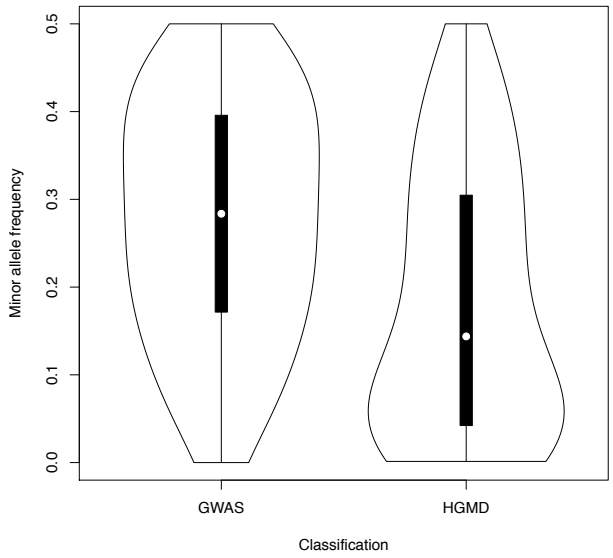


Figure 1.1. Violin plots depicting minor allele frequency distributions for GWAS Catalogue versus HGMD variants

GWAS = autosomal variants present in the GWAS Catalogue (with $p<5 \times 10^{-8}$) downloaded August 7, 2014 (n=3,618); HGMD= autosomal variants in the HGMD database as of the 4th quarter of 2013 provided to Ensembl that are found with an rsID identifier in the 1000 Genomes Project (n=4,862). (Note that HGMD variants without chromosomal and base position information provided were not considered.) Minor allele frequencies were obtained from the European population of the Phase 1, version 3 of the 1000 Genomes Project (n=379). The violin plot shows the density distribution of the variants, and the summary statistics presented in a box plot. The density is shown by the smooth lines that make up the "body", and the box plot is the black box inside the "body". The white dot is the median, and the box outlines the 25% and 75% percentiles. The lower and upper whiskers on the plot represent the 25% percentile minus 1.5*IQR and the 75% percentile plus 1.5*IQR, respectively. If the data does not extend as far as those calculated ranges, then the whisker is plotted at the value of the minimum or maximum data point. [IQR= interquartile range]

Variants in these two databases differ with regard to position: GWAS Catalogue variants are vastly non-exonic (>70%), whereas HGMD variants are vastly exonic (~70%).

However, there are some similarities. The GWAS Catalogue variants and HGMD variants shown in **Figure 1.1** fell into 1,510 (42% genes/number of SNPs) and 1,835 (38% genes/number of SNPs) RefSeq genes, respectively, and of those genes 308 were in common. However, there is nearly no overlap between the actual variants in the GWAS Catalogue with either HGMD or also with ClinVar pathogenic variants, likely due to the frequency of the variants in the GWAS Catalogue compared to the latter two.

Databases of variants have been used in various papers in order to define genetic risk variants. In my work described later in this thesis (**Chapter 3**) (Gagliano et al., 2014a), in my best performing models I defined risk variants as those variants present in the GWAS Catalogue with an association p-value lower than the accepted threshold for genome-wide significance, 5×10^{-8} (Pe'er et al., 2008). Iversen et al. (2014) also used variants from the GWAS Catalogue, regardless of their association p-value, but confined to studies that used an Affymetrix and/or Illumina array. Moving away from GWAS, Ritchie et al. (2014) was specifically interested in regulatory variants, and defined such variants as those present in the public version of HGMD that are regulatory (n=1,614). They used variants labelled as pathogenic from ClinVar that do not overlap with HGMD as a validation of their tool, called GWAVA. Shihab et al. (2015) also used variants in HGMD.

The above briefly highlights that current databases of risk variants have different characteristics and overlap with functional annotations with different frequencies. The implications of these differences will be considered further in the analysis of **Chapter 5** and in the discussion, **Chapter 7**.

1.7 Gap in variant identification with GWAS

There are many as yet uncharacterized risk variants. There are still two points surrounding the detection of disease-associated variants from GWAS that my thesis will aim to address:

(1) Still undiscovered loci (i.e. "missing heritability") (Manolio et al., 2009)
(2) Causal variant identification (i.e. GWAS-implicated loci comprise of multiple variants in high LD, identifying which variant(s) in the locus is disease causing: "causal"/"functional"/directly influencing the phenotype)?

What needs to be done (applicable to both of the above points) is prioritization of variants. Prioritizing which variants are potentially disease-causing, provides researchers with a smaller set of variants on which to follow-up (for instance, to attempt replication of findings or to perform *in vitro* or *in vivo* studies to determine the functionality of the variants).

To illustrate the first point, missing heritability, height will be used an example. 16% of the phenotypic variability in height is explained by 697 known GWAS loci (Wood et al., 2014). 45% is explained by all genotyped variants (imputation was not considered) (Yang et al., 2010), but 80% is explained by twin studies (Silventoinen et al., 2013). The missing heritability lies between the all genotyped variants' contribution (45%) to the variance calculated through twin studies (80%). This gap begs the need for larger sample sizes or new approaches.

The second point relates to fine-mapping (Edwards et al., 2013): determining the causal variant in a locus (where locus refers to a region of high LD in the DNA sequence) that is associated with the phenotype of interest. The need for fine-mapping is a limitation of GWAS. GWAS identify associated loci, such variants are not necessarily the disease-causing variant; indeed, any variant in high LD with the associated variant may be causal. There could be more than one causal variant at a locus as well. This need for fine-mapping motivated the creation and use of specialized genotyping chips, for example using the immunochip (Lenz et al., 2015), and sequencing.

For prioritization, many methods focused on variants (often nonsynonymous) within genes because these variants have an easily explained biological rationale: a direct effect on the protein or gene function. Example of such methods include SIFT (Ng and Henikoff, 2003) and PolyPhen (Adzhubei et al., 2010). SIFT predicts whether an amino acid substitution will have an effect on the protein function based on evolutionary conservation and how much the predicted biochemical properties differ between the altered amino acid from the expected one. PolyPhen uses a combination of conservation and three-dimensional structure to predict damaging mutations. Another method looked at the functionality of synonymous variants (SilVA) (Buske et al., 2013), albeit it is rare to have synononymous changes that are harmful in comparison to nonsynonymous changes (Buske et al., 2013).

For genetic variants that do not fall into the coding sequence, adding additional information to genotype can offer biological explanations as to why these non-coding variants are associated with a phenotype. Epigenetic and other functional genomic information may be useful in prioritizing which variants are risk variants. Such data will be discussed in the next section.

My work aims to create *in silico* tools to help researchers either fill some of the void of missing heritability or to select the best variants for follow-up by functional studies. I will be prioritizing SNPs from GWAS studies combining statistical and functional genomic information together to address both points. What makes my work novel is that it incorporates more functional genomic data than previously published methods, and also investigates the use of phenotype-specific prioritization models.

1.8 Functional genomic information

There are a number of types of functional annotations that are not within the "boundaries" of a gene (loosely defined). A well-known example is the promoter region. Promoters are regions upstream of a gene, which recruit the proteins required for that

gene to be transcribed (Baumann et al., 2010). Core promoters have been identified based on the location in relation to genes (e.g. the 30 base pairs upstream from the transcription start site) (Griffiths et al., 2008). Another example is enhancer regions. Enhancers are regions in the DNA sequence that recruit transcription factors through specific motifs binding in order to accelerate transcription (Spitz and Furlong, 2012). These locations can also be defined based on epigenetic marks.

Epigenetic modification covers a broad range of functional annotations. The term signifies "over" genetics, and encompasses chemical modifications to the DNA that do not alter the DNA base sequence itself (Griffiths et al., 2008). In some cases, DNA regions are identified to have a regulatory signature based on the proteins that bind to them. Histones, for instance, are proteins that the DNA wraps around to maintain its conformation, and they play an active role in transcription. Histone modifications are chemical groups added to the histone proteins. Depending on the histone modifications, the adjoining DNA sequence has different roles in transcription. Such modifications include H3K27Ac (acetylation of the twenty-seventh lysine of H3, which is associated with active enhancers), H3K4Me1 (monomethylation of the fourth lysine residue of H3, which is associated with poised enhancers or with active enhancers if it is in combination with H3K27Ac), and H3K4Me3 (trimethylation of the fourth lysine of H3, which is associated with active promoters if it is in combination with H3K27Ac) (Shlyueva et al., 2014).

Another example of an epigenetic modification is DNA methylation, which involves the enzymatic addition of a methyl group to the carbon-five position on cytosine residues (Griffiths et al., 2008). Furthermore, there are other forms of epigenetic DNA modifications (e.g. hydroxymethylation), and methylation is not specific to cytosine bases (Lister et al., 2013).

There are many sources of publically available functional genomic information. There are large consortiums that have generated a range of data such as the Encyclopedia of DNA

Elements (ENCODE) and the Epigenomics Roadmap Projects. There are also other specific types of functional annotations that have been generated by a variety of different groups and published, such as expression quantitative trait loci (eQTLs) and conservation measures.

1.9 ENCODE

The goal of the ENCODE project was to map the functional elements in the genome: a segment of the genome that either encodes a defined product such as a protein, or has a biochemical signature (e.g. transcription factor binding site or some other protein binding site) or a specific chromatin structure (e.g. accessible open chromatin) (The ENCODE Project Consortium, 2011). ENCODE was the first large collaborative international project to undertake such an ambitious task. Experiments have been performed by many groups in numerous human immortal cell lines and tissues, and also in mouse. A limitation is that most of the ENCODE data are from (immortal) cell lines ("tier 1" cell types, see **section 1.10.1**), with a limited amount of data from actual tissue. Immortal cell lines may not reflect the actual biology in normal cells and tissue (Kashyap et al., 2011).

ENCODE data (https://www.encodeproject.org/) have been generated following standardized guidelines, and the data have been uniformly processed to ensure robustness. Some key insights from this project include: many non-coding variants fall into ENCODE-annotated functional regions, many associated variants identified through GWAS are enriched in non-coding functional elements, and there is conservation of these elements among primates (The ENCODE Project Consortium, 2012).

The UCSC Genome Browser (Meyer et al., 2013) Table Browser tool (Karolchik et al., 2004) and the FTP site (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/) provide access to data related to mapping and sequencing, genes, expression, regulation, comparative genomics and variation and repeats, many of which are from ENCODE.

Various wet laboratory methodologies were employed to determine the genomic sites of these functional annotations. DNase I hypersensitivity can be detected by FAIRE or DNase-seq, for example. The histone modifications and transcription factor binding sites are detected by ChIP-seq. For FAIRE, chromatin is cross-linked with formaldehyde *in vivo*, sheared by sonication, and phenol-chloroform extracted. The DNA recovered is fluorescently labelled and hybridized to a microarray (or sequenced) to get its sequence (Giresi et al., 2007). For ChIP-seq, formaldehyde is used to cross-link proteins to DNA. Sonication shears the chromatin to a target size of 100 to 300 base pairs, and the protein of interest bound to DNA is then isolated with an antibody specific for the factor (e.g. transcription factor or histone modification). Those DNA fragments can then be sequenced (Landt et al., 2012).

1.10 Evaluation of functional annotations from ENCODE

This section delves into some ENCODE data available. I highlight some issues relating to the data including cell lines, and measure choice. I also provide some details about two annotations below: transcription factor binding sites and DNase I hypersensitive sites.

1.10.1 ENCODE cell lines

The ENCODE Project has categorized various cell lines into three tiers, Tiers 1 through 3, where Tier 1 cell lines have the highest priority with regard to designing the functional experiments. There are three Tier 1 cells (GM12878, H1-hESC, K562), and around 15 Tier 2 cells. The original Tier 2 cell lines were HeLa-S3, HepG2, and HUVEC, and the remaining (A549, CD20+, CD20+_RO01778, CD20+_RO01794, H1-neurons, IMR90, LHCN-M2, MCF-7, Monocytes-CD14+, Monocytes-CD14+_RO01746, Monocytes-CD14+_RO01826, SK-N-SH) were added afterwards. Most of the experiments have data from all Tier 1 cells that can be accessed as separate from the other cell types. However, the presence of the Tier 2 cells is sparser (see **Table 1.1**).

Table 1.1. Cell types (tiers) for some ENCODE Regulation data tables/tracks

Table	Cell Types *
DNase Clusters (v2) (wgEncodeRegDnaseClusteredV2.bed.gz)	GM12878 H1-hESC K562 A549 HeLa-S3 HepG2 HUVEC Monocytes-CD14+_RO01746 CD20+ HMEC AG04449 8988T AG04450 AG09309 AG09319 AG10803 Adult_CD4_Th0 AoAF AoSMC BE2_C BJ CD34+_Mobilized CLL CMK Caco-2 Chorion FibroP Fibroblasts GM06990 GM12864 GM12865 GM12891 GM12892 GM18507 GM19238 GM19239 GM19240 Gliobla H7-hESC H9ES HA-h HA-sp HAEpiC HAc HBMEC HCF HCFaa HCM HCPEpiC HCT-116 HConF HEEpiC HFF HFF-Myc HGF HIPEpiC HL-60 HMF HMVEC-LBl HMVEC-LLy HMVEC-dAd HMVEC-dBl-Ad HMVEC-dBl-Neo HMVEC-dLy-Ad HMVEC-dLy-Neo HMVEC-dNeo HNPCEpiC HPAEC HPAF HPDE6-E6E7 HPF HPdLF HRCEpiC HRE HRGEC HRPEpiC HSMM HSMM_emb HSMMtube HTR8svn HVMF Hepatocytes Huh-7 Huh-7.5 Ishikawa Jurkat LNCaP MCF-7 Medullo Melano Myometr NB4 NH-A NHDF-Ad NHDF-neo NHEK NHLF NT2-D1 Osteoblasts PANC-1 PanIsletD PanIslets PrEC ProgFib RPTEC RWPE1 SAEC SK-N-MC SK-N-SH_RA SKMC Stellate T-47D Th1 Th2 Urothelia WERI-Rb-1 WI-38 iPS pHTE
DNase Clusters (v1) (wgEncodeRegDnaseClustered.bed.gz)	GM12878 H1-hESC K562 A549 HUVEC HeLa-S3 HepG2 MCF-7 Monocytes-CD14+ SK-N-SH_RA AG04449 AG04450 AG09309 AG09319 AG10803 AoAF BE2_C BJ Caco-2 GM06990 GM12864 GM12865 H7-hESC HA-h HA-sp HAEpiC HAc HBMEC HCF HCFaa HCM HCPEpiC HCT-116 HConF HEEpiC HFF HFF-Myc HGF HIPEpiC HL-60 HMEC HMF HMVEC-LBl HMVEC-LLy HMVEC-dBl-Ad HMVEC-dBl-Neo HMVEC-dLy-Ad HMVEC-dLy-Neo HMVEC-dNeo HNPCEpiC HPAF HPF HPdLF HRCEpiC HRE HRGEC HRPEpiC HSMM HSMMtube HVMF Jurkat LNCaP NB4 NH-A NHDF-Ad NHDF-neo NHEK NHLF PANC-1 RPTEC SAEC SK-N-MC SKMC WERI-Rb-1 WI-38
UW DNase I HS	Gm12878 H1hESC K562 A549 CD20+_RO01778 Hela-S3 HepG2 HUVEC LHCN-M2 Monocd14 Monocd14ro1746 Ag04449 Ag04450 Ag09309 Ag09319 Ag10803 Aoaf Be2c Bj Caco2 Cd34mobilized Cd4naivewb11970640 Cd4naivewb78495824 Cmk Gm04503 Gm04504 Gm06990 Gm12864 Gm12865 c H7es H7esDiffa14d H7esDiffa2d H7esDiffa5d H7esDiffa9d Hac Hae Hah Hasp Hbmec Hbvp Hbvsmc Hcf Hcfaa Hcm Hconf Hcpe Hct116 Hff Hffmyc Hgf Hipe Hl60 Hmec Hmf Hmvecdad Hmvecdblad Hmvecdblneo Hmvecdlyad Hmvecdlyneo Hmvecdneo Hmveclbl Hmveclly Hnpce Hpaec Hpaf Hpdlf Hpf Hrce Hre Hrgec Hrpe Hs27a Hs5 Hsmm Hsmmt Hvmf Jurkat K562Znf2c10c5 K562Znf4c50c4 K562Znf4g7d3 K562Znfa41c6 K562Znfb34a8 K562Znfe103c6 K562Znff41b2 K562Znfg54a11 K562Znfp5 Lhcnm2Diff4d Lncap M059j Mcf7 Mcf7Est100nm1h Mcf7Estctrl0h Msc Nb4 Nha Nhbera Nhdfad Nhdfneo Nhek Nhlf Nt2d1 Panc1 Prec Rpmi7951 Rptec Saec Skmc Sknmc Sknshra T47d Th1 Th17 Th1wb33676984 Th1wb54553204 Th2 Th2wb33676984 Th2wb54553204 Tregwb78495824 Tregwb83319432 Werirb1 Wi38 Wi38Ohtam
Duke DNase I	Gm12878 H1-hesc K562 A549 CD20+_RO01794 HeLa-S3 HepG2 HUVEC Monocd14 SK-N-SH 8988t Adultcd4th0 Adultcd4th1 AosmcSerumfree Cerebellumoc Cerebrumfrontaloc Chorion Cll Colo829 Ecc1Dm002p1h Ecc1Est10nm30m Fibroblgm03348Lenticon Fibroblgm03348Lentimyod Fibroblgm03348 Fibrobl Fibropag08395 Fibropag08396 Fibropag20443 Fibrop Frontalcortexoc Gcbcell Gliobla Gm10248

	Gm10266 Gm12891 Gm12892 Gm13976 Gm13977 Gm18507 Gm19238 Gm19239 Gm19240 Gm20000 H7es H9es Heartoc Hek293t Helas3Ifna4h Hepatocytes Hmec Hpde6e6e7 Hsmmemb Hsmmfshd Hsmm Hsmmt Htr8 Huh75 Huh7 Imr90 Ipscwru1 Ipsnihi11 Ipsnihi7 Ips IshikawaEst10nm30m IshikawaTam10030 K562G1phase K562G2mphase K562Nabut K562Saha1u72hr K562Sahactrl LncapAndro Lncap Mcf7Ctcfshrna Mcf7Hypoxlaccon Mcf7Hypoxlac Mcf7 Mcf7Randshrna Medullod341 Medullo Mel2183 Melano Myometr Naivebcell Nhek Olfneurosphere Osteobl Panisd Panislets Phte Progfib Psoasmuscleoc Rwpe1 Stellate T47dEst10nm30m T47d UrothelV2 UrothelUt189V2 Urotsa UrotsaUt189
Txn factor ChIP (wgEncodeRegTfbsClustered.bed.gz)	Transcription factors: AP-2alpha AP-2gamma ATF3 BAF155 BAF170 BATF BCL11A BCL3 BCLAF1M33-P5B11 BDP1 BHLHE40 BRCA1C-1863 BRF1 BRF2 Brg1 CCNT2 CEBPB c-Fos CHD2N-1250 c-Jun c-Myc CtBP2 CTCF CTCFC-20 CTCFLSC-98982 CTCFSC-5916 E2F1 E2F4 E2F6 E2F6H-50 EBF EBF1C-8 eGFP-FOS eGFP-GATA2 eGFP-HDAC8 eGFP-JunB eGFP-JunD eGFP-NR4A1 Egr-1 ELF1SC-631 ELK4 ERalphaa ERRA ETS1 FOSL1SC-183 FOSL2 FOXA1C-20 FOXA1SC-101058 FOXA2SC-6554 GABP GATA-1 GATA-2 GATA2CG2-96 GATA3SC-268 GCN5 GR GRp20 GTF2B GTF2F1RAP-74 HA-E2F1 HDAC2SC-6296 HEY1 HMGN3 HNF4A HNF4AH-171 HNF4GSC-6558 HSF1 Ini1 IRF1 IRF3 IRF4M-17 JunD KAP1 MafFM8194 MafKab50322 MafKSC-477 Max MEF2A MEF2CSC-13268 Mxi1bHLH NANOGSC-33759 NELFe NF-E2 NF-E2H-230 NFKB NF-YA NF-YB Nrf1 NRSF Oct p300 p300F-4 p300N-15 PAX5-C20 PAX5-N19 Pbx3 PGC1A Pol2 Pol2-4H8 Pol2b Pol2phosphoS2 Pol3 POU2F2 POU5F1SC-9081 PRDM1Val90 PU.1 Rad21 RFX5N-494 RPC155 RXRA SETDB1 Sin3Ak-20 SIRT6 SIX5 SMC3ab9263 SP1 SP2SC-643 SPT20 SREBP1 SREBP2 SRF STAT1 STAT2 STAT3 SUZ12 TAF1 TAF7SQ-8 TAL1SC-12984 TBP TCF12 TCF4 TFIIIC-110 THAP1SC-98174 TR4 USF-1 USF1SC-8983 USF2 WHIP XRCC4 YY1 YY1C-20 ZBTB33 ZBTB7ASC-34508 ZEB1SC-25388 Znf14316618-1-AP ZNF263 ZNF274 ZZZ3
Layered H3K4Me1/ H3K4Me3/ H3K27Ac	GM12878 H1-hESC K562 HUVEC HSMM NHEK NHLF
Broad Histone-H3K4Me1, H3K4Me3, H3K27Ac	GM1278 H1-hESC K562 A549 (conditions: Dex^{+} or EtOH) HeLa-S3 HepG2 HUVEC Monocytes-CD14+_ RO01746 Dnd41 HMEC HSMM HSMM tubule NH-A NHDF-Ad NHEK NHLF Osteoblasts
Transcription (RNA-seq)	GM12878 H1-hESC K562 HeLa-S3 HepG2 HUVEC LHCN-M2 Myoblast LHCN-M2_Myocyte_7d MCF-7 GM12891 GM12892 HSMM NHEK NHLF

* Tier 1; Tier 2; Tier 3^{+} dexamethasone. Credits for each data set available on the UCSC site.

ENCODE accession numbers for UW DNase I, HS, Duke DNase I, and Broad Histone- HeK4Me1, H3K4Me3 and H3K27Ac are listed in **Appendix D**.

The cell types that have data vary depending on the functional annotation. Limiting the analysis to certain cell types will limit the data available for each annotation.

1.10.2 "Peaks" versus "Signals"

Histone data are available in tables for peaks (the "BroadHistone" tracks in the table browser) and signals ("Layered" tracks in the table browser). Details are found here http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=wgEncodeBroadHistone, but in brief signals are based on density and are given for each base pair position while peak scores are based on regions of statistically significant enrichment based on the signal from controls (measurement of background abundance in the genome). The signal is a function of the cell counts that contain the modification of interest. The peak scores are more informative than the signal data (i.e. density) in our application of these data as a predictor of SNP functionality. In this analysis we are most interested in genomic regions enriched with the functional annotation, which would be the peak scores as they are based on regions of statistical significance from comparing the signals in the experiments to the signals from the corresponding control set. Moreover, other tracks, including DNase clusters as well as Txn Factor ChIP also used standardized scores (on a scale of 0-1000) based on peaks.[1]

1.10.3 DNase Hypersensitivity- DNase Clusters, UW DNase I HS, Duke DNase I HS

There are several tracks available for DNase I hypersensitivity: two UW (UW DNase I HS and DNase Clusters) and one from Duke (Duke DNase I HS).

---

[1] The peak scores have been standardized to fall between 0 and 1000. The input signal values were multiplied by a normalization factor: the ratio of the maximum score value (1000) to the signal value at one standard deviation from the mean, and values exceeding 1000 were assigned to 1000. The peak score for the interval is the mean signal value across the interval.

UW DNase I HS and Duke DNase I HS gives individual tables for each cell type, while DNase Clusters amalgamate the cell types together. The UW HS track shows DNase I sensitivity measured in different cell lines using Digital DNase I methodology (in brief, DNaseI digestion of intact nuclei, isolating DNaseI fragments, and direct sequencing of fragment ends).

The Duke DNase I HS shows the locations of regulatory elements identified as open chromatin in multiple cell types using DNase I HS assays. There is more coverage compared to UW HS as assessed by the total length of base pair regions present in each of these tracks.

The DNase Clusters track contains a score based on peaks for genomic regions. See **Figure 1.2** for the score distribution. This track additionally provides the number of experiments or cell lines in which the results were significant (range: 2-148). There is no correlation between number of experiments and score although the latter distribution may be influenced by the cut-off of 1000.

With regard to coverage among the tracks, the DNase Clusters table combines information from all the cell lines from both the UW and Duke groups and has the most genomic coverage (13%). However, as mentioned this track provides peak scores for all of the cell types together rather than a peak score for each cell type as do the tracks for the UW and Duke groups.

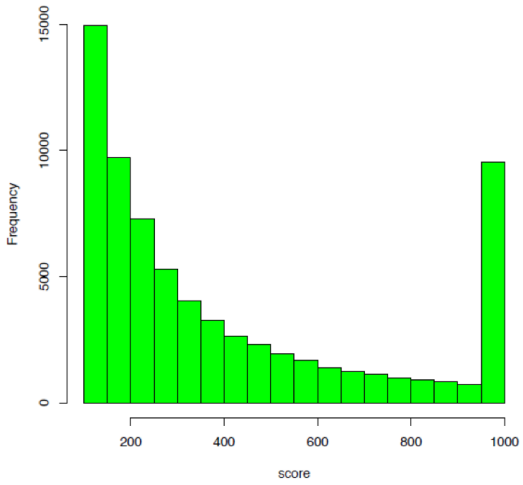


Figure 1.2. Peak score distributions for the DNase I Clusters table for human chromosome 3

1.10.4 Txn Factor ChIP

Peak scores are provided for several cell lines, and the overall score reports the highest peak score from among all the cell lines for the particular transcription factor. See **Figure 1.3** for the distribution of the transcription factor binding peak scores on chromosome 3.

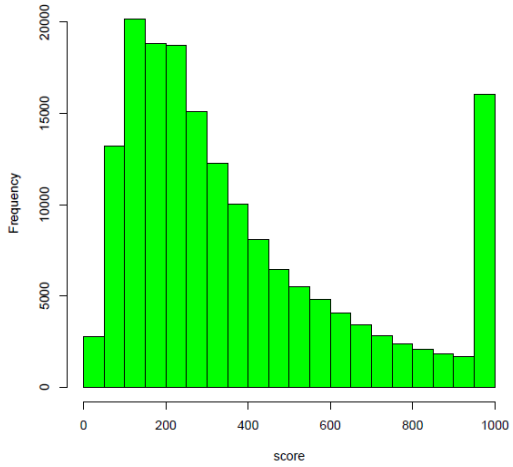


Figure 1.3. Transcription factor binding sites peak scores for human chromosome 3

Analyzing the scores according to each cell type shows that all of the cell types have data available (e.g. none of the cell lines have complete missing data), and the ranges vary for the various transcription factors, and some factors are more represented than others. Interestingly, most (76%; 128,928 of the 170,219 results) of the chromosome 3 data have transcription factor binding site data from only one experiment (i.e. one cell line). There is no preference as to which cell line has the most non-zero scores, and so the presence of the epigenetic mark only in other cell types will be lost if only certain cell types are considered or if each transcription factor is assessed by a per cell line basis.

1.11 Roadmap Epigenomics Project

The NIH Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015) http://www.roadmapepigenomics.org/ is a large consortium to map the epigenome, specifically DNA methylation, DNA accessibility (e.g. histone modifications and DNase I hypersensitivity), and RNA expression in humans (n=111). There are differences

between ENCODE and Roadmap. ENCODE tends to use cell lines; for instance, for brain-level results, ENCODE uses two cancerous cell lines both in Tier 3: glioblastoma and neuroblastoma (http://genome-mirror.duhs.duke.edu/ENCODE/cellTypes.html), which may not reflect the epigenetic patterns found in non-tumor cells. Roadmap assesses functional elements in stem cells and primary *ex vivo* tissues. For stem cells, there is evidence of stochastic random changes in the epigenome as stem cells divide (Yatabe et al., 2001), and thus again such cells are not ideal for investigating the epigenome in a living system. The tissue-level data available through Roadmap is a closer source to the patterns exhibited in a living system. There are still factors to consider, for instance when post-mortem samples are used to acquire brain tissue samples, the cells are dead, and thus the amount of time after death the tissue was collected and analyzed is important (the postmortem interval) (Birdsill et al., 2011; Dodd et al., 1988). Although there are advantages to the Roadmap data compared to ENCODE, since tissue-level data more accurately represent the epigenomic architecture in living systems, there are still limitations such as epigenomes may differ in the different cell types within the tissue, and the use of post-mortem brain tissue. Additionally, epigenetic marks can be missed in cells that have low numbers in the tissue.

1.12 eQTLs

There are a number of expression quantitative trait loci (eQTLs) databases. eQTLs are regions in the DNA sequence that affect expression of nearby genes (cis-eQTLs) or distant genes (trans-eQTLs). Older GTEx (Genotype-Tissue Expression) eQTL Browser data can be accessed through http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi, and the more recent data on dbGAP or through their new portal at http://www.gtexportal.org/home/ (The GTEx Consortium, 2013). Most of the data from the older studies are from microarray gene expression experiments. Expression studies commonly use microarrays to measure gene expression, but there are limitations to this methodology that RNA-sequencing can overcome (e.g. novel genes and non-coding or

microRNAs cannot be assessed by arrays, and alternative splicing is generally not taken into account). The older version of GTEx contains data, primarily microarray data, from four studies (Montgomery et al. 2010, Schadt et al. 2008, Gibbs et al. 2010, Stranger et al. 2007) in lymphoblastoid cells, liver, or four brain regions (cerebellum, frontal cortex, pons, or temporal cortex). The newer data are RNA-sequencing data from a variety of human tissue (n>40) including whole blood, brain, lung and stomach from a total of 1,421 samples (The GTEx Consortium, 2013).

A tissue-specific dataset is available through the UK Brain eQTL Consortium (UKBEC) www.braineac.org (Trabzuni et al., 2011), which identifies eQTLs in brain tissue. UKBEC data are based on microarray experiments. The consortium is currently generating RNA-sequencing data that will also be made publically available. Many eQTL studies perform their analyses on whole tissue, rather than specific regions. UKBEC, however, has performed RNA-sequencing on targeted regions in the brain: substantia nigra, putamen, and hippocampus in a large number of post-mortem unaffected brains (N=150).

1.13 Conservation measures

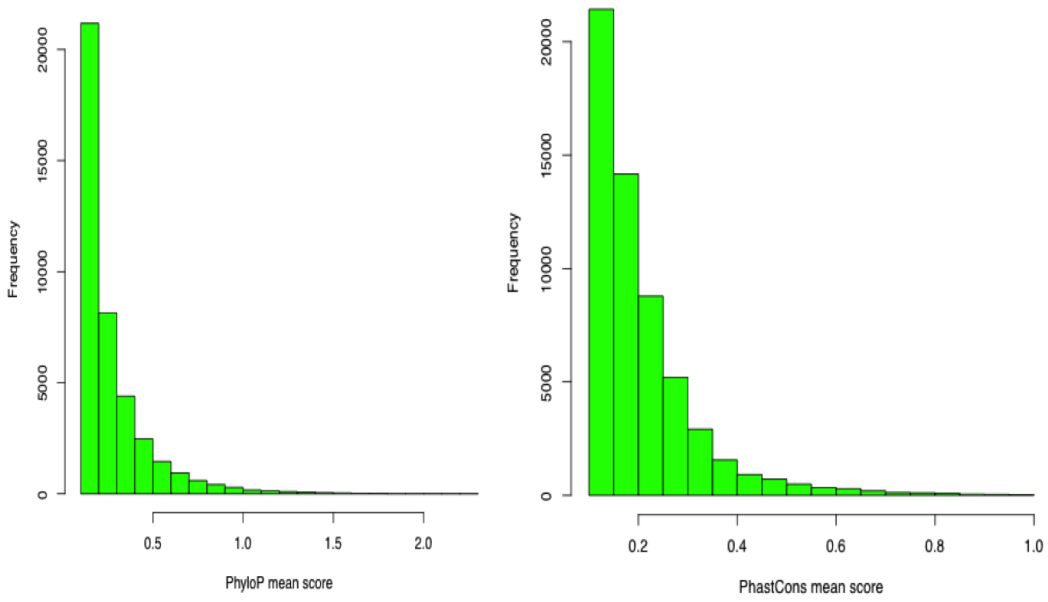
Conservation of a stretch of DNA sequence among ancestrally-related species (for instance among placental mammals) could suggest that that region of DNA plays an essential role in normal function. Thus, variants in conserved areas may be more likely to have functional consequences than variants outside of such areas (Frazer et al., 2003).

Common measures of conservation are PhyloP (Pollard et al., 2010), PhastCons (Siepel et al., 2005) and GERP (Cooper et al., 2005). PhlyoP and GERP are conservation measures for a single DNA nucleotide, whereas PhastCons provides a score for a small region of DNA. Genomic Evolutionary Rate Profiling (GERP) is a score referring to the conservation of each DNA nucleotide in multi-species alignment. Positive scores indicate

a site is under evolutionary constraint, whereas negative scores may suggest accelerated rates of evolution.

Both PhyloP and PhastCons scores are derived from the PHAST package, which makes use of phylogenetic hidden Markov models. According to the UCSC website, these two measures have their own advantages. PhyloP scores do not take into account conservation at neighbouring sites, whereas PhastCons estimates the probability that each nucleotide belongs to a conserved element. PhyloP is more effective at analyzing "signatures of selection" whereas PhastCons' strength is in detecting conserved elements (http://genome.ucsc.edu/cgi-bin/hgTables).

Regarding the actual data, I compared the base coverage and score distribution for PhyloP and PhastCons scores for 46 placental mammals. Both datasets have the same coverage of the genome (98.20%). Data points, or in other words: scores at specific SNPs, are not available for download. Instead, for both measures, the downloadable file provides the lower limit, range, and sum of all the data points in regions. The average score for each region was calculated by dividing the sum of all the data points by the number of valid data values in the block. These distributions are both positively skewed (**Figure 1.4**).



a **b**

Figure 1.4. Distribution of mean conservation scores for human chromosome 3 for placental mammals

[**a**] Distribution of PhyloP mean scores. [**b**] Distribution of PhastCons mean scores.

1.14 Dimension reduction for functional annotations

The above outlines some of the available functional data. There have been methods proposed to integrate these data and thus reduce the dimensionality of the functional data. Ernst et al. (2011) divided the genome into chromatin states based on several histone modifications through the use of a multivariate hidden Markov model. They focused on cell type-specific patterns of promoters and enhancers to define a map of chromatin states across nine human cell types in six general categories: enhancer, promoter, insulator, transcribed, repressed, and inactive states. These chromatin states can be visualized using the webserver ChroMoS (Barenboim and Manke, 2013). The knowledge of chromatin state can help inform the functional impact of the variant, but a limitation is that other types of annotations that may be important for function (e.g. DNase I hypersensitive sites or transcription factor binding sites, for instance) are not included.

Another tool is Segway (semi-automated genomic annotation), which proposes that DNA segments fall into seven "flavours" (Hoffman et al., 2012). The authors trained a dynamic Bayesian network method, simultaneously on chromatin data from multiple experiments to categorize the genome into the flavours. Unlike the chromatin states described above, Segway uses multiple sources of functional annotations: histone modifications and transcription factor binding sites, and DNaseI hypersensitive sties.

1.15 Rationale for uses of regulatory genomic information

The rationale for believing that epigenetic and other genomic information can be useful for identifying risk variants among all variants is that numerous studies have demonstrated the enrichment of associated variants from GWAS and other trait or disease-associated variants with such characteristics. Emerging experimental data from various sources have suggested that the functional annotations of specific genomic regions, such as histone modifications, DNase I hypersensitive sites, transcription factor binding sites, and expression quantitative trait loci (eQTL) among others, could offer biological explanations for many variants found to be associated with disease (Hindorff et al., 2009; Knight et al., 2011; Nicolae et al., 2010). This evidence all suggests that functional information has the potential to be included in statistical learning algorithms to differentiate genetic risk variants from non-risk variants based on their overlap with various functional annotations.

Below I will highlight a few key papers published shortly after the publication of data from the ENCODE Project featuring those ENCODE results that demonstrate an enrichment of genetic risk variants for various functional genomic characteristics.

Schaub et al. (2012) showed that putative disease-associated variants (GWAS Catalogue SNPs) and variants in high linkage disequilibrium (LD) with those variants show significant enrichment for multiple functional annotations from the ENCODE Project. Maurano et al. (2012) also found enrichment in GWAS variants or variants with which

they are in high LD. The authors looked specifically at DNase I hypersensitive sites, and found that the GWAS variants are more frequently localized to DNase I hypersensitive sites than would be expected by chance. Maurano et al. also showed that the level of enrichment for subsets of GWAS Catalogue variants associated with a particular trait depends on the cell/tissue type considered. Further evidence for varying level of enrichment was presented in Farh et al. (2015). They created an algorithm and used permutation to estimate the posterior probability that an individual SNP is a causal variant given the haplotype structure and observed pattern of association at the locus for autoimmune-associated loci. They observed that their identified causal SNPs were enriched in enhancers (i.e. H3K4Me1 and H3K27Ac histone marks) that were mapped in immune cells (Farh et al., 2015).

The enrichment of GWAS variants has been found in other functional sources in addition to ENCODE data. Hnisz et al. (2013) showed that trait-associated genetic variants from GWAS are enriched in super-enhancers (large clusters of enhancers associated with genes involved in cell identity, for instance encoding cell-type-specific transcription factors) and to a lesser degree in enhancers in general. Furthermore, the Roadmap Epigenomics Consortium also showed an enrichment of GWAS Catalogue variants with this consortium's data (e.g. histone marks and DNase I) across all of their epigenomes interrogated (Roadmap Epigenomics Consortium et al., 2015).

Hindorff et al. (2009) and Knight et al. (2011) showed enrichment of SNPs from the GWAS Catalogue for several functional annotations using a random sampling of SNPs from the HapMap II European-ancestry (CEU) population or from GWAS genotyping arrays, respectively.

Similarly, enrichment of risk variants from sources other than the GWAS Catalogue with such characteristics have been demonstrated, such as enrichment of variants in the HGMD (Ritchie et al., 2014).

Given that risk variants are enriched in functional information, these data can be used to help with the two points that remain outstanding for disease-implicated loci: to identify novel risk variants and to identify the causal variant at a disease-associated locus.

The next section describes the evolution of methods used to incorporate functional genomic data to prioritize genetic risk variants.

1.16 Using functional genomic information to prioritize genetic risk variants

Originally ad hoc methods were utilized for incorporating functional information, from which investigators could make their own conclusions on the functionality of a variant. For instance, user-friendly tools that process data from ENCODE and other sources were developed that show the overlap of variants with various genomic annotations, and based on that one can comment on the variants' causality. Examples of such tools include HaploReg (Ward and Kellis, 2012) and RegulomeDB (Boyle et al., 2012) (see **Table 1.2**). HaploReg shows the overlap of the variant of interest (and also variants at user-defined pairwise-LD cutoffs with that variant) with annotations from ENCODE and other sources. RegulomeDB also incorporates several annotations from ENCODE and other sources. The latter uses a categorical scoring system, but the scale is crude. Likely causal variants are those that are expression quantitative trait loci (eQTL) and at the same time fall in transcription factor binding sites and DNase I hypersensitive sites. These SNPs are more highly ranked with regard to likely having an effect (i.e. affect binding of factors and expression of a gene). SNPs that are not eQTLs, regardless of whether they fall into a transcription factor binding site or DNase I hypersensitive site, are placed in a category of SNPs less likely to be functional. Variant identifiers can be inputted into these tools in order to either decide which are suitable candidates for follow-up or which should be included in an association study. For example, in a candidate gene study on antipsychotic-induced weight gain (Gagliano et al., 2014b) (see **Appendix B**), I inputted

into HaploReg the variant that showed the highest evidence for association with the phenotype to examine its functional potential.

Table 1.2. Selection of online tools that are available for showing overlap of variants, including noncoding variants, with functional annotations

Goal	Input	Output	Annotations used	Utility	Caveat
RegulomeDB (Boyle et al., 2012)					
A database that annotates SNPs with known and predicted regulatory elements in the intergenic regions of the human genome	Multiple including: dbSNP IDs, BED or VCF files, hg19 coordinates;	Categorical score where the highest scoring SNPs are likely to affect binding and gene expression	DNAse I, transcription factors, and promoter regions (sources: GEO, ENCODE)	Can download all the dbSNP 137 SNPs for each category	Categorical outcome limited; LD between SNPs not taken into account
HaploReg (Ward and Kellis, 2012)					
Tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as SNPs at disease-associated loci.	List of rsIDs; single region; select a GWAS	Annotates inputted SNPs (and proxies) based on location	ENCODE (histone marks, proteins bound, DNase I), conservation, motifs changed, etc.	LD threshold available from r^2>0.2 (based on 1KG phase 1)	Annotates SNPs but does not provide a score/ prediction

1KG= 1000 Genomes Project

There are also publically available databases specifically designed to look at transcription factor binding sites, such as MAPPER2 (Riva, 2012), and JASPAR (Mathelier et al., 2013; Sandelin et al., 2004). These tools either contain transcription factor binding sites that are predicted computationally or have been observed experimentally. MAPPER2 contains putative transcription factor binding sites (upstream of genes in the promoter and the initial introns) in the genomes of human, mouse, and drosophila. JASPAR contains

curated experimentally-derived transcription factor binding motifs from many eukaryotes, including human.

Then came methods that produce a score or rank describing how likely the variant is to be functional or in other words is a genetic risk variant by combining lots of functional data together. Some of these methods are specifically for the integration of functional information with statistical association data from conducted GWAS.

Schork et al. (2013), for example, looked at enrichment of genic elements (e.g. intergenic, intron, exon, etc.) in various GWAS using summary statistics taking into account LD. They suggest the use of stratified False Discovery Rate (sFDR) to rank variants. A limitation to this methodology is that the FDR is dependent on the study's data and thus ranks cannot necessarily be extrapolated to other studies.

Some of these methods provide a posterior probability to rank the variants in the locus. For instance, Knight et al. (2011) reported Bayes factors for annotation (based on three annotations: eQTLs in open chromatin, nonsynonymous SNPs, SNPs in promoters) for each SNP. They propose that these Bayes factors should be combined with the corresponding Bayes Factor for association from a GWAS. This study had a limited number of annotations. Thompson et al. (2013) looked at binary predictor variables (such as whether or not a variant is in a functional protein domain or whether or not the variant is in a gene expressed in tissue relevant to the phenotype) using a logistic regression model, and they incorporate GWAS data. A limitation is that some of their predictor variables were subjective (e.g. in a gene with protein-protein interactions relevant to the phenotype) based on expert GWAS investigators' opinions described in Minelli et al. (2013), and also they had a limited number of predictor variables (n=15) (Thompson et al., 2013).

The online tools in **Table 1.3** are additional tools that all give some sort of score or posterior probability to SNPs. A downside to these methods in the table is that they are

only applicable to GWAS that have already been conducted since they require summary statistics (information summarizing the strength of the association with the phenotype for each SNP such as odds ratios, test statistics and p-values).

Table 1.3. Selection of online tools that are available for prioritizing genetic variants, including noncoding variants, requiring either association study data or summary statistics

Goal	Input	Output	Annotations used	Utility	Caveat
Multi-threshold (and Multi-marker) Association Study Analysis: MASA (Darnell et al., 2012)					
To compute an association statistic taking into account prior information (multi-thresholding akin to varying the significance threshold at each marker depending on prior info)	Case/control GWAS data, reference haplotype file, marker file	Z-score and p-values for each SNP	Annotations used as prior information – ENCODE data	Provides an association p-value for each SNP corrected for multiple testing (either Bonferroni or permutation)	Data must be in Beagle (Browning and Browning, 2007) format

Table 1.3. Selection of online tools that are available for prioritizing genetic variants, including noncoding variants, requiring either association study data or summary statistics (**continued from previous page**)

Goal	Input	Output	Annotations used	Utility	Caveat
Probabilistic Annotation INTegratOR (PAINTOR) (Kichaev et al., 2014)					
Fine-mapping- prioritize causal GWAS variants using association stats and genomic functional info (maximum likelihood estimation using an application of Bayes Theorem)	Association info (i.e. Z score), LD info (e.g. from 1KG) & annotations (e.g. ENCODE)	Posterior probabilities; Gamma (effect size) estimates	Need to add your own annotation columns	Estimates the contribution of each annotation from summary stats; accounts for LD; allows multiple causal variants at a locus	Restricted to empirical GWAS data
fgwas software (Pickrell, 2014)					
Test whether SNPs that influence a trait are enriched or depleted in certain genomic annotations (using a penalized likelihood to get posterior probability that a SNP in a given genomic region is causal)	GWAS data (SNP IDs, allele frequency, Z-score, sample size of study) + genomic data input	Posterior probabilities; the association statistics for each SNP in the genome and in each region as estimated by the model	DNase I HS, Chromatin state data, gene annotations used in the paper; for fgwas, need to add own annotations	Input own GWAS data and annotations to get posterior probabilities for genomic regions and/or each SNP in the genome	Assumes only a single causal SNP in a given genomic region; restricted to empirical GWAS data
Phenotype Driven Variant Ontological Re-ranking tool (Phevor) (Singleton et al., 2014)					
Integrate phenotype, gene function, & disease data with genomic data for improved power to identify disease-causing alleles by using both variant prioritization tools and biomedical ontologies	Phenotypes; output from other variant prioritization tools (e.g. PhastCons)	Phevor score for each gene	Ontologies: Human Phenotype, Mammalian Phenotype, Disease, & Gene Ontologies	Not limited to known disease-associated genes/variants; useful for single exome and trio-based diagnostic analyses (i.e. clinical scenarios)	Individual diagnostic analysis; depends on reliability of input (e.g. ontologies)

Additional methods are trained directly on known risk variants from databases through employing supervised statistical/machine learning algorithms that output a score/probability inferring the likelihood of a SNP to be functional. These methods are most versatile since they can be used to score SNPs without requiring GWAS summary statistics, and thus their utility is not limited to following up GWAS signals from an existing study.

Kindt et al. (2013) published a permutation approach examining the enrichment or depletion of a subset of GWAS Catalogue SNPs ($p<5x10^{-8}$) in the annotations investigated in two previous papers (Hindorff et al., 2009; Knight et al., 2011), and also added in a number of genic and regulatory features, conserved elements and chromatin states. They report odds ratios of the annotations (from logistic regression) signifying which annotations are more likely to contain significant associated SNPs, which can be used to prioritize GWAS hits for further studies. Although the Kindt et al. method uses risk variants from a database, a limitation is that a SNP is not actually given a score/probability as to how likely it is have a functional consequence.

In a Bayesian framework, Iversen et al. (2014) incorporated multiple annotations (for example, genomic location, DNase I hypersensitivity, and scores from databases such as RegulomeDB (Boyle et al., 2012)) and was able to improve the ranks of known associated variants in a GWAS of ovarian cancer. This method produces posterior probabilities for each SNP, but a limitation is that a script or program to implement the method is not made available.

None of the studies mentioned above considered using a phenotype-specific analysis: creating a model to specifically identify risk variants for a particular disease. Although Iversen et al. (2014) tested their model on a GWAS of ovarian cancer, their model was not specifically trained to identify variants specific to such a phenotype since they trained their model on all GWAS Catalogue variants. Additionally, none of the studies considered the issue of cell/tissue-specificity for the annotations. These studies used

annotations that integrated all of the cell types together as a unified annotation. For instance, the DNase Clusters track provided by ENCODE unifies the cell types together to define DNase I hypersensitive sites (see **section 1.10.3**). Of the selection of data-trained online tools in **Table 1.4**, the first three do consider cell/tissue-specificity for the annotations.

Table 1.4. Selection of online tools that are available for prioritizing variants, including noncoding variants, based on data-trained algorithms

Goal	Input	Output	Annotations used	Utility	Caveat
(Gagliano et al., 2014a)					
To prioritize GWAS SNPs for follow-up based on functional data (Used a version of elastic net to train data on genome-wide significant SNPs in the GWAS Catalogue ("hits") vs. SNPs not present in the Catalogue ("non-hit"))	List of SNPs (or GWAS summary data if want to apply the method directly to GWAS)	Bayes factors for annotation (and Bayes factors for association if using GWAS summary data)	14 with cell types amalgamated together: ENCODE (DNase I, TFBS, histone marks, conservation, eQTLs, etc.)	LD between SNPs taken into account for annotating; precomputed Bayes factors for 1KG SNPs available on website	Model needs to be rerun to include new annotations
Combined Annotation-Dependent Depletion (CADD) (Kircher et al., 2014) [DANN- uses the published CADD training data to train a neural net (Quang et al., 2014)]					
To prioritize functional, deleterious and pathogenic variants across many functional categories, effect sizes and genetic architectures (Used support vector machine to train data–half human derived allele variants, half simulated; DANN uses identical training set, but employs a deep neural net instead.)	VCF file containing up to 100,000 variants	C score (raw and scaled) for each variant with option to include the underlying annotations	63 distinct: Ensembl Variant Effect Predictor16 (VEP), data from the ENCODE Project, information from UCSC Genome Browser tracks	Webserver to get precomputed C scores for 8.6 billion human SNPs	Arbitrary C score cut-off to define deleterious; Model needs to be rerun to include new annotations; >1 line of output for variants in multiple genes

Table 1.4 Selection of online tools that are available for prioritizing variants, including noncoding variants, based on data-trained algorithms (**continued from previous page**)

Goal	Input	Output	Annotations used	Utility	Caveat
Genome-wide annotation of variants (GWAVA) (Ritchie et al., 2014)					
Tool for prioritizing non-coding variants by integrating genomic and epigenomic data (Used modified random forest to train data on HGMD SNPs vs. matched or unmatched control sets)	rsIDs, regions	Prediction scores from 3 different versions of the classifier (based on different control sets)	174: ENCODE (DNase I, Txn factors, histone marks), conservation, genic & sequence contexts	Interactive webserver to get scores; Python scripts and data available on FTP site	Classifier based on HGMD SNPs, so not as effective for GWAS SNPs
Silent Variant Analyzer (SilVA) (Buske et al., 2013)					
Random-forest based method for prioritizing ranking (and scoring) synonymous variants that are likely to be functional	VCF file of the variants (SilVA will only analyze synonymous)	Variant rank out of all synonymous variants considered; SilVA score, between 0 and 1	All related to synonymous SNPs: Sequence conservation, splice sites/factor motifs, RNA folding energy, codon usage and CpG content	Score provided, but authors stress that the rank is the more important output	Only for synonymous SNPs; run on local computer, but need wget software, etc.
FunSeq2 (Fu et al., 2014)*					
To identify noncoding genetic somatic drivers in cancer; 2 steps: creation of data context, and variant prioritization	Cancer variants (BED/VCF); gene list (optional) differential gene expression data (optional)	Variant reports that identify novel sensitive/ultra-sensitive regions based on networks; Candidates File with potential candidates	7 binary: functional annotations (DNase HS, etc.) 4 continuous: motif-breaking/gaining score, GERP score, etc.	Can provide own features, and own gene networks or use those supplied	Intended for somatic cancer variants in genes (can download a file with scores for all noncoding variants)

*Not completely data-trained because weights are derived for each variant independently based on its annotations, i.e. a model is not created *per se* in a training set and then applied to the test set variants

There are numerous statistical learning algorithms from which to choose to create data-trained prioritization models. These algorithms must be able to handle the features of the functional data: correlations among predictor variables, and a large quantity of both samples and predictor variables. A few of the algorithms that have such characteristics include: penalized regression, random forest, and support vector machine.

For regression models, to prevent overfitting, a penalty needs to be incorporated to prevent the coefficients from getting too large due to the correlated functional data. In the case of logistic regression, there is a binary outcome variable, for instance risk versus non-risk variants. A continuous probability outcome can also be obtained.

Random forest constructs a series of decision trees to separate two classes (risk versus non-risk variants. The resulting model is created by averaging the decision trees together (Malley et al., 2011). A subset of features (functional annotations in the context of genetic variant prioritization) is considered at each node in the tree. In the case of a simple presence or absence of the sample with the feature, there are only two decisions at the node. A simple example could be at a node, if a variant falls into a splice site, it will go to one side, and if it does not then it will go to the other side. The algorithm will rank the features based on how many times they appear in the tree, and thus how important they are in differentiating the two classes.

Support vector machine separates data using a hyperplane in multi-dimensional space. The shape of the decision boundary depends on the kernel function (Malley et al., 2011). The most basic kernel is linear, where the samples are separated linearly (for instance, the risk separated from the non-risk variants in the realm of genetic variant prioritization). However, more mathematical functions, such as polynomials, can be used to separate data as well (Ben-hur and Weston, 2007).

All algorithms have their advantages and disadvantages. Regression has the advantage over other algorithms that the importance of the predictor variables are easy to determine

by means of the magnitude of the beta coefficient assigned to each predictor variable. However, that being said, regression is not scale-invariant, and thus scaling or not scaling the predictor variables will affect the model (Abdi et al., 2013).

Random forest has a bias to include continuous features into the model (Strobl et al., 2007). However, this bias can be mitigated by selecting appropriate parameters (for instance, the minimum number of samples at which to stop constructing the tree).

There are packages written in freely available coding languages to perform all of these algorithms (see **Table 1.5**).

Table 1.5. Non-exhaustive selection of available packages for performing some statistical learning algorithms in R and Python

	R package	Python package
Penalized regression	glmnet	LogisticRegression in scikit-learn
Random forest	e1071, party, randomforest	RandomForestClassifier in scikit-learn
Support vector machine	e1071	svm in scikit-learn

These algorithms can be applied to genetic variant prioritization. The input can be a set of variants: some labelled as risk variants and other labelled as non-risk variants, and all the variants are annotated with their functional information. These data can then be fed to the algorithm, which will consequently produce a prediction score for each variant (the probability of it being a risk variant) and a variable importance measure for each annotation demonstrating how important it is in differentiating the risk from the non-risk variants (**Figure 1.5**).

Figure 1.5. Input and output variables for statistical learning algorithms in the context of genetic variant prioritization

For all of these algorithms, it is important to train the data (i.e. create the model) in one dataset, and then apply it and test its accuracy in an independent dataset (Smialowski et al., 2010). A model may be highly accurate in differentiating the risk variants from non-risk variants in the training dataset, but that does not necessarily mean that such a model is flexible enough to be applied to new data. A model that has high accuracy in training data, but does very poorly when applied to a novel dataset, is referred to as being over-fit. This model is too specific and sensitive to the fine-scale characteristics of the training set, which makes it uninformative in any other dataset. Thus, over-fit models are not useful as they do not have broad applicability.

For the test set, there are certain predictive accuracy measures (statistical tests and visualization techniques) that are most appropriate for evaluating data-trained models for prioritizing genetic risk variants. These data tend to have the characteristic of consisting of imbalanced classes: a very high proportion of non-risk variants and a small proportion of risk variants. This class imbalance, and other factors unique to genetic data (for instance linkage disequilibrium, allele frequency, etc.), warrant exercising caution when interpreting the results of predictive accuracy measures that are applied to such models. I undertook a thorough investigation of such measures (**Chapter 4**).

Referring back to the methods in **Table 1.4**, Gagliano et al. (2014a), Ritchie et al. (2014), and Kircher et al. (2014) all have data-trained classifiers. They use a supervised statistical learning algorithm (i.e. algorithm is given the task to differentiate between assigned risk

variants and non-risk variants) to create a model that assigns the functional annotations various degrees of importance relative to each other, which is based on an annotated dataset containing both risk and non-risk variants. The model can then be used to generate prediction values or scores for each genetic variant on which the model is applied (probabilities of how likely the variant will belong in the risk variant class). These methods differ in the algorithm, annotation set, and how the risk and non-risk are defined. These methods are described in detail in **Chapter 5**. The method cited as Gagliano et al. (2014a) is described in **Chapter 3**.

Iversen et al. (2014), and Pickrell (2014) are in the context of a Bayesian framework. Both consider two Bayes factors: Bayes factors for annotation and Bayes factors for association. My method Gagliano et al. (2014a) (extending on the backbone of the method first presented in Knight et al. (2011)) can also be applied in a Bayesian framework. However, there are fundamental differences in the Bayesian methods for my work compared to these two others. Gagliano et al. and Iversen et al. calculate the Bayes factors for annotation and the Bayes factors association in separate data, whereas Pickrell calculates both sets of Bayes factors on the same dataset. With regard to Gagliano et al. and Iversen et al., the former uses a regularized logistic regression called elastic net, whereas the latter employs a Bayesian shrinkage method. For dealing with LD among the genetic variants, Gagliano et al. applied the annotations from variants in LD to the GWAS variant, whereas Iversen et al. tested each LD-block separately. Iversen et al. defined LD-blocks as the SNP plus its LD partners. Again, my method will be further described in **Chapter 3**.

In summary, many of these genetic variant functional annotation and/or prioritization methods have been made available as either online or downloadable tools to be run on a local system, making these tools accessible for researchers to integrate into their association analyses. Some of these methods simply show the overlap of variants with various functional annotations, while others are specifically meant to be applied to

GWAS summary data, and still others are data-trained producing a prediction score applicable to numerous contexts.

1.17 Impact

A better understanding of the genetic architecture of complex disease leads to a more comprehensive understanding of the biological pathways responsible for the pathology. This enhanced knowledge is the driving force enabling the development of novel therapies and personalized treatments to provide relief for millions of people who suffer worldwide. The evidence discussed here of enrichment of known risk variants with functional data suggests that the use of existing functional data can help illuminate the genetic factors involved in complex disease *in silico*. More variants are being identified (for instance, through sequencing projects such as the 1000 Genomes Project), and more functional genomic data is constantly being made available (for instance, through the Roadmap Epigenomics Project). The challenge now is to integrate these data together in order to identify novel risk variants.

Chapter 2
Thesis Aims and Hypotheses

2

2.1 Aims and Hypotheses

The primary aim of this thesis is to develop a prediction model using statistical learning that is able to differentiate between genetic variants that increase or decrease one's chance of developing a complex illness or trait from those that are not associated with such an outcome. Each genetic variant is given a probability (between 0 and 1) for how likely it is to be a disease-associated variant. This aim can be used to identify novel disease-implicated loci, as well as the variant causing the phenotypic effect at a known locus. Alongside, I compare my method to other similar existing methods (which use different statistical learning algorithms, different functional annotations, and different definitions of risk variants). I conduct a thorough comparison of the respective algorithms and functional annotation sets to determine the combination with the best predictive accuracy by exploring various predictive accuracy measures. Finally, I perform analyses of a new annotation for prioritizing associated variants in the GWAS.

The specific hypotheses tested are the following:

1) A method can be developed to incorporate functional annotations to predict risk genetic variants defined as those that are associated with a complex disease/trait in humans.

2) By combining different statistical learning algorithms and functional annotation sets that exist in the literature, a more accurate model for genetic risk variant prioritization can be created.

3) A novel annotation based on allele-specific methylation is a relevant annotation to include for genetic variant prioritization.

2.2 Structure of the thesis

This thesis is composed of four studies. In the first study, I describe in detail my method for prioritizing genetic risk variants. I then investigate statistical and visualization techniques that are appropriate in the context of assessing the accuracy of methods for genetic variant prioritization based on functional genomic information. Following that, I provide a comparison of my prioritization method with two other methods. I use my observations of the most informative measures from my predictive accuracy investigation to assess the various models. The final study focuses on a novel type of functional information that can be incorporated into the prioritization procedure: allele-specific methylation.

Chapter 3
A New Method to Prioritize Genetic Risk Variants using Functional Information

This chapter is modified from the following: **Gagliano SA,** Barnes MR, Weale ME, Knight J (2014) A method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. PLoS ONE 9: e98122.

3

3.1 Abstract

The increasing quantity and quality of functional genomic information motivate the assessment and integration of these data with association data, including data originating from genome-wide association studies (GWAS). We used previously described GWAS signals ("hits") to train a regularized logistic model in order to predict SNP causality on the basis of a large multivariate functional dataset. We show how this model can be used to derive Bayes factors for integrating functional and association data into a combined Bayesian analysis. Functional annotations were obtained from the Encyclopedia of DNA Elements (ENCODE), from published expression quantitative trait loci (eQTL), and from other sources of genome-wide characteristics. We trained the model using all GWAS signals combined, and also using phenotype specific signals for autoimmune, brain-related, cancer, and cardiovascular disorders. The non-phenotype specific and the autoimmune GWAS signals gave the most reliable results. We found SNPs with higher probabilities of causality from functional annotations showed an enrichment of more significant p-values compared to all GWAS SNPs in three large GWAS studies of complex traits. We investigated the ability of our Bayesian method to improve the identification of true causal signals in a psoriasis GWAS dataset and found that combining functional data with association data improves the ability to prioritize novel hits. We used the predictions from the penalized logistic regression model to calculate Bayes factors relating to functional annotations and supply these online alongside resources to integrate these data with association data.

3.2 Introduction

Genome-wide association studies (GWAS), which investigate the association between genetic variation and phenotypic traits, have identified many loci associated with human diseases (Hindorff et al., 2010). However, despite considerable advances, much of the

estimated heritability remains unaccounted for. Purcell et al. (International Schizophrenia Consortium et al., 2009) showed that single nucleotide polymorphisms (SNPs) from GWAS with sub-genome-wide significant p-values account for a considerable proportion of the variance in independent samples suggesting that they are enriched for causal SNPs or their proxies. The issues of small sample size, low minor allele frequency, and lack of linkage disequilibrium (LD) between genotyped SNPs and the un-genotyped causal SNPs present challenges to detecting truly causal variants among near-significant genetic associations.

The central challenge in the interpretation of genetic associations lies in the processing and meaningful integration of a hugely diverse range of information. Having derived a score for a region containing a candidate variant, it has to be integrated with association evidence. We proposed the use of empirically derived weightings within a Bayesian framework (Knight et al., 2011). More recently Schork et al. suggested the use of stratified False Discovery Rate (sFDR) and Darnell et al. proposed multi-thresholding in a manner that they say is equivalent to varying the significance threshold at each marker depending on the prior information (Darnell et al., 2012; Schork et al., 2013). In order to implement these approaches one needs to define appropriate weights. For instance, Schork et al. (2013) used an LD-weighted scoring algorithm, and Kindt et al. (2013) recently published a multivariate logistic regression approach. However, neither of these approaches is easily scalable to the very large number of functional annotations that are becoming available.

The primary objectives of this study are to describe an empirically justified method to identify which functional annotations are best correlated with GWAS hit SNPs, to provide clues to the etiology of such traits, and to develop and implement a method to incorporate functional annotations with statistical information in association studies. To achieve these objectives we use a machine learning approach, elastic net (a regularized logistic regression), to predict causality of a SNP based on information from 439

functional annotations. We explore models based on all GWAS significant SNPs and also subsets of significant SNPs selected on the basis of phenotype and p-value. Functional annotations are considered individually or in groups. We report a) the accuracy of the predictions to demonstrate the utility of the method and to investigate the behaviour of the different models, b) the frequency, correlation between and coefficients of the functional annotations providing insight about their functional relevance to disease, c) a prediction score for each SNP, and d) details of how to combine this score with association statistics in a formal Bayesian framework.

We provide online scripts that can be employed so the method can be used by other researchers using additional functional annotations (http://www.camh.ca/en/research/research_areas/genetics_and_epigenetics/Pages/Statistical-Genetics.aspx). For the best models we provide the probability of causality (the prediction score) for each SNP, the corresponding Bayes factor (BF_{annot}) and scripts to combine BF_{annot} with GWAS association signals.

3.3 Methods

3.3.1 Representative GWAS SNPs

To represent the characteristics of a typical GWAS panel, markers from the Affymetrix Genome-Wide Human SNP Array 6.0, the Illumina Human1M-Duo Genotyping BeadChip, and the Illumina HumanOmni1-Quad BeadChip were downloaded from the UCSC genome browser, using the table browser tool (Karolchik et al., 2004). The union of these three arrays consisted of 1,936,864 unique SNPs from the 22 autosomes. Because of its unique LD and genic properties, the MHC region (chr6:29,624,809 - 33,160,245 on build 37) was excluded from downstream analyses.

LD proxies or "tagging" SNPs ($r^2 \geq 0.8$) for the GWAS panel SNPs were identified using VCFtools (Danecek et al., 2011) based on data from the (N=379) Europeans (Phase I,

version 3, March 14, 2012) in the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010).

GWAS "non-hits" were defined as all those SNPs in our union GWAS set, which were neither a GWAS "hit" (see below), nor in high LD ($r^2 \geq 0.8$) with a GWAS hit.

3.3.2 GWAS hits

To obtain a set of SNPs (and their LD proxies) with good prior evidence of causality, we downloaded the Catalogue of Published Genome-wide Association Studies from the National Human Genome Research Institute (NHGRI) (http://www.genome.gov/gwastudies) (Hindorff et al., 2010) on August 6, 2013. This catalogue contains a list of SNPs that have been shown to be associated with a particular trait in a GWAS at a suggestive p-value $<10^{-5}$. There were 13,708 entries from a total of 1,664 different studies with publication year ranging from 2005 to the date of download (**Figure 3.1**). We removed SNPs in the Catalogue that were not present in the representative GWAS set defined above, and similarly removed SNPs on the sex chromosomes or in the MHC region, and a total of 8,405 SNPs remained.

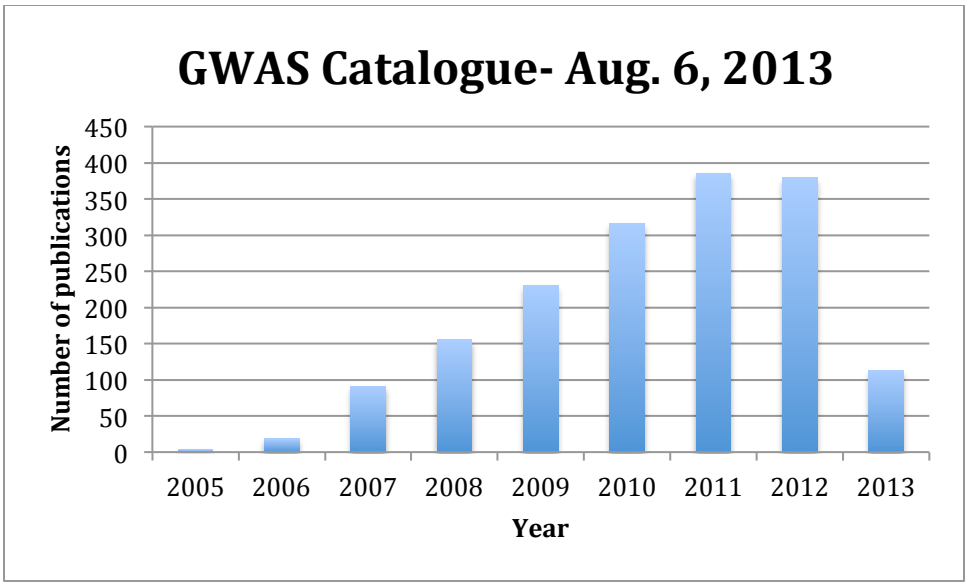


Figure 3.1. Number of publications with data in the GWAS Catalogue.

Regardless of whether a publication had one or several variants in the Catalogue it was only counted once.

All SNPs in our GWAS hit and GWAS non-hit sets, along with all their LD proxies, were annotated with all the functional annotations defined below. Each GWAS hit and non-hit SNP was then given the maximum value for each functional annotation found across all its LD proxies.

3.3.3 Functional annotations

We acquired functional data from a variety of sources (**Table 3.1**). A full list is provided in **Table S1** available from the online PLOS ONE publication: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0098122. In brief, the GTEx eQTLs have been separated into 7 samples (separated by study and for one of the studies, also by tissue). The three histone marks are separated into 18 cell types each. There are 148 transcription factor binding sites. There are DNase I data from 100 cell types from Duke University data and 122 from the University of Washington. Much

of the data was downloaded from the UCSC Genome Browser using the table browser tool (Karolchik et al., 2004). Additionally, a substantial proportion of the data was derived from the Encyclopedia of DNA Elements (ENCODE) Project Consortium, which developed and implemented a range of experimental techniques with the aim of identifying the functional regions of the human genome, particularly including non-coding regions (The ENCODE Project Consortium, 2011). Data from this project that were used included transcription factor binding sites (TFBSs), three histone modifications (H3K4Me1, H3K4Me3, H3K27Ac), and DNase I hypersensitive sites. H3K4Me1 is associated with enhancers and DNA regions downstream of transcription starts, and often found near regulatory elements; H3K4Me3 is associated with promoters active or poised to be active, and often found near promoters; H3K27Ac thought to enhance transcription possibly by blocking repressive histone mark H3K27Me3, and often found near active regulatory elements. The technologies for identifying the functional annotations mentioned above were chromatin immunoprecipitation followed by sequencing (ChIP-seq).

DNase I hypersensitive sites are regions in the genome with high affinity of being cleaved by the DNase I enzyme. The University of Washington (UW) group identified DNase I hypersensitive sites using Digital DNase I. This method involves DNase I digestion of intact nuclei, isolation of DNase I "double-hit" fragments, and direct sequencing of fragment ends. Peaks are regions that are enriched in the captured fraction of the DNA suggesting they are occupied by the protein of interest (any score > 0). The DNase I hypersensitive sites from the Duke University group were identified using DNase I assays. We used a binary variable to indicate whether a SNP was within a peak.

Two types of conservation scores from 46 placental mammals (PhyloP and PhastCons) were incorporated. Both PhyloP and PhastCons scores are derived using phylogenetic hidden Markov models. These two measures have their own advantages. PhyloP scores

do not take into account conservation at neighbouring sites, whereas PhastCons estimates the probability that each nucleotide belongs to a conserved element.

Expression quantitative trait loci (eQTLs), which are variants that are correlated with gene expression, were included. In particular those that fall within 2Mb (+/-1Mb upstream and downstream) (cis-eQTLs) of the gene of interest were used. These data were derived from the NCBI-hosted GTEx Browser (http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi) (Montgomery et al. 2010, Schadt et al. 2008, Gibbs et al. 2010, Stranger et al. 2007) and the UK Brain Expression Consortium (www.braineac.org) (Trabzuni et al., 2011).

Summary information concerning the location or function within a gene (coding-non-synonymous, coding-synonymous, splice site, untranslated regions, etc.) was derived from dbSNP (version 137). Non-synonymous SNPs, were classified as those SNPs with one of the following annotations: stop-gain (nonsense), missense, stop-lost, frameshift or inframe indel. Splice site regions were defined as being within five base pairs upstream and five base pairs downstream of the exon start site or the exon end site. The UCSC gene table was used to determine the exon start and end sites. The UCSC gene table is comprised of a set of gene predictions based on data from RefSeq, GenBank, the Consensus Coding Sequence (CCDS) variable, Rfam, and the Transfer RNA Genes variable. (This track has since been replaced by Gencode tracks.) Additional annotations used were 3' targets for microRNA (miRNA), and also transcription start sites as described by Gencode (Harrow et al., 2012). As miRNA targets are known to be substantially over-predicted, we used a conservative miRNA target dataset based on conserved mammalian microRNA regulatory target sites in the 3' UTR regions of Refseq Genes, as predicted by the TargetScan algorithm (Human 5.1) (Lewis et al., 2005).

Table 3.1. Summary of functional annotations

Functional characteristic analysed	Description	Number and detail of measures used in the analysis*	
		Clumped	Separated
ENCODE data			
UW DNase I hypersensitive sites	Data from digital DNaseI methodology, Replication 1 samples; ("peaks")	N/A	122
Duke DNase I hypersensitive sites	Positions of open chromatin by FAIRE and ChIP-seq experiments; ("peaks")	N/A	100
DNase Clusters (v2)**	Stringent (FDR 1% threshold) for "peaks" of DNase I hypersensitivity from uniform processing by the ENCODE Analysis Working Group of data from UW and Duke	1	N/A
Txn Factor ChIP	Transcription factor binding sites (TFBS) from ChIP Seq experiments; ("peaks")	1 (presence or absence in any TFBS)	148 (separated by TF, but not by cell type due to sparse data)
Broad Histone – H3K4Me1, H3K4Me3, H3K27Ac	All are assayed using ChIP-Seq; ("peaks")	3 (each histone mark grouped by the 18 cell types and/or conditions)	54 (each histone mark separated by cell type and/or conditions)
Conservation			
PhyloP	Average scores can be calculated as the sum of scores divided by the number of valid data values in the block (scores range from 0.1 to 2.2910)	1	1
PhastCons	Average scores can be calculated as for PhyloP (scores range from 0.1 to 1.0 in this dataset)	1	1
Expression quantitative trait loci			
eQTL- GTEx	cis-eQTLs, $p<1x10^{-5}$ cut-off for variants within 2Mb of the expressed gene.	1 (any eQTL)	7 (separated by dataset)
eQTLs - UK Brain	cis-eQTLs, FDR<1% cut-off for variants within 2Mb of the expressed gene.	1	1
Other characteristics			
UCSC Genes	UCSC known Gene	1	1
Splice sites	Splice site region defined as -5 to +5 range around exon starts & exon ends of UCSC Genes	1	1
Nonsynonymous SNPs	Coding Nonsynonymous SNPs defined as stop-gain (nonsense), missense, stop-lost, frameshift or inframe indel	1	1
TS miRNA sites	Conserved mammalian microRNA regulatory target sites for conserved microRNA families	1	1
Gencode transcription start sites	Based on the GENCODE Genes variable (version 17, June 2013)	1	1

* All SNPs are annotated in a binary fashion indicating the presence or absence of a functional annotation, except for the conservation scores, for which the SNPs are assigned a quantitative score.
** The DNase Clusters v2 file was created by combining the UW and Duke DNase I data that have been uniformly processed and replicates merged. Stringent (FDR 1% thresholded) peaks of DNase I hypersensitivity from uniform processing by the ENCODE Analysis Working Group were applied. Grouping the UW and the Duke DNase I hypersensitive variables are not equivalent to the DNase Clusters v2 file, and thus we used the latter to represent DNase I hypersensitive sites in the clumped analysis due to the substantial efforts made to combine the data meaningfully.

All SNPs in our GWAS hit and GWAS non-hit sets, along with all their LD proxies, were annotated with all the functional annotations defined above. Each GWAS hit and non-hit SNP was then given the maximum value for each functional annotation found across of all its LD proxies.

3.3.4 Tests for functional enrichment

Counts of GWAS hits and non-hits were categorized by annotation value and compared using Fisher's exact test. To verify that results were not unduly influenced by correlations (LD) among observations, we also conducted analyses in which genetic variants were "pruned" so that all SNPs have $r^2<0.8$ with all other SNPs. The results of these analyses were very similar (data not shown).

Heat maps were constructed using R (R Core Development Team, 2008) to compare correlations among the various functional annotation.

3.3.5 Regularized logistic regression via elastic net

As a start, we performed a univariate analysis for the 14 clumped functional annotations, and found that all were significantly related to the status of a GWAS hit or not ($p<0.005$). We used a regularized form of logistic regression known as elastic net to predict GWAS hit versus non-hit status on the basis of the functional annotations we had collected. Elastic net is a form of machine learning first described by Zou and Hastie (2005), and is implemented in the glmnet package (Friedman et al., 2010) in R. Briefly, regularization is achieved via the subtraction of a penalty term from the log-likelihood prior to maximization. The penalty term includes both a "lasso-like" L1 component (the sum of the absolute values of all fitted coefficients) and a "ridge-like" L2 component (the sum of squares of all fitted coefficients). Two parameters, alpha and lambda, determine the relative importance of the L1 versus the L2 term (alpha), and the overall importance of the penalty term in the maximization (lambda). Appropriate values for these parameters were found by 10-fold cross-validation of the training set (see below).

Due to the unbalanced nature of the data (many more GWAS non-hits than hits) we employed a weighting procedure in the logistic regression to balance the accuracy of prediction in both types of markers. We weighted all hits by (Nhits+Nnon-hits)/2Nhits and all non-hits by (Nhits+Nnon-hits)/2Nnon-hits, where Nhits and Nnon-hits denote the number of hits and non-hits, respectively, in the training set. This procedure has the effect of equalizing the importance of hits and non-hits in the logistic regression.

We randomly selected 60% of our GWAS hits and non-hits to form our training set. The remaining 40% of the data (the test set) was used to assess the performance of the model using ROC curves and other measures. We repeated the machine learning modifying the percentage of the data used in the training and test sets, and all splits produced similar results (**Figure 3.2**).

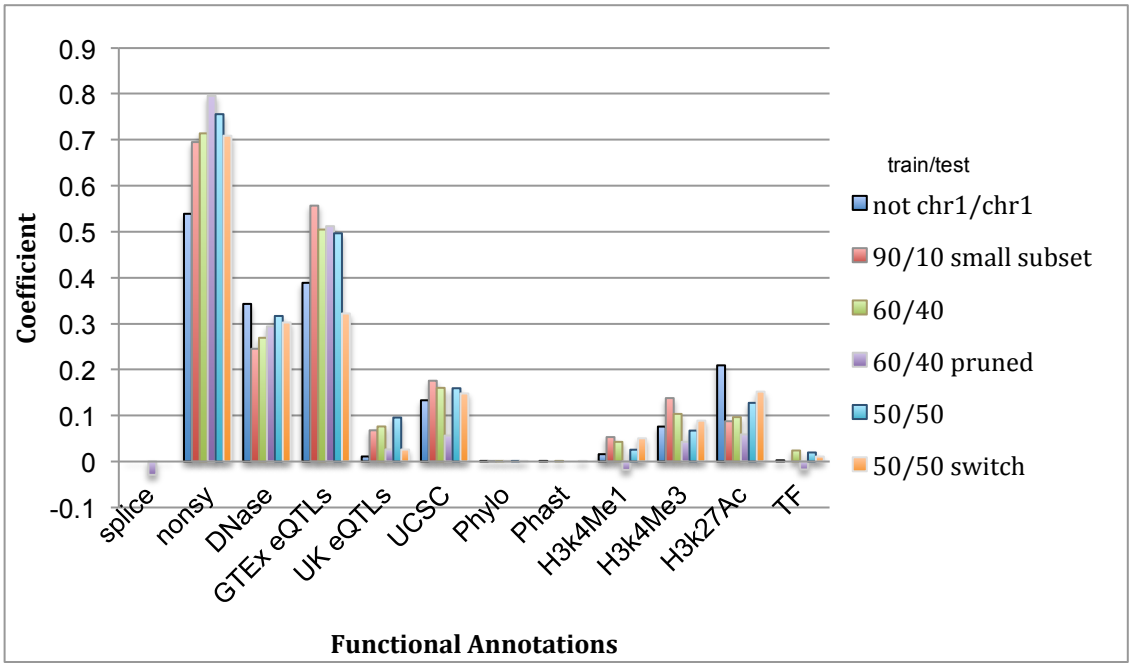


Figure 3.2. Coefficients for functional annotations in the clumped analysis for different training and test set proportions

Comparison of beta coefficients that resulted from machine learning in the clumped non-phenotype specific analysis for various classifications of the training and test sets. [splice= splice sites, Nonsy= nonsynonymous SNPs, DNase= DNase I hypersensitive sites, GTEx eQTLs= cis-eQTL data from the GTEx Consortium, UK eQTLs= cis-eQTL data from the UK Brain Consortium, Phylo= PhyloP conservation, Phast= PhastCons conservation, H3K4Me1= H3K4Me1 histone modification, H3K4Me3= H3K4Me3 histone modification, H3K27Ac=H3K27Ac histone modification, TF= transcription factor binding sites]

3.3.6 Sensitivity analysis- elastic net

To diminish the possibility that the models are over-fit since the training of the data and tuning of the parameters were conducted on the same set, we created a 70%/30%, split where the 70% was further split into 60% and 40% for training the coefficients and tuning the parameters, respectively. The remaining 30% was used to test the model. Additionally, we examined the stability of the beta coefficients when assigning the data to training the test sets using different random number generators.

3.3.7 Predictive accuracy

We employed three methods to determine which models had the best predictive accuracy: ROC curves, positive predictive values, and histograms of the predicted values from the models.

ROC curves show the sensitivity and specificity of a fitted model. Sensitivity is the probability of the model providing a true positive result (identifying a true GWAS hit in the test set). Specificity is the probability of the model providing a true negative result (identifying a true GWAS non-hit in the test set). An AUC of 0.5 indicates a model of no predictive value, while an AUC of 1 indicates perfect predictive power. The ROC curves were created using the ROCR package (Sing et al., 2005) in R.

ROC curves do not reflect how well a model performs within each class given unbalanced data (a very large number of non-hit SNPs compared to hits). To capture this aspect we also investigated positive predictive values (PPVs), the proportion of SNPs with predicted probabilities of causality above a certain threshold (we investigated thresholds of 0.5, 0.6, 0.7, 0.8 or 0.9) that are true GWAS hits in the test set. Finally, we visualized class separation with histograms of the predicted probabilities of causality by class.

3.3.8 Definition of functional variables and GWAS hits

A variety of functional annotations were investigated as input variables. One, defined as the "clumped" analysis, featured groups of functional annotations, which were collapsed into a single summary variable. The "separated" analysis worked on all functional annotations individually.

We performed phenotype specific analyses in which the analyses outlined above were carried out using phenotype specific GWAS hits as classifiers. An autoimmune list, a brain-related list and a cardiovascular list were created using the GWAS Catalogue

searching for terms relating to those phenotypes. Each list was then verified by an expert in the field.

Additionally, the GWAS Catalogue was divided up into categories specified by the Experimental Factor Ontology (EFO) definitions; however, due to small numbers of SNPs in each category this mode of classification is not currently feasible for most of the subsets. Only the cancer list, which was the largest disease-relevant list, was used.

Due to the small size of the lists (not including "other disease" or "other measurement", which both lack biological relevance), it is not feasible to use the EFO classifications. **Table 3.2** shows the number of GWAS hits that fall into each category. The numbers provided in the table are inflated as they assume that all of those SNPs are present on the GWAS arrays in our analysis and that none of them are in the MHC region (which was excluded for the machine learning). Thus, the lists for training and testing are around 100 SNPs less than the listed values.

There were no results for the GWAS list for "biological processes" (i.e. the betas were all zero), so machine learning on other lists with a smaller number of SNPs was not performed. Machine learning was also not run on the lists that lacked biological relevance even if they were larger than the list for "biological processes": for example: "other disease", "other measurement", and "other trait".

Table 3.2. EFO phenotype specific GWAS lists

Phenotype	N in GWAS Catalogue (Aug. 6, 2013)
Biological process	616
Metabolic disease	389
Mental disease	827
Immune disease	349
Hematological Measurement	284
Digestive disease	468
Cardiovascular disease	356
Cancer	685
Body measurement	639
Nervous system	680
Other Disease	1231
Other measurement	3216
Other trait	211
Drug response	593

We defined two sets of GWAS hits for downstream analysis, one based on a weak significance threshold of $p<10^{-5}$ and one based on a strong significance threshold of $p<5x10^{-8}$, as reported in the NHGRI GWAS Catalogue. An additional analysis was undertaken in which hits were defined as the subset of the hits from the $5x10^{-8}$ non-phenotype specific analysis that were not also defined as hits in at least one of the phenotype-specific analyses assessed. Note, to view the distribution of the hits used in the $5x10^{-8}$ non-phenotype specific analysis, a Manhattan plot was constructed (**Figure 3.3**).

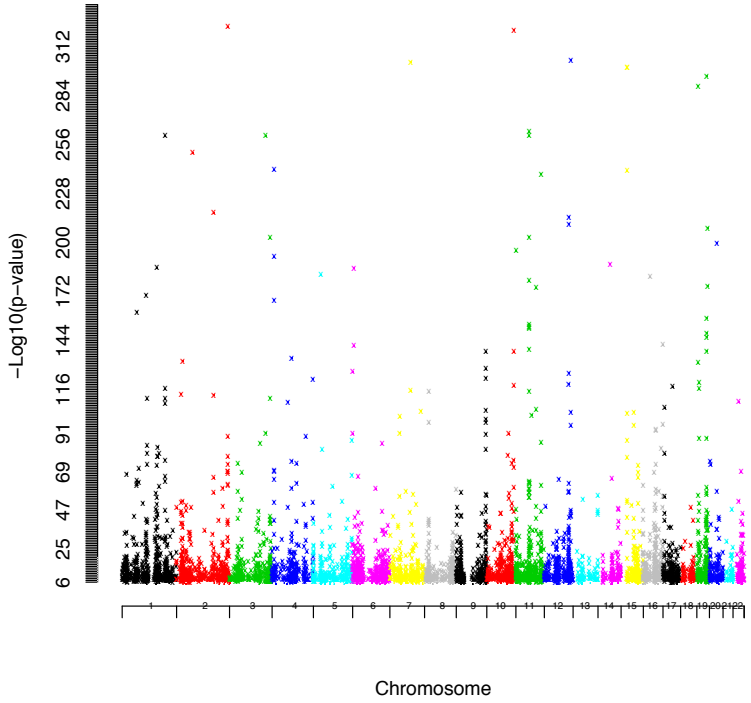


Figure 3.3. Manhattan plot of the hits used for the non-phenotype specific analysis ($p<5 \times 10^{-8}$)

SNPs are pruned and MHC region hits have been removed, as described.

3.3.9 Sensitivity analysis- classification

An analysis was also undertaken in which hits were defined as the subset of the non-phenotype specific 5×10^{-8} hits minus those hits used in the phenotype-specific analyses (autoimmune, brain-related, cancer and cardiovascular).

3.3.10 Derivation of Bayes Factors

Bayesian analysis provides the most suitable framework for combining functional annotations (here referred to as "annotation data"), with evidence from an association

study ("association data") (Stephens and Balding, 2009). We expand on our previous empirically-based approach to the calculation of Bayes factors for annotation (Knight et al., 2011) to allow multiple functional annotations to be considered simultaneously. The posterior odds (O post) of causality for a trait of interest at a given SNP are given by the ratio of the conditional probability of causality, given the annotation and the association data, to the conditional probability of non-causality:

$$O_{post} = \frac{P(Causal \mid AnnotData, AssocData)}{P(NotCausal \mid AnnotData, AssocData)}$$

If we assume the annotation data and association data are independent once conditioned on causality, then the posterior odds become:

$$\frac{P(Causal)}{P(NotCausal)} \times \frac{P(AnnotData \mid Causal)}{P(AnnotData \mid NotCausal)} \times \frac{P(AssocData \mid Causal)}{P(AssocData \mid NotCausal)}$$

These three products are, respectively, the prior odds before seeing any association and annotation data (O prior), the Bayes factor for annotation data (BF_{annot}) and the Bayes factor for association data (BF_{assoc}). We note that this factorization implies that, while functional annotations are allowed to be enriched (or impoverished) for causal SNPs relative to non-causal SNPs, the enrichment pattern is assumed to be the same for rare versus common causal SNPs, and for low-effect size versus high effect size causal SNPs. We accept that this is an imperfect approximation, and it assumes among other things that SNPs are either causal or non-causal when in reality their effect size can be arbitrarily close to zero, but we note that the main limitation of our approach lies with the small number of GWAS hits available to us, and subdividing these still further according to allele frequency and effect size would be problematic. We also note that by "causal" what we actually mean is "causal or in high LD with a causal variant", as both the association data and the annotation data (as defined in our study) are affected by LD proxies.

In our previous study (Knight et al., 2011), we noted that if one assumed that (1) all hits in the NHGRI GWAS Catalogue were truly causal; and (2) functional annotation enrichment patterns were the same for these known hits as for future undiscovered truly causal SNPs; then an empirically based estimate for BF_{annot} for a single binary functional annotation would simply be the ratio of its frequency in GWAS hit versus non-hit data. Here we note that if we start with the same two assumptions, and further assume that a true (but unknown) logistic model exists that relates a set of functional annotations (which can be either binary or quantitative) to the probability that a SNP is truly causal, then one reasonable approach to estimating that logistic model would be via regularized logistic regression as described above. Once fitted, the estimated odds of causality to non-causality, obtained from the GWAS hit and non-hit datasets, need only be multiplied by the prior odds of non-causality in these dataset (i.e. the ratio of the weighted sample sizes of GWAS non-hits to GWAS hits in these data) in order to obtain the Bayes factor for annotation. Here, we chose to weight hits and non-hits to appear of equal size, and thus our estimate for BF_{annot} is obtained directly as the estimated odds of causality to non-causality from the regularized logistic regression.

Methods for estimating BF_{assoc} from association data are reviewed by Stephens & Balding (2009). Here, we use the convenient approximation described by Wakefield (Wakefield, 2007).

3.3.11 Investigating the model in the context of known GWAS

To investigate the relevance of the predictions in a variety of disorders we looked at the p-value distribution of SNPs according to their functional class in large GWAS datasets with a substantial fraction of GWAS significant findings. Quantile-quantile plots were constructed for each study with multiple lines corresponding to SNPs binned according to their predicted value. Predicted values were those derived from the non-phenotype specific clumped model in which GWAS hits were defined as those SNPs in the GWAS Catalogue with p-values of less than $5x10^{-8}$. We expected those SNPs with higher

predicted values to be enriched with GWAS SNPs with more significant p-values, whereas those SNPs with lower predicted values would be enriched with less significant p-values compared to all SNPs in the GWAS.

We also selected some SNPs shown to be associated in a large psoriasis meta-analysis which had not been identified in a previous GWAS study (Strange et al., 2010; Tsoi et al., 2012). We then determined the effect on the rank of their Bayes Factors in the previous study derived either using association data or both association data and functional annotations.

3.4 Results

3.4.1 Functional enrichment in GWAS hits

Frequencies of functional annotations in GWAS hits compared to non-hits were compared using Fisher's exact test. Our analyses indicate that GWAS hits are enriched for most functional annotations compared to GWAS non-hits, except for splice sites and micro RNA (miRNA) targets, perhaps due to the very low frequency of these two classes of functional annotations compared to the others (**Table 3.3**).

Table 3.3. Summary statistics for the functional annotations in the clumped non-phenotype specific analysis

Description	Frequency of annotation in GWAS hits	Frequency of annotation in GWAS non-hits	p value (Fisher's exact test)	Odds Ratio	95% Confidence interval
splice	0.002	0.002	0.142	1.26	0.78 – 2.02
non-synonymous	0.022	0.007	2.38E-38	3.10	2.67 – 3.59
DNase Clusters	0.193	0.141	1.87E-39	1.46	1.38 – 1.54
GTEx eQTLs (all 7 experiments together)	0.020	0.007	1.69E-31	2.92	2.50 – 3.41
UK brain eQTLs	0.108	0.081	2.19E-18	1.37	1.28 – 1.47
UCSC Genes	0.422	0.357	7.36E-35	1.31	1.26 – 1.27
PhyloP*	0.217	0.172	6.56E-27	1.34	1.27 – 1.41
PhastCons*	0.243	0.202	3.63E-20	1.27	1.20 – 1.33
BroadHistone-H3k4Me1	0.637	0.566	2.20E-40	1.35	1.29 – 1.41
BroadHistone-H3k4Me3	0.509	0.434	1.63E-43	1.35	1.30 – 1.41
BroadHistone-H3k27ac	0.587	0.503	1.28E-53	1.48	1.34 – 1.46
Txn Factor ChIP (if annotation for any TF)	0.511	0.456	5.26E-24	1.25	1.10 – 1.14
miRNA	1.12E-4	7.00E-5	0.116	1.70	0.24 –12.15
Gencode-Txn start sites	0.003	0.002	0.012	1.64	1.08 – 2.49

*As PhlyloP and PhastCons conservation scores were left as continuous measures, the frequencies reported for those characteristics represent the presence of a conservation score (i.e. score > 0).

The histone modification data from the Broad Institute had the highest frequencies in GWAS hits, and the lowest p-values for enrichment. Many functional annotations, most notably miRNA, were very infrequent, but the general picture was that their frequency in GWAS hits was greater than in GWAS non-hits.

We examined the correlations among the various functional annotations (**Figure 3.4** and **Figure 3.5**). The separated-variable analysis included measures of functional annotations from different cell lines as individual factors, whereas the clumped-variable analysis grouped data from different cell lines for the same functional annotation. The clumped analysis showed a strong correlation between the two conservation measures (PhyloP and PhastCons), as well as strong positive correlations among the three histone marks (H3k4Me1, H3k4Me3 and H3k27Ac), and to a lesser degree among the histone marks and transcription factor binding sites. The separated analysis revealed additional correlations among cell types investigated for the DNase I hypersensitive annotations from Duke University, and to a lesser degree among the DNase I hypersensitive annotations from the University of Washington, and between these two groups. These results highlight the issue of correlations among functional annotations, many of which simply represent the same genomic feature, for example a promoter element measured by different technologies. One advantage of elastic net as a regularized logistic regression method is its ability to accommodate highly correlated variables.

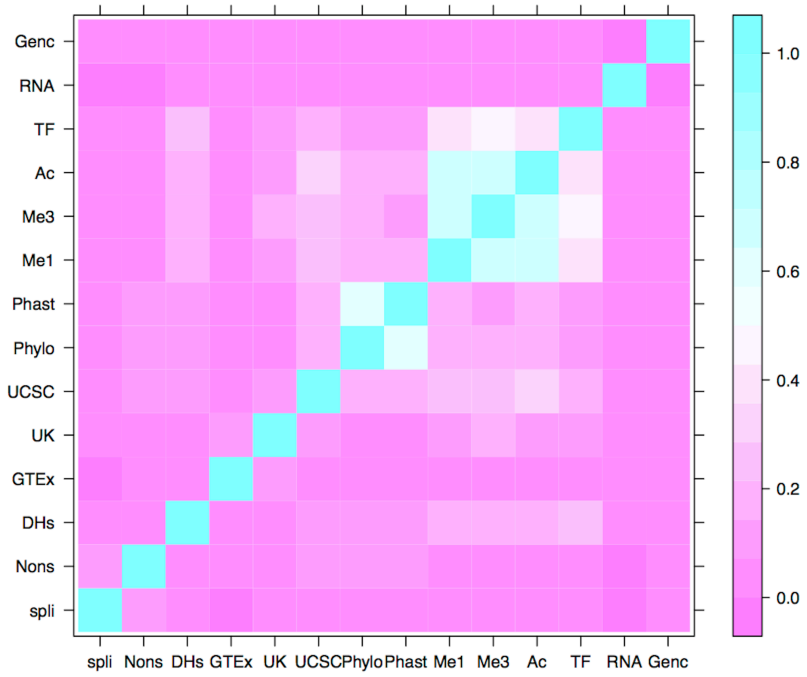


Figure 3.4. Heat map of correlations among the clumped functional annotations for 79,821 variants.

High correlations are seen between the two conservation measures PhyloP and PhastCons (represented as Phylo and Phast, respectively). Correlations are also seen among the histone modifications, H3k4Me1, H3k4Me3 and H3k27Ac (Me1, Me3 and Ac, respectively.) Transcription factor binding sites also show a correlation with the histone modifications. Note that there are negative correlations, but are all close to zero (i.e. the most negative correlation was around -0.002). [spli= splice sites, Nons= nonsynonymous SNPs, DHs= DNase I hypersensitive sites, GTEx= cis-eQTL data from the GTEx Consortium, UK= cis-eQTL data from the UK Brain Consortium, Phylo= PhyloP conservation, Phast= PhastCons conservation, Me1= H3K4Me1 histone modification, Me3= H3K4Me3 histone modification, Ac=H3K27Ac histone modification, TF= transcription factor binding sites, RNA= micro RNA targets, Genc= transcription start sites from Gencode]

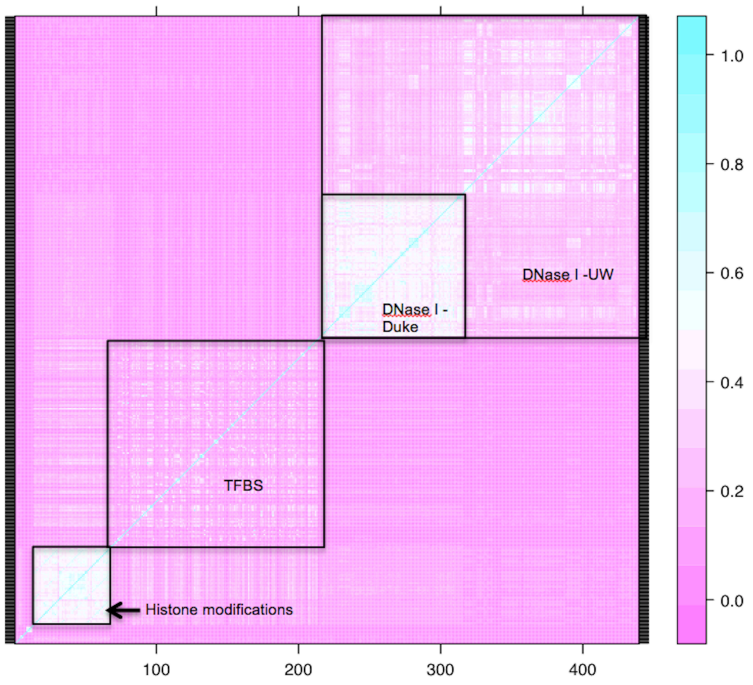


Figure 3.5. Heat map of correlations among the separated functional annotations

A full list of the numbered annotations is provided in **Table S1** (available from the online PLOS ONE publication: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0098122). The white box in the bottom left corner corresponds to high correlation among the histone modifications. The less defined white area spanning from 72 to 219 on the x axis corresponds to correlation among the transcription factor binding sites, which also show some correlation with the histone modifications. The white box from 220 to 319 on the x axis corresponds to a high correlation among the different cell types for the DNase I hypersensitivity annotation from Duke University. The less refined white box from around 320 and onwards on the x axis corresponds to the DNase I annotations from the University of Washington. The plot also shows some correlation among the DNase I annotations from both groups.

3.4.2 Sensitivity analysis- elastic net

Similar results were produced when the training and tuning were conducted in independent subsets (**Figure 3.6**), and so the 60%/40% training/test set split was pursued for the remaining analyses.

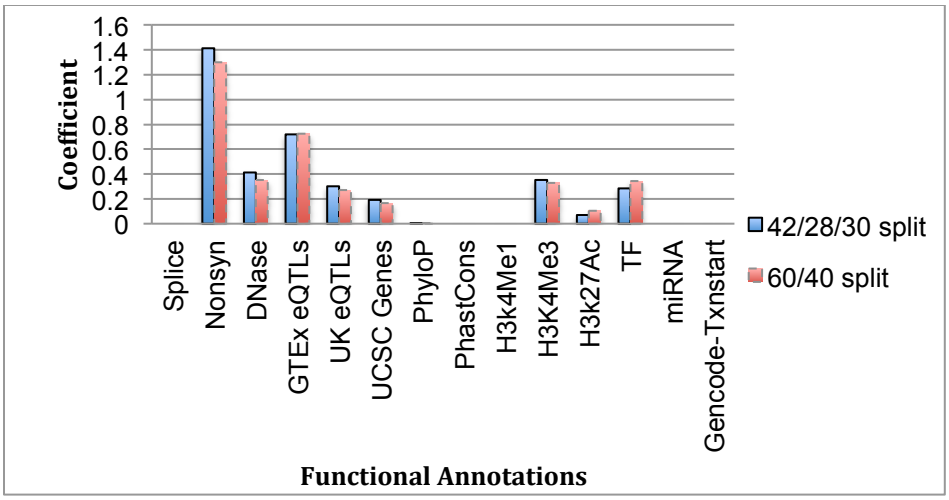


Figure 3.6. Coefficients for functional annotations in the clumped analysis when trained the model and tuned the parameters on independent sets

Comparison of beta coefficients that resulted from machine learning in the clumped non-phenotype specific analysis when using a 42%/28%/30% split for training the model, tuning the parameters, and testing the model, respectively. (The 42% and 28% refer to 60% and 40% of 70%, respectively.) This model was compared to using a 60%/40% split where the training and tuning were conducted on the same set.

The data was split into the training and test sets ten times using a random number generator, and the beta coefficients were examined. We conducted this procedure multiple times using different random numbers (i.e. starting with a different "seed"). We found that the beta coefficients were consistent for all of the functional annotations with the exception of those with the lowest frequencies. For splice sites in the autoimmune analysis (**Table 3.4**), seed2 only had one splice site that was also a GWAS hit in the training set. Thus, betas are not always reliable for the low frequency annotations. This conclusion is a caveat for the separated analysis since the frequencies for many of the annotated SNPs are very small.

Table 3.4. Beta values for "splice sites" for autoimmune clumped analysis

seed1	seed2	seed3
0.18	0	0.34

In further investigation, we assessed the relationship between the variance of betas and the frequency of the annotation in the GWAS hits for the clumped non-phenotype specific analysis. Generally, the lower the frequency in the hits, the larger the variability of the beta coefficients for that particular functional annotation (**Figure 3.7**).

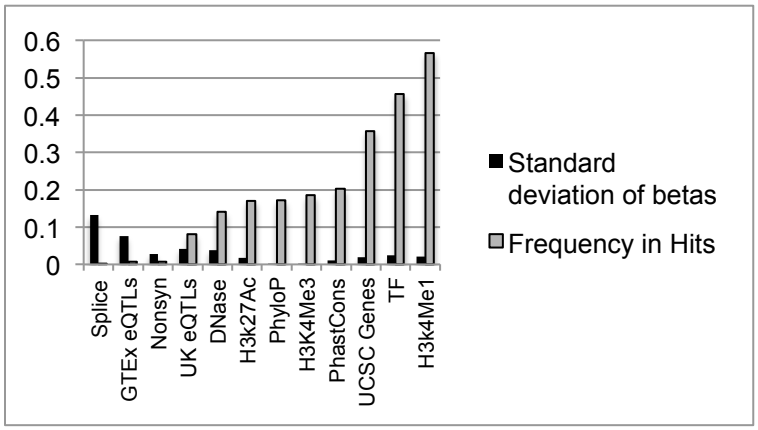


Figure 3.7. Standard deviation and frequency of functional annotations

Relationship between the standard deviation of the beta coefficients (square root of the variance of the coefficients) derived from the machine learning performed 10 times using 10 different seeds in the random number generator that distributes the SNPs into the training and test sets, and the frequency of the functional annotations in the GWAS hits. Note that the two lowest frequency annotations are not shown.

Next, we investigated whether the betas would be stabilized among the different seeds if all functional annotations were forced to be included in the model, which can be achieved through ridge regression. Ridge regression was performed for 10 different seeds, but the variability of the betas seen when using elastic net persisted (**Figure 3.8**).

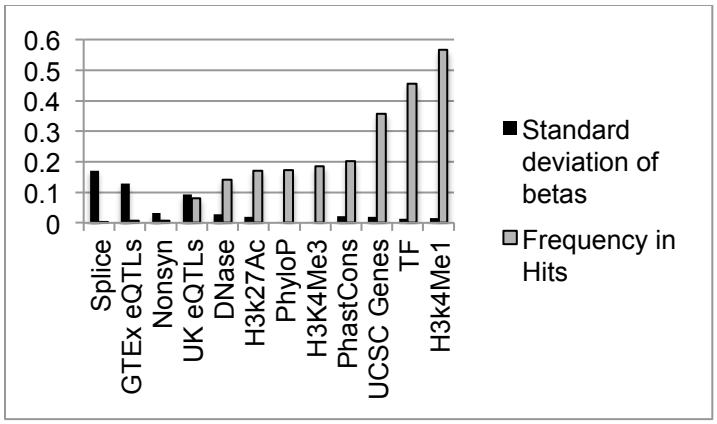


Figure 3.8. Standard deviation from ridge regression and frequency of functional annotations

Relationship between the standard deviation of the beta coefficients (square root of the variance of the coefficients)derived from the ridge regression performed 10 times using 10 different seeds in the random number generator that distributes the SNPs into the training and test sets, and the frequency of the functional annotations in the GWAS hits. Note that the two lowest frequency annotations are not shown.

3.4.3 Predictive accuracy of functional annotations

We fitted predictive models for GWAS hit status via elastic net, using clumped and separated functional variable sets, using high-confidence ($p<5x10^{-8}$) and low-confidence ($p<10^{-5}$) GWAS hits, and using all GWAS hits ("non-phenotype specific") as well as hits classified according broad phenotype areas. We primarily investigated predictive accuracy in a separate test set that was not involved in the fitting of the models. Variants were randomly split between the training and test sets.

For all of our fitted models, the area under the curve (AUC) of a receiver operating characteristic (ROC) curve was similar in the test and training sets, suggesting that the models had not been over-fitted. (**Figure 3.9** plots the AUCs derived from the training set, and **Figure 3.10** plots the AUCs derived from the test set.)

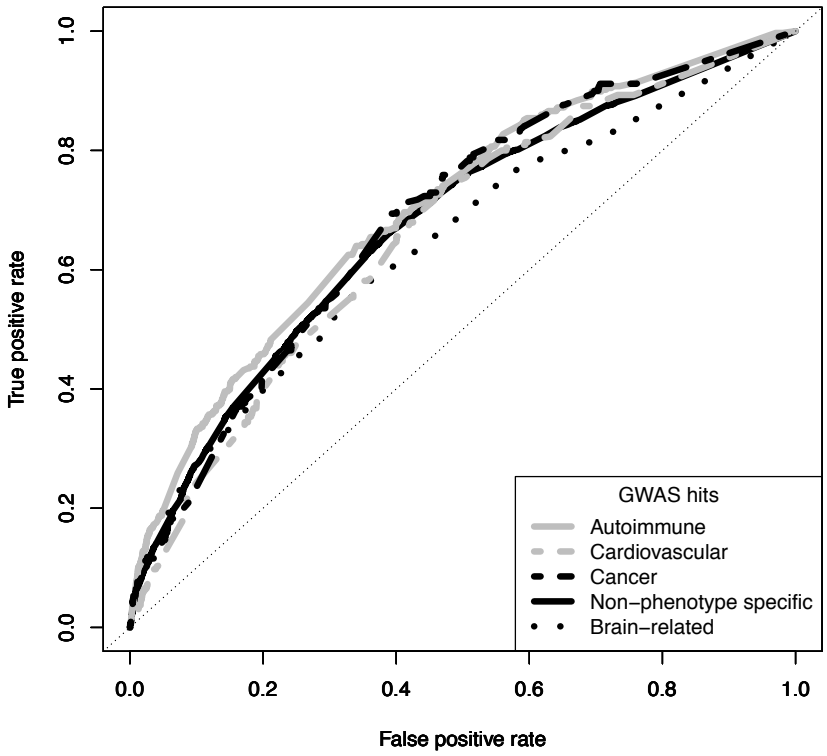


Figure 3.9. Receiver operating characteristic (ROC) curves for analyses of clumped functional variables and high-confidence GWAS hits using the training set

This plot is similar to the plot obtained from the separate test set, **Figure 3.10**.

We found that the ROC curves for both the separated and clumped analyses had similar AUCs: for instance 0.58 in the test set for the non-phenotype specific clumped analysis and 0.59 in the test set for the separated analysis.

Two analyses emerged as most predictive based on integrating results from ROC curves, positive predictive values, and histograms of the probabilities of causality (the prediction scores). These were the analyses based on non-phenotype specific and the autoimmune

GWAS analyses. Best results were obtained from analyses using high-confidence GWAS hits. Results for clumped and separated functional variables were very similar (**Table 3.5** and **Figure 3.10**).

Table 3.5. Areas under fitted ROC curves

AUCs for analyses using the high-confidence GWAS hits. Values in parentheses are for all SNPs in the GWAS Catalogue.

	Non-phenotype specific	Brain-related	Cancer	Cardiovascular	Autoimmune
N	4480 (8219)	530 (1741)	300 (607)	369 (716)	570 (863)
AUC clumped	0.68 (0.58)	0.62 (0.52)	0.67 (0.60)	0.69 (0.61)	0.71 (0.67)
AUC separated	0.70 (0.59)	0.61 (0.51)	0.68 (0.60)	0.66 (0.61)	0.75 (0.71)

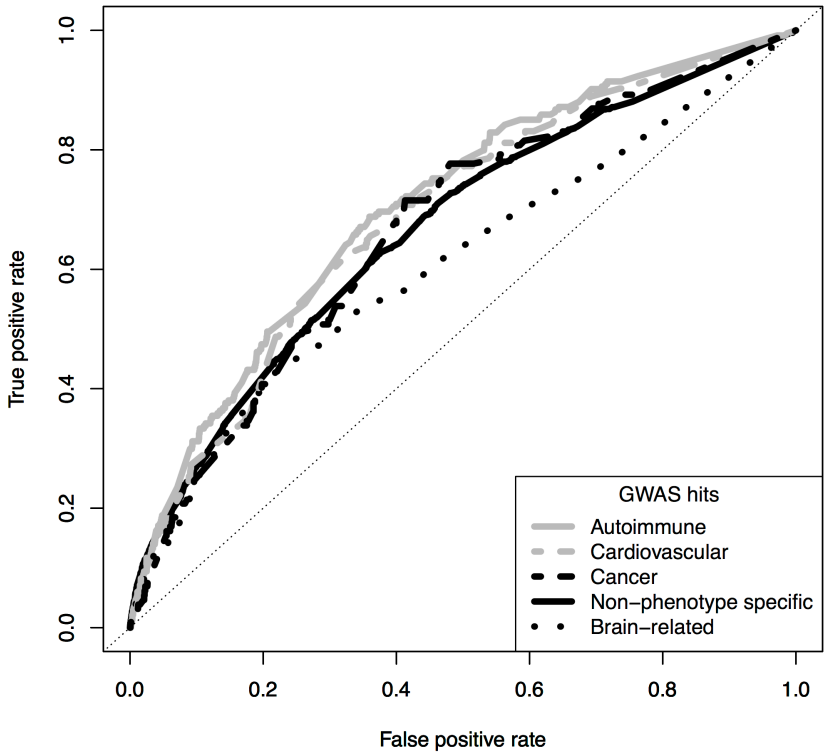


Figure 3.10. Receiver operating characteristic (ROC) curves for analyses of clumped functional variables and high-confidence GWAS hits

ROC curves were obtained from a separate test set.

The numbers of hits and non-hits in the test sets are reported in **Table 3.6**.

Table 3.6. The number of hits and non-hits in the test set sets for the analyses of clumped functional variables and high-confidence GWAS hits.

	Hits	Non-hits
Brain	144	32723
Cardiovascular	154	33346
Cancer	130	33370
Autoimmune	234	33266
Non-phenotype specific	1292	30135
Non-phenotype specific- all Catalogue	3405	30039

We also investigated positive predictive values (PPVs) and histograms of the probability of causality (prediction score). PPV estimates could not be obtained due to insufficient data (a limited number of true hits correctly identified as hits at a particular prediction value threshold) for the phenotype specific analyses since these analyses contain only a subset of all GWAS hits. As a result, PPVs were only plotted for the non-phenotype specific analyses (**Figure 3.11**). PPVs appear to be highest for the analysis using all GWAS hits compared to the analysis using the high-confidence hits when defining hits as those variants with a prediction score of greater than 0.5, 0.6, or 0.7. There was insufficient data at the higher thresholds for declaring a positive hit for the analysis based on all GWAS hits. Yet sufficient data was available at the higher prediction value thresholds for the analysis using the subset of high-confidence hits, demonstrating a broader spread in prediction values for that analysis compared to the analysis on all GWAS hits.

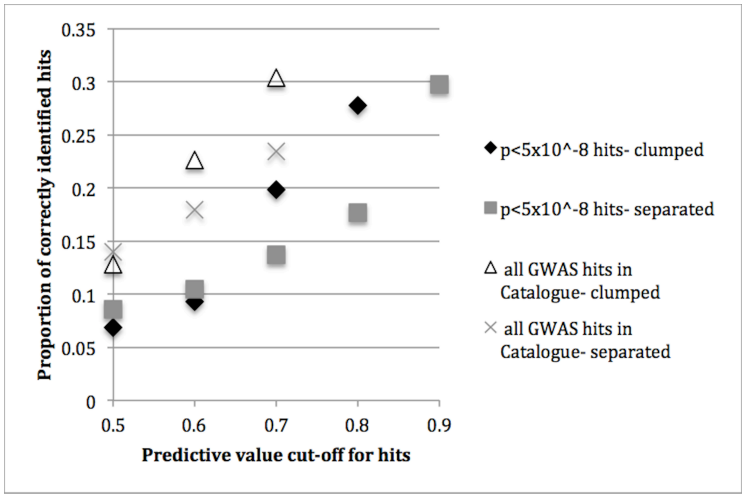


Figure 3.11. Proportion of correctly identified hits in the test data (positive predictive values)

In the non-phenotype specific analyses at various cut-offs for defining hits: SNPs with predictive values of greater than 0.5, 0.6, 0.7, 0.8, or 0.9. Note that results are only plotted for those predictive value thresholds in which there are at least 11 hits correctly identified.

Histograms of the probability of causality in the test data allowed visualization of the separation (or non-separation) of true hits versus non-hits. We found that for the non-phenotype specific analysis and for the autoimmune analysis, the use of high-confidence GWAS hits in the training data improved the separation of true hits from non-hits in the test data (**Figure 3.12**).

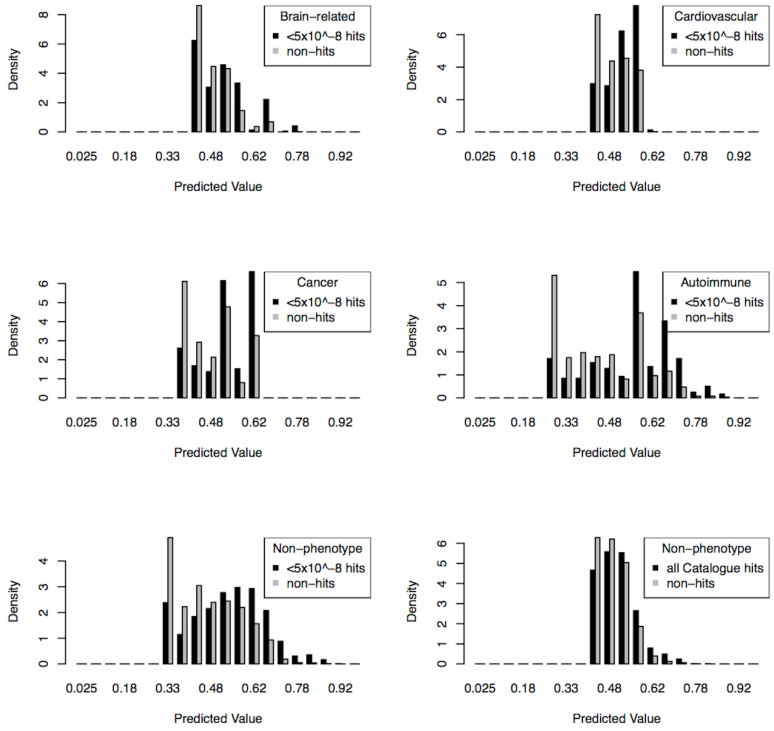


Figure 3.12. Predicted values for true GWAS hits and non-hits in the test data

Panels show results of clumped-variable analyses on high-confidence GWAS hits for brain-related [**a**], cardiovascular [**b**], cancer [**c**], autoimmune [**d**], and non-phenotype specific hit sets [**e**], and for all hits in the GWAS Catalogue for the non-phenotype specific hit set [**f**].

The results from the histograms of the predicted values showed a broader spread in the non-phenotype specific clumped analysis on high-confidence GWAS hits compared to the analysis using all hits. The former separated true hits from non-hits better than the latter, with the modes of the two distributions distinct. These results suggest that the weighted elastic net procedure was successful in producing models that performed well in identifying true hits as well as in identifying true non-hits. While we could not obtain reliable PPV estimates for the autoimmune analysis due to insufficient data, the separation of non-hits from hits in the histogram was taken as sufficient evidence that the

high area under the ROC curve for the autoimmune clumped analysis was also due to positive predictive power.

Results will only be provided for the non-phenotype specific and the autoimmune clumped analyses, the two models that were deemed to be reliable based on the predictive accuracy measures. For the non-phenotype specific clumped analysis, the highest Bayes factor for annotation (11.95) was obtained for rs11177, which is a known GWAS hit associated with osteoarthritis on chromosome 3. It had a predicted value of 0.93. This SNP or its proxies held all functional annotations except three low-frequency annotations: splice sites, miRNA targets, and Gencode transcription start sites. This SNP, which results in a missense change in the GNL3 gene, has 218 LD proxies (defined as SNPs with an r^2 of $\geq$0.8 with rs11177 that are present in Phase I of the 1000 Genomes Project). Of the proxies, the majority of them (203; 93%) are intronic.

Nine percent of the variants with the top 500 Bayes factors were known GWAS hits. The frequency of hits in the test set data was 4.1%. The mean and median of the predicted values for the true hits in the test set were higher than those for the true non-hits (for hits: mean= 0.54, standard deviation=0.13 and median= 0.54; for non-hits: mean= 0.46, standard deviation=0.12 and median= 0.44).

For the autoimmune clumped analysis, the SNP with the highest Bayes factor was the same as for the non-phenotype specific clumped analysis, rs11177.

3.4.4 Investigation of the relative importance of different functional annotations

The importance of a particular functional annotation in predicting whether or not a SNP is more probable to be a GWAS hit is assessed by means of the magnitude of the coefficient assigned to the annotation. In both the non-phenotype specific and

autoimmune analyses we note that the nonsynonymous SNP functional annotation had one of the highest coefficients (**Figure 3.13**).

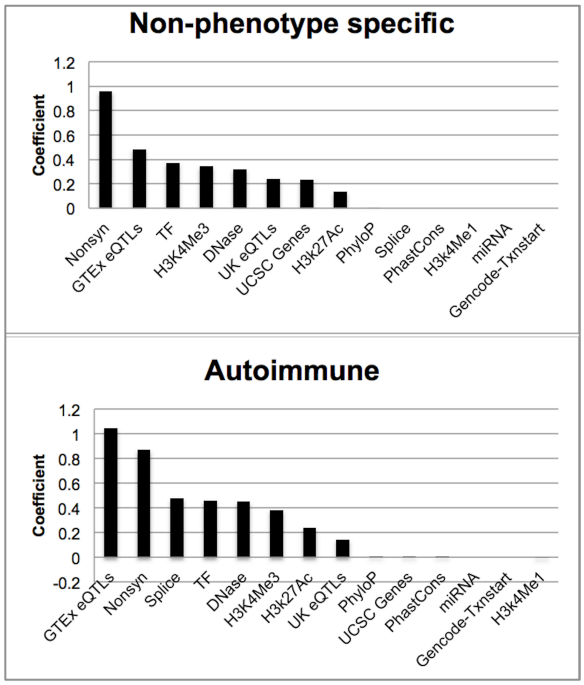


Figure 3.13. Coefficients of the functional annotations for the two best analyses

The figure shows the coefficients from the clumped analysis on high-confidence GWAS hits for the non-phenotype specific versus the autoimmune model.

The coefficients for the non-phenotype specific model are provided in **Table 3.7**, and the coefficients for the autoimmune model are provided in **Table 3.8**. Confidence intervals cannot be easily calculated for coefficients from elastic net, and so to estimate standard error for the coefficients we performed multivariate logistic regression (see the right columns in **Table 3.7** and **Table 3.8**). GTEx eQTLs had the highest coefficient in the autoimmune analysis.

Table 3.7. Coefficients from elastic net and multivariate logistic regression for the non-phenotype-specific analysis

Coefficients for the non-phenotype-specific analysis defining hit SNPs as those SNPs in the GWAS Catalogue with a p-value of less than $5x10^{-8}$. The coefficients for the multivariate logistic regression are shown in order to provide estimates of error for the coefficients, which is not possible for elastic net.

	Non-phenotype specific			
	Elastic net	**Multivariate logistic regression**		
	Coefficient	Coefficient	p-value	Standard error
Splice	0	-3.45E-02	0.0556	1.80E-02
PhastCons	0	1.86E-04	0.94	2.48E-05
H3k4Me1	0	-8.87E-03	3.30E-04	2.47E-03
miRNA	0	7.77E-03	0.92	7.53E-02
Gencode-Txnstart	0	-4.22E-02	0.62	8.55E-02
PhyloP	2.70E-03	1.98E-04	6.34E-14	2.63E-05
H3k27Ac	0.1	8.16E-03	1.10E-03	2.50E-03
UCSC Genes	0.16	9.48E-03	7.08E-08	1.76E-03
UK Brain eQTLs	0.27	2.84E-02	< 2.0E-16	2.96E-03
H3K4Me3	0.33	2.06E-02	< 2.0E-16	2.32E-03
TFBS	0.34	1.88E-02	< 2.0E-16	1.84E-03
DNase I	0.35	3.10E-02	< 2.0E-16	2.26E-03
GTEx eQTLs	0.72	0.13	< 2.0E-16	9.54E-03
Nonsynonymous	1.3	0.26	< 2.0E-16	7.99E-03

Table 3.8. Coefficients from elastic net and multivariate logistic regression for the autoimmune-specific analysis

Coefficients for the autoimmune-specific analysis defining hit SNPs as those SNPs in the GWAS Catalogue with a p-value of less than $5x10^{-8}$. The coefficients for the multivariate logistic regression are shown in order to provide estimates of error for the coefficients, which is not possible for elastic net.

	Autoimmune			
	Elastic net	**Multivariate logistic regression**		
	Coefficient	Coefficient	p-value	Standard error
miRNA	0	-1.39E-02	0.61	2.74E-02
Gencode-Txnstart	0	-2.54E-02	0.41	3.11E-02
PhastCons	2.00E-04	1.25E-05	0.16	8.85E-06
H3k4Me1	-6.20E-03	-2.24E-03	0.01	8.79E-04
PhyloP	2.00E-03	1.92E-05	0.84	9.41E-06
UCSC Genes	1.20E-03	-2.64E-04	0.67	6.27E-04
UK Brain eQTLs	0.14	3.50E-03	9.90E-04	1.06E-03
H3k27Ac	0.24	2.00E-03	0.02	8.88E-04
H3K4Me3	0.38	3.95E-03	1.60E-06	8.24E-04
DNase I	0.45	5.89E-03	3.30E-13	8.09E-04
TFBS	0.46	3.36E-03	2.80E-07	6.54E-04
Splice	0.48	8.23E-03	0.21	6.53E-03
Nonsynonymous	0.87	2.71E-02	< 2.0E-16	3.06E-03
GTEx eQTLs	1.04	2.70E-02	4.30E-15	3.44E-03

3.4.5 Sensitivity analysis- classification

The resulting AUCs and Beta coefficients from the analysis in which hits were defined as the subset of the non-phenotype specific $5x10^{-8}$ hits minus those hits used in the phenotype-specific analyses (autoimmune, brain-related, cancer and cardiovascular) were very similar to the results from the $5x10^{-8}$ non-phenotype specific analysis. The results suggest that the non-phenotype specific analysis was not being driven variants from one of the larger phenotypes.

3.4.6 Investigating functional predictions in the context of known GWAS

We investigated: schizophrenia (SZ) from a meta-analysis GWAS involving the first sample from the Psychiatric Genomics Consortium (PGC1) combined with a Swedish sample (Ripke et al., 2013), systolic blood pressure (SBP) from the International Consortium for Blood Pressure (ICBP) (Ehret et al., 2011), and height from Genetic Investigation of Anthropomorphic Traits (GIANT) Consortium (Lango Allen et al., 2010). The studies analyzed over 35,000 cases and 47,000 controls, 200,000 individuals, and over 180,000 individuals, respectively. (The significant hits from these studies were not included in the respective models.)

For each study, we stratified the quantile-quantile plots according to predicted value bins (**Figure 3.14**). We found that SNPs with higher predicted values from the non-phenotype specific clumped analysis tended to deviate more from the line corresponding to the overall GWAS, in favour of more association signals. Similar results were obtained for all three GWAS analyzed: schizophrenia, systolic blood pressure and height.

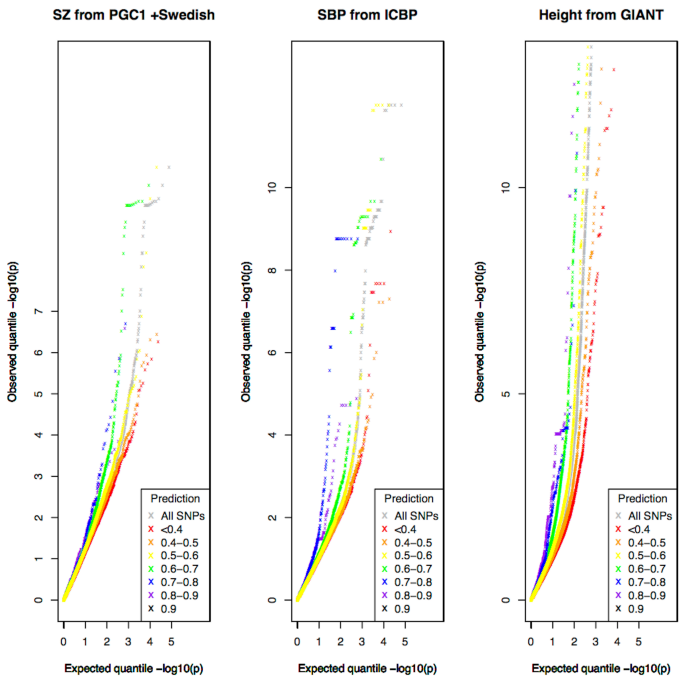


Figure 3.14. Quantile-quantile plots stratified by predicted values for SNPs in real GWAS

All GWAS SNPs (in grey) for a schizophrenia GWAS from PGC1 with a Swedish sample [**a**], a systolic blood pressure GWAS from ICBP [**b**], and a height GWAS from GIANT [**c**]. The non-grey lines show plots for SNPs binned according to their predicted value from the non-phenotype specific model.

The pattern remained when only the GWAS SNPs present in the test set were plotted, and also when prediction values were obtained from models derived from excluding the genome-wide significant SNPs in the training set for each GWAS respectively.

We obtained summary data obtained from a psoriasis GWAS study from Strange et al. (2010). We then selected 15 SNPs that were subsequently discovered in a meta-analysis (Tsoi et al., 2012). Using summary association statistics from the Strange et al. study we derived Bayes factors for association (BF_{assoc}) and Bayes factors based on association data combined with the annotation of functional annotations (BF_{assoc}*BF_{annot}) for each SNP. We ranked the SNPs according to BF_{assoc}, and ranked them again according to BF_{assoc}*BF_{annot} to determine whether annotating SNPs with their functional annotations

improved their rank (larger Bayes factors were assigned smaller ranks). BF_{annot} values were derived from the non-phenotype specific clumped analysis using high-confidence GWAS hits. As negative controls, we took 12 independent sets of a random 15 SNPs (which were not in high LD with any of the 15 hits and had similar p-values to the hits) and compared the difference in the sum of ranks based on BF_{assoc} versus BF_{assoc}*BF_{annot}. The procedure was repeated using BF_{annot} derived from the autoimmune clumped analysis.

Of the 15 true psoriasis hit SNPs, 7 had better ranks based on BF_{assoc}*BF_{annot} compared to association information on its own (BF_{assoc}). The difference of the sum of ranks assigned to the 15 hits was nearly 48,000 based on BF_{assoc}*BF_{annot} compared to BF_{assoc}, with the former having the lower sum (better ranks). Many of the hit SNPs had very large ranks based merely on the association data (>3,000), which was also the case for ranks based on BF_{assoc}*BF_{annot}, but the trend was in the right direction with better ranks obtained when combing the association information with the annotation of functional annotations. Of the 12 random sets of 15 independent SNPs, the trend was in the opposite direction for 10 of the sets (with SNPs having better ranks based on BF_{assoc} alone). Of the remaining 2 sets, one of them had the same number of the SNPs with improved ranks based on BF_{assoc}*BF_{annot} compared to BF_{assoc} as did the analysis with the actual hits (7 out of 15), and the other random set had 8 SNPs that showed improvement. However, for those random SNP lists the difference in the sum of ranks from BF_{assoc} compared to BF_{assoc}*BF_{annot} was less than half of the improvement of ranks seen for the 15 hits. Comparable results were seen when using BF_{assoc} based on the autoimmune clumped analysis. The difference between the sum of the ranks for BF_{annot} compared to BF_{assoc}*BF_{annot} was over 49,000, with improved ranks of the hits based on the BF_{assoc}*BF_{annot} ranks. Of the random lists the largest difference in the sum of ranks from BF_{assoc} compared to BF_{assoc}*BF_{annot} was less than a third of the improvement of ranks seen for the 15 hits.

3.5 Discussion

The release of major genome-wide datasets such as ENCODE and NIH Roadmap projects, offers an excellent opportunity to re-assess the existing GWAS corpus and draw conclusions about which functional annotations in the human genome are most likely to indicate causality in association studies. We previously considered Bayes factors based on a limited set of functional annotations, considering each functional annotation separately (Knight et al., 2011). Here we have extended our Bayesian framework by developing Bayes factors for multiple functional annotations, considering all functional annotations jointly. We used a regularized logistic regression to fit predictive models allowing for large numbers of both qualitative and quantitative functional annotation data. We performed our analysis under a wide variety of conditions, including phenotype specific analysis for autoimmune, brain-related, cancer, and cardiovascular disorders.

Our results confirm previous findings of differences in functional enrichment in GWAS hits compared to non-hits, which provided a rationale for utilizing functional annotations as predictors of SNP causality. We found that using high-confidence GWAS hits ($p<5x10^{-8}$) as a classifier resulted in more predictive power. However, if the number of GWAS hits that are available for training are too low, then the predictions become imprecise. This was a reoccurring theme for many of the phenotype specific analyses. The separation between true GWAS hits and non-hits in the test set, in addition to the AUC, should be used to assess the predictive power of a model. Using those methods we found that the non-phenotype specific and the autoimmune analyses on clumped variables using high-confidence GWAS hits were most reliable. For instance, although the AUCs were slightly higher for the separated analyses, the classification of true GWAS hits and non-hits was better in the clumped analysis, suggesting that the clumped analysis may provide more accurate predictions. The benefit of the separated analysis is that it allows researchers to identify annotations specific to certain conditions, for example specific cell types, which can be useful for planning further investigations, but

the increased number of variables and sparsity of the data reduces the power of this type of analysis.

While our study has demonstrated that relevant functional information is indeed predictive for identifying GWAS hits, and that Bayes factors incorporating this functional information rank known GWAS hits better than Bayes factors based on association information alone, the improvements based on current information (for example, in the psoriasis GWAS we analyze) are marginal. However, we outline reasons below to argue that the benefit of adding functional information to analyses of causal variant discovery will increase in the future.

A limitation to the study is the restricted amount of tissue- or cell-specific data, especially in light of the findings that enrichment of disease-specific GWAS hits can differ in certain cell types, for example for DNase I hypersensitive sites (Maurano et al., 2012). Incorporating additional functional annotations, for example those from relevant tissue types, will likely improve the understanding of which annotations are associated with GWAS hit SNPs, especially for the phenotype specific analyses. Furthermore, other functional annotations, such as further histone marks and other epigenetic modifications, could be incorporated to improve the models.

Another limitation is that the hits and non-hits were not matched by minor allele frequency or base pair distance, which may partially drive differences between the functional annotations of the hits compared to the non-hits. As discussed the non-hit selection was chosen from the group of variants not in LD with a GWAS hit. A subsequent analysis showed that the selection of non-hits tended to have lower allele frequencies compared to the hits (**Figure 3.15**).

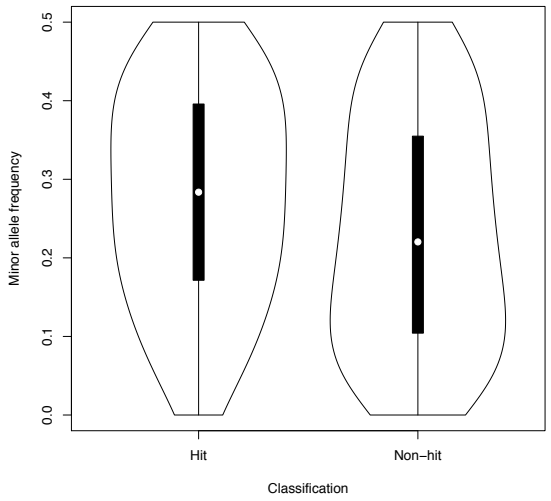


Figure 3.15. Violin plot showing the minor allele frequency distribution between the hits and non-hits.

This plot shows data for 4,480 GWAS hits and 75,341 randomly selected non-hits, defined as not being in LD with a hit. Mann-Whitney U p-value < 2.2×10^{-16}.

Furthermore, SNPs with higher MAF may be thought to have more LD proxies. However, an investigating this hypothesis showed that there is no correlation between the number of LD proxies and MAF (**Figure 3.16**).

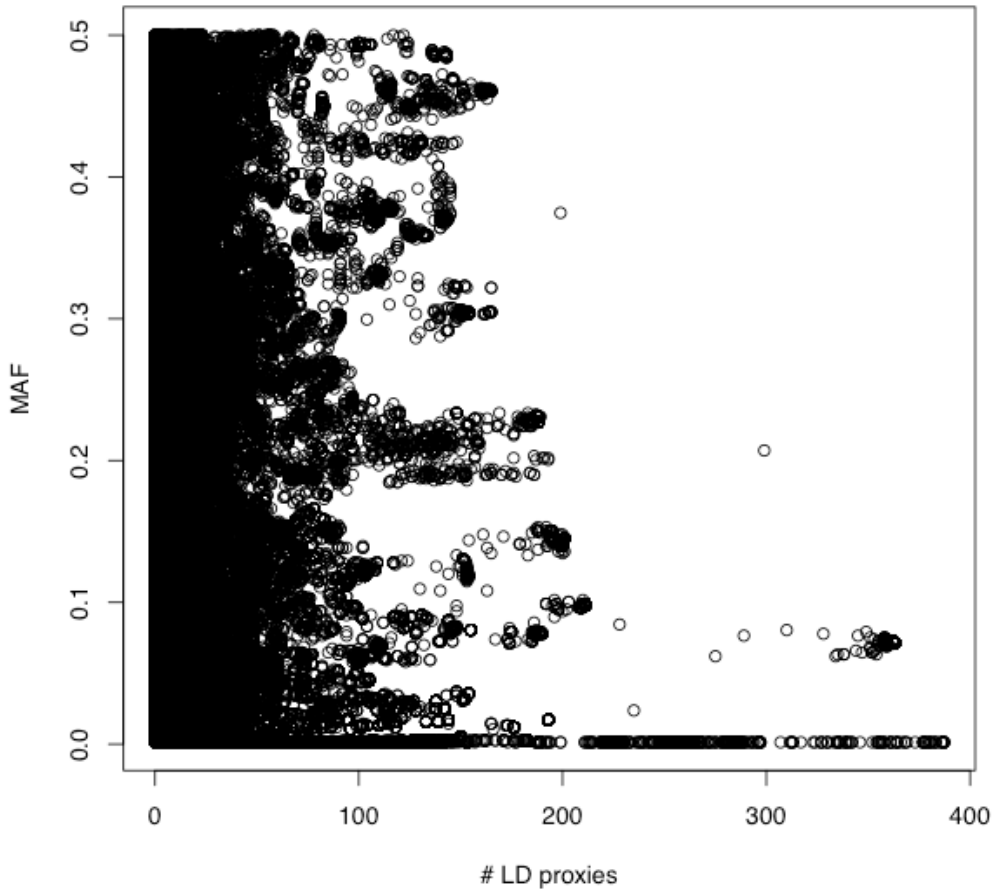


Figure 3.16. Number of LD proxies versus minor allele frequency distribution for SNPs on chromosome 22.

The correlation between the two measures was 0.03. Only chromosome 22 shown for computational efficiency.

The current number of GWAS hits in the GWAS Catalogue makes it challenging to sub-divide hits into phenotype specific traits. However, preliminary results showing differences in the coefficients for the functional annotations suggest that as the number of GWAS hits grows, a phenotype specific approach from which to derive Bayes factors for

prioritization could be more biologically relevant than simply an approach that combines all GWAS hits together. Interestingly, although it was one of the largest lists, the brain-related list did not have a greater predictive power than expected by chance. This finding only serves to reinforce the widely appreciated complexity of brain-related disorders. Nevertheless, schizophrenia GWAS significant SNPs showed enrichment of SNPs with high predicted values from the model, as did SNPs associated with systolic blood pressure or height.

Using manually curated phenotype lists as done here may not be the best option. Using lists that are more reproducible, such as those based on the Experimental Factor Ontology (EFO) definitions, may be more appealing. However, most of the lists created using the EFO definitions were relatively small, covering less than 10% of the total GWAS hits on the common genotyping arrays, and thus this method of classifying GWAS hits was deemed to be not feasible, but may be possible in the future as the size of GWAS Catalogue grows still larger.

The coefficient for SNPs was the highest in the non-phenotype specific analysis and a close second in the autoimmune analysis. This result suggests that being a variant in a gene that causes a protein alteration is an important indicator of whether or not a genetic variant will be truly associated with a phenotype. The result agrees with the findings that the top associated SNPs and also those that are nominally associated with a phenotype are more likely to overlap genes than non-GWAS SNPs (Tang and Ferreira, 2012). Our analysis appears to underscore the primacy of variation as a leading mediator of functional variation in the human genome. Although this result is perhaps unsurprising, it lends support to many of the gene-focused, rare-variant strategies that have been recently employed (for example: Barrans and Liew, 2006; Cortes and Brown, 2011; Voight et al., 2012). However, depending on the inclusiveness of promoter regions in chip design, these strategies may or may not capture other high scoring variant types, such as eQTLs and histone marks, which collectively account for more GWAS hits than variants alone.

These patterns highlight a possible need for follow-up on non-coding variation chips. GTEx eQTLs came up as the most important factor in the autoimmune analysis. Two of the experiments analyzed eQTLs from lymphoblastoid cells, which may explain the importance of this functional annotation in the autoimmune traits.

We have shown that our method can be used to calculate Bayes factors for annotation (BF_{annot}). These can be applied to GWAS data to prioritize near-significant variants for follow-up based on the likelihood of being causal in light of their functional annotations. The method takes LD into account, and uses information from the March 2012 release of the 1000 Genomes Project to map relevant annotation information from all variants in high LD, including both SNPs and indels. In addition to being used for variant prioritization of GWAS data, the methodology could be applied in the future to the prioritization of variants from fine mapping and sequencing studies. Here, the question arises as to whether the models described here, which were created based on common variation, could be applied to rare variation. In time, larger databases of true causal variation, including rare variation, will allow our method to be applied with increasing accuracy.

3.6 Subsequent Developments

Further work has involved incorporated some additional annotations into the non-phenotype specific model using the GWAS hits with a p-value $< 5 \times 10^{-8}$: synonymous SNPs (since synonymous SNPs too can have a phenotypic effect, for instance see Buske et al. (2013), albeit an effect is more rare than for nonsynonymous SNPs), and super-enhancers associated in 86 human cell and tissue samples (Hnisz et al., 2013). However, the addition of neither of these two annotations altered the accuracy of the model. The lack of effect of the synonymous annotation was not due to low frequency of synonymous SNPs in the full dataset, since the frequency of synonymous SNPs (0.06) was 10-fold higher than for nonsynonymous SNPs (0.007), and the latter was the most important predictor in the model. Super-enhancers (0.001) were not included in the model

(i.e. it was assigned a beta coefficient of 0), which may have been in part due to a low frequency in the full data (0.001; compare to another low frequency annotation: splice sites at 0.002).

3.7 Supporting Data

Three files are provided, not including the "README.txt", which describes the files similarly to as below. "Non-phenotypespecific_BFannot.txt" is a space-delimited text file of Bayes Factors for Annotation (BF_{annot}) for the non-phenotype specific analysis. The first row contains the headers. The rest of the rows contain information for SNPs in 1000 Genomes EUR phase I. The meaning of the column names are as follows: rs: SNP ID, BFannot: Bayes Factors for Annotation (based on 14 functional annotations).

"Non-phenotypespecific_assoc+pred.txt" is a space-delimited text file of the functional annotations and the prediction value for the non-phenotype specific analysis derived from 14 functional annotations. The first row contains the headers. The rest of the rows contain information for SNPs in the 1000 Genomes EUR phase I. The meaning of the column names are as follows: bp: base position, hg19, chr: chromosome number, rs: SNP ID, bp: base position, hg19 (same as column 1).

The next 14 columns are the functional annotations (splice, nonsynonymous, DNase_I, GTEx_eQTLs, UK_Brain_eQTLs, PhyloP, PhastCons, H3K4Me1, H3K4Me3, H3K27Ac, TFBS, miRNA, Gencode_Txnstart). 1= the SNP has the functional annotation or it is in high LD ($r^2 \geq 0.8$) with a SNP that does; 0= neither the SNP nor its high LD proxies have the functional annotation.

The second last column (cls) is classifier where 1= GWAS "hit" ($p<5x10^{-8}$ in NHGRI GWAS Catalogue http://www.genome.gov/gwastudies/ as of Aug. 6, 2013) and 0 = "non-hit". The final column in the file (pred) is the prediction score (ranging from 0 to 1, where 1 is likely to be a GWAS "hit") from the non-phenotype specific analysis.

"PLINK2wakefieldBF_2013.R" is an R script to calculate Bayes Factors for Association (BF_{assoc}) based on GWAS summary data.

All files and also the elastic net R code are available on GitHub (and linked to Zenodo at **http://dx.doi.org/10.5281/zenodo.34268).**

Chapter 4
A Review of Predictive Accuracy Measures that can be Applied to Models for Prioritizing Risk Variants Based on Functional Information

This section is modified from the following: **Gagliano SA,** Paterson AD, Weale ME, Knight J (2015). Assessing models for genetic prediction of complex traits: a comparison of visualization and quantitative methods. *BMC Genomics* 16(1):405.

4

4.1 Abstract

Background: *In silico* models have recently been created in order to predict which genetic variants are more likely to contribute to the risk of a complex trait given their functional annotations. However, there has been no comprehensive review as to which type of predictive accuracy measures and data visualization techniques are most useful for assessing these models.

Methods: We assessed the performance of the models for predicting risk using various methodologies, some of which include: receiver operating characteristic (ROC) curves, histograms of classification probability, and the novel use of the quantile-quantile plot. These measures have variable interpretability depending on factors such as whether the dataset is balanced in terms of numbers of genetic variants classified as risk variants versus those that are not.

Results: We conclude that the area under the curve (AUC) is a suitable starting place, and for models with similar AUCs, violin plots are particularly useful for examining the distribution of the risk scores.

4.2 Introduction

The risk of developing a complex trait is influenced by many genetic variants, possibly hundreds, in combination with environmental factors. Genome-wide association studies (GWAS) have had success in identifying some of the genetic risk factors involved in complex traits, but more remain to be discovered. Recently, there have been several *in silico* attempts at utilizing epigenetic and genomic data to prioritize genetic risk variants. These methods simultaneously incorporate multiple lines of genomic and epigenomic data to identify potential risk variants from all variants (Gagliano et al., 2014a; Iversen et al., 2014; Kindt et al., 2013; Kircher et al., 2014; Pickrell, 2014; Ritchie et al., 2014).

A variety of predictive accuracy measures and data visualization techniques have been used (**Table 4.1**) to assess these models for prioritizing genetic variants. An example is the area under the curve (AUC) from the receiver operating characteristic (ROC) curve, which is generally accepted as a measure of how closely the prediction values reflect the true class. Such methods have previously been employed to predict diagnosis of an individual (risk of developing Type II Diabetes (Janipalli et al., 2012; Lango et al., 2008; Xu et al., 2010), for example), but have only recently been applied to predict whether genetic variants are likely to be risk variants.

Table 4.1. Predictive accuracy measures in the literature for models for prediction of variants associated with complex traits.

			Predictive accuracy measures employed					
	Algorithm	**Classifier**	Area under ROC curve	Positive Predictive value	Box plot	Histo-Gram	Violin plot	Mann-Whitney U / Wilcoxon Rank Sum test
Gagliano et al. 2014	Modified Elastic net	GWAS hits vs. non-hits	x	x		x		
Iversen et al. 2014	Penalized logistic regression	GWAS hits vs. non-hits	x*					
Kircher et al. 2014	Support Vector Machines	High-frequency human-derived alleles vs. simulated variants	x				x	x
Ritchie et al. 2014	Modified Random Forest	HGMD hits vs. non-hits	x		x			x

* reports "Concordance index", which is equivalent to the area under the ROC curve

We will utilize test set data from a regularized logistic model that predicts genetic risk variants on the basis of a large multivariate functional dataset (Gagliano et al., 2014a). We investigate the utility of several approaches for assessing predictive accuracy and data visualization. Based on observations from this work we conclude with suggested guidelines to aid researchers when assessing models for genetic variant prediction.

Three broad categories of predictive accuracy measures will be discussed here: (1) concepts in describing predictive accuracy, including ROC, AUC and the confusion matrix (2) visualization of the distribution of prediction values, and (3) statistical tests. All the methods described below were conducted in R, version 3.0.2 (Hothorn et al., 2006; Lemon, J., 2006; R Core Development Team, 2008; Sing et al., 2005). See **Table 4.2**. Sample R code is available in **Additional_File_1**. Code and data to reproduce the results in this chapter are provided in **Additional_File_2**. Further details are embedded in the results. Additional files are available in **Appendix C**.

Table 4.2. Predictive accuracy measures and the corresponding R package in which they can be computed.

Predictive Accuracy Measure	R package	Version
(1) The confusion matrix		
Receiver Operating Characteristic Curve and area under the curve	prediction and performance in ROCR (Sing et al., 2005) performance(prediction.object, "auc")	1.0-7
Positive predictive value and negative predictive value	prediction and performance in ROCR performance(prediction.object, "ppv") performance(prediction.object, "npv")	1.0-7
(2) Visualization of the distribution of prediction values		
Histograms of the prediction values separated by class	multhist in plotrix (Lemon, J., 2006)	3.5-11
Box plots	boxplot in graphics	Base package
Violin plots	vioplot in vioplot	
Quantile-quantile plots	qqplot in stats	Base package
(3) Statistical tests		
Hypergeometric test	phyper in stats	Base package
Mann-Whitney U test	wilcox.test in stats	Base package
Asymptotic Generalized Cochran-Mantel-Haenszel Test	cmh_test in coin (Hothorn et al., 2006)	1.0-24

4.3 Dataset and models

The example dataset and model are described in detail previously (Gagliano et al., 2014a) and are only described briefly here. Genetic variants from common genotyping arrays were annotated for 14 functional annotations (twelve of which are binary and two are

quantitative), many of which are from the ENCODE Project, with data from various cell types merged (un-weighted) into a single variable for each annotation. All functional annotations could be presented in a binary presence/absence format with the exception of two types conservation scores, which remained on a quantitative scale. A regularized logistic model, capable of handling correlated predictor variables, was used. A random 60% of the genetic variants were assigned to the training set to determine the parameters of the model, and the remaining variants were reserved for the independent test set to evaluate the accuracy of the model. All models produced a prediction value ranging from 0 to 1 for each genetic variant, with values close to 1 implying high probability of the variant contributing to risk. Due to the unbalanced nature of the data a weighting procedure that equalizes the importance of hits and non-hits in the training set was employed. Hits were weighted by $(N_{hits}+N_{non\text{-}hits})/2N_{hits}$ and all non-hits by $(N_{hits}+N_{non\text{-}hits})/2N_{non\text{-}hits}$, where N_{hits} and $N_{non\text{-}hits}$ denote the number of hits and non-hits, respectively, in the training set (Gagliano et al., 2014a). Without this weighting scheme, all variants are assigned low prediction values although the model still retains comparable overall accuracy. Overall accuracy may not be representative of accuracy within classification groups, which is the main problem with unbalanced data. As well as using the weighting scheme to ameliorate this issue in our example data we discuss other matters to be considered in relation to the accuracy and data visualization methods described.

For model 1, variants were classified as being hits if present in the genome-wide association study (GWAS) Catalogue published by the National Human Genome Research Institute (Hindorff et al., 2010) downloaded on August 6, 2013. The GWAS Catalogue reports variants found to be associated with disease or quantitative trait in a GWAS study with a p-value $<1x10^{-6}$. Variants not present in the Catalogue but present on common genotyping arrays were assumed to be non-hits. Three alternate classifiers were used to designate hits: (a) p-value $< 5x10^{-8}$ (model 2), and (b) p-value $< 5x10^{-8}$ for only a subset of phenotype specific hits namely an autoimmune (model 3) and a brain-related analysis (model 4).

In our previous work, six models were created using the alterations to the classifier described above. The four assessed here are the two models with the highest AUC (models 2 and 3) and two models with the lowest AUC (models 1 and 4). (See **Table 4.3** for descriptive statistics for the test sets of the various models.)

Table 4.3. Descriptive statistics of the causality predictive values for the various genetic prediction models from Chapter 3 to be used as examples here.

Phenotype-specific analyses		N	Minimum	25% Percentile	Median	Mean	75% Percentile	Maximum	Standard Deviation	N outliers*
Brain-related	Hits	144	0.40	0.42	0.51	0.51	0.57	0.77	0.09	3
	Non-hits	32723	0.40	0.40	0.46	0.48	0.53	0.79	0.07	61
Autoimmune	Hits	234	0.29	0.45	0.55	0.55	0.66	0.86	0.14	0
	Non-hits	33266	0.29	0.30	0.44	0.45	0.55	0.93	0.13	0
All phenotype analyses										
p<5E-8	Hits	1292	0.32	0.44	0.54	0.54	0.62	0.92	0.13	4
	Non-hits	30135	0.32	0.35	0.44	0.46	0.55	0.91	0.12	7
all GWAS Catalogue	Hits	3405	0.44	0.45	0.50	0.51	0.54	0.81	0.06	144
	Non-hits	30039	0.44	0.44	0.48	0.49	0.52	0.80	0.05	336

*Outliers are defined as data points outside 1.5x interquartile range (interquartile range= 75% percentile - 25% percentile).

4.4 Results

4.4.1 Concepts in describing predictive accuracy

4.4.1.1 The Confusion Matrix

Predictive accuracy is derived from a confusion matrix (**Figure 4.1**). The cells in the diagonal of the matrix are the correctly identified genetic variants. (See Chapter 4 in "*An Introduction to Statistical Learning with Applications in R*" (James et al., 2013) and Chapter 11 in "*Statistical Learning for Biomedical Data*" (Malley et al., 2011) for more details.) The effects of unbalanced data in un-weighted models can be detected in such a matrix. There would be a much larger proportion of negatives compared to positives. The effects on false positive rate (FPR), true negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV) are described in further detail below. The

confusion matrix itself is not often studied as it represents data at only one threshold. However both the ROC curve and PPV and NPV are used to consider model accuracy.

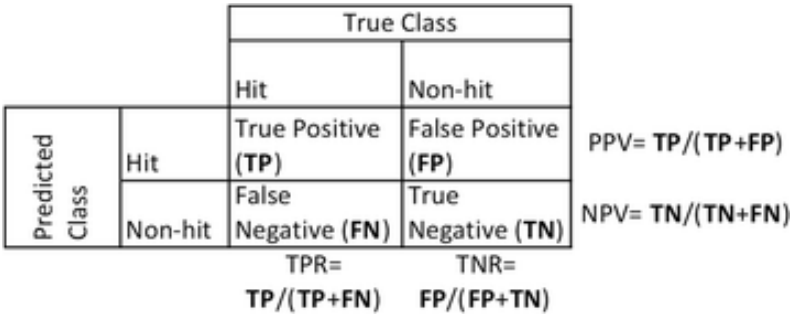


Figure 4.1. A Confusion matrix and its relation to predictive accuracy terms.

TPR = True Positive Rate, TNR=True Negative Rate, PPV = Positive Predictive Value, NPV= Negative Predictive Value.

4.4.1.2 Receiver operating characteristic curves and area under the curve

The use of ROC curves is a common way for assessing binary outcome models (Davis and Goadrich, 2006). ROC curves offer a global summary of machine performance at all possible cut-offs of prediction values for defining the two classes. In this way, the ROC is a summary of the model's overall performance. ROC curves reflect the columns of the confusion matrix by presenting FPR (equivalent to 1-TNR)) by true positive rate (TPR), with the advantage of depicting these values at every threshold for defining a hit. An AUC = 0.5 means that the predictive accuracy of the model is not better than chance, whereas an AUC = 1 implies perfect predictive accuracy. (See Chapter 4 in "*Road to Statistical Bioinformatics*" (Lee, 2010) and Chapter 11 in "*Statistical Learning for Biomedical Data*" (Malley et al., 2011) for more details.)

There typically is not just one confusion matrix (see previous section), but rather there is an infinite number: one for each point along the x-axis of the ROC. Thus in the context of a model that outputs prediction values measured on a continuous scale rather than binary

categories (e.g. a logistic regression model among others) one needs to decide at what probability level one "declares" a hit to be a hit. One could use the arbitrary value of greater than 0.5 as the cut-off to declare hits from non-hits, but there are other probability thresholds one could use, which can be summed up in a ROC curve. That is the conceptual difference between the AUC (average over all possible thresholds) and the confusion matrix itself (considers the ROC "frozen" at one particular probability threshold).

It should be noted that unless a weighting scheme such as the one we employed in our modeling or an equal subset of both classes is chosen, ROC curves can present an overly optimistic view of performance for unbalanced data (Davis and Goadrich, 2006). If the model simply assigns all variants to the non-hit class then it will appear to do well, for instance with an AUC much larger than 0.5. In this way, the larger class (non-hits) can overwhelm the smaller class (hits). The TPR thus tends to be low throughout the thresholds.

In the example data, the AUC of two of the models (autoimmune and all phenotype for the high confidence hits) were very similar and reasonably good (between 0.67 and 0.71) (see **Figure 4.2**). The AUC for the other two models (the all phenotype using all Catalogue hits and the brain-related models) were also similar to each other, but poor (less than 0.61). Thus, the AUC seems to categorize models as either good or poor, but is not particularly useful for finer discrimination between models. (See Chapter 11 in "*Statistical Learning for Biomedical Data*" (Malley et al., 2011) for details on the limitations of ROC curves.) Below we demonstrate that additional investigation provides further insight into the results.

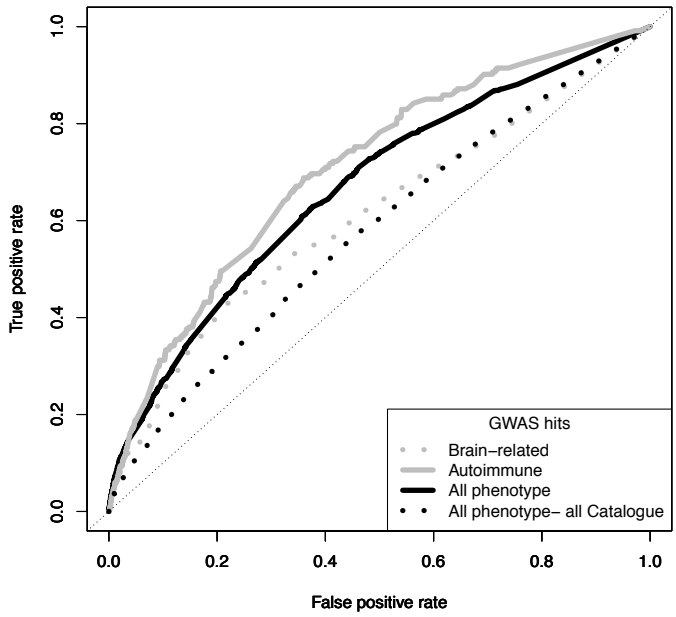


Figure 4.2. ROC curves for the four models.

The precision-recall curve has been proposed to be more appropriate than the ROC for unbalanced data (Davis and Goadrich, 2006). Precision is equivalent to positive predictive value (discussed in the next section) and recall is equivalent to true positive rate (Vihinen, 2012). In this way, the curve depicts information from three of the four cells in the confusion matrix, all of the cells except the true negative cell. An ideal precision-recall curve has data in the top right corner of the plot. Results with the data here (**Figure 4.3**) suggest that none of the models are performing particularly well, suggesting that the ROC AUCs may be driven by the correct identification of the larger class (non-hits).

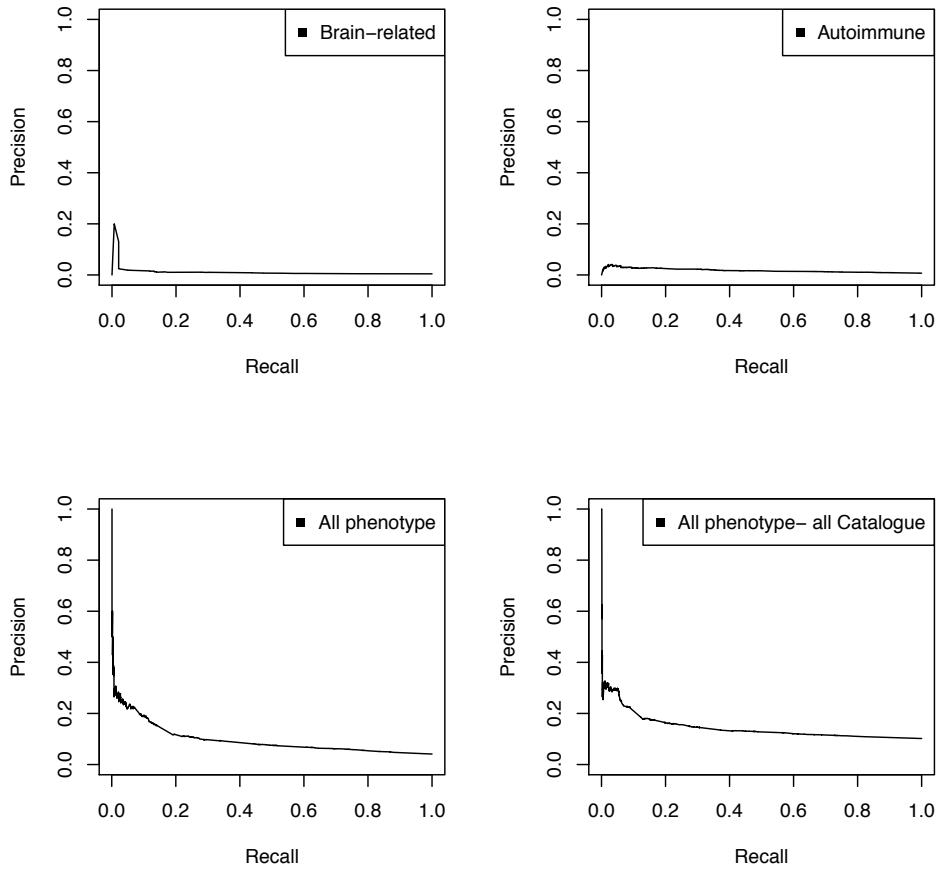


Figure 4.3. Precision-recall curves for the four models.

4.4.1.3 Positive and negative predictive values

The rows of the confusion matrix are represented by PPV and NPV. PPV is the probability of variants that are true hits being correctly classified as hits, and NPV is the probability of variants that are true non-hits being correctly classified as non-hits at any one given threshold. (See Chapter 4 in "*Road to Statistical Bioinformatics*" (Lee, 2010) for details.) PPV and NPV are also affected by the class imbalance inherent in real genetic association data. The effect of imbalanced data on PPV and NPV has been

previously described (Vihinen, 2012). In scenarios where the negative class is larger than the positive class, NPV is inflated and PPV is lower compared to the corresponding model where the class sizes are equal and the negative and predictive classes have the same rate of correct predictions (Vihinen, 2012). These values are best when there are equal amounts of data in each category (Vihinen, 2012). The issue is that cell sizes of the confusion matrix can become too small for the smaller class (hits). One needs to ensure that there is a large enough quantity of hits and/or non-hits per cell in the confusion matrix to draw conclusions. Otherwise, results will be driven by a very small unrepresentative subset of the data. For the models considered here, only the two all phenotype analyses had an adequate amount of samples in each cell, and thus PPV and NPV were only calculated for those models. The NPV tended to be high (>0.899) at all the various prediction value thresholds chosen to define the two classes. See **Table 4.4**. However, it is the accuracy of predicting the hits, not the non-hits, which is of interest in this work. Hence, the PPV provides more interesting results. Overall, the all phenotype analysis using all hits in the GWAS Catalogue produced the highest PPVs as the threshold for declaring a positive hit increased. The highest PPV (30.4%) was achieved for this model at the threshold defining hits as those variants with prediction values greater than 0.7. PPV results conflict between the AUC results. For the two all phenotype models, the one with the higher AUC (the model for the GWAS hits in the Catalogue with the stringent p-value cut-off) had overall lower PPV compared to the model using all GWAS hits in the Catalogue. NPV results for the two models were similar, but the model based on all GWAS hits in the Catalogue had slightly lower NPV compared to the stringent p-value model.

Table 4.4. Positive predictive and negative predictive values at various prediction value cut-offs for the two all phenotype analyses.

	Positive Predictive Values		Negative Predictive Values	
Prediction value cut-off	p<5E-08 hits	all GWAS hits in Catalogue	p<5E-08 hits	all GWAS hits in Catalogue
0.5	0.069	0.128	0.968	0.915
0.6	0.094	0.226	0.956	0.903
0.7	0.198	0.304	0.948	0.899

4.4.2 Visualization of the distribution of prediction values

4.4.2.1 Histograms

Next, class separation was investigated through histograms of the prediction values outputted from the models, which display differences in the density distribution between the two classes. Known hits were plotted in black and non-hits in grey on the same plot, with the y-axis being probability densities, rather than numerical quantity, which masks the data imbalance and thus allows for comparison between the two classes. The all phenotype model with high confidence hits (**Figure 4.4**) and the autoimmune model showed the most evidence of having two separate distributions. Although the distributions of the prediction values for the hits and the non-hits overlap, the distribution of the non-hits has the majority of its values closer to the 0 end of the prediction value range. Confirming the AUC results, the brain-related model and all phenotype model using all Catalogue hits (**Figure 4.4**) do poorly with regard to class separation.

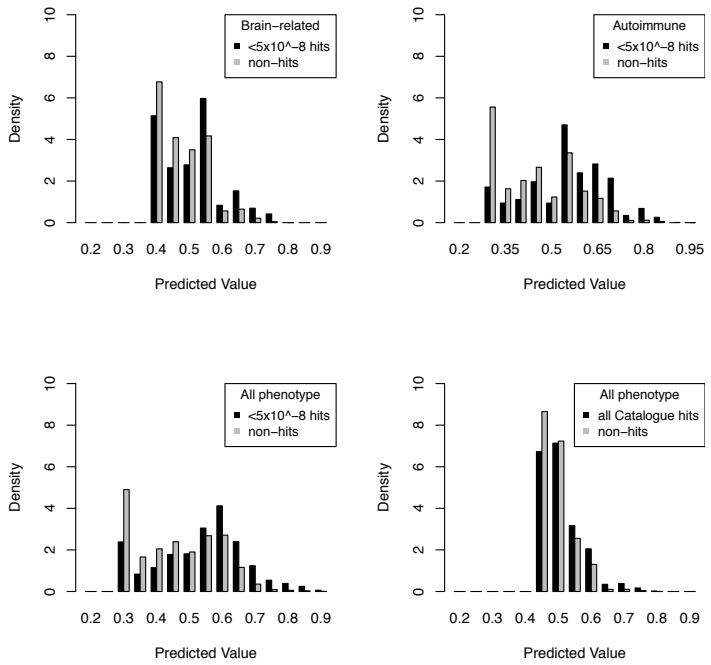


Figure 4.4. Histogram of predictive values for the all phenotype models with a bin size of 0.05.

Compare to Figure 4.5 with a bin size of 0.1. For the probability densities, the sum of the area under the black bars adds up to one. The same is true for the grey bars. The ideal plot would have two non-overlapping distributions with the distribution of the grey bars closest to 0 and the distribution of the black bars close to 1.

As always, caution is warranted since the visualization of the distributions differ depending on the bin size chosen (compare **Figure 4.4** to **Figure 4.5**). For the histograms with a larger bin size differences in distributions between hits and non-hits at a finer scale is less apparent, and the distributions look more similar compared to if a smaller bin size is used.

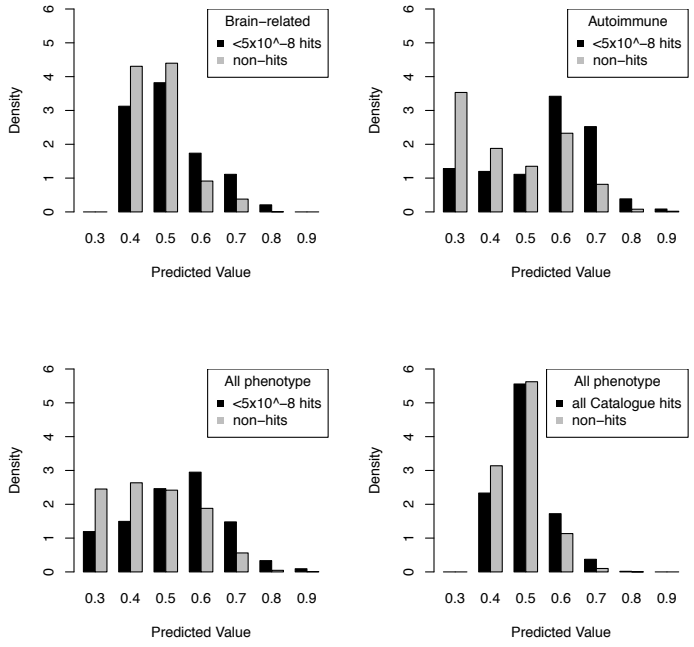


Figure 4.5. Histogram of predictive values for the all phenotype models with a bin size of 0.1.

Compare to Figure 4.4 with a bin size of 0.05. For the probability densities, the sum of the area under the black bars adds up to one. The same is true for the grey bars. The ideal plot would have two non-overlapping distributions with the distribution of the grey bars closest to 0 and the distribution of the black bars close to 1. The bin size is 0.1.

4.4.2.2 Box and whisker plots

Box plots were constructed to visually compare the distributions of the hits versus the non-hits in an alternate way (**Figure 4.6**). These plots visually depict much of the descriptive data present in **Table 4.4 (above)**, notably differences in the median between the two classes. Again the data imbalance is masked as the summaries presented in the plot are from within each class. As visualized in the histograms, the box plots also show that for all of the models the distributions of the prediction values for the hits and non-hits overlapped, but to different degrees. The plots for the brain-related model and the all phenotype model for all variants in the GWAS Catalogue had many outliers for both

classes, signifying that for both hits and non-hits had predictions that were a large distance from the predictions of other variants in the respective class. Additionally, the mean prediction scores for the hits and the non-hits appear very close for the all phenotype model for all variants in the GWAS Catalogue.

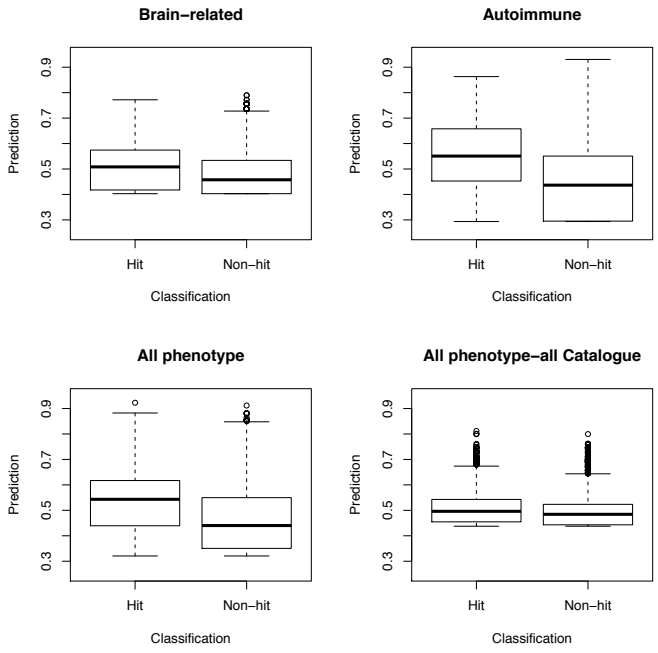


Figure 4.6. Box and whisker plots for the four models.

The line in the box is the median, and the box outlines the 25% and 75% percentiles. Outliers are shown as individual data points if the value is 1.5 times the interquartile range (IQR). The lower and upper whiskers on the plot represent the 25% percentile minus 1.5*IQR and the 75% percentile plus 1.5*IQR, respectively. If the data does not extend as far as those calculated ranges, then the whisker is plotted at the value of the minimum or maximum data point.

4.4.2.3 Violin plots

Violin plots visually combine the density differences depicted in the histograms and the median differences depicted in the box plots into one plot. These plots summarize the

results of the histograms and box plots. Furthermore, they are comparable to a histogram with infinitely small bin sizes. See **Figure 4.7**.

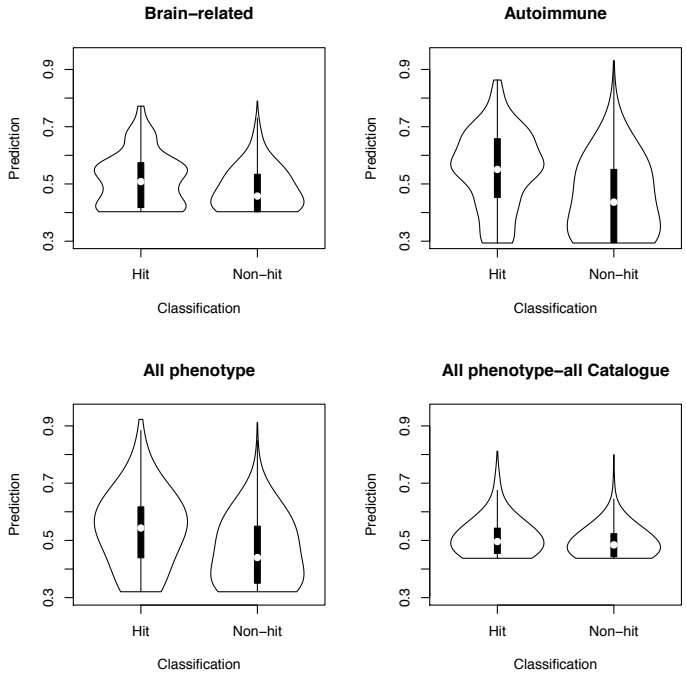


Figure 4.7. Violin plots of the predictive values for the four models.

4.4.2.4 Quantile-quantile plots

A final visualization method, the quantile-quantile plot was explored. See **Figure 4.8**. The quantile-quantile plot is often used in the context of GWAS, but it also has the potential to be useful as a predictive accuracy measures. Instead of expected and observed p-values on the axes as is done in GWAS, we plotted the prediction values for non-hits on the x-axis and the values for the hits on the y-axis. Plotted in this way, the plot compares the quantiles of the hits to the non-hits. When the data points on the plot deviate above the diagonal, the hits have higher prediction values compared to non-hits in that quantile. Due to a limited number of hits, the quantile-quantile plots for the

phenotype-specific analyses produced a staircase pattern. This pattern suggests two characteristics: those models are assigning the same prediction value to several variants, and also there are not enough hits to create a smooth curve. The former could be due to there being different variants that have been assigned identical or similar functional annotations. The models are binning variants together and are not able to differentiate them on a finer scale. The small sample size for the phenotype specific analyses, makes it difficult to draw conclusions from those quantile-quantile plots. For the two all phenotype analyses, the quantile-quantile plots supported the findings from the other visualization methods that the high confidence all phenotype analysis separated hits from non-hits better than the analysis based on hits from the GWAS Catalogue. For the all phenotype model based on the high confidence hits, the distribution consistently deviated from the diagonal. The distribution demonstrates that the hits had higher prediction values than non-hits in the same quantiles. The all phenotype analysis based on all hits in the GWAS Catalogue produced a quantile-quantile plot that closely followed the line for prediction values less than 0.6. This group of prediction values contained most of the data since from the histograms it was determined that the distribution of the prediction values is skewed so that most of the data fall in the lower percentiles. The distribution deviated from the diagonal roughly in the prediction value range of 0.6 and 0.7.

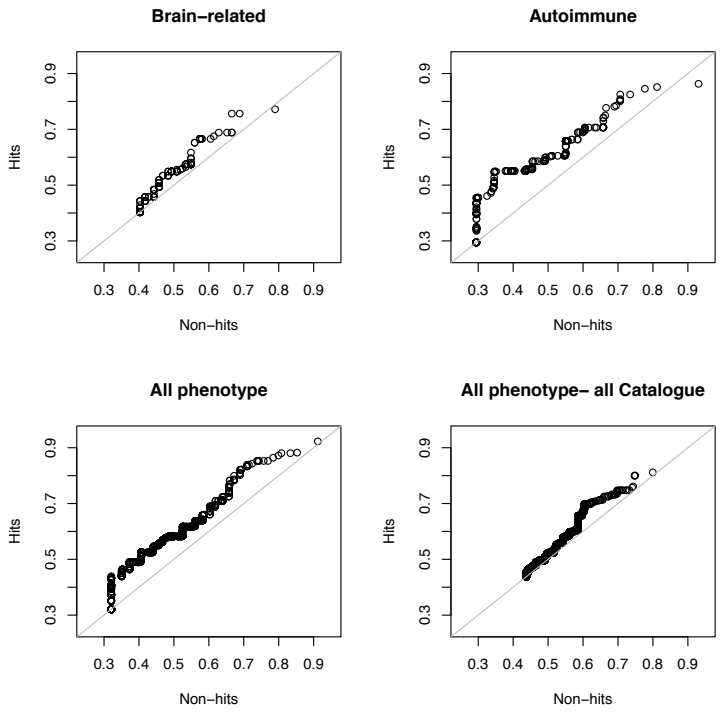


Figure 4.8. Quantile-quantile plots for the four models.

4.4.3 Statistical tests

4.4.3.1 Hypergeometric test

The hypergeometric test was also used to identify significant enrichment of hits compared to non-hits in particular prediction value bins by splitting the data into bin sizes of 0.05 ranging from less than 0.35 up to 0.95. For each model, there were effectively 13 tests performed, one test per prediction value bin. Based on this resulting contingency table, significant enrichment of hits was seen for all of the models in at least one bin greater than 0.55 (with significant p-values ranging from 0.01 to 5.58×10^{-29}), while no enrichment (all p-values greater than 0.2) was seen in bins less than 0.55.

4.4.3.2 Cochran-Mantel-Haenszel test

Another test was investigated, the asymptotic generalized Cochran-Mantel-Haenszel test, which tests the independence of two possibly ordered factors (prediction values of hits vs. non-hits). As with the hypergeometric, a contingency table for hits and non-hits stratified by prediction value was created. Hits and non-hits were stratified independently by prediction values by splitting the data into bin sizes of 0.05 ranging from less than 0.35 up to 0.95. Rather than a single test per prediction value bin as in the hypergeometric, the generalized Cochran-Mantel-Haenszel test is a single omnibus test per model. It looks for a trend across the span of prediction values. Similar to the other statistical tests explored in this section, significant p-values were produced for all models ($p < 5.3x10^{-9}$).

4.4.3.3 Mann-Whitney U test

A two-sided Mann-Whitney U test can be used to determine whether or not the distributions of the prediction values for the hits differs significantly from that of the non-hits. The Mann-Whitney U tests whether the ranks of the variants in the hit and non-hit sets differ. Significant p-values were obtained for all analyses, including those with poor AUCs and poor class separation; most notably the all phenotype analysis not refined to the high confidence hits had a Mann-Whitney p-value of $7.17x10^{-50}$. It was hypothesized that this significant p-value was due to the class imbalance and/or outliers. To explore these hypotheses, only a random subset of non-hits equal in size to the number of hits were selected for the Mann-Whitney U test, and in other test only outliers were removed. In both situations, the p-values tended to remain highly significant (**Table 4.5**).

Table 4.5. Mann-Whitney U p-values for the four models.

Mann Whitney U p value			
	Unaltered	n(hits)= n(nonhits)	No outliers (1.5x outside 25% or 75% percentiles)
Phenotype-specific analyses			
Brain-related	3.49E-06	0.007447	1.76E-05
Autoimmune	8.63E-28	5.26E-15	8.63E-28
All phenotype analyses			
p<5E-08	2.08E-93	3.01E-52	3.53E-92
all Catalogue	7.17E-50	7.26E-27	1.37E-34

The significant Mann-Whitney U p-values do not necessarily suggest that the hits and non-hits are well separated by their prediction values. Instead, the p-values are highlighting differences in ranks between the hits and the non-hits, which may or may not imply class separation. We plotted the hits and non-hits according to their ranks. In all of the plots, the non-hits follow a uniform distribution, whereas the hits follow a different distribution, roughly negatively skewed (**Figure 4.9**). Thus, as with enrichment according to the hypergeometric, and the Cochran-Mantel-Haenszel test for independence, differences in rank according to the Mann-Whitney U are not particularly informative with regard to class separation between the hits and non-hits according to their prediction values.

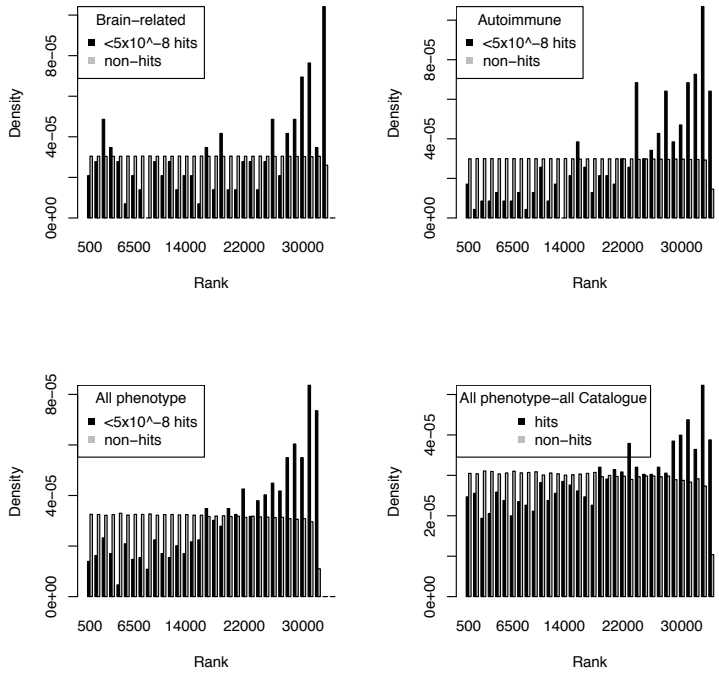


Figure 4.9. Ranked Mann-Whitney U ranks plotted separately for the hits and non-hits.

The non-hits follow a uniform distribution, whereas the hits do not. The same pattern was observed for all four models.

The statistical tests mentioned above do not explicitly measure class separation between hits and non-hits based on their prediction values, which is a key outcome for investigating the predictive accuracy of models for variant prioritization. The hypergeometric assesses enrichment of hits, the Mann-Whitney U tests for differences in ranks between the hits and non-hits, and the generalized Cochran-Mantel-Haenszel test evaluates independence of the hits and non-hits. Thus, significant p-values from these statistical tests cannot alone be taken as proof of class separation or model performance.

4.5 Discussion

In this review we summarized various predictive accuracy measures related to the confusion matrix, visualization methods, and some statistical tests. These methods were described in the context of genetic models for prediction of risk variants in complex traits in which a class imbalance between the hits and non-hits is often inherent.

The choice of predictive accuracy measures was partially motivated by the measures found in the publications described in the background as well as other measures. Note that two of the mentioned papers, (Kindt et al., 2013; Pickrell, 2014), both focused on investigating enrichment or depletion of disease- or trait-associated variants with particular functional and genomic features. Since the predictive accuracy measures in those papers did not relate to an output of a prediction value for each variant, those methods were not discussed further.

In summary, the investigation above emphasizes the importance of visualizing the underlying distributions of the classes. The ROC curve is a good starting place, but visualization measures, especially violin plots, are valuable for differentiating models with similar AUCs. A downside of histograms is that depending on the bin size, the interpretation of the results may vary. With regard to box plots, these plots do not offer any information about density. On the other hand, violin plots are able to show density without the need of binning and at the same time depict the summary statistics that would be seen from a box plot. Caution is needed when making conclusions about model performance based on p-values, such as from the Mann-Whitney U test. Significant p-values cannot necessarily be attributed to a good separation between hits and non-hits. Visualizing the class distribution seems to be the most informative for determining the predictive accuracy in these scenarios.

All of the papers mentioned in the introduction apply their model(s) to real data to assess the accuracy of identifying disease-relevant genetic variants. Predictive accuracy

measures and visualization of the prediction values can only show model performance in theory. When evaluating model performance it is also vital to assess the model in real applications.

4.6 Supporting Data

The R code referred to below can be found in **Appendix C**, and the data files are available on the online version of this chapter that has been published as a paper in *BMC Genomics*: http://www.biomedcentral.com/1471-2164/16/405

File name: Additional_File_1

Sample R code to perform the tests mentioned in this chapter. MyData.txt: Sample output data from a model on which to run the code.

File name: Additional_File_2

R code to reproduce the results in this chapter. Autoimmune-testset.csv, Brain-testset.csv, Nonpheno-5e-8-testset.csv, Nonpheno-allCat-testset.csv: data files required for Code-for-paper.R; they contain five columns: the identifier for the genetic variant, base position,

Chapter 5
Comparison of Statistical Learning Methods Using Functional Annotations for Prioritizing Risk Variants

This chapter is modified from the following: **Gagliano SA,** Ravji R, Barnes MR, Weale ME, Knight J (2015) Smoking Gun or Circumstantial Evidence? Comparison of Statistical Learning Methods using Functional Annotations for Prioritizing Risk Variants. *Scientific Reports* 5:13373.

5

5.1 Abstract

Although technology has triumphed in facilitating routine genome sequencing, new challenges have been created for the data-analyst. Genome-scale surveys of human variation generate volumes of data that far exceed capabilities for laboratory characterization. By incorporating functional annotations as predictors, statistical learning has been widely investigated for prioritizing genetic variants likely to be associated with complex disease. We compared three published prioritization procedures, which use different statistical learning algorithms and different predictors with regard to the quantity, type and coding. We also explored different combinations of algorithm and annotation set. As an application, we tested which methodology performed best for prioritizing variants using data from a large schizophrenia meta-analysis by the Psychiatric Genomics Consortium. Results suggest that all methods have considerable (and similar) predictive accuracies (AUCs 0.64-0.71) in test set data, but there is more variability in the application to the schizophrenia GWAS. In conclusion, a variety of algorithms and annotations seem to have a similar potential to effectively enrich true risk variants in genome-scale datasets, however none offer more than incremental improvement in prediction. We discuss how methods might be evolved for risk variant prediction to address the impending bottleneck of the new generation of genome re-sequencing studies.

5.2 Introduction

Complex diseases are caused by the interplay of many genetic variants and the environment, and represent a considerable health burden. Genome-wide association studies (GWAS) have had success in identifying some genetic risk factors involved in complex diseases such as inflammatory bowel disease (Jostins et al., 2012) and schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium,

2014). Interrogating the entire genome, exome or even selected genes through next generation sequencing technologies have also identified further risk variants (De Rubeis et al., 2014; Epi4K Consortium et al., 2013; Neale et al., 2012; Rivas et al., 2011). However, more disease-associated variants, hereafter called risk variants or hits, remain to be discovered. Some risk variants are difficult to detect by current techniques due to limited sample sizes and low effect size of the variants. *In silico* methodologies that integrate evidence over multiple data sources have the potential to unearth some of these risk variants in a cost-effective manner. The novel risk variants that are identified will help illuminate the genetic risk factors involved in complex diseases, which in turn could lead to earlier or more accurate diagnoses, and the development of personalized treatment options.

Risk variants show enrichment in functional annotations, such as DNase I hypersensitive sites, transcription factor binding sites, and histone modifications; for example, Disanto et al. (2014), Maurano et al. (2012), and Schaub et al. (2012). Several groups have gone further with the results of enrichment by incorporating functional annotations as predictor variables in statistical learning frameworks to prioritize genetic variants for further study (Gagliano et al., 2014a; Kircher et al., 2014; Ritchie et al., 2014). These statistical learning algorithms use the functional annotations to define a model that provides some measure of whether a variant is likely to increase the risk of manifesting a complex trait. However, understanding the relative merits of these approaches requires a thorough investigation into which statistical learning algorithm and/or which combination of functional annotations most effectively identifies novel risk variants.

There are many aspects to consider in the statistical learning framework (**Figure 5.1**). The genetic data input consists of both known risk variants and corresponding control variants (those with no evidence for risk effect); the classifier is used to discriminate between the two. Known risk variants may be identified from sources, such as the GWAS Catalogue (Hindorff et al., 2010), the ClinVar database (Landrum et al., 2014), and the

Human Gene Mutation Database (HGMD) (Stenson et al., 2009) as mentioned above. In addition, the variants can be simulated; for example, Kircher et al. used an empirical model of sequence evolution with local adjustment of mutation rates (Kircher et al., 2014). In this way, the simulated variants would contain *de novo* pathogenic mutations. The goal of these methods is to identify disease-causing variants, but their application can differ depending on whether the data under consideration consist of densely mapped variants, as in sequence data, or coarsely mapped variants, as in GWAS data. The use of different classifiers has the effect of refining the goal, in that coarsely mapped variants may tag other variants in high linkage disequilibrium, and so the functional characteristics of these other variants should be taken into account. The methods we investigate have been applied to both types of data (Griswold et al., 2014; Parra et al., 2014).

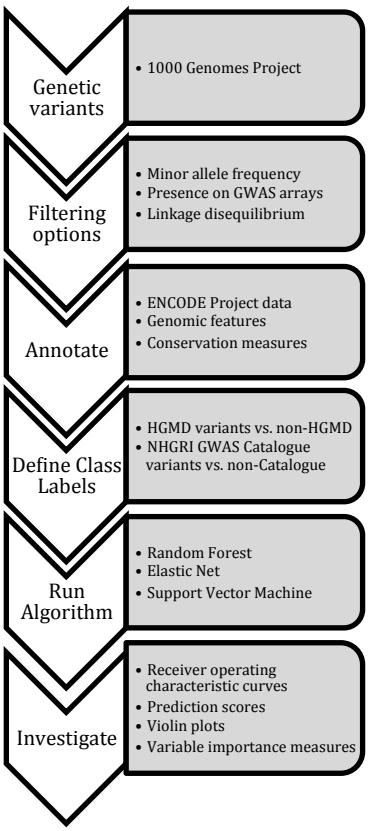


Figure 5.1. Various steps in the statistical learning pipeline for genetic variant prioritization using functional annotations, with examples outlined for each

GWAS=Genome-wide association studies; ENCODE= Encyclopedia of DNA Elements; NHGRI= National Human Genome Research Institute; HGMD= Human Gene Mutation Database

With regard to the functional annotations, some come from experimental procedures while others are predicted computationally. Examples include genomic and epigenomic annotations that can be incorporated from various online browsers and collections such as the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2010) and the Encyclopedia of DNA Elements (ENCODE) Project (The ENCODE Project Consortium, 2011). Whether a variant is assigned the annotations that can be attributed to itself only or to other variants with which it is in linkage disequilibrium can also refine the goal of the method.

In this chapter, we compared the performance of three published methods that differ in annotation set, algorithm and genetic variants, including the classifier: a regularized regression called elastic net from Gagliano et al. (14 annotations) discussed in Chapter 3, a modified random forest from Ritchie et al. (174 annotations) (Ritchie et al., 2014) called GWAVA and a support vector machine from Kircher et al. (949 annotations, expanded from 63 unique annotations) called CADD, v.1.0 (major release) (Kircher et al., 2014). These three papers describe algorithms capable of incorporating a large number of genetic variants labeled with multiple functional annotations, and can output a prediction score for each variant; hence, they are highly comparable. Although other methods exist to prioritize genetic risk variants, such as through the use hierarchical Bayesian analysis (Kichaev et al., 2014; Pickrell, 2014), these require genetic association statistics for each variant for prioritization, and thus were beyond the scope of the comparisons in this paper. We investigate nine model types: combinations of the three different statistical learning algorithms and the three different functional annotation sets (summarized in **Table 5.1**). All model types were created for different classifications of hits: the NHGRI GWAS Catalogue (Hindorff et al., 2010) and the Human Gene Mutation Database (HGMD) (Stenson et al., 2009).

Table 5.1. Comparison of the three data-trained genetic variant prioritization papers

	Gagliano *et al.* (PLoS ONE 2014)	**Ritchie *et al.* (Nat Methods 2014) "GWAVA"**	**Kircher *et al.* (Nat Genetics 2014) "CADD"**
Functional annotations	n= 14 (ENCODE, eQTLs, PhastCons, Genic context…)	n= 174 (ENCODE, GERP, Genic context…)	n=63 (expanded to 949) (Ensembl VEP, ENCODE, PolyPhen…)
Risk variants ("Hits") [N]	NHGRI GWAS Catalogue (p-value ≤ $5x10^{-8}$) [3227 in non-phenotype specific model]	HGMD - "regulatory" [1614 in most stringent matched by gene region model"]	Simulated mutations under neutral model - "gap" sites [14.7 million]
Non-risk variants ("Non-hits") [N]	union of common Illumina and Affymetrix GWAS panels [75,341 in non-phenotype specific model]	Other variants in 1000 Genomes Project (for example, within 1kb of each HGMD variant) [5027 in gene region model]	high-frequency derived human alleles from 1000 Genomes [14.7 million]
Classifier algorithm	Elastic net	Random forest	Support vector machine
Training protocol	60% training 40% reserved for testing	100% training	99% training 1% reserved for testing

Models based on GWAS data can be tested effectively in current data (we apply those models to the schizophrenia GWAS from the Psychiatric Genomics Consortium). For the purpose of this thesis we have kept this chapter largely in the format in which it was submitted; hence Methods appear at the end of this chapter in **Section 5.5**.

5.3 Results

Our primary analysis used the NHGRI GWAS Catalogue as the classifier. Risk variants/hits were defined as those variants present in the NHGRI GWAS Catalogue (www.genome.gov/gwastudies, downloaded on August 7, 2014) (Hindorff et al., 2010) with a p-value of equal to or less than the accepted threshold for genome-wide significance, $5x10^{-8}$. A subset of non-hits (that are not in high linkage disequilibrium with the hits) was selected from common GWAS arrays for comparability. For the three annotation sets described above, when working with different classifiers some rare

annotations have no variability and hence were not used to build the model. In this analysis none of the 14 annotations from Gagliano et al. were invariable, three of the 174 annotations from Ritchie et al. were invariable, and 509 of the 949 annotations from Kircher et al. were invariable. An independent test set was used to determine accuracy of the models for discriminating hits from non-hits based on the predictive score output from each model. These results are presented below.

5.3.1 Area under the ROC curve

All the models had similar accuracy as demonstrated by the area under the curve (AUC) in the test set data (**Table 5.2**). Models using Kircher et al.'s annotations produced slightly higher AUCs compared to the other two annotation sets for the elastic net and random forest algorithms. In particular the combination of elastic net and Kircher et al.'s annotations was the only model that produced an AUC with confidence intervals that do not overlap with any of the other models.

Table 5.2. The area under the curve (AUC) for the GWAS Catalogue comparisons, holding data and classifier constant, while varying algorithm and annotations.

The 95% confidence interval based on 2000 bootstrap replicates (generated using the R package pROC) is shown in square brackets. The AUC in the training set is in parentheses.

Annotations →	**Gagliano et al.**	**Ritchie et al.**	**Kircher et al.**
Elastic Net	0.67 [0.65-0.68] (0.67)	0.65 [0.63-0.66] (0.67)	0.71 [0.69-0.73] (0.74)
Random Forest (altered minimum node size)	0.67 [0.65-0.68] (0.69)	0.68 [0.66-0.69] (0.72)	0.70 [0.68-0.72] (0.79)
Support Vector Machine (with prior feature selection)	0.66 [0.65-0.68] (0.66)	0.64 [0.63-0.66] (0.66)	0.64 [0.61-0.66] (0.68)

The AUC results for the training set were also computed to investigate whether the models were over-fit; that is to say, whether the training set AUC is much higher than the

test set AUC. We found that for the Ritchie et al. and Kircher et al. annotation sets, the random forest models with node size equal to one were prone to over-fitting. For instance, for the random forest model based on the Ritchie et al. annotations, the test set and training set AUCs were 0.687 and 0.998, respectively (further data available on request). The over-fitting in the random forest models was solved when the minimum node size was set to 10% of the total sample size. Therefore only the random forest models with the minimum node size equal to 10% of the data are presented in **Table 5.2** and discussed further in the results. These results highlight the importance of ensuring that appropriate parameters are chosen for the algorithms.

5.3.2 Density and distribution of prediction scores

Violin plots were constructed by plotting the prediction scores for hits (risk variants) and non-hits separately in order to visualize how well the two classes separated (**Figure 5.2** and **Table 5.3**). The two models with the best AUCs (Kircher et al. annotations with elastic net (0.71) and with random forest (0.70)) have comparatively well separated means and relatively normal distributions. In one of the two models with the lowest AUC (Ritchie et al. annotations with support vector machine (0.64)), the median prediction score between hits and non-hits is most similar and the distribution is very skewed. Interestingly, one of the mid-range performance models, the Gagliano et al. annotations for the support vector machine (0.66) showed evidence of a multimodal distribution where one mode is more common for hits and another for non-hits. However, this effect may simply be due to the comparatively small number of annotations, which lead to a smaller number of possible scores.

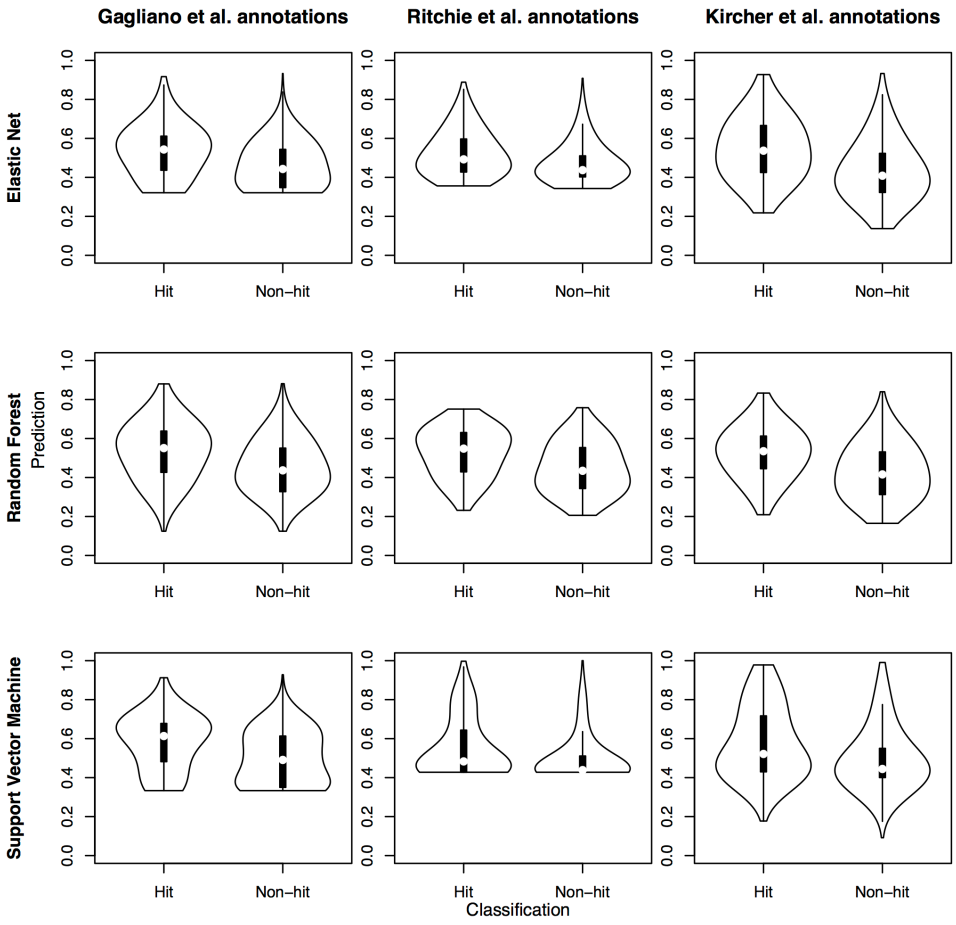


Figure 5.2. Violin plots showing class separation by prediction scores for the various comparisons using the GWAS Catalogue as the classifier

Hits are variants in the GWAS Catalogue with a genome-wide significant p-value ($p<5x10^{-8}$) and non-hits are those not present in the GWAS Catalogue, but are found on common GWAS arrays for comparison purposes. The non-scaled elastic net models are plotted. The adjusted minimum node size (10%) random forest models are plotted.

Table 5.3. Summary statistics of the prediction score distributions for the various models based on the GWAS Catalogue classifier

For a visual representation see the violin plots (**Figure 5.2**). [SD=standard deviation]

Functional Annotations		Gagliano et al.		Ritchie et al.		Kircher et al.	
		Hits	Non-hits	Hits	Non-hits	Hits	Non-hits
Elastic Net (not scaled)	Minimum	0.32	0.32	0.36	0.34	0.22	0.14
	Median	0.54	0.44	0.49	0.44	0.54	0.41
	Mean	0.54	0.46	0.52	0.47	0.55	0.43
	Maximum	0.92	0.93	0.89	0.91	0.93	0.93
	SD	0.13	0.12	0.11	0.09	0.15	0.15
Random Forest (altered minimum node size)	Minimum	0.12	0.12	0.23	0.21	0.21	0.16
	Median	0.55	0.44	0.55	0.43	0.53	0.44
	Mean	0.54	0.46	0.53	0.45	0.42	0.43
	Maximum	0.88	0.88	0.75	0.76	0.83	0.84
	SD	0.13	0.12	0.12	0.13	0.12	0.14
Support Vector Machine (with prior feature selection)	Minimum	0.33	0.33	0.43	0.43	0.18	0.09
	Median	0.61	0.49	0.48	0.44	0.52	0.44
	Mean	0.58	0.50	0.55	0.49	0.58	0.50
	Maximum	0.91	0.93	1.00	1.00	0.98	0.99
	SD	0.14	0.14	0.15	0.11	0.18	0.14

Generally, the models created using the Kircher et al. annotations showed the largest spread of prediction scores for both hits and non-hits. We have also reported the proportion of hits in the top versus the bottom quartiles of the prediction scores in the test set (**Table 5.4**). In summary the violin plots show that the distributions for hits and non-hits overlapped for all models. However, we see from **Table 5.4** that of the variants in the top quartile of prediction scores, there are significantly more hits compared to the lower quartile for all models assessed ($p < 2.2 \times 10^{-16}$, chi-square test).

Table 5.4. Proportion of GWAS Catalogue hits for the various models

Results are shown for the variants in the test set data that were assigned the highest prediction scores (top quartile) and the lowest scored variants (lower quartile). The difference row shown corresponds to the proportion of GWAS significant variants in the top quartile minus that of the lower quartile, so a positive difference suggests that the quartile of the most highly scored variants (top quartile) contains more GWAS significant variants compared to the lowest scored variants (lower quartile). The number of variants present in each quartile are in parentheses. Note that quartiles can vary in size where prediction scores are identical across many variants, and all those variants with that particular score were included in the quartile.

Annotation set						
	Gagliano et al.		Ritchie et al.		Kircher et al.	
	Elastic Net					
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	8.8% (7872)		7.4% (7823)		10% (2656)	
lower quartile	2.2% (8261)	< 2.2e-16	2.1% (7837)	< 2.2e-16	1.1% (2655)	< 2.2e-16
Difference	6.6%		5.3%		9.3%	
	Random Forest					
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	8.8% (7956)		7.8% (7826)		10% (2654)	
lower quartile	2.2% (7889)	< 2.2e-16	1.4% (7825)	< 2.2e-16	1.0% (2654)	< 2.2e-16
Difference	6.6%		6.4%		9.1%	
	Support Vector Machine					
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	8.1% (7873)		7.3% (8150)		8.1% (2655)	
lower quartile	2.2% (7807)	< 2.2e-16	2.2% (7555)	< 2.2e-16	2.9% (2654)	< 2.2e-16
Difference	5.8%		5.1%		5.2%	

To investigate the consistency of the models we calculated pairwise correlations of the prediction scores in the test set for the various models either holding the algorithm or the annotation set constant. We found that the models with the most correlated scores were those using the Gagliano et al. annotation set. Furthermore, the degree of correlation when holding the algorithm constant, but varying the annotation set, was generally not as high as when holding the annotation set constant (**Table 5.5**).

Table 5.5. Pairwise correlation between prediction scores in the test set between models either holding the annotation set or the algorithm constant in the primary analysis

EN= elastic net, RF=random forest, SVM= support vector machine

			Annotation set								
			Gagliano et al.			Ritchie et al.			Kircher et al.		
		Algorithm	EN	RF	SVM	EN	RF	SVM	EN	RF	SVM
Annotation set	Gagliano et al.	EN	--	0.95	0.98	0.41	--	--	0.47	--	--
		RF	0.95	--	0.93	--	0.47	--	--	0.51	--
		SVM	0.98	0.93	--	--	--	0.28	--	--	0.35
	Ritchie et al.	EN	0.41	--	--	--	0.84	0.79	0.71	--	--
		RF	--	0.47	--	0.84	--	0.66	--	0.82	--
		SVM	--	--	0.28	0.79	0.66	--	--	--	0.69
	Kircher et al.	EN	0.47	--	--	0.71	--	--	--	0.84	0.72
		RF	--	0.51	--	--	0.82	--	0.84	--	0.69
		SVM	--	--	0.35	--	--	0.69	0.72	0.69	--

5.3.3 Feature selection within elastic net and random forest

More does not necessarily equal better as not all the annotations may be relevant to predicting risk variants. Generally, not all of the functional annotations in the annotation sets were used to create the various models. For instance of the variable features, elastic net assigned non-zero Beta coefficients to 9 out of 14 annotations, 12 out of 171, and 16 out of 432. Random forest assigned non-zero Gini importance values to all of the 14, 131 out of 171, and 239 out of 432. All of these models had similar performance in the test sets (AUCs ranging from 0.68 to 0.70 for the random forest models and 0.65 to 0.71 for the elastic net models). The results suggest that elastic net has a more stringent feature selection implementation than random forest. The support vector machine models always assigned non-zero feature weights, as support vector machine does not intrinsically perform feature selection, as does elastic net and random forest. Thus, we inputted only

those annotations with a non-zero Beta coefficient from the elastic net models into the support vector machine models (see **Methods**).

5.3.4 Importance of the functional annotations

Different combinations of annotations can be used to obtain models with similar predictive accuracy. Furthermore, it is difficult to interpret the importance of the annotations for numerous reasons, some of which are discussed below.

All three annotation sets contained a mixture of binary variables and continuous variables. For Kircher et al.'s annotations, background selection (the annotation with the widest continuous scale that ranged from 0 to 1000) came up as most important for predicting the class label in the random forest model. This bias for random forest preferentially selecting annotations measured on a continuous scale has been previously described (Strobl et al., 2007). When making a decision at a node, continuous annotations can be used multiple times at varying cut-offs to split the data. In this way, functional annotations measured on a continuous scale are incorporated more often into the forest compared to non-continuous annotations, and thus obtain higher variable importance measures (Boulesteix et al., 2014; Strobl et al., 2007).

It is also difficult to interpret the variable importance measures derived from elastic net because this algorithm is not scale invariant. Using Gagliano et al.'s annotations with elastic net, we compared the models created with scaled (all annotations have a standard deviation of 1 and a mean of 0) versus non-scaled annotations. Although the AUCs for both models were nearly identical, the assigned Beta coefficients differed (**Figure 5.3**). When we do standardize the scale, we find that the order of importance of coefficients replicates that of the random forest model. However, standardizing a set of largely binary variables removes the effect linked to the frequency, and thus skews the biological representation. So it is not clear that scaling is the best approach.

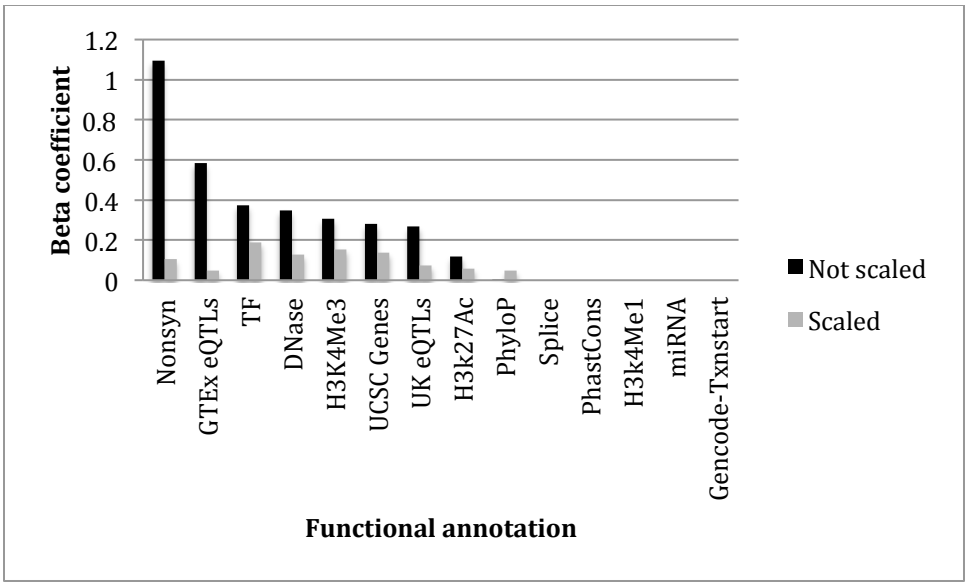


Figure 5.3. Feature importance for elastic net models using the Gagliano et al. annotations based on the GWAS Catalogue classifier

The importance of annotations differed when using scaled versus non-scaled annotations in elastic net [splice= splice sites, Nonsyn= nonsynonymous SNPs, DNase= DNase I hypersensitive sites, GTEx eQTLs= cis-eQTL data from the GTEx Consortium, UK eQTLs= cis-eQTL data from the UK Brain Consortium, Phylo= PhyloP conservation, PhastCons= PhastCons conservation, H3K4MeMe1= H3K4Me1 histone modification, H3K4Me3= H3K4Me3 histone modification, H3K27Ac=H3K27Ac histone modification, TF= transcription factor binding sites, miRNA= micro RNA targets, Gencode-Txnstart= transcription start sites from Gencode]

Although the focus is not about annotations we have provided details of the various importance measures in **Appendix A** for the feature importance measures from all the models based on the GWAS Catalogue as the classifier. In the primary analysis transcription factor binding sites were consistently in the top three annotations for the Gagliano et al. annotations for all three algorithms, but there were no other clear patterns with regard to important annotations for the Ritchie et al. or Kircher et al. annotation sets. In summary, different annotations came up as most important for the various models regardless of predictive accuracy.

5.3.5 Performance for complex disease variants: Application to Schizophrenia GWAS

Various quantile-quantile plots were constructed in order to compare which models showed greater separation of the schizophrenia GWAS p-values for high scoring and low scoring functional variants. For all of the models, scores were obtained for the sub-genome-wide-significant variants ($5x10^{-8}<p<1x10^{-6}$) from the first round of the GWAS by the Psychiatric Genomics Consortium (PGC1) (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). The PGC1 p-values were plotted on the x-axis and the p-values from the second larger round of the schizophrenia GWAS (PGC2) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) were plotted on the y-axis (**Figure 5.4**). (The results from PGC2 were not used to train the model.) Plots were constructed where annotations were held constant but the algorithm differed. For instance, for the 14 annotations from Gagliano et al. we plotted the models from the three algorithms in one plot. Furthermore, models from the same algorithm but varying by annotation set were compared (**Figure 5.5**).

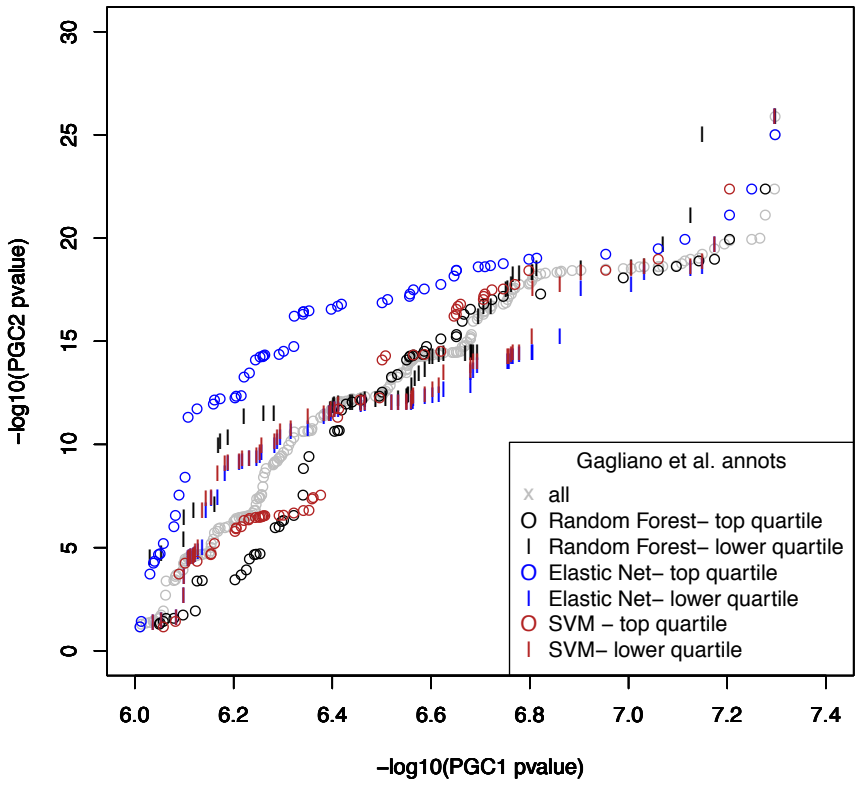

−log10(PGC2 pvalue)
0
5
10
15
20
25
30
6.0
6.2
6.4
6.6
6.8
7.0
7.2
7.4
−log10(PGC1 pvalue)
Gagliano et al. annots
all
Random Forest− top quartile
Random Forest− lower quartile
Elastic Net− top quartile
Elastic Net− lower quartile
SVM − top quartile
SVM− lower quartile

Figure 5.4a

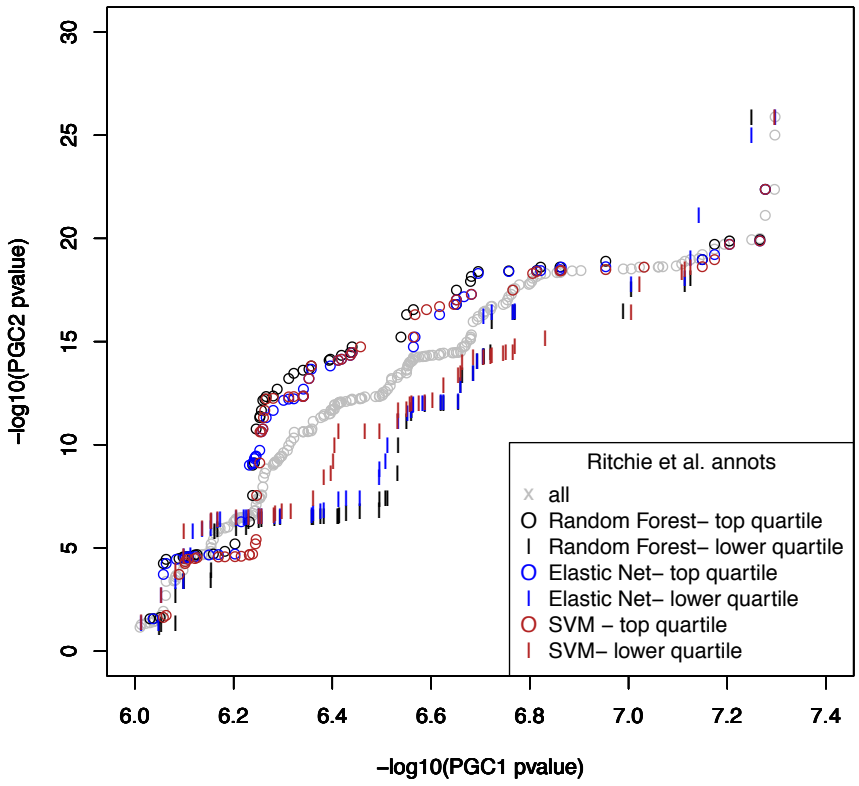

−log10(PGC2 pvalue)
0
5
10
15
20
25
30
6.0
6.2
6.4
6.6
6.8
7.0
7.2
7.4
−log10(PGC1 pvalue)
Ritchie et al. annots
all
Random Forest− top quartile
Random Forest− lower quartile
Elastic Net− top quartile
Elastic Net− lower quartile
SVM − top quartile
SVM− lower quartile

Figure 5.4b

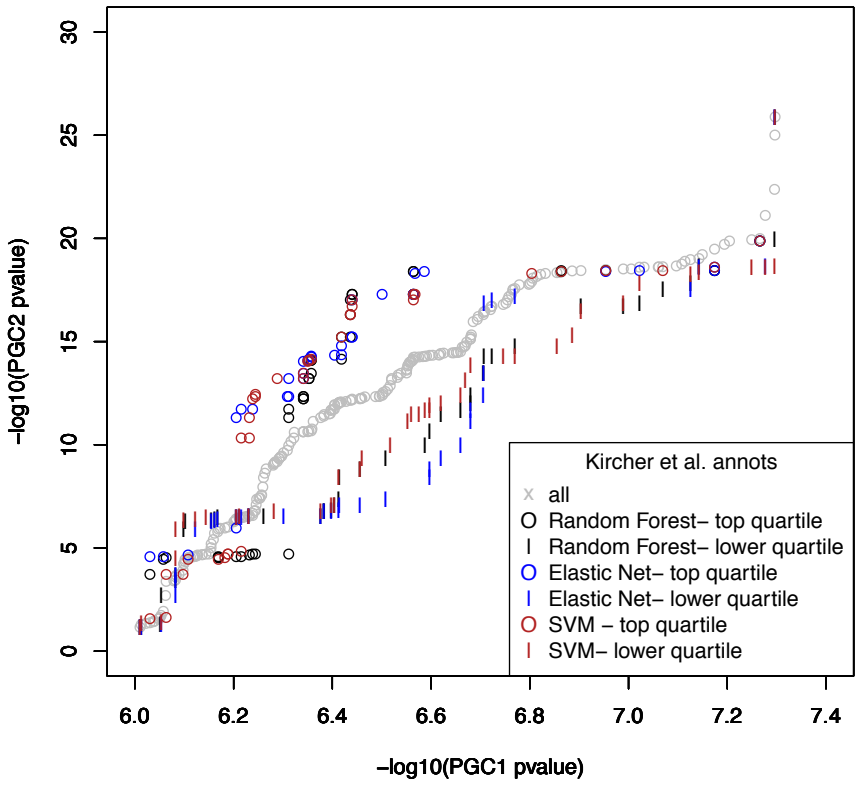


Figure 5.4c

Figure 5.4. Quantile-quantile plots of PGC1 sub-genome-wide-significant variants ($5x10^{-8}<p<1x10^{-6}$) stratified by prediction score for the various models based on the GWAS Catalogue classifier, and plotted by PGC2 p-values

PGC1 p-values are plotted on the x-axis and PGC2 p-values are plotted on the y-axis. Models grouped by annotation set: Gagliano et al. [**a**], Ritchie et al. [**b**], and Kircher et al. annotations [**c**]. The lower quartile genetic variants are those PGC1 sub-genome-wide-significant variants that were assigned the lowest prediction scores (in the first quartile), and the top quartile variants are those with the highest prediction scores (in the fourth quartile).

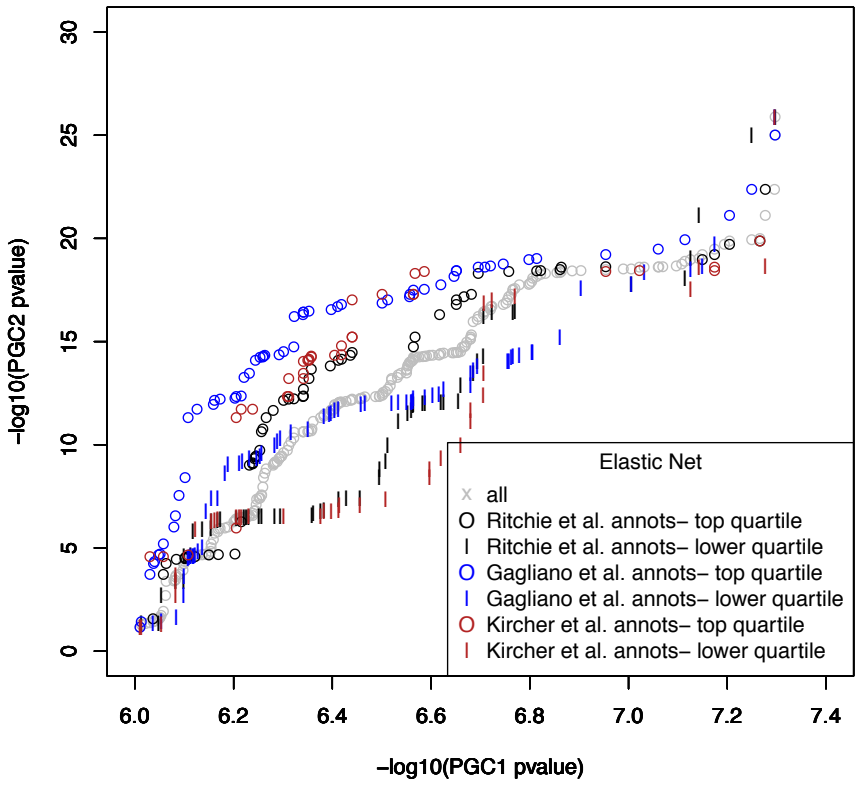

−log10(PGC2 pvalue)
0
5
10
15
20
25
30
6.0
6.2
6.4
6.6
6.8
7.0
7.2
7.4
−log10(PGC1 pvalue)
Elastic Net
all
Ritchie et al. annots− top quartile
Ritchie et al. annots− lower quartile
Gagliano et al. annots− top quartile
Gagliano et al. annots− lower quartile
Kircher et al. annots− top quartile
Kircher et al. annots− lower quartile

Figure 5.5a

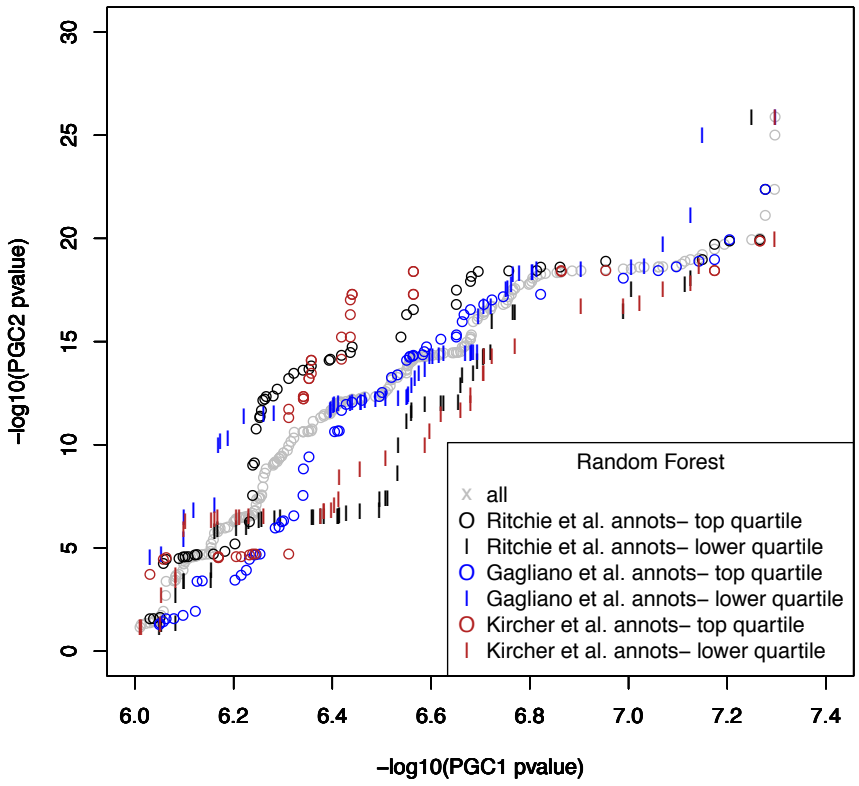

−log10(PGC2 pvalue)
0
5
10
15
20
25
30
6.0
6.2
6.4
6.6
6.8
7.0
7.2
7.4
−log10(PGC1 pvalue)
Random Forest
all
Ritchie et al. annots− top quartile
Ritchie et al. annots− lower quartile
Gagliano et al. annots− top quartile
Gagliano et al. annots− lower quartile
Kircher et al. annots− top quartile
Kircher et al. annots− lower quartile

Figure 5.5b

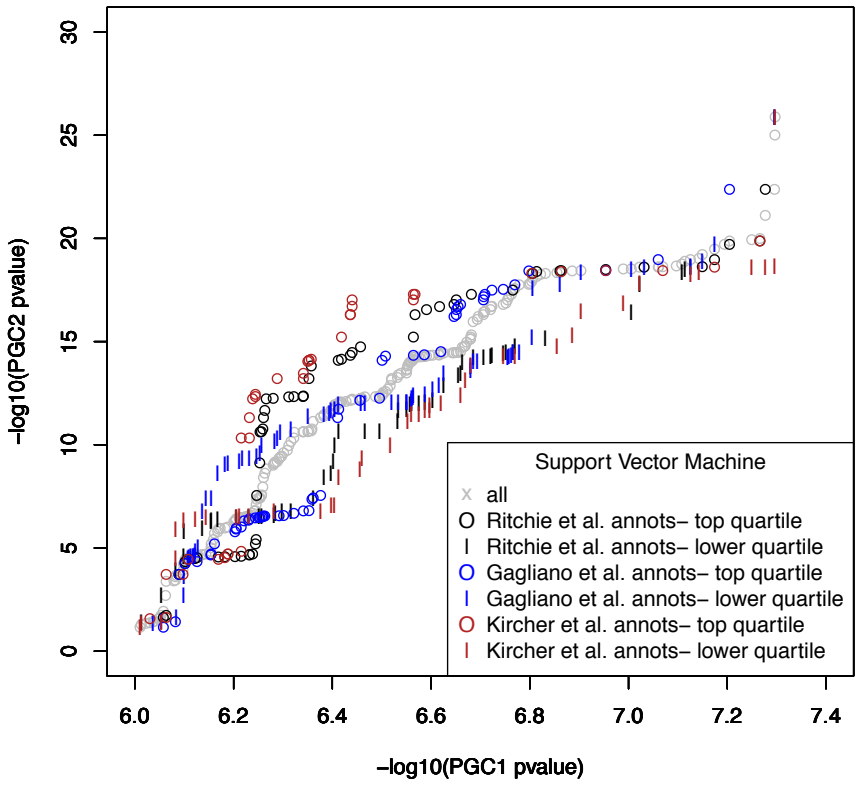


Figure 5.5c

Figure 5.5. Quantile-quantile plots of PGC1 sub-genome-wide-significant variants ($5x10^{-8}<p<1x10^{-6}$) stratified by prediction scores for the various models based on the GWAS Catalogue classifier, and plotted by -log10(PGC1 p-values) versus -log10(PGC2 p-values)

Models grouped by algorithm: elastic net (non-scaled annotations) [**a**], random forest (adjusted minimum node size) [**b**], and support vector machine (with prior feature selection) [**c**]. The lower quartile genetic variants are those PGC1 sub-genome-wide significant variants that were assigned the lowest prediction scores (in the first quartile), and the top quartile variants are those with the highest prediction scores (in the fourth quartile).

We have also reported the proportion of hits in the top versus the bottom quartiles of the prediction scores in the test set (**Table 5.6**). With regard to the functional annotation set, the separation of the novel associated variants from the non-associated in the sub-genome-wide-significant variants was best exhibited in the quantile-quantile plots when using either the Kircher et al. or Ritchie et al. annotation sets. Regardless of annotation set, the elastic net models consistently showed good separation. For all algorithms using either the Ritchie et al. or Kircher et al. annotations, the PGC1 sub-genome-wide-significant variants that have the highest prediction scores (within the top quartile) consistently contain a higher proportion of GWAS significant variants from the second round of the schizophrenia GWAS ($p<5x10^{-8}$) compared to the variants that have scores in the lower quartile. The elastic net models too, regardless of annotation set, showed this pattern. Although these patterns are not all statistically significant, it is notable that the biggest positive difference comes from using the Ritchie et al. annotations with the elastic net algorithm, and the most significant difference between the proportion of GWAS significant variants in the top quartile compared to the proportion in the lower quartile comes from the Kircher et al. annotations using the elastic net algorithm; (there are more variants available in the Kircher et al. model than the Ritchie et al. model). The Gagliano et al. annotations performed very poorly with both the random forest and support vector machine algorithms since the variants with low prediction scores were more likely to be hits than those with high scores. This is a result of the PGC2 hits not being enriched in two of the top annotations for the Gagliano et al. models using either the random forest or support vector machine algorithms, H3K4Me3 and H3K27Ac. In the GWAS Catalogue analysis of the variants that possess the H3K4Me3 and H3K27Ac marks, nearly 70% are hits and the remainder are non-hits. In comparison, of the PGC1 sub-genome-wide-threshold variants that possess those two annotations, only 21% are PGC2 hits, and the remaining variants are non-hits.

Table 5.6. Pairwise correlation between prediction scores in the test set between models either holding the annotation set or the algorithm constant in the primary analysis

Results are shown for the variants that were assigned the highest scores (top quartile) and the lowest scored variants (lower quartile). The difference row shown corresponds to the proportion of GWAS significant variants in the top quartile minus that of the lower quartile, so a positive difference suggests that the quartile of the most highly scored PGC1 sub-genome-wide significant variants (top quartile) contains more GWAS significant variants from PGC2 compared to the lowest scored PGC1 sub-genome-wide significant variants (lower quartile). The number of variants present in each quartile are in parentheses. Note that quartiles can vary in size where prediction scores are identical across many variants, and all those variants with that particular score were included in the quartile.

Annotation set						
	Gagliano et al.		Ritchie et al.		Kircher et al.	
	Elastic Net					
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	83% (60)		77% (56)		54% (34)	
lower quartile	79% (66)	0.52	55% (56)	0.02	43% (37)	7.30E-05
Difference	4%		22%		11%	
	Random Forest					
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	65% (60)		72% (55)		71% (41)	
lower quartile	90% (59)	1.20E-03	51% (55)	0.02	53% (43)	0.10
Difference	-25%		21%		18%	
	Support Vector Machine					
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	50% (54)		70% (56)		73% (37)	
lower quartile	79% (68)	6.30E-04	67% (52)	0.79	64% (42)	0.41
Difference	-29%		3%		9%	

The results for the application to the schizophrenia GWAS did not always reflect the AUCs from the training data. For instance, a poor performing model in terms of AUC based on the test set, elastic net with the Ritchie et al. annotations, performed well in the GWAS application. All in all, the accuracy of the resulting models should be assessed by various means, including (but not limited to) theoretical models such as the ROC curve,

as well as empirical approaches such as applying the model using data from one study and evaluating its performance on independent data with gold standard answers.

5.3.6 HGMD Analysis

In an attempt to apply the algorithms and annotation set combinations to whole genome sequencing data, and indeed fine-mapping studies, rather than just GWAS, a different classifier was used to identify hits and non-hits, the Human Gene Mutation Database (HGMD). We conducted two analyses with subsets of the public release of HGMD. In the first, we took all the variants (single nucleotide polymorphisms) in HGMD and chose controls that fell within a kilobase of either side from the HGMD variant. In this analysis one of the 14 annotations from Gagliano et al. was invariable, eight of the 174 annotations from Ritchie et al. were invariable, and 396 of the 949 annotations from Kircher et al. were invariable. Secondly, models based on the subset of non-exonic HGMD variants and non-exonic control variants were assessed. This second set of models was created in an effort to overcome the ascertainment bias inherent in HGMD related to genes. In this analysis two of the 14 annotations from Gagliano et al. were invariable, 16 of the 174 annotations from Ritchie et al. were invariable, and 756 of the 949 annotations from Kircher et al. were invariable.

The models for the analysis using all of the HGMD variants using either the Ritchie et al. or Kircher et al. annotations had high predictive accuracy (**Table 5.7**).

Table 5.7. The area under the curve (AUC) for the HGMD comparisons, holding data and classifier constant, while varying algorithm and annotations

The 95% confidence interval based on 2000 bootstrap replicates (generated using the R package pROC) is shown in square brackets. The AUC in the training set is in parentheses.

Annotations →	**Gagliano et al.**	**Ritchie et al.**	**Kircher et al.**
Elastic Net	0.66 [0.64-0.67] (0.65)	0.87 [0.86-0.88] (0.88)	0.88 [0.87-0.89] (0.88)
Random Forest (altered minimum node size)	0.65 [0.64-0.66] (0.66)	0.91 [0.90-0.92] (0.91)	0.87 [0.86-0.88] (0.89)
Support Vector Machine (with prior feature selection)	0.63 [0.62-0.64] (0.66)	0.85 [0.83-0.86] (0.86)	0.85 [0.84-0.86] (0.87)

The AUCs for the non-exonic HGMD analysis were more comparable to the ones obtained for the primary analysis using the GWAS Catalogue as the classifier (**Table 5.8**), but again the annotations from Ritchie et al. and Kircher et al. performed better.

Table 5.8. The area under the curve (AUC) for the non-exonic HGMD comparisons, holding data and classifier constant, while varying algorithm and annotations

The 95% confidence interval based on 2000 bootstrap replicates (generated using the R package pROC) is shown in square brackets. The AUC in the training set is in parentheses.

Annotations →	**Gagliano et al.**	**Ritchie et al.**	**Kircher et al.**
Elastic Net	0.65 [0.61-0.68] (0.66)	0.77 [0.74-0.80] (0.78)	0.79 [0.76-0.81] (0.80)
Random Forest (altered minimum node size)	0.65 [0.61-0.68] (0.65)	0.80 [0.77-0.82] (0.86)	0.78 [0.75-0.80] (0.85)
Support Vector Machine (with prior feature selection)	0.61 [0.58-0.65] (0.68)	0.68 [0.65-0.72] (0.78)	0.76 [0.73-0.78] (0.82)

Similar to the analysis using the GWAS Catalogue as the classifier, for the HGMD analysis models the features that came up as most important tended to vary depending on the algorithm and are difficult to interpret. It is however notable that genic annotations featured highly (see **Appendix A)**. For the Gagliano et al. annotations, the top annotation (or the second most important in the case of support vector machine) was nonsynonymous SNPs. For the Kircher et al. annotations, the top annotations for the random forest and support vector machine models were related to the coding sequence or nonsynonymous SNPs. The top annotation for elastic net was CpG. For the Ritchie et al. annotations, the top two annotations were coding sequence and exon for both the random forest and support vector machine models. For elastic net, the top two annotations were donor and coding sequence. The importance of genic features is likely linked to bias in the data, which will be examined further in the **Discussion.**

The HGMD analysis in which only non-exonic HGMD and control variants were considered seemed to overcome this bias towards genes or positions relative to genes. Interestingly, for all algorithms, the top annotation for the Gagliano et al. annotation set was DNase I hypersensitive sites, but we caution against making biological inferences on the top annotations for the reasons outlined above (see **Appendix A**).

5.3.7 Comparison of scores from the three papers: Application to Schizophrenia GWAS

When using the actual prediction scores made available in the three papers, the quantile-quantile plot suggested that the Gagliano et al. scores best identified the novel hits from the second round of the schizophrenia GWAS that were not significant in the first round (**Figure 5.6**). The proportion of hits in the top versus the bottom quartiles of prediction scores are significantly different for the Gagliano et al. method ($p<0.03$, chi-square test), whereas the difference between the quartiles for the Ritchie et al. and Kircher et al. methods were not significant ($p\sim0.4$ for both methods) (**Table 5.9**).

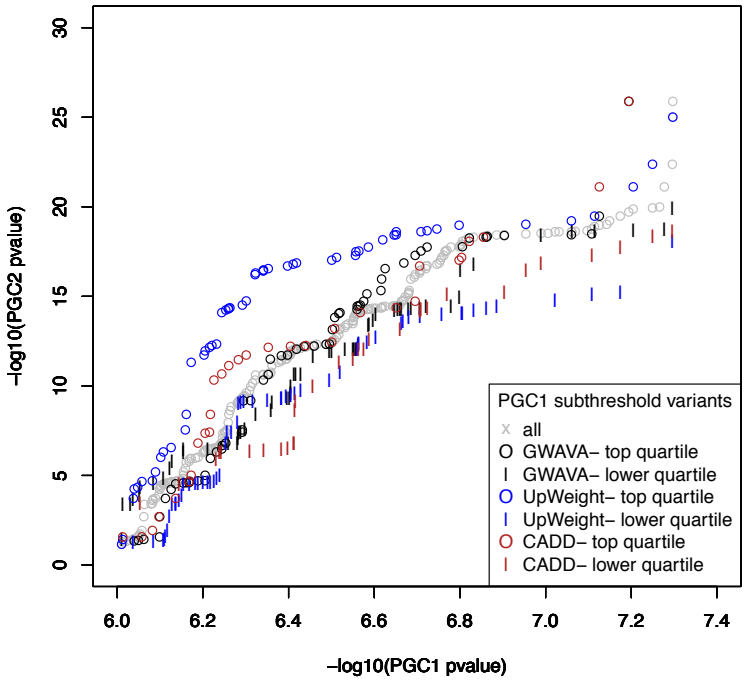


Figure 5.6. Quantile-quantile plots of PGC1 sub-genome-wide-significant variants ($5x10^{-8}<p<1x10^{-6}$) stratified by prediction scores obtained from the three papers, and plotted by -log10(PGC1 p-values) versus -log10(PGC2 p-values)

"GWAVA" corresponds to the scores obtained from the method published in Ritchie et al. 2014, "UpWeight" corresponds to the method in Chapter 3 and "CADD" corresponds to the method in Kircher et al. 2014. The lower quartile genetic variants are those with a prediction score in the first quartile, and the top quartile variants are those with prediction values in the fourth quartile.

Table 5.9. Using the scores from the actual published models, the proportion of sub-genome-wide-significant variants ($5x10^{-8}<p<1x10^{-6}$) variants from the first round of the schizophrenia GWAS (PGC1) that are GWAS significant (p<5e-8) in the second round (PGC2) for the various models

Results are shown for the variants that were assigned the highest scores (top quartile) and the lowest scored variants (lower quartile). The difference row shown corresponds to the proportion of GWAS significant variants in the top quartile minus that of the lower quartile, so a positive difference suggests that the quartile of the most highly scored PGC1 sub-genome-wide significant variants (top quartile) contains more GWAS significant variants from PGC2 compared to the lowest scored PGC1 sub-genome-wide significant variants (lower quartile). The number of variants present in each quartile are in parentheses. Note that quartiles can vary in size where prediction scores are identical across many variants, and all those variants with that particular score were included in the quartile. "UpWeight" corresponds to the method in Chapter 3, "GWAVA" corresponds to the scores obtained from the method published in Ritchie et al. 2014, and "CADD" corresponds to the method in Kircher et al. 2014.

	Method					
	UpWeight		GWAVA		CADD	
		Chi-sq p-val		Chi-sq p-val		Chi-sq p-val
top quartile	80% (55)		67% (60)		74% (31)	
lower quartile	61% (59)	0.03	73% (62)	0.48	65% (31)	0.41
Difference	19%		-6%		9%	

Of the variants in the top quartile for the Gagliano et al. scores, most (80%) were GWAS significant variants ($p<5x10^{-8}$) from the second round of the GWAS. Of the variants in the top quartile for the Ritchie et al. scores and the Kircher et al. scores there were fewer significant variants: 67% and 74% respectively. Only a small percentage of variants in the top quartiles were nonsynonymous SNPs (i.e. missense, nonsense, frameshift, inframe indel, or stop-lost mutations): 9%, 2% and 4% for the Gagliano et al. scores, Ritchie et al. scores and Kircher et al. scores, respectively. Of the sub-genome-wide significant PGC1 SNPs, only 5% are nonsynonymous, and of those, most (83%) become PGC2 hits.

5.4 Discussion

We found that the three algorithms assessed here, elastic net, random forest and the linear support vector machine show comparable accuracy in the GWAS test data. The Kircher et al. annotations trained using the elastic net algorithm have the highest AUC. When applied to real data, several models show the potential to prioritize novel hits, with the exception of the random forest and support vector machine models using the Gagliano et al. annotations. However, this was just one real dataset and further studies would need to be assessed to validate this conclusion. Under the conditions employed in our analysis, none of the models were over-fitted, as demonstrated by verifying that the training set AUC is similar in magnitude to that of the test set.

Furthermore, our results show that various combinations of annotations can create models with similar predictive ability when it comes to identifying risk variants from non-risk variants. One must be wary of making strong conclusions about the relevance of the annotations because of the difficulty in interpretation. The coefficients or variable importance measures are differentially affected by issues such as correlation between the attributes, and whether variables are normalized (for elastic net and support vector machine). This observation makes it difficult to differentiate the predictive power of the functional annotation sets used by each study, at least in the case of GWAS risk variants.

As mentioned in the **Introduction**, the main goals of these methods are to identify those variants that are important for disease risk, which can be applied to identifying novel loci or for fine-mapping at previously implicated loci. The HGMD is designed to contain disease variants, whereas the GWAS Catalogue contains variants associated with disease, but those variants may only be tagging the "causal" variant. GWAS are undertaken to identify the loci containing the variant and may identify the actual causal variant but will more often identify variant in high linkage disequilibrium with the causal variant. Thus, the primary analyses in this paper (using the GWAS Catalogue) may be considered to be about identifying novel loci rather than fine-mapping, and the HGMD analyses may be

considered to be more about fine-mapping a specific locus. Furthermore, the Gagliano et al. method may be considered to be better suited to identifying novel loci (rather than fine-mapping) because it annotates variants on whether or not the variant itself falls into the base pair range for the functional annotation, but also if that variant has is in linkage disequilibrium ($r^2>0.8$) with a variant that falls into the range. The Ritchie et al. and Kircher et al. methods annotate the variants just based on whether the variant itself falls into the base sequence for the functional annotation, and do not look at their linkage disequilibrium proxies. That being said, we also performed the analyses for the Gagliano et al. annotations only considering whether the variant itself falls into the sequence for the functional annotation as an additional analysis. The resulting models had very similar accuracy to those models created when the linkage disequilibrium proxies were taken into account (data available on request).

To apply the methods in next generation sequencing data and fine-mapping studies we would ideally use risk variants identified from such studies. Unfortunately, there are not a sufficient number available. We used the HGMD to attempt to extrapolate our findings. However, we believe the high accuracies achieved for the all HGMD models (i.e. not the models looking just at non-exonic variants) are driven by the inherent bias of the HGMD data, in that it is largely focused on genes. For the models using only non-exonic HGMD and control variants, the AUCs were considerably lower, with the Kircher et al. and Ritchie et al. annotation sets clearly out-performing the annotations used by Gagliano et al. Yet, this subset of HGMD is a highly derived and filtered set of variants, emphasizing the need for empirical data. The simulation employed by Kircher et al. to consider all variants, in which the functional annotations were used to differentiate between millions of high frequency human-derived alleles from the same number of simulated alleles, (Kircher et al., 2014) showed considerable accuracy; further adaptions to this strategy may prove useful.

Compared to the corresponding elastic net or random forest models, the support vector machine models consistently produced slightly lower AUCs for the GWAS Catalogue and all HGMD analyses. This poorer performance may be attributed to the fact that we implemented the most basic kernel type for the support vector machine, a linear kernel. This kernel was chosen in an effort to be consistent with the type of kernel that was utilized by Kircher et al., and with the advantage that computational time remains comparable with the other algorithms. All of the models run in this paper took under 130 minutes to complete. Note that for the support vector machine, in addition to the linear kernel, we also tried using the radial basis function kernel (the type of kernel one step more complex than linear). We could not achieve convergence using the radial basis function kernel within a reasonable amount of time (i.e. still no convergence after running 48 hours on a high performance computing cluster). However, a linear kernel may not be best to separate the data. Furthermore, as support vector machine does not intrinsically perform feature selection, we selected a subset of features with a non-zero Beta coefficient from the corresponding analysis using the elastic net algorithm. Use of another method of feature selection may have yielded different results. Our results do not necessarily suggest that the elastic net and random forest algorithms out-perform the support vector machine algorithm, since altering either the kernel type or the functional annotations in the support vector machine models may produce results comparable to the other two algorithms.

There are limitations to this comparison. For example, other statistical learning algorithms, such as a deep neural network (Quang et al., 2015), and other annotation sets could be explored. Annotation sets could be phenotype specific, as there is evidence that the level of enrichment of functional information can differ depending on the subset of risk variants selected (Farh et al., 2015). For instance, enrichment of disease-specific variants in the GWAS Catalogue can differ in certain cell types, for example for DNase I hypersensitive sites (Maurano et al., 2012).

Identifying which algorithm and/or annotations identify risk variants with the highest accuracy will help researchers develop a better understanding of the genetic factors involved in complex disease in a cost-effective manner making use of a rich set of publically available functional data. This work helps illuminate the genetic factors involved in disease by making use of existing functional data *in silico*. Increasing knowledge on the etiology of complex disease will allow for earlier or better diagnoses, and the development of personalized treatment and novel therapies.

5.5 Methods

We explored the utility of each of the three algorithms with each of the three functional annotation sets in order to attribute performance differences to the algorithm and/or annotations. A total of nine model types were created.

In the primary analysis, the set of risk variants used for training all the models were based on whether or not a genetic variant is a hit or a non-hit from a genome-wide association study (GWAS). Hits were defined as those variants present in the NHGRI GWAS Catalogue (www.genome.gov/gwastudies, downloaded on August 7, 2014) (Hindorff et al., 2010) with a p-value of equal to or less than $5x10^{-8}$. There were 3,618 unique genetic variants that met these criteria. (Note that at the time of download the novel hits from the second phase of the schizophrenia GWAS from the Psychiatric Genomics Consortium (PGC2) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) had not yet been included.) A subset of non-hits was selected from common GWAS arrays (Affymetrix Genome-Wide Human SNP Array 6.0, the Illumina Human1M-Duo Genotyping BeadChip, and the Illumina HumanOmni1-Quad BeadChip). Those non-hits in high linkage disequilibrium ($r^2 > 0.8$) with hits were removed from the analyses, and a random subset of these non-hits was utilized as controls (n= 75,319).

5.5.1 Functional annotation sets

The data was then annotated using three distinct protocols outlined in each of the three respective papers. The variants were marked with the Gagliano et al. annotations available on the website (http://www.camh.ca/en/research/research_areas/genetics_and_epigenetics/Pages/Statistical-Genetics.aspx). Fourteen functional annotations were used by Gagliano et al., two of which were on a continuous scale (two conservation measures, PhyloP and PhastCons), and the remaining were binary, signifying the presence or absence. The binary annotations included those related to genomic context such as the presence in a gene, a splice site or a transcription start site, as well as those from the ENCODE Project (The ENCODE Project Consortium, 2011) such as three types of histone modifications and DNase I hypersensitivity. For the ENCODE data, functional annotations present in multiple cell lines were grouped together, and genetic variants were annotated accordingly in a binary, present or absent, fashion. Variants were marked with an annotation if they or their linkage disequilibrium proxies fall into the base pair range of the annotation.

To annotate the variants using Ritchie et al.'s annotations, the data were entered into the online GWAVA webserver (https://www.sanger.ac.uk/resources/software/gwava/). Ritchie et al. investigated 174 functional annotations, some binary and others continuous. They also used ENCODE Project tracks including those investigated in Gagliano et al. but not necessarily coded as presence or absence. For instance, for transcription factor binding sites, the number of cell types in which the site was present was used as the annotation. Additionally, variation such as mean heterozygosity and genic and sequence contexts were included. Variants were marked with an annotation if they fall into the base pair range of the annotation.

To obtain Kircher et al.'s annotations, the data were entered into the online CADD webserver (http://cadd.gs.washington.edu). However, Kircher et al. also imputed missing

values, expanded categorical variables, added indicator variables, and included interaction terms. Martin Kircher provided scripts to run on the webserver output to prepare our dataset in accordance with the complete protocol. Kircher et al. looked at 63 unique functional annotations, which totaled to 949 once the categorical variables were expanded, and the indicator variables and interaction terms were included. A mixture of continuous, categorical, and binary functional annotations was included. Similar annotations to those used by Gagliano et al. and/or Ritchie et al. were included, such as ENCODE Project annotations and genic context. Additionally, data from online variant prediction programs (e.g. Sift (Ng and Henikoff, 2003) and PolyPhen (Adzhubei et al., 2010) were incorporated. Variants were marked with an annotation if they fall into the base pair range of the annotation.

5.5.2 Statistical learning algorithms

The variants were randomly divided; 60% was used for training the models, and the remaining 40% was reserved for testing. Elastic net is a regularized logistic regression, and those models were constructed using the glmnet package in R (R Core Development Team, 2008). A weighting procedure was included to up-weight hits, as described in Knight et al. (2011); in brief, the weighting has the effect of equalizing the number of hits and non-hits in the training set. Optimal values of the parameters lambda and alpha were selected for each elastic net model using 10-fold cross validation. (The corresponding values that are one standard deviation from the values that produce the lowest binomial deviance.) Lambda is an overall penalty parameter. Alpha controls the proportion of weight assigned to both the sum of the absolute value of the coefficients and the sum of the squared value of the coefficients, which affects the degree of their sparsity. A range of combinations of lambda and alpha were investigated. The lambda and corresponding alpha that give a model a deviance one standard deviation above the model with the lowest deviance was selected.

Random forest is a collection of decision trees. The random forest models were implemented in Python using the scikit-learn package (Pedregosa et al., 2011). Two sets of random forest models were created, both using 10-fold cross validation. For the first set, we replicated Ritchie et al.'s random forest implementation by using scripts (e.g. gwava.py) provided on their online GWAVA FTP site (ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/). For instance, bootstrap sampling was employed to form decision trees from bootstrap subset samples. To address the class imbalance in the datasets, non-hits were down-weighted through the balance_classes function created by Ritchie et al. and included in their random forest implementation. The balance_classes function selects a subset of non-hits that is equal to the number of hits in order to grow a tree. Furthermore, the subset of annotations used to determine the node split was set to the square root of the total number of annotations. This setting is the default setting for classification problems to determine the best split at each node of the decision tree (Malley et al., 2012). Additionally, as done by Ritchie et al., we used 100 decision trees since we determined that the prediction scores and variable importance measures did not significantly differ past 100 trees.

Ritchie et al. used a minimum node size (min_samples_split) of 1. The minimum node size is the minimum number of samples required to split an internal node. We created another set of random forest models in which we adjusted the minimum node size. This parameter is dataset specific, and a recommended setting is 10% of the total dataset (Malley et al., 2012). Consider n to be the number of hits in the training dataset. For the second set of random forest models, we set the minimum node size to approximately 10% of 2n.

Support vector machine creates a hyperplane within a decision boundary space defined by support vectors to separate the classes in multidimensional space. The support vector machine models were implemented in Python through the scikit-learn package (Pedregosa et al., 2011). Kircher et al. did not use a weighting procedure as their training

set was already balanced. To compare protocols in an unbiased manner, we used a subset of the training set in which we chose all hits, and randomly selected an equal amount of non-hits. We performed a grid search using the tune function in order to determine the optimal cost parameter for a linear kernel. The cost parameter is a penalty (see chapter 9 in James et al. (2013) for details). Feature selection is critical to improving model performance and is intrinsically incorporated by the elastic net and random forest algorithms (Appavu et al., 2011). Feature selection must be implemented before using support vector machine, as there is no feature selection protocol built in. Kircher et al. utilized univariate logistic regression among other methods to select features that best predict genetic risk variants. In this paper our support vector machine models included those annotations that had a non-zero Beta coefficient from the corresponding elastic net models. We chose the annotations found to be important from elastic net, since this algorithm implements a more stringent feature selection protocol compared to random forest (see **Results**).

5.5.3 Assessment of model performance

We assessed model performance in the test set data by calculating the area under the receiver operating characteristic (ROC) curve using the R package ROCR (Sing et al., 2005) (and verified using the R package pROC (Robin et al., 2011)). 95% confidence intervals were generated using 2000 bootstrap replicates also using pROC (Robin et al., 2011). As another measure of model performance, we also examined the distribution of prediction scores assigned to the test set data with the aid of violin plots.

We investigated importance of the functional annotations through the Beta coefficient for elastic net. Similar to the output from a simple logistic regression, the larger coefficients are interpreted as more important to predicting genetic risk variants. For random forest we used Gini importance, which was also used in Ritchie et al. Gini importance is a scaled measure of Gini impurity averaged over all trees; it represents the improved capacity for correctly predicting variants that can be directly attributed to the annotation

(Hastie et al., 2009). For support vector machine, feature weights can be obtained related to the construction of the hyperplane when a linear kernel is used (Rosenbaum et al., 2011).

5.5.4 Performance for complex disease variants: Application to Schizophrenia GWAS

We tested the performance of the nine models based on the GWAS classifier in a schizophrenia GWAS context. We selected all sub-genome-wide-significant variants ($5x10^{-8}<p<1x10^{-6}$) from the first round of the GWAS by the Psychiatric Genomics Consortium (PGC1) (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). For each of the nine models we obtained prediction scores for these variants and selected the variants from the first and fourth prediction score quartiles. For these variants we extracted the p-values from the larger second round of the GWAS (PGC2) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) and plotted these in quantile-quantile plots. Note that there is sample overlap in the discovery cohort (about 30%) of the smaller PGC1 in the larger PGC2. Sample details are provided as a Supplementary Table in the PGC2 paper (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014)**.** We were able to determine for all models whether variants assigned higher scores were enriched in the variants with more significant p-values compared to variants with less significant p-values.

5.5.5 HGMD analysis

The nine models created by combinations of annotation sets and algorithms were assessed using two sets of the public release of the Human Gene Mutation Database (HGMD) variants provided to Ensembl in the fourth quarter of 2013 (provided by Graham Ritchie). In the first, we took all the variants (single nucleotide polymorphisms) in HGMD (N= 3,391) and chose non-hits/controls (n= 24,408) that fell within a kilobase of either side from the HGMD variant (for consistency with the way the controls were selected in Ritchie et al. (2014)). Secondly, models based on the subset of non-exonic

HGMD variants (N= 689) and non-exonic control variants present in the 1000 Genomes Project (Phase 1, version 3) that are within +/-1 kilobase from any of the HGMD variants (n= 16,527). were assessed. Additionally, the data was randomly split into 60% for training and 40% for testing. The same procedures for elastic net, random forest and support vector machine used in the GWAS Catalogue analysis were also conducted for the HGMD analyses.

5.5.6 Comparison of scores from the three papers: Application to Schizophrenia GWAS

In the effort for a more general comparison of the published methods as is, rather than looking specifically at the algorithm and annotations as done above, we additionally conducted the schizophrenia GWAS application using scores for the variants obtained directly from the published papers. Gagliano et al. makes available prediction scores from the non-phenotype specific analysis (which defined risk variants as variants present in the NHGRI GWAS Catalogue (Hindorff et al., 2010) downloaded on August 6, 2013 with a p-value of less than or equal to $5x10^{-8}$, and controls as variants on common GWAS platforms that are not in linkage disequilibrium ($r^2 >= 0.8$) with the GWAS Catalogue variants). Ritchie et al. makes available prediction scores from three models. We used the most stringent, the scores from the "region" model (which defined risk variants as "regulatory mutations" in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009) public database, and the control variants as all those variants in the 1000 Genomes Project within a kilobase distance from each HGMD variant. Regulatory mutations are those variants that fall into regions that do not encode for a protein. For both Gagliano et al. and Ritchie et al. the prediction scores range from 0 to 1, where a value closer to one assigned to a variant suggests that that variant is more likely to be a risk variant as defined in the models. Kircher et al. defined phred-like scores (scaled C scores) in addition to raw scores. We plotted based on the raw scores.

Chapter 6
Allele-specific DNA Methylation: A Functional Annotation with Potential for Risk Variant Prioritization in GWAS

6

6.1 Abstract

It has been hypothesized that allele-specific DNA methylation (ASM) can supplement GWAS of complex diseases and traits. We provide the first confirmation of this hypothesis by showing that single nucleotide polymorphisms exhibiting significant methylation intensity differences between the two alleles (ASM-SNPs) in the brain were consistently enriched in the GWAS sub-genome-wide significant SNPs of several phenotypes, with the strongest effect in schizophrenia. Our data also indicate that ASM-SNPs are over-represented in functional genomic regions, and that the association between ASM and disease could be causal.

6.2 Introduction

Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) associated with psychiatric disease, but more associated SNPs remain to be discovered. SNPs from GWAS with nominal but sub-genome-wide significant p-values account for a considerable proportion of the variance in independent psychiatric samples (International Schizophrenia Consortium et al., 2009), suggesting they are enriched for causal SNPs. Obtaining larger sample sizes (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) or using sub-phenotypes (Mahon et al., 2011) has been used to discover additional risk SNPs for psychiatric diseases. Other options for identification of novel risk loci should be explored.

DNA methylation may play a role in disease. For instance, work has been done on investigating the implications of methylation patterns resulting in imprinting or parent-of-origin bias of alleles, as reviewed in Falls et al. (1999) and Butler et al. (2009). Another type of methylation phenomenon is that some SNPs exhibit allele-specific methylation

(ASM): where one allele shows significantly different methylation levels compared to another allele. ASM can be determined by detecting methylation at SNPs in individuals and then comparing the methylation levels between alleles at each SNP in the sample. An initial ASM study used Affymetrix 250K *StyI* SNP arrays to assess ASM in various human tissues, and they showed that ASM can occur outside of imprinted regions (Kerkel et al., 2008). ASM may play a role in disease etiology through the regulation of gene expression since ASM has been shown to be associated with expression changes in nearby genes (Gertz et al., 2011; Schalkwyk et al., 2010). However, there has been a limited number of studies (all in small sample sizes) (see **Table 6.1**), which investigated ASM (e.g. n=10 (Schalkwyk et al., 2010) and n=42 (Hutchinson et al., 2014)). Larger studies to detect ASM effects are warranted.

Table 6.1. Comparison of allele-specific DNA methylation studies**.**

Study	Sample size	DNA tissue source	ASM lab detection method	ASM statistical detection method
Schalkwyk et al. 2010	10 (5 twin pairs)	Whole blood Buccal (verification)	Affymetrix SNP 6.0 + MSRE (HpaII, HlaI, AciI)	For heterozygotes, relative allelic score difference between genotyping and MSRE-digested arrays
Gertz et al. 2011	8 (6 family members from a 3 generation family and 2 unrelateds)	Leukocytes	RRBS (validated 4 loci through Sanger sequencing)	For heterozygotes, compared methylation status on the variant allele and reference allele for each SNP-CpG pair by performing a Fisher's Exact Test and calculated q-values.
Hutchinson 2014	42 (12 twin pairs and 18 singletons)	Whole blood	Affymetrix SNP 6.0 + MSRE (AciI, BsaH, HhaI, HpaII, HpyCH4IV)	Heterozygous SNPs with the MPRs with values lower than the 2.5 and 97.5 percentiles of the MNR distribution

ASM= allelic-specific methylation; MSRE= methylation-specific restriction enzymes; MPR= MSRE positive region; MNR= MSRE negative region; RRBS= reduced representation bisulphite sequencing

Previous studies investigated ASM only in heterozygous individuals, where the intensity at one allele was compared to the intensity of the other allele after digestion with a cocktail of methylation-specific restriction enzymes to enrich for the hypomethylated fraction on the genotyping array (**Figure 6.1**).

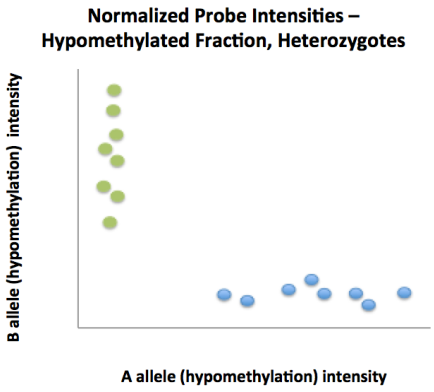


Figure 6.1. Example of ASM detection for heterozygote SNPs after digestion with MSRE.

MSRE= methylation-specific restriction enzymes

With regard to methylation and psychiatric diseases, there is evidence that this epigenetic phenomenon of ASM plays a role in such diseases. For instance, differences in DNA methylation at numerous loci has been shown to be associated with schizophrenia and bipolar disorder in the frontal cortex (Mill et al., 2008).

ASM may help identify the causal SNPs for psychiatric diseases from among other SNPs with sub-genome-wide significant p-values. SNPs exhibiting allele-specific methylation will be referred to as ASM-SNPs from here in. We hypothesized that SNPs from psychiatric GWAS with nominal sub-genome-wide significant p-values are enriched for brain ASM-SNPs compared to SNPs in less significant bins.

6.3 Methods

6.3.1 Samples

Analyses were performed using DNA from human post-mortem prefrontal cortex, Brodmann area 10, were analyzed from control (N=74), bipolar disorder (BPD) (N=65) and schizophrenia (SCZ) (N=64) European-ancestry individuals from the Stanley Medical Research Institute and the Harvard Brain Tissue Resource Center. Sperm samples from BPD (n=24) and control samples (n=24) collected at the Centre for Addiction and Mental Health (Toronto) were also available. Ethnicity of the samples was determined using principal components analysis using super populations from the 1000 Genomes Project (Phase 1). DNA samples from both brain tissues and sperm were extracted using standard phenol-chloroform methods. Demographic data for the samples are summarized in **Table 6.2**.

Table 6.2. Demographics for the samples**.**

	STANLEY (brain)					
	N=90	*Female (N=33, 37%)*	*Male (N=57, 63%)*	*Age (yrs; mean ± SD)*	*Ethnicity-Caucasian*	*Ethnicity-Other*
Controls	27	7	19	42.7±7.3	27	0
SCZ	31	7	23	42.5±8.6	31	0
BPD	32	18	14	45.2±10.3	30	2
	HARVARD (brain)					
	N=118	*Female (N=54, 46%)*	*Male (N=64, 54%)*	*Age (yrs; mean ± SD)*	*Ethnicity-Caucasian*	*Ethnicity-Other*
Controls	49	20	29	57.9±15.9	47	2
SCZ	34	13	21	58.5±13.7	33	1
BPD	35	21	14	62.6±17.4	35	0
	CAMH (sperm)					
	N=48	*Female (N/A)*	*Male (N=48, 100%)*	*Age (yrs; mean ± SD)*	*Ethnicity-Caucasian*	*Ethnicity-Other*
Controls	24	N/A	24	38.5±11.3	16	8
SCZ	0	N/A	N/A	N/A	0	N/A
BPD	24	N/A	24	38.5±12.4	21	3

SCZ= schizophrenia; BPD= bipolar disorder; Age was only provided as decade ranges (*e.g.* 11-20, 21-30, etc.) for the Harvard samples, so to calculate the mean age, the decade was replaced by the median age for that decade. Ethnicity determined by principal component analysis using genetic data. Only the "Caucasian"/European samples (n=203 brains) were utilized for the identification of ASM.

6.3.2 Identification of ASM-SNPs

The samples described above were interrogated twice on Affymetrix SNP 6.0 (Affy6) microarrays: once for genotyping and the other for detecting the methylation levels for the genotypes (**Figure 6.2**). The genotyping was undertaken using standard procedures following the manufacturer's instructions, and possible batch effects were tested for and not found. As cases and controls were run separately on two batches of arrays, a subset of 10 cases and 10 controls was re-run in the second batch to ensure comparability. These

technical replicates were enriched separately versus the original cases and controls, which were enriched together. For the detection of methylation levels, in brief, DNA samples were separately digested with three methylation-specific restriction enzymes: *HpaII*, *HinP1I*, and *HpyCH4IV*. The three digests per sample were then pooled in equal amounts, and adaptors were ligated onto the ends of DNA fragments. To eliminate the fragments containing methylated cytosines between the restriction enzyme targets, ligation products were additionally digested with *McrBC*. Samples were then PCR-amplified using primers complementary to the adaptor sequences, fragmented, labelled, and hybridized to Affy6 microarrays. The crlmm R package (v1.8.11) was used to background correct, normalize and summarize (via RMA) the SNP probes, and to make genotype calls. Individual genotypes were assigned based on the individual's hybridization score for each allele separately.

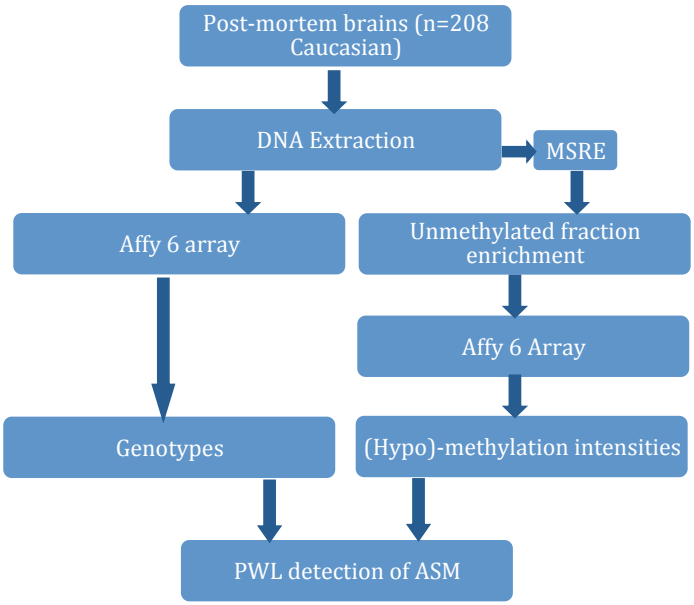
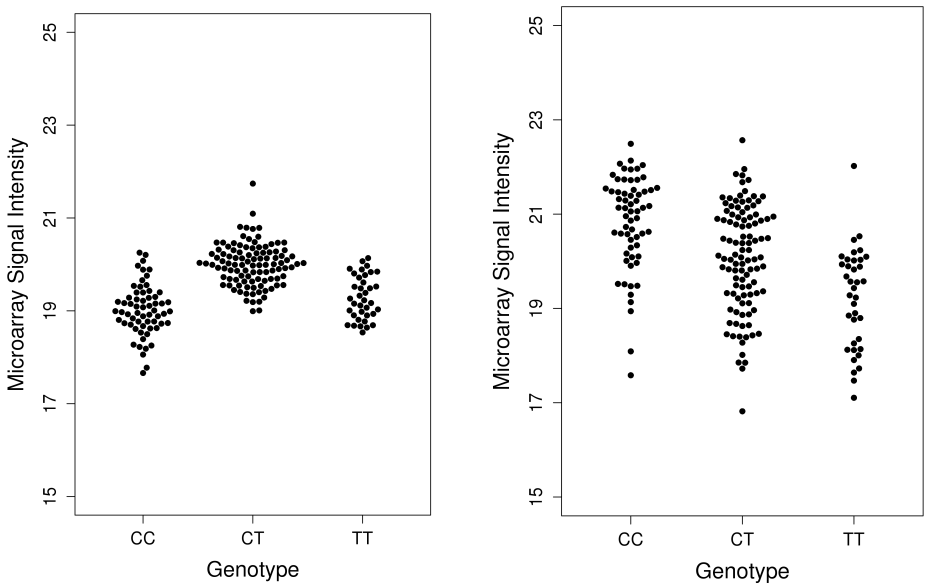


Figure 6.2. Wet lab methodology for ASM detection.

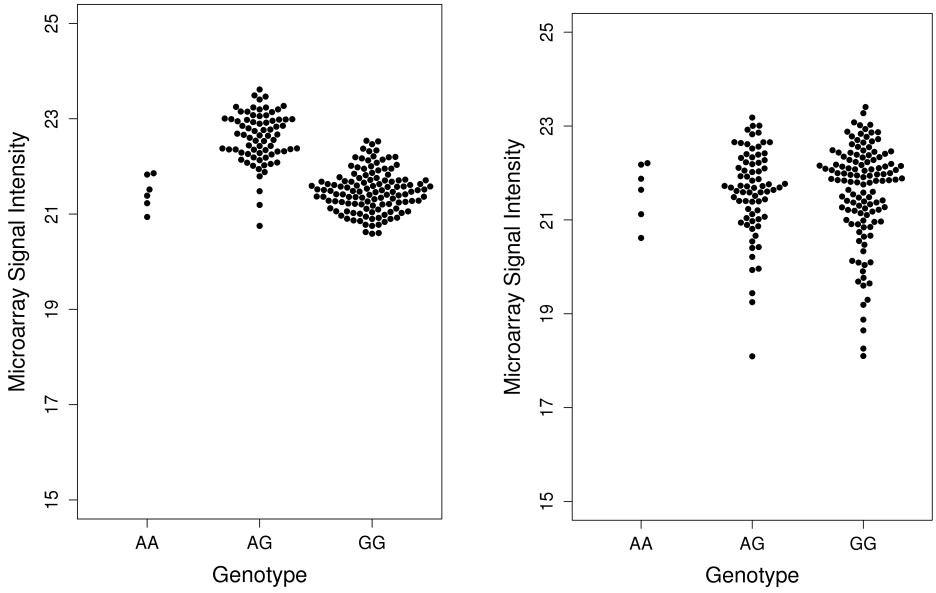
MSRE= methylation-specific restriction enzymes (HpaII, HinP1I, and HpyCH4IV were used here); Affy 6 = Affymetrix SNP 6.0; PWL= piecewise linear regression

ASM-SNPs are detected by establishing whether there is a difference between the total hybridization score (sum of intensities from both alleles) between groups of individuals with different genotypes. Four ASM-SNP lists were derived: all brain, BPD, SCZ, and control using piecewise linear regression (PWL) at $q<0.01$ on the total hybridization score. PWL is a two step linear regression model, first between genotypes AA and AB, and then between genotypes AB and BB. The genotypes were determined from the allelic intensities from the normal genotyping array (i.e. no methylation restriction enzymes added). No covariates were included into the model. For the array to which the methylation specific restriction enzyme digested fragments (i.e. the hypomethylated fraction) were bound, the microarray intensity can be interpreted as hypomethylation level. SNPs that demonstrated one or two significant slopes (the slope between AA-AB and/or AB-BB with a FDR<0.01) were classified as ASM (see **Figure 6.3**). This procedure was done for four ASM cohorts: SCZ, BPD, controls and all brains to get the four ASM-SNP lists.

Genotyping array (no MSRE) Hypomethylated fraction (after MSRE)



a rs9587163, an ASM-SNP b



c rs481818, a non-ASM-SNP d

Figure 6.3. Methylation signal intensity plots from the Affymetric SNP 6.0 array before and after MSRE digestion using all brain samples.

[a] signal intensity for genotyping array for rs9587163, an ASM SNP. [b] signal intensity for the same ASM-SNP as in [a] for the hypomethylated fraction (i.e. MSRE digestion) on which the PWL was conducted to derive the all brain ASM-SNP list in this example. [c] signal intensity for genotyping array for rs481818, a non-ASM SNP. [d] signal intensity for the same non-ASM-SNP as in [c] for the hypomethylated fraction (i.e. MSRE digestion) on which the PWL was conducted to derive the all brain ASM-SNP list in this example. [MSRE= methylation-specific restriction enzymes (HpaII, HinP1I, and HpyCH4IV were used here); PWL= piecewise linear regression]

6.3.3 Quality control

Standard quality control procedures were implemented for SNPs on the genotyping arrays. Hardy-Weinberg equilibrium (HWE) in the control samples was assessed using PLINK (Purcell et al., 2007), and we removed those SNPs with HWE $p < 10^{-10}$. SNPs with low minimum allele frequencies (MAF < 0.05) were also excluded from the analysis.

6.3.4 Analysis of ASM-SNPs in GWAS

We investigated whether ASM-SNPs were enriched in sub-genome-wide significant p-value bins from GWAS. We analyzed brain ASM-SNPs in the context of an SCZ GWAS, which consisted of 34,417 SCZ cases and 45,674 controls and 1,235 parent affected-offspring trios (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). ASM-SNPs were also assessed in publically available summary statistics from 17 large GWAS conducted from 2010 onwards for non-psychiatric diseases and normal traits with a minimum of N>10k cases or N>20k individuals for continuous traits (**Table 6.3**). We began our search for GWAS that meet such criteria starting with the list from the Psychiatric Genomics Consortium (PGC) website. If the same study conducted more than one GWAS on correlated traits, then in order to attempt to make the results more independent, only one GWAS per study (the largest in terms of sample size) was selected (with the exception of the height and body mass index GWAS, which were published in the same study but were deemed as uncorrelated traits so both were assessed).

Table 6.3. Sample information for the schizophrenia GWAS and large non-psychiatric GWAS assessed for enrichment of ASM-SNPs**.**

GWAS	Reference	Sample
Schizophrenia	(Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014 Nature)	49 ancestry matched, non-overlapping case-control samples (46 of European and three of east Asian ancestry, 34,241 cases and 45,604 controls) and 3 family-based samples of European ancestry (1,235 parent affected-offspring trios)
Height	(Yang et al. 2012 Nature Genetics)	~170,000 individuals, European ancestry
BMI	(Yang et al. 2012 Nature Genetics)	~170,000 individuals, European ancestry
Type 2 Diabetes	(Morris et al. 2012 Nature Genetics)	34,840 cases and 114,981 controls, overwhelmingly European ancestry
Age-related macular degeneration	(Fritsche et al. 2013 Nature Genetics)	>70,000 cases >60,000 controls of European or Asian ancestry
College Completion	(Rietveld et al. 2013 Science)	101,069 individuals
Waist to Hip Ratio	(Heid et al. 2010 Nature Genetics)	up to 123,865 individuals, European ancestry
HDL	(Teslovich et al. 2010 Nature)	~88,754 individuals, European ancestry
Coronary Heart disease	(Schunkert et al. 2011 Nature Genetics)	22,233 cases and 64,762 controls, European ancestry
Crohn's disease (all IBD samples)	(Jostins et al. 2012 Nature)	13,510 cases and 20,783 controls, European ancestry
Cigarettes per day	(Tobacco and Genetics Consortium 2010 Nature Genetics)	74,053 individuals, European ancestry
Systolic blood pressure	(Ehret et al. Nature 2011)	69,395 individuals, European ancestry
Platelet count	(Gieger et al. 2011 Nature)	Up to 66,867 individuals, European ancestry

Alzheimer's disease	(Lambert et al. 2013 Nature Genetics)	17,008 Alzheimer's disease cases and 37,154 controls, European ancestry
Hemoglobin level	(van der Harst et al. 2012 Nature)	up to 51,711 individuals of European or South Asian ancestry
Fasting insulin level	(Dupuis et al. 2010 Nature Genetics)	up to 46,186 non-diabetics, European ancestry
Bone mineral density-femoral neck	(Estrada et al. 2012 Nature Genetics)	32,961 individuals, European or East Asian ancestry
2 hour glucose level	(Saxena et al. 2010 Nature Genetics)	15,234 non-diabetic individuals, European ancestry

Enrichment of ASM-SNPs in GWAS p-value bins ($p \leq 0.1$; $0.1 < p \leq 0.2$; $0.2 < p \leq 0.3$; *etc.*) was assessed using the hypergeometric test. For the hypergeometric test, the ASM and non-ASM-SNPs are pooled together. At a particular GWAS p-value bin, the test assesses whether more ASM-SNPs are present in that bin compared to non-ASM-SNPs on the Affymetrix array than what would be expected by chance with sampling from the pool of SNPs without replacement. As a negative control, two independent random SNP lists similar in size to the ASM-SNP lists were compared to the other SNPs on the Affy6 array.

In ASM-SNP analysis of the 17 non-psychiatric GWAS plus SCZ GWAS, 720 tests were performed in total (4 ASM-SNP lists, 10 GWAS p-value bins, and 18 GWAS), and Bonferroni correction for multiple testing was applied accordingly.

Both GWAS and ASM-SNP lists were pruned to ensure our observations were not confounded by correlated SNPs. Pruning was implemented in PLINK (Purcell et al., 2007), and was conducted using the LD structure from the HapMap Project European-ancestry (CEU) samples from the phase containing the most SNPs to ensure maximum

overlap of SNPs (Phase 2, release 23) (Frazer et al., 2007). The filtered SNP set (SNPs that have MAF > 0.01 and genotyping rate greater than 0.95 in the 60 CEU founders) available on the PLINK website (http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml) was utilized. This sample was used for pruning in order to reflect the European-derived ASM-SNP lists. The parameters for pruning were as follows: a 500 kbp window was considered, and the number of SNPs to shift the window at each step was five. For pairs of SNPs with an $r^2 > 0.25$, one SNP was randomly selected for removal. For all of the GWAS, enrichment was assessed for the four ASM-SNP lists derived from: subjects affected with SCZ, subjects affected with BPD, control subjects and all brain samples assessed in the study (All brain).

6.3.5 Ruling out possible confounders

Given the use of restriction enzymes in the ASM detection procedure, we also tested to see if there is over-representation of ASM-SNPs in linkage disequilibrium (LD) with nearby restriction enzyme target overlapping SNPs across various LD thresholds. We investigated LD effects between ASM-SNPs and SNPs that fall within any of the bases of the MSRE sites. LD values were calculated between SNPs and MSRE SNPs in PLINK (Purcell et al., 2007), and r^2 values ranging from 0 to 1 were calculated.

We conducted a few analyses to ensure that the enrichment of ASM-SNPs seen in the $p \leq 0.1$ schizophrenia GWAS bin is not due to confounding factors. ASM SNPs have significantly higher minor allele frequency (MAF) compared to non-ASM SNPs on the Affymetrix array (mean ASM MAF= 0.28; mean non-ASM MAF= 0.24; $p < 2.2 \times 10^{-16}$, Mann-Whitney U test). In order to exclude the possibility that enrichment results are driven due to differing MAF in the ASM-SNP lists compared to non-ASM-SNPs, we created a "MAF-filtered pseudo ASM-SNP list" containing the same number of SNPs in minor allele frequency categories as the ASM-SNPs. We tested for enrichment this pseudo list in the schizophrenia GWAS.

Additionally, we conducted work to demonstrate that the identification of ASM-SNPs is not a hybridization artifact due to differing hybridization of alleles regardless of the methylation status. If there is unequal hybridization at the probes (for example, A alleles give off a greater signal), then there would be a difference between the total hybridization signals of different genotypes even at non-ASM-SNPs, and thus SNPs that exhibit differential hybridization would be detected in this manner. We aimed (1) to establish if SNPs that exhibit differential hybridization exist, and (2) to see if they are enriched in any of the schizophrenia GWAS p-value bins. To answer the first aim, we ran PWL on the raw intensity data. For the second aim, we assessed for enrichment of the resulting pseudo ASM-SNPs in the schizophrenia GWAS p-value bins using the hypergeometric test as previously described for the actual ASM-SNP lists.

6.3.6 Functional genomic characterization of ASM-SNPs

To further elucidate the roles of ASM-SNPs in disease, we explored functional features of the genomic regions in which they are located, using functional genomic data from the Encyclopedia of DNA Elements (ENCODE), for instance. Functional genomic characterization of ASM-SNPs with functional genomic characterization (*e.g.* DNase hypersensitivity, histone modifications, transcription factor binding sites, *etc.*) was performed by comparing frequencies for ASM-SNPs to frequencies of SNPs that did not exhibit ASM, using the hypergeometric test. Splice sites and nonsynonymous SNPs were taken from the UCSC Genome Browser (Meyer et al., 2013). Splice site boundaries were defined as a window of 5 bases up and 5 bases downstream a splice site. Nonsynonymous variants (coding SNPs that fall into one of the following categories: stop-gained/nonsense, missense, stop-lost, frameshift or inframe indel) were defined as a single base pair. Cis eQTLs were defined as single base pairs from the GTEx Project (http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi) (Gibbs JR, 2010; Montgomery SB, 2010; Schadt et al., 2008; Stranger et al., 2007), and from the UK Brain Expression Consortium (www.braineac.org) (Trabzuni et al., 2011). DNase clusters are DNase

hypersensitivity data from all available cell types from the ENCODE Project have been uniformly processed and replicates merged, and peaks are defined by a FDR 1% threshold. UCSC Genes was available from the UCSC Genome Browser (Meyer et al., 2013). Three histone marks (H3K4Me1, H3K4Me3, H3K27Ac) and transcription factor binding sites were based on regions identified by chromatin immunoprecipitation followed by sequencing (ChIP-seq). The peaks data available on UCSC Genome Browser (Meyer et al., 2013) were used: regions of statistically significant signal enrichment where scores associated with each enriched interval is the mean signal value across the interval.

6.4 Results

6.4.1 Samples

Ancestry of the samples was determined by principal components analysis using 1000 Genomes Project super populations as a reference (**Figure 6.4**).

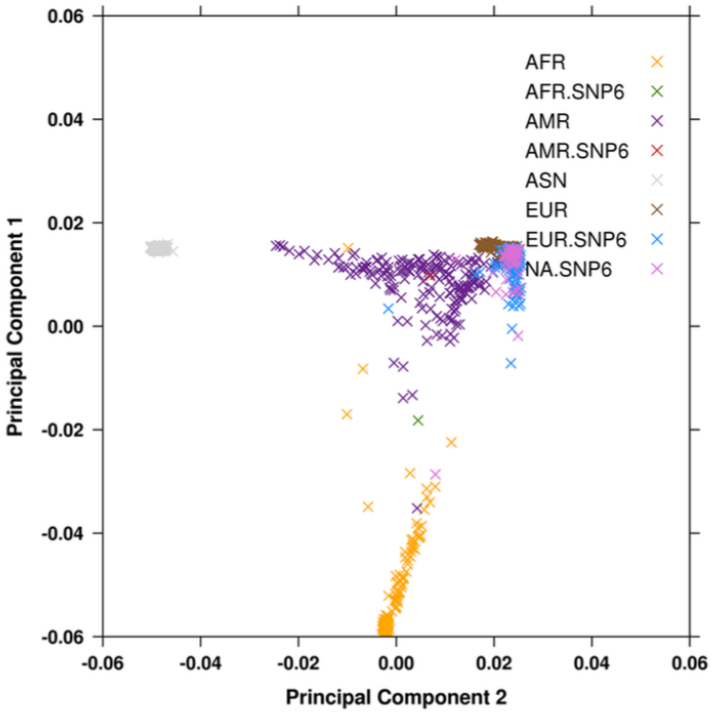


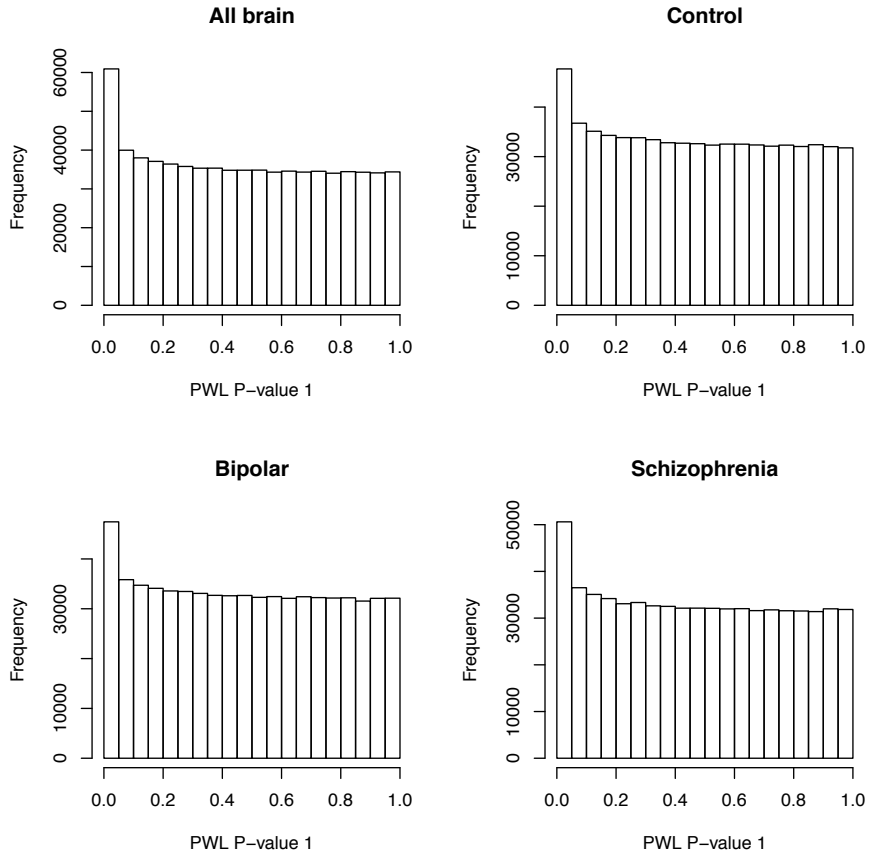
Figure 6.4. Ancestry clusters using principal component analysis.

AFR= 1000 Genomes Project Africans; AFR.SNP6= Samples with self-reported African ancestry; AMR= 1000 Genomes Project Admixed-American; AMR.SNP6= Samples with self-reported Admixed-American ancestry; ASN= 1000 Genomes Project Asians; EUR= 1000 Genomes Project Europeans; EUR.SNP6= Samples with self-reported European ancestry; NA.SNP6= Samples without self-reported ancestry.

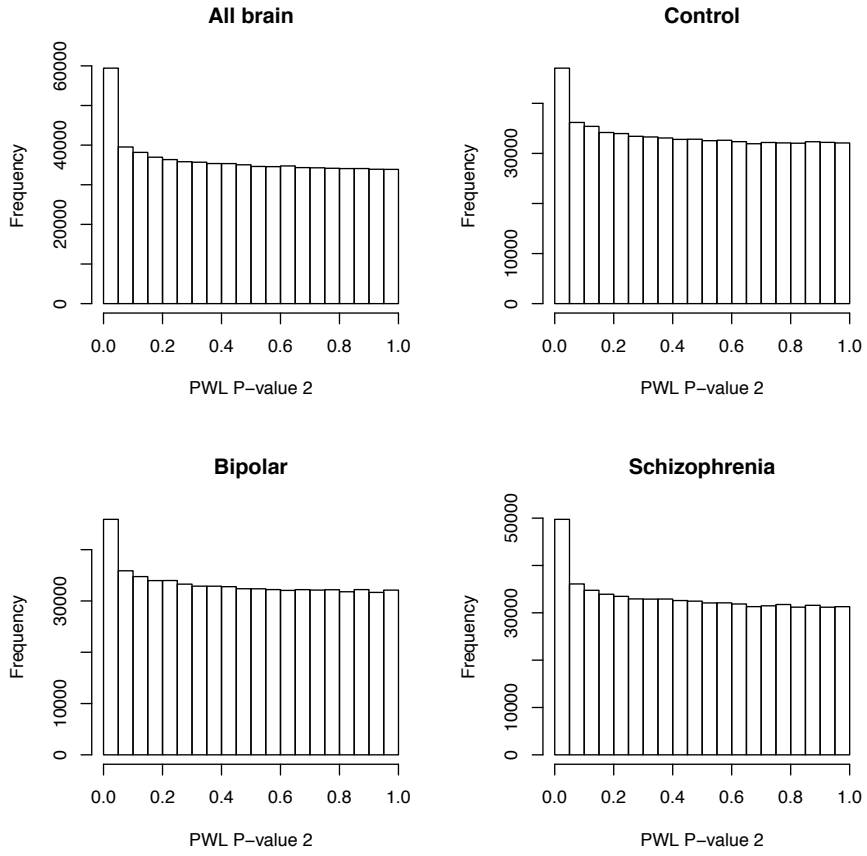
6.4.2 Identification of ASM-SNPs

1,374 ASM-SNPs detected in the control brains (1.31% of all SNPs investigated after removing those in linkage disequilibrium with one another, r^2>0.25); 2,921 in SCZ brains (2.79%); 1,313 in BPD brains (1.25%); and 7,744 in all brain samples (major psychosis cases plus controls; 7.40%). The different sized lists depending on the cohort is likely due

to power differences. The p-values for the two sets of slopes from the piecewise linear regression are shown in **Figure 6.5**.



a



b

Figure 6.5. Distribution of p-values for piecewise linear regression among the cohorts.

SNPs assessed for ASM from the various brain sample cohorts. [**a**] P-values for the first slope (between genotypes AA and AB) [**b**] P-values for the second slope (between genotypes AB and BB)

We also looked at these p-values by constructing Manhattan plots to see the distribution of the SNPs across the genome according to their p-value for the piecewise linear regression (not shown due to large file sizes). There were no particular patterns or preferences for p-value distributions by chromosome.

All pairwise correlations among the four ASM-SNP lists were significantly higher pairwise overlap than expected by chance alone ($p < 2.2 \times 10^{-16}$, hypergeometric test; **Figure 6.6**).

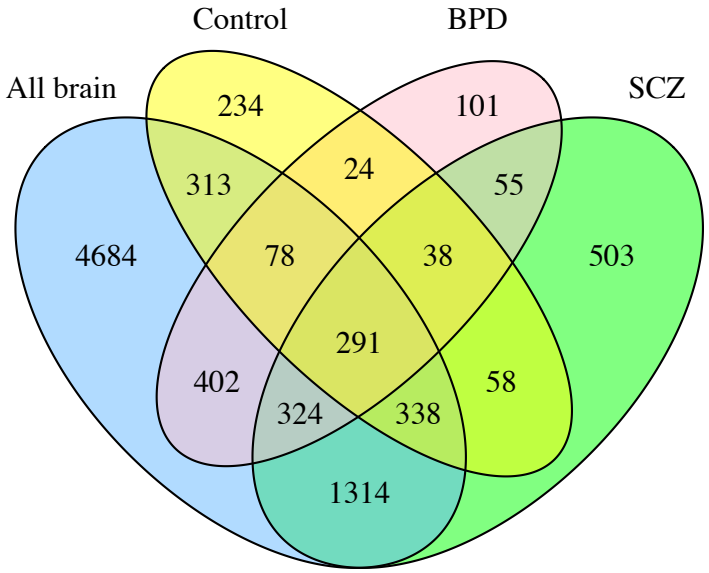


Figure 6.6. Overlap of identified ASM-SNPs among cohorts.

Venn diagram showing overlap of identified LD-pruned ASM-SNPs from the various brain sample cohorts. All brain= ASM-SNPs identified in all the brains; SCZ= ASM-SNPs identified in the brains of schizophrenia patients; control= ASM-SNPs identified in the control brains; BPD= ASM-SNPs identified in the brains of bipolar disorder patients.

6.4.3 Quality control

We generated four ASM-SNP lists using piecewise linear regression. Depending on the cohort being examined, we removed a set of SNPs that failed our quality control tests

described below. Such SNPs were not found on autosomes or sex chromosomes, were not genetically diverse (genetically diverse SNPs defined as SNPs with at least two samples in each of the three genotype categories), diverged from Hardy-Weinberg equilibrium (HWE), exhibited low minor allele frequency (MAF) or had limited genotype confidence call rates. A threshold of $p < 10^{-10}$ was used to filter SNPs that failed HWE (based on the controls), and the vast majority of SNPs were in even stronger agreement with HWE: 97% of SNPs with $p > 10^{-10}$ also exhibited $p > 10^{-7}$. Of the 906,600 SNPs assessed on the Affy6 array, there were 1,140 SNPs that were not found on autosomes or sex chromosomes. The other quality control procedures were implemented for each cohort separately (**Table 6.4**).

Table 6.4. Quality Control filtering of SNPs**.**

Number of SNPs that remain after various quality control procedures before and after piecewise linear regression (PWL). MAF= Minor Allele Frequency; HWE= Hardy-Weinberg Equilibrium; LD= Linkage Disequilibrium

	Control	BPD	SCZ	All brain	Control-sperm	BPD-sperm
Genetically diverse	797,776	795,945	792,343	845,139	710,646	690,085
After PWL (FDR q>1%)	2,546	2,294	4,919	15,514	300	81
MAF and HWE cut-offs	2,025	1,926	4,431	13,795	279	79
LD pruning (r2<0.25)	1,374	1,313	2,921	7,744	222	62

6.4.4 Analysis of ASM-SNPs in GWAS

All four brain ASM-SNP lists showed significant enrichment in the $p \leq 0.1$ schizophrenia GWAS bin, but not in any of the remaining bins ($p > 0.1$) (**Figure 6.7** and **Table 6.5**). The most significant ASM-SNP enrichment was for the all brains ASM-SNP list ($p = 2.0 \times 10^{-19}$). Random SNP lists from the Affymetrix array that passed the quality control procedures showed no effect.

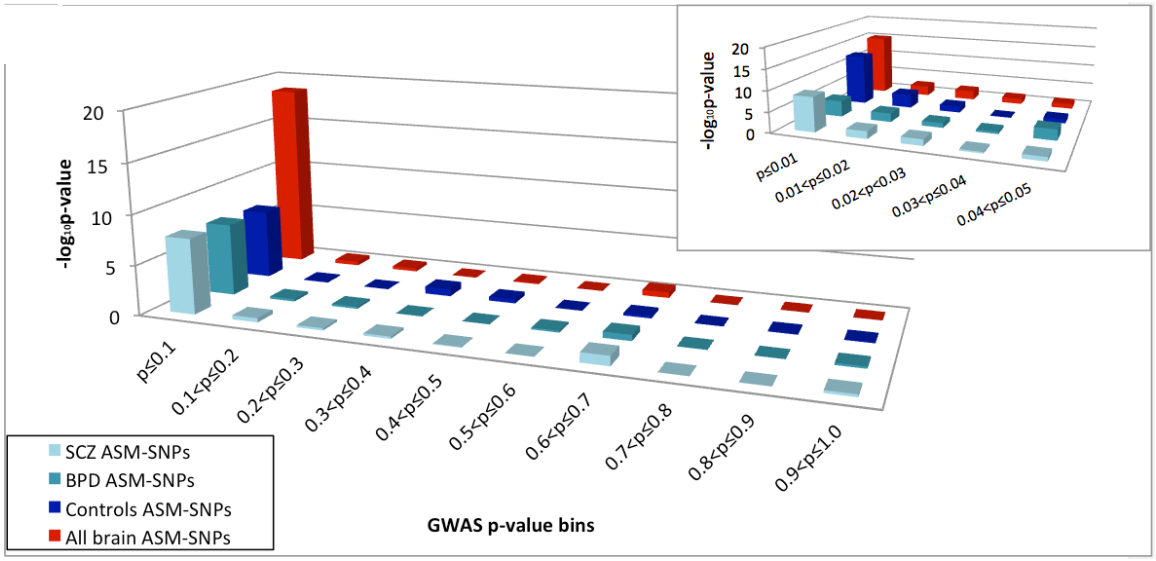


Figure 6.7. Distribution of ASM-SNPs in GWAS p-value bins.

ASM-SNPs detected in the brains of controls, SCZ and BPD patients are overrepresented in the sub-genome-wide significant $p \leq 0.1$ SCZ GWAS SNP group. SCZ GWAS p-value bins are plotted on the x-axis, negative $\log_{10}$ p-values are on the y-axis. The inset shows the further division of the $p \leq 0.1$ bin, revealing the highest density of ASM-SNPs in the SCZ GWAS $p \leq 0.01$ sub-bin.

ASM detection in sperm samples may suggest causal association between ASM-SNPs and psychiatric disease. Although not sufficiently robust to withstand multiple-testing correction, both control-sperm and BPD-sperm ASM-SNPs showed enrichment in the schizophrenia GWAS $p \leq 0.1$ bin (1.38-fold and 2-fold enrichment, respectively), but not in any other bin (**Table 6.4**). There was some overlap between the sperm ASM-SNP lists and the all brain ASM-SNP list. 41 (56%) of the BPD-sperm ASM-SNPs are also all brain ASM-SNPs, and 134 (49%) of the control-sperm ASM-SNPs are also all brain ASM-SNPs.

Table 6.5. Enrichment of ASM-SNPs in Schizophrenia GWAS p-value bins**.**

Hypergeometric p-values (uncorrected for multiple testing) comparing the proportion of ASM-SNPs to all SNPs in GWAS p-value bins. Counts (after overlap with the Affymetrix array SNPs and LD pruning) in parentheses.

ASM-SNP CATEGORY	P- VALUE BINS									
	≤0.1	**>0.1 ≤0.2**	**>0.2 ≤0.3**	**>0.3 ≤0.4**	**>0.4 ≤0.5**	**>0.5 ≤0.6**	**>0.6 ≤0.7**	**>0.7 ≤0.8**	**>0.8 ≤0.9**	**>0.9**
	Schizophrenia GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014 Nature)									
All brains	2.03×10^{-19} (1584)	0.39 (865)	0.47 (771)	0.95 (692)	0.77 (689)	0.98 (647)	0.26 (685)	1.00 (576)	0.98 (620)	0.98 (615)
SCZ brains	2.68×10^{-8} (599)	0.39 (328)	0.59 (286)	0.59 (272)	0.84 (251)	0.97 (234)	0.11 (271)	1.00 (207)	0.97 (225)	0.55 (248)
BPD brains	5.87×10^{-8} (293)	0.57 (143)	0.61 (127)	0.92 (109)	0.93 (104)	0.66 (113)	0.25 (120)	0.88 (103)	0.98 (92)	0.62 (109)
Control brains	1.14×10^{-7} (303)	0.78 (143)	0.95 (118)	0.19 (139)	0.39 (128)	0.95 (106)	0.55 (117)	0.98 (99)	0.77 (111)	0.76 (110)
	Randomly selected SNPs in Schizophrenia GWAS									
Sample 1	0.96 (762)	0.22 (719)	0.10 (716)	0.31 (718)	0.05 (723)	0.77 (682)	0.34 (689)	0.74 (726)	0.83 (683)	0.28 (676)
Sample 2	0.74 (747)	0.59 (740)	0.08 (726)	0.33 (703)	0.21 (701)	0.35 (690)	0.74 (716)	0.59 (688)	0.44 (707)	0.81 (676)
	Schizophrenia GWAS									
BPD sperm	7.7×10^{-4} (20)	0.08 (10)	0.59 (5)	0.85 (3)	0.68 (4)	0.98 (1)	0.91 (2)	0.92 (2)	0.09 (8)	0.16 (7)
Control sperm	6.9×10^{-3} (51)	0.34 (26)	0.79 (18)	0.90 (15)	0.57 (19)	0.71 (17)	0.29 (21)	1.00 (7)	0.05 (26)	0.21 (22)

Enrichment of ASM-SNPs in the more significant p-value bins held when the $p \leq 0.1$ bin was sub-divided into five bins between p-values 0 to 0.05, and the strongest enrichment was observed in the $p \leq 0.01$ bin (**Table 6.6**, and inset of **Figure 6.7**).

Table 6.6. Enrichment of ASM-SNPs in SCZ GWAS p-value bins ($p \leq 0.05$).

Partitioning the $p \leq 0.1$ bin from Table 6.5. Hypergeometric p-values (uncorrected for multiple testing) comparing the proportion of ASM-SNPs to all SNPs in SCZ GWAS; OR- Odds ratios and their corresponding 95% confidence intervals.

ASM-SNP CATEGORY	P-VALUE BINS				
	p≤0.01	**0.01<p≤0.02**	**0.02<p≤0.03**	**0.03<p≤0.04**	**0.04<p≤0.05**
BPD ASM-SNPs					
p	3.84E-09	0.01	0.03	0.44	0.10
OR	1.5	1.0	1.0	1.0	1.0
95% CI	1.3-1.7	0.9-1.2	0.9-1.2	0.8-1.2	0.8-1.2
Controls ASM-SNPs					
p	1.19E-04	0.01	0.07	0.28	2.45E-03
OR	1.2	1.1	0.9	0.8	1.2
95% CI	0.9-1.5	0.7-1.5	0.6-1.3	0.5-1.1	0.8-1.7
SCZ ASM-SNPs					
p	2.85E-13	3.73E-04	0.03	0.92	0.11
OR	1.6	1.0	0.9	0.6	0.8
95% CI	1.2-2.0	0.7-1.4	0.6-1.4	0.4-1.0	0.5-1.3
All brain ASM-SNPs					
p	6.56E-16	3.35E-03	0.01	0.04	0.06
OR	1.6	1.1	0.9	0.5	0.9
95% CI	1.3-1.9	0.9-1.4	0.7-1.2	0.4-0.8	0.6-1.1

In order to more clearly assess the potential of ASM-SNPs to prioritize sub-genome-wide significant GWAS SNPs, we looked at the effect size for schizophrenia GWAS bins ranging all the way from GWAS $p<10^{-7}$ to p=1. There is a clear gradient of ASM enrichment across these bins: the more the significant p-value, the higher the proportion of ASM-SNPs in that bin; for example, schizophrenia ASM-SNPs in the schizophrenia GWAS $p<10^{-7}$ bin exhibits odds ratio of 7.3, while it is only 1.4 for $0.001 \leq p < 0.01$ (**Figure 6.8**). This finding supports the use of ASM to prioritize sub-genome-wide significant GWAS SNPs.

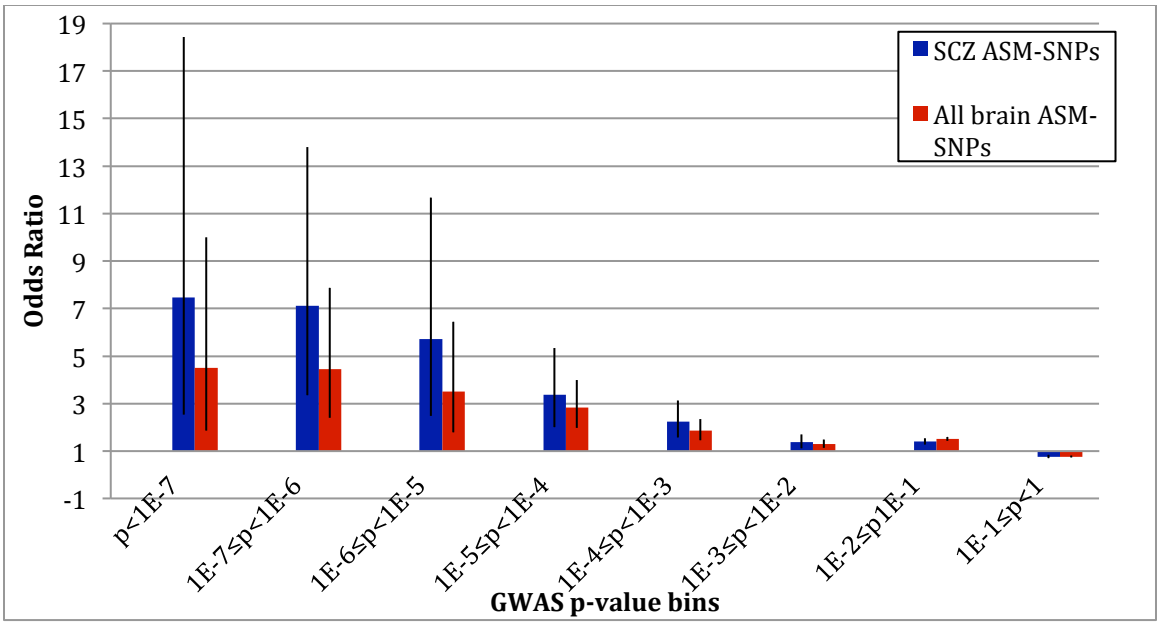


Figure 6.8. Odds ratios (with 95% confidence intervals) for the enrichment of ASM-SNPs in various GWAS p-value bins in the schizophrenia GWAS.

Odds ratios and confidence intervals calculated from a 2x2 contingency table. Blue bars – ASM-SNPs detected in the post-mortem brains from schizophrenia patients; red bars – ASM-SNPs detected in the entire sample of brains (schizophrenia, bipolar disorder, and controls). Control and Bipolar disorder ASM-SNP lists are not shown for clarity due to a small number of SNPs (<10), in the smaller p-value bins, which resulted in very wide confidence intervals.

We then investigated the enrichment of ASM-SNPs in 17 non-psychiatric GWAS (**Figure 6.9**).

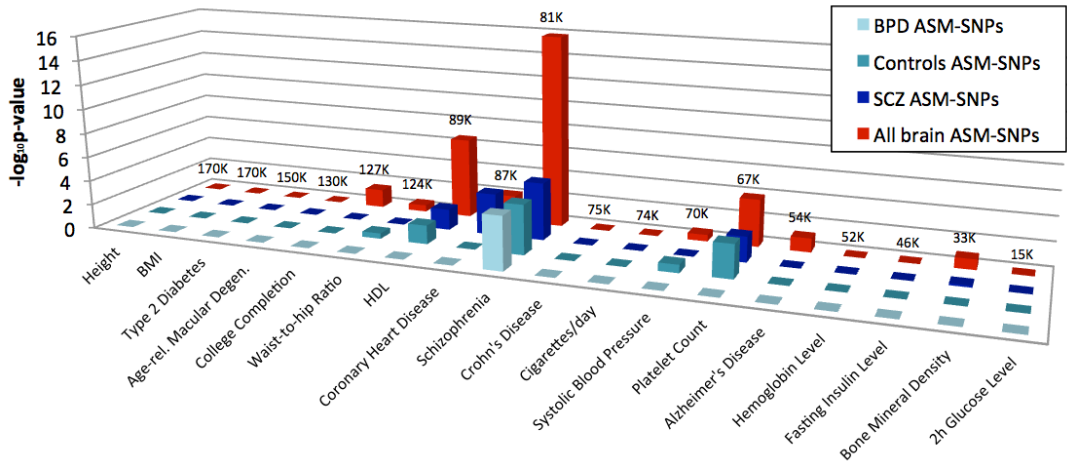


Figure 6.9. Distribution of ASM-SNPs in GWAS p-value bins.

Distribution of -$\log_{10}$ p-values (corrected for multiple testing) for 4 lists of brain ASM-SNPs interrogated in 18 large GWAS. Only GWAS SNP $p \leq 0.1$ bins are presented here. Total sample size of each GWAS in thousands (k) is presented above each row of ASM-SNP p-values.

Enrichment in the GWAS p≤0.1 bin was seen to a lesser degree for some of the four ASM-SNP lists than in three blood/cardiovascular-related GWAS: platelet count, high density lipoprotein (HDL) and coronary heart disease. None of the odds ratios for these cardiovascular-related traits surpassed the odds ratios observed for the enrichment of the corresponding ASM-SNP list in the SCZ GWAS. Significant enrichment was seen neither in any other GWAS investigated nor in any other p-value bin (**Table 6.7**).

Table 6.7. Enrichment of ASM-SNPS in GWAS p-value bins ≤0.1 of large GWAS**.**

Hypergeometric p-values (uncorrected for multiple testing) comparing the proportion of ASM-SNPs to all SNPs in GWAS p-value bins. OR- Odds ratios followed by the corresponding 95% confidence intervals.

ASM SNP brain list	Height	Body Mass Index (BMI)	Type 2 Diabetes	Age-related macular degeneration	College Completion	Waist-to-hip ratio	High Density Lipoprotein (HDL)	Coronary heart disease	Schizophrenia
All brains									
p	1	1	1	1	0.03	0.29	2.4e-7	7.2e-3	1.5e-16
OR	1.0	0.9	1.0	1.1	1.2	1.1	1.3	1.2	1.3
95% CI	0.9-1.0	0.9-1.0	1.0-1.1	1.0-1.2	1.1-1.2	1.0- 1.2	1.2-1.4	1.1-1.2	1.2-1.4
SCZ brains									
p	1	1	1	1	1	1	0.02	3.9e-4	1.9e-5
OR	0.9	1.0	1.1	1.2	1.1	1.2	1.3	1.3	1.3
95% CI	0.8-1.1	0.9-1.1	1.0-1.2	1.0-1.3	1.0-1.3	1.0-1.3	1.1-1.4	1.2-1.5	1.2-1.4
BPD brains									
p	1	1	1	1	1	1	1	1	4.2e-5
OR	1.0	1.0	0.9	1.2	1.1	1.1	1.2	1.3	1.4
95% CI	0.8-1.2	0.8-1.2	0.8-1.1	1.0-1.4	0.9-1.3	0.9-1.3	1.0-1.4	1.1-1.5	1.2-1.6
Control brains									
p	1	1	1	1	1	0.36	0.03	1	8.2e-5
OR	1.0	0.9	1.1	0.9	1.2	1.3	1.4	1.2	1.4
95% CI	0.8-1.2	0.7-1.1	0.9-1.2	0.7-1.1	1.0-1.4	1.1-1.6	1.2-1.6	1.1-1.4	1.2-1.6

Table 6.7 Enrichment of ASM-SNPs in GWAS p-value bins ≤0.1 of large GWAS (**continued**)

Hypergeometric p-values (uncorrected for multiple testing) comparing the proportion of ASM-SNPs to all SNPs in GWAS p-value bins. OR- Odds ratios followed by the corresponding 95% confidence intervals.

ASM SNP brain list	Crohn's disease	Cigarettes /day	Systolic Blood Pressure	Platelet count	Alzheimer's disease	Hemoglobin level	Fasting insulin	Bone mineral density-Femoral neck	2h glucose level
All brains									
p	0.07	1	0.29	1.3e-4	0.07	1	1	0.14	1
OR	0.6	1.1	1.1	1.3	1.1	1.1	1.1	1.1	1.0
95% CI	0.5-0.6	1.0-1.2	1.1-1.2	1.2-1.4	1.1-1.2	1.0-1.2	1.0-1.2	1.1-1.2	0.9-1.0
SCZ brains									
p	1	1	1	8.4e-3	1	1	1	0.72	1
OR	1.1	1.1	1.0	1.3	1.1	1.0	1.1	1.2	1.0
95% CI	1.0-1.3	1.0- 1.2	0.9-1.2	1.1-1.4	1.0-1.2	0.9-1.2	0.9-1.2	1.1-1.4	0.8-1.1
BPD brains									
p	1	1	1	1	1	1	1	1	1
OR	1.1	1.3	1.1	1.2	1.2	1.2	1.0	1.2	1.0
95% CI	1.0-1.4	1.0-1.4	0.9-1.3	1.0-1.4	1.0-1.4	1.0-1.4	0.8-1.2	1.0-1.4	0.8-1.2
Control brains									
p	1	1	0.22	1.8e-3	1	1	1	1	1
OR	1.2	1.2	1.3	1.4	1.2	1.2	1.0	1.1	0.9
95% CI	1.0-1.4	1.0-1.4	1.1-1.6	1.2-1.6	1.0-1.4	1.1-1.4	0.8-1.2	0.9-1.3	0.8-1.1

To further demonstrate that ASM-SNP analysis can identify those sub-genome-wide significant GWAS SNPs most likely to be disease-associated, we analyzed a 52k-individual SCZ GWAS (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011), which was a subset of the 81k-individual SCZ GWAS. We categorized sub-genome-wide significant GWAS SNPs in the 52k-individual study ($5 \times 10^{-8} < p < 0.1$) as either ASM-SNPs or non-ASM-SNPs. For these SNPs we created a quantile-quantile plot of the p-values in the 81k-individual SCZ GWAS (observed p-values *vs.* expected p-values; **Figure 6.10**).

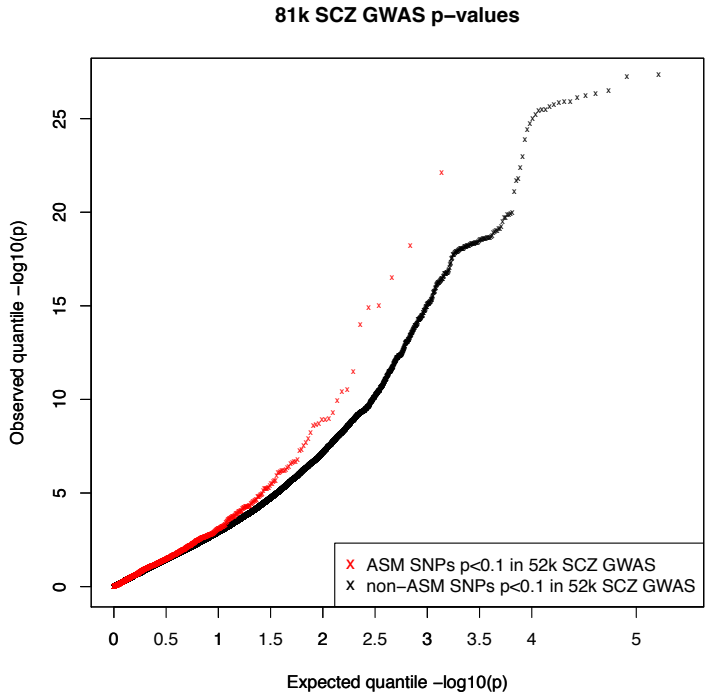


Figure 6.10. The quantile-quantile plot shows ASM-SNPs and non-ASM-SNPs with a $p \leq 0.1$ in the 52k SCZ GWAS plotted by their p-value in the 81k GWAS.

The observed quantiles were derived from the 81k SCZ GWAS p-values for the respective SNPs, while the expected quantiles were from a continuous uniform distribution of p-values. The steeper slope of the ASM-SNPs indicates that these SNPs have lower p-values in the 81k SCZ GWAS, where both the sample size and power is greater, compared to the non-ASM-SNPs. The plotted ASM-SNPs are those from all brains in the $p \leq 0.1$ bin of the 52k SCZ GWAS (n = 1,376) and the plotted non-ASM-SNPs are those in the 52k SCZ GWAS $p \leq 0.1$ bin (n = 163,592 from the total of n = 1,252,902 SNPs tested in 52K SCZ GWAS).

6.4.5 Ruling out possible confounders

We found no significant over-representation of ASM-SNPs in LD with SNPs in nearby restriction enzyme sequences across all LD threshold values ($p > 0.1$, hypergeometric test).

We ensured that the enrichment of ASM-SNPs seen in the $p \leq 0.1$ schizophrenia GWAS bin is not due to differing minor allele frequencies (MAF) between the ASM and non-

ASM-SNPs or due to differing hybridization of alleles regardless of methylation status. We found that for the "MAF-filtered pseudo ASM-SNP list" with the same allele frequency distribution as the ASM-SNPs, there was no enrichment (uncorrected $p>0.039$, hypergeometric test) for any of the schizophrenia GWAS p-value bins, suggesting that the ASM-SNP enrichment seen in the $p\leq0.1$ schizophrenia GWAS bin is not due to MAF differences between ASM and non-ASM-SNPs.

With regard to the hybridization, we detected "differential hybridization SNPs" by running PWL on the genotyping intensity data (the array for the normal genotyping without the use of the methylation specific restriction enzymes). We found no correlation between the p-values for the first slope (AA and AB) with that of second slope (AB and BB) (correlation= 0.028). We also found that the p-values obtained to detect differential hybridization from the normal genotyping array were not correlated with the p-values obtained from the hypomethylated fraction from which the ASM effects were detected (correlation= 0.016 for the first slope for the two arrays; correlation= 0.019 for the second slope for the two arrays). We defined those SNPs with a $q<0.01$ as SNPs that exhibit differential hybridization. We tested these SNPs in the context of the schizophrenia GWAS. Most SNPs (n= 104,688) were classified as differential hybridization SNPs by this method, but these SNPs are not significantly enriched in any of the schizophrenia GWAS p-value bins (uncorrected $p> 0.0002$, hypergeometric test) (**Figure 6.11**). Although unequal hybridization of alleles is evident and creates pseudo ASM-SNPs, these are not enriched in GWAS bins of interest, and therefore the enrichment results are likely due to a true ASM effect.

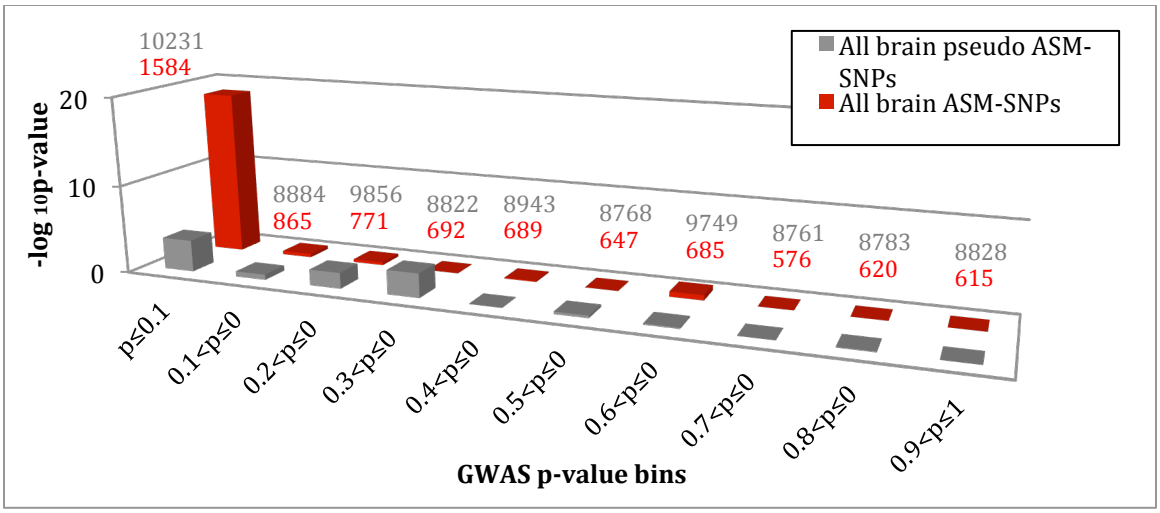


Figure 6.11. Distribution of SNPs that exhibit differential hybridization and ASM-SNPs in SCZ GWAS p-value bins.

ASM-SNPs detected in the all brains cohort are overrepresented in the sub-genome-wide significant $p \leq 0.1$ SCZ GWAS SNP group compared to SNPs that exhibit differential hybridization detected in the same cohort. SCZ GWAS p-value bins are plotted on the x-axis, negative $\log_{10}$ p-values are on the y-axis. The numbers on top of the bars give the number of SNPs in each of the two lists.

6.4.6 Functional genomic characterization of ASM-SNPs

13% (1,036 of the 7,743) of the all brains ASM-SNP list are in CpG islands, and 7.6% (586 out of the 7,743) are in coding regions. None of the non-ASM-SNPs (subset selected with the same MAF distribution as the all brain ASM-SNP list) fall into CpG islands, and 4.1% are in coding regions.

ASM-SNPs in the schizophrenia GWAS p≤0.1 bin showed significant enrichment in functional genomic categories (for example, transcription factor binding sites, DNase I hypersensitive sites, regulatory histone modifications) compared to all GWAS SNPs p>0.1 that did not exhibit ASM effects (**Table 6.8**).

Table 6.8. ASM-SNPs in the SCZ GWAS p ≤ 0.1 bin are found in functional regions of the genome more than expected by chance alone (uncorrected hypergeometric test p-values).

There were 2,351 ASM-SNPs (the union of the four brain ASM-SNP lists after pruning based on linkage disequilibrium) and 122,186 non-ASM-SNPs. Frequencies of ASM-SNPs with GWAS p > 0.1 (n = 8,829) and non-ASM-SNPs with GWAS p>0.1 (n = 511,636) shown for comparison purposes. All SNPs are annotated in a binary fashion indicating the presence or absence of a functional characteristic for the SNP itself. OR= odds ratio for the 2x2 contingency table; and 95% CI is the corresponding 95% confidence interval.

	SNPs with GWAS P ≤ 0.1			**SNPs with GWAS P > 0.1**		
Functional Characteristic	Proportion		P (OR; 95% CI)	Proportion		P (OR; 95% CI)
	ASM-SNPs	non-ASM-SNPs		ASM-SNPs	non-ASM-SNPs	
splice	0.0021	0.0009	0.0249 (2.4; 0.8, 5.2)	0.0015	0.0009	0.0361 (6.1; 3.2, 10)
non-synonymous	0.0030	0.0039	0.7032 (0.8; 0.3,1.5)	0.0044	0.0037	0.1063 (4.2; 3.0, 5.8)
DNase Clusters	0.4122	0.1483	2.02E-205 (4.0; 3.7, 4.4)	0.3906	0.1479	< 1E-205 (14; 13, 15)
GTEx eQTLs (all 7 experiments together)	0.0285	0.0109	4.51E-12 (2.6; 2.0, 3.4)	0.0134	0.0067	9.64E-12 (4.7; 3.8, 5.6)
UK brain eQTLs	0.1438	0.0885	3.80E-10 (1.5; 1.3, 1.6)	0.0832	0.0646	7.67E-12 (3.8; 3.6, 4.2)
UCSC Genes	0.5070	0.4207	3.06E-17 (1.4; 1.3, 1.5)	0.4646	0.3821	1.65E-55 (4.9; 4.6, 5.2)
BroadHistone-H3k4Me1	0.6508	0.4392	3.00E-93 (1.3; 1.2, 1.4)	0.6026	0.4272	2.12E-236 (4.6; 4.3, 5.0)
BroadHistone-H3k4Me3	0.4785	0.2419	7.22E-134 (1.7; 1.5, 1.8)	0.4493	0.4756	<1E-205 (5.7; 5.3, 6.2)
BroadHistone-H3k27ac	0.6159	0.3931	1.27E-103 (1.6; 1.5, 1.7)	0.5537	0.4272	2.48E-239 (5.4; 5.0, 5.9)
Txn Factor ChIP (if annotation for any TF)	0.6159	0.0821	4.54E-109 (1.5; 1.4, 1.6)	0.2235	0.0815	<1E-205 (5.1; 4.7, 5.5)

ASM-SNPs are distributed throughout the genome, and only a few are SNPs that are significantly associated with SCZ in the GWAS (**Figure 6.12**).

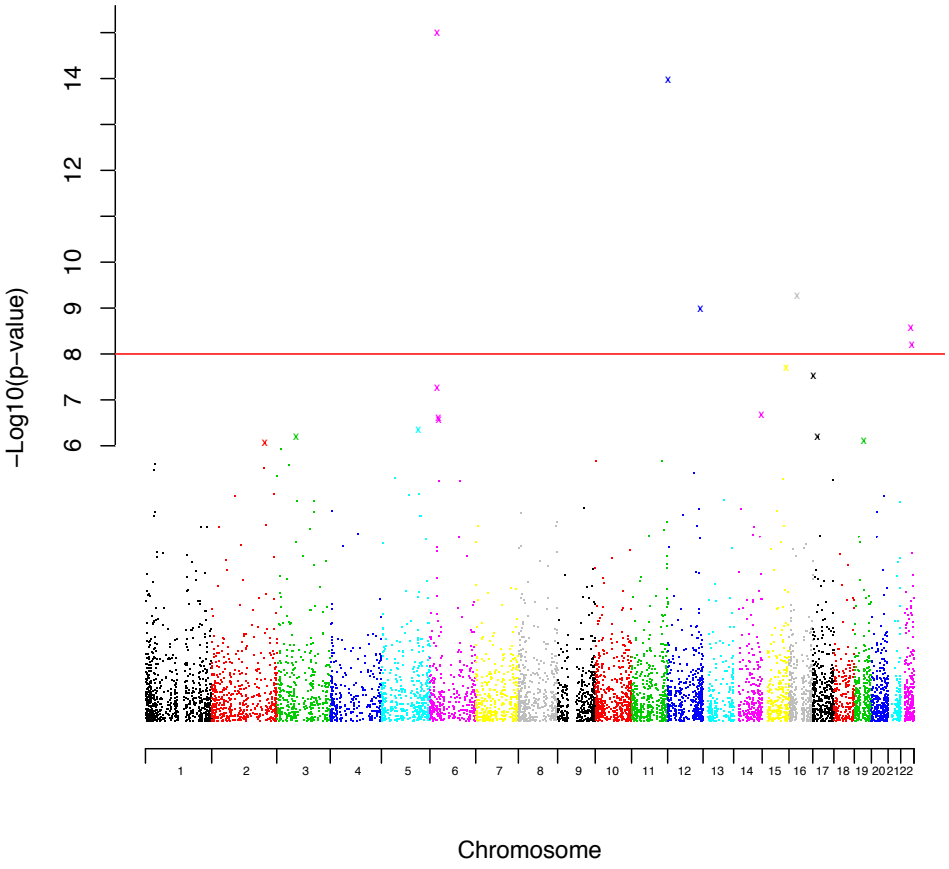


Figure 6.12. Manhattan plot of ASM-SNPs plotted by their SCZ GWAS p-values.

The LD-pruned all brain ASM-SNP list (n=7,744 SNPs is plotted) using data from the second round of the SCZ GWAS.

6.5 Discussion

We demonstrate that ASM in the brain is relevant to psychiatric GWAS by demonstrating that brain ASM-SNPs were consistently enriched in schizophrenia GWAS sub-genome-wide significant SNPs, with a lesser degree of enrichment in the HDL, platelet count and, coronary heart disease GWAS. The degree of enrichment seen in the $p \leq 0.1$ bin for these three cardiovascular related GWAS may point to a sharing of genetic factors between psychiatric and cardiovascular disorders. Yet it is difficult to disentangle whether this relationship is primarily environmental or genetic. Furthermore, ASM-SNPs are over-represented in functional genomic regions, and thus ASM may be important in prioritizing which sub-genome-wide significant GWAS SNPs are causal.

Unlike previous ASM studies, in this work we assessed ASM at all SNPs rather than just in heterozygous individuals by considering methylation differences among genotypes rather than between the two alleles of a heterozygous individual. However, similar to previous work we used a cocktail of methylation-specific restriction enzymes (MSRE) to enrich for the hypomethylated fraction and assess this fraction on an Affymetrix SNP 6.0 array taking the allele intensities as a measure of hypomethylation intensity. We compared the all brain ASM-SNP list (before LD pruning) to the ASM-SNP lists in Schalkwyk et al. (2010) and Hutchinson et al. (2014), two papers in which MSRE and was combined with Affymetrix SNP 6.0 arrays to detect ASM. ASM-SNPs were only detected in heterozygous individuals in those two studies. Three ASM-SNPs (rs220030, rs9366927, rs943049) listed in Schalkwyk et al. (2010) in either of Tables 1,2,3 or S3 (n=204) were also identified as an ASM-SNPs in Hutchinson et al. (2014) in Figure 2b (n=30). These two groups (see **Table 6.1**) used a different cocktail of enzymes, but they both used whole blood. Comparing these ASM-SNP lists to the all brain ASM-SNP list described here, 28/204 (14%) (see **Table 6.9**) of the ASM-SNPs detected by Schalkwyk were also detected in our all brain ASM-SNP list, and 2/30 (7%) (rs11761231,

rs4689713) of the ASM-SNPs detected by Hutchinson were also detected in our brain ASM-SNP list.

Table 6.9. ASM-SNPs identified in this study and also in Schalkwyk et al.

Common SNPs between the all brain ASM-SNP list here and ASM-SNPs in either Tables 1, 2, 3 or S3 in Schalkwyk et al. RAS= relative allelic score

SNP	Schalkwyk-average RAS change	All brain ASM-SNPs			
		p-value 1	direction 1	p-value 2	direction 2
rs10234308	0.34	0.003	positive	9.46E-05	negative
rs1043509	0.11	NA	NA	4.20E-05	negative
rs11211481	0.22	9.65E-05	positive	0.004	negative
rs13099918	0.23	0.885	positive	2.60E-06	negative
rs1378942	0.11	3.80E-10	positive	6.66E-07	negative
rs1889364	0.15	4.99E-09	positive	0.098	negative
rs1953211	0.1	0.001	positive	3.35E-10	negative
rs2143346	0.23	0.392	positive	2.35E-05	negative
rs2234211	0.17	1.20E-06	positive	0.766	positive
rs2272554	0.14	6.54E-06	positive	0.011	negative
rs2731826	0.38	1.56E-09	positive	0.002	negative
rs2824493	0.1	0.013	positive	7.87E-05	negative
rs3821023	0.31	1.65E-05	positive	0.187	negative
rs391467	0.21	0.011	negative	9.63E-07	negative
rs4556786	0.18	1.50E-22	positive	0.006	negative
rs4653164	0.11	0.003	positive	1.51E-07	negative
rs4828524	0.1	0.002	positive	4.99E-05	negative
rs4837866	0.13	1.80E-06	positive	7.98E-13	negative
rs553161	0.13	0.143	negative	8.37E-12	negative
rs6441992	0.16	0.567	negative	3.81E-06	negative
rs6760544	0.36	1.75E-05	positive	0.010	negative
rs6864309	0.1	7.74E-05	positive	0.526	positive
rs7146315	0.16	0.458	positive	1.33E-19	negative
rs7209653	0.11	0.000	positive	0.018	negative
rs734380	0.18	2.96E-10	positive	0.012	negative
rs7534271	0.22	0.290	positive	2.47E-10	negative
rs762982	0.14	0.542	negative	9.81E-05	negative
rs822625	0.26	0.017	positive	9.52E-05	negative

A limitation in this study is not taking into account the differential hybridization seen between alleles on the genotyping array in the ASM detection procedure even though these differential hybridization pseudo ASM-SNPs did not exhibit the enrichment in the SCZ GWAS $p \leq 0.1$ bin as seen in the ASM-SNP lists. SNPs that demonstrate differential hybridization from the genotyping array do not exhibit the enrichment in the SCZ sub-genome-wide significant SNPs, as was seen with the ASM-SNPs (those SNPs that show differences in allele intensities on the hypomethylation arrays). To background correct for underlining differential hybridization we could have, for each SNP, divided its hypomethylation intensity by its genotyping array intensity. Furthermore, there are some issues with the Affymetrix array platforms that could lead to incorrect calls. For instance, the genotyping call rate is reduced for SNPs in probes with high GC content (>70%), and variants in probes with low sequence complexity are more likely to be called incorrectly (Kothiyal et al., 2009).

Additionally, we have not investigated other confounding factors that could be interpreted as ASM by our method such as there being nearby SNPs interfering with the methylation specific restriction enzyme sites. One could impute to a reference panel such as the 1000 Genomes Project data or perform whole-genome sequencing to test whether SNPs are interfering with restriction enzyme sites.

Other considerations surround ethnic heterogeneity. ASM may differ between different populations. We had a largely European population, and thus derived ASM-SNP lists from the genetically-determined European samples. Due to a limited number of non-European samples, we were unable to assess ASM in different populations, but comparing ASM in different populations would be interesting to investigate in the future. That being said, although our ASM-SNP lists were derived from European individuals, not all of the GWAS we investigated were composed of solely European subjects (see **Table 6.2**).

Next steps would also be to replicate ASM results using another detection methodology such as bisulphite sequencing as there are limitations with using the Affymetrix arrays to detect ASM. For instance, different types of methylation (e.g. hydroxylmethylation) cannot be differentiated using this methodology.

Chapter 7
Overall Conclusion and Future Directions

7

7.1 Conclusion

This thesis has investigated the potential of using functional genomic annotations in a statistical learning framework in order to identify novel disease-associated loci, and/or to prioritize the actual causal genetic variant at identified loci. I used elastic net, a type of penalized logistic regression. My work was unique because I created a score for each SNP using hundreds more annotations than previous publications in the field, and also created phenotype-specific models (for autoimmune, brain-related, and cardiovascular diseases, and also for cancer) in addition to a general non-phenotype specific model differentiating GWAS Catalogue variants from variants on common genotyping arrays as the classifier (Gagliano et al., 2014a). These models were able to identify genetic risk variants; the models with the highest accuracies were the non-phenotype specific model and the autoimmune model both trained using variants in the GWAS Catalogue below the accepted threshold for genome-wide significance, $p < 5x10^{-8}$.

The timeliness of my prioritization method (Gagliano et al., 2014a) was demonstrated by it being published within weeks of two others (Kircher et al., 2014; Ritchie et al., 2014). These methods all use different functional annotations as predictor variables, a different classification of disease-associated from benign variants, and different statistical learning algorithms. I investigated which combination of predictor variables, classifier and algorithm produced the model with the best predictive accuracy (Gagliano et al., 2015a). I assessed the accuracy of these models through the use of AUCs and violin plots, two measures deemed as informative from my investigation of predictive accuracy measures (Gagliano et al., 2015b). Additionally, I explored which of the published models are best at prioritizing genetic variants by applying the models to a schizophrenia (SCZ) GWAS for which there were two studies conducted by the Psychiatric Genomics Consortium

(Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). I applied the models to the first SCZ GWAS and evaluated which model best prioritized the novel associated variants from the second study. Results suggested that all methods have considerable (and similar) predictive accuracies (AUCs 0.64-0.71) in test set data, but there is more variability in the application to the schizophrenia GWAS. With regard to the functional annotation set, the Kircher et al. or Ritchie et al. annotation sets performed the best in identifying schizophrenia-associated variants. Regardless of annotation set, the elastic net models consistently showed good separation of GWAS significant SNPs from other SNPs. I found that using both the same algorithm and annotation set, but a different database as the classifier (GWAS Catalogue or HGMD) resulted in vastly different models with regard to overall accuracy. Additionally, which annotations were included in the models differed between the two databases, and the models exhibited similar accuracy within a database. Finally, in **Chapter 6** I showed that a new annotation, allele-specific methylation (ASM) is useful for prioritizing GWAS hits. Variants that exhibit ASM (ASM-SNPs) showed enrichment in functional annotations, and also the most significant enrichment in the sub-genome-wide significant SNPs in the largest to date schizophrenia GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) as well as other traits.

With regard to my initial hypotheses on page 50:

1) I developed a method to incorporate multiple functional annotations that is able to predict genetic risk variants for various complex diseases/traits generally and also for phenotype-specific outcomes with some accuracy.

2) I evaluated the performance of different statistical learning algorithms, functional annotation sets and classifiers that exist in the literature. I found that accuracy tends to be similar when the same classifier is used, but the annotations that are identified as most important vary. No one model was found to out-perform the others.

3) I assessed the functional enrichment and enrichment in GWAS sub-genome-wide significant variants of a novel annotation based on allele-specific methylation (ASM). The results suggest that ASM is a relevant annotation to include for genetic variant prioritization.

Broadly speaking, the use of statistical learning to prioritize genetic risk variants is very timely and relevant in the age where genome-wide genetic information and a vast amount of functional genomic information are available. This work has potential for improved understanding of common health conditions; identifying novel risk variants by the use of computers is cost effective and may ultimately result in the development of better treatment options for people who suffer from a variety of devastating diseases around the world.

More specifically, *in silico* prioritization of variants has several applications in genetic association analysis pipelines, and can be used for several purposes in the context of association studies. For instance, at the completion of a GWAS, the top findings can be prioritized to determine which will be either subjected to functional studies for further follow-up or assessed in an independent replication sample. In this way, the prioritization can be useful for fine-mapping associated loci. Furthermore, it may be useful to use this methodology to select likely functional SNPs for a custom array. Prioritization may also be used in the middle of a two-stage GWAS, where a proportion of the individuals in the study are genotyped on all available variants in the first stage, and a proportion of these variants are genotyped on the remaining samples in the second stage (Skol et al., 2007). Rather than selecting variants for the second stage based solely on their association p-value from the first stage, their prediction score (based on functional genomic information) can also be used to help select those variants that move on to the next stage. Additionally, prioritization could allow for more informative pathway analyses.

7.2 Limitations

There are limitations to the work described in this thesis. First of all, risk variants are difficult to define. This challenge is clear from the inherent differences between variants in databases such as in the GWAS Catalogue and in HGMD (as discussed in **Chapter 1**) that are being used to create models for prioritization of risk variants. One notable difference is that variants in the GWAS Catalogue have higher minor allele frequencies compared to variants in HGMD. The two databases mostly contain different types of variation, and so it unclear whether a model trained using GWAS Catalogue variants as the classifier will effectively prioritize low frequency predominantly coding variants such as those in HGMD.

The differences between variants in different databases are further supported by my work in the methods comparison chapter (**Chapter 5**). I explored the differences between the GWAS Catalogue and HGMD variants further by testing to see if I could use statistical learning algorithms to predict variants from one database from the other. For the annotation set, I used the 14 discussed in Chapter 3. The models created had an AUC of 82% in the independent test set for random forest, and 80% for elastic net. The accuracy could be attributed to the underlying frequency differences between the GWAS Catalogue and HGMD variants. Those frequency differences for the Gagliano et al. annotations are shown in **Figure 7.1**.

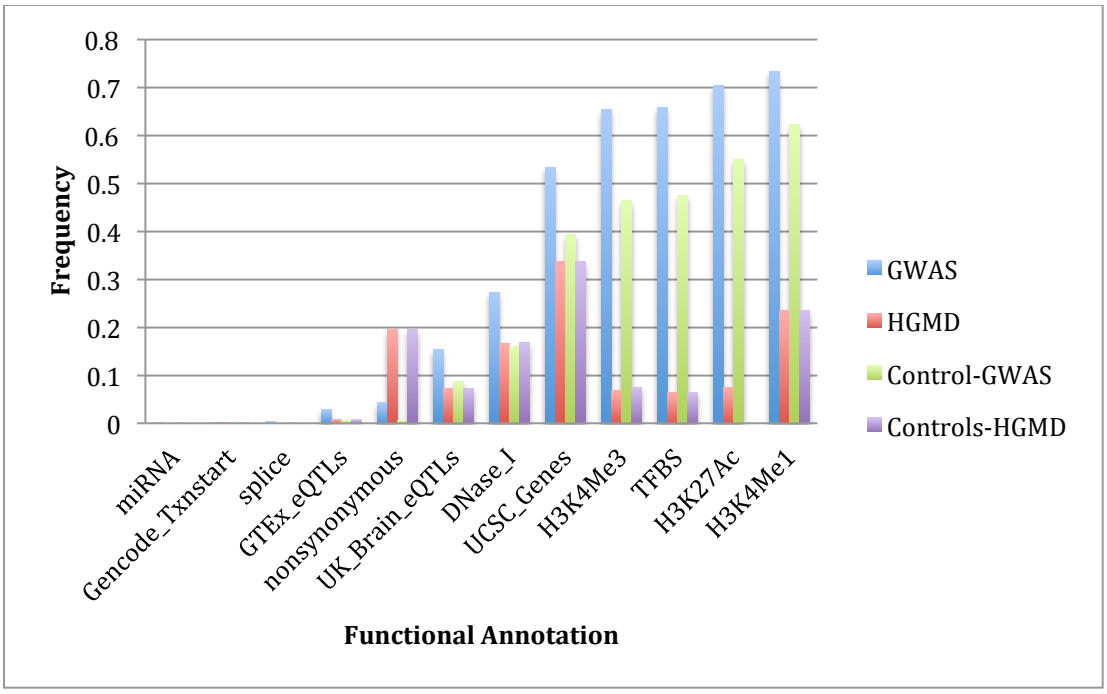


Figure 7.1. Frequency of GWAS Catalogue and HGMD variants that overlap with the binary annotations from Chapter 3.

GWAS Catalogue variants are those with a p-value of $< 5x10^{-8}$ as of May 15, 2015 (n=3607). HGMD variants in the public version provided to Ensembl in the fourth quarter of 2013 (n=3963). The control-GWAS are SNPs (n= 31,663) selected that are not in LD with the selected GWAS Catalogue SNPs but have the same minor allele frequency distribution as the GWAS Catalogue SNPs. The control-HGMD have the same minor allele frequency distribution as the HGMD SNPs (n= 3971). All the selected variants are autosomal variants present in the 1000 Genomes Project. GTEx_eQTLs= cis-eQTL data from the GTEx Consortium, nonsynonymous= nonsynonymous SNP, UK_Brain_eQTLs= cis-eQTL data from the UK Brain Consortium, DNase_I= DNase I hypersensitive sites, UCSC_Genes= UCSC Genes, H3K4Me3= H3K4Me3 histone modification, TFBS= transcription factor binding site, H3K27Ac=H3K27Ac histone modification, H3K4Me1= H3K4Me1 histone modification

This high accuracy held true when a different set of functional genomic annotations were utilized, those utilized by the program GWAVA (Ritchie et al., 2014). The two sets of risk variants could be separated with high accuracy through random forest and elastic net.

Ritchie et al. (2014) applied their random forest model trained using regulatory HGMD variants as a classifier to the non-coding variants in the GWAS Catalogue. They conclude that their model works (slightly but significantly) in scoring GWAS Catalogue variants higher than control variants (Mann-Whitney U test $p=3.6x10^{-29}$) (Ritchie et al., 2014). However, (as I discussed in the predictive accuracy chapter, **Chapter 4**), p-values from statistical tests can be misleading with regard to accuracy of the model. Visualizing the distribution of the two classes is important. Indeed, Ritchie et al. provide a box plot in their Supplementary Material, which demonstrates a strong overlap between the prediction scores for the GWAS Catalogue and control variants, suggesting that their HGMD classifier model was not very effective in identifying GWAS variants.

Missing heritability is likely explained by both common and rare variants (and also other factors such as interactions between genes and between genes and the environment, for instance), and thus databases containing either of these variants are relevant. Future work could involve applying my methodology with the GWAS Catalogue variants to rare variants. It would also be interesting to look at creating a model in which risk variants were defined from various databases considered together rather than just one database.

Furthermore, there are limitations to all of the machine learning methods as discussed in **Chapter 1**. All of the papers also have methodological limitations. For instance, there were several non-standard methodological procedures utilized in the Ritchie et al. paper. For instance, it is common practice to test the accuracy of a model in an independent test set. Ritchie et al. did not reserve any of their samples to create a separate test set. What is more, in random forest, it is recommended to set the minimum sample size at a node to 10% of the overall sample in order to avoid overfitting (Malley et al., 2012). However, Ritchie et al. set the minimum sample size to 1.

For my methodology, a limitation surrounds the selection of control variants from which to differentiate the GWAS Catalogue variants. I selected control variants as those that are on common genotyping arrays. However, imputation has become commonplace in

GWAS, with papers that imputed using HapMap Project data starting in around 2010 (Dupuis et al., 2010; Franke et al., 2010). As a result, the whole genome (or at least the reference genome to which the variants are being imputed: HapMap and/or 1000 Genomes Projects' variants) is being interrogated in GWAS. It may no longer make sense to limit the controls to only variants on genotyping arrays now that more variants in the genome are beginning to be interrogated through imputation. Given this consideration, the use of annotating SNPs with their proxy information when all variants have been assessed may reduce accuracy. However, regardless of imputation, the fact remains that variants present in the GWAS Catalogue may not themselves be the causal variant. A SNP that is in LD with the SNP in the GWAS Catalogue may be the causal SNP, and that SNP may not have the same functional annotations as the GWAS Catalogue SNP. Annotating SNPs with the annotations of their proxies accounts for the uncertainty of the causal SNP in the LD block, as was implemented in **Chapter 3**. Furthermore, work of others has demonstrated that SNPs on genotyping arrays (e.g. 1M Illumina that are not present in the GWAS Catalogue) show a similar pattern to that of the GWAS SNPs, possibly reflecting a bias in the array SNPs for functional regions (Hoffman et al., 2013).

The sample size of known risk variants is also a limitation. A small number of known associated loci with a particular disease makes it challenging to create disease-specific models. However, a more homogenous subset of variants may be required to make more accurate models.

Another limitation to the GWAS Catalogue is that it does not include CNVs. CNVs may contribute to the genetic component of complex disease as well. For instance, there is strong evidence for CNVs contributing to autism spectrum disorders (Devlin and Scherer, 2012; Glessner et al., 2009; Pinto et al., 2010; Sebat et al., 2007). That being said, the inclusion of CNVs may require a consideration of new annotations. For instance, one of the annotations I included, nonsynonmous SNPs, would not apply to CNVs. In addition, Kircher et al.'s annotations for the reference allele and alternate allele or previous amino

acid and new amino acid would not apply to CNVs. Moreover, the effect of having a CNV fall into a regulatory region is not necessarily the same effect as that of having a SNP in that region. For instance, take the case of a transcription factor binding site. A SNP in such a site may lead to reduced or increased binding of the appropriate transcription factor, which could affect the binding of the other factors that interact with that factor. A CNV in that same region, say having more copies of a sequence than in the wild-type, may result in a drastic and copy-number-dependent increase of gene expression. On the other hand, a CNV with fewer copies of a sequence than in the wild-type, can result in decreased expression. Although CNVs may be contributing to the missing heritability, new models may need to be created that are specific to CNVs.

Furthermore, it is important to look at epigenetic marks at various developmental time-points. It is becoming clear that the establishment of epigenetic marks is crucial early in development, and that these functional marks alter throughout development. Even *in utero* environmental differences can modify epigenetic marks, resulting in increased risk of developing a particular trait. An example is malnutrition in the mother (e.g. Dutch Famine in the winter of 1944-1945). Malnutrition can modify DNA methylation, and the prevalence of a trait may be increased in that population (e.g. schizophrenia) (Heijmans et al., 2008; Tobi et al., 2009). The mechanisms underlying these methylation changes due to malnutrition are not known (Tobi et al., 2015). The binding of transcription factors also changes throughout the course of development, and these changes are necessary for normal development (Spitz and Furlong, 2012). Furthermore, DNA methylation patterns change throughout the lifespan; for instance, in the frontal cortex, changes in DNA methylation are important for brain development (Lister et al., 2013). However, all of the functional data considered for the statistical learning presented in this thesis have been from one developmental time point (i.e. adult). There are some data for developmental time points (albeit limited) from the Roadmap Epigenomics Project. Incorporating data from various developmental time points or perhaps variables representing the change in marks between developmental time points may be informative to identify variants

associated with disease. Other limitations to the current work are discussed in the next section along with steps that I could take to overcome them.

7.3 Future directions

Models for genetic variant prioritization can be improved by incorporating more functional annotations from additional tissues/cell types, other functional genomic annotations, and data derived from laboratory techniques that suggest more direct functionality rather than only sequence overlap. Considering rare variant analysis and also the use of more homogenous sets of variants of which to use as a classifier in machine learning algorithms are also relevant.

7.3.1 Tissue-specificity

Tissue-specificity is important in regulation, and applies to many of the functional annotations considered in my statistical learning framework. As discussed in **Chapter 1**, demonstrated disease-associated variants have different functional annotations depending on the tissue, including DNase I hypersensitive sites, transcription factor binding sites, histone modifications, and expression quantitative trait loci (eQTLs) to name a few (Farh et al., 2015; Gagliano et al., 2014a; Maurano et al., 2012; Nicolae et al., 2010). It is understood that epigenetic profiles are tissue-specific. Several groups have shown that there is tissue-specific enrichment of variants in functional annotations, and that subsets of variants show different patterns of enrichment. For example, as mentioned in **Chapter 1**, Maurano et al. (2012) showed that the enrichment of subsets of disease-associated variants in DNase I hypersensitive sites varies depending on the tissue. Although it is well known that tissue-specificity plays an important role in the function of genetic variants dependent on the set of variants considered, tissue-specificity has only been a minor consideration in data-driven genetic variant prioritization models to date. Taking these points into consideration may be key in developing more accurate models for prioritization.

In my analyses I found that all models performed better than chance, except for the brain-related psychiatric analysis, which had limited predictive power. As more data from additional tissue and cell types become available, they can be incorporated into prediction models to improve the accuracy. I started working on a tissue-specificity model for prioritizing psychiatric risk variants.

Pilot Work - prioritizing brain-related psychiatric risk variants

I started incorporating newly available brain data to better prioritize brain-related variants. I hypothesized that brain tissue-specific functional annotations would improve prediction of risk variants in this particular phenotype-specific analysis. Since the publication of my method (Gagliano et al., 2014a), more brain tissue data have become available through the Roadmap Epigenomics Project, as well as an extensive eQTL meta-analysis study that also collected data from the brain (Kim et al., 2014).

I added some additional tissue-relevant regulatory features, and used a more homogenous subset of risk variants (psychiatric-related) into the elastic net algorithm discussed in **Chapter 3**. I downloaded the histone marks for H3K4Me1, H3K4Me3, and H3K27Ac for all of the brain regions from the Roadmap Epigenomics Project from the FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/). Peaks had not been called, and so I used the program MACS (Feng et al., 2011) to compute the ChIP-seq peaks from corresponding background control files of the abundance of reads that were also available.[2]

I downloaded a more recent version of the GWAS Catalogue (May 15, 2015) that contained the additional loci identified by the large meta-analysis for schizophrenia

[2] For parameters, I set the size of the sequencing tags to 35, and scaled the smaller dataset towards the larger. In the case of replicates for a particular tissue and histone mark, which replicate to select is arbitrary. I visually inspected the input files on the UCSC Genome Browser, and if both had adequate signals I chose the largest replicate that also had the corresponding background control file.

(Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). In order to maximize my sample size of disease-variants, while keeping them as homogenous as possible, I selected all variants associated with any of the five psychiatric diseases (schizophrenia, bipolar disorder, major depressive disorder, autism and attention deficit hypersensitivity disorder) shown to share a proportion of common variants (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). I excluded other brain-related or neurodegenerative disorders (e.g. Alzheimer's disease or Parkinson's disease). Recall that the GWAS Catalogue reports variants with association p-values of up to 10^{-5}. In order to have an adequate number of variants for the statistical learning procedure, I used all the GWAS Catalogue variants that met the above criteria, rather than restricting to only the subset that reached genome-wide significance with a p-value of less than $5x10^{-8}$. There were a total of 915 independent variants that met these criteria. I used elastic net as previously described in the Methods of **Chapter 3**, but only annotated the variants with the brain-specific functional annotations: the histone marks from the Roadmap Epigenomics Project described above, brain eQTLs from Kim et al. (2014) and brain eQTLs from the UK Brain Expression Consortium (UKBEC) (Trabzuni et al., 2011). Unfortunately, the results from this larger and more homogenous set of variants with the brain-specific functional annotations, did not offer much better predictive accuracy than by chance: the AUC in the test set was 0.534 (and a similar AUC was observed in the training set, 0.535 demonstrating that the model was not over-fitted).

This result suggests that there is still work to be done in improving the accuracy of a psychiatric-specific prioritization model, which could involve adding more functional annotations. A logical next step would be to incorporate more brain-level data into the model, which will be further discussed below.

Next steps- prioritizing brain-related psychiatric risk variants

One could use brain-level data soon to be available from the new PsychENCODE project, http://psychencode.org/. The goal of this project is to look at regulatory elements (e.g.

transcription factor binding sites) as was done by the ENCODE Project, but in either schizophrenia or control post-mortem brains.

Furthermore, UKBEC has recently generated new RNA-sequencing data, soon to be made publically available. Studies to identify eQTLs previously used microarrays to measure gene expression, but there are limitations to this methodology that RNA-sequencing can overcome (e.g. novel genes and non-coding or microRNAs, allele-specificity, and alternative splicing are taken into account in the latter). Additionally, many eQTL studies perform their analyses on whole tissue, rather than specific regions. UKBEC, however, has performed RNA-sequencing on targeted regions in the brain: substantia nigra, putamen, and hippocampus in a large number of post-mortem unaffected brains (N=150). These data represent a unique resource that could be useful to incorporate in the brain-related psychiatric model. Additionally, there are new RNA-sequencing data from the GTEx Project (Ardlie et al., 2015), albeit not yet from those specific brain regions as for the UKBEC data.

7.3.2 Incorporating additional functional genomic annotations

There are also other functional annotations that may prove to be relevant to include that provide observational evidence that suggests functionality, for example, splicing QTLs (sQTLs), which are genetic variants that affect the generation of transcript isoforms of the same genes (Ardlie et al., 2015; Zhang et al., 2015). Again, these are tissue specific, and the authors who coined the term show that sQTLs are significantly enriched for SNPs associated with traits in previous GWAS (that is to say SNPs present in the GWAS Catalogue).

Another option is to look at allele-specific epigenetic effects, which could indicate a potential regulatory role for the variant exhibiting this allele-specific effect. For instance, Peralta et al. (2014) at the Genetic Analysis Workshop 19 investigated changes in allele-specific chromatin accessibility (measured as DNase-seq read depth of each allele at a

heterozygous locus). They mapped genome-wide genotypes from a reference sample to sequencing reads for DNase I hypersensitive sites (DHS) for heterozygous SNPs. SNPs that show a significant difference in chromatin accessibility between the alleles may suggest that that SNP can compromise DHSs.

7.3.3 Annotating not based solely on location overlap

Another possible way forward would be annotating variants based on data from laboratory methods with results that imply an actual function due to the physical interaction between the DNA sequence and the protein of interest. DNA variants falling into a sequence that is part of a protein's recognition sequence does not necessarily mean those variants are functional. The variant itself may not fall precisely within the consensus region for binding, and also an effect may not be seen due to redundancy of function with another site (Spitz and Furlong, 2012). Furthermore, the interaction between a protein and a stretch of DNA (for instance, detected through ChIP-seq) does not necessarily imply that that region of DNA is functional, meaning that there are effects resulting in alterations downstream. For instance, binding of a transcription factor can occur without influencing the transcription of any genes (Shlyueva et al., 2014). However, there are methods that confirm an interaction between two stretches of DNA as a result of a bound protein, and data from such methods suggest functionality. Annotating based on evidence for functionality from a DNA-protein-DNA interaction, would make the annotations less noisy. There has been an evolution of variations and extensions of the chromosome conformation capture (3C) method to detecting such physical interactions between fragments of DNA (for instance, between promoter and enhancer regions).

Essentially, all of these 3C-based methods involve creating a one-dimensional image of a three-dimensional structure. The chromatin is fixed, and then digested. Afterwards, the sticky-ends of the cross-linked DNA fragments are allowed to ligate together. This procedure can detect which fragments are far away on the linear chromosome template,

but co-localize in space (Wit and Laat, 2012). In the 3C procedure, PCR primers are designed for the ends of the fragments, so that the frequency and sequence of those fragments can then be quantified by quantitative polymerase chain reaction (qPCR) (Dekker et al., 2002). Rather than qPCR, chromosome conformation capture-on-chip (4C) applies next-generation sequencing or microarrays to the 3C procedure, and it uses restriction enzymes to digest the DNA before the ligation step. Chromosome conformation capture carbon copy (5C) and Hi-C offer interaction frequency, a high throughput, and less PCR bias compared to 3C (Wit and Laat, 2012). 5C does not have as good a resolution as Hi-C since the former is based on distances between oligonucleotides whereas the latter depends on the sequencing depth (Wit and Laat, 2012). However, unlike ChIP-seq (the method used for the ENCODE and Roadmap Epigenomics Projects histone modification data), both 5C and Hi-C methods are able to concurrently observe many or all interactions of one DNA sequence with multiple sequences elsewhere. These data are useful to observe with which genes the regulatory element interacts. These experimental observations can subsequently be used to infer biological pathways that may be relevant to understanding the disease of interest.

Furthermore, DNase footprinting can be used to get a more precise location of where the protein of interest binds to the DNA sequence compared to ChIP-seq. For ChIP-seq, formaldehyde is used to cross-link proteins to DNA. Sonication shears the chromatin to a target size of 100 to 300 base pairs, and the protein of interest bound to DNA is then isolated with an antibody specific for the factor. Those DNA fragments that were cross-linked with the factor of interest in a ChIP-seq experiment can be used as the input for DNase footprinting. In this technique, labelled DNA sequences are fragmented by DNase I. The location in the sequence that is bound to the protein is protected from being cleaved, and thus one can infer that that is where the protein is bound. In this way, through the use of restriction enzymes, highly occupied binding sites can be detected at high resolution (Hesselberth et al., 2009).

7.3.4 Incorporating prediction scores into rare variant analysis

In rare variant analysis, variants can be grouped together based on genes or sliding windows. Rare variant association tests will weight variants based on features, for example minor allele frequency, where the weight assigned to a variant is the inverse of the minor allele frequency, and in that way the rarer the variant the higher the weight. Other weights that can be included reflect the impact on amino acid sequence, such as PolyPhen category ("benign," "possibly damaging," or "probably damaging"), and other sequence-based annotations (Lee et al., 2014).

During my PhD, I briefly explored a similar idea of up-weighting rare variants (only those found in genes) using sequence-based weights. I did this work using real (i.e. not simulated) hypertension phenotype data and sequencing data of chromosome 3 from the Genetic Analysis Workshop 18 (GAW18) meeting in Stevenson, Washington (October 2012). For the weights, I used the simple model of whether a SNP is nonsynonymous and whether or not it falls into a DNase I hypersensitive site. Tests for association were conducted in SKAT-O, one analysis without functional weights and the other with the weights. The use of weights based on those two functional annotations did not improve power in the analysis, which is likely due to the simplicity of the model.

I propose that a new weighting scheme can be to use the prediction scores from the prioritization model using the functional annotations to weight SNPs in rare variant association analysis. The higher the prediction score, the larger the weight. In this way, more weight is assigned to those variants that are more likely to have functional consequences that result in a non-wild-type phenotype.

7.3.5 Using a homogenous set of genetic risk variants for training

Some key findings that I would like to bring back up are that I found that different annotations came up as important for different sets of variants (Gagliano et al., 2014a), and that the predictive accuracy of the models varied (Gagliano et al., 2015b). I also

found that the use of variants from other databases, such as variants in HGMD (Stenson et al., 2009), produced models with varying results as well (Gagliano et al., 2015a). These observations suggest that use of a homogenous ascertained set of the disease-associated variants may create models with higher accuracy. Ritchie et al. (2014) tried using a homogenous subset of regulatory variants in the HGMD Catalogue, and I (Gagliano et al., 2014a) tried using phenotype-specific variants from the GWAS Catalogue. However, both of these subsets are based only on current knowledge of variants, and thus are limited.

As discussed in this thesis, I performed a supervised statistical learning method on phenotype-specific sets of disease-associated variants (which were subjectively categorized based on descriptions provided in the GWAS Catalogue). In order to identify novel disease-associated loci objectively, I propose to identify more homogenous subsets of disease-associated variants through unsupervised learning. The unsupervised learning methods that can be employed are *K*-means clustering and principal components analysis. Those subsets can be used as classifiers in supervised learning, which would include penalized regression like elastic net, and decision-tree methods for example. Recall that in unsupervised learning, the algorithm is unaware of which variants are disease-associated; this method is employed to discover any patterns inherent in the data on which the algorithm is trained. In supervised learning, the algorithm is aware of which variants are disease-associated (for instance knowledge derived from GWAS Catalogue as in my work); this method can be employed to develop and test the accuracy of the models derived to predict novel disease-associated variants and identify novel structures in genomic data.

7.4 To the future

In this thesis the focus has been on using functional genomic information to prioritize which genetic variants are functional or are likely associated with a complex disease or trait of interest. First of all, there needs to be an unbiased large set of genetic risk variants

from which to make the predictions (for instance, not primarily common variants as in the GWAS Catalogue or coding sequence biases as in HGMD).

The precision and quality of the features inputted into the models is also important. Functional data is becoming more abundant and technologies for quantifying these data are improving. Predicting the functionality of genetic variants using high-quality data (e.g. at single base pair resolution, and in a tissue-specific manner) in phenotype-specific models will allow the predictions for each variant to be incorporated together to predict the risk of a particular person to develop a particular trait.

In the perfect world every SNP in the human genome will be completely characterized from observations conducted in hundreds of individuals in every available cell type. In this way, the entire DNA sequence will be available for searching for novel disease-associated loci, as well as for fine-mapping variants at disease-associated loci in relevant tissue for the disease. For rare Mendelian disorders, it would be necessary to sequence hundreds (which may be all) of the cases.

I predict that a big leap in the future will be to use the scoring of genetic variants in order to predict the status of a person for numerous diseases/traits based on genome-wide genetic variants and functional information while they are in the prodromal phase, and this knowledge can then be used for earlier treatment or preventative measures. When such procedures are successful, the consequences could look a lot like the fictional film GATTACA (Niccol, 1997). In one of the earlier scenes in the film, when a baby is born at the hospital, the nurse takes a blood sample, and from the DNA sequence is immediately able to tell the parents the probabilities of their child having a whole array of diseases, and even the baby's estimated age of death. However, one can defy their odds as in the case of the main character in GATTACA; he does not experience his apparently highly probable heart deficits, outlives his premature estimated age of death, and ultimately succeeds in his dreams that should have been impossible for a person with his genetic "imperfections".

These probabilities determined in GATTACA are presumably based solely on the genomic sequence itself, and examples of being able to confidently make disease-risk predictions currently exist. For instance, in the domain of genetic testing, tests exist for disorders with strong genetic components. For example, the presence of 40 or more CAG repeats in the first exon of the huntingtin gene (*HTT*) results in Huntington's disease (Lench et al., 2013), or the deletion of the codon that encodes phenylalanine at position 508 in the cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) gene (*CFTR*) in homozygous state, among other mutations, results in Cystic Fibrosis. Direct-to-consumer companies (e.g. 23andMe) have looked at specific genetic variants to predict simple non-medical traits such as whether or not one is likely to be able to smell asparagus in his/her urine, and also (more controversially) to predict the risk of developing complex diseases (e.g. Alzheimer's disease, diabetes, or cardiovascular disease).

Much work is being done in the area of prediction, but the scores are generally used for other purposes such as exploring disease overlap (International Schizophrenia Consortium et al., 2009) or for the prediction of benign versus malignant tumors (Steyerberg et al., 1995). In late 2013, the USA's Food and Drug Association ordered 23andMe to stop providing consumers with health-related data (but they can still use the genetic data to investigate ancestry) (The Associated Press, 2013).

However, I envision that in the future, the algorithm responsible for determining these probabilities will be based upon a number of factors in addition to the actual genotypes, including: epigenetic data from the actual individual at single cell resolution (i.e. instead of using publically available ENCODE data for instance), biochemical biomarkers such as blood levels of a particular protein), gene expression data, and other childhood environmental factors known to be important for health outcomes (including socioeconomic status). After all, with regard to the latter point, there is strong evidence that early exposures to adversity (such as maltreatment or neglect) can alter epigenetic

modifications (for example, (Boyce and Kobor, 2015)), which have downstream effects on phenotype, and so it is logical to be able to make predictions based on more than just genetic factors, but rather both genetics and the environment. From these inputs, one will obtain all the probabilities of the person's risk of developing a number of diseases and traits.

Large challenges will be presented to society with the algorithm that I am envisioning for the future that will be responsible for determining the probabilities of one developing a particular complex disease or trait will be based upon a number of factors in addition to the actual genotypes. There may be some people who choose that they would rather not know their risks. Additionally, the challenge will also come for healthcare professionals to explain to the public that these risks are only probabilities, and not certainties. Yet, as beautifully depicted in GATTACA, these probabilities do not and should never define the worth and value of a human being.

References

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abdi, H., Chin, W.W., Vinzi, V.E., Russolillo, G., and Trinchera, L. (2013). New Perspectives in Partial Least Squares and Related Methods (Springer Science & Business Media).

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). Molecular Biology of the Cell (New York: Garland Science).

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.

Antonarakis, S.E., and Beckmann, J.S. (2006). Mendelian disorders deserve more attention. Nat. Rev. Genet. *7*, 277–282.

Appavu, S., Rajaram, R., Nagammai, M., Priyanga, N., and Priyanka, S. (2011). Bayes Theorem and Information Gain Based Feature Selection for Maximizing the Performance

of Classifiers. In Advances in Computer Science and Information Technology, N. Meghanathan, B.K. Kaushik, and D. Nagamalai, eds. (Springer Berlin Heidelberg), pp. 501–511.

Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Barenboim, M., and Manke, T. (2013). ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. Bioinformatics *29*, 2197–2198.

Barrans, J.D., and Liew, C.-C. (2006). "Chip"ping away at heart failure. Methods Mol. Med. *126*, 157–169.

Baumann, M., Pontiller, J., and Ernst, W. (2010). Structure and Basal Transcription Complex of RNA Polymerase II Core Promoters in the Mammalian Genome: An Overview. Mol. Biotechnol. *45*, 241–247.

Ben-hur, A., and Weston, J. (2007). A user's guide to support vector machines.

Birdsill, A.C., Walker, D.G., Lue, L., Sue, L.I., and Beach, T.G. (2011). Postmortem interval effect on RNA and gene expression in human brain tissue. Cell Tissue Bank. *12*, 311–318.

Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. Nat. Rev. Genet. *3*, 872–882.

Boulesteix, A.-L., Janitza, S., Hapfelmeier, A., Van Steen, K., and Strobl, C. (2014). Letter to the Editor: On the term "interaction" and related phrases in the literature on Random Forests. Brief. Bioinform.

Boyce, W.T., and Kobor, M.S. (2015). Development and the epigenome: the “synapse” of gene-environment interplay. Dev. Sci. *18*, 1–23.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.

Browning, B.L., and Browning, S.R. (2007). Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet. Epidemiol. *31*, 365–375.

Buske, O.J., Manickaraj, A., Mital, S., Ray, P.N., and Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. Bioinformatics *29*, 1843–1850.

Butler, M.G. (2009). Genomic imprinting disorders in humans: a mini-review. J. Assist. Reprod. Genet. *26*, 477–486.

Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.-F., Wang, Q., Krueger, B.J., et al. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science *347*, 1436–1441.

Clancy, S., and Brown, W. (2008). Translation: DNA to mRNA to Protein. Nat. Educ. *1*, 101.

Combarros, O., Alvarez-Arcaya, A., Sánchez-Guerra, M., Infante, J., and Berciano, J. (2002). Candidate gene association studies in sporadic Alzheimer’s disease. Dement. Geriatr. Cogn. Disord. *14*, 41–54.

Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. Arthritis Res. Ther. *13*, 101.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet *381*, 1371–1379.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. Bioinformatics *28*, i147–i153.

Davis, J., and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning, (New York, NY, USA: ACM), pp. 233–240.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

Devlin, B., and Scherer, S.W. (2012). Genetic architecture in autism spectrum disorder. Curr. Opin. Genet. Dev. *22*, 229–237.

Disanto, G., Kjetil Sandve, G., Ricigliano, V.A., Pakpoor, J., Berlanga-Taylor, A.J., Handel, A.E., Kuhle, J., Holden, L., Watson, C.T., Giovannoni, G., et al. (2014). DNase hypersensitive sites and association with multiple sclerosis. Hum Mol Genet *23*, 942–948.

Dodd, P.R., Hambley, J.W., Cowburn, R.F., and Hardy, J.A. (1988). A comparison of methodologies for the study of functional transmitter neurochemistry in human brain. J. Neurochem. *50*, 1333–1345.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet *42*, 105–116.

Eddy, S.R. (2001). Non–coding RNA genes and the modern RNA world. Nat. Rev. Genet. *2*, 919–929.

Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: Illuminating the Dark Road from Association to Function. Am J Hum Genet *93*, 779–797.

Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., Hwang, S.J., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature *478*, 103–109.

Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., et al. (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217–221.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Falls, J.G., Pulford, D.J., Wylie, A.A., and Jirtle, R.L. (1999). Genomic Imprinting: Implications for Human Disease. Am. J. Pathol. *154*, 635–647.

Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to Identify Peaks from ChIP-Seq Data. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *CHAPTER*, Unit2.14.

Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet *42*, 1118–1125.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. (2003). Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. Genome Res. *13*, 1–12.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. *15*.

Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2014). minimac2: faster genotype imputation. Bioinformatics btu704.

Gagliano, S.A., Barnes, M.R., Weale, M.E., and Knight, J. (2014a). A Bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. PLoS ONE *9*, e98122.

Gagliano, S.A., Tiwari, A.K., Freeman, N., Lieberman, J.A., Meltzer, H.Y., Kennedy, J.L., Knight, J., and Muller, D.J. (2014b). Protein kinase cAMP-dependent regulatory type II beta (PRKAR2B) gene variants in antipsychotic-induced weight gain. Hum Psychopharmacol *29*, 330–335.

Gagliano, S.A., Ravji, R., Barnes, M.R., Weale, M.E., and Knight, J. (2015a). Smoking Gun or Circumstantial Evidence? Comparison of Statistical Learning Methods using Functional Annotations for Prioritizing Risk Variants. Sci. Rep. *5*, 13373.

Gagliano, S.A., Paterson, A.D., Weale, M.E., and Knight, J. (2015b). Assessing models for genetic prediction of complex traits: a comparison of visualization and quantitative methods. BMC Genomics *16*, 405.

Gertz, J., Varley, K.E., Reddy, T.E., Bowling, K.M., Pauli, F., Parker, S.L., Kucera, K.S., Willard, H.F., and Myers, R.M. (2011). Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. PLoS Genet *7*, e1002228.

Gibbs JR, van der B.M. Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genet. *6*, e1400952.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877–885.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature *459*, 569–573.

Griffiths, A., Wessler, S., Lewontin, R., and Carroll, S. (2008). Introduction to Genetic Analysis (W. H. Freeman).

Griswold, A., Van Booven, D., Dueker, N., Martin, E., Cuccaro, M., Gilbert, J., Haines, J., Hussman, J., and Pericak-Vance, M. (2014). Computational evaluation of the pathogenicity of noncoding sequence variants in autism spectrum disorder (San Diego, CA).

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760–1774.

Hart, A.B., de Wit, H., and Palmer, A.A. (2013). Candidate gene studies of a promising intermediate phenotype: failure to replicate. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol. *38*, 802–816.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (New York, NY: Springer-Verlag New York).

Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E., and Lumey, L.H. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc. Natl. Acad. Sci. U. S. A. *105*, 17046–17049.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods *6*, 283–289.

Hindorff, L., Junkins, H., Mehta, J., and Manolio, T. (2010). A catalog of published genome-wide association studies.

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

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods *9*, 473–476.

Hothorn, T., Hornik, K., van de Wiel, M.A., and Zeileis, A. (2006). A Lego System for Conditional Inference. Am. Stat. *60*, 257–263.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet *5*, e1000529.

Hutchinson, J.N., Raj, T., Fagerness, J., Stahl, E., Viloria, F.T., Gimelbrant, A., Seddon, J., Daly, M., Chess, A., and Plenge, R. (2014). Allele-Specific Methylation Occurs at Genetic Variants Associated with Complex Disease. PLoS ONE *9*, e98464.

International Human Genome Sequencing Consortium, Adekoya, E., Ait-Zahra, M., Allen, N., Anderson, M., Anderson, S., Anufriev, F., Ambruster, J., Ayele, K., Baker, J., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

Iversen, E.S., Lipton, G., Clyde, M.A., and Monteiro, A.N. (2014). Functional annotation signatures of disease susceptibility loci improve SNP association analysis. BMC Genomics *15*, 398.

James, G., Witten, D.M., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning with applications in R (New York, NY: Springer).

Janipalli, C.S., Kumar, M.V.K., Vinay, D.G., Sandeep, M.N., Bhaskar, S., Kulkarni, S.R., Aruna, M., Joglekar, C.V., Priyadharshini, S., and Maheshwari, N. (2012). Analysis of 32 common susceptibility genetic variants and their combined effect in predicting risk of Type 2 diabetes and related traits in Indians. Diabet. Med. *29*, 121–127.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Philip Schumm, L., Sharma, Y., Anderson, C.A., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493–D496.

Kashyap, C., Tikka, B., Sharma, S., Kumari, S., Verma, P., Sharma, S., and Arya, V. (2011). Human cancer cell lines- A brief communication. J. Chem. Pharm. Res. *3*, 514–520.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLoS Genet. *10*.

Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., van den Oord, E., and Sullivan, P.F. (2014). A meta-analysis of gene expression quantitative trait loci in brain. Transl. Psychiatry *4*, e459.

Kindt, A.S., Navarro, P., Semple, C.A., and Haley, C.S. (2013). The genomic signature of trait-associated variants. BMC Genomics *14*, 108.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Knight, J., Barnes, M.R., Breen, G., and Weale, M.E. (2011). Using Functional Annotation for the Empirical Determination of Bayes Factors for Genome-Wide Association Study Analysis. PLoS ONE *6*, e14808.

Kothiyal, P., Cox, S., Ebert, J., Aronow, B.J., Greinwald, J.H., and Rehm, H.L. (2009). An Overview of Custom Array Sequencing. Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al *0 7*, Unit – 7.17.

Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.-X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nat. Genet. *47*, 582–588.

Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet *45*, 1452–1458.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res *42*, D980–D985.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

Lango, H., Palmer, C.N.A., Morris, A.D., Zeggini, E., Hattersley, A.T., McCarthy, M.I., Frayling, T.M., and Weedon, M.N. (2008). Assessing the Combined Impact of 18 Common Genetic Variants of Modest Effect Sizes on Type 2 Diabetes Risk. Diabetes *57*, 3129–3135.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Lathrop, G.M., Lalouel, J.M., Julier, C., and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. Proc. Natl. Acad. Sci. U. S. A. *81*, 3443–3446.

Leask, S.J. (2004). Environmental influences in schizophrenia: the known and the unknown. Adv. Psychiatr. Treat. *10*, 323–330.

Lee, J.K. (2010). Road to Statistical Bioinformatics. In Statistical Bioinformatics, J.K. Lee, ed. (John Wiley & Sons, Inc.), pp. 1–6.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am. J. Hum. Genet. *95*, 5–23.

Lemon, J. (2006). Plotrix: a package in the red light district of R. R-News *6*, 8–12.

Lench, N., Barrett, A., Fielding, S., McKay, F., Hill, M., Jenkins, L., White, H., and Chitty, L.S. (2013). The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made. Prenat. Diagn. *33*, 555–562.

Lenz, T.L., Deutsch, A.J., Han, B., Hu, X., Okada, Y., Eyre, S., Knapp, M., Zhernakova, A., Huizinga, T.W.J., Abecasis, G., et al. (2015). Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat. Genet. *47*, 1085–1090.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

Li, B., and Leal, S.M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am. J. Hum. Genet. *83*, 311–321.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.

Mahon, P.B., Pirooznia, M., Goes, F.S., Seifuddin, F., Steele, J., Lee, P.H., Huang, J., Hamshere, M., DePaulo, J.R., Kelsoe, J.R., et al. (2011). Genome-wide association analysis of age at onset and psychotic symptoms in bipolar disorder. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. *156B*, 370–378.

Malley, J.D., Malley, K.G., and Pajevic, S. (2011). Statistical learning for biomedical data (Cambridge: Cambridge University Press).

Malley, J.D., Kruppa, J., Dasgupta, A., Malley, K.G., and Ziegler, A. (2012). Probability machines: consistent probability estimation using nonparametric learning machines. Methods Inf Med *51*, 74–81.

Manolio, T.A. (2013). Bringing genome-wide association findings into clinical use. Nat. Rev. Genet. *14*, 549–558.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., et al. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res.

Mattick, J.S., and Makunin, I.V. (2006). Non-coding RNA. Hum. Mol. Genet. *15 Spec No 1*, R17–R29.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190–1195.

McGhee, J.D., and Felsenfeld, G. (1980). Nucleosome Structure. Annu. Rev. Biochem. *49*, 1115–1156.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinforma. Oxf. Engl. *26*, 2069–2070.

Mendel, G. (1866). Versuche über pflanzen-hybriden. Verhandlungen Naturforschenden Vereines Brünn *4*, 3–47.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

Mill, J., Tang, T., Kaminsky, Z., Khare, T., Yazdanpanah, S., Bouchard, L., Jia, P., Assadzadeh, A., Flanagan, J., Schumacher, A., et al. (2008). Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis. Am. J. Hum. Genet. *82*, 696–711.

Montgomery SB, S.M. Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Mullaney, J.M., Mills, R.E., Pittard, W.S., and Devine, S.E. (2010). Small insertions and deletions (INDELs) in human genomes. Hum. Mol. Genet. *19*, R131–R136.

Neale, M.C. (1992). Methodology for genetic studies of twins and families (Boston: Kluwer Academic Publishers).

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an Unusual Distribution of Rare Variants. PLoS Genet *7*, e1001322.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Niccol, A. (1997). Gattaca.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet *6*, e1000888.

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. *12*, 465–474.

Ozeki, T., Mushiroda, T., Yowang, A., Takahashi, A., Kubo, M., Shirakata, Y., Ikezawa, Z., Iijima, M., Shiohara, T., Hashimoto, K., et al. (2011). Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. Hum. Mol. Genet. *20*, 1034–1041.

Parra, E., Eaton, K., Kavanagh, P., Edwards, M., and Krithika, S. (2014). Association study confirms that two OCA2 polymorphisms are involved in normal skin pigmentation variation in East Asian populations (San Diego, CA).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825−2830.

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

Peralta, J.M., Almeida, M., Abraham, L., Moses, E., and Blangero, J. (2014). Finding potential cis-regulatory loci using allele-specific chromatin accessibility as weights in a kernel-based variance component test (Vienna, Austria).

Petes, T.D. (2001). Meiotic recombination hot spots and cold spots. Nat. Rev. Genet. *2*, 360–369.

Pickering, T.G. (1997). The effects of environmental and lifestyle factors on blood pressure and the intermediary role of the sympathetic nervous system. J. Hum. Hypertens. *11 Suppl 1*, S9–S18.

Pickrell, J.K. (2014). Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet *94*, 559–573.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorder. Nature *466*, 368–372.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinforma. Oxf. Engl. *19*, 149–150.

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

Quang, D., Chen, Y., and Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics btu703.

Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinforma. Oxf. Engl. *31*, 761–763.

Race, R., Sanger, R., Lawler, S.D., and Bertinshaw, D. (1949). The inheritance of the MNS blood groups: A second series of families. Heredity *3*, 205–213.

R Core Development Team (2008). A language and environment for statistical computing. R Found. Stat. Comput.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen — The Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.

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

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.

Rice, T., Rankinen, T., Province, M.A., Chagnon, Y.C., Pérusse, L., Borecki, I.B., Bouchard, C., and Rao, D.C. (2000). Genome-wide linkage analysis of systolic and diastolic blood pressure: the Québec Family Study. Circulation *102*, 1956–1963.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet *45*, 1150–1159.

Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.

Riva, A. (2012). The MAPPER2 Database: a multi-genome catalog of putative transcription factor binding sites. Nucleic Acids Res. *40*, D155–D161.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Rosenbaum, L., Hinselmann, G., Jahn, A., and Zell, A. (2011). Interpreting linear support vector machine models with heat map molecule coloring. J. Cheminformatics *3*, 11.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol *6*, e107.

Schalkwyk, L.C., Meaburn, E.L., Smith, R., Dempster, E.L., Jeffries, A.R., Davies, M.N., Plomin, R., and Mill, J. (2010). Allelic skewing of DNA methylation is widespread across the genome. Am. J. Hum. Genet. *86*, 196–212.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res *22*, 1748–1759.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

Schork, A.J., Thompson, W.K., Pham, P., Torkamani, A., Roddey, J.C., Sullivan, P.F., Kelsoe, J.R., O'Donovan, M.C., Furberg, H., Schork, N.J., et al. (2013). All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. PLoS Genet *9*, e1003449.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. Rare Allele Hypotheses for Complex Diseases. Curr. Opin. Genet. Dev. *19*, 212–219.

Schwarz, J.M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods *7*, 575–576.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong Association of De Novo Copy Number Mutations with Autism. Science *316*, 445–449.

Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinforma. Oxf. Engl. *31*, 1536–1543.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034–1050.

Silventoinen, K., Helle, S., Nisén, J., Martikainen, P., and Kaprio, J. (2013). Height, age at first birth, and lifetime reproductive success: a prospective cohort study of Finnish male and female twins. Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud. *16*, 581–589.

Sinclair, D.C. (1989). Human growth after birth (Oxford University Press, Incorporated).

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am. J. Hum. Genet. *94*, 599–610.

Skol, A.D., Scott, L.J., Abecasis, G.R., and Boehnke, M. (2007). Optimal designs for two-stage genome-wide association studies. Genet. Epidemiol. *31*, 776–788.

Smialowski, P., Frishman, D., and Kramer, S. (2010). Pitfalls of supervised feature selection. Bioinformatics *26*, 440–443.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613–626.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., and Cooper, D.N. (2009). The Human Gene Mutation Database: 2008 update. Genome Med. *1*, 13.

Stephens, J., and Balding, D. (2009). Bayesian statistical methods for genetic association studies. Nat Rev Genet *10*, 681–690.

Steyerberg, E.W., Keizer, H.J., Fosså, S.D., Sleijfer, D.T., Toner, G.C., Schraffordt Koops, H., Mulders, P.F., Messemer, J.E., Ney, K., and Donohue, J.P. (1995). Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *13*, 1177–1187.

Strange, A., Capon, F., Spencer, C.C., Knight, J., Weale, M.E., Allen, M.H., Barton, A., Band, G., Bellenguez, C., Bergboer, J.G., et al. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet *42*, 985–990.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

Strobl, C., Boulesteix, A.L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics *8*, 25.

Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S., et al. (2015). The impact of low-frequency and rare variants on lipid levels. Nat. Genet. *47*, 589–597.

Tabor, H.K., Risch, N.J., and Myers, R.M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat. Rev. Genet. *3*, 391–397.

Tang, C.S., and Ferreira, M.A.R. (2012). GENOVA: gene overlap analysis of GWAS results. Stat. Appl. Genet. Mol. Biol. *11*, Article 6.

Tenesa, A., and Haley, C.S. (2013). The heritability of human disease: estimation, uses and abuses. Nat. Rev. Genet. *14*, 139–149.

Terwilliger, J.D., and Ott, J. (1994). Handbook of Human Genetic Linkage (Baltimore: The John Hopkins University Press).

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1016–1073.

The Associated Press (2013). Genetic test maker 23andMe told to halt sales.

The ENCODE Project Consortium (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol *9*, e1001046.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

The GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. Nat Genet *45*, 580–585.

Thompson, J.R., Gogele, M., Weichenberger, C.X., Modenese, M., Attia, J., Barrett, J.H., Boehnke, M., De Grandi, A., Domingues, F.S., Hicks, A.A., et al. (2013). SNP prioritization using a Bayesian probability of association. Genet Epidemiol *37*, 214–221.

Tobi, E.W., Lumey, L.H., Talens, R.P., Kremer, D., Putter, H., Stein, A.D., Slagboom, P.E., and Heijmans, B.T. (2009). DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. Hum. Mol. Genet. *18*, 4046–4053.

Tobi, E.W., Slieker, R.C., Stein, A.D., Suchiman, H.E.D., Slagboom, P.E., Zwet, E.W. van, Heijmans, B.T., and Lumey, L.H. (2015). Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome. Int. J. Epidemiol. dyv043.

Trabzuni, D., Ryten, M., Walker, R., Smith, C., Imran, S., Ramasamy, A., Weale, M.E., and Hardy, J. (2011). Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. J Neurochem *119*, 275–282.

Tsoi, L.C., Spain, S.L., Knight, J., Ellinghaus, E., Stuart, P.E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J.E., et al. (2012). Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat Genet *44*, 1341–1348.

Verbeek, E.C., Bakker, I.M.C., Bevova, M.R., Bochdanovits, Z., Rizzu, P., Sondervan, D., Willemsen, G., de Geus, E.J., Smit, J.H., Penninx, B.W., et al. (2012). A Fine-Mapping Study of 7 Top Scoring Genes from a GWAS for Major Depressive Disorder. PLoS ONE *7*, e37384.

Vesell, E.S. (1991). Genetic and environmental factors causing variation in drug response. Mutat. Res. *247*, 241–257.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics *13 Suppl 4*, S2.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genet *8*, e1002793.

Wain, L.V., Armour, J.A.L., and Tobin, M.D. (2009). Genomic copy number variation, human health, and disease. Lancet Lond. Engl. *374*, 340–350.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930–D934.

Watson, J., and Crick, F. (1953). Molecular Structure of Nucleic Acids. Nature *171*, 737–738.

Weir, B.S. (1990). Genetic Data Analysis: Methods for Discrete Population Genetic Data (Sinauer Associates Incorporated).

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

Wit, E. de, and Laat, W. de (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev. *26*, 11–24.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

Xavier, R.J., Huett, A., and Rioux, J.D. (2008). Autophagy as an important process in gut homeostasis and crohn's disease pathogenesis. Gut *57*, 717–720.

Xu, M., Bi, Y., Xu, Y., Yu, B., Huang, Y., Gu, L., Wu, Y., Zhu, X., Li, M., and Wang, T. (2010). Combined effects of 19 common variations on type 2 diabetes in Chinese: results from two community-based studies. PLoS One *5*, e14022.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

Yatabe, Y., Tavaré, S., and Shibata, D. (2001). Investigating stem cells in human colon by using methylation patterns. Proc. Natl. Acad. Sci. U. S. A. *98*, 10839–10844.

Zhang, X., Joehanes, R., Chen, B.H., Huan, T., Ying, S., Munson, P.J., Johnson, A.D., Levy, D., and O'Donnell, C.J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat. Genet. *47*, 345–352.

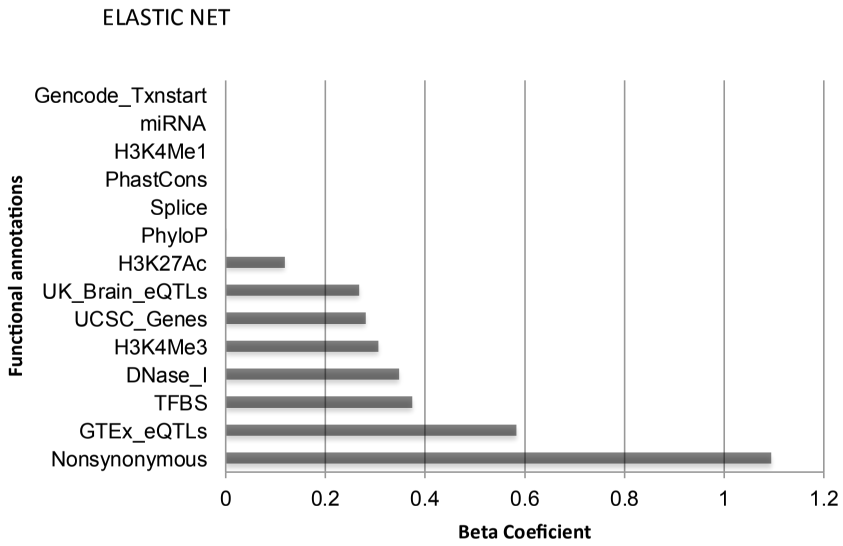
Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B- Stat. Methodol. *67*, 301–320.

Appendices

A Supplementary Tables and Figures for Chapter 5 Comparison of Statistical Learning Methods Using Functional Annotations for Prioritizing Risk Variants

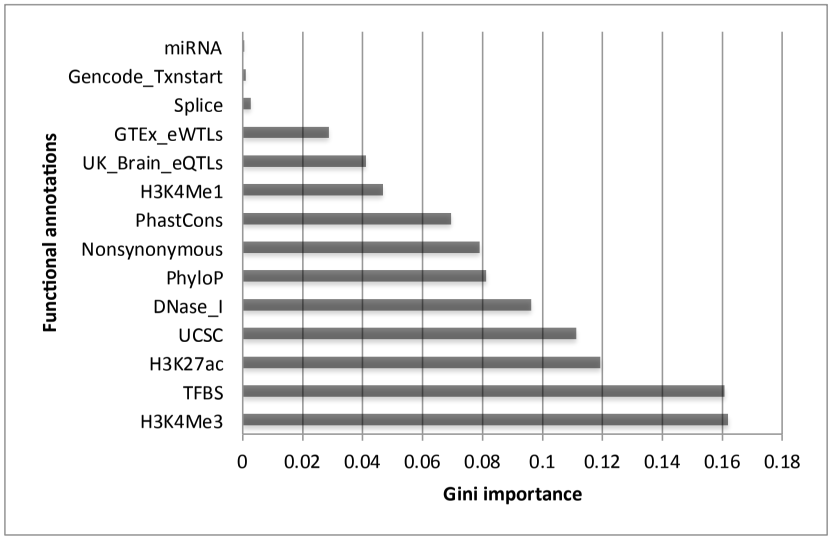
These models are based on the following classifier: variants in the GWAS Catalogue with a p-value < 5E-8 and a random subset of control variants from common genotyping arrays. The annotations from Gagliano et al. were used.

Nonsynonymous	1.09457232
GTEx_eQTLs	0.58352715
TFBS	0.37469007
DNase_I	0.34800772
H3K4Me3	0.30673014
UCSC_Genes	0.2806802
UK_Brain_eQTLs	0.26813544
H3K27Ac	0.11850658
PhyloP	0.00121327
Splice	0
PhastCons	0
H3K4Me1	0
miRNA	0
Gencode_Txnstart	0



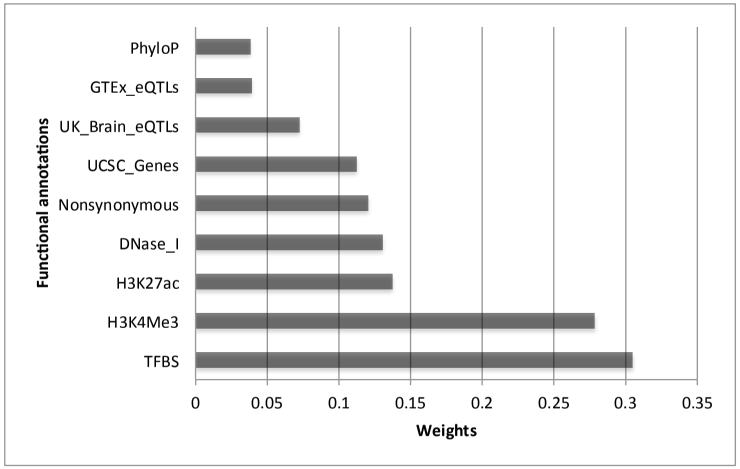
H3K4Me3	0.16194218
TFBS	0.16076837
H3K27ac	0.11930863
UCSC	0.11125026
DNase_I	0.09617033
PhyloP	0.08126218
Nonsynonymous	0.07904821
PhastCons	0.06953112
H3K4Me1	0.04671044
UK_Brain_eQTLs	0.04111495
GTEx_eWTLs	0.0287619
Splice	0.0027309
Gencode_Txnstart	0.00094908
miRNA	0.00045145

RANDOM FOREST



TFBS	0.30471134
H3K4Me3	0.27845552
H3K27ac	0.1374529
DNase_I	0.13058098
Nonsynonymous	0.12064785
UCSC_Genes	0.11260425
UK_Brain_eQTLs	0.07259671
GTEx_eQTLs	0.0394731
PhyloP	0.03845896

SUPPORT VECTOR MACHINE

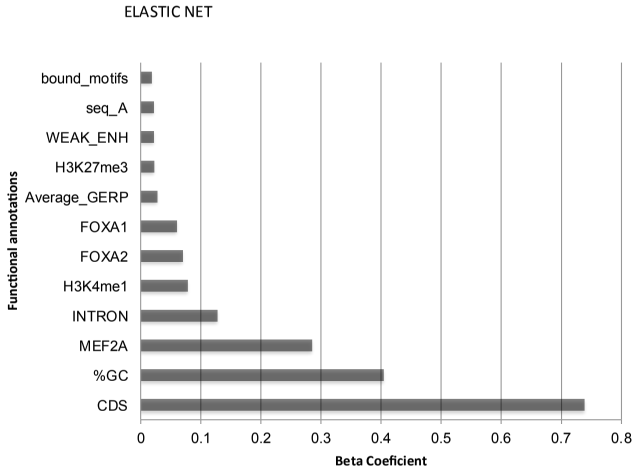


Annotation Legend: See Gagliano et al. for further details.

Nonsynonymous= Nonsynonymous SNP
GTEx_eQTLs= cis eQTL from the GTEx Project
TFBS= Transcription factor binding site
DNase_I= DNaseI hypersensitive site
H3K4Me3= H4K4Me3 histone modification
UCSC_Genes= UCSC Gene
UK_Brain_eQTLs= cis eQTL from the UK Brain Consortium
H3K27Ac= H3K27Ac histone modification
PhyloP= PhyloP conservation score
Splice= +/-5 base pairs from a splice site
PhastCons= PhastCons conservation score
H3K4Me1= H3K4Me1 histone modification
miRNA= microRNA target as defined by TargetScan
Gencode_Txnstart= Transcription start site as defined by Gencode

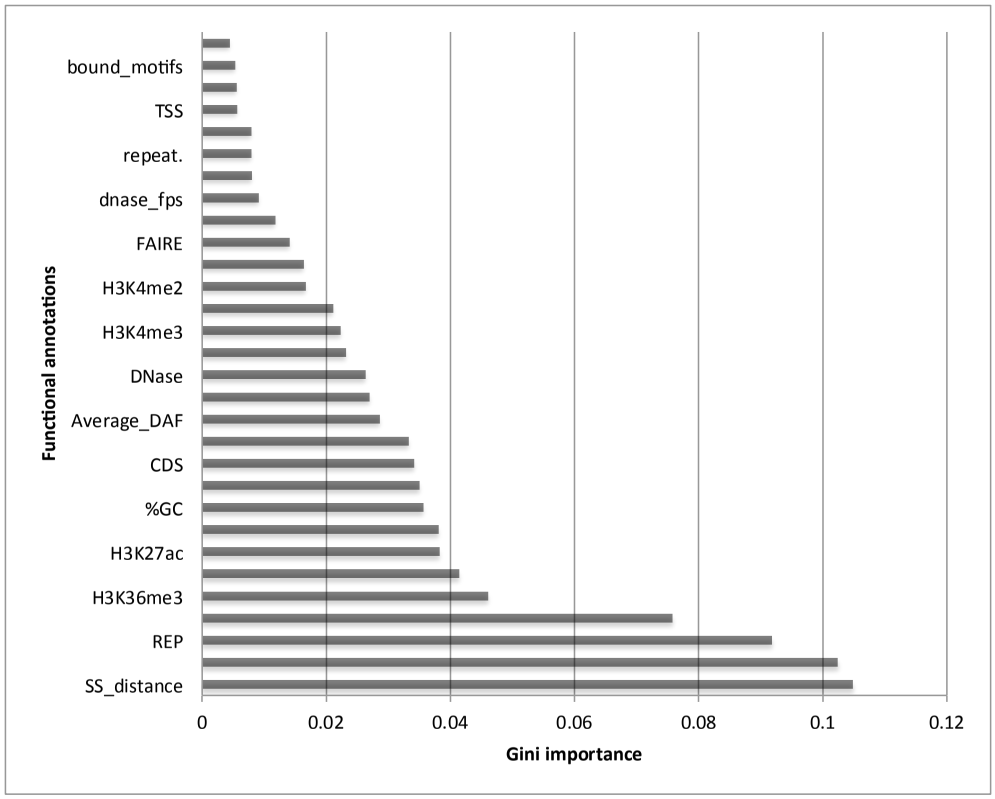
These models are based on the following classifier: variants in the GWAS Catalogue with a p-value < 5E-8 and a random subset of control variants from common genotyping arrays. The annotations from Ritchie et al. were used.

CDS	0.73835991
%GC	0.40444405
MEF2A	0.28525379
INTRON	0.12824043
H3K4me1	0.0787302
FOXA2	0.07059569
FOXA1	0.0605902
Average_GERP	0.02813259
H3K27me3	0.02267571
WEAK_ENH	0.02244694
seq_A	0.02225467
bound_motifs	0.01886099



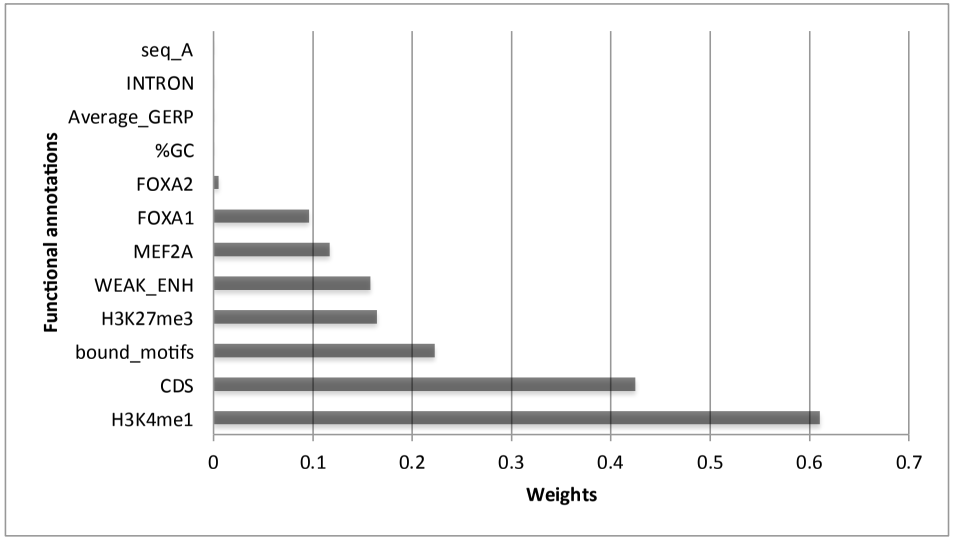
SS_distance	0.10484422
H3K4me1	0.10242725
REP	0.09179542
TSS_distance	0.0758117
H3K36me3	0.04611345
H3K9ac	0.04144153
H3K27ac	0.03827775
Average_GERP	0.03809685
%GC	0.03564533
Average_het	0.0350499
CDS	0.03418113
TRAN	0.0332855
Average_DAF	0.02864317
GERP	0.02700529
DNase	0.02631869
H3K79me2	0.02315817
H3K4me3	0.02233333
EXON	0.02109917
H3K4me2	0.0166956
H2AFZ	0.01635564
FAIRE	0.0141235
H3K27me3	0.01177772
dnase_fps	0.00912907
ENH	0.00799977
repeat.	0.00797199
POLR2A	0.00793303
TSS	0.00564932
INTRON	0.00556233
bound_motifs	0.00531464
UTR3	0.00446204

RANDOM FOREST --> too many non-zero annots for a graph (131). The top 30 are displayed here



H3K4me1	0.61016666
CDS	0.42464906
bound_motifs	0.22274777
H3K27me3	0.16459611
WEAK_ENH	0.15813353
MEF2A	0.11712527
FOXA1	0.09599154
FOXA2	0.00525437
%GC	0.00010008
Average_GERP	6.13E-05
INTRON	5.17E-05
seq_A	4.37E-05

SUPPORT VECTOR MACHINE



Annotation Legend: See Ritchie et al. for further details.

%GC= GC content of 100bp flanking region

Average_DAF= mean derived allele frequency of variants in 1kb flanking region

Average_GERP= mean GERP score of 100bp flanking region

Average_het= mean heterozygosity of 1kb flanking region

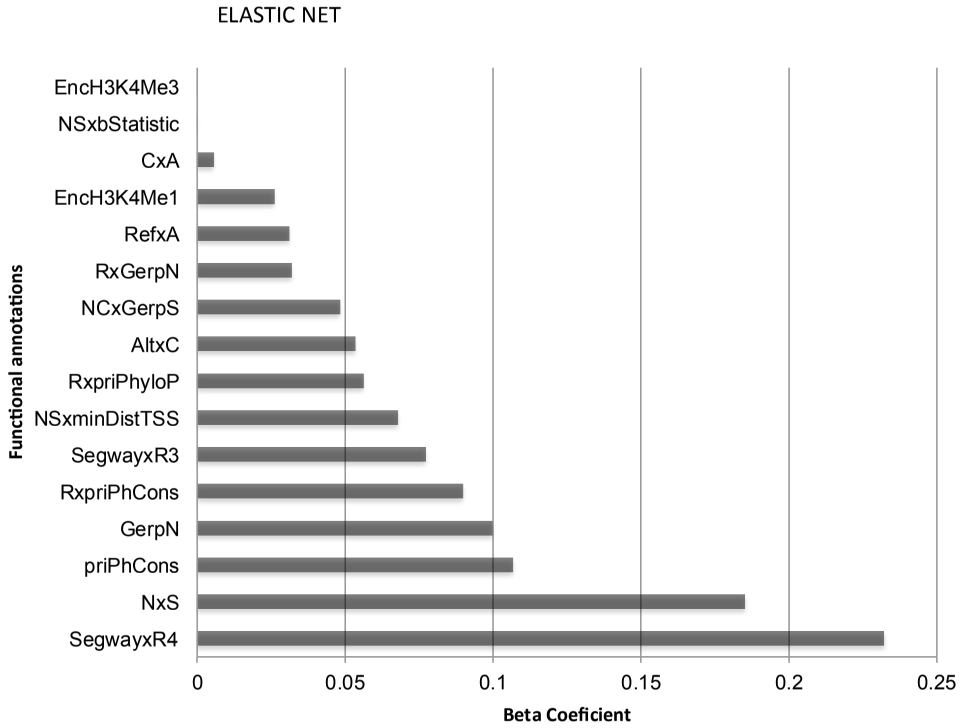
CDS= coding sequence

DNase= DNase1-seq peak

ENH= predicted enhancer segment
EXON= exonic region
FAIRE= FAIRE-seq peak
FOXA1= FOXA1 Transcription Factor ChIP-seq peaks
FOXA2= FOXA2 Transcription Factor ChIP-seq peaks
GC= GC content of 100bp flanking region
GERP= GERP score at the variant locus
H2AFZ= H2AFZ Histone modification ChIP-seq peaks
H3K27ac= H3K27ac Histone modification ChIP-seq peaks
H3K27me3= H3K27me3 Histone modification ChIP-seq peaks
H3K36me3= H3K36me3 Histone modification ChIP-seq peaks
H3K4me1= H3K4me1 Histone modification ChIP-seq peaks
H3K4me2= H3K4me2 Histone modification ChIP-seq peaks
H3K4me3= H3K4me3 Histone modification ChIP-seq peaks
H3K79me2= H3K79me2 Histone modification ChIP-seq peaks
H3K9ac= H3K9ac Histone modification ChIP-seq peaks
INTRON= intronic region
MEF2A= MEF2A Transcription Factor ChIP-seq peaks
POLR2A= POLR2A Transcription Factor ChIP-seq peaks
REP= predicted repressed sequence
SS_distance= distance to the nearest splice site
TRAN= predicted transcribed segment
TSS= predicted promoter segment
TSS_distance= distance to the nearest TSS
UTR3= 3 prime untranslated region
WEAK_ENH= predicted weak enhancer segment
bound_motifs= bound transcription factor motifs
dnase_fps= DNase1-seq footprint
repeat.= annotated repeat element
seq_A= reference base at variant locus is A

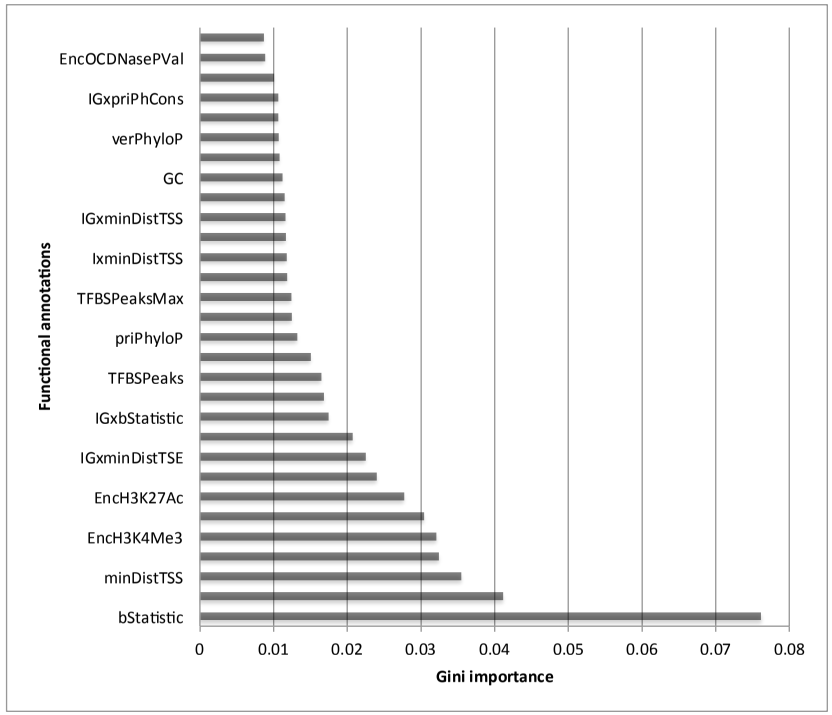
These models are based on the following classifier: variants in the GWAS Catalogue with a p-value < 5E-8 and a random subset of control variants from common genotyping arrays. The annotations from Kircher et al. were used.

SegwayxR4	0.23214744
NxS	0.18512232
priPhCons	0.1068047
GerpN	0.09977634
RxpriPhCons	0.0898631
SegwayxR3	0.07726528
NSxminDistTSS	0.06787831
RxpriPhyloP	0.05626706
AltxC	0.05350694
NCxGerpS	0.04830807
RxGerpN	0.03203313
RefxA	0.03125206
EncH3K4Me1	0.02614708
CxA	0.00567947
NSxbStatistic	0.00030707
EncH3K4Me3	0.00015327



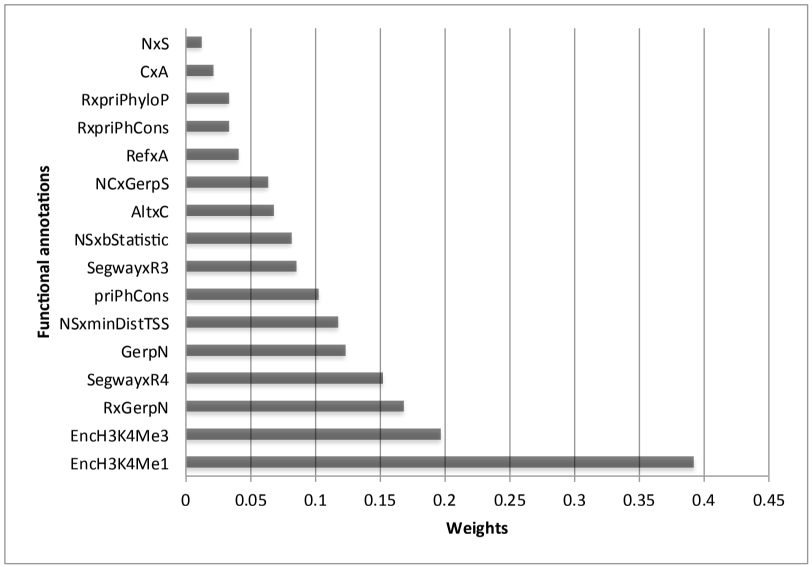
bStatistic	0.07612596
EncH3K4Me1	0.041142
minDistTSS	0.03549463
minDistTSE	0.0324394
EncH3K4Me3	0.03210379
IxbStatistic	0.03040545
EncH3K27Ac	0.02770442
GerpN	0.02402458
IGxminDistTSE	0.02248406
EncExp	0.0207314
IGxbStatistic	0.01747853
GerpS	0.01685097
TFBSPeaks	0.01645934
EncOCFaireSig	0.01505465
priPhyloP	0.01321063
IxminDistTSE	0.01247181
TFBSPeaksMax	0.0123973
priPhCons	0.01186039
IxminDistTSS	0.01176682
EncOCmycSig	0.01164866
IGxminDistTSS	0.01163399
EncOCCombPVal	0.01150561
GC	0.01120606
EncOCpolIISig	0.01082852
verPhyloP	0.01070643
TFBS	0.01065238
IGxpriPhCons	0.01062909
EncOCctcfSig	0.01006817
EncOCDNasePVal	0.00883529
EncOCDNaseSig	0.00868473

RANDOM FOREST --> too many non-zero annots for a graph (239). The top 30 are displayed here.



EncH3K4Me1	0.39204676
EncH3K4Me3	0.19670888
RxGerpN	0.16826805
SegwayxR4	0.15207001
GerpN	0.12315151
NSxminDistTSS	0.11758948
priPhCons	0.10252227
SegwayxR3	0.08543625
NSxbStatistic	0.08159819
AltxC	0.06789917
NCxGerpS	0.06358947
RefxA	0.04072076
RxpriPhCons	0.03338482
RxpriPhyloP	0.03325345
CxA	0.02117921
NxS	0.01230448

SUPPORT VECTOR MACHINE



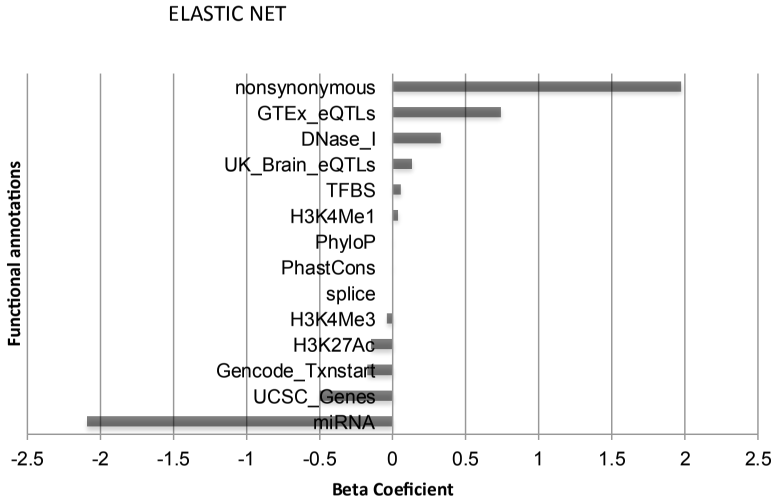
Annotation Legend: See Kircher et al. for further details.

AltxC= Interaction between observed allele and the new amino acid cysteine
CxA= Interaction between the previous amino acid cysteine and the new amino acid alanine
EncExp= Maximum ENCODE expression value
EncH3K27Ac= Maximum ENCODE H3K27 acetylation level
EncH3K4Me1= Maximum ENCODE H3K4 methylation level
EncH3K4Me3= Maximum ENCODE H3K4 trimethylation level
EncOCCombPVal= ENCODE combined p-Value (PHRED-scale) of Faire, Dnase, polII, CTCF, Myc evidence for open chromatin
EncOCDNasePVal= p-Value (PHRED-scale) of Dnase evidence for open chromatin
EncOCDNaseSig= Peak signal for Dnase evidence of open chromatin
EncOCFaireSig= Peak signal for Faire evidence of open chromatin
EncOCctcfSig= Peak signal for CTCF evidence of open chromatin
EncOCmycSig= Peak signal for Myc evidence of open chromatin
EncOCpolIISig= Peak signal for polII evidence of open chromatin
GC= Percent GC in a window of +/- 75bp
GerpN= Neutral evolution score defined by GERP++
GerpS= Rejected Substitution' score defined by GERP++
IGxbStatistic= interaction between intergenic and Background selection score
IGxminDistTSE= interaction between intergenic and Distance to closest Transcribed Sequence End (TSE)
IGxminDistTSS= interaction between intergenic and Distance to closest Transcribed Sequence Start (TSS)
IGxpriPhCons= interaction between intergenic and Primate PhastCons conservation score (excl. human)
IxbStatistic= interaction between intronic and Background selection score
IxminDistTSE= interaction between intronic and Distance to closest Transcribed Sequence End (TSE)
IxminDistTSS= interaction between intronic and Distance to closest Transcribed Sequence End (TSE)
NCxGerpS= interaction between noncoding and Rejected Substitution' score defined by GERP++
NSxbStatistic= interaction between nonsynonymous and Background selection score
NSxminDistTSS= interaction between nonsynonymous and Distance to closest Transcribed Sequence End (TSE)
NxS= interaction between previous amino acid aspargine and synonymous
RefxA= interaction between reference allele and the new amino acid alanine
RxGerpN= interaction between previous amino acid arginine and Neutral evolution score defined by GERP++

RxpriPhCons= interaction between previous amino acid arginine and Primate PhastCons conservation score (excl. human)
RxpriPhyloP= interaction between previous amino acid arginine and Primate PhyloP score (excl. human)
SegwayxR3= Result of genomic segmentation algorithm, R3 category
SegwayxR4= Result of genomic segmentation algorithm, R4 category
TFBS= Number of different overlapping ChIP transcription factor binding sites
TFBSpeaks= Number of overlapping ChIP transcription factor binding site peaks summed over different cell types/tissue
TFBSpeaksMax= Maximum value of overlapping ChIP transcription factor binding site peaks across cell types/tissue
bstatistic= Background selection score
minDistTSE= Distance to closest Transcribed Sequence End (TSE)
minDistTSS= Distance to closest Transcribed Sequence Start (TSS)
priPhCons= Primate PhastCons conservation score (excl. human)
priPhyloP= Primate PhyloP score (excl. human)
verPhyloP= Vertebrate PhyloP score (excl. human)

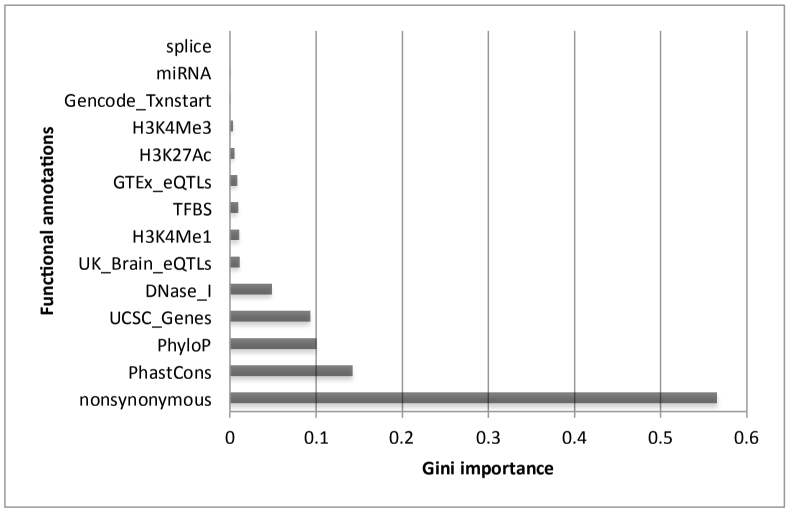
These models are based on the following classifier: HGMD and control variants within 1KB of the HGMD variant.
The annotations from Gagliano et al. were used.

miRNA	-2.0913612
UCSC_Genes	-0.4987777
Gencode_Txnstart	-0.1735627
H3K27Ac	-0.1485844
H3K4Me3	-0.0402223
splice	0
PhastCons	0.0020501
PhyloP	0.00248018
H3K4Me1	0.03728084
TFBS	0.05476116
UK_Brain_eQTLs	0.13337447
DNase_I	0.33042212
GTEx_eQTLs	0.73931183
nonsynonymous	1.97196766



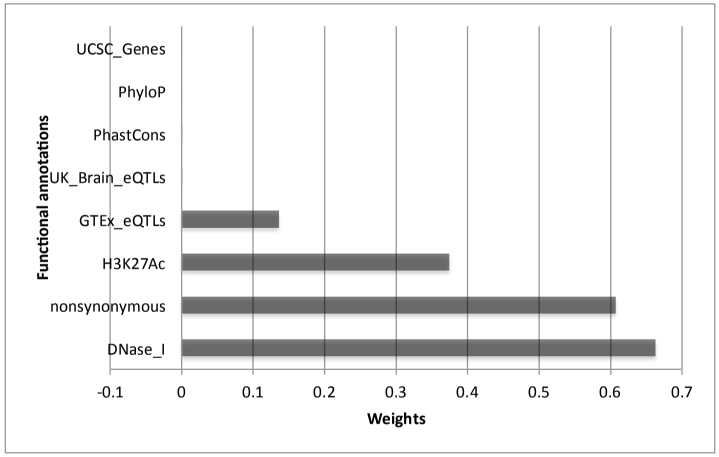
nonsynonymous	0.56512023
PhastCons	0.1424801
PhyloP	0.10031994
UCSC_Genes	0.09342726
DNase_I	0.0485705
UK_Brain_eQTLs	0.01121601
H3K4Me1	0.01100313
TFBS	0.00949715
GTEx_eQTLs	0.00874107
H3K27Ac	0.00536329
H3K4Me3	0.0036272
Gencode_Txnstart	0.00052891
miRNA	0.00010522
splice	0

RANDOM FOREST



SUPPORT VECTOR MACHINE

DNase_I	0.66280725
nonsynonymous	0.60714531
H3K27Ac	0.37433641
GTEx_eQTLs	0.13604832
UK_Brain_eQTLs	0.0001785
PhastCons	3.09E-05
PhyloP	-2.83E-05
UCSC_Genes	-0.0001683

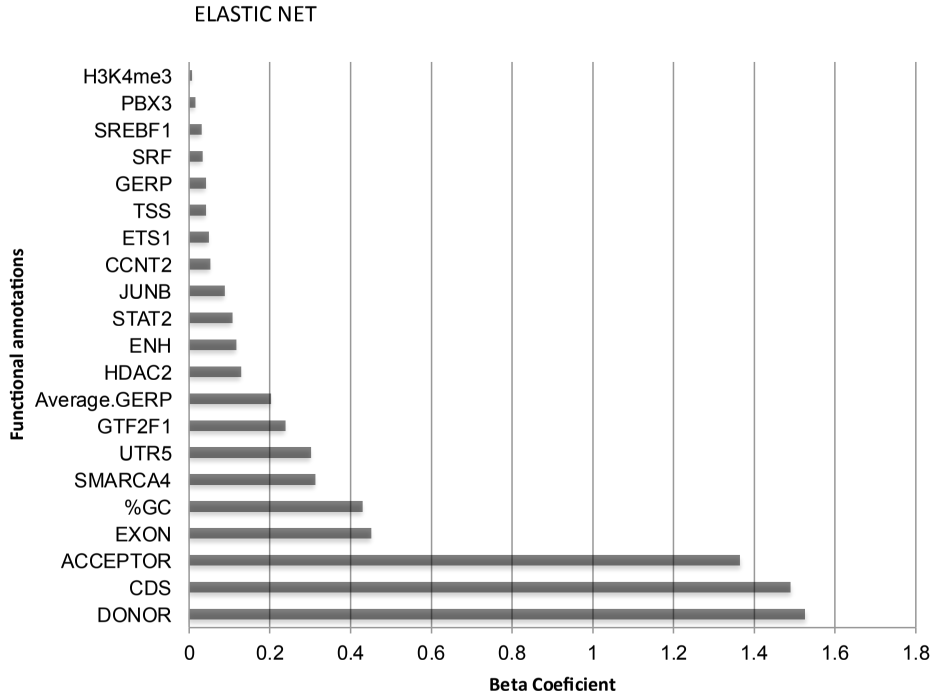


Annotation Legend: See Gagliano et al. for further details.

Nonsynonymous= Nonsynonymous SNP
GTEx_eQTLs= cis eQTL from the GTEx Project
TFBS= Transcription factor binding site
DNase_I= DNaseI hypersensitive site
H3K4Me3= H4K4Me3 histone modification
UCSC_Genes= UCSC Gene
UK_Brain_eQTLs= cis eQTL from the UK Brain Consortium
H3K27Ac= H3K27Ac histone modification
PhyloP= PhyloP conservation score
Splice= +/-5 base pairs from a splice site
PhastCons= PhastCons conservation score
H3K4Me1= H3K4Me1 histone modification
miRNA= microRNA target as defined by TargetScan
Gencode_Txnstart= Transcription start site as defined by Gencode

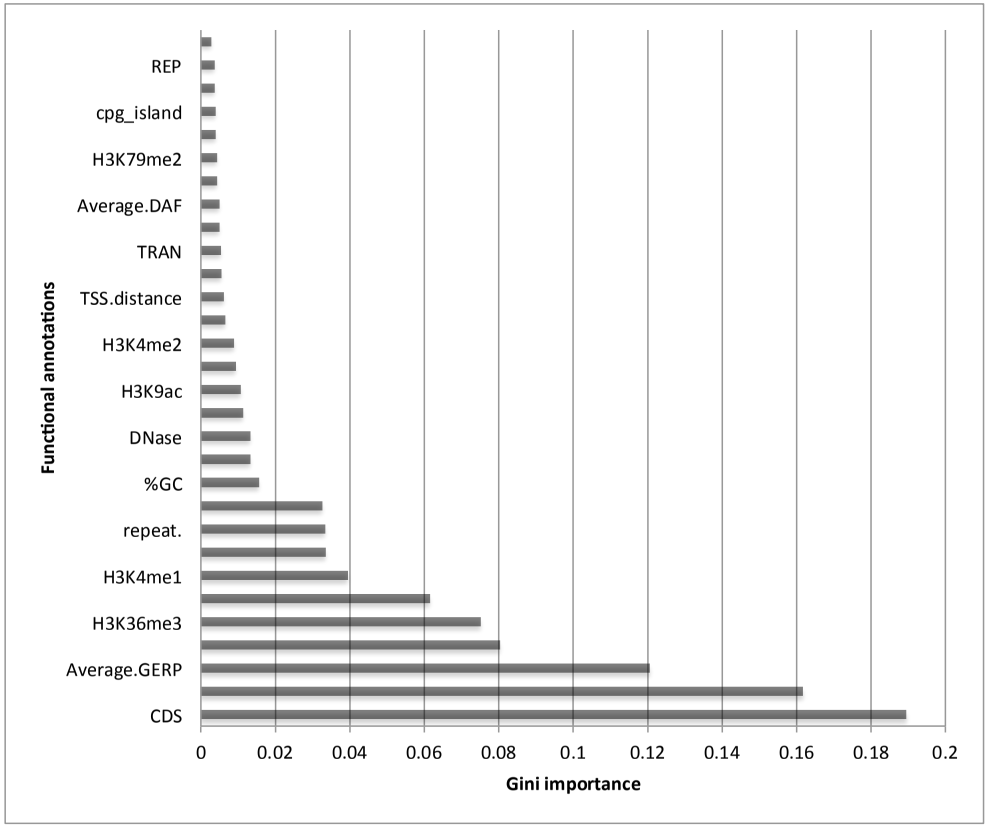
These models are based on the following classifier: HGMD and control variants within 1KB of the HGMD variant. The annotations from Ritchie et al. were used.

DONOR	1.5258467
CDS	1.48920315
ACCEPTOR	1.36461534
EXON	0.4510464
%GC	0.43019703
SMARCA4	0.31248539
UTR5	0.30175308
GTF2F1	0.23865927
Average.GERP	0.20291391
HDAC2	0.12905712
ENH	0.11773293
STAT2	0.1072002
JUNB	0.08893068
CCNT2	0.05291225
ETS1	0.0489323
TSS	0.04240205
GERP	0.04200631
SRF	0.03347607
SREBF1	0.03091549
PBX3	0.01549044
H3K4me3	0.00772507



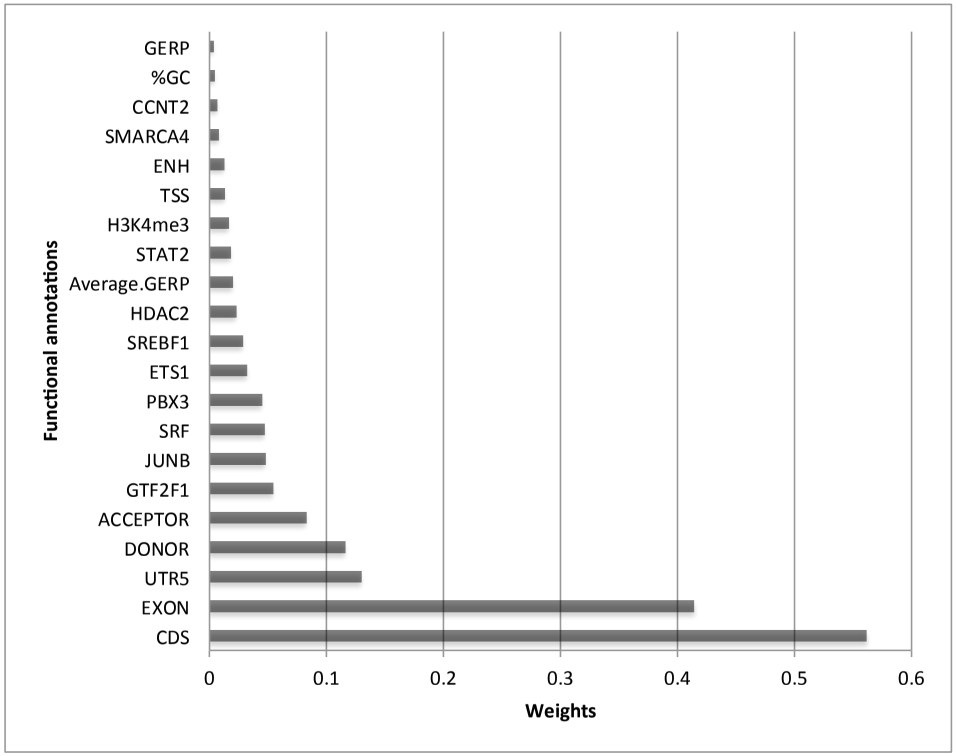
CDS	0.18943882
EXON	0.16168397
Average.GERP	0.12058937
INTRON	0.08031842
H3K36me3	0.07516592
GERP	0.06150859
H3K4me1	0.03945476
H3K27me3	0.03356697
repeat.	0.03339007
SS.distance	0.03264821
%GC	0.01561323
H3K27ac	0.0133447
DNase	0.01328737
in_cpg	0.01135456
H3K9ac	0.01074734
H3K4me3	0.00938739
H3K4me2	0.00896775
H2AFZ	0.0065741
TSS.distance	0.00619892
H4K20me1	0.00548802
TRAN	0.00541612
UTR5	0.00499614
Average.DAF	0.00497807
UTR3	0.00434988
H3K79me2	0.00434417
Average.het	0.00400752
cpg_island	0.003979
TSS	0.00373025
REP	0.00366999
H3K9me3	0.00281866

RANDOM FOREST --> too many non-zero annots for a graph (134). The top 30 are displayed here.



CDS	0.56149539
EXON	0.41436719
UTR5	0.12996327
DONOR	0.11619774
ACCEPTOR	0.0829269
GTF2F1	0.05462709
JUNB	0.04795783
SRF	0.04726239
PBX3	0.04495562
ETS1	0.0323004
SREBF1	0.02882632
HDAC2	0.02325905
Average.GERP	0.02004804
STAT2	0.01835382
H3K4me3	0.01657522
TSS	0.01337685
ENH	0.01261234
SMARCA4	0.00820819
CCNT2	0.0068465
%GC	0.00449145
GERP	0.00378191

SUPPORT VECTOR MACHINE



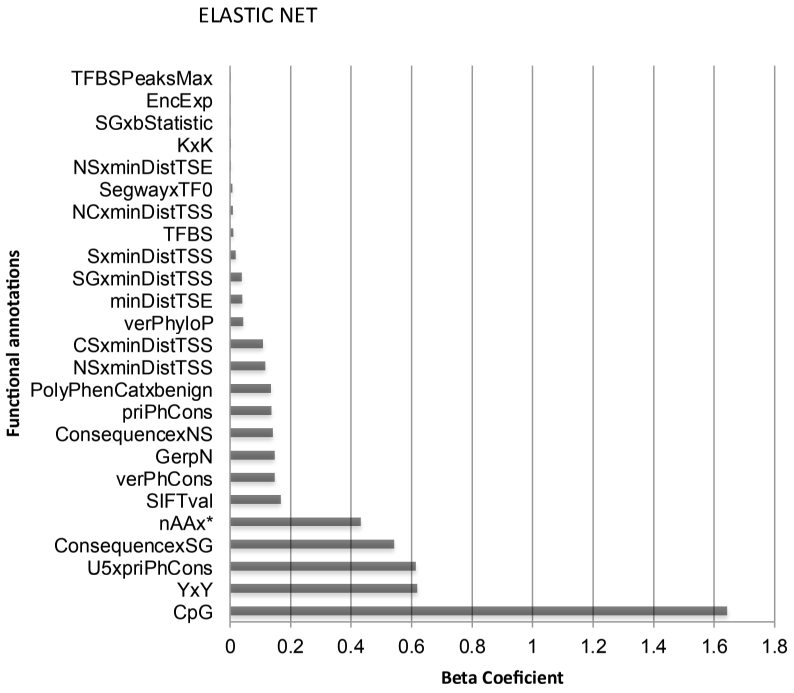
Annotation Legend: See Ritchie et al. for further details.

ACCEPTOR= acceptor splice site
Average.DAF= mean derived allele frequency of variants in 1kb flanking region
Average.GERP= mean GERP score of 100bp flanking region
Average.het= mean heterozygosity of 1kb flanking region
CCNT2= CCNT2 Transcription Factor ChIP-seq peaks
CDS= coding sequence
DNase= DNase1-seq peak
DONOR= donor splice site
ENH= predicted enhancer segment
ETS1= ETS1 Transcription Factor ChIP-seq peaks
EXON= exonic region
GERP= GERP score at the variant locus
GTF2F1= GTF2F1 Transcription Factor ChiP-seq peaks
H2AFZ= H2AFZ Histone modification ChIP-seq peaks
H3K27ac= H3K27ac Histone modification ChIP-seq peaks
H3K27me3= H3K27me3 Histone modification ChIP-seq peaks
H3K36me3= H3K36me3 Histone modification ChIP-seq peaks
H3K4me1= H3K4me1 Histone modification ChIP-seq peaks
H3K4me2= H3K4me2 Histone modification ChIP-seq peaks
H3K4me3= H3K4me3 Histone modification ChIP-seq peaks
H3K79me2= H3K79me2 Histone modification ChIP-seq peaks
H3K9ac= H3K9ac Histone modification ChIP-seq peaks
H3K9me3= H3K9me3 Histone modification ChIP-seq peaks
H4K20me1= H4K20me1 Histone modification ChIP-seq peaks
HDAC2= HDAC2 Transcription Factor ChIP-seq peaks
INTRON= intronic region
JUNB= JUNB Transcription Factor ChIP-seq peaks
PBX3= PBX3 Transcription Factor ChIP-seq peaks
REP= predicted repressed sequence

SMARCA4= SMARCA4 Transcription Factor ChIP-seq peaks
SREBF1= SREBF1 Transcription Factor ChIP-seq peaks
SRF= SRF Transcription Factor ChIP-seq peaks
SS.distance= distance to the nearest splice site
STAT2= STAT2 Transcription Factor ChIP-seq peaks
TRAN= predicted transcribed segment
TSS= predicted promoter segment
TSS.distance= distance to the nearest TSS
UTR3= 3 prime UTR
UTR5= 5 prime UTR
%GC= GC content of 100bp flanking region
cpg_island= Predicted CpG island
in_cpg= reference sequence at variant locus is a CpG dinucleotide
repeat.= annotated repeat element

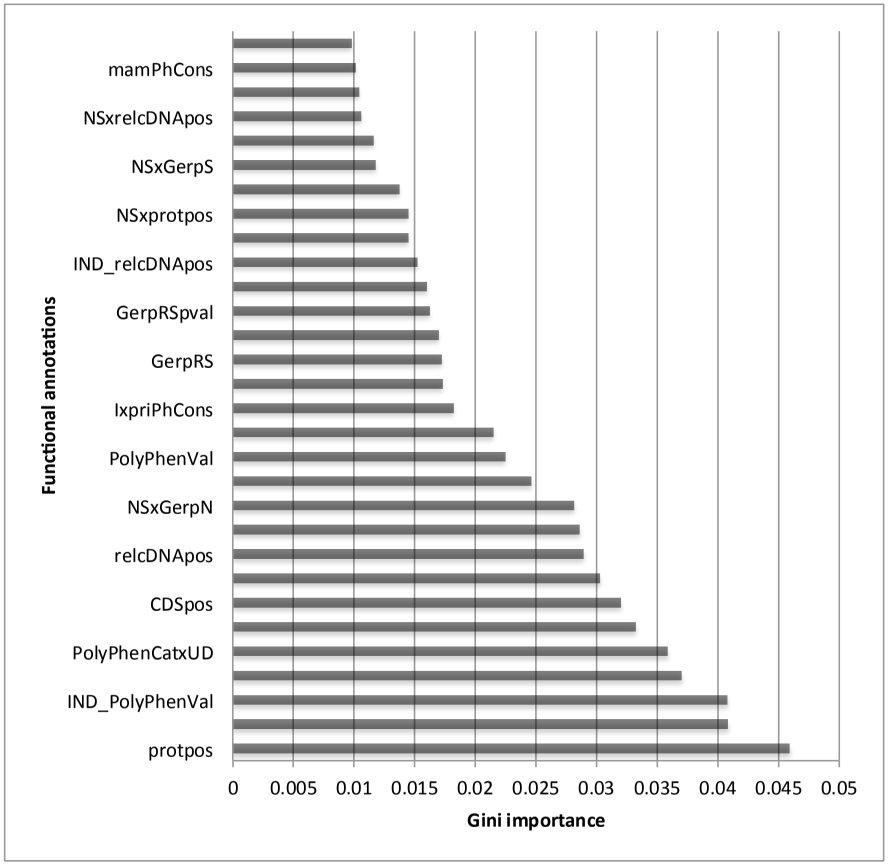
These models are based on the following classifier: HGMD and control variants within 1KB of the HGMD variant.
The annotations from Kircher et al. were used.

CpG	1.64257081
YxY	0.61902895
U5xpriPhCons	0.61425523
ConsequencexSG	0.54174906
nAAx*	0.43276376
SIFTval	0.16734979
verPhCons	0.14840771
GerpN	0.14715183
ConsequencexNS	0.14189243
priPhCons	0.13669347
PolyPhenCatxbenign	0.13499441
NSxminDistTSS	0.11727677
CSxminDistTSS	0.10941404
verPhyloP	0.04275962
minDistTSE	0.0403301
SGxminDistTSS	0.03900904
SxminDistTSS	0.0192986
TFBS	0.0111528
NCxminDistTSS	0.00937678
SegwayxTF0	0.00856464
NSxminDistTSE	0.00389292
KxK	0.0024366
SGxbStatistic	0.00050175
EncExp	0.00023775
TFBSPeaksMax	0.00014412



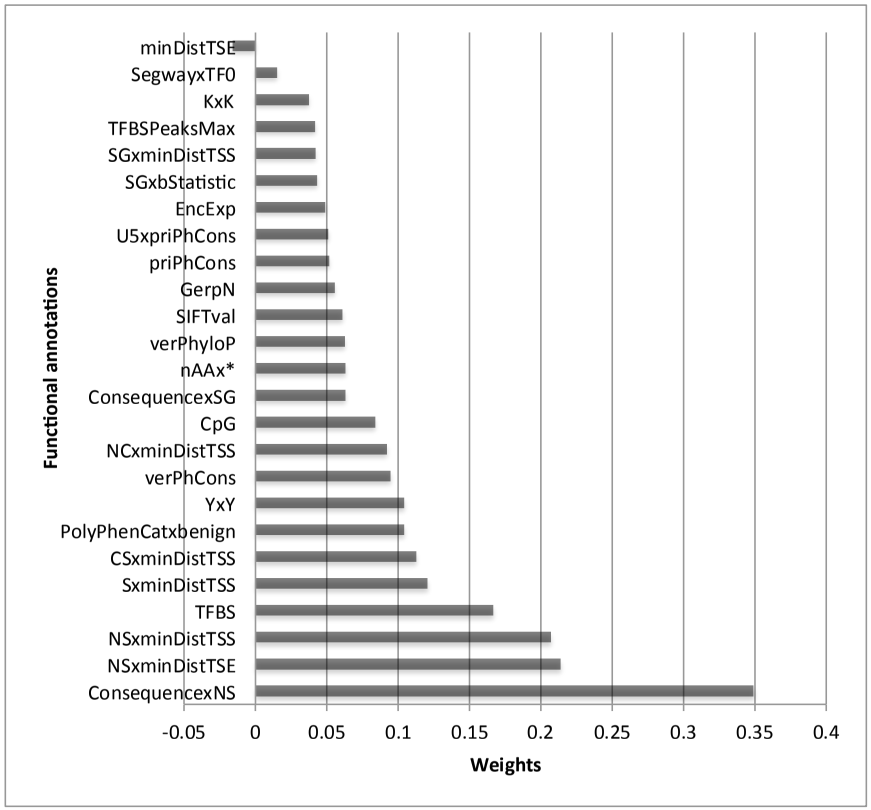
protpos	0.04590905
NSxCDSpos	0.04081936
IND_PolyPhenVal	0.04077479
IND_protpos	0.03701
PolyPhenCatxUD	0.0358595
NSxrelCDSpos	0.03321642
CDSpos	0.03199467
NSxcDNApos	0.03029051
relcDNApos	0.02894276
IND_relCDSpos	0.02860594
NSxGerpN	0.02815068
IND_Grantham	0.02462413
PolyPhenVal	0.02247068
SIFTcatxUD	0.02148823
IxpriPhCons	0.01823818
NSxbStatistic	0.01731354
GerpRS	0.0172512
NSxminDistTSE	0.01699031
GerpRSpval	0.01624278
NSxminDistTSS	0.0160178
IND_relcDNApos	0.0152251
nAAxUD	0.01447571
NSxprotpos	0.0144745
IxbStatistic	0.01376113
NSxGerpS	0.01180528
NSxverPhyloP	0.01163979
NSxrelcDNApos	0.01059885
SIFTval	0.01045299
mamPhCons	0.01013812
IxGerpN	0.00981919

RANDOM FOREST --> too many non-zero annots for a graph (430).
The top 30 are displayed here.



ConsequencexNS	0.34914625
NSxminDistTSE	0.213846
NSxminDistTSS	0.20693293
TFBS	0.1665795
SxminDistTSS	0.12067178
CSxminDistTSS	0.11274411
PolyPhenCatxbenign	0.10430502
YxY	0.10426699
verPhCons	0.09479932
NCxminDistTSS	0.09210029
CpG	0.08386596
ConsequencexSG	0.06312334
nAAx*	0.06312334
verPhyloP	0.06262857
SIFTval	0.06106331
GerpN	0.0555322
priPhCons	0.05166973
U5xpriPhCons	0.05113698
EncExp	0.04900535
SGxbStatistic	0.04323714
SGxminDistTSS	0.04206416
TFBSPeaksMax	0.04194938
KxK	0.03740145
SegwayxTF0	0.01519715
minDistTSE	-0.0160579

SUPPORT VECTOR MACHINE



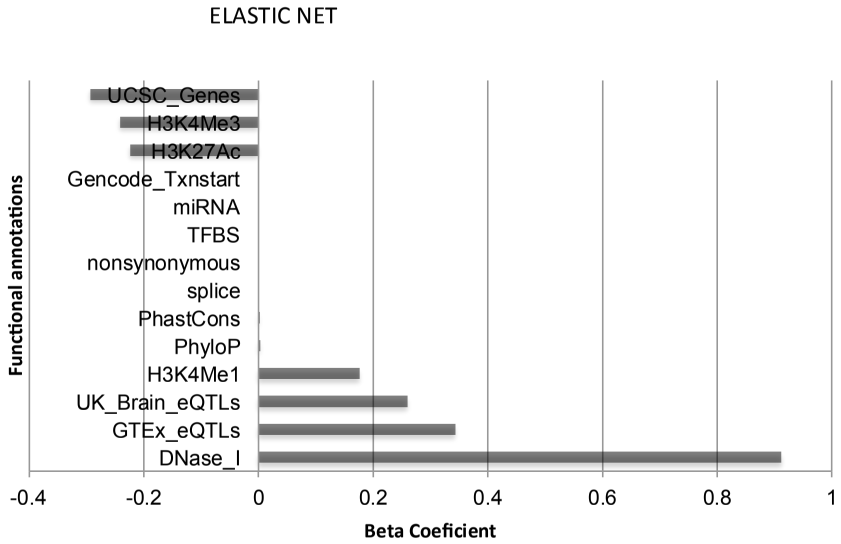
Annotation Legend: See Kircher et al. for further details.

CDSpos= Base position from coding start
CSxminDistTSS= interaction between canonical splice and Distance to closest Transcribed Sequence Start (TSS)
ConsequencexNS= nonsynonymous
ConsequencexSG= stop-gained
CpG= Percent CpG in a window of +/- 75bp
EncExp= Maximum ENCODE expression value
GerpN= Neutral evolution score defined by GERP++
GerpRS= Gerp element score
GerpRSpval= Gerp element p-Value
IND_Grantham= indicator variable for Grantham score: oAA,nAA
IND_PolyPhenVal= PolyPhen score
IND_protpos= indicator variable for Amino acid position from coding start
IND_relCDSpos= indicator variable for Relative position in coding sequence
IND_relcDNApos= indicator variable for Relative position in transcript
IxGerpN= interaction between intronic and Neutral evolution score defined by GERP++
IxbStatistic= interaction between intronic and Background selection score
IxpriPhCons= interaction between intronic and Primate PhastCons conservation score (excl. human)
KxK= interaction between previous amino acid lysine and new amino acid lysine
NCxminDistTSS= interaction between noncoding and Distance to closest Transcribed Sequence Start (TSS)
NSxCDSpos= interaction between nonsynonymous and Base position from transcription start
NSxGerpN= interaction between nonsynonymous and Neutral evolution score defined by GERP++
NSxGerpS= interaction between nonsynonymous and Rejected Substitution' score defined by GERP++
NSxbStatistic= interaction between nonsynonymous and Background selection score
NSxcDNApos= interaction between nonsynonymous and Base position from transcription start
NSxminDistTSE= interaction between nonsynonymous and Distance to closest Transcribed Sequence End (TSE)
NSxminDistTSS= interaction between nonsynonymous and Distance to closest Transcribed Sequence Start (TSS)
Nsxprotpos= interaction between nonsynoymous and Amino acid position from coding start
NSxrelCDSpos= interaction between nonsynonymous and Relative position in coding sequence
NSxrelcDNApos= interaction between nonsynonymous and Relative position in transcript

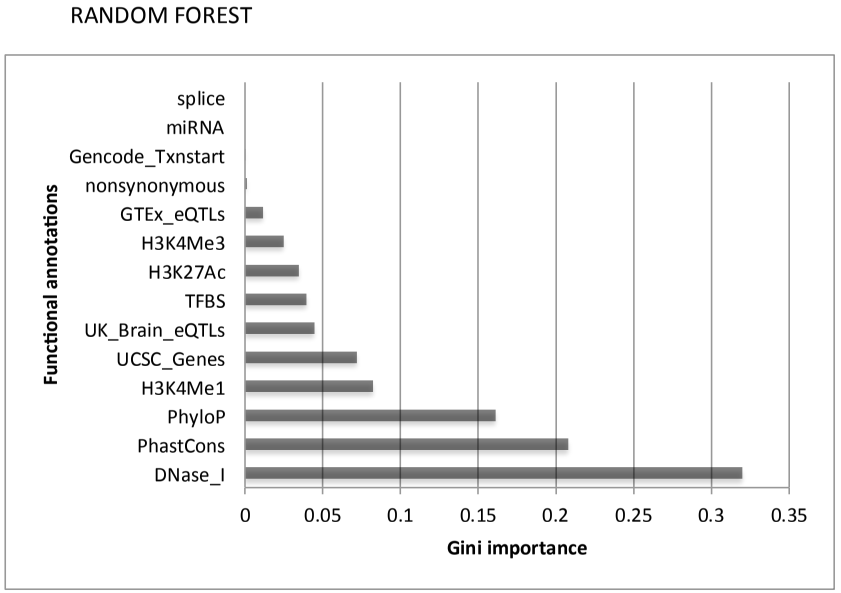
NSxverPhyloP= interaction between nonsynonymous and Vertebrate PhyloP (excl. human)
PolyPhenCatxUD= PolyPhen category, undefined
PolyPhenCatxbenign= PolyPhen category, benign
PolyPhenVal= PolyPhen scorel
SGxbStatistic= interaction between stop-gained and Background selection score
SGxminDistTSS= interaction between stop-gained and Distance to closest Transcribed Sequence Start (TSS)
SIFTcatxUD= SIFT category, undefined
SIFTval= SIFT score
SegwayxTF0= Segway, TFO category
SxminDistTSS= interaction between synonymous and Distance to closest Transcribed Sequence Start (TSS)
TFBS= Number of different overlapping ChIP transcription factor binding sites
TFBSPeaksMax= Maximum value of overlapping ChIP transcription factor binding site peaks across cell types/tissue
U5xpriPhCons= interaction between 5Prime UTR and Primate PhastCons conservation score (excl. human)

These models are based on the following classifier: non-exonic HGMD and non-exonic control variants within 1KB of the HGMD variant. The annotations from Gagliano et al. were used.

DNase_I	0.91216267
GTEx_eQTLs	0.34362132
UK_Brain_eQTLs	0.26032036
H3K4Me1	0.17681986
PhyloP	0.00330053
PhastCons	0.00195327
splice	0
nonsynonymous	0
TFBS	0
miRNA	0
Gencode_Txnstart	0
H3K27Ac	-0.2236143
H3K4Me3	-0.2411067
UCSC_Genes	-0.2931357

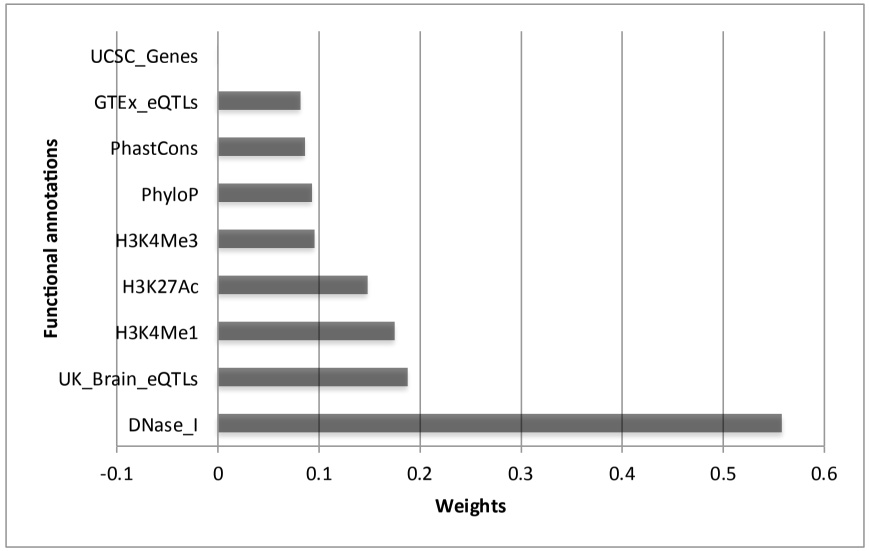


DNase_I	0.31980572
PhastCons	0.20783459
PhyloP	0.16117035
H3K4Me1	0.08249287
UCSC_Genes	0.07202067
UK_Brain_eQTLs	0.04456005
TFBS	0.03965993
H3K27Ac	0.03478116
H3K4Me3	0.02484832
GTEx_eQTLs	0.01158394
nonsynonymous	0.00111109
Gencode_Txnstart	0.00013131
miRNA	0
splice	0



SUPPORT VECTOR MACHINE

DNase_I	0.55783189
UK_Brain_eQTLs	0.18788467
H3K4Me1	0.17503961
H3K27Ac	0.14812591
H3K4Me3	0.09582921
PhyloP	0.09328432
PhastCons	0.08604586
GTEx_eQTLs	0.08190499
UCSC_Genes	-0.0003697

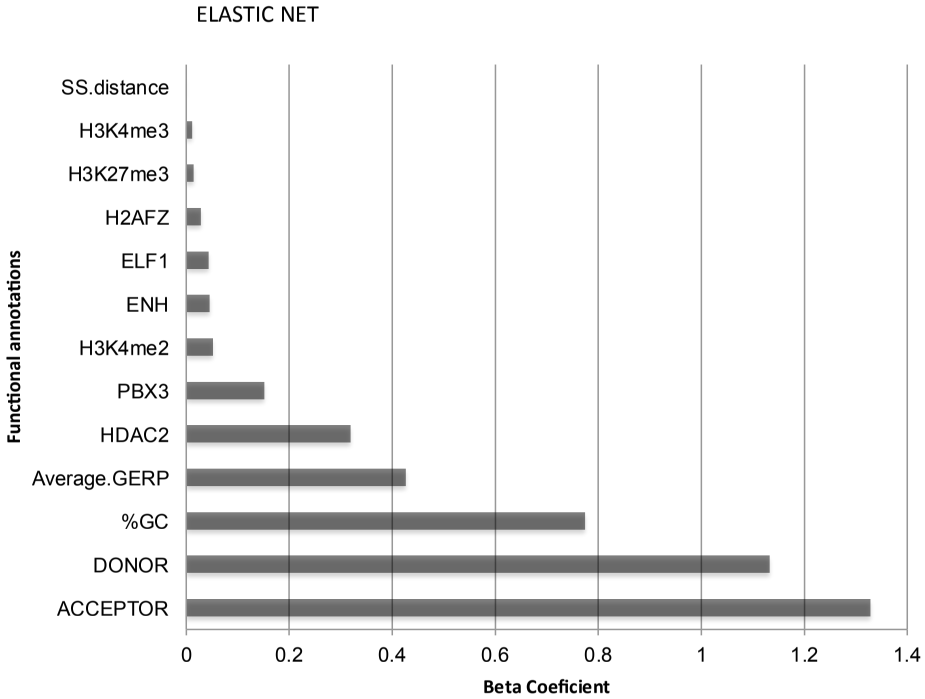


Annotation Legend: See Gagliano et al. for further details.

Nonsynonymous= Nonsynonymous SNP
GTEx_eQTLs= cis eQTL from the GTEx Project
TFBS= Transcription factor binding site
DNase_I= DNaseI hypersensitive site
H3K4Me3= H4K4Me3 histone modification
UCSC_Genes= UCSC Gene
UK_Brain_eQTLs= cis eQTL from the UK Brain Consortium
H3K27Ac= H3K27Ac histone modification
PhyloP= PhyloP conservation score
Splice= +/-5 base pairs from a splice site
PhastCons= PhastCons conservation score
H3K4Me1= H3K4Me1 histone modification
miRNA= microRNA target as defined by TargetScan
Gencode_Txnstart= Transcription start site as defined by Gencode

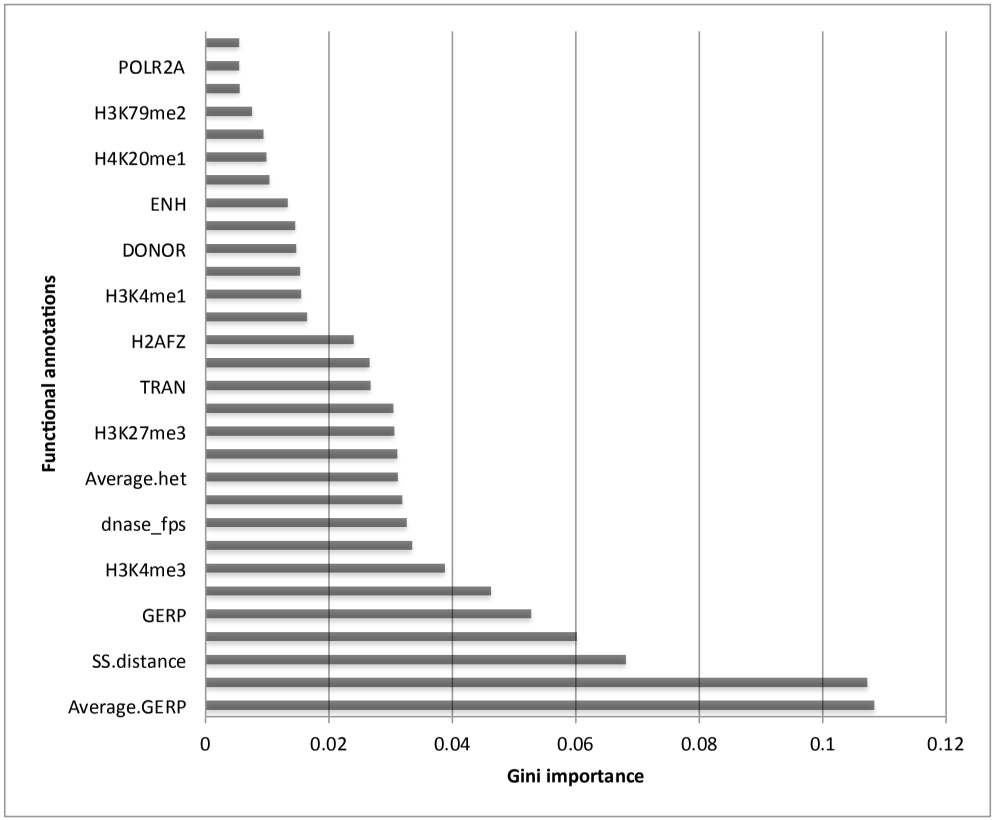
These models are based on the following classifier: non-exonic HGMD and non-exonic control variants within 1KB of the HGMD variant. The annotations from Ritchie et al. were used.

ACCEPTOR	1.328016
DONOR	1.13231198
%GC	0.77416304
Average.GERP	0.42610847
HDAC2	0.31861927
PBX3	0.15134738
H3K4me2	0.05150363
ENH	0.04571758
ELF1	0.04362891
H2AFZ	0.0286725
H3K27me3	0.01442794
H3K4me3	0.01127367
SS.distance	7.46E-07



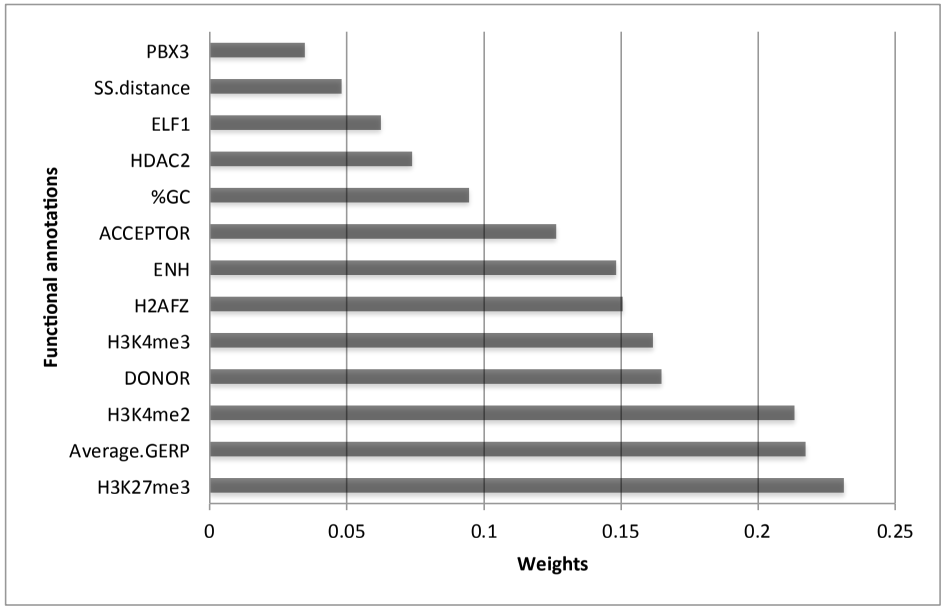
Average.GERP	0.10839647
TSS.distance	0.10727704
SS.distance	0.06808475
INTRON	0.0601164
GERP	0.05273117
H3K36me3	0.04625738
H3K4me3	0.03874786
DNase	0.03345831
dnase_fps	0.03257359
repeat.	0.03186024
Average.het	0.03112989
H3K4me2	0.0310369
H3K27me3	0.03055748
Average.DAF	0.03039181
TRAN	0.02670498
%GC	0.02655676
H2AFZ	0.02403689
FAIRE	0.01643756
H3K4me1	0.01545784
TSS	0.01528586
DONOR	0.01467283
H3K27ac	0.01451522
ENH	0.01331986
H3K9ac	0.0103722
H4K20me1	0.00982251
REP	0.00941113
H3K79me2	0.00756666
bound_motifs	0.00549367
POLR2A	0.00542883
cpg_island	0.00542017

RANDOM FOREST --> too many non-zero annots for a graph (124). The top 30 are displayed here



H3K27me3	0.23132127
Average.GERP	0.21730765
H3K4me2	0.21324897
DONOR	0.16474043
H3K4me3	0.16167312
H2AFZ	0.15065681
ENH	0.14814975
ACCEPTOR	0.12629488
%GC	0.0945848
HDAC2	0.07381445
ELF1	0.06236036
SS.distance	0.04800924
PBX3	0.03475798

SUPPORT VECTOR MACHINE



Annotation Legend: See Ritchie et al. for further details.

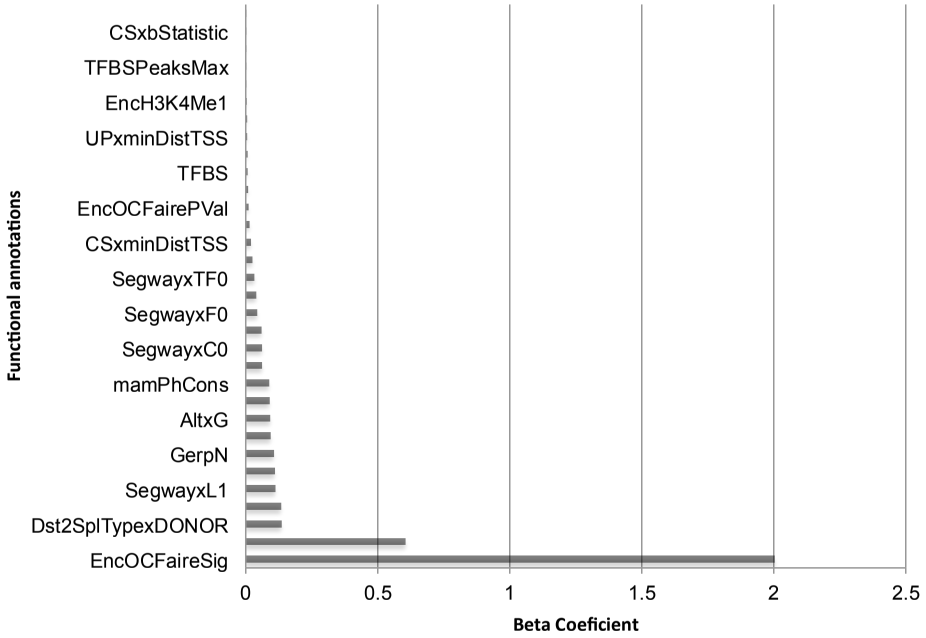
ACCEPTOR= acceptor splice site
Average.DAF= mean derived allele frequency of variants in 1kb flanking region
Average.GERP= mean GERP score of 100bp flanking region
Average.het= mean heterozygosity of 1kb flanking region
DNase= DNase1-seq peak
DONOR= donor splice site
ELF1= ELF1 Transcription Factor ChIP-seq peaks
ENH= predicted enhancer segment
FAIRE= FAIRE-seq peak
GERP= GERP score at the variant locus
H2AFZ= H2AFZ Histone modification ChIP-seq peaks
H3K27ac= H3K27ac Histone modification ChIP-seq peaks
H3K27me3= H3K27me3 Histone modification ChIP-seq peaks
H3K36me3= H3K36me3 Histone modification ChIP-seq peaks
H3K4me1= H3K4me1 Histone modification ChIP-seq peaks
H3K4me2= H3K4me2 Histone modification ChIP-seq peaks
H3K4me3= H3K4me4 Histone modification ChIP-seq peaks
H3K79me2= H3K79me2 Histone modification ChIP-seq peaks
H3K9ac= H3K9ac Histone modification ChIP-seq peaks
H4K20me1= H3K20me1 Histone modification ChIP-seq peaks
HDAC2= HDAC2 Transcription Factor ChIP-seq peaks
INTRON= intronic region
PBX3= PBX3 Transcription Factor ChIP-seq peaks
POLR2A= POLR2A Transcription Factor ChIP-seq peaks
REP= predicted repressed sequence
SS.distance= distance to the nearest splice site
TRAN= predicted transcribed segment
TSS= predicted promoter segment
TSS.distance= distance to the nearest TSS

%GC= GC content of 100bp flanking region
bound_motifs= bound transcription factor motifs
cpg_island= Predicted CpG island
dnase_fps= DNase1-seq footprint
repeat.= annotated repeat element

These models are based on the following classifier: non-exonic HGMD and non-exonic control variants within 1KB of the HGMD variant. The annotations from Kircher et al. were used.

EncOCFaireSig	2.00275077	ELASTIC NET
EncOCpolIISig	0.60623754	
Dst2SplTypexDONOR	0.13618017	
DNxpriPhyloP	0.13403555	
SegwayxL1	0.11259132	
ConsequencexCS	0.1105286	
GerpN	0.1065717	
minDistTSE	0.09535823	
AltxG	0.09345319	
priPhCons	0.09003158	
mamPhCons	0.08869554	
Dst2SplTypexACCEPTOR	0.06263028	
SegwayxC0	0.06199754	
ConsequencexUP	0.06007579	
SegwayxF0	0.04442006	
UPxGerpN	0.0397594	
SegwayxTF0	0.03347192	
CxG	0.0258295	
CSxminDistTSS	0.01964801	
UPxminDistTSE	0.01457942	
EncOCFairePVal	0.01010597	
EncOCCombPVal	0.00856037	
TFBS	0.00690565	
EncOCDNasePVal	0.00642198	
UPxminDistTSS	0.00577863	
EncOCctcfPVal	0.0053397	
EncH3K4Me1	0.00402159	

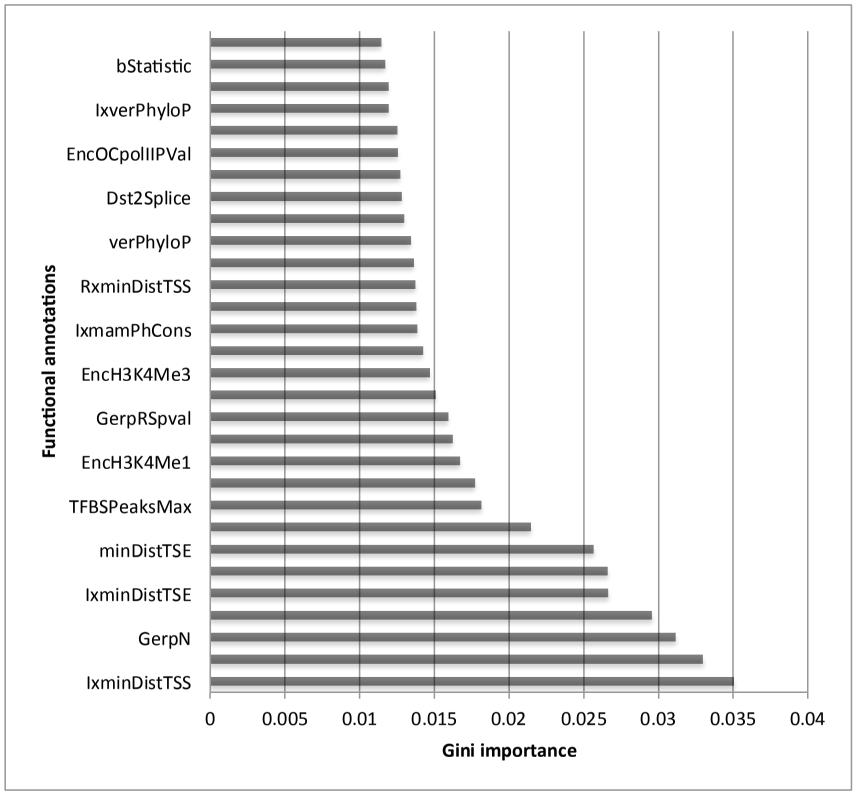
EncExp	0.00118442
TFBSPeaksMax	0.00062729
TFBSPeaks	0.00024418
CSxbStatistic	0.00014575
UPxbStatistic	0.00011621



IxminDistTSS	0.03505276
ConsequencexI	0.03298746
GerpN	0.03114855
minDistTSS	0.02957615
IxminDistTSE	0.02664504
IxGerpN	0.02660732
minDistTSE	0.02567448
IxpriPhCons	0.02146826
TFBSPeaksMax	0.01816294
UPxminDistTSE	0.01772384
EncH3K4Me1	0.01672854
GerpRS	0.01622944
GerpRSpval	0.01592789
mamPhCons	0.01510546
EncH3K4Me3	0.01470354
Dst2SplTypexUD	0.01424108
IxmamPhCons	0.01386881
verPhCons	0.01378669
RxminDistTSS	0.01374052
EncOCctcfSig	0.01362472
verPhyloP	0.01343313
IxbStatistic	0.01299017
Dst2Splice	0.01282226
UPxGerpN	0.01271164
EncOCpolIIPVal	0.01256621
priPhyloP	0.01252237
IxverPhyloP	0.01195523
EncOCDNasePVal	0.01193051
bStatistic	0.01172126
GC	0.01144353

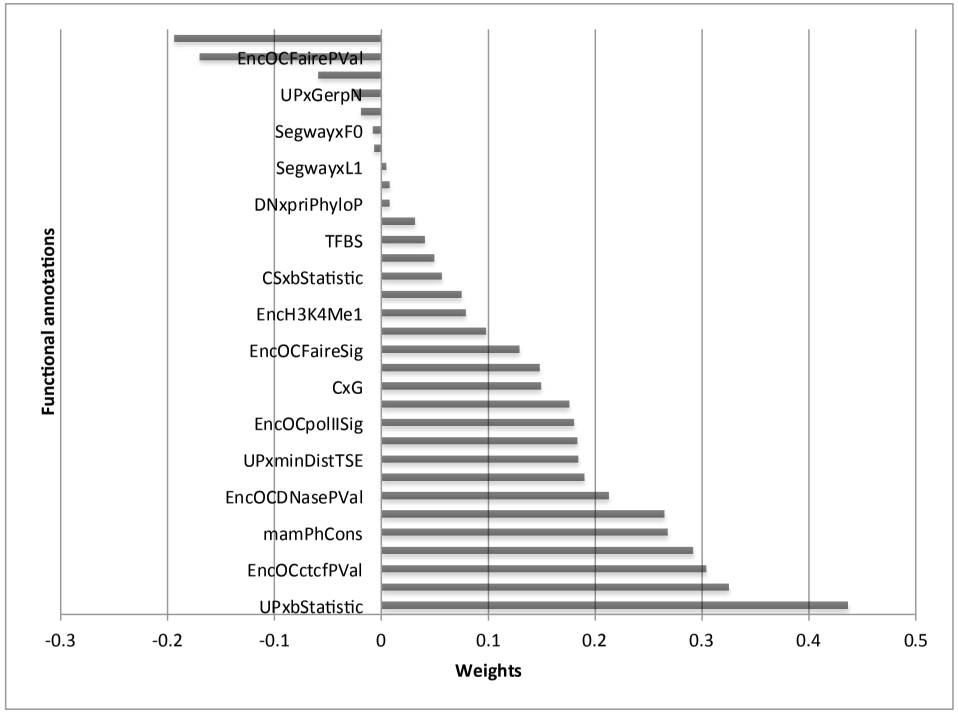
RANDOM FOREST

--> too many non-zero annots for a graph (175). The top 30 are displayed here.



UPxbStatistic	0.43664711	SUPPORT VECTOR MACHINE
TFBSPeaksMax	0.32498967	
EncOCctcfPVal	0.30378731	
Dst2SplTypexACCEPTOR	0.29170818	
mamPhCons	0.26782301	
SegwayxTF0	0.2649652	
EncOCDNasePVal	0.21291776	
ConsequencexUP	0.18995155	
UPxminDistTSE	0.18435745	
EncExp	0.1832309	
EncOCpolIISig	0.18029964	
Dst2SplTypexDONOR	0.17574117	
CxG	0.14940578	
EncOCCombPVal	0.14830761	
EncOCFaireSig	0.12933384	
SegwayxC0	0.09815048	
EncH3K4Me1	0.07896724	
CSxminDistTSS	0.07502759	
CSxbStatistic	0.05672659	
GerpN	0.04950504	
TFBS	0.04091979	
minDistTSE	0.03162832	
DNxpriPhyloP	0.00788203	
priPhCons	0.0078501	
SegwayxL1	0.0047192	
AltxG	-0.0067476	
SegwayxF0	-0.0078445	
ConsequencexCS	-0.0189846	

UPxGerpN	-0.0258417
TFBSPeaks	-0.0589407
EncOCFairePVal	-0.169988
UPxminDistTSS	-0.1939936



Annotation Legend: See Kircher et al. for further details.

AltxG= interaction between Observed allele and new amino acid glycine
CSxbStatistic= interaction between canonical splice and Background selection score
CSxminDistTSS= interaction between canonical splice and Distance to closest Transcribed Sequence Start (TSS)
ConsequencexCS= canonical splice
ConsequencexI= intronic
ConsequencexUP= upstream
CxG= interaction between previous amino acid cysteine and new amino acid glycine
DNxpriPhyloP= interaction between downstream and Primate PhyloP score (excl. human)
Dst2SplTypexACCEPTOR= Closest splice site is ACCEPTOR
Dst2SplTypexDONOR= Closest splice site is DONOR
Dst2SplTypexUD= Closest splice site is undefined
Dst2Splice= Distance to splice site in 20bp; positive: exonic, negative: intronic
EncExp= Maximum ENCODE expression value
EncH3K4Me1= Maximum ENCODE H3K4 methylation level
EncH3K4Me3= Maximum ENCODE H3K4tri methylation level
EncOCCombPVal= ENCODE combined p-Value (PHRED-scale) of Faire, Dnase, polII, CTCF, Myc evidence for open chromatin
EncOCDNasePVal= p-Value (PHRED-scale) of Dnase evidence for open chromatin
EncOCFairePVal= p-Value (PHRED-scale) of Faire evidence for open chromatin
EncOCFaireSig= Peak signal for Faire evidence of open chromatin
EncOCctcfPVal= p-Value (PHRED-scale) of CTCF evidence for open chromatin
EncOCctcfSig= Peak signal for CTCF evidence of open chromatin
EncOCpolIIIPVal= p-Value (PHRED-scale) of polII evidence for open chromatin
EncOCpolIIISig= Peak signal for polII evidence of open chromatin
GC= Percent GC in a window of +/- 75bp
GerpN= Neutral evolution score defined by GERP++
GerpRS= Gerp element score
GerpRSpval= Gerp element p-Value
IxGerpN= interaction between intronic Neutral evolution score defined by GERP++
IxbStatistic= interaction between intronic and Background selection score

B Protein kinase cAMP-dependent regulatory type II beta (PRKAR2B) gene variants in antipsychotic-induced weight gain

Content in this chapter is published in **Gagliano SA**, Tiwari AK, Freeman N, Lieberman JA, Meltzer HY, et al. (2014) Protein kinase cAMP-dependent regulatory type II beta (PRKAR2B) gene variants in antipsychotic-induced weight gain. Hum Psychopharmacol 29: 330-335.

HUMAN PSYCHOPHARMACOLOGY
Hum. Psychopharmacol Clin Exp 2014; **29**: 330–335.
Published online 16 April 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/hup.2407

Protein kinase cAMP-dependent regulatory type II beta (*PRKAR2B*) gene variants in antipsychotic-induced weight gain

Sarah A. Gagliano[1,2], Arun K. Tiwari[1], Natalie Freeman[1], Jeffrey A. Lieberman[3,6], Herbert Y. Meltzer[4], James L. Kennedy[1,2,5], Jo Knight[1,2,5]* and Daniel J. Müller[1,2,5]

[1]*Neurogenetics Section, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada*
[2]*Institute of Medical Science, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada*
[3]*Department of Psychiatry, College of Physicians and Surgeons, Columbia University, New York, NY, USA*
[4]*Department of Psychiatry and Behavioral Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA*
[5]*Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada*
[6]*The New York State Psychiatric Institute, New York, NY, USA*

Objective Antipsychotics are effective in treating schizophrenia symptoms. However, the use of clozapine and olanzapine in particular are associated with significant weight gain. Mouse and human studies suggest that the protein kinase cAMP-dependent regulatory type II beta (*PRKAR2B*) gene may be involved in energy metabolism, and there is evidence that it is associated with clozapine's effects on triglyceride levels. We aimed at assessing *PRKAR2B*'s role in antipsychotic-induced weight gain in schizophrenia patients.
Methods DNA samples from adult schizophrenia or schizoaffective disorder patients of mixed ancestry were genotyped, and weight gain was assessed. We analyzed 16 tag single-nucleotide polymorphisms across the *PRKAR2B* gene in a Caucasian subset treated either with clozapine or olanzapine ($N = 99$). Linear regression based on an additive model was performed with the inclusion of relevant covariates.
Results Normalized per cent weight change was analyzed, revealing that patients with the minor allele at rs9656135 had a mean weight increase of 4.1%, whereas patients without this allele had an increase of 3.4%. This association is not significant after correcting for multiple testing.
Conclusions Because of limited power, *PRKAR2B*'s role in antipsychotic-induced weight gain is unclear, but biological evidence suggests that *PRKAR2B* may be involved. Further research in larger sample sizes is warranted. Copyright © 2014 John Wiley & Sons, Ltd.

KEY WORDS—*PRKAR2B*; antipsychotic-induced weight gain; schizophrenia; pharmacogenetics; polymorphisms

INTRODUCTION

The use of antipsychotics, such as clozapine and olanzapine, has been effective in treating schizophrenia patients but is often associated with severe metabolic side effects, particularly significant weight gain. Weight gain itself is a serious health concern due to comorbidities such as cardiovascular disease and type II diabetes (Reynolds, 2012). With regard to the genetic component of antipsychotic-induced weight gain (AIWG), there is a heritable component. In a monozygotic twin and sibling pair study, Gebhardt *et al.* (2010) estimated the contribution of genetic factors in AIWG to be 60–80%. Additionally, numerous genes, some of which have been replicated, have been shown to be associated with AIWG (Müller and Kennedy, 2006, Lett *et al.*, 2012). A recent example of a replicated finding is with a locus near the melanocortin 4 receptor gene (Malhotra *et al.*, 2012). Other replicated findings involve variants in leptin genes and others in the promoter of the 5-hydroxytryptamine (serotonin) receptor 2C gene (Reynolds, 2012). In this study, we investigate another likely candidate gene to be involved in AIWG, the protein kinase cAMP-dependent regulatory type II beta (*PRKAR2B*) gene. Other protein kinase genes, particularly the subunits of AMP-activated protein kinase, have been previously studied in AIWG (Jassim *et al.*, 2011; Souza *et al.*, 2012). However, *PRKAR2B* has so far only been investigated in one study that looked at phenotypic outcomes related to AIWG. A variant in this gene was shown to be associated with clozapine's effects on triglyceride levels in a genome-wide pharmacogenomics study of metabolic side effects using participants from the Clinical Antipsychotic Trial of Intervention Effectiveness (Adkins *et al.*, 2011). *PRKAR2B* codes for one of the several regulatory subunits of cAMP-dependent protein

*Correspondence to: J. Knight, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. Tel: +1 416 535 8501 E-mail: jo.knight@camh.ca

Received 30 July 2013
Accepted 13 March 2014

kinase. It is expressed in all tissue, including the hypothalamus, which could suggest a role that is linked to appetite.

Furthermore, the *PRKAR2B* gene is a plausible candidate for being implicated in antipsychotic-induced metabolic outcomes as supported by animal studies. For instance, with regard to the metabolic phenotype, Czyzyk *et al*. (2008) showed that disruption of the RII-beta subunit (coded by *PRKAR2B*) reverses elevated body weight, hyperphagia, and obesity of agouti lethal yellow mice. In that paper, Czyzyk *et al*. (2008) also discuss that *PRKAR2B* may be one of the cAMP effector molecules working downstream of the melanocortin 4 receptor gene. As for being implicated in antipsychotic effects, Adams *et al*. (1997) found that the cataleptic response to haloperidol is blocked in mice with a targeted disruption in the RII-beta subunit. In addition, mice lacking this regulatory subunit exhibit a 10% reduction in body weight and a 50% decrease in white adipose tissue and are resistant to diet-induced obesity and hyperglycemia (Adams *et al*., 1997). Altogether, these previous studies support the hypothesis that variants of the *PRKAR2B* gene may be implicated in AIWG. Thus, we aimed at studying the contribution of *PRKAR2B* to AIWG in a sample of schizophrenia or schizoaffective disorder patients.

METHODS

Samples

Patients were recruited from four sites. Within each site, patients were from various ethnic backgrounds. For the first three sites, 226 clinically diagnosed schizophrenia or schizoaffective disorder patients were recruited and are summarized in the succeeding texts. In the first sample (DJM-1), schizophrenia patients ($N = 99$; Berlin) were given different antipsychotics and assessed up to 6 weeks. Patients ($N = 77$) from the second sample (HYM; Ohio) were treated with clozapine for up to 6 weeks, and patients ($N = 55$) from the third (JAL; New York) were treated with clozapine, haloperidol, olanzapine, or risperidone for up to 14 weeks. Demographic details on these subjects have been previously described (Tiwari *et al*., 2013), but refer to Table S1 for a summary. For the fourth sample, 21 patients were recruited from an ongoing study at the Centre for Addiction and Mental Health in Toronto (DJM-2; Toronto) study. Patients were included when either starting or switching to a new second-generation antipsychotic (clozapine, olanzapine, risperidone, or quetiapine) and were prospectively assessed for AIWG and treatment response for a minimum of 6 months. All patients were assessed for research diagnosis and comorbid conditions using the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (First *et al*., 1997) and symptom severity using the Positive and Negative Syndrome Scale (Kay *et al*., 1987). Metabolic assessments included weight at baseline, week two and week six. Exclusion criteria included severe medical conditions (e.g., hepatitis C, HIV, and diabetes), substance abuse/dependence, significant mental retardation, or severe personality disorder. Ethylenediaminetetraacetic acid tubes with a minimum of 10 ml venous blood were drawn from each subject. Approval from the institutional ethics committees and informed consent were obtained for all patients.

Genotyping

A total of 16 tag single-nucleotide polymorphisms (SNPs) were selected in the *PRKAR2B* gene for association with AIWG. Additional genotyped SNPs were available for quality control procedures. DNA samples were genotyped using the GoldenGate Genotyping Assay (Illumina Inc. San Diego, CA, USA) as per the manufacturers' protocol (Fan *et al*., 2006) at The Centre for Applied Genomics (Toronto, Ontario, Canada). Briefly, SNPs were uploaded to Illumina's Assay Design Tool (http://www.illumina.com/) for probe design resulting in a custom panel (GS0013427-OPA) of 384 SNPs. A total of 5 µl of 50 ng/µl in 10 mM Tris–HCl pH 8.0, 1 mM ethylenediaminetetraacetic acid of genomic DNA underwent an allele-specific oligonucleotide hybridization followed by extension and ligation. A universal polymerase chain reaction step for all 384 loci followed with primers labeled with either Cy3 (primer 1) or Cy2 (primer 2). The amplified products were then hybridized to GoldenGate Genotyping Universal-32, 384-plex beadchips, and scanned using the Illumina iScan (Illumina Inc.). The resulting data was analyzed with GenomeStudio v2011 using the default parameters. SNPs were clustered on the sample dataset and manually inspected. SNPs were discarded if call rates were less than 90%. A total of seven SNPs failed, leading to 377 SNPs of good quality for further use.

Genetic data quality control

Quality control procedures and association analyses were performed using PLINK (version 1.07, http://pngu.mgh.harvard.edu/~purcell/plink/) (Purcell *et al*., 2007). Plots for call rate distributions and ancestry mapping based on principal component analysis (PCA) were created using *R* (http://cran.r-project.org/) (*R*, 2008). Quality control measures were applied to both

individuals and markers. Duplicate samples, individuals with less than 95% call rates, and individuals with outlying heterozygosity were removed from the analysis. As for the quality control measures applied to the markers, standard thresholds were chosen: rare variants defined as markers with a minor allele frequency (MAF) of less than 1%, and markers with a missing data rate of greater than 5% were excluded. Thresholds for other quality control measures, such as Hardy–Weinberg equilibrium (HWE), were decided on the basis of the number of markers. The HWE threshold of 0.0001 was determined by dividing the alpha value of 0.05 into the total number of markers available on the array ($N = 377$). None of the *PRKAR2B* markers failed HWE (see the first set of columns in Table 1 for summary statistics for the 16 *PRKAR2B* SNPs).

Statistical analysis

Considering the samples were of mixed ethnicity, an analysis option would have been to conduct multiple association studies (each of which analyze a single ethnicity) and then combine the results in a meta-analysis. However, this option was not feasible because of small sizes of some samples. Instead, the association analysis was performed on the largest ethnic subset, Caucasians. Those Caucasians treated with either clozapine or olanzapine ($N = 99$) were included (refer to Table 2 for the demographics). Those individuals who self-reported as Caucasian and also clustered with the HapMap (Frazer *et al*., 2007) CEU population after PCA using the independent genotyped markers ($N = 123$) available from all samples were considered to be Caucasian ($N = 99$). The rationale behind choosing the subset of individuals treated with either clozapine or olanzapine was that in literature reviews of AIWG, the highest weight gain is typically observed in individuals taking those medications (e.g., Lett *et al*., 2012). The trade-off involved in choosing to use this subset with less noise is that a smaller sample size also results.

Table 1. Summary statistics and regression results for *PRKAR2B* single-nucleotide polymorphisms (SNPs)

	Summary statistics		Regression results	
PRKAR2B SNP	Minor allele frequency	Hardy–Weinberg equilibrium *p*-value	Beta	Uncorrected *p*-value
rs1544582	0.44 (G)	0.23	−0.15	0.17
rs2237648	0.32 (A)	0.65	0.069	0.57
rs2237649	0.43 (A)	0.36	−0.15	0.23
rs11766415	0.40 (G)	0.53	−0.09	0.41
rs2536504	0.34 (T)	1	−0.02	0.88
rs2536505	0.12 (C)	1	0.37	0.045
rs6960842	0.45 (G)	0.84	−0.05	0.66
rs2536508	0.27 (G)	0.45	−0.006	0.96
rs13311274	0.38 (C)	0.47	0.05	0.69
rs17153823	0.13 (G)	1	−0.12	0.50
rs13224682	0.07 (G)	0.39	0.008	0.97
rs9656135	0.07 (T)	1	0.72	0.0015
rs2302453	0.45 (A)	0.42	−0.086	0.44
rs12705406	0.16 (A)	0.70	−0.09	0.58
rs257376	0.39 (G)	0.83	−0.047	0.69
rs257378	0.22 (G)	0.39	0.086	0.57

Table 2. Demographics of Caucasian subset used in the analysis

Characteristic	Median (range)
Sex	55 women 44 men
Age (years)	34 (18–65)
Baseline weight (kg)	78.20 (49.50–185.40)
Treatment duration (weeks)[b]	6 (1–14)
Per cent weight change (%)	2.91 (−7.59 to 26.85)
Normalized per cent weight change (%)[a]	3.49 (1.00–5.94)

[a]Used as the outcome variable in the linear regression association analysis where genotype at each locus is the predictor variable.
[b]For most patients (87%), the treatment duration was 6 weeks.

In PLINK, linear regression was performed on the subset described with the inclusion of the following variables as covariates: baseline weight, study duration, and the first principal component from the PCA on the subset of individuals analyzed. Per cent weight change rather than absolute weight change was used as the outcome variable since the US Food and Drug Administration defines clinically significant weight gain using a percentage (≥7% of baseline weight) on US package inserts for these antipsychotics (Casey *et al*., 2004). Because linear regression assumes that the continuous variable follows a normal distribution, the Shapiro–Wilk normality test in *R* was applied, and the null hypothesis that the distribution is normal was rejected ($p = 3.1\text{e-}05$). Data were consequently normalized using a square root transformation to follow a normal distribution according to the Shapiro–Wilk normality test.

Correction for multiple testing was performed in two ways: by adjusting for the *PRKAR2B* SNPs and also by taking into account all of the SNPs genotyped on the same array as the *PRKAR2B* SNPs. The number of independent tests was determined taking into account the linkage disequilibrium structure of the *PRKAR2B* SNPs using matrix spectral decomposition (Nyholt, 2004). Specifically, the Li and Ji method (2005) that is recommended by Nyholt was employed.

Statistical power calculation

In order to assess statistical power, calculations were performed using the Genetic Power Calculator (http://pngu.mgh.harvard.edu/~purcell/gpc/) (Purcell *et al*., 2003). The calculations, on the basis of quantitative trait loci for singletons, were conducted using the following

DOI: 10.1002/hup

assumptions: The marker allele is in perfect linkage disequilibrium with the high risk allele, the quantitative trait accounts for 5% of the total variance under an additive model, and the MAF of these alleles is 0.3. The first two assumptions make the estimated sample size conservative because the frequency and percentage of the heritability accounted for by the quantitative trait may be lower than specified. The MAF of 0.3 was chosen as it is the average MAF for the *PRKAR2B* SNPs analyzed.

In silico functional analysis

An *in silico* functional analysis was performed in HaploReg v2 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward and Kellis, 2012). HaploReg is a resource that incorporates data from the ENCODE Project, Roadmap Epigenome Mapping Consortium, and also expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression eQTL Browser in order to explore the annotations of noncoding SNPs.

RESULTS

There were 99 individuals who belonged to the Caucasian subset being treated with either clozapine or olanzapine: 22 from DJM-1, 50 from HYM, 8 from JAL, and 19 from DJM-2. In the association analysis, one of the SNPs in *PRKAR2B* (rs9656135) was significantly associated with AIWG before correcting for multiple testing (uncorrected $p = 0.0015$, odds ratio = 2.05). There were no significant associations between the genotype and any of the covariates. At SNP rs9656135, the predicted values from the fitted model $y = 0.72x - 0.01\beta1 + 0.02\beta2 - 0.29\beta3 + 4.01$ where y is the normalized per cent weight change, x is the genotype, $\beta1$ is the baseline weight, $\beta2$ is the study duration, and $\beta3$ is the first principal component were plotted (Figure 1). (See the final set of columns in Table 1 for the regression results for the 16 *PRKAR2B* SNPs.) The MAF for this marker is 7%. A closer inspection at the number of individuals per genotype at this marker (Table 3) showed that there were no individuals homozygous for the minor allele (T), and thus, the linear regression was only comparing the heterozygotes with those homozygous for the major allele (C). With the lack of homozygotes for the minor allele, it cannot be determined whether the trend seen between SNP rs9656135 and per cent weight change follows an allelic model.

Single-nucleotide polymorphism rs9656135 is in close proximity to the SNP (rs13224682) found in Adkins *et al*. (2011) to be associated with clozapine's effects on triglyceride levels; however, these SNPs are not in linkage disequilibrium with each other ($r^2 = 0$) (Figure 2).

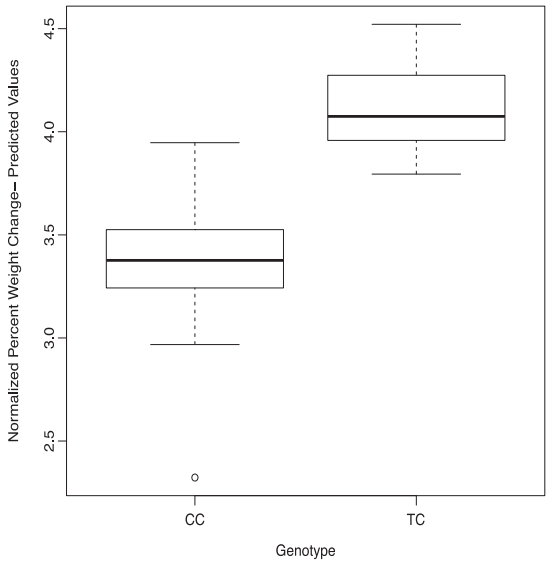


Figure 1. Box plots of normalized weight change distributions for the various genotypes at single-nucleotide polymorphism rs9656135. The black line in each box represents the median. The lower line of the box is the 25% quartile, and the upper line is the 75% quartile. The lower and upper whiskers represent the minimum and maximum values, respectively, but these do not include outliers. Outliers, represented as isolated circles drawn outside of the boxes, are those values that are either 1.5 times less than or greater than the interquartile range (the difference between the 75% and 25% quartiles)

Table 3. Genotype counts summary for the most significant single-nucleotide polymorphism from the association analysis, rs9656135

Genotype	T/T	T/C	C/C
Counts	0	15	84
Frequency	0	0.15	0.85
Normalized mean per cent weight gain (%)	N/A	4.1	3.4

We assessed the significance of the association by adjusting for multiple comparisons based on two strategies. One only accounted for the SNPs in the *PRKAR2B* gene, whereas the other accounted for all of the SNPs genotyped on the array that were selected as possible candidates for AIWG. According to the method in Li and Ji (2005), there are nine effective tests. Implementing the same procedure, but taking into account all of the SNPs ($N = 377$) successfully genotyped on the same array as the *PRKAR2B* SNPs, there are 176 effective tests. The association between SNP rs9656135 and AIWG remained significant ($p = 0.01$) when correcting for just the SNPs in *PRKAR2B*. Using all of the hypothesized SNPs that were genotyped on the array, SNP

DOI: 10.1002/hup

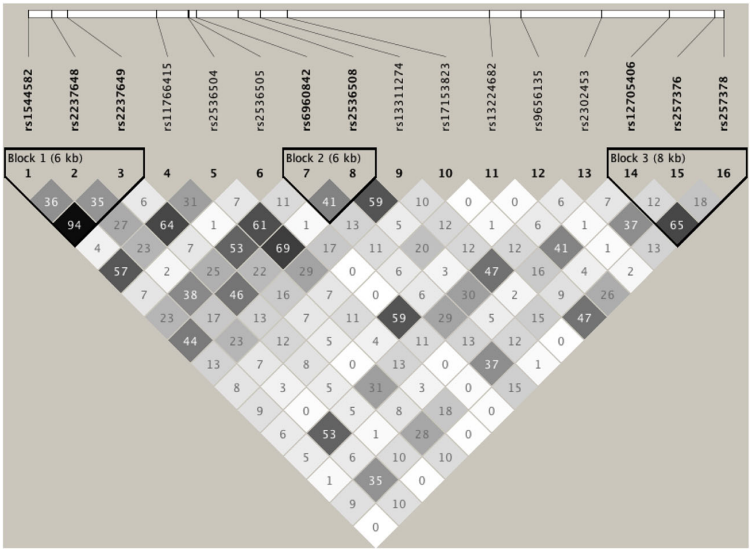


Figure 2. Linkage disequilibrium plot displaying r^2 values for the 16 *PRKAR2B* single-nucleotide polymorphisms analyzed. The plot was constructed in Haploview version 4.2 (Barrett *et al*., 2005)

rs9656135 is no longer statistically significant ($p = 0.26$) when adjusting for the 176 effective tests.

Power analyses revealed that our study was underpowered, requiring 153 individuals instead of 99 to achieve 80% power (assuming MAF 3%, and 5% of the variance accounted for).

PLINK was used to calculate the inflation factor based on the median chi-squared value for the linear regression model to ensure that the sample did not contain admixture. An inflation factor of one suggests that there is no stratification in the sample, whereas values greater than one indicate stratification effects. Using the additional typed markers (total $N = 377$), the inflation factor in the Caucasian subset in the clozapine and olanzapine group resulted in one with the inclusion of the first principal component in the analysis, suggesting that that principal component effectively corrected for stratification.

DISCUSSION

We investigated the role of the *PRKAR2B* gene in AIWG in a sample of Caucasian schizophrenia patients being treated with clozapine or olanzapine ($N = 99$). We tested for associations between the 16 genotyped tag SNPs in this gene and per cent weight gain. One SNP (rs9656135) showed an association with AIWG. The odds ratio was 2.05. This SNP remained statistically significant after adjusting for multiple testing by taking into account only the SNPs in *PRKAR2B*; however, it was no longer significant when adjusting for all of the SNPs genotyped on the same array used for ancillary analyses outlined earlier and for the investigation of other AIWG candidate genes.

Single-nucleotide polymorphism rs9656135 is an intronic SNP, and there is no currently available functional evidence to support the role of this SNP in AIWG. There is a possibility that SNP rs9656135 may tag a functional variant, which has a more significant association signal with AIWG. Inputting rs9656135 into HaploReg (Ward and Kellis, 2012) showed that this SNP does not overlap with any DNase I hypersensitive sites, binding sites for proteins, promoter or enhancer annotations, or eQTL. However, using a lower linkage disequilibrium (LD) threshold of $r^2 = 0.6$, rather than the default 0.8 in HaploReg, shows that there are a large number of SNPs with LD between $r^2 = 0.6$ and 0.8 that show extensive enhancer histone and promoter marks and a few with DNase protection and multiple proteins bound as well.

The other investigated SNPs yielded no significant results prior to correction for multiple testing, and overall, our study suggests that the *PRKAR2B* gene may not play a major role in AIWG. As for the rationale to investigate the *PRKAR2B* gene in AIWG, evidence was provided by animal studies suggesting a role in energy metabolism. For example, *PRKAR2B* mouse knockouts are lean, with increased activity and resting metabolic rate. These mice are protected from diet-induced obesity and fatty livers (Cummings *et al*., 1996). In addition, one variant in *PRKAR2B* was found to be significantly associated

DOI: 10.1002/hup

with trigylceride levels, a variable related to AIWG (Adkins *et al*., 2011). This SNP was not associated with AIWG in our study. However, the study by Adkins *et al*. (2011) corrected for multiple testing using a false discovery rate approach, which gives rise to a higher number of false positive results. Additionally, the study included principal components into their regression model from a PCA that was performed on an admixed sample, and it is not clear if PCA is able to adjust for such extensive population stratification. Thus, *PRKAR2B* association findings of Adkins *et al*. are difficult to interpret. A limitation of our study is limited power because of a small sample size. Additional limitations involve the heterogeneity of the sample with regard to potential confounding variables that may affect weight gain but were not included in the model, such as calorie intake, inpatient versus outpatient status, concomitant therapy, and study duration. In light of these described limitations, the *PRKAR2B* gene's involvement in AIWG cannot be conclusively determined at the present time.

Larger samples are required for further analysis; however, *PRKAR2B* remains a biologically plausible candidate as a contributor to AIWG. Association analysis approaches extending beyond genes to investigate biological pathways could be conducted in the future to investigate the influence of this gene and others on AIWG.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank the following individuals at CAMH: Eva Brandl, Shannon Collinson, Vanessa Goncalves, and Clement Zai. D. J. M. would like to thank the Canadian Institutes of Health Research (CIHR operating grant: Genetics of antipsychotics-induced metabolic syndrome, MOP 89853), Brain & Behavior Research Foundation (NARSAD), CIHR Michael Smith New Investigator Salary Prize for Research in Schizophrenia, Ontario Mental Health Foundation New Investigator Fellowship, and the Ministry of Research and Innovation of Ontario for the Early Researcher Award.

REFERENCES

Adams MR, Brandon E, Chartoff EH, *et al*. 1997. Loss of haloperidol induced gene expression and catalepsy in protein kinase A-deficient mice. *Proc Natl Acad Sci U S A* **94**: 12157–12161.

Adkins MR, Aberg K, McClay JL, *et al*. 2011. Genomewide pharmacogenetics study of metabolic side effects to antipsychotic drugs. *Mol Psychiatry* **16**: 321–332.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**(2): 263–265.

Casey DE, Haupt DW, Newcomer JW, *et al*. 2004. Antipsychotic-induced weight gain and metabolic abnormalities: implications for increased mortality in patients with schizophrenia. *J Clin Psychiatry* **65**(Suppl 7): 4–18; quiz 19–20.

Cummings DE, Brandon EP, Planas JV, Motamed K, Idzerda RL, McKnight GS. 1996. Genetically lean mice result from targeted disruption of the RII beta subunit of protein kinase A. *Nature* **382**: 622–626.

Czyzyk TH, Sikorski MA, Yang L, McKnight GS. 2008. Disruption of the RIIbeta subunit of PKA reverses the obesity syndrome of agouti lethal yellow mice. *Proc Natl Acad Sci U S A* **8**: 276–281.

Fan JB, Chee MS, Gunderson KL. 2006. Highly parallel genomic assays. *Nat Rev Genet* **7**(8):632–644.

First MB, Gibbon M, Spitzer RL, Williams JBW, Benjamin LS. 1997. Structured Clinical Interview for DSM-IV Axis II Personality Disorders, (SCID-II). American Psychiatric Press, Inc.: Washington, D.C..

Frazer KA, Ballinger DG, Cox DR, *et al*. 2007. A second generation human haplotype map of over 3.1 million SNPs. [Research Support, N.I.H., Extramural Research Support, Non-US Gov't]. *Nature* **449**(7164): 851–861. DOI: 10.1038/nature06258

Gebhardt S Theisen FM, Haberhausen M, *et al*. 2010. Body weight gain induced by atypical antipsychotics: an extension of the monozygotic twin and sib pair study. *J Clin Pharm Ther* **35**: 207–211.

Jassim G, Fern J, Theisen FM, *et al*. 2011. Association study of energy homeostasis genes and antipsychotic-induced weight gain in patients with schizophrenia. *Pharmacopsychiatry* **44**: 15–20.

Kay SR, Fiszbein A, Opler LA. 1987. The positive and negative symptom scale (PANSS) for schizophrenia. *Schizophr Bull* **13**(2): 261–276.

Lett TA, Wallace TJ, Chowdhury NI, Tiwari AK, Kennedy JL, Müller DJ. 2012. Pharmacogenetics of antipsychotic-induced weight gain: review and clinical implications. *Mol Psychiatry* **17**: 242–266.

Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**: 221–227.

Malhotra AK, Correll CU, Chowdhury NI, *et al*. 2012. Association between common variants near the melanocortin 4 receptor gene and severe antipsychotic drug-induced weight gain. *Arch Gen Psychiatry* **69**(9): 904–912.

Müller DJ, Kennedy JL. 2006. Genetics of antipsychotic treatment emergent weight gain in schizophrenia. *Pharmacogenomics* **7**(6): 863–887

Nyholt DR. 2004. A simple correction for multiple testing for SNPs in linkage disequilibrium with each other. *Am J Hum Genet* **74**(4): 765–769.

Purcell S, Neale B, Todd-Brown K, *et al*. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**(3): 559–575.

Purcell S, Cherny SS, Sham PC. 2003. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**(1): 149–150.

R Development Core Team. 2008. R: a language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reynolds G. 2012. Pharmacogenetic aspects of antipsychotic drug-induced weight gain—a critical review. *Clin Psychopharmacol Neurosci* **10**(2): 71–77.

Souza R, Tiwari AK, Chowdhury N, *et al*. 2012. Association study between variants of AMP-activated protein kinase catalytic and regulatory subunit genes with antipsychotic-induced weight gain. *J Psychiatr Res* **46**(4): 462–468.

Tiwari AK, Brandl EJ, Weber C, *et al*. 2013. Association of a functional polymorphism in neuropeptide Y with antipsychotic-induced weight gain in schizophrenia patients. [Comparative Study Multicenter Study Research Support, Non-US Gov't]. *J Clin Psychopharmacol* **33**(1): 11–17. DOI: 10.1097/JCP.0b013e31827d145a

Ward LD, Kellis M. 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**(D1): D930–D934.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website.

C R code for Chapter 4 A Review of Predictive Accuracy Measures that can be Applied to Models for Prioritizing Risk Variants Based on Functional Information

Additional_File_1.R

```
#Sample R Code
#MyData.txt is a text file (either space or tab-delimited) with a header. It contains a
 list of genetic variants (one variant per line) with at least the following two labelled
 columns: cls (a 0/1 binary indicator: 1=hit and 0=non-hit) and score (contains the
 prediction value).

#Receiver operator characteristic curve
pdf("ROC.pdf")
x<-read.table("MyData.txt", h=T, as.is=T)
library(ROCR)
pred<-prediction(x$score, x$cls)
perf <- performance( pred, "tpr", "fpr" )
plot(perf, lwd=5)
abline(0,1,lty=3)
dev.off()
#display area under the curve
performance(pred, "auc")
#display positive predictive values
performance(pred, "ppv")
#display negative predictive values
performance(pred, "npv")

#Histogram
require(plotrix)
x<-read.table("MyData.txt", h=T, as.is=T)
hits<-subset(x, x$cls==1)
nonhits<-subset(x, x$cls==0)
l<-list(hits$score, nonhits$score)
#adjust the start and end position and bin increments below
bins<-seq(0,1, by=0.05)
pdf("Histogram.pdf")
multhist(l, freq=F, xlab="Predicted Value", breaks=bins, col=c("black","grey"))
legend("top", title="Classifier", c("Hits", "Non-hits"), pch=c(15, 15),
 col=c("black","grey"))
dev.off()

#Box plot
pdf("Boxplot.pdf")
x<-read.table("MyData.txt", h=T, as.is=T)
hits<-subset(x, x$cls=="1")
nonhits<- subset(x, x$cls=="0")
boxplot(hits$score, nonhits$score, xlab="Classification", ylab="Prediction",
 names=c("Hit","Non-hit"), ylim=c(0,1))
dev.off()

#Violin plot
library(vioplot)
pdf("Violinplot.pdf")
x<-read.table("MyData.txt", h=T, as.is=T)
hits<-subset(x, x$cls=="1")
```

```
nonhits<- subset(x, x$cls=="0")
vioplot(hits$score, nonhits$score, names=c("Hit","Non-hit"), col="white", ylim=c(0,1))
title(xlab="Classification", ylab="Prediction")
dev.off()

#Quantile-quantile plot
pdf("Qqplot.pdf")
x<-read.table("MyData.txt", h=T, as.is=T)
hits<-subset(x, x$cls=="1")
nonhits<- subset(x, x$cls=="0")
qqplot(nonhits$score, hits$score, ylab="Hits", xlab="Non-hits", ylim=c(0,1), xlim=c(0,1))
abline(0,1, col="grey")
dev.off()

#Hypergeometric test
x<-read.table("MyData.txt", h=T, as.is=T)
hits<-subset(x, x$cls==1)
nonhits<-subset(x, x$cls==0)
res<-matrix(nrow=3,ncol=13)
row=1
col=0
BD<-length(nonhits[,1])
j<-length(hits[,1])
#prediction value bins ranging from less than 0.35 to between 0.9 and 0.95, increasing by
 increments of 0.5
for (i in seq(0.3,0.9,0.05))
{
col<-col+1
c<-length(subset(nonhits$score,nonhits$score<i+0.05 & nonhits$score>i))
a<-length(subset(hits$score, hits$score<i+0.05 & hits$score>i))
res[row,col]<-c/dim(nonhits)[1]
res[row+1,col]<-a/dim(hits)[1]
res[row+2,col]<-sum(phyper(a,j,BD-j,c, lower.tail=F))
}
#write a table to read in Excel
head<-c("p<0.35", "0.35<p<0.4", "0.4<p<0.45", "0.45<p<0.5", "0.5<p<0.55",
 "0.55<p<0.6","0.6<p<0.65", "0.65<p<0.7", "0.7<p<0.75","0.75<p<0.8", "0.8<p<0.85",
 "0.85<p<0.9","0.9<p<0.95")
table<-rbind(head, res)
write.table(table, "Hypergeometric.csv", sep=",", row.names=F, col.names=F, quote=F)
#the first row is the frequency of non-hits
#the second row is the frequency of the hits
#the third row is the hypergemoetric p-value

#Mann-Whitney U test
x<-read.table("MyData.txt", h=T, as.is=T)
nonhits<-subset(x, x$cls==0)
hits<-subset(x, x$cls==1)
wilcox.test(nonhits$score, hits$score)

#Asymptotic Generalized Cochran-Mantel-Haenszel Test
library("coin")
x<-read.table("MyData.txt", h=T, as.is=T)
nonhits<-subset(x, x$cls==0)
hits<-subset(x, x[$cls==1)
```

```
counts<-matrix(nrow=2,ncol=13)
row=1
col=0
for (i in seq(0.3,0.9,0.05))
{
col<-col+1
c<-length(subset(nonhits$score,nonhits$score<i+0.05 & nonhits$score>i))
a<-length(subset(hits$score, hits$score<i+0.05 & hits$score>i))
counts[row,col]<-c
counts[row+1,col]<-a
}
counts<-as.table(counts)
cmh_test(counts)
```

Additional_File_2.R

```
#Code for the plots in the paper: "Assessing models for genetic prediction of complex
 traits: a comparison of visualization and quantitative methods" Sarah A Gagliano, Andrew
 D Paterson, Michael E Weale and Jo Knight
#Figure 1- the confusion matrix,
#no data in this figure

#Figure 2- ROC curves
library("ROCR")
pdf("ROC-clumped-4models.pdf")
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",", h=F)
pred<-prediction(x[,5], x[,4])
perf <- performance( pred, "tpr", "fpr" )
plot(perf, lwd=5)
par(new=T)
x<-read.table("Autoimmune-testset.csv", sep=",", h=F)
pred<-prediction(x[,5], x[,4])
perf <- performance( pred, "tpr", "fpr" )
plot(perf, lwd=5, col="grey")
par(new=T)
x<-read.table("Brain-testset.csv", sep=",", h=F)
pred<-prediction(x[,5], x[,4])
perf <- performance( pred, "tpr", "fpr" )
plot(perf, lwd=5, lty=3, col="grey")
par(new=T)
x<-read.table("Nonpheno-allCat-testset.csv", sep=",", h=F)
pred<-prediction(x[,5], x[,4])
perf <- performance( pred, "tpr", "fpr" )
plot(perf, lwd=5, lty=3)
abline(0,1, lty=3)
legend("bottomright", title="GWAS hits", c("Autoimmune", "Non-phenotype specific", "Non-
 phenotype specific- all Catalogue", "Brain-related"), lty=c(1, 1, 3, 3), lwd=c(5, 5, 5,
 5), col=c("grey", "black", "black", "grey"))
dev.off()

#Figure 3- Histograms
require(plotrix)
pdf("Histograms-clumped-4models.pdf")
#make PDF first (better quality);then use Preview to convert to TIFF
```

```
par(mfrow=c(2,2))
x<-read.table("Brain-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.175,0.95, by=0.05)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density", breaks=bins,
 col=c("black","grey"), ylim=c(0,10))
legend("topright", title="Brain-related", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"), cex=0.9)
x<-read.table("Autoimmune-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.175,0.975, by=0.05)#0.27 as starting works but starts at 0.3
multhist(l, freq=F, xlab="Predicted Value", ylab="Density",  breaks=bins,
 col=c("black","grey"), ylim=c(0,10))
legend("topright", title="Autoimmune", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"), cex=0.9)
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.175,0.95, by=0.05)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density", breaks=bins,
 col=c("black","grey"), ylim=c(0,10))
legend("topright", title="All phenotype", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"), cex=0.9)
x<-read.table("Nonpheno-allCat-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.175,0.95, by=0.05)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density", breaks=bins,
 col=c("black","grey"), ylim=c(0,10))
legend("topright", title="All phenotype", c("all Catalogue hits", "non-hits"), pch=c(15,
 15), col=c("black","grey"), cex=0.9)
dev.off()

#Figure 4- Histograms (bin size of 0.1)
require(plotrix)
pdf("Histograms0.1bins-clumped-4models.pdf")
par(mfrow=c(2,2))
x<-read.table("Brain-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.25,0.95, by=0.1)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density", breaks=bins,
 col=c("black","grey"), ylim=c(0,6))
legend("topright", title="Brain-related", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"))
x<-read.table("Autoimmune-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
```

```
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.25,0.95, by=0.1)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density",  breaks=bins,
 col=c("black","grey"), ylim=c(0,6))
legend("topright", title="Autoimmune", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"))
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.25,0.95, by=0.1)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density", breaks=bins,
 col=c("black","grey"), ylim=c(0,6))
legend("topright", title="All phenotype", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"))
x<-read.table("Nonpheno-allCat-testset.csv", sep=",")
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
l<-list(hits[,5], nonhits[,5])
bins<-seq(0.25,0.95, by=0.1)
multhist(l, freq=F, xlab="Predicted Value", ylab="Density", breaks=bins,
 col=c("black","grey"), ylim=c(0,6))
legend("topright", title="All phenotype", c("all Catalogue hits", "non-hits"), pch=c(15,
 15), col=c("black","grey"))
dev.off()

#Figure 5- Box plots
pdf("boxplots-clumped-testset-4models.pdf")
par(mfrow=c(2,2))
x<-read.table("Brain-testset.csv",sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
boxplot(hits[,5], nonhits[,5], xlab="Classification", ylab="Prediction", main="Brain-
 related", names=c("Hit","Non-hit"), ylim=c(0.25,0.95))
x<-read.table("Autoimmune-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
boxplot(hits[,5], nonhits[,5], xlab="Classification", ylab="Prediction",
 main="Autoimmune", names=c("Hit","Non-hit"), ylim=c(0.25,0.95))
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
boxplot(hits[,5], nonhits[,5], xlab="Classification", ylab="Prediction", main="Non-
 phenotype specific", names=c("Hit","Non-hit"), ylim=c(0.25,0.95))
x<-read.table("Nonpheno-allCat-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
boxplot(hits[,5], nonhits[,5], xlab="Classification", ylab="Prediction", main="Non-
 phenotype specific-all Catalogue", names=c("Hit","Non-hit"), ylim=c(0.25,0.95))
dev.off()

#Figure 6- Violin plots
library(vioplot)
pdf("vioplots-clumped-testset-4models.pdf")
par(mfrow=c(2,2))
```

```
x<-read.table("Brain-testset.csv",sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
vioplot(hits[,5], nonhits[,5], names=c("Hit","Non-hit"), col="white", ylim=c(0.25,0.95))
title("Brain-related", xlab="Classification", ylab="Prediction")
x<-read.table("Autoimmune-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
vioplot(hits[,5], nonhits[,5], names=c("Hit","Non-hit"), col="white", ylim=c(0.25,0.95))
title("Autoimmune", xlab="Classification", ylab="Prediction")
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
vioplot(hits[,5], nonhits[,5], names=c("Hit","Non-hit"), col="white", ylim=c(0.25,0.95))
title("Non-phenotype specific", xlab="Classification", ylab="Prediction")
x<-read.table("Nonpheno-allCat-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
vioplot(hits[,5], nonhits[,5], names=c("Hit","Non-hit"), col="white", ylim=c(0.25,0.95))
title("Non-phenotype specific-all Catalogue", xlab="Classification", ylab="Prediction")
dev.off()

#Figure 7- Quantile-quantile plots
pdf("qqplots-clumped-4models.pdf")
par(mfrow=c(2,2))
x<-read.table("Brain-testset.csv",sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
qqplot(nonhits[,5], hits[,5], ylab="Hits", xlab="Non-hits", main="Brain-related",
 xlim=c(0.25,0.95), ylim=c(0.25,0.95))
abline(0,1, col="grey")
x<-read.table("Autoimmune-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
qqplot(nonhits[,5], hits[,5], ylab="Hits", xlab="Non-hits",
 main="Autoimmune",xlim=c(0.25,0.95), ylim=c(0.25,0.95))
abline(0,1, col="grey")
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
#nrow(hits) #4480
#nrow(nonhits) #75341
qqplot(nonhits[,5], hits[,5], ylab="Hits", xlab="Non-hits", main="Non-phenotype
 specific",xlim=c(0.25,0.95), ylim=c(0.25,0.95))
abline(0,1, col="grey")
x<-read.table("Nonpheno-allCat-testset.csv", sep=",")
hits<-subset(x, x[,4]=="1")
nonhits<- subset(x, x[,4]=="0")
qqplot(nonhits[,5], hits[,5], ylab="Hits", xlab="Non-hits", main="Non-phenotype specific-
 all Catalogue",xlim=c(0.25,0.95), ylim=c(0.25,0.95))
abline(0,1, col="grey")
dev.off()

#Figure 8- Ranks
require(plotrix)
```

```
pdf("Ranks-clumped-4models.pdf")
par(mfrow=c(2,2))
x<-read.table("Brain-testset.csv", sep=",")
dim(x) # 32867
sortbypred<-x[with(x, order(V5)), ]
sortbypred$rank<-seq(1, 32867,1)
hitsforplot<-subset(sortbypred, sortbypred$V4==1)
nonhitsforplot<-subset(sortbypred, sortbypred$V4==0)
l<-list(hitsforplot$rank, nonhitsforplot$rank)
bins<-seq(0, 34000, by=1000)
multhist(l, freq=F, xlab="Rank", ylab="Density", breaks=bins, col=c("black","grey"))
legend("topleft", title="Brain-related", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"))
x<-read.table("Autoimmune-testset.csv", sep=",")
dim(x) # 33500
sortbypred<-x[with(x, order(V5)), ]
sortbypred$rank<-seq(1, 33500,1)
hitsforplot<-subset(sortbypred, sortbypred$V4==1)
nonhitsforplot<-subset(sortbypred, sortbypred$V4==0)
l<-list(hitsforplot$rank, nonhitsforplot$rank)
bins<-seq(0,34000, by=1000)
multhist(l, freq=F, xlab="Rank", ylab="Density", breaks=bins, col=c("black","grey"))
legend("topleft", title="Autoimmune", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"))
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",")
dim(x) # 31427
sortbypred<-x[with(x, order(V5)), ]
sortbypred$rank<-seq(1, 31427,1)
hitsforplot<-subset(sortbypred, sortbypred$V4==1)
nonhitsforplot<-subset(sortbypred, sortbypred$V4==0)
l<-list(hitsforplot$rank, nonhitsforplot$rank)
bins<-seq(0,34000, by=1000)
multhist(l, freq=F, xlab="Rank", ylab="Density",  breaks=bins, col=c("black","grey"))
legend("topleft", title="All phenotype", c("<5x10^-8 hits", "non-hits"), pch=c(15, 15),
 col=c("black","grey"))
x<-read.table("Nonpheno-allCat-testset.csv", sep=",")
dim(x) # 33444
sortbypred<-x[with(x, order(V5)), ]
sortbypred$rank<-seq(1,33444,1)
hitsforplot<-subset(sortbypred, sortbypred$V4==1)
nonhitsforplot<-subset(sortbypred, sortbypred$V4==0)
l<-list(hitsforplot$rank, nonhitsforplot$rank)
bins<-seq(0,34000, by=1000)
multhist(l, freq=F, xlab="Rank", ylab="Density", breaks=bins, col=c("black","grey"))
legend("topleft", title="All phenotype-all Catalogue", c("hits", "non-hits"), pch=c(15,
 15), col=c("black","grey"))
dev.off()

##Statistical tests
#Mann-Whitney U p-value
x<-read.table("Brain-testset.csv", sep=",")
nonhits<-subset(x, x[,4]==0)
hits<-subset(x, x[,4]==1)
wilcox.test(nonhits[,5], hits[,5])$p.value
x<-read.table("Autoimmune-testset.csv", sep=",")
```

```
nonhits<-subset(x, x[,4]==0)
hits<-subset(x, x[,4]==1)
wilcox.test(nonhits[,5], hits[,5])$p.value
x<-read.table("Non-pheno-5e-8-testset.csv", sep=",")
nonhits<-subset(x, x[,4]==0)
hits<-subset(x, x[,4]==1)
wilcox.test(nonhits[,5], hits[,5])$p.value
x<-read.table("Non-pheno-allCat-testset.csv", sep=",")
nonhits<-subset(x, x[,4]==0)
hits<-subset(x, x[,4]==1)
wilcox.test(nonhits[,5], hits[,5])$p.value

#Hypergeometric test p-value
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",", as.is=T) #repeat for other data sets
hits<-subset(x, x[,4]==1)
nonhits<-subset(x, x[,4]==0)
res<-matrix(nrow=3,ncol=13)
row=1
col=0
BD<-length(nonhits[,1])
j<-length(hits[,1])
#prediction value bins ranging from less than 0.35 to between 0.9 and 0.95, increasing by
 increments of 0.5
for (i in seq(0.3,0.9,0.05))
{
col<-col+1
c<-length(subset(nonhits[,5],nonhits[,5]<i+0.05 & nonhits[,5]>i))
a<-length(subset(hits[,5], hits[,5]<i+0.05 & hits[,5]>i))
res[row,col]<-c/dim(nonhits)[1]
res[row+1,col]<-a/dim(hits)[1]
res[row+2,col]<-sum(phyper(a,j,BD-j,c, lower.tail=F))
}
#write a table to read in Excel
head<-c("p<0.35", "0.35<p<0.4", "0.4<p<0.45", "0.45<p<0.5", "0.5<p<0.55",
 "0.55<p<0.6","0.6<p<0.65", "0.65<p<0.7", "0.7<p<0.75","0.75<p<0.8", "0.8<p<0.85",
 "0.85<p<0.9","0.9<p<0.95")
table<-rbind(head, res)
write.table(table, "Hypergeometric.csv", sep=",", row.names=F, col.names=F, quote=F)
#the first row is the frequency of non-hits
#the second row is the frequency of the hits
#the third row is the hypergemoetric p-value

#Asymptotic Generalized Cochran-Mantel-Haenszel Test
library("coin")
x<-read.table("Nonpheno-5e-8-testset.csv", sep=",", as.is=T) #repeat for other data sets
nonhits<-subset(x, x[,4]==0)
hits<-subset(x, x[,4]==1)
counts<-matrix(nrow=2,ncol=13)
row=1
col=0
for (i in seq(0.3,0.9,0.05))
{
col<-col+1
c<-length(subset(nonhits[,5],nonhits[,5]<i+0.05 & nonhits[,5]>i))
a<-length(subset(hits[,5], hits[,5]<i+0.05 & hits[,5]>i))
```

```
counts[row,col]<-c
counts[row+1,col]<-a
}
counts<-as.table(counts)
cmh_test(counts)
```

D ENCODE accession numbers

Histone Marks:

ENCSR000AKF

ENCSR000AOT

ENCSR000AKS

ENCSR000AMJ

ENCSR000ANA

ENCSR000AMU

ENCSR000APJ

ENCSR000ANI

ENCSR000ANX

ENCSR000ALI

ENCSR000AKL

ENCSR000EXV

ENCSR000EWC

ENCSR000EXJ

ENCSR000DWJ

ENCSR000DVU

ENCSR000DUA

ENCSR000DUO

ENCSR000DWD

ENCSR000DRY

ENCSR000DQH

ENCSR000DTU

ENCSR000AKA

ENCSR000AMP

ENCSR000DUF

ENCSR000AKU

ENCSR000AOF

ENCSR000DTQ

ENCSR000DQV

ENCSR000DQM

ENCSR000DXR

ENCSR000DWP

ENCSR000AOC

ENCSR000AKC

ENCSR000AKP

ENCSR000AMO

ENCSR000ALB

ENCSR000APH

DNase I:

ENCSR000ENO

ENCSR000EPC

ENCSR000EMI

ENCSR000ENP

ENCSR000EPL

ENCSR000EMN

ENCSR000EMS

ENCSR000ENM

ENCSR000ENU

ENCSR000EPT

ENCSR000EPZ

ENCSR000EPS

ENCSR000EMQ

ENCSR000ELF

ENCSR000EJS

ENCSR000EJI

ENCSR000EJK

ENCSR000ELT

ENCSR000EJL

ENCSR000ELA

ENCSR000EJA

ENCSR000EKE

ENCSR000EKS

ENCSR000EKD

ENCSR000EJT

ENCSR000EID

ENCSR000EKZ

ENCSR000EJX

ENCSR000EJJ

ENCSR000EKC

ENCSR000EJD

ENCSR000ELU

ENCSR000EJF

ENCSR000ELV

ENCSR000EJE

ENCSR000EKU

ENCSR000EKT

ENCSR000EIE

ENCSR000EPS

ENCSR000CZZ

ENCSR000DBG

ENCSR000DBP

ENCSR000DAB

ENCSR000DBK

ENCSR000DBN

ENCSR000DAS

ENCSR000DBD

ENCSR000CZG

ENCSR000DBO

ENCSR000DBL

ENCSR000DBM

ENCSR000DBB

ENCSR000DAD

ENCSR000DAZ

ENCSR000DBC

ENCSR000CZK

ENCSR000CZE

ENCSR000CZJ

ENCSR000CZD

ENCSR000DBH

Copyright Acknowledgements