THE UNIVERSITY
of EDINBURGH

The design and application of SuRFR: an R package to prioritise candidate functional DNA sequence variants

Niamh Margaret Ryan

Doctor of Philosophy – The University of Edinburgh – 2015

Declaration

I declare that this thesis has been composed by me, that the work described in this thesis is my own, except where otherwise stated, and that the work described in this thesis has not been submitted for any other degree or professional qualification.

_____

Niamh Margaret Ryan

Table of Contents

Acknowledgements

This thesis is dedicated to my family: to my parents, Pat and Ben Ryan, and to my brother and sister, David and Megan. In particular this thesis is for my mum, Pat, for putting up with me for all of these years and for always being there with an encouraging message or a shoulder to cry on. My family are my rock and I honestly could not have done this without them.

There are so many people I need to thank for helping me get through this and helping me complete this thesis. First and foremost, I would like to thank my first supervisor, Kathy Evans. Kathy not only helped guide me through this project, sharing her knowledge, expertise and experience, but made me feel that we were a partnership, and we were doing this together. Kathy always encouraged me to push myself and look at my project from different angles. She was also there for me in the second year of my PhD when I contracted glandular fever. Almost five months of recovery time left me wondering if it was even going to be possible to finish my PhD, but I did, and it was largely due to Kathy. I have learnt so much from this incredible role model and I will be ever grateful to her. I would also like to thank my second supervisor, Martin Taylor, again, for all of his guidance and expertise, but also for putting up with my at times … shall we say enthusiastic (?!) personality. I also really need to thank Stewart Morris, who has been the other rock in my life – so much of the bioinformatics I have learnt, I have learnt from him, and I always know, no matter what problem I am having, Stewart will think of a solution! For a long time it was just myself, Stewart and Steve in our office, and I'd like to thank them both for putting up with me.

I would also like to thank David Porteous, Pippa Thomson and Rosie Walker for all of their input into my project. To everyone else in my group: Susan, Dan, Alan and all of MedGen – thank you for being such a wonderfully warm and inclusive group of people who made my time as a PhD student here very easy.

I would also like to thank Rosie again, separately, along with Faith Davies, for being there when I needed to talk. Whether it was to gossip, or vent, or cry, or to try and work out some ideas out loud, you were both always there for me. Thank you so much! Not to forget two other very important people, my flatmate Pete, and Nidhi Sharma. Thank you for helping me forget the PhD at times, for all the fun we have had, and for all of your support and friendship (and food!).

There are a host of other people in Edinburgh who have been my family away from home: Ciara and Ronan, Leagh and Mitch, all my choir folk… you all know who you are!

Lastly to Pete and Marian Humphries, for being the people who first directed me down this path when I did work experience with you at the age of 15. I've finally done it!

I am so, so grateful to all of the wonderful people in my life who have helped me through the good times and the bad. I am beyond lucky.

Figures and Tables

Abstract

Genetic analyses such as linkage and genome wide association studies (GWAS) have been extremely successful at identifying genomic regions that harbour genetic variants contributing to complex disorders. Over 90% of disease-associated variants from GWAS fall within non-coding regions (Maurano et al., 2012). However, pinpointing the causal variants has proven a major bottleneck to genetic research.

To address this I have developed SuRFR, an R package for the ranked prioritisation of candidate causal variants by predicted function. SuRFR produces rank orderings of variants based upon functional genomic annotations, including DNase hypersensitivity signal, chromatin state, minor allele frequency, and conservation. The ranks for each annotation are combined into a final prioritisation rank using a weighting system that has been parametrised and tested through ten-fold cross-validation.

SuRFR has been tested extensively upon a combination of synthetic and real datasets and has been shown to perform with high sensitivity and specificity. These analyses have provided insight into the extent to which different classes of functional annotation are most useful for the identification of known regulatory variants: the most important factor for identifying a true variant across all classes of regulatory variants is position relative to genes. I have also shown that SuRFR performs at least as well as its nearest competitors whilst benefiting from the advantages that come from being part of the R environment.

I have applied SuRFR to several genomics projects, particularly the study of psychiatric illness, including genome sequencing of a large Scottish family with bipolar disorder. This has resulted in the prioritisation of such variants for future study.

Thesis Lay Summary

The vast majority of our DNA sequence is identical between humans. The small fraction that is different reflects the differences between individuals, from how different you are to your siblings and parents to population level differences. Some of these variations in our DNA contribute to our risk of developing hereditary diseases. Most of the known variants that cause human disease are found in the regions of our DNA that encode proteins. Proteins are the building blocks of life and when the DNA that encodes them contains a variation or mistake, the structure of the protein can be altered, or the protein is not made at all, meaning it is no longer able to do its job, potentially leading to disease. The stretches of DNA between the protein coding regions, the non-coding regions, often contain molecular switches that control how much protein is made and when. These molecular switches are very important for the correct function of proteins, but our knowledge of what defines them is very limited. This makes it difficult to predict whether a variant is located in a molecular switch and if it has a functional role in human disease. However, we have lots of different ways of characterising and annotating DNA sequence, for example the extent to which a stretch of DNA is conserved between humans and other organisms (highly conserved DNA sequences often having important functions), which we can layer together to find unique patterns of annotation associated with different types of variants.

The aim of my PhD was to use computational techniques to piece together some of the information we have on the characteristics of these molecular switches to predict the likelihood of a non-coding variants playing a role in human disease. To do this, I used a dataset of known disease-causing non-coding variants and a machine learning method to identify patterns associated with these variants. I used a statistical method called cross-validation to show that the patterns I identified are truly associated with the disease variants rather than being due to chance or to over-fitting of my data. I then used this information to build a computational tool to prioritise variants on their likelihood of being functional variants (and in turn likely to play a role in human disease). This tool is called SuRFR (SNP Ranking by Function R package) and is

freely available as part of a widely used computer programming language. I have run SuRFR on a range of test datasets, where SuRFR was correctly able to prioritise the known disease variant(s) above the background variants in a reproducible manner. This shows that SuRFR is likely to work well on novel data where the disease causing variant is not known. As such, SuRFR can be used in the search for functional and disease-causing non-coding variants.

Abbreviations

AAS : Amino acid substitution

ADHD : Attention deficit-hyperactivity disorder

AFR : African

AMR : American

ASD : Autism spectrum disorder

ASN : Asian

AUC : Area under the (ROC) curve

BD : Bipolar disorder

CD : Conduct disorder

CP : Combined p-value

CRAN : Comprehensive R archive network

CSHL : Cold Spring Harbour Laboratories

DAF : Derived allele frequency

DNA : Deoxyribonucleic acid

DNase HS : Dnase I hypersensitive site

EAS : East Asian

ENCODE : Encyclopedia of DNA Elements

eQTL : expression quantitative trait loci

EUR : European

FunSeq : Function based prioritisation of sequence variation

GD : Gene density

GERP : Genomic evolutionary rate profile

GIN : Genomic information network

GTR : General time reversible

GWAS : Genome-wide association study

GWAVA : Genome-wide annotation of variants

h2 : Heritability

HGMD : Human gene mutation database

indel : insertion-deletion

IO : Idiopathic oedema

LD : Linkage disequilibrium

lincRNA : Long intergenic non-coding RNA

LOD : Log-odds

LOOCV : Leave-one-out cross-validation

LSDB : Locus specific database

MAF : Minor allele frequency

MDD : Major depressive disorder

meQTL : methylation quantitative trait loci

miRNA : micro RNA

mLOD : maximum LOD

MooDS : Systematic Investigation of the Molecular Causes of Major Mood Disorders and Schizophrenia

NEC : Non-exonic conservation score

NGS : Next generation sequencing

NMD : Normalised mutational proportion

OMIM : Online Inheritance in Man

PGC : Psychiatric Genomics Consortium

PRS : Polygenic risk score

RAVEN : Regulatory analysis of variation in enhancers

RNA : Ribonucleic acid

ROC : Receiver operating characteristic

RS : Rejection substitution

rSNP : regulatory SNP

SAS : South Asian

SBF2 : Scottish bipolar family 2

SCZ : Schizophrenia

SIFT : Sorting tolerant from intolerant

SM : Stewart Morris

SNP : Single nucleotide polymorphism

SuRFR : SNP ranking by function R package

SVM : Support vector machine

T2D : Type 2 diabetes

TF : Transcription factor

TFBS : Transcription factor binding site

TSS : Transcription start site

UCSC : University of California, Santa Cruz

UTR : Untranslated region

WES : Whole exome sequencing

WGS : Whole genome sequencing

WHO : World Health Organisation

WTCCC : Wellcome Trust Case Control Consortium

1000 Genomes : Thousand Genomes project

Chapter 1: Introduction

1.1 The study of human disease and complex traits

Human genetics is the study of genetic variation in human genomes and the impact of this variation on phenotypes, including complex traits and disease. Fu et al. (2013) defined the genetic architecture of a trait as being "a comprehensive description of how genes and the environment conspire to produce phenotypes" (Fu et al., 2013). Improving our understanding of the genetic architecture and heritability of complex traits and diseases, and how our genotypes biologically connect to phenotypes, is a major goal of genomics projects. This is not a straightforward task, as there is a large amount of genetic complexity across different diseases and disorders.

The simplest form of inheritance, commonly known as "Mendelian inheritance", was discovered by Gregor Mendel at the end of the 19th century. Mendel's laws describe the relationship between genotype and phenotype, where a single variant drives the expression of a particular phenotype or disease. Under this model, alleles follow either a dominant or recessive pattern of inheritance, further complicated by whether the variant is autosomal or X-linked. Mendel studied simple genetic traits where two alleles (A and a) generated three possible genotypes (AA, Aa, aa). When inheritance is dominant, the phenotypic effect of the dominant allele (A) will mask the phenotypic effect of the second, recessive, allele (a), so that both the dominant homozygous genotype (AA) and heterozygous genotype (Aa) will have the same phenotype. The phenotype of the recessive allele (a) will only be seen when the genotype is homozygous for the recessive allele (aa). If one parent carries an autosomal dominant disease allele, the probability is that 50% of the offspring will also receive the allele and be affected by the disease or phenotype. For a recessive disorder or trait, both parents will each have to carry a copy of the recessive allele. When both parents are heterozygous for the two alleles (Aa), 25% of the offspring will be homozygous for the recessive allele and be affected by the disease (aa), 25% will be homozygous for the dominant allele (AA) and so will not express the phenotype and 50% will be heterozygous (Aa) and will not express the phenotype (Aa), but will be carriers of the disease allele (and are therefore capable of passing it on to their own offspring).

An example of a disease that follows a Mendelian autosomal-recessive pattern of inheritance is cystic fibrosis, which is caused by large effect size (where effect size is defined as the ratio of the odds of disease manifestation in carriers vs. non-carriers (Zollner and Pritchard, 2007)) mutations in both copies of the *CFTR* gene. However, even this is not as simple an example as originally thought; to date, over 2,000 variants have been catalogued located in and around the *CFTR* gene that lead to cystic fibrosis (Drumm et al., 2012), with at least eight regulatory variants in the promoter region of *CFTR* also known to contribute to a cystic fibrosis phenotype (Giordano et al., 2013). Additional loci across the genome have also been identified that harbour variants that modify clinical outcomes of cystic fibrosis disease (Drumm et al., 2012). This suggests that even the architecture of 'simple' disorders is not particularly simple.

A catalogue of known Mendelian diseases (genes, mutations and associated phenotypes) can be found in the online database, the 'Online Mendelian Inheritance in Man' (OMIM) (Amberger et al., 2015). It is estimated that there are ~7,000 rare monogenic human diseases. Over the past 25 years ~50% of the genes responsible for these diseases have been identified (Boycott et al., 2013, Deciphering Developmental Disorders, 2015). The advances of next generation sequencing (NGS), both whole genome and whole exome sequencing, are predicted to aid the identification of the remaining genes causing Mendelian diseases by 2020 (Boycott et al., 2013).

At the other end of the spectrum are complex traits and disorders that do not follow Mendelian patterns of inheritance. These traits and disorders are caused by susceptibility variants with much smaller effect sizes that together (along with environmental factors) are associated with a phenotype. Such diseases have been the focus of genome wide association studies (GWAS), which assay the genotypes of hundreds of thousands of common markers (present in at least 1% of the population) across the genome, in thousands of cases and controls, to test for the association of variants with a phenotype of interest (Hardy and Singleton, 2009).

Height is an example of a complex trait that is affected by both an individual's genotype at multiple loci and environmental factors. Heritability (h^2, the proportion of phenotypic variation in a trait or disease that is due to genetic factors (Wray et al., 2013)) estimates for height range from ~70-90% (Silventoinen et al., 2003). To date, nearly 200 loci have been associated with height in GWASs, which together explain roughly 20% of the heritability (Berndt et al., 2013, Lango Allen et al., 2010). A complex disorder with a genetic architecture that is similar to height is type 2 diabetes (T2D). Heritability estimates for T2D range from 30 – 70% (Wellcome Trust Case Control, 2007). To date, over 70 loci have been identified that are associated with T2D (Replication et al., 2014), each of small individual effect size, together explaining roughly 10% of the heritability of T2D (Voight et al., 2010).

NGS methods have led the discovery of millions of genetic variants identified through the sequencing of thousands of individuals. Comprehensive catalogues of human variation, such as the 1000 genomes database (Genomes Project et al., 2010), contain data on a range of variant classes in the human genome, including single nucleotide polymorphisms (SNPs), short insertions and deletions (indels) and structural variants (copy number variants (CNVs) and chromosomal rearrangements). These variants range in allele frequency from common to very rare or unique variants. The vast majority of human variation (90%) is common and ancient, dating back to before the out-of-Africa migration (at least 50,000 – 60,000 years ago) (McClellan and King, 2010a). In contrast, the majority of rare variants are very new, tend to be population, family or individual specific, and are more likely to be deleterious than older, more common variants (Henn et al., 2015). This is due to the actions of genetic drift and purifying selection, which work to remove deleterious variants from a population, preventing them from reaching high frequency. Variants causing Mendelian diseases tend to fall into this category of high effect-size rare variation. In contrast, variants associated with complex traits and variants of weaker deleterious effect can become common due to random drift (Figure 1.1).

Occasionally deleterious mutations of larger effect can become common in a population if a fast population expansion occurs (Henn et al., 2015), or if the disease has a later age

of onset (thereby not affecting reproductive fitness). An example of such a disease is Huntington's disease, which segregates in an autosomal dominant fashion and presents later in life (~30 -50 years of age)(Gusella et al., 1983). The *Huntingtin* gene was identified through investigation of an extremely large Venezuelan family (the pedigree dating back to the 1800s and containing over 3,000 members all related to a single common ancestor).

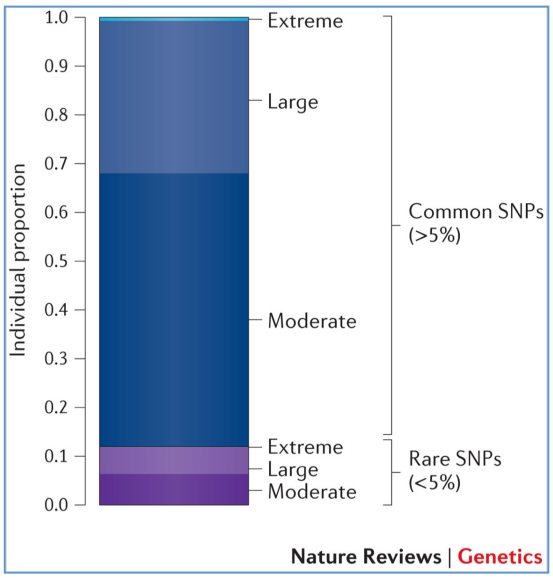


Figure 1.1 Proportion of deleterious variants found in the average genome classified by their frequency in the population (common (in blue) versus rare (in purple)). This figure shows that the proportion of rare variants in an individual's genome will be equally likely to have a medium or a large effect size. The vast majority of deleterious variants in an individual's genome (70%) are common variants, most of which have only a small effect ('moderate'). A very small fraction of common variants will also have extreme effects; however, the majority of these will have been purged by purifying selection. Taken from Henn et al., 2015 (Henn et al., 2015).

To summarise, deleterious variants contributing to disease occur across the full spectrum of allele frequencies and with a range of effect sizes (Figure 1.1). The genetic effect size of a variant is related to both the penetrance (the number of individuals who carry a particular genotype that also express the associated phenotype) and frequency of the variant (Zollner and Pritchard, 2007). Figure 1.2 summarises the relationship between effect size (penetrance), allele frequency and genetic architecture of disease variants.

Variants of different effect sizes and frequencies play different roles in human disease. Rare variants with high effect sizes generally lead to Mendelian diseases like cystic fibrosis, while more common variants with smaller effect sizes playing a role in susceptibility to more complex diseases and disorders, such as T2D. Variable penetrance, which can be due to the influence of genetic modifiers and environment, can make it difficult to identify susceptibility variants.

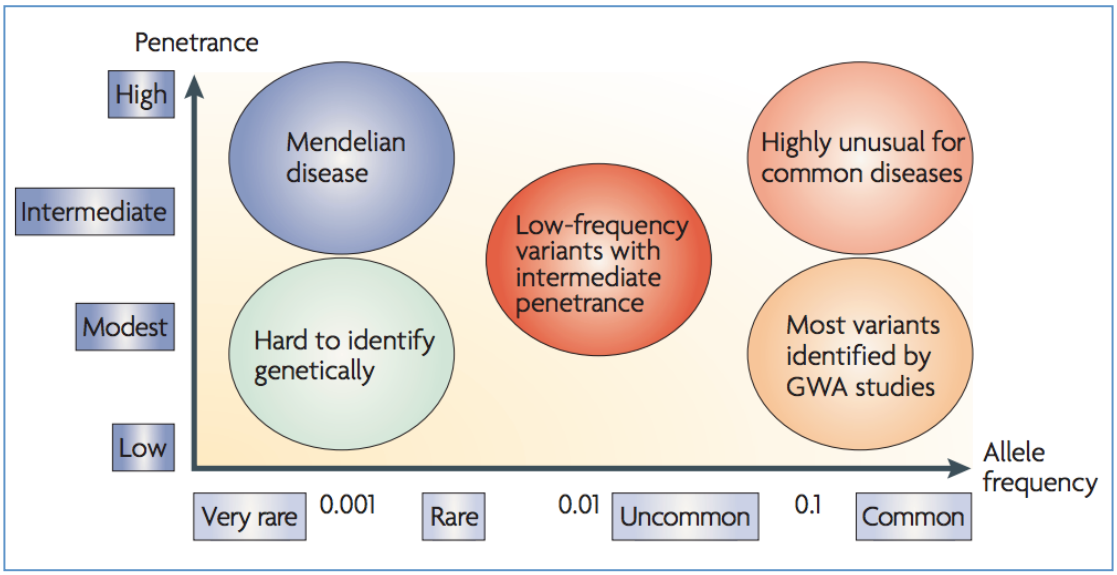


Figure 1.2 This figure outlines potential classes of human disease, stratified by the frequency of underlying genetic variation on the x-axis and the penetrance of these variants on the y-axis. Taken from McCarthy et al. (2008) (McCarthy et al., 2008).

1.2 Psychiatric illness

Psychiatric disorders are debilitating clinical syndromes that are characterised by psychological symptoms that impact multiple life areas, creating distress for the person experiencing these symptoms as well as for their family and friends ((WHO, 2015) accessed June 2015). Although they have largely unknown aetiology and pathophysiology (psychiatric syndromes hence being referred to as 'disorders', as

opposed to 'diseases'), such conditions are associated with high morbitity, accounting for roughly one-third of disability world wide, negatively impacting the lives of both sufferers and their families and causing considerable personal and societal burden (reviewed in (Sullivan et al., 2012)). In addition, such disorders are associated with increased mortality rates, from suicide and other causes (Eaton et al., 2008).

To date, over 300 psychiatric disorders have been defined. Nine of these, described by Sullivan et al. (2012) as the 'cardinal psychiatric disorders', are summarised in Table 1.1, taken from (Sullivan et al., 2012)). Of these, the Psychiatric Genomics Consortium (PGC) has defined autism spectrum disorder (ASD), attention deficit-hyperactivity disorder (ADHD), bipolar disorder (BD), major depressive disorder (MDD), and schizophrenia (SCZ) as being the 'major' psychiatric syndromes. The aim of the PGC is to conduct statistically rigorous and comprehensive GWAS meta-analyses on each of these major psychiatric illnesses, as well comparative analyses across the five disorders (Sullivan, 2010).

Family, twin and adoption studies have shown that there is also a strong heritable component to psychiatric disorders (see lifetime prevalence in table 1.1). In addition, several environmental risk factors have been identified for many of these disorders (summarised in Table 1.2, taken from the review by Uher, (2014) (Uher, 2014)). The genetic architecture of these disorders has yet to be determined, but is likely to be complex. De *novo* mutations, structural rearrangements, rare variants, common variants have all been implicated in the aetiology of these disorders. Each of these, along with methods to identify pathogenic variants, will be described briefly in the following sections. More details on the potential genetic architectures of BD and MDD can be found in the Introduction to Chapter 5.

Abbrev.	Name	Life prev	Heritability	Essential characteristics	Notable feature
AD	Alzheimer's disease	0.132	0.58	Dementia, defining neuropathology	Of the top 10 causes of death in the US, AD alone has increasing mortality
ADHD	Attention-deficit hyperactivity disorder	0.053	0.75	Persistent inattention, hyperactivity, impulsivity	Costs estimated at ~$US 100×10^9/year
ALC	Alcohol dependence	0.178	0.57	Persistent ethanol use despite tolerance, withdrawal, dysfunction	Most expensive psychiatric disorder (total costs exceed $US 225×10^9/year)
AN	Anorexia nervosa	0.006	0.56	Dangerously low weight from self-starvation	Notably high standardized mortality ratio
ASD	Autism spectrum disorder	0.001	0.80	Markedly abnormal social interaction and communication beginning before age 3	Huge range of function, from people requiring complete daily care to exceptional occupational achievement
BIP	Bipolar disorder	0.007	0.75	Manic-depressive illness, episodes of mania usually with MDD	As a group, nearly as disabling as SCZ
MDD	Major depressive disorder	0.130	0.37	Unipolar depression, marked and persistent dysphoria with physical/cognitive symptoms	Ranks #1 in burden of disease in world
NIC	Nicotine dependence	0.240	0.67	Persistent nicotine use with physical dependence (usually cigarettes)	Major preventable risk factor for many diseases
SCZ	Schizophrenia	0.004	0.81	Long-standing delusions and hallucinations	Life expectancy decreased by 12–15 years

Table 1.1 This table describes nine psychiatric disorders, defining their lifetime prevalence, heritability, essential characteristics and notable features. Taken from Sullivan et al. (2012) (Sullivan et al., 2012).

	Exposure	Schizophrenia	Bipolar disorder	Major depressive disorder
Prenatal	Season of birth	+++ (17)	++ (19)	+ (19)
	Inadequate nutrition	++ (20)	++ (21)	+ (21)
	Vitamin D levels	+++ (15)		
	Lead	+ (22)		
	Herpes simplex virus-2	++ (23)		
	Rubella	+ (24)		
	Prenatal stress		+ (25)	+ (25)
Perinatal	Preterm birth	++ (26)	+++ (26)	+ (26)
	Obstetric complications	+ (27)	– (28)	
	Hypoxia			
Childhood	Cytomegalovirus	+ (29)		
	Maltreatment	+++ (16)	+ (30–32)	+++ (33, 34)
	Loss of a parent			++ (35)
	Social disadvantage	+++ (36, 37)	– (36)	+++ (36, 38)
	Bullying	++ (39)		+ (40)
	Urbanicity	+++ (71)		
	Minority status	+++ (41)		++ (42)
Adolescence	Cannabis	+++ (18)	+ (30)	+ (18)
Adulthood	Stressful life events	+ (43)	++ (44)	+++ (45)
	Toxoplasma		+ (46)	

The number of plus signs indicates the strength of evidence for association: +++, consistent evidence from multiple studies or a meta-analysis; ++, evidence from several studies or a strong association in a high-quality study; +, evidence from a single study or multiple low quality studies; –, evidence for no association; blank fields reflect lack of evidence for or against association. The list is limited to environmental factors and excludes risk factors that reflect condition of the individual (e.g., birth weight).

Table 1.2. Summary of environmental risk factors for schizophrenia, bipolar disorder and major depressive disorder. Taken from Uher (2014) (Uher, 2014).

1.3 A summary of potential disease models

1.3.1 De novo mutations

There is growing evidence of the role of *de novo* variants (SNPs, indels and CNVs) in the genetics of psychiatric illness (Gratten et al., 2013). *De novo* CNVs have been implicated in a range of nervous system disorders (Lee and Lupski, 2006), including ASD (Marshall et al., 2008), SCZ ((Karayiorgou et al., 1995); (Xu et al., 2008); (International Schizophrenia, 2008); (Stefansson et al., 2008); (Vacic et al., 2011); (Bassett and Chow, 2008)) and BD (Malhotra et al., 2011).

Two studies, in 2007 and 2008 respectively, showed that *de novo* variants were more commonly found in ASD cases than controls ((Marshall et al., 2008); (Sebat et al., 2007)). Several additional studies have since shown an increase in the frequency (difference of 6%) of *de novo* variants in children with ASD compared to unaffected siblings ((Levy et al., 2011); (Sanders et al., 2011)). Most recently, two large exome sequencing studies (published in 2014) of thousands of families with a history of autism implicated over 400 genes with *de novo* loss of function variants, or likely gene-disrupting variants, as contributing to ASD ((Iossifov et al., 2014); (De Rubeis et al., 2014)). Many of the genes implicated in these studies were found to encode proteins involved in neuronal processes such as synaptic formation and voltage-gated ion channels, as well as transcription regulation and chromatin remodelling pathways (De Rubeis et al., 2014). A review by Ronemus et al. (2014) proposed that ASD is most commonly caused by parental germ line *de novo* mutations in a two-class risk model (Ronemus et al., 2014). The model suggests most cases (99%) are low-risk, with *de novo* mutations contributing risk of 0.5% for males and 0.15% for females; in contrast, for high-risk families, one parent carries a highly penetrant *de novo* mutation which confers a 50% risk of ASD in males and 12.5% risk in females (Ronemus et al., 2014).

De novo mutations have also been shown to play a role in SCZ, being shown in one study to be eight times more common in patients with sporadic SCZ than in controls ((McClellan and King, 2010b); (Xu et al., 2008)). This class of variation have been identified at specific loci across the genome, including chr22q11.2, chr15q13.3 and

chr1q.21.1 (Purcell et al., 2014) and have been shown to converge on sets of functionally related proteins (including synaptic proteins and genes that have been implicated in the study of other psychiatric conditions). This was shown by Fromer et al. (2014), who performed exome sequencing in 617 schizophrenia trios and an independent set of 731 controls (Fromer et al., 2014). This study found no increase in the rate of *de novo* mutations between probands and controls. However, it did identify an enrichment (corrected p = 0.0007) of *de novo* nonsynonymous substitutions in "SCZ genes" (identified by independent evidence in the literature). Similarly, this study identified an enrichment (corrected p = 0.0098) of mutations in synaptic genes, which were defined as being known associates of the N-methyl-D-aspartate (NMDA) receptor or proteins that interact with the activity-regulated cytoskeleton-associated protein (ARC) complex.

Georgieva et al. (2014), compared the role of *de novo* variants in SCZ versus BD (368 BD and 76 SCZ probands respectively and all parents) (Georgieva et al., 2014). This study identified a significant increase in the rate of *de novo* CNVs in SCZ probands vs. BD probands, and SCZ and controls. They also concluded that although there was a higher rate of *de novo* CNVs in BD patients versus controls, the difference was not significant, and therefore *de novo* CNVs are likely play a smaller role in the aetiology of BD than SCZ (Georgieva et al., 2014).

1.3.2 Rare variants

McClellan and King are advocates of the common disease rare variant hypothesis (McClellan et al., 2007), which proposes that complex disorders such as BD and SCZ are caused by rare variants of intermediate or large effect size, which segregate in families with incomplete penetrance (Figure 1.2). This theory suggests that these variants, although individually rare (often family specific), occur at multiple loci, each locus explaining a fraction of cases, which collectively explain a substantial proportion of heritability of these disorders (McClellan and King, 2010b).

Most human variation is very ancient (roughly 90%), having occurred millennia before humans first migrated out of Africa (McClellan and King, 2010a). The recent

exponential growth of the human population has resulted in many rare (present in a specific population, sub-population or only on a family level) alleles (Keinan and Clark, 2012). As many disease variants of intermediate and high effect confer reduced reproductive viability, as in SCZ (Visscher et al., 2012), these variants would be less likely to be transmitted down generations. Such variants could therefore be disproportionately rare compared to other complex traits.

While Mendelian forms of AD have been identified ((Blennow et al., 2006); (Bertram and Tanzi, 2008)) and ASD is a co-morbid feature of over one hundred Mendelian diseases ((Sullivan et al., 2012); (Betancur, 2011)), no examples of Mendelian forms of SCZ, MDD, or BD have yet been discovered. This could be because these diseases do not follow that particular genetic architecture, or because limited study designs have prevented the discovery of this form of disease. However, examples of families multiply affected with psychiatric disorders have been described in the literature. Most recently, a study of 40 families multiply affected by BD (Ament et al., 2015) identified uncommon and rare variants that influence disease risk. Other examples include the study of BD in a large old order Amish family (Georgi et al., 2014) and a large Scottish family multiply affected by SCZ, MDD and BD ((Millar et al., 2000); (St Clair et al., 1990)). While the large Scottish family study identified a balanced translocation that segregates with psychiatric illness (maximum LOD = 6.0), both of the other examples require further work to identify the susceptibility variants.

A study by Need et al. (Am J Hum Gen, 2012) analysed sequencing data (whole genome sequencing (WGS) and whole exome sequencing (WES) data) for 166 cases of SCZ and schizoaffective disorder (Need et al., 2012). 5,155 of the variants identified (restricted to nonsynonymous, nonsense or splice variants with MAFs < 0.05 (or <0.3 for recessive model)) were then genotyped in an independent cohort of 2,617 cases and 1,800 controls (cases and controls being of either African American or European ancestry). The first round association study consisted of 337,312 variants identified using sequencing in 166 cases and 307 controls. As no SNPs passed Bonferroni correction ($p < 1.5 \times 10^{-7}$), the authors developed a two-step process: i) selecting a subset of variants (5,155) with either a p <0.05 or being present in >1 cases and no controls followed by ii) genotyping in

additional cases and controls (2,617 cases and 1,800 controls). Once again, no SNPs survived the Bonferroni correction; the best p-value from the combined dataset was 0.0003, for the African American only analysis was p = 0.0006, and the best for European only was $p = 5.9 \times 10^{-6}$. However, while this study had 99% power to detect moderately rare (1%-5%) variants with a relative risk between 2 and 6, most very rare highly penetrant SCZ associated genotypes would not be expected to show a significant association in a dataset the size of their discovery cohort. Very rare variants will also not reach genome-wide significant in the expanded cohort. Need et al. concluded several genetic architectures for SCZ could be excluded based on their analysis:

1. A small number of highly penetrant loci explaining the majority of cases.
2. A moderate number (less than several hundred) of common variants with a low relative risk underlying most cases.
3. Moderately rare variants that have moderate relative risk explaining most cases (goldilocks alleles).

It is likely that there is a high level of locus and allele heterogeneity in SCZ. Need et al. suggested that the majority of SCZ associated variants will be of very low frequency and will be identified through common pathways and genes; however, these results could also imply oligogenic, polygenic or epistatic models (Need et al., 2012).

Rare, family-specific variants of intermediate effect size are not identifiable using current GWAS methods, even with increased sample size. However, studying families, using techniques such as linkage analysis and WGS, can help identify the causal variants unique to specific families and lead to the identification of candidate genes to be studied in other families.

1.3.3 Common variants

The common disease, common variant model assumes that disease is caused by common variants of small to medium affect size (Figure 1.2). These variants are most likely to be identified by GWAS. A large number of psychiatric GWASs have been performed to date (which can be found in the NHGRI GWAS Catalogue ((Welter et al., 2014). Available at: www.genome.gov/gwastudies).

The first successful psychiatric association study was for AD. Variants in and around the APOE gene have been consistently identified as being significantly associated with AD, with a very large effect size (odds ratios 3-4) ((Strittmatter et al., 1993); (Bertram and Tanzi, 2008); (Jonsson et al., 2013)). To date, there are 28 AD GWAS in the NHGRI GWAS Catalogue (www.genome.gov/gwastudies. Accessed [14th August 2015].), which have identified at least 10 additional regions as being significantly associated with AD ((Harold et al., 2009); (Lambert et al., 2009); (Hollingworth et al., 2011); (Naj et al., 2011); (Jonsson et al., 2013); (Perez-Palma et al., 2014)). A comprehensive database of AD associations can be found at www.AlzGene.org (Bertram et al., 2007).

A smaller number of GWASs have been undertaken for ASD, the NHGRI GWAS Catalogue currently listing seven GWAS for ASD (www.genome.gov/gwastudies. Accessed [14th August 2015].). Of these, three identified genome-wide significant associations:

1. Wang et al. (Nature, 2009) performed a GWAS on two discovery cohorts and two replication cohorts of European Ancestry, totalling 3,115 cases and 8,619 controls. This study identified a genome-wide significant hit at chr5p14.1 (combined $p = 2.1 \times 10^{-10}$)(Wang et al., 2009).
2. Anney et al. (Hum Mol Genet, 2010) performed a ASD GWAS on a discovery cohort of 1,385 cases (from 1,369 families) and a replication cohort of 1,086 cases (from 595 families) and 1,965 controls. A genome-wide significant association ($p = 4 \times 10^{-8}$) was found on chr20p12.1 (Anney et al., 2010).
3. Xia et al. (Mol Psychiatry, 2014) performed a meta-analysis on two ASD GWASs for two Chinese cohorts (275 and 136 cases respectively, with 550 and 984 controls respectively) and a replication cohort of European ancestry (1,299 trios), which returned three genome-wide significantly associated variants (p values ranging from $3 \times 10^{-8} – 4 \times 10^{-8}$), all of which map to chr1p13.2 (Xia et al., 2014).

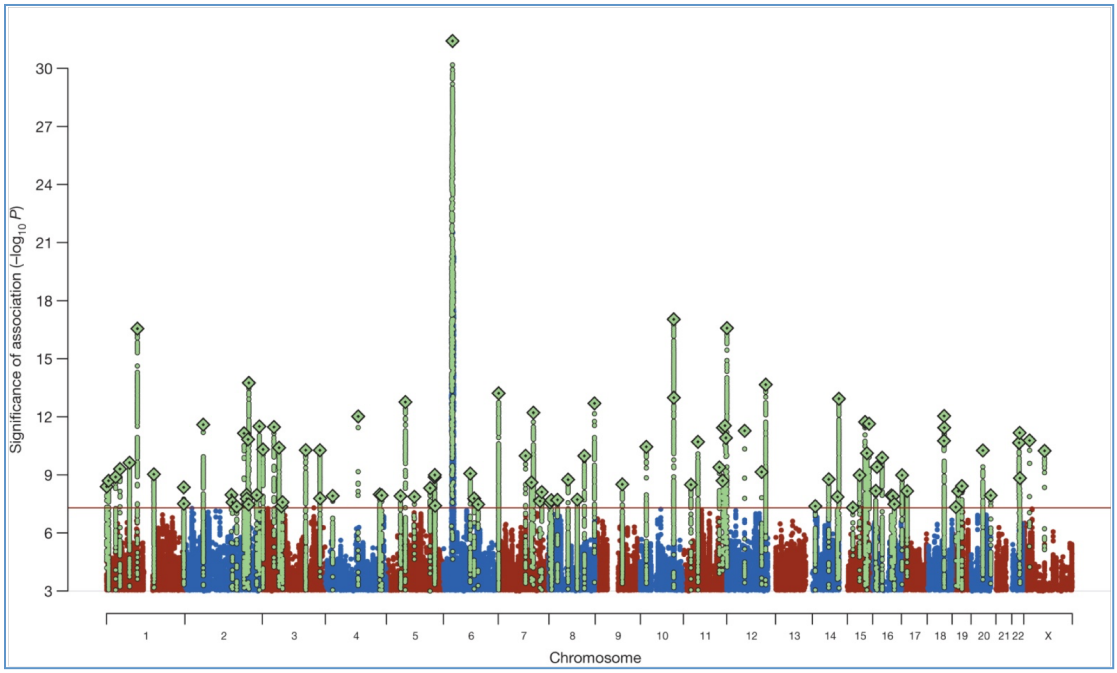


Figure 1.3 Manhattan plot from the largest schizophrenia GWAS to date, showing 108 genome-wide significant loci. X-axis plots SNPs across the genome from chr1 to chrX. Y-axis plots the $-\log_{10}$ p value. Bonferroni correction threshold is marked with a red line. All variants that pass Bonferroni correction are marked as green diamonds. All SNPs in linkage-disequilibrium with the significantly associated SNPs are also coloured in green. Taken from Ripke et al. (2014)(Schizophrenia Working Group of the Psychiatric Genomics, 2014).

In 2014, the PGC published a SCZ GWAS, which studied 36,989 cases and 113,075 controls, the largest psychiatric GWAS to date. This study identified 108 loci significantly associated with SCZ (Figure 1.3)(Schizophrenia Working Group of the Psychiatric Genomics, 2014). Previous to this, ~30 SCZ associated loci had been identified by GWAS (see references 10-23 from (Schizophrenia Working Group of the Psychiatric Genomics, 2014)). While still far behind the success of the SCZ GWAS, five genome-wide significant loci have also been identified for BD (Muhleisen et al., 2014). However, to date only two loci have been associated with MD (consortium, 2015), which have yet to be replicated. The results of GWAS studies for BD and MDD are discussed in detail in the Introduction to Chapter 5.

The lack of any replicated-associations for MDD and the potential causes for this has been discussed extensively in the literature ((Flint and Kendler, 2014); (Levinson et al., 2014); (Major Depressive Disorder Working Group of the Psychiatric et al., 2013); (Wray et al., 2012)). These discussions focus on three main factors, which are potentially applicable to all psychiatric GWASs:

1.3.3.1 GWAS sample sizes are too small

GWAS are designed to capture association signals for common variants. The number of samples affects the ability to detect loci with different effect sizes (Flint and Kendler, 2014). When the effect sizes are small (less than 1.2), more samples are needed to achieve the power to detect significantly associated loci (See Figure 1, (Flint and Kendler, 2014), for a description of the relationship between effect size and sample size for common variants). An example of this has already been seen in the literature for schizophrenia, where ~9,000 cases had the power to detect five genome-wide significant associations. This number increased to 108 when the number of cases was increased to ~35,000 (Levinson et al., 2014). Once the number of cases passed a critical inflection point, as demonstrated in Figure 1.4 (in the case of schizophrenia, this number is ~13,000-18,000 cases), ~4 new hits per 1000 additional cases were observed (Levinson et al., 2014). The MDD working group postulated that a sample size 2.4 times greater than that used for schizophrenia (prevalence 0.007) would be needed for MDD, as this disorder is more common (prevalence 0.15) (Major Depressive Disorder Working Group of the Psychiatric et al., 2013). Furthermore, as MDD has a lower heritability than schizophrenia (0.37 vs. 0.81), as many as 5 times the number of samples used for the schizophrenia GWAS might be needed. This number is debatable; the MDD working group suggested that at least 100,000 MDD samples (and an equal number of controls) would be required (Major Depressive Disorder Working Group of the Psychiatric et al., 2013), while Flint & Kendler (2014) suggested as few as 50,000 MDD cases would be sufficient to detect genome-wide significant associations. Ultimately, these estimates will be dependent on the level of heterogeneity of MDD.

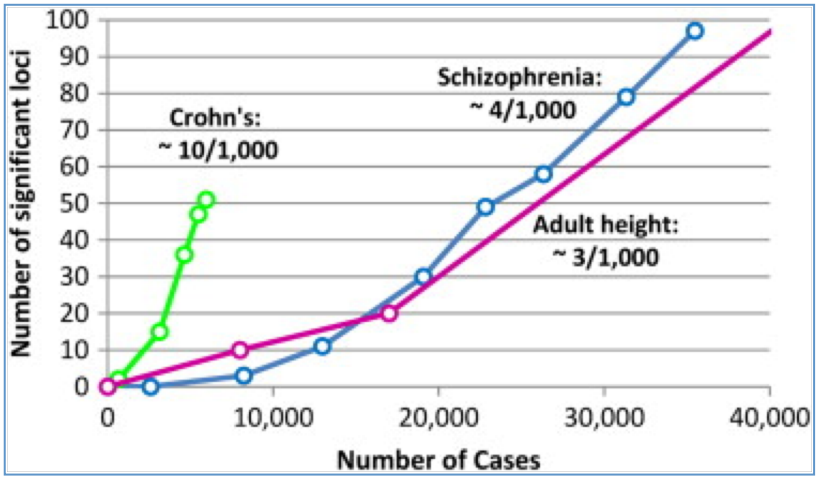


Figure 1.4. A graphical representation of the critical inflection point required for significant associations for different diseases. This figure shows how the number of discoveries is directly related to sample size and that this is not a fixed relationship, but specific to each genetic architecture. Taken from Levinson et al., 2014 (Levinson et al., 2014).

1.3.3.2 The causal variant is not in linkage disequilibrium with any of the markers on the genotyping arrays

If the causal variants are rare variants, they may not be in sufficient linkage disequilibrium (LD) with genotyping variants and therefore they will be below the level of detection of current GWAS methods ((Sullivan et al., 2012); (McClellan and King, 2010b)). In addition, different populations having different genomic patterns of recombination events and haplotypes (Flint and Kendler, 2014); if multiple causal variants occur in the same gene, but these variants are located on different haplotypes, the signal would similarly not be seen in a GWAS. Therefore, population stratification could result in the causal variant not being in sufficient LD with a tagging SNP for the association to be detected by GWAS.

1.3.3.3 Psychiatric disorders are a genetically heterogeneous disorder

Based on the different rates of MDD between men and women (which suggests that although there are shared genetics between men and woman affected by MDD, there are also gender specific genetic determinants), there are likely to be genetically different forms of MDD. Similarly, sub-types of SCZ and BD, which differ in the combination of symptoms or severity of symptoms, are likely to exist. If psychiatric disorders consist of yet to be established subclasses, studying them as a group could reduce power. Identifying a cohort of more phenotypically homogeneous cases could identify a more genetically homogeneous subset of cases and therefore reduce the sample size needed to detect significant associations. One such example would be to focus on hospital based MDD samples, as these individuals tend to represent a more extreme phenotype, with lower prevalence and higher heritability (Wray et al., 2012). Similarly, stratifying cases based on symptoms or shared environment (such as traumatic life events, environmental exposures such as pregnancy (cases of prenatal and post-partum MDD)) can potentially classify MDD cases into sub-types that are more genetically similar to each other. This has proven to be a strong strategy, as the only GWAS for MDD that has successfully identified genome-wide significant loci for MDD was for a homogeneous cohort, consisting of women with recurrent MDD of Han Chinese ancestry (consortium, 2015). This dataset only consisted of ~5,000 cases and ~5,000 controls, showing that a more homogeneous cohort can improve the ability to detect genome-wide significant loci.

1.3.4 Polygenic model

GWAS of complex diseases have identified large numbers of associated loci (over 70 loci for diabetes (Replication et al., 2014) and over 100 loci for schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics, 2014), each of small individual effect size ((Voight et al., 2010); Supplementary Table 2. (Schizophrenia Working Group of the Psychiatric Genomics, 2014)). Lee et al. (2012) analysed SCZ GWAS data (Schizophrenia Working Group of the Psychiatric Genomics, 2014) and calculated the variance explained by autosomal SNPs in a chromosome-by-chromosome manner (Lee et al., 2012). These authors reported that the variance explained by chromosomes is linearly proportional to chromosome length. This is considered

consistent with a polygenic model (Lee et al., 2012). In a polygenic model the coinheritance of multiple common variants of individually small effect size can together push an individual above a particular 'threshold', leading, with environmental triggers, to disease phenotype. Individuals sharing some of these variants who fall below the threshold do not become ill.

Wray et al. (2007) proposed that polygenic risk profiles could be generated from GWAS data, and this can be used to predict disease risk (Wray et al., 2007). Polygenic risk scores (PRS) are a measure of the association of a combination of markers with a trait within an individual. To generate a PRS, markers are selected in a training sample (often ranked on association p-value) using a cut-off (e.g. a p-value threshold). The weighted sum of associated alleles is used to calculate the PRS for each individual, based on the top ranked variants in an independent dataset (Dudbridge, 2013). This approach has been used to generate PRSs for a range of complex diseases including T2D (Lango Allen et al., 2010) and SCZ ((Schizophrenia Psychiatric Genome-Wide Association Study, 2011); (Ripke et al., 2013)). The polygenic component of SCZ has been predicted to be derived from large numbers (potentially over a thousand) variants, which together could account for roughly a third of the genetic liability of SCZ (Kavanagh et al., 2015). In addition, this polygenic component to SCZ has been shown to also contribute to the risk of BD (International Schizophrenia et al., 2009). As with GWAS, the power, sensitivity, and specificity of PRSs are affected by sample sizes (Dudbridge, 2013).

1.3.5 Epistasis

While most GWASs to date have focused on the identification of simple additive effects, the hypothesis being that SNPs exhibit additive, independent and cumulative effects on the trait or phenotype under investigation, there is a lot of debate over the contribution of epistasis to complex traits and diseases (Phillips, 2008). In contrast to the polygenic risk model, which suggests that multiple genes or variants of small effect contribute additively to trait variation, epistasis can be defined as the statistical or functional interaction between two or more loci, where the impact of a genotype at one locus is dependent on the genotype of another locus (or several other loci in the case of multi-

locus epistasis) (Wei et al., 2014). A recent review by Wei, Hemani and Haley (Wei et al., 2014), discusses the likely contribution of epistasis to disease, current methods to detect epistasis in GWAS data, and provides some examples of epistatic interactions associated with disease phenotypes.

The discrepancy between the sum of known genetic effects and the estimate of narrow-sense heritability, also known as the problem of 'missing heritability', may in part be accounted for by epistasis, as regions which may individually fail to pass significance thresholds in GWASs studies, could be shown by their interaction term to contribute to the variance of a trait or disease (Wei et al., 2014). However, caution must be advised before undertaking a search for epistatic interactions, as there are many confounding factors affecting such an analysis. Hypothesis-free methods, which search the full parameter space and compare all pair-wise interactions, are computationally intense and have the potential to suffer from both model complexity and the curse of dimensionality. The solution to both of these problems is very large sample sizes. Hypothesis driven approaches, in contrast, make use of biological priors (candidate gene analysis, pathway analysis, and subsets of GWAS SNPs chosen based on significance thresholds) and can reduce both the search space and the Bonferroni-corrected threshold (by reducing the number of multiple tests ((Carlborg and Haley, 2004); (Liu et al., 2011)).

Wan et al. (2010), Liu et al. (2011), and Lippert et al. (2013) all performed hypothesis-free studies, using the Wellcome Trust Case Control Consortium (WTCCC) study dataset ((Wellcome Trust Case Control, 2007)) to identify genome-wide interaction based associations across seven traits ((Wan et al., 2010); (Liu et al., 2013); (Lippert et al., 2013)). While the methods used by Wan et al. and Lippert et al. reported significant interactions, neither of these studies was able to replicate their findings in independent cohorts. However, the method implemented by Liu et al. identified several pair-wise interactions with Bonferroni corrected $P < 0.05$, which also replicated in independent datasets. In particular, they identified an interaction between *C1orf106* and a novel locus, *TEC*, which was significantly associated with Crohn's disease and which replicated in an independent dataset (Liu et al., 2013). The candidate loci approach has

also successfully been used to identify epistatic interactions influencing the risk of AD ((Combarros et al., 2009); (Rhinn et al., 2013)).

Prabhu & Pe'er (2012) performed a similar analysis on the Wellcome Trust bipolar cohort and were able to identify a significant interaction between SNPs within two genes encoding calcium channel subunits: *RYR2* and *CACNA2D4*. Although they were not able to replicate the exact same SNP-pair interaction, the interaction between these two genes was replicated, suggesting epistasis may exist between these two genes and this interaction could play a role in the aetiology of bipolar disorder (Prabhu and Pe'er, 2012).

While these examples show the potential use of studying the role of epistasis in human disease, it is clear that such analyses suffer from background noise and are affected by large false positive rates. Either increased sample size or reduced candidate variant sets must be used to improve the ability of such methods to identify true epistatic interactions contributing to disease and trait variance (Wei et al., 2014).

1.3.6 Overlap between psychiatric disorders

It is becoming apparent that psychiatric disorders are aetiologically complex, with substantial genetic, locus and allelic heterogeneity ((Sullivan et al., 2012); (McClellan and King, 2010a); (Visscher et al., 2012)), leading to blurring between different diagnoses. Reasons for this are two-fold: firstly, phenotypic similarity and secondly, shared genetic aetiology ((Cardno and Owen, 2014); (Serretti and Fabbri, 2013)).

Defining the phenotypes for each psychiatric illness is difficult. To date, no diagnostic tests, such as those available for many other illnesses (such as heart disease, diabetes and cancer), exist for psychiatric disorders. Instead, the diagnosis of psychiatric illness is based on clinical features. Furthermore, clinical features including psychosis, mood dysregulation and cognitive impairment are diagnostic for several psychiatric illnesses (including SCZ, BD and MDD) (discussed by (Cross-Disorder Group of the Psychiatric

Genomics, 2013)). Greenwood et al. (2012) suggest that the diagnostic systems currently used to define psychiatric phenotypes (such as the DSM-IV system (Wilson and Skodol, 1994)) are not optimised for identifying genetic contributors to psychiatric illness (Greenwood et al., 2012). In addition, the diagnosis of a patient can change over time, based on changes to their symptoms.

A meta-analysis of BD and SCZ (653 BD cases and 13,034 controls; 1,172 SCZ cases and 1,379 controls) identified two genomic loci (9q33.1 and 6q15) that reached genome-wide significance (5.56×10^{-9} and 3.88×10^{-8} respectively), but were not significant for either individual GWAS (Wang et al., 2010). More recently, the Cross Disorder Group of the PGC performed a meta-analysis of the five major psychiatric illnesses (SCZ, BD, MDD, ASD, ADHD) in 33,332 cases and 27,888 controls. This meta-analysis identified four loci that passed genome-wide significance ($p < 5 \times 10^{-8}$) at the chr3p21, 10q24, *CACNA1C* and *CACNB2* gene regions (Cross-Disorder Group of the Psychiatric Genomics, 2013). A second analysis by the cross-disorder group of the PGC compared the genetic variation and the covariance between these five disorders (Cross-Disorder Group of the Psychiatric Genomics et al., 2013). This study made use of genome-wide genotype data for the GWAS meta-analysis of the five disorders previously described (Cross-Disorder Group of the Psychiatric Genomics, 2013) and reported a high correlation of common SNPs between SCZ and BD (0.68 ± 0.04 standard error), moderate correlation between SCZ and MDD (0.43 ± 0.04 standard error), BD and MDD (0.47 ± 0.04 standard error), and ADHD and MDD (0.32 ± 0.04 standard error). Only a low correlation was found between SCZ and ADHD (0.16 ± 0.04 standard error), while none was found between the remaining pair-wise combinations, or against Crohn's disease (Cross-Disorder Group of the Psychiatric Genomics et al., 2013). Similar analyses have been attempted, using the polygenic component of these disorders to predict disease risk across different disorders with mixed success ((International Schizophrenia et al., 2009); (Schulze et al., 2014); (Wiste et al., 2014); (Maier et al., 2015)).

In addition to a high correlation of common variants between unrelated cases of psychiatric disorders, there is evidence that relations of probands affected with one disorder (such as SCZ) are more likely to suffer themselves from another psychiatric

illness (such as BD) compared to population controls ((Lichtenstein et al., 2009); (Wray and Gottesman, 2012)). This further supports the possibility that these disorders share common genetic determinants (Figure 1.5).

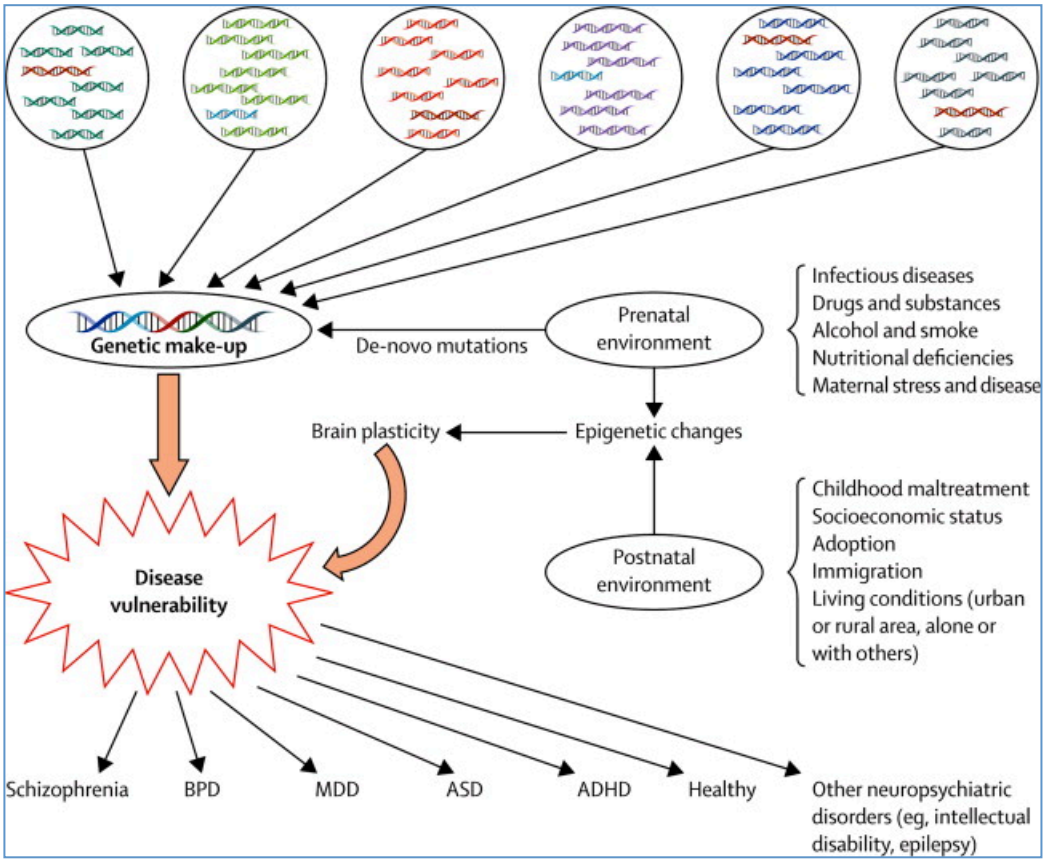


Figure 1.5. This figure outlines a potential shared pathogenesis and aetiology of psychiatric illnesses. The top section of this figure represents six genetic profiles, containing variants that are specific to one profile (one colour) or shared with other profiles (different colour). These profiles represent individuals with different genetic susceptibility to psychiatric illness. These genetic factors, in combination with environmental factors, can lead to disease vulnerability. Different combinations of genetic and environmental factors can present as different psychiatric illnesses, as shown by the black arrows (BPD = bipolar disorder). Taken from (Serretti and Fabbri, 2013).

1.3.7 Complex Architecture

As psychiatric illnesses such as SCZ are associated with increased mortality and reduced reproductive rates, variants with a large effect on the incidence of SCZ could be selected against in the population and so would remain uncommon (low allele frequency). However, it is more difficult to predict the allele frequency of variants with a modest effect on susceptibility, as these variants individually may have only a small effect on fitness, and could in fact have a positive effect on fitness due to their role in other traits (Visscher et al., 2012).

Visscher et al. (2012) proposed two models that might explain the distribution of variants contributing to SCZ susceptibility: the neutral model and the Eyre-Walker model (Visscher et al., 2012). Under the first model, although most variants are rare (roughly 70% of variants having a MAF < 0.05), these variants only explain 10% of genetic variation. Therefore, the majority of variation is due to common variants of small effect size. In contrast, the Eyre-Walker model shows that most of the variance on fitness can be explained by very rare mutations of large effect size, most of which have a MAF < 0.05. These two models largely encapsulate the two sides of the allelic spectrum argument for the genetic architecture of psychiatric illness. These two theories are not mutually exclusive; risk variants are likely to be both common and rare with a range of effect sizes. The limiting factor of current GWAS studies of psychiatric illness is sample size: many variants having population level effect sizes that are too small to pass genome-wide significance thresholds (Baker, 2014). To rectify this, the PGC has focused its efforts on obtaining as many cases and controls as possible. Comparing the most recent PGC GWAS for schizophrenia to the older GWAS study shows a marked increase in the number of significant loci (Schizophrenia Working Group of the Psychiatric Genomics, 2014), which supports the theory that additional loci will be identified for psychiatric disorders in the future.

As can be seen from Section 1.3.3, particularly the SCZ association example, there is a lot of evidence suggesting that more information will be obtained from GWASs of BD and MDD by increasing sample sizes and/or decreasing hetergozygosity. However, the latest GWAS identified 108 significant loci for SCZ, the overall amount of variance

explained is still very small, the odds ratio of each locus being low (in the range of 1-2)(Schizophrenia Working Group of the Psychiatric Genomics, 2014). Therefore, it is likely that only a fraction of the total variance for SCZ, BP and MDD will ever be explained by GWAS and alternative methods are needed to identify the other portion of variation (likely to be caused in part by rare variants of moderate or greater effect). Family studies will be particularly important for this. Linkage analysis can be combined with whole genome sequencing to improve the filtering of candidate variants (Ott et al., 2015). It is also important to remember that variants identified by GWAS mark significantly associated regions and are not necessarily themselves risk variants. Therefore, these regions need to be followed up to identify the true causal variants at these loci.

To summarise, it is very likely that psychiatric illnesses such as SCZ, MDD and BD will be shown to be caused by complex genetic architectures, comprising a range of genetic models. These disorders are likely genetically heterogeneous, consisting of both common and rare variants, both within the same gene (allelic heterogeneity) and across many genes (locus heterogeneity), with a range of effect sizes. It will therefore be important for the study of psychiatric illness not to focus too heavily on any single methodology. Both GWAS and NGS will be important and will play complementary roles in elucidating the aetiology of these disorders.

1.4 Contribution of regulatory variants to disease

The vast majority of variants associated with disease that have been identified by GWAS (over 90%) lie within non-coding regions (Maurano et al., 2012), Similarly, WGS projects (including the 1000 Genomes Project (Genomes Project et al., 2012) have shown that the vast majority of human variation is non-coding (Elgar and Vavouri, 2008). The identification of the phenotypically causal fraction of variants is a major challenge to the study of the genetic basis of human disease. Cooper and Shendure state: “The primary roadblock faced by the field is increasingly one of variant interpretation, rather than data acquisition” (Cooper and Shendure, 2011).

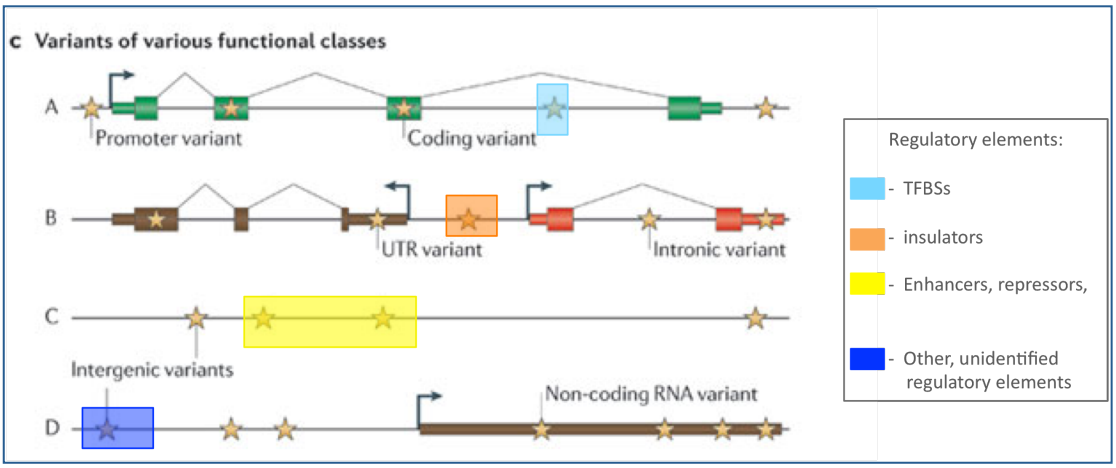


Figure 1.6. This cartoon summarises the various functional classes of variants that can occur in the genome. Functional variants are represented as stars. The top sequence (A) shows an example of a protein-coding gene, containing a SNP within the promoter, two SNPs in exons and an intronic variant, which overlaps a transcription factor binding site. The second sequence (B) shows two genes, the first with a variant in the 5'UTR, the second, with a variant in the 3'UTR. Between the two genes is an insulator, which is modified by a SNP. The third sequence (C) shows an intergenic region, without any genes nearby. Within this sequence there are variants that overlap long-range enhancers, which modify the expression of genes elsewhere in the genome. The last sequence (D) shows variants within a non-coding RNA sequence as well as an intergenic variant shown to alter gene expression but acting via some unknown function. Modified from (Cooper and Shendure, 2011).

By what means can a genetic variant be deleterious? A range of classes of functional variation is summarised in Figure 1.6. Changes to protein coding sequences can result in changes to amino acids, which can affect both the structure and function of a protein. Many tools exist that predict the deleterious consequence of protein changing variation (Table 1.3, taken from (Cooper and Shendure, 2011)). Furthermore, all variants within mRNA coding sequences (untranslated regions (UTRs) as well as protein coding sequence) can affect RNA structure, which in turn can affect RNA stability, localisation, translation efficiency and gene regulation by small RNAs ((Brest et al., 2011); (Mortimer et al., 2014)). However, as non-coding variants do not directly alter an amino acid in the mature protein, it is more difficult to identify the functional from benign variants. Some mechanisms by which non-coding variants may have a functional affect include: changes to exon/intron splicing ((Ward and Cooper, 2010)); or by disrupting

microRNA (miRNAs), long intergenic non-coding RNAs (lincRNAs) and other non-coding RNAs ((Carbonell et al., 2012); (Kumar et al., 2013)). In addition, non-coding variants can function by modulating gene expression, by modifying regulatory elements such as promoter elements (De Gobbi et al., 2006), transcription factor binding sites (TFBSs) ((Zhang et al., 2012b); (Pomerantz et al., 2009)), insulators and enhancers ((Schodel et al., 2012); (Bauer et al., 2013)).

Name	Type	Information	URL	Refs
MAPP	Constraint-based predictor	Evolutionary and biochemical	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	27
SIFT	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://sift.bii.a-star.edu.sg/	39
PANTHER	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://www.pantherdb.org/	41
MutationTaster*	Trained classifier	Evolutionary, biochemical and structural	http://www.mutationtaster.org/	40
nsSNP Analyzer	Trained classifier	Evolutionary, biochemical and structural	http://snpanalyzer.uthsc.edu/	44
PMUT	Trained classifier	Evolutionary, biochemical and structural	http://mmb2.pcb.ub.es:8080/PMut/	38
polyPhen	Trained classifier	Evolutionary, biochemical and structural	http://genetics.bwh.harvard.edu/pph2/	35
SAPRED	Trained classifier	Evolutionary, biochemical and structural	http://sapred.cbi.pku.edu.cn/	42
SNAP	Trained classifier	Evolutionary, biochemical and structural	http://www.rostlab.org/services/SNAP/	36
SNPs3D	Trained classifier	Evolutionary, biochemical and structural	http://www.snps3d.org/	51
PhD-SNP	Trained classifier	Evolutionary and biochemical (indirect)	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html	37

*Also makes predictions for synonymous and non-coding variant effects: for example, splicing. MAPP, Multivariate Analysis of Protein Polymorphism; polyPhen, polymorphism phenotyping.

Table 1.3 Summary table of tools that predict the deleterious impact of protein variants, showing the name of each tool, the type of predictive method utilised, additional information on how the tool predicts deleterious impact and the URL. Taken from (Cooper and Shendure, 2011)

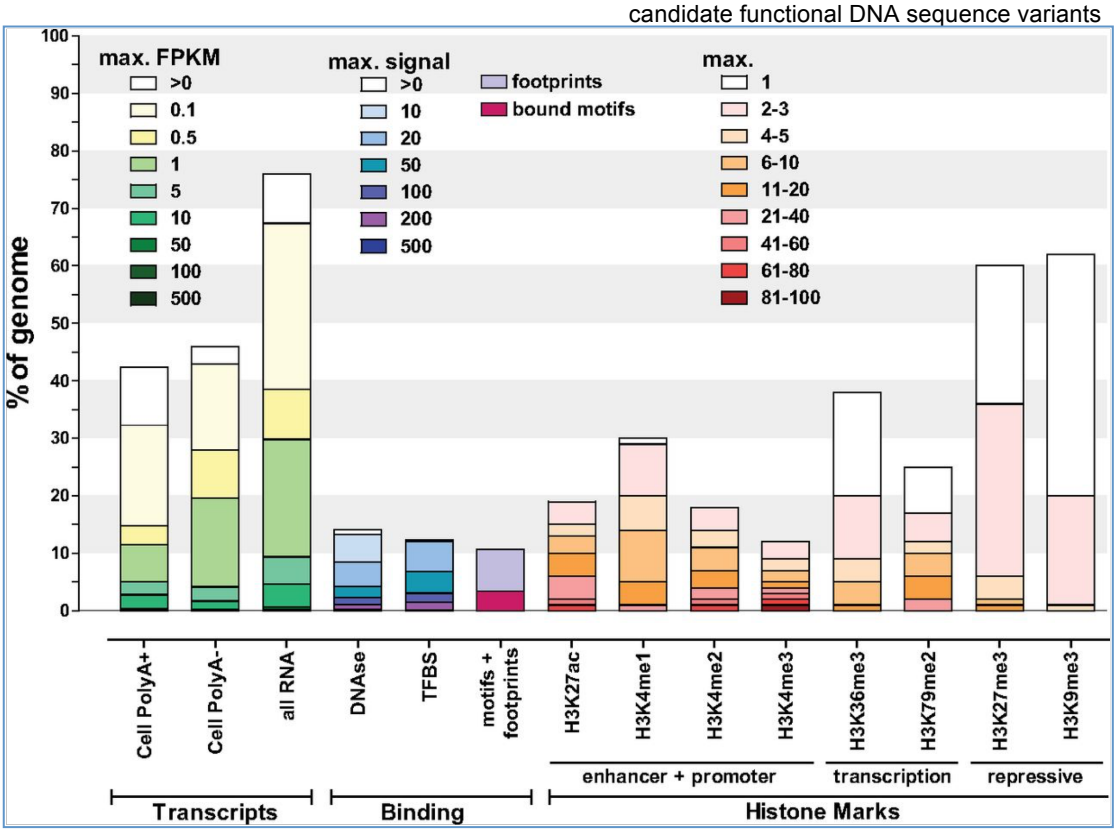


Figure 1.7 This graph shows the proportion of the genome that is covered by biochemically functional elements including transcribed regions, regions bound by DNA binding proteins, and with histone marks known to be associated with functional elements. Taken from (Kellis et al., 2014).

While there is substantial evidence that regulatory variants contribute to human disease (Li and Montgomery, 2013), our ability to detect the functional portion of the genome is limited by both our knowledge of what constitutes a functional non-coding variant and methods to identify those with a deleterious impact.

In 2012 a paper was published that described 80% of genome as being biochemically functional (the ENCODE project Consortium, 2012), defining functional as participating in at least one RNA and/or chromatin associated event in at least one cell type. Kellis et al. (2014) discussed the merits and limitations of this and other definitions of functionality in greater depth (Kellis et al., 2014). Figure 1.7 summarises the proportion of the human genome that is covered by functional elements including transcripts

(RNA), DNA binding sites and histone modifications that mark sites of DNA regulation (including promoters and enhancers). If this data were taken to represent the fraction of the genome that is functional, then indeed, 80% would be an accurate estimation. However, at the other end of the functional spectrum are regions of the genome that are evolutionary constrained. If a region of the genome is conserved between species it is said to be under purifying selection, suggestion mutations at that site will be deleterious (Kellis et al., 2014). If only the portion of the genome that is under evolutionary constraint is considered as functional, only 5% of the genome would be included.

The difference between the upper and lower bounds of predicted functional genomic elements (80% vs 5%) is substantial and highlights our limited understanding of the non-coding portion of the genome. In addition, as so much of the genome can be assigned as being "functional" based on biochemically functional, the search for disease variants amongst these functional candidates could be compared to searching for a needle in a haystack, where, once the haystack has been removed, there remains a stack of needles (Cooper and Shendure, 2011). These needles will need to be further whittled down to identify the deleterious non-coding variants. The first step (removing the haystack) is identifying whether a variant has a functional effect and the second, is discovering if this functional effect is deleterious (sorting through the needles). Therefore, currently, all variants implicated by both NGS methods and GWASs must be functionally evaluated, before their role in disease can be confirmed. Experimental methods to predict the functional effect of a variant include *in vitro* investigation to determine the molecular consequences of a variant (for instance, whether it alters protein structure, stability, localisation or expression) and *in vivo* modelling in another organism (Cooper and Shendure, 2011). However, *in vivo* and *in vitro* methods are both time-consuming and expensive to perform on large numbers of variants.

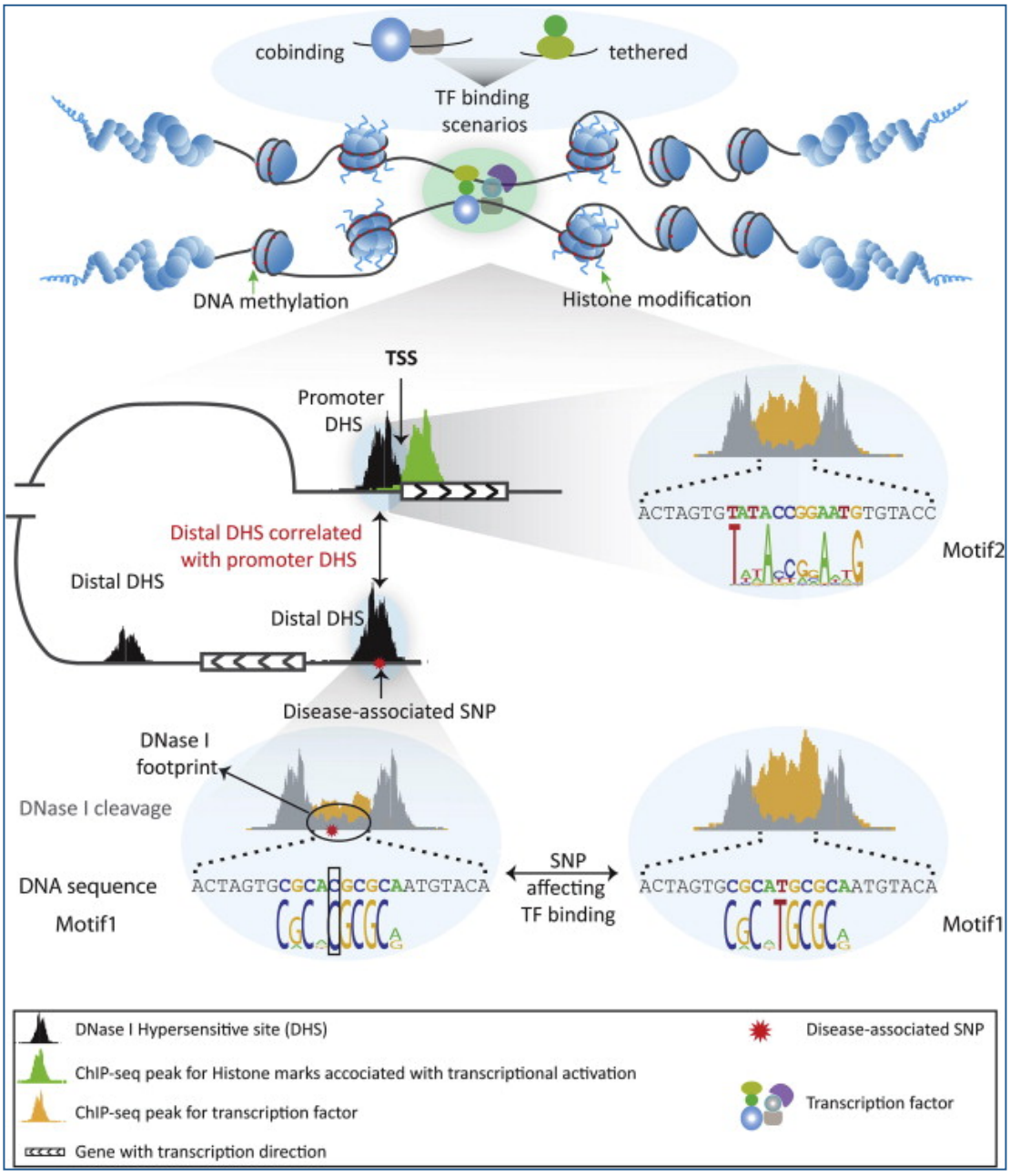


Figure 1.7. This image describes the variety of genomic features that are altered during the regulation of gene expression. Regulatory elements overlap a range of features including TFBSs, DNase HS, ChIP-seq peaks for a range of histone modifications. These data can be used to predict whether a variant overlaps a regulatory element and to predict the likelihood of that variant having a functional or deleterious consequence. Taken from (Qu and Fang, 2013).

Bioinformatics methods can be used to filter candidate variants and reduce the number of variants to test in laboratory-based analyses. However, while predictive methods exist

for assessing the consequence of protein coding variation, tools to analyse the functional and potentially pathogenic consequence of non-coding variants are not as common and have been limited by our knowledge of the mechanisms regulating gene expression. We do know that transcriptional regulation is controlled by a complicated interaction of regulatory elements and that these elements are correlated with certain genomic features (Figure 1.8). Transcription factors and other DNA binding proteins bind to regions of open chromatin, marked by specific patterns of histone modifications, DNase HS sites and other genomic features. When a variant maps to a regulatory element, such as a TFBS, it can impact upon the binding ability of the transcription factor, thus altering gene expression, potentially leading to a deleterious effect. These genomic features can be used to help identify functional variants. This topic is discussed in more detail in the Introduction to Chapter 2.

1.5 Thesis Aims

At the time I undertook this PhD project, no suitable bioinformatics methods were available to prioritise non-coding variants from genomics projects.

The first aim of my PhD was to develop a method that would prioritise candidate variants on the basis of their putative functional consequence. This will be covered in Chapters 2 and 3.

Once this method was designed and tested, the second aim of my PhD was to perform a comparative analysis between my tool and the most comparable methods available from the literature. This analysis is described in Chapter 4.

The final aim of my PhD was to apply this method to the study of psychiatric illness, which is detailed in Chapter 5.

Chapter 2: Design of a SNP prioritisation method and a spiking strategy

2.1 Introduction

2.1.1 The problem: identifying regulatory variants

Linkage analysis and genome wide association studies (GWAS) have been extremely successful at identifying genomic regions that harbour genetic variants contributing to a phenotype of interest (Manolio et al., 2009). Technological advances such as next generation sequencing and genotyping arrays have aided this, allowing the fine mapping of regions of interest. Over 90% of disease-associated single nucleotide polymorphisms (SNPs) from GWAS fall within non-coding regions (Maurano et al., 2012), underlining the importance of the regulatory genome and the need for bioinformatics methods to identify regulatory variants. However, the ability to distinguish disease-predisposing non-coding variants from background variants is impeded by our incomplete understanding of regulatory architecture and the fact that genomic signals that characterise functional regulatory variants are not fully defined (Li and Montgomery, 2013). In addition, the molecular consequences of such variants are more difficult to evaluate than those of variants that change the sequence of encoded proteins, leading to a bias in the characterisation of putative causal coding vs non-coding variants. Furthermore, the relatively lower cost of sequencing exomes rather than whole genomes has also played a part in biasing the identification and characterisation of disease variants towards coding SNPs. Nevertheless, recent improvements in sequencing platforms and methodologies are reducing the cost of whole genome sequencing (WGS) compared to exome sequencing, whilst also comparably improving its accuracy (Meynert et al., 2014), leading to what will soon be a tipping point in favour of WGS and therefore the identification of a greater number of candidate non-coding variants.

These technological advances put a growing pressure on our ability to characterise regulatory variants. In particular there will be an increased demand to prioritise candidate causal variants for their likelihood to be pathogenic via computational methodologies, as current experimental assays are too costly and time consuming to perform on large numbers of variants. One method commonly used to computationally

characterise variants is to annotate SNPs using pre-existing genome annotation data, from sources such as UCSC Genome Browser ((Kuhn et al., 2013)) and the Ensembl Genome Browser (Cunningham et al., 2015), and use this information to prioritise putative pathogenic variants.

2.1.2 Publically available genomic annotation data

A limiting factor within the field has been the lack of genomic and epigenomic annotation data to aid the identification of functional non-coding SNPs (Cooper and Shendure, 2011). This issue is currently being addressed; several large consortia have been established with the aim of developing techniques and producing data for the systematic identification and characterisation of functional elements on a genome-wide scale (Cooper and Shendure, 2011). By using a variety of biochemical techniques, such as ChIP-seq, in combination with novel computational approaches, these projects are producing annotation datasets for genomic and epigenomic markers including post translational modifications of histone proteins (including acetylation, phosphorylation and methylation) (Consortium, 2012); DNase hypersensitive sites (DNase HS) ((Degner et al., 2012); (Thurman et al., 2012)]; DNase footprints ((Hager, 2009, Hesselberth et al., 2009)); transcription factor binding sites (TFBSs) ((Neph et al., 2012);(Wang et al., 2012)]; chromatin states (Ernst and Kellis, 2010); enhancers (Ernst et al., 2011); conserved sequences (Davydov et al., 2010); as well as catalogues of SNPs, indels and copy number variants (CNVs) (Genomes Project et al., 2010). A fundamental aim of these projects is to provide the global scientific community with open-source, freely accessible data, promoting a vast wealth of downstream analyses. These data are available through genome browsers such as the Ensembl Genome Browser (Cunningham et al., 2015), the Epigenome Roadmap (Bernstein et al., 2010) and the UCSC Genome Browser (Kuhn et al., 2013).

2.1.2.1 ENCODE ChIP-seq data

The Encyclopedia of DNA elements (ENCODE) project was embarked on in 2003 with the aim of increasing our knowledge and understanding of human biology and disease by delineating all of the functional elements encoded by the human genome (Consortium,

2011)). This ambitious project began with a pilot phase, which took place over 4 years from 2003-2007. The aims of the ENCODE project pilot phase were to: i) functionally characterise 1% of the genome; ii) develop and advance methods for annotating the functionality of the genome; and iii) if successful, scale up for the whole genome, thereby improving our understanding of the genome (organisation, regulation and functionality).

During the pilot phase, over 200 experimental and computational datasets were generated by 35 groups across the ENCODE consortia, with emphasis placed on the development and implementation of standards to ensure high data quality. This work provided a model for the next phase of the project and comprehensive annotation of the entire human genome, while also providing new tools and techniques to analyse the data efficiently, accurately and cost-effectively, in a high throughput approach. The ENCODE project provides genome-wide annotations of candidate functional elements to help better our ability to interpret the human genome ((Qu and Fang, 2013, Consortium, 2012).

Figure 2.1 describes the variety of methods used in the ENCODE project to characterise the genome, including ChiP-seq, which was used to generate data for histone modifications such as H3K27ac, H3K27me3 and H3K36me3, as well as RNA polymerases and certain transcription factors; and DNase-seq and FAIRE-seq, which were used to define regions of DNase hypersensitivity.

These data can be used to identify putative functional elements and regulatory regions such as promoters, enhancers, repressors and insulators, by characterising the chromatin signatures of known elements, as changes to histone methylation and acetylation change the accessibility of the genome. These data can be used to detect putative regulatory elements.

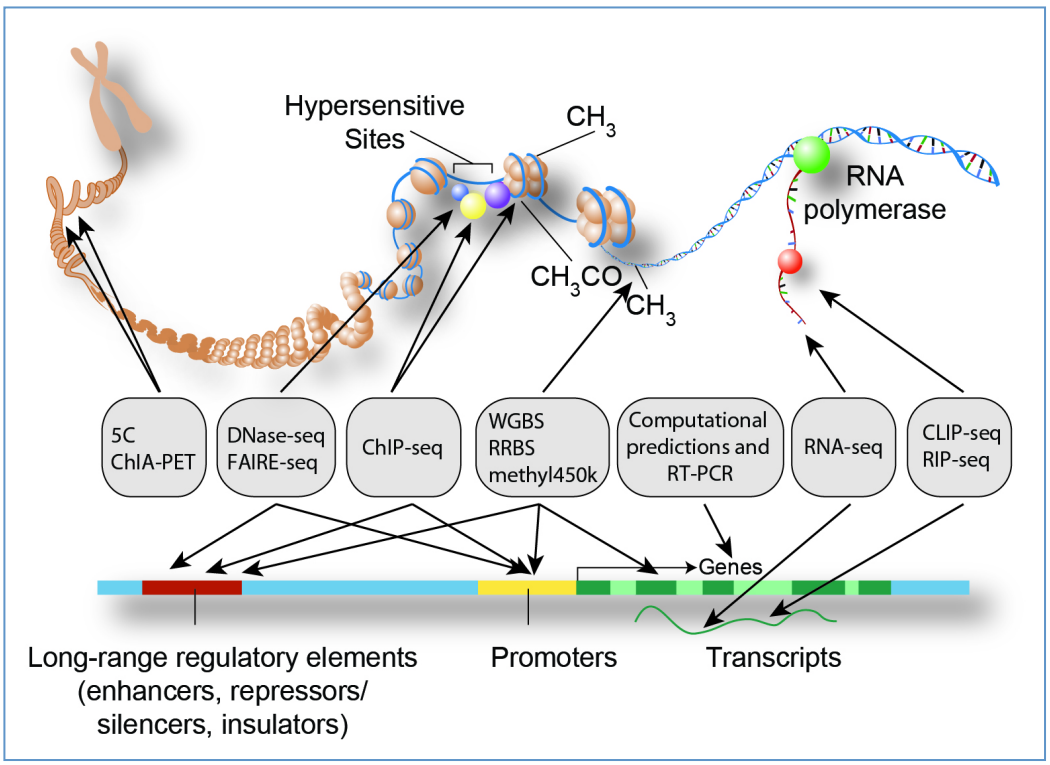


Figure 2.1: Cartoon taken from the User's guide to ENCODE, representing the methods used across the ENCODE consortia to detect functional elements (Consortium, 2011).

2.1.2.2 UCSC genome browser:

The University of California Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/) provides access to a large range of genomic, epigenomic, conservation and sequence annotation data through a series of annotation tracks, which can be used to assign function to both individual nucleotide positions and larger genomic regions. Many different UCSC tracks are available, including information on assembly data, genes, predicted genes and mRNAs, expression, regulation and comparative genomics, and the ENCODE project data (Karolchik et al., 2011). The UCSC table browser allows users to query and manipulate the Genome Browser annotation tables in a flexible, user-oriented manner. It also provides access to the full datasets via an ftp site and MySQL queries.

Population Code	Population Description	Super Population Code	Sequence Data Available	Alignment Data Available	Variant Data Available
CHB	Han Chinese in Bejing, China	EAS	1	1	1
JPT	Japanese in Tokyo, Japan	EAS	1	1	1
CHS	Southern Han Chinese	EAS	1	1	1
CDX	Chinese Dai in Xishuangbanna, China	EAS	1	1	1
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	1	1	1
CEU	Utah Residents (CEPH) with Northern and Western European ancestry	EUR	1	1	1
TSI	Toscani in Italia	EUR	1	1	1
FIN	Finnish in Finland	EUR	1	1	1
GBR	British in England and Scotland	EUR	1	1	1
IBS	Iberian population in Spain	EUR	1	1	1
YRI	Yoruba in Ibadan, Nigera	AFR	1	1	1
LWK	Luhya in Webuye, Kenya	AFR	1	1	1
GWD	Gambian in Western Divisons in The Gambia	AFR	1	1	1
MSL	Mende in Sierra Leone	AFR	1	1	1
ESN	Esan in Nigera	AFR	1	1	1
ASW	Americans of African Ancestry in SW USA	AFR	1	1	1
ACB	African Carribbeans in Barbados	AFR	1	1	1
MXL	Mexican Ancestry from Los Angeles USA	AMR	1	1	1
PUR	Puerto Ricans from Puerto Rico	AMR	1	1	1
CLM	Colombians from Medellin, Colombia	AMR	1	1	1
PEL	Peruvians from Lima, Peru	AMR	1	1	1
GIH	Gujarati Indian from Houston, Texas	SAS	1	1	1
PJL	Punjabi from Lahore, Pakistan	SAS	1	1	1
BEB	Bengali from Bangladesh	SAS	1	1	1
STU	Sri Lankan Tamil from the UK	SAS	1	1	1
ITU	Indian Telugu from the UK	SAS	1	1	1

Table 2.1. The 26 populations included in the 1000 Genomes project. These populations can be grouped into five super populations: African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). Columns 4, 5 and 6 describe whether data is available or not for each population (1 = data is available, 0 = data is not available). This table was modified from the population table provided by the 1000 Genomes project (http://www.1000genomes.org/faq/which-populations-are-part-your-study).

2.1.2.3 1000 Genomes Project:

The 1000 Genomes project is a catalogue of human variation. Human genetic variation was mapped using whole genome and exome sequencing to sequence the genomes of 1,092 individuals from 26 populations (see Table 2.1)(Genomes Project et al., 2012).

This allowed a comprehensive review of the variation that exists across population groups to aid our understanding of how variation contributes to human phenotypes and disease. These data are available through a variety of sources, including a web-interface (http://browser.1000genomes.org/index.html) as well as downloadable bam and vcf format files from the 1000 Genomes FTP site (http://www.1000genomes.org/ftpsearch) and from Ensemble's FTP site (ftp://ftp.ensembl.org/pub/).

2.1.3 Lack of an appropriate method to prioritise variants

The genomic annotation data from these public resources can be used to make informal and adhoc functional predictions. However, manual interrogation of such resources for multiple functional annotations simultaneously does not scale well for large numbers of SNPs spread across a broad genomic region (or genome wide), is unsystematic, lacks reproducibility and is difficult to benchmark (Ryan et al., 2014).

A number of tools have been developed for analysing SNPs. However, the vast majority focus on coding variants, incorporate limited annotation data, are designed for particular analyses (exome data, GWAS) and require highly specific input data (such as rs numbers, p-values or linkage disequilibrium (LD) data). These tools implement a variety of strategies to identify functional SNPs, including use of comparative genomics ((Chun and Fay, 2009)), predicted transcription factor binding sites (Conde et al., 2006), positional information ((Adie et al., 2005); (Xu et al., 2005); (Calabria et al., 2010)), amino acid substitutions (Yandell et al., 2011), chromatin state markers ((Ernst et al., 2011); (Barenboim and Manke, 2013)), p-values from GWAS ((Merelli et al., 2013)) and linkage disequilibrium (LD)((Ward and Kellis, 2012)).

SNP analysis tools can be divided into five main categories: i) variant annotation approaches; ii) GWAS tools; iii) gene based prioritisation methods; iv) exonic variant tools; and v) non-coding variant tools. Here I provide a brief description of some of the SNP analysis methods that were available at the start of my PhD:

2.1.3.1 Variant annotation:

F-SNP: a web based database, integrating information from 16 bioinformatics tools and databases to allow the user to predict the function of SNPs (Lee and Shatkay, 2008). This method does not provide a ranking measure; instead it implements a binary logic where SNPs are classified as being either functional or not functional. To analyse the output, the user must scroll through a tabular output and look at every SNP individually to see their predicted effect, an impractical approach for large SNP sets.

AnnTools: a toolkit for annotating novel and known SNPs and indels that integrates 15 annotation sources (Makarov et al., 2012), including dbSNP minor allele frequencies (MAFs), gene annotation data from UCSC, conserved TFBSs, miRNA binding sites, and promoter predictions. Although this method is presented as being a fast and versatile approach to annotate the full spectrum of coding and non-coding variants (being capable of annotating both SNPs and indels), it does not prioritise variants on putative pathogenicity.

SNPnexus: an annotation tool that links functional annotation data with SNPs across a range of annotation categories including gene annotation (from sources such as Refseq, Ensembl and UCSC); gene consequences (coding, intronic, splice site, untranslated regions, upstream, downstream); protein consequences (synonymous, non-synonymous, stop-gain/loss, frameshift); HapMap population frequencies; conservation scores; and whether variants overlap regulatory elements (predicted TFBSs, Vista enhancers, CpG islands, etc.) (Dayem Ullah et al., 2013). The output table links each query SNP (one SNP per row) with each annotation (series of columns), allowing SNPs to be filtered on any single annotation or combination of annotations. While this method provides useful information, the number of SNPs analysed affects the efficacy of this method: the larger the numbers of SNPs in the input dataset, the larger, and less manageable, the output table. Although filtering on annotations would make the table smaller and easier to handle, filtering could also potentially remove borderline SNPs that may actually be functional. Filtering is therefore less constructive than prioritisation.

HaploReg: an online resource combining conservation, histone modifications, chromatin states and linkage disequilibrium data with regulatory motif prediction algorithms to predict the impact of non-coding variants (Ward and Kellis, 2012). This method is designed for GWAS data, to explore the potential functionality of non-coding variants within disease-associated loci.

rSNPBase: a database of regulatory SNPs (rSNPs) supported by experimental evidence (Guo et al., 2014). This online resource provides annotation data ranging from ENCODE-generated experimental data, TFBSs, DNase hypersensitive sites (DNase HS), miRNA regulatory sites, to SNPs in strong LD (r^2 >0.8) with the rSNPs. This data can be used to identify regulatory SNPs and the genes that they regulate. This method is useful on a SNP by SNP level, where the SNPs of interest are already curated in human variation catalogues such as dbSNP (Sherry et al., 2001). However, this approach does not scale well for comparing multiple variants simultaneously and cannot be used for novel variants.

RegulomeDB: an online database integrating annotation data from six main categories (protein binding, motifs, chromatin structure, expression quantitative trait loci (eQTLs), histone modifications and related data) to assign regulatory information onto any variants (both from sequencing projects and GWAS) (Boyle et al., 2012). As with other annotation methods, this is useful for assigning functional data to SNPs, however it does not report which SNPs are most likely to be putative functional candidates and cannot be used to prioritise variants.

2.1.3.2 GWAS tools:

SPOT: a web tool for the prioritisation of GWAS SNPs for replication studies. This method integrates data from several biological databases and uses the genomic information networks (GIN) prioritisation method to combine the information from these databases, along with GWAS p-values, to prioritise SNPs for further investigation (Saccone et al., 2010). This process is performed for both the original tagging SNP and all LD proxies.

SNPselector: a web tool, designed to select the most appropriate SNPs for association studies (Xu et al., 2005). This method prioritises SNPs across multiple categories including: allele frequency; whether they are the tagging SNP in an LD block; regulatory potential (does the SNP overlap conserved sequences, transcription factor binding sites or CpG islands?); and if it is located within a repetitive element. SNPs are scored on these categories and others and the SNPs are prioritised on this score.

FunctSNP: an R package that links SNPs with functional knowledge, scoring GWAS SNPs based on the functional information associated with them, the total score being the sum of factors including SNP location, type of amino acid substitution (Goodswen et al., 2010). This method was trained on GWAS data, which is inherently a mixed data source, containing many false positives, which could affect the accuracy of the method. In addition, when a GWAS significant SNP does not lie within a gene, FunctSNP automatically links the SNP to the nearest gene and the focus is shifted to the nearest SNP within that gene. This method assumes that the functional variant is a coding variant and ignores the possibility of regulatory variants having a role in disease.

ChroMoS: an integrated web tool for GWAS SNP classification, prioritisation and functional interpretation. This method utilises a MySQL database to provide chromatin state annotations for SNPs from the National Human Genome Research Institute GWAS catalogue. SNPs can also be passed to two additional tools, sTRAP and microSNiPer, which predict differential transcription factor and micro-RNA binding respectively (Barenboim and Manke, 2013).

FunciSNP: and R package designed to move beyond GWAS tagging SNPs to identify candidate regulatory variants (Coetzee et al., 2012). Putatively functional surrogate SNPs in high LD with GWAS tagging SNPs are identified by taking all SNPs in LD with GWAS tagging SNPs and overlapping annotation data for a range of user-defined "biofeatures" (including ENCODE ChIP-seq data for transcription factors; DNase HS sites; CFCF binding sites and annotated promoters).

2.1.3.3 Gene based tools

SNPRanker: a data-mining tool for disease associated SNPs that focuses on target gene prediction (ranks SNPs associated with target genes based on functional evidence) (Calabria et al., 2010). This tool no longer appears to be available from the source website. This could either be because the method is being upgraded and they have removed the old version until the new one is ready, or this method is now defunct.

Residual Variation Intolerance Score: a gene-based score designed for the assessment of how well genes tolerate functional genetic variation, to aid the identification of pathogenic coding mutations (Petrovski et al., 2013). This method ranks genes on the amount of purifying selection acting against functional variation in genes, taking into account both gene size and total mutation rate (both the number of common variants and the number of protein-coding variants), to assess if genes have more or less functional genetic variation than expected compared to the calculated neutral variation rate for that gene.

2.1.3.4 Exonic variant tools:

VAAST: a probabilistic approach that combines elements of aggregative scoring methods and amino acid substitution (AAS) data in a unified framework (Yandell et al., 2011). VAAST can be used to prioritise coding and non-coding variants; however, as non-coding variants cannot be scored using the AAS approach (as they do not encode amino acids), VAAST uses two different approaches to assess the deleteriousness of coding and non-coding variant. Instead, non-coding variants are scored using a log-likelihood ratio combining allele frequencies in cases and controls; an estimate of the impact of non-coding and synonymous substitutions called Normalized Mutational Proportion (NMP), based on the frequency of codons in the human genome aligning with primate genomes and the proportion of occurrences of each of these codon pairs occurring across primate alignments; conservation estimates around DNase hypersensitive sites; and transcription factor binding sites defined by ENCODE regulation data, focusing on elements conserved across primate alignments.

EXtasy: a ranking method for the prioritisation of non-synonymous SNPs (Sifrim et al., 2013), making use of annotation data including but not limited to: allele frequency; conservation; sorting tolerant from intolerant (SIFT) scores and PolyPhen scores; deleteriousness prediction scores from the dbSNP database; and haplo-insufficiency scores. This method is available both as a web interface and as a downloadable, stand-alone program. The developers of this method have shown that it performs well compared to its individual component parts, suggesting that combining multiple annotation data can provide better sensitivity and specificity than each annotation in isolation.

2.1.3.5 Non-coding variant tools

RAVEN: regulatory analysis of variation in enhancers (RAVEN) is a web-based application that combines phylogenetic footprinting and TFBS prediction methods to aid the selection of candidate regulatory variants for follow-up analysis (Andersen et al., 2008). The user selects a gene of interest and RAVEN provides a graphical view of the region proximal to the chosen gene, highlighting the dbSNP variants within this region, as well as predicted TFBSs, conservation scores (as defined by PhastCons scores) and any repeat sequences in the region. This method does not rank or prioritise SNPs of interest. Instead, it provides information on all of the potentially regulatory SNPs within a selected locus.

Pupasuite: a web based SNP analysis tool that prioritises SNPs on factors including LD, MAF, validation status, variant type, and a small selection of putative functional properties including if the SNPs are known to be pathological (compared against a reference list of confirmed pathogenic variants), or occur at exon/intron boundaries (Conde et al., 2006). This tool uses TransFac (Wingender, 2008) to predict if non-coding variants overlap TFBSs, however, this feature is limited to SNPs within 10kb upstream of transcription start sites. This method is therefore not suitable for the analysis of variants further than 10kb upstream of known genes.

Weka: this method was trained using machine learning and a true positive dataset of real, biologically active regulatory variants and a background set of non-coding variants to train a model for predicting regulatory polymorphisms (Torkamani and Schork, 2008). Torkamani and Schork (2008) were among the first researchers in the field to make use of the ENCODE data to predict functionality. These authors compared over three hundred ENCODE feature sets and used machine learning to reduce this to a smaller, more usable set. In this study they used a good statistical framework consisting of hold out datasets and cross validation; however, the true positive data set was small (104 true variants), affecting the accuracy of the measured performance. In addition, this method is not formatted as a tool for predicting the functionality of regulatory polymorphisms. Instead the authors provide the software platform (Weka) used in their analysis and the input data to allow their method to be reproduced. This prohibits this method from becoming a field standard approach and makes it inaccessible to the majority of inexperienced scientists (particularly for bench scientists with limited bioinformatics experience who wish to test their data).

To summarise, many bioinformatics approaches have been designed to identify (or prioritise) tagging SNPs from GWAS studies, exonic variants, or purely to annotate genomic variants. However, at the start of this project, there was no effective method for the prioritisation of regulatory variants from NGS projects. Nor was there a gold standard protocol for testing and comparing methods, making a comparison between different strategies subjective (focusing on limited data, biased to a particular variant class, genomic region, or disease).

Therefore, there existed a need for a simple, robust system that can combine a range of annotation datasets, along with other genomic functional measures, to prioritise candidate variants for follow up analyses. I proposed to address this by designing my own method. This method would be appropriate for all genomics projects, capable of handling both coding and non-coding variants in a single analysis, and not be limited by distance to the nearest transcription start site (TSS).

Three questions to ask when developing any new prioritisation approach are:

1. Which annotation features should be included in the model?
2. How should these features be combined into a single pipeline?
3. How can this method be tested to assess its ability to prioritise functional variants?

The following sections will deal with these points.

2.1.4 Annotation features:

2.1.4.1 Conservation

Regions of the genome that are conserved across species are said to be under purifying selection, suggesting that mutations at these sites may have a deleterious impact to the organism (Kellis et al., 2014). Therefore, DNA elements with important functions are often conserved across species (Cooper and Shendure, 2011).

Many algorithms have been designed to take advantage of this feature, making use of the availability of fully sequenced genomes of over 46 species (Cooper and Shendure, 2011). A key decision to be made when performing a cross-species analysis is the phylogenetic scope one should use . If this is too broad (e.g. humans to yeast) many true functional sites are likely to be missed as they are unlikely to be conserved over such a great evolutionary distance; if too narrow (e.g. humans to primates) many non-functional sites will appear to be conserved as they have had insufficient time to diverge. In addition, a choice must be made as to which conservation methods to use. Conservation algorithms can be broadly divided into two groups: those that assign a conservation score to individual nucleotide positions and those that use a sliding window to assign a score to a small region.

Genomic Evolutionary Rate Profiling (**GERP**) calculates levels of evolutionary constraint on a position specific level, based on an alignment of 35 mammals to the hg19 release of the human genome (https://genome.ucsc.edu/cgi-bin/hgTables). Candidate constrained elements are identified by annotating regions that show a lower number of substitutions than expected. Each element is assigned a rejected substitution (RS) score,

in proportion to the magnitude of the substitution deficit obtained ((Davydov et al., 2010); (Siepel et al., 2005)).

PhastCons and **PlyloP** are part of the PHAST, 46way conservation package, based on hidden Markov algorithms. The first looks at conservation of the region containing each variant while the second looks at the conservation of the specific base. Both tools can be run on three subsets of organisms: primates, placental mammals and vertebrates (King et al., 2005).

2.1.4.2 Chromatin States

While the genomic signatures of each histone modification can provide a certain amount of data, layering them together can increase precision and specificity (Ernst et al., 2011).However, our knowledge of how best to combine these data is limited. To correctly define the combinations of raw histone marker ChIP-seq data would require a lot of additional experimental and computational work far beyond the scope of this project. An alternative option was to use predefined data. One such source of data comes from the labs of Kellis and Berstein (Ernst et al., 2011), who systematically mapped nine chromatin marks across nine cell lines, and developed a multivariate hidden Markov model to distinguish different chromatin states, through recognition of combinatorial patterns of the chromatin marks. These data have been rigorously tested and confirmed by in vitro assays (Ernst et al., 2011). 15 chromatin states were predicted, including active promoter, weak promoter, strong enhancer, transcriptional elongation, polycomb repressed, and repetitive /copy-number-variant.

2.1.4.3 DNase hypersensitivity

DNase I hypersensitivity is a universal feature of active cis-regulatory sequence and has long been used to map general chromatin accessibility**.** The use of this method has led to the discovery of functional regulatory elements that include enhancers, insulators, promoters, locus control regions and novel elements (Thurman et al., 2012). The DNase

Clusters track on the UCSC Table Browser contains genome wide data assayed in 74 cell types, pooled together into a single value, as part of the data generated by the University of Washington ENCODE group (Sabo et al., 2004).

2.1.4.4 Repetitive elements

Repetitive sequences present many technical challenges for the alignment and assembly of next-generation sequencing data (Treangen and Salzberg, 2012). As nearly 50% of the human genome is derived from repeats this can be a major influence on the reliability of downstream analyses including SNP calling. Various tools have been designed to overcome this issue at the sequence alignment and assembly level and recommendations have been made for quality control settings to improve the accuracy of the SNP calling. The RepeatMasker track on the UCSC Table Browser was created using the program RepeatMasker, written by Arian Smit (Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org. 1996-2010.). This tool is used to annotate repetitive elements present within the query sequence. The data generated includes a column describing the type of repeat element identified (repClass).

2.1.4.5 Mapability

Mapability provides information on the align-ability and uniqueness of sequences based on the hg19 release of the human genome. Each 20bp sequence is assigned a score between 0 and 1: a sequence will score “1” is it is unique and “0” if it occurs four or more times in the genome (UCSC Table Browser Schema, (Karolchik et al., 2011)), and a score of 0.5 indicates the sequence occurs exactly twice, while a score of 0.33 indicates three times and 0.25 four times.

2.1.4.6 Position

The position of a SNP relative to that of genes is an important factor to be considered when scoring SNPs, as studies have shown that the position of SNPs relative to genes affect the likelihood of that SNP being causal: i) a regulatory site’s influence on

expression falls off almost linearly with distance from the TSS in a 10kb range (MacIsaac et al., 2010); ii) disease associated SNPs are significantly enriched within strong enhancers (Ernst et al., 2011); iii) non-coding disease-associated GWAS variants are concentrated in DNase HS sites (Maurano et al., 2012); and iv) DNase HS are known to overlap with regulatory elements (Vernot et al., 2012). This information can therefore be used to predict the probability of a variant being pathogenic.

2.1.4.7 Allele frequency

The frequency of a variant within a population or across populations can aid our understanding of the phenotypic impact of that SNP. For instance, a common SNP is unlikely to be a causal variant for a highly penetrant, Mendelian disease. Understanding the disease model under investigation can allow scientists to hypothesize the frequency of a causal variant in the population and use this information to reduce the number of potential variants to investigate.

2.1.4.8 Chromosome region score

Some studies, such as GWAS and linkage studies, provide information on which parts of the genome are most likely to contain the causal variant for a disease of interest. These data can be used to filter out candidate SNPs that are not located in the region of interest, or to give them lower priority.

2.1.5 Combining features into a model framework

Defining the features to be included in a prioritisation method is an important aspect of developing such a method. Equally important is the selection of a model framework to combine these features. Simply combining them together in a 1:1:1 ratio is not a viable approach, as the data are all on different scales (for instance, 0-1 versus 0-1000). The SNP analysis tools described in section 2.1.2 were all developed using different approaches and do not provide a consensus on the best way to design a variant prioritisation method. An aim of this chapter is to develop an appropriate model framework and test it.

2.1.6 Test datasets

A universal challenge facing developers of new bioinformatics tools is how to test the performance and accuracy of their methods. Meaningful benchmarking requires test datasets comprising large numbers of both "true positives" and "true negatives". For the development of a SNP prioritisation approach, these would be, respectively, variants that have been shown experimentally to be deleterious and variants that have been shown not to have a negative effect, both datasets containing both coding and non-coding SNPs. A frequent problem with designing such a dataset is that it is difficult to determine with absolute certainty whether a variant is a true negative.

Potential sources of true positive variants for this project fall into three main categories: i) large-scale databases (Cooper and Shendure, 2011); ii) single locus and/or disease specific databases; and iii) repositories of experimentally validated variants compiled by single research groups. There is much debate over which type of database contains the best data.

Many authors have argued for a focus on locus specific databases (LSDB's) (reviewed by (Samuels and Rouleau, 2011)), containing variants experimentally proven to be associated with a specific disease. These databases are usually fully accessible to the public and, being primarily maintained by academic researchers, tend to be more comprehensive and accurate than their larger scale counterparts, which contain variants with mixed levels of evidence supporting their functional and pathogenic effects (Samuels and Rouleau, 2011). They do, nevertheless have drawbacks, most pressing being they tend to suffer from limited sample sizes, an unavoidable consequence of focusing on a single disease or locus. There is also no universal standard regulating the design and curation of such databases, meaning there is a lot of variability in the standard of these databases. As with larger databases, they are also biased towards coding variants. Some good examples of LSDB's are the Cystic Fibrosis Mutation Database, which catalogues variants associated with the cystic fibrosis trans-membrane conductance regulator gene (CFTR) locus (http://www.genet.sickkids.on.ca/app); the HBB variants catalogued in the HbVar database of Human Haemoglobin Variants and

thalassaemia mutations (Giardine et al., 2007); and the Fanconi Anaemia Mutation database compiled by the Rockefeller University ((Samuels and Rouleau, 2011); (Cotton et al., 2008)).

The advantages of large-scale databases such as the Online Mendelian Inheritance in Man (OMIM) database ((Amberger et al., 2015)) and the Human Gene Mutation Database (HGMD)(Stenson et al., 2003) are that they contain much larger numbers of variants; however a lot of variation exists between such databases, each missing some data (genes or variants) that the other provides. For instance, OMIM only includes select mutations for each gene, chosen (amongst other criteria) based on phenotypic impact, population frequency and historical significance. HGMD is available in two forms: a comprehensive, professional version, which requires a (expensive) license to access; and a smaller, more outdated public version. In contrast to the professional version, the public database cannot be downloaded as a whole from the HGMD website, most likely to encourage use of the professional version. Both versions report variants that are not necessarily disease causing, for instance GWAS SNPs and SNPs with experimental evidence of functionality but no known link to disease.

2.1.6.1 HBB dataset

Haemoglobinopathies, which affect the structure and function of haemoglobin molecules, are among the most common hereditary disorders in humans and are caused by mutations in the α- and β-globin gene clusters (Giardine et al., 2007). One example is beta-thalassaemia, a Mendelian disorder characterised by changes to the synthesis of the β-globin chain, causing either a structural change, affecting how it binds to α-globin, or by changing the quantity produced. This results in an imbalance in globin chain production and a reduction in the amount of mature haemoglobin A produced, leading to abnormal erythropoiesis (Amberger et al., 2015). Mutations at the HBB locus lead to a variety of phenotypes, including the Beta-thalassaemias and sickle cell anaemia.

The HbVar database is a locus-specific database of human haemoglobin variants that underlie thalassaemias and haemoglobinopathies (Giardine et al., 2007). Developed in

2001, HbVar is known for being well maintained, comprehensive and simple to use (Samuels and Rouleau, 2011). The phenotypic heterogeneity reflects the heterogeneity of mutations at this locus. The variants catalogued in this database can be queried using several approaches including: by type of thalassaemia; gene name; globin chain; where the variant is located within gene (exon, intron, un-translated region (UTR); non-coding); ethnic occurrence; or a combination of factors. Compared to other LSDBs, HbVar has the added advantage of allowing variants to be batch downloaded. In addition, the variants catalogued in HbVar are supported by extensive documentation, including the biochemical and phenotypic effects of each variant (including references), allele frequency. This database has a limited number of variants as it only focuses on a specific disease locus; however, as it is a highly accurate and well-maintained database, it is a very good candidate database for my analysis (Giardine et al., 2007).

A shortcoming of HbVar is that the majority of variants are coding variants. This is largely due to the genetic architecture of thalassaemias, although it is also in part due to acquisition bias and the greater ease in identifying coding variants. Some non-coding variants included within the database may have been identified concurrently with coding variants (i.e. being secondary or tertiary modifiers) or leading to specific functional effects, such as only affecting TFBSs or strong enhancers, and so may not provide a good representation of the global diversity of regulatory polymorphisms.

2.1.6.2 RAVEN dataset

Andersen et al. (2008) performed a literature search to identify regulatory SNPs to build their own training dataset to test their bioinformatics approach, RAVEN (see Section 2.1.2)(Andersen et al., 2008). By searching for regulatory variants themselves, they were able to specify strict search criteria and so ensure all variants included in their dataset were experimentally verified (either by luciferase assay, in vitro electrophoretic shift assays or by showing allele-specific binding to nuclear extracts); likewise, they were able to restrict the variants to those within 10kb upstream of the TSS of human genes with available human-mouse orthologs. In this manner they identified 104 variants that matched their strict criteria and which could be used to train their application. In addition, they developed a background variant dataset, consisting of SNPs similarly

restricted to within the 10kb region upstream and downstream of genes with known mouse orthologs. All 4,000 background SNPs were common SNPs, with a MAF >0.05 from dbSNP (Andersen et al., 2008).

2.1.6.3 Spiking strategy

The HBB and RAVEN datasets both represent very specific classes of test data. I also designed a strategy for building a new class of test datasets, whereby known variants could be spiked into an unrelated background variant set. This spiking strategy would allow me to assess how well the model can discriminate real functional variants against any arbitrary background. As I already had at my disposal two positive (known functional) variant datasets, I only required additional background data to spike these variants into.

2.1.6.4 SBF2 4p background dataset

I had access to WGS data for five individuals from a family with multiple instances of bipolar disorder and evidence of linkage to a locus on chromosome 4 (see Chapter 5 for more details): three affected carriers of the disease haplotype and two unaffected, married-in relations. The SBF2 linkage-region is defined as an approximately 20 Mb region on chromosome 4 (Le Hellard et al., 2007). Using this data as a background variant dataset allowed me to assess the ability of my model to prioritise the HBB and RAVEN variants in a novel genomic context. As this region is likely to contain at least one putative candidate SNP for bipolar disorder, the "known" SNPs to be spiked into this background set will have to compete with other functional variants. This would suggest that the performance of the prioritisation model on this data would be negatively affected and the performance measures would be lower than for the two known variant datasets against their own background variants.

2.1.6.5 ENCODE pilot project background dataset

The ENCODE pilot project generated annotation data across 44 genomic regions, covering roughly 30 Mb of DNA, 1% of the genome. These regions were chosen to

cover a wide range of genomic contexts. 50% were chosen around medically important genes (or other known sequence elements) whose biology has been well studied, while the other regions were selected using a stratified random-sampling strategy, based on two parameters: gene density and non-exonic conservation scores.

I generated a background dataset consisting of all of the 1000 Genomes EUR variants from within the coordinate boundaries of the 44 ENCODE pilot project regions. Details of the 44 regions can be found in Table 2.4. Each of the 44 regions can be used individually as background datasets with different genomic contexts to spike true positive variants into. Similarly, these regions can be merged to test performance on a single, large, heterogeneous dataset.

2.1.7 Summary of chapter aims

The first aim of this chapter was to develop a method to prioritise candidate pathogenic variants.

The second aim was to develop a performance evaluation strategy that can assess the prioritisation method and predict how well it will perform on novel data, across a range of genomic contexts.

2.2 Methods

2.2.1 Annotation data sources and data management

The annotation data pertaining to the HBB, RAVEN and Scottish bipolar family 2 (SBF2) datasets were obtained and processed by Stewart Morris (SM) into spreadsheet format, the first column containing the SNP reference (coordinates based on the hg19 build of the human genome) and each additional column containing the annotation data for each SNP for all of the collected annotation features. I developed a shell script combining MySQL, awk and bedtools commands to generate the data tables for each feature annotation (see Appendix A).

These features were obtained from a variety of sources including the 1000 Genomes repository of human variation, the ENCODE project, and UCSC Genome Browser and included four conservation tools (GERP, PhastCons, 7xregpotential and PhyloP), DNase HS data, minor allele frequencies, chromatin states, repetitive elements, mapability and position relative to genes.

2.2.2 Correlation analysis

To select the best conservation tools to include in the prioritisation method, I performed correlation analysis using the statistical software R (version 2.14.0). Using the function *cor()* and the method "spearman", I calculated Spearman's rank correlation coefficients, rho scores. In addition to producing rho values for this correlation analysis, this method also reports p-values on the significance of the relationship.

2.2.3 Data preparation

2.2.3.1 HBB

I searched the HbVar database (accessed: November 2011) for variants specific to the HBB gene using the command:

```
>"name like 'HBB' AND any substitution"
```

This returned 767 variants, which I downloaded as a tab-separated text file and converted into an excel spreadsheet. Of the 767 disease-associated variants available for

this region, I excluded roughly half because they were not SNPs (either small indels or other mutations), leaving 363 disease SNPs. An additional 141 background SNPs were identified within this region using the dbSNP archive of genetic variation, giving a total benchmarking dataset of 504 SNPs. Functional annotation data for these variants were collected and processed by SM.

2.2.3.2 RAVEN

The RAVEN dataset was made available as part of the supplementary material for the paper by Andersen et al. (2008), in the format described in Table 2.2. From this file, I extracted the "chromosome position in Hg17" column, which I separated into a chromosome column and a coordinate column for conversion into Hg19. Conversion from Hg17 to Hg19 was performed by SM. Due to changes between Human Genome releases, my final list of true positive regulatory SNPs was 95 and my total SNP set was 4,085.

Gene	PubMed ID	dbSNP_ID	Chromosome_position_in_HG17	Allele_1	Allele_2	source
AGER	11375354	rs1800624	chr6:32260365	CCCAGCCTTGCCTTCATGATGCAGGCCCAAATGC ACCCTTGCAGACAACAGTCTGGCCTGA	CCCAGCCTTGCCTTCATGATGCAGGCCCAATTGC ACCCTTGCAGACAACAGTCTGGCCTGA	Rockman and Wray
APOC3	8675624	rs2854117	Chr11:116205352	TAACCAGGCCTTGCCGGAGCCACTGATGCCTGG TCTTCTGTGCCTTTACTCCAAACACCCC	TAACCAGGCCTTGCCGGAGCCACTGATGCCCGG TCTTCTGTGCCTTTACTCCAAACACCCC	Rockman and Wray
APOC3	8675624	rs2854116	chr11:116205379	GCCTGGTCTTCTGTGCCTTTACTCCAAACACCCC CCAGCCCAAGCCACCCACTTGTTCTCA	GCCTGGTCTTCTGTGCCTTTACTCCAAACATCCCC CAGCCCAAGCCACCCACTTGTTCTCA	Rockman and Wray
APOE	9468288	rs449647	chr19:50100404	GTTTCACCATGTTGGCCAGGCTGGTCTCAATCTC CTGACCTTAAGTGATTCGCCCACTGTG	GTTTCACCATGTTGGCCAGGCTGGTCTCAAACTC CTGACCTTAAGTGATTCGCCCACTGTG	Rockman and Wray
APOE	9468288	rs405509	chr19:50100676	CCAGGAAGGGAGGACACCTCGCCCAGTAATACA GACACCCTCCTCCATTCTGGGGGCCAAG	CCAGGAAGGGAGGACACCTCGCCCAGTAATCCA GACACCCTCCTCCATTCTGGGGGCCAAG	Rockman and Wray
CD14	11698458	rs2569190	chr5:139993100	AGATGCCCTGCAGAATCCTTCCTGTTACGGCCCC CCTCCCTGAAACATCCTTCATTGCAAT	AGATGCCCTGCAGAATCCTTCCTGTTACGGTCCC CCTCCCTGAAACATCCTTCATTGCAAT	Rockman and Wray

Table 2.2. This table represents an example of the RAVEN SNP file format. Column 1 lists the gene names associated with each variant; columns 2 shows the PubMed IDs for the analyses where each SNP was functionally assessed; column 3 lists the dbSNP IDs for each SNP; columns 4 and 5 describe the two alleles and the proximal sequence for each SNP; and column 6 relates the reference source for each SNP.

ENCODE names	pick method	size (Mb):	interest	# genes	# genes/reg size	# SNPs	# SNPs/Reg size	NEC:	GD:
ENm001	manual	1.9	CFTR	25	13.16	6286	3308.421053	0.07	0.03
ENm002	manual	1	Interleukin_Cluster	28	28.00	3501	3501	0.06	0.07
ENm003	manual	0.5	Apo_Cluster	9	18.00	2206	4412	0.09	0.03
ENm004	manual	1.7	Chr22	26	15.29	7419	4364.117647	0.05	0.03
ENm005	manual	1.7	Chr21	29	17.06	5834	3431.764706	0.07	0.05
ENm006	manual	1.2	ChrX	53	44.17	2419	2015.833333	0.05	0.10
ENm007	manual	1	Chr19	50	50.00	4733	4733	0.01	0.10
ENm008	manual	0.5	Alpha_Globin	26	52.00	2481	4962	0.03	0.11
ENm009	manual	1	Beta_Globin	47	47.00	4336	4336	0.01	0.06
ENm010	manual	0.5	HOXA_cluster	24	48.00	1625	3250	0.22	0.05
ENm011	manual	0.6	IGF2/H19	23	38.33	2728	4546.666667	0.05	0.05
ENm012	manual	1	FOXP2	5	5.00	3858	3858	0.20	0.01
ENm013	semi-manual	1.1	7q21.13	6	5.45	3808	3461.818182	0.10	0.04
ENm014	semi-manual	1.2	7q31.33	7	5.83	4120	3433.333333	0.10	0.02
ENm111	Random	0.5		2	4.00	2116	4232	0.0280	0.0050
ENm112	Random	0.5		0	0.00	2150	4300	0.0380	0.0000
ENm113	Random	0.5		0	0.00	2101	4202	0.0390	0.0000
ENm114	Random	0.5		1	2.00	1915	3830	0.0280	0.0120
ENm121	Random	0.5		2	4.00	2008	4016	0.0620	0.0230
ENm122	Random	0.5		9	18.00	2151	4302	0.0340	0.0340
ENm123	Random	0.5		3	6.00	1940	3880	0.0170	0.0310
ENm131	Random	0.5		19	38.00	2234	4468	0.0130	0.0460
ENm132	Random	0.5		10	20.00	2251	4502	0.0110	0.0550
ENm133	Random	0.5		7	14.00	1630	3260	0.0230	0.0520
ENm211	Random	0.5		1	2.00	1911	3822	0.0970	0.0050
ENm212	Random	0.5		5	10.00	2408	4816	0.0670	0.0170
ENm213	Random	0.5		2	4.00	1532	3064	0.0740	0.0090
ENm221	Random	0.5		3	6.00	1791	3582	0.0790	0.0220
ENm222	Random	0.5		5	10.00	2000	4000	0.0690	0.0210
ENm223	Random	0.5		14	28.00	2202	4404	0.0640	0.0360
ENm231	Random	0.5		14	28.00	1560	3120	0.1020	0.0840
ENm232	Random	0.5		12	24.00	2144	4288	0.0830	0.0590
ENm233	Random	0.5		19	38.00	1251	2502	0.0970	0.1060
ENm311	Random	0.5		0	0.00	1735	3470	0.1490	0.0010
ENm312	Random	0.5		3	6.00	2093	4186	0.1350	0.0030
ENm313	Random	0.5		0	0.00	1711	3422	0.1540	0.0000
ENm321	Random	0.5		2	4.00	1739	3478	0.1140	0.0320
ENm322	Random	0.5		2	4.00	1865	3730	0.1590	0.0290
ENm323	Random	0.5		6	12.00	1919	3838	0.1860	0.0230
ENm324	Random	0.5		2	4.00	872	1744	0.1070	0.0200
ENm331	Random	0.5		12	24.00	1949	3898	0.1330	0.0910
ENm332	Random	0.5		17	34.00	2089	4178	0.1340	0.0900
ENm333	Random	0.5		17	34.00	1589	3178	0.1150	0.0920
ENm334	Random	0.5		12	24.00	2030	4060	0.1520	0.0480

Table 2.3.Table of information on the ENCODE pilot regions, including information on non exonic conservation score (NEC), % gene density (GD), pick method, number of genes in each region and the number of SNPs (MAF<5% from the 1000 genomes EUR database). The ENCODE pilot regions cover 30Mb of the genome (~1%) and were picked either manually (based around well studied genes or other well-known sequence elements, in regions where a high amount of comparative sequence data had been collected) or according to a stratified random-sampling strategy so as to include representative regions varying in the number of genes and functional elements based on gene density score (percentage of bases covered by exons) and non-exonic conservation score (sharing at least 80% base alignment with the mouse genome).

2.2.3.3 ENCODE pilot project background dataset

To build the ENCODE datasets I extracted SNPs from the 1000 Genomes European (EUR) subpopulation. Table 2.3 describes each of the ENCODE pilot regions. I assembled the SNP functional annotations using a shell script that made use of tools such as mySQL, bedtools, bedops, and awk commands (The full shell script can be found in Appendix A). I spiked several subsets of HBB and RAVEN into the 44 ENCODE background datasets.

2.2.3.4 SBF2 4p linkage region

Chapter 5 describes the analysis of sequencing data from a large Scottish family (SBF2) with multiple cases of bipolar disorder. SM extracted variants located on the SBF2 chr4p15-16 disease-linked haplotype (see Chapter 5 for more details). The annotations for the roughly 5,000 SNPs were collected and formatted into a spreadsheet as described in section 2.3.1. These data were used as a background spiking dataset to test the prioritisation model.

2.2.4 Model implementation

2.2.4.1 Perl

The models were initially implemented in procedural Perl language (version 5.10) and designed to run as UNIX command-line applications. All bioinformatics work was run using the server "Ironhide", which has an Intel quad core (2 threads per core) i7 Processor running at 2.67 GHz per core, 12Gb of RAM and 5Tb of RAID storage, with a 64 bit Fedora Linux operating system.

2.2.4.2 R

I re-implemented my prototype Perl codes in R as a series of functions. Over the course of this analysis, the version of R was updated from 2.14 to 2.15.

2.2.5 Performance measures: ROC Curves and AUCs

The performance measures I used to assess my method were Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve. Originally the ROC curves and AUCs were produced using SPSS (Figures 2.3 – 2.6). However, later ROC curves and AUCs were produced using the R package ROCR using the commands:

```
Set discriminator score (1/0 for regulatory/ background variants):
ROCR.simple <- ...
ROCR commands to generate ROC Curves:
library(ROCR)
data(ROCR.simple)
pred <- prediction( ROCR.simple$predictions, ROCR.simple$labels )
perf <- performance( pred, "tpr", "fpr" )
plot( perf )
```

AUCs were calculated using the following R function:

```
TP_function <- function (x){
        numRows <- length(x$Pos)
# build a dataframe of the correct dimensions:
TP_HBB <- as.data.frame(matrix(data=0,nrow=numRows, ncol = 3), stringsAsFactors = FALSE)
colum_headers <- c("Pos", "GrandTotal", "Score")
colnames(TP_HBB) <- colum_headers
TP_HBB$Pos <- x$Pos
TP_HBB$GrandTotal <- x$Grand_Total
TP_HBB$Score[TP_HBB$Pos %in% positives] = 1
return(TP_HBB)
}
setwd("~")
TrPos_File <- "6.5.14.HBB_coding.final.bed"
TrPos<-read.table(TrPos_File,header=T, sep = "\t", stringsAsFactors = FALSE, na = ".")
setwd("~/R")
positives <- TrPos$Pos
TEST_ROC <- TP_function(OUT_TEST)
TEST_pred <- prediction(TEST_ROC$GrandTotal, TEST_ROC$Score)
TEST_perf <- performance(TEST_pred,"tpr","fpr")
TEST_AUC <- performance( TEST_pred, 'auc')
TEST_AUC_details <- slot(TEST_AUC,"y.values")
TEST_AUC_value <- TEST_AUC_details[[1]][1]
```

2.3 Results

2.3.1 Summary

The aim of this chapter was to design a method, combining multiple types of functional annotation data into a single measure, to prioritise candidate causal SNPs for further investigation. Central to the development of this method was the selection of appropriate functional annotations; the choice of model structure; and lastly, development of a stringent model assessment protocol to gauge model performance.

2.3.2 Feature Selection

The functional annotations I chose to assess for inclusion in the SNP prioritisation pipeline were: minor allele frequency (MAF); SNP position relative to gene features; conservation; DNase HSs; repetitive elements; mapability; and chromatin states. For background on each of these features see section 2.1.3.2. These annotations were chosen based on: i) data availability and ease of access; ii) evidence from the literature showing that these annotations overlap functional elements; iii) data quality (genome coverage and accuracy); and iv) redundancy, as there are many annotations that perfectly correlate with each other.

Several conservation annotations were available from the UCSC genome browser. Before deciding how to combine the various annotations into an integrated pipeline, I first compared the different conservation annotations to identify the most informative data to include in the model.

2.3.2.1 Correlation analysis of conservation tools

I compared the utility of four conservation tools available from the UCSC table browser: GERP, PhastCons, PhyloP and 7xRegPotential ((Cooper et al., 2005); (Siepel et al., 2005); (Kolbe et al., 2004)). As PhyloP and GERP have been shown to have similar performances, incorporation of both tools in the method would be redundant. As GERP was shown to perform marginally better (Pollard et al., 2010), I chose to only include GERP in the next step of my analysis. I then calculated pair-wise correlation coefficients for the remaining conservation tools in an all-against-all approach. The aim of this analysis was to identify the combination of tools with the highest individual predictive

value, whilst controlling for redundancy. Using R, I calculated Spearman's rank correlation coefficient, rho, and the associated p-values for each of the SNPs within the SBF2 linkage region (see section 2.2.5.1). Table 2.4 shows the rho scores and p-values for each of these combinations.

The three variations of PhastCons (Primate, Placental and Vertebrate) had medium to high correlation with each other, ranging from ~0.68 – 0.92, with very low p-values. As these tools are so similar, differing only in the phylogenetic scope of their training data, it would be redundant to use more than one in the pipeline. PhastCons placental has the best phylogenetic scope of the three PhastCons methods, as it has good power (more so than PhastCons primate) but also captures non-coding conservation which is often poorly conserved when going out as far as the "vertebrate" scope.

Pairwise_combination	p_value	rho
GERP.GERP	0.00000	1
REG.POTENTIAL.7X.REG.POTENTIAL.7X	0.00000	1
X46.WAY.Placental.X46.WAY.Placental	0.00000	1
X46.WAY.Primate.X46.WAY.Primate	0.00000	1
X46.WAY.Vertebrate.X46.WAY.Vertebrate	0.00000	1
X46.WAY.Placental.X46.WAY.Vertebrate	0.00000	0.918956753
X46.WAY.Vertebrate.X46.WAY.Placental	0.00000	0.918956753
X46.WAY.Placental.X46.WAY.Primate	0.00000	0.743818484
X46.WAY.Primate.X46.WAY.Placental	0.00000	0.743818484
X46.WAY.Primate.X46.WAY.Vertebrate	0.00000	0.683952597
X46.WAY.Vertebrate.X46.WAY.Primate	0.00000	0.683952597
GERP.X46.WAY.Placental	0.00000	0.187720539
X46.WAY.Placental.GERP	0.00000	0.187720539
REG.POTENTIAL.7X.X46.WAY.Placental	0.46565	0.127444865
X46.WAY.Placental.REG.POTENTIAL.7X	0.46565	0.127444865
GERP.X46.WAY.Vertebrate	0.00345	0.116904158
X46.WAY.Vertebrate.GERP	0.00345	0.116904158
GERP.X46.WAY.Primate	0.00190	0.09777688
X46.WAY.Primate.GERP	0.00190	0.09777688
REG.POTENTIAL.7X.X46.WAY.Primate	0.63068	0.055287884
X46.WAY.Primate.REG.POTENTIAL.7X	0.63068	0.055287884
REG.POTENTIAL.7X.X46.WAY.Vertebrate	0.92771	-0.015014748
X46.WAY.Vertebrate.REG.POTENTIAL.7X	0.92771	-0.015014748
GERP.REG.POTENTIAL.7X	0.82685	-0.019842987
REG.POTENTIAL.7X.GERP	0.82685	-0.019842987

Table 2.4. Correlation coefficients (rho) and associated p-values for the pair wise comparison of all the conservation tools I included in my analysis, ranked highest to lowest rho score.

GERP and PhastCons placental produced the next highest score (0.19), the remaining pair-wise correlations performing with rho's ranging from 0.13 to -0.02 (Table 2.). The correlations of 7xRegPotential with all of the other tools (PhastCons Placental (D), PhastCons Primate (G), PhastCons Vertebrate (H) and GERP (I)) produced very high p-values, making these combinations untrustworthy. This appears to be due to the low number of SNPs with a 7xRegPotential score and the even lower number of SNPs with both a 7xRegPotential score and another conservation score.

Of these four tools, the two most commonly used methods are PhastCons and GERP (Pollard et al., 2010). As they do not correlate perfectly, I chose to include both GERP and PhastCons in my method. As PhastCons placental had the best correlation with GERP (without being redundant) and had the best phylogenetic scope of the three PhastCons models, I chose to use this version in my pipeline.

2.3.3 Comparison of model frameworks

I considered several models for combining functional annotation data to best prioritise putative functional over non-functional variants. The simplest method would have been to add the raw scores together in a 1:1:1 ratio. However, this method relies on the assumptions that all features are perfectly correlated with functionality, they are all equally important to the discrimination of functional vs. non-functional variants, and they are all independent features. This is not to the case for annotations such as the chromatin states, where studies have shown different combinations of histone acetylation and methylation mark different types of regulatory elements (Ernst et al., 2011). Furthermore, scoring variants on their raw scores across multiple functional categories is problematic, as different annotations are scored on different scales.

The second method I considered was a ranking system: i.e. each SNP would be ranked on their score from each functional annotation and these ranks would be combined into a single measure, re-ranking SNPs on this score. This method avoids the use of arbitrary weightings and would make it easy to add new annotation data in the future. As with any system, there are also disadvantages to this method. By prioritising SNPs based on a

rank of ranks, each individual score is "smoothed" and any interesting features (such as uneven distribution of scores) is lost. For instance, if an annotation contains a sudden shift from high scores to much lower scores, the SNPs on the boundary of this shift will appear more closely related in a ranked system then if the actual score was included. Like the first method, a pure rank-of-ranks would also fail to take into account the relative ability of each annotation to discriminate between functional and non-functional variants.

The third model I considered involved weighting the different features against each other, based on the premise that different annotation sets have different levels of predictive value and should therefore be weighted with respect to each other to produce a final prioritisation score. This is favourable compared to the blind additive method; however caution must be taken as to how to define weightings, as this method could be considered arbitrary and biased. In addition, weighting the "raw" scores, all on different scales, is also problematic.

I chose to implement a fourth, hybrid, model, combining the best aspects of both the feature weighting and ranking models. This model would rank continuous features, score categorical features and weight the annotations against each other and combined them into a final rank-of-ranks (see Figure 2.2).

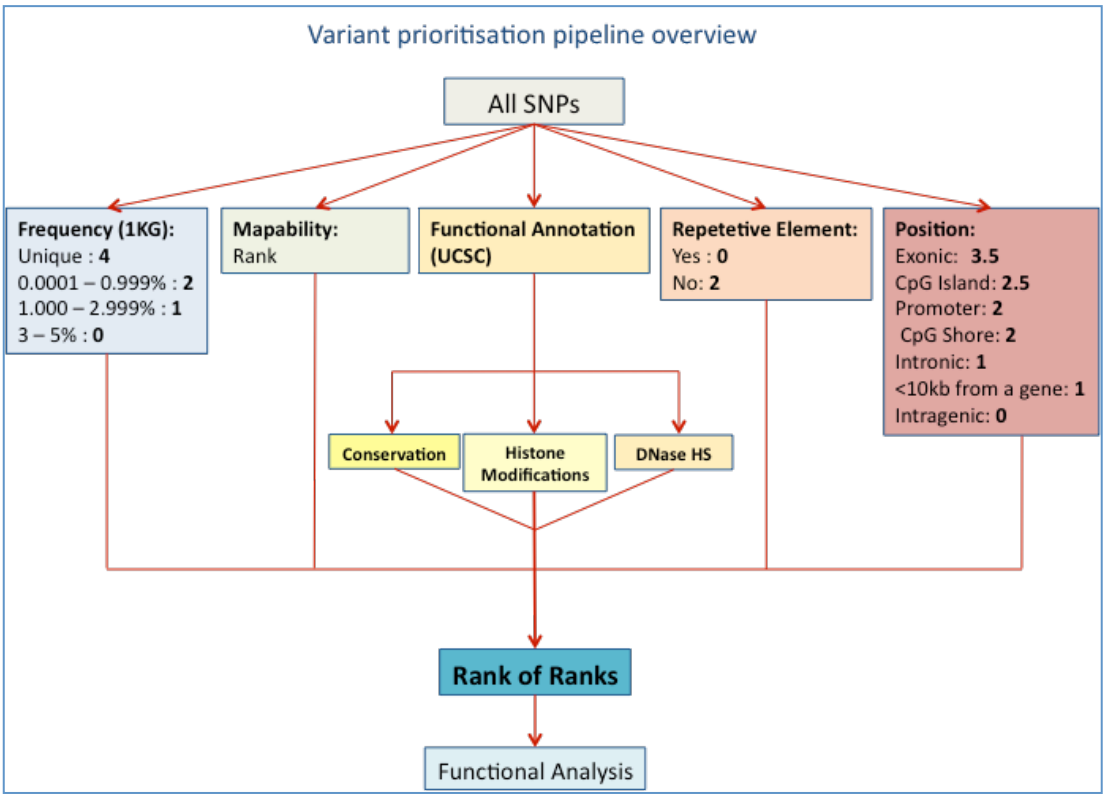


Figure 2.2. Graphical representation of the model described and tested in this chapter. This model combines a ranking system for the functional annotation features conservation, chromatin states and DNase HS, whilst scoring SNPs on a number of other categories including position, frequency, chromosome region and repetitive. All these factors are then combined into a cumulative "rank-of-ranks" to prioritise SNPs from most to least likely to be functional.

2.3.4 Feature scores

In this model framework, SNPs were ranked on their scores for the features DNase HS and the two conservation features, GERP and PhastCons. For the other features, including MAF and Position, SNPs were stratified into subclasses and scored according to which subclass they fell into (a more detailed description of each annotation can be found in sections 2.3.3.3.1-5). The feature ranks and scores were then weighted against each other and combined into a final collective score, which was used to generate a "rank-of-ranks" to prioritise SNPs on the likelihood of functionality. Figure 2.2 provides an overview of the variant prioritisation pipeline, including the scores used for categorical features and the values used to weight the annotation categories. The results were saved to a tab-delimited output file. This process was originally performed by a

Perl script, but was later re-implemented in R. Model performance was assessed using a selection of test datasets and ROC curve analysis.

2.3.4.1 MAF

MAFs were divided into 4 bins (unique, not unique but less than 1%, greater than 1% but less than 3% and greater than 3%), each bin corresponding to a different score. I implemented a 4,2,1,0 scoring system, arbitrarily chosen based on doubling in importance and so doubling in score from bin to bin: anything >3% scores 0, >1% and <3% scores 1, <0% and <1% scores 2 and anything unique (= 0%) scores 4 (see figure 2.3). This method was based on the assumption that unique variants (not present in the 1000 Genomes database) are most likely to be pathogenic, and the more common a SNP is, the less likely it is to be pathogenic.

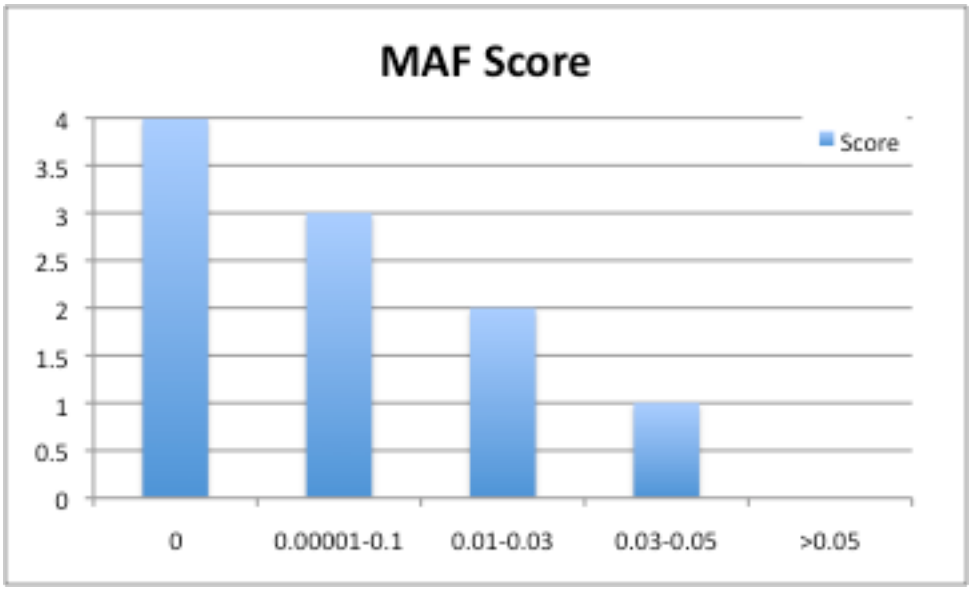


Figure 2.3. Graph illustrating the scores assigned by the model to SNPs with different MAFs.

2.3.4.2 Position

For the position score, SNPs were binned into distinct position classes: exonic, splice site, promoter, 10kb upstream or down stream of a gene, within a CpG island or CpG shore or beyond 10kb from any gene (i.e. intergenic). These classes were assigned different scores based on the relative importance of each positional class and the likelihood of SNPs in that class being causal variants, over SNPs in other classes (see Table 2.5 for more details). The scores were chosen based on knowledge from the literature ((MacIsaac et al., 2010); (Ernst et al., 2011); (Maurano et al., 2012)) and the example set by other existing methods (SNPselector: (Xu et al., 2005); SNPRanker: (Calabria et al., 2010)).

Position category	Score
Exonic	3.5
Splice site	3.5
CpG island	2.5
CpG shore	2
Promoter	2
Intronic	1
10kb upstream	1
10kb downstream	1
Intragenic	0

Table 2.5. Position scores. Each SNP is assessed using data from RefSeq and UCSC to define the position category it belongs to. If a SNP belongs to multiple categories (exonic variant for one transcript but an intronic variant for a second gene transcript), it is assigned to the highest scoring category (ie exonic over intronic).

2.3.4.3 UCSC genome browser annotation data

The conservation and DNase HS site annotations from UCSC were all ranked on the total number of SNPs in the dataset, the ranks normalised to a 0-1 scale, ensuring each analysis is comparable, irrespective of the number of SNPs in the dataset.

Ernst et al. showed that chromatin state classes are correlated across cell lines and that a region that is predicted to be a strong enhancer (or promoter) in one cell line tended to be an enhancer in other cell lines (although in the other cell lines, it might only be a weak enhancer or promoter) (Ernst et al., 2011). The Ernst-defined chromatin states were converted to a binary classifier: all marks of active regulation being converted to "1" and the marks of closed chromatin (no regulation) being scored "0". This was done for all nine Ernst cell lines and the 9 binary scores were added together and used to rank SNPs (SNPs having active scores across all nine cell lines ranking highest and those with none ranking lowest).

2.3.4.4 Repetitive elements

I included two methods in my pipeline as *posterior* WGS quality control settings: mapability and RepeatMasker data.

Mapability: SNPs were scored on mapability, a score of "1" indicating the sequence is unique, a score of less than 1 indicating an increasing numbers of occurrences in the genome, and a score of "0" suggesting it occurs four or more times in the genome.

RepeatMasker: SNPs were separated into two classes based on their RepeatMasker annotation: SNPs overlapping any repetitive element or not overlapping any repetitive element. SNPs not in repetitive elements were scored above those overlapping repetitive elements.

2.3.4.5 Chromosome region score

For studies where linkage data is available (such as from a pedigree analysis) or where significantly associated regions are known, I wanted to allow SNPs within the linkage/associated loci to be prioritised above SNPs outside of these regions, as such *a priori* information can reduce the search space for putative causal variants. Therefore, SNPs within known linkage regions were boosted with an additional score. However, as none of the test datasets provided this type of annotation, none of the studies outlined in later sections of this chapter benefited from this score.

2.3.5 Assessment of model framework

Validation of the prioritisation method required the use of test datasets to assess model performance. A good benchmarking dataset is characterised by being validated, accessible and containing a good number of variants. I have compiled multiple test datasets, each of which fulfils these criteria.

2.3.5.1 HBB dataset

The first benchmarking dataset I used consisted of *HBB* gene mutations associated with Beta thalassemia (variants from the HbVar database: see section 3.2.5.2). This dataset contained 363 disease associated SNPs and a further 141 SNPs control SNPs (benign). As I was particularly interested in prioritising non-coding variants, I filtered out all coding variants from this dataset. This left me with 39 non-coding disease variants and 141 background variants. I tested the prioritisation method on this data and drew ROC curves and calculated the AUC, shown in Figure 2.4. This result (AUC of 0.988) showed that my method was able to correctly prioritise the non-coding disease variants over the background variants with very high specificity and sensitivity.

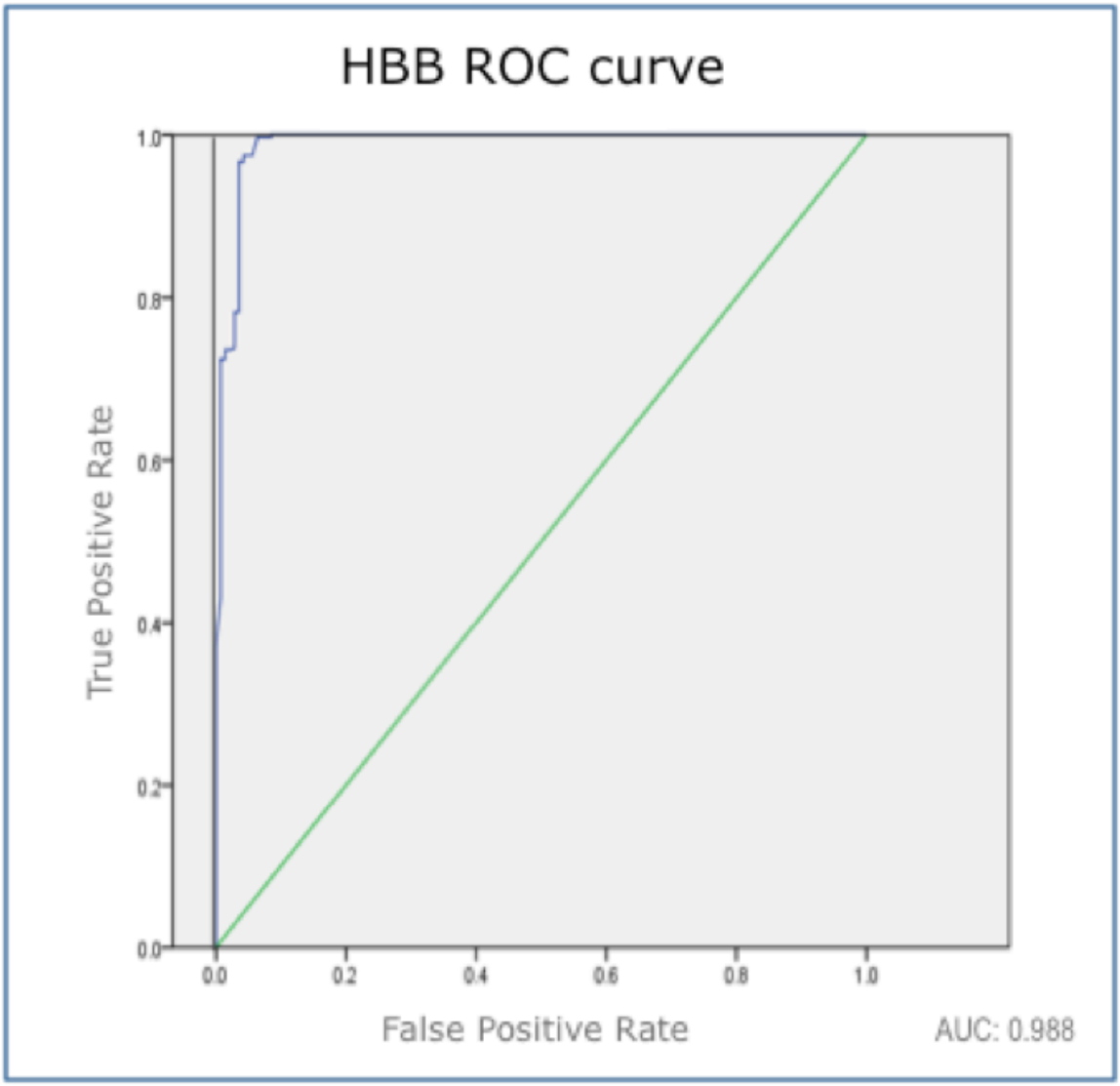


Figure 2.4. ROC curve and AUC for the model run on the HBB non-coding disease variants against the control set of SNPs from the same HBB locus. ROC curve (blue) shows the true positive rate plotted against the false positive rate, the green line representing the result expected by chance. Both the AUC and the ROC curve show that the model is able to prioritise the disease variants over background variants with almost perfect specificity and sensitivity.

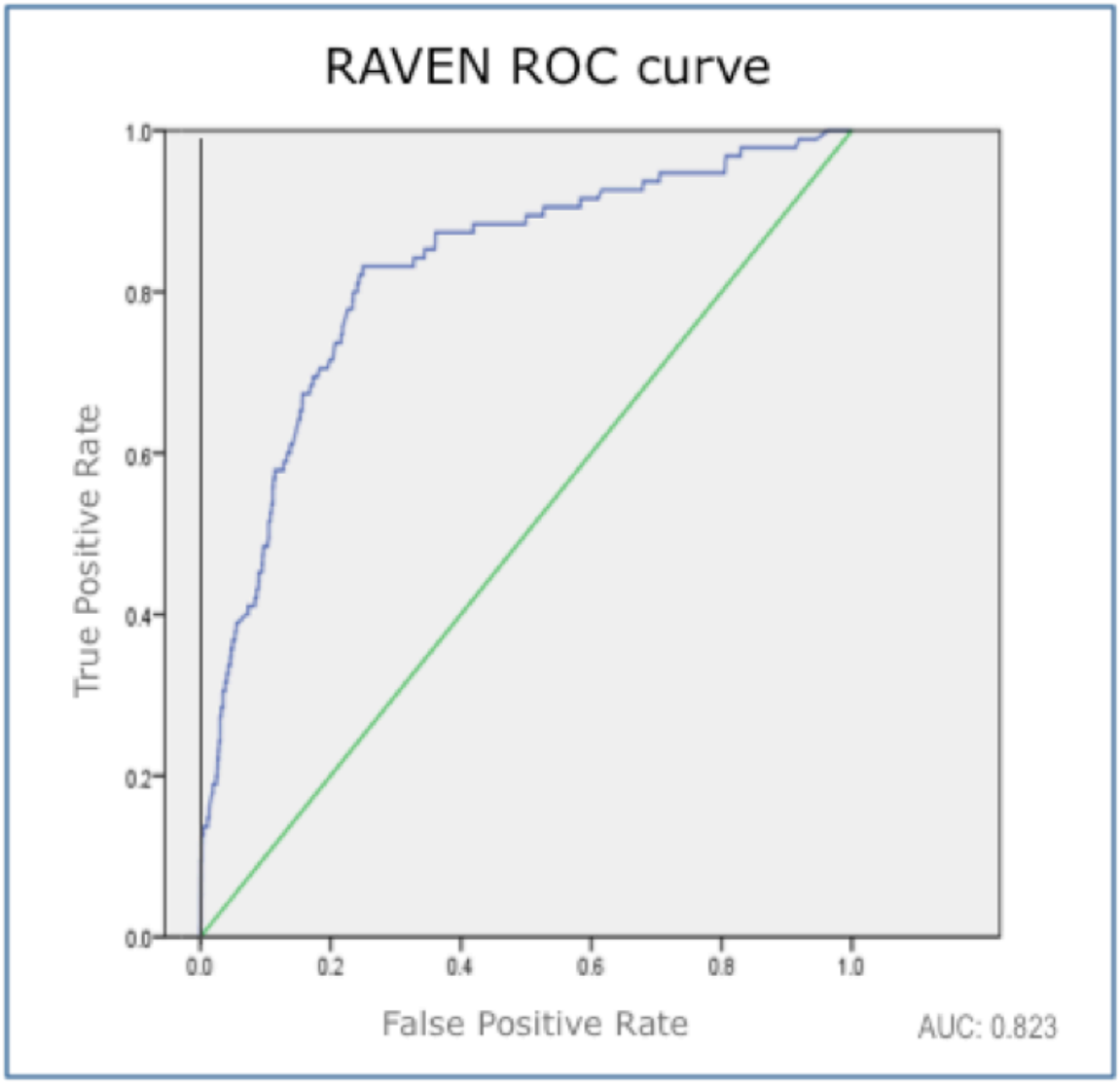


Figure 2.5 ROC curve and AUC for the model run on the RAVEN experimentally validated regulatory variants against the control set of SNPs matched to the true variants to within 10kb of human genes with mouse homologs. ROC curve shows the true positive rate plotted against the false positive rate, the green line representing the result expected by chance. Both the AUC and the ROC curve show that the model is able to prioritise the regulatory variants over background variants with high specificity and sensitivity.

2.3.5.2 The RAVEN dataset

As functional characterisation of non-coding SNPs is particularly challenging, it was important that I identify an additional dataset that contained such SNPs. My second test dataset was constructed by Anderson et al. (2008) to assess the performance of their method RAVEN, and which I named the RAVEN dataset. The dataset contained 95 experimentally validated regulatory SNPs and 3990 control (background) SNPs (Andersen et al., 2008).

My model achieved an AUC of 0.823 on this dataset (Figure 2.4). This lower performance compared to the HBB dataset is to be expected for several reasons. Firstly, the RAVEN dataset contains a much higher proportion of control SNPs to case SNPs. Secondly, as the control SNPs have not been proven to be non-functional, it is possible that some are in fact functional. As I could not control for such false negatives, it could be assumed that the specificity and sensitivity achieved by the scoring methods were the minimum the model will achieve on these data.

2.3.6 Spiking analysis

2.3.6.1 SBF2 background dataset

The first spiking background set chosen consisted of variants on a 20Mb region of chromosome 4p16 locus that has been shown to be linked to BP disorder in a large Scottish family (SBF2) (see Chapter 5). The spiking approach was first used to spike the 39 regulatory SNPs from the HBB dataset into the SBF2 SNP file. I hypothesised that this would provide a greater measure of the pipeline's performance than comparing the known HBB variants against the HBB background variants. My reasons were: i) the SBF2 background dataset consists of a much larger number of SNPs; ii) these SNPs have already been filtered on MAF and consist of only SNPs with a MAF < 5%, decreasing the advantage of the HBB variants (which are all rare) over the background set; iii) this region is much larger than the HBB locus, including over 100 genes producing proteins and non-coding RNAs with a broad range of cellular functions and therefore likely to have diverse forms of regulation (and features); and iv) the pipeline would have to be

able to prioritise the known HBB variants over variants that may also be causal but have yet to be identified, thus providing competition for the highest ranking positions.

As with the RAVEN dataset, this background set contained true positives that would be classed as false positives, which would be predicted to lower the performance of any tested pipeline. However, an advantage of this method was that any background SNPs that performed better then the true HBB SNPs could be considered excellent candidates for my analysis of variants segregating with illness. In addition, as the HBB regulatory SNPs were taken out of context, there would be no question of linkage between the case and control SNPs.

Despite the added level of stringency and complexity, the model was still able to prioritise the HBB variants over the majority of the SBF2 variants, with an AUC of 0.993 (Figure 2.6). I also spiked the RAVEN regulatory variants into the SBF2 background set. This proved more of a challenge for the model and the performance dropped, with an AUC of 0.797 (Figure 2.7).

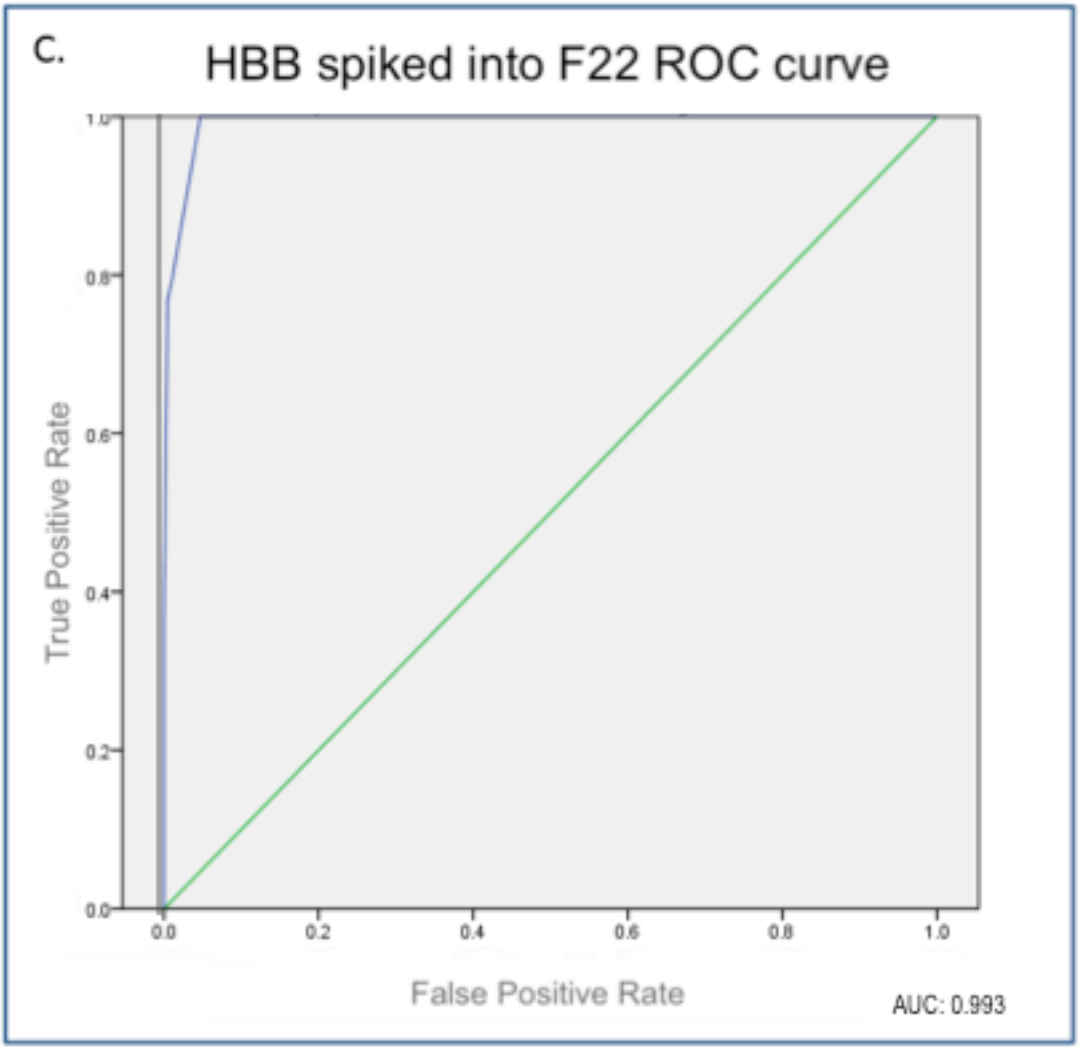


Figure 2.6 ROC curve and AUC for the model run on the HBB non-coding disease variants against the control set of SNPs from the F22 chr4p16 locus. ROC curve shows the true positive rate plotted against the false positive rate, the green line representing the result expected by chance. Both the AUC and the ROC curve show that despite being spiked against a novel unrelated background set, the model is able to prioritise the disease variants over background variants with almost perfect specificity and sensitivity.

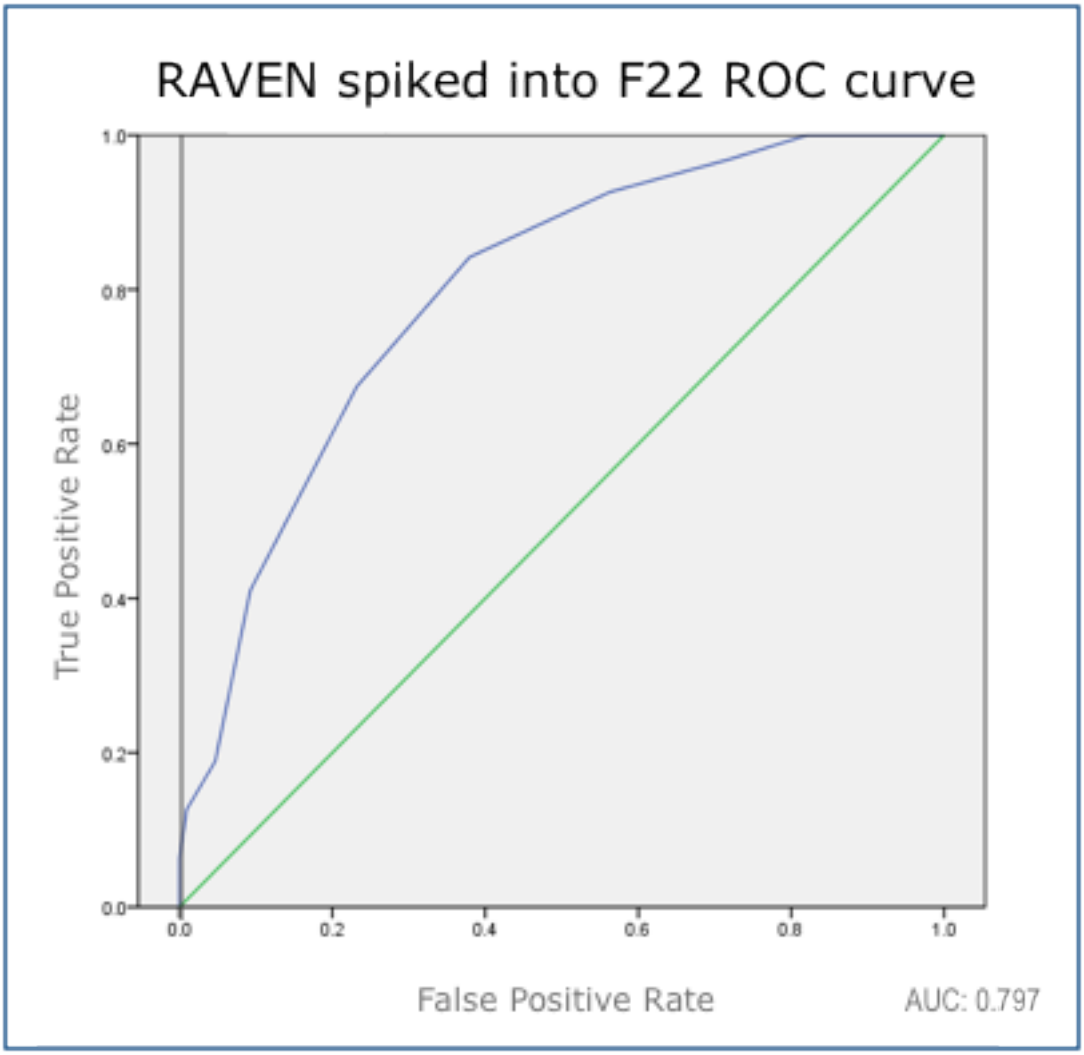


Figure 2.7 ROC curve and AUC for the model run on the RAVEN non-coding disease variants against the control set of SNPs from the F22 chr4p16 locus. ROC curve shows the true positive rate plotted against the false positive rate, the green line representing the result expected by chance. Both the AUC and the ROC curve show that the model is able to prioritise the disease variants over background variants with high specificity and sensitivity and this result is not dissimilar to the model's performance on the RAVEN analysis shown in Figure 2.4.

2.3.6.2 The ENCODE pilot project background dataset

I also spiked various subsets of HBB and RAVEN into the 44 ENCODE background datasets in order to assess the impact that varying the genomic context of the background set would have on the model's performance. Table 2.6 shows the full results for this analysis, which can also be seen in summary form in Table 2.7.

	AUCS					
ENCODE Region	HBB			RAVEN		
Filenames	All True Positives	Noncoding True Positives	All False Positives	Full	Noncoding	Noncoding less 5% AF
ENm001	0.997496514	0.986942901	0.377176118	0.718052648	0.689147686	0.92155849
ENm002	0.994388459	0.972297292	0.343755887	0.663867466	0.632938488	0.886445143
ENm003	0.996860553	0.983279866	0.400857108	0.718545116	0.691043144	0.914752695
ENm004	0.997081984	0.984697986	0.396493477	0.732538078	0.706036964	0.924151203
ENm005	0.995335834	0.979663863	0.391670739	0.715450264	0.688323015	0.91293376
ENm006	0.991758883	0.960621575	0.370638474	0.670461913	0.644583323	0.873754077
ENm007	0.995545024	0.975418096	0.393482909	0.706541973	0.681823492	0.901372162
ENm008	0.990386441	0.955595862	0.287621384	0.572080443	0.536896887	0.835292669
ENm009	0.99701342	0.983915224	0.401134722	0.728030928	0.702528761	0.92145859
ENm010	0.992703539	0.965357002	0.411301691	0.720612146	0.699355656	0.897521368
ENm011	0.994910953	0.973743327	0.347833604	0.679850286	0.650797826	0.892951694
ENm012	0.999126712	0.995380894	0.440228281	0.793626368	0.771265209	0.960046944
ENm013	0.998507564	0.991057962	0.395607046	0.747951681	0.721417758	0.933976716
ENm014	0.999117053	0.995394573	0.461933657	0.814494124	0.794623073	0.968466828
ENm111	0.998875809	0.992923271	0.461204065	0.805506915	0.785310797	0.955550305
ENm112	1	1	0.594509319	0.923096695	0.914125855	0.999095607
ENm113	1	1	0.591538646	0.920053108	0.910681188	0.998677878
ENm114	0.999682081	0.997322086	0.500099994	0.851614676	0.835880817	0.982898172
ENm121	0.997879007	0.989554602	0.51113454	0.823880793	0.80730607	0.953602258
ENm122	0.997906029	0.988204651	0.428509253	0.770471507	0.747808133	0.939924583
ENm123	0.998616199	0.993008195	0.434124808	0.782216495	0.759824136	0.952004582
ENm131	0.997989992	0.987822234	0.421119132	0.76051218	0.737437464	0.93781707
ENm132	0.994714317	0.971010035	0.245218043	0.568378966	0.526181828	0.851053853
ENm133	0.995916781	0.980682712	0.398803463	0.735521472	0.710685673	0.919904567
ENm211	0.999913506	0.999262032	0.509421008	0.855771296	0.839486564	0.98403977
ENm212	0.998988107	0.992194821	0.396847977	0.737541528	0.709155755	0.929240495
ENm213	0.998270145	0.990928567	0.431089477	0.771272502	0.747270005	0.941325791
ENm221	0.997806603	0.990264714	0.470785765	0.805739222	0.78624495	0.953703704
ENm222	0.997115014	0.990589744	0.425801418	0.787805263	0.765435294	0.947305556
ENm223	0.9972408	0.984943757	0.401754369	0.747100722	0.722383395	0.928663336
ENm231	0.988376775	0.955802104	0.332467267	0.630958165	0.600788084	0.854487179
ENm232	0.991953327	0.969132702	0.398307664	0.714886096	0.690301251	0.899940402
ENm233	0.990764413	0.954026522	0.301914497	0.599840128	0.567122772	0.84228173
ENm311	0.999931725	0.999527082	0.518241462	0.87038374	0.855324631	0.98639129
ENm312	0.999854559	0.998725912	0.405908923	0.768375789	0.742283803	0.955207836
ENm313	1	1	0.589460769	0.917158325	0.907460377	0.99821417
ENm321	0.998010319	0.986125241	0.335974046	0.654538301	0.619098197	0.897322855
ENm322	0.999158044	0.994211865	0.472747324	0.812225201	0.792373443	0.958355675
ENm323	0.998254371	0.98953782	0.424894393	0.771158224	0.74778224	0.942084998
ENm324	0.998572042	0.990884498	0.445267259	0.775307822	0.752050729	0.941227064
ENm331	0.989793452	0.95878886	0.380014847	0.692047204	0.668674735	0.883857819
ENm332	0.991755318	0.960274208	0.34127938	0.648791918	0.620507983	0.867679911
ENm333	0.988524758	0.950025012	0.308642306	0.597615183	0.565486988	0.836112859
ENm334	0.995172278	0.976190476	0.385272333	0.70674099	0.680173863	0.905336617
Average AUC:	0.996392471	0.98262114	0.415502019	0.74519577	0.721032461	0.924727052

Table 2.6. Table of AUCs and Average AUCs for each spiking analysis. Each ENCODE pilot region dataset consists of SNPs within each region with a MAF <5% based on the 1000 Genomes Eur subpopulation database. Bottom row (highlighted in green) shows the average AUCs for each spiking analysis. Across all analyses, decreasing AUCs correlate with increasing gene density and number of genes corrected for region size. AUC is not affected by the size of the pilot regions, the NEC or the number of SNPs.

I first used the non-coding, true positive HBB SNPs used for the initial analysis described in Figures 2.4 and 2.6 (see section 2.3.5.1 and 2.3.5.2). Table 2.6 shows a very narrow range of AUCs for this spiking set across the 44 regions, ranging from 0.950 – 1.000, with an average AUC of 0.983.

I next ran the pipeline on the full true positive HBB dataset (including coding variants) and tested the method's ability to rank coding as well as non-coding true positives above background variants. As expected, the average AUC was higher than for the non-coding true variants alone, increasing to 0.996.

The third spiking analysis I ran was a spiking set consisting of the 141 control, non-disease causing, HBB SNPs (All control HBB SNPs). The average AUC produced was 0.416, much lower than for the non-coding true positive HBB dataset and the "ALL' (coding and non-coding variant dataset) spiked into the ENCODE background SNP sets. This result shows that disease SNPs are more likely to rank above control SNPs. An average AUC of less than 0.5 also implies that the control variants have ranked lower than expected by chance. This result may be explained by the fact that although we are treating while the ENCODE datasets as "control" sets, they will likely contain true functional variants which could be prioritised above the non-functional (non-disease causing) HBB SNPs. This is another indication of the method's ability to distinguish functional from non-functional variants

I then ran the Full RAVEN true positive set against the ENCODE background variants. These SNPs were used for the initial analysis described in Figure 2.5. The average AUC for this dataset was 0.745.

One drawback of the RAVEN dataset is the proportion of 'true' SNPs that have a MAF greater than 5%. These will perform worse than the ENCODE background SNPs, which were selected to have a MAF less than 5%. I therefore tested a final RAVEN set

consisting of only true SNPs with a MAF of less than 5%. This dramatically improved the average AUC to 0.925, which is more comparable with the HBB true non-coding SNP dataset.

25.7.13: Average DAF's and AUCs across each dataset:		
	Average DAF.G1K.EUR	AUCs (spiked into ENCODE)
HBB noncoding TRUE	0.000	0.9826
HBB ALL TRUE	0.000	0.9964
HBB ALL FALSE	0.265	0.4155
RAVEN full	0.277	0.7452
RAVEN noncoding	0.277	0.721
Raven noncoding less 5% freq	0.009	0.9247

Table 2.7. This table summarizes the average allele frequencies (Average DAF.G1K.EUR) and average AUCs for each of the spiking analyses.

2.4 Summary and Discussion

2.4.1 Summary of chapter

To prioritise candidate pathogenic variants I have designed a pipeline, which utilises a combination of functional annotation, frequency, and positional information for each variant, and combines them into a single rank score.

The aims of this chapter were: i) to assess the different model designs that could be used for my prioritisation method, identify the most appropriate one and test how well this model works on a variety of test datasets; ii) to explore the wide range of functional data that is available and select those features which would be most informative in our model; and iii) discuss the availability of test datasets and different methods of constructing new ones.

2.4.2 Comparison of different features:

I compared functional annotation data from a variety of sources to identify those features that would best aid the discrimination of functional from non-functional variants. Features selected for this model included MAF, conservation (GERP and PhastCons scores), position relative to genic elements, DNase HS, chromatin states (Ernst data), repetitive elements, mapability and a weighting for being located within a chromosome region of interest.

2.4.2.1 The advantages of using chromatin states versus raw histone data

As described in the Introduction to this chapter, the acetylation and methylation of different combinations of histone markers have been shown to correlate with different regulatory elements. These data can be used to aid the prioritisation of functional SNPs by providing information on any overlaps with putative regulatory elements. Ernst et al. looked at the specific relationship between these markers and applied this data to predict whether a variant overlaps regulatory elements. This data has been made publically available in the form of chromatin states (Ernst et al., 2011). An advantage of this data is that the complex job of combining individual histone modification data has already been

performed, tested, peer reviewed and shown to be accurate. A disadvantage is that this data currently relates to a very small set of cell lines, whereas the equivalent raw data from the ENCODE project can be obtained for a much wider range of cell lines.

As part of the prioritisation model, the chromatin state data was averaged across the nine cell lines to get an idea of the regulatory effect of each variant under investigation across all available cell lines. However, this method puts greatest emphasis on SNPs that overlap high scoring chromatin states across all nine cell-lines. This quantitative approach might not best represent nature, as a SNP that overlaps a high scoring chromatin state in one cell line still provides valuable information. Similarly, the chromatin states were divided into two bins and ranked accordingly, a method that does not take into account differences between the chromatin states within the two bins. I therefore consider in the next chapter a more justifiable method of scoring SNPs based on their chromatin state data.

2.4.2.2 Cross-species conservation

Conservation has historically been the most commonly used feature for identifying regulatory elements ((Boffelli et al., 2003); (Brugger et al., 2004); (Ghanem et al., 2003); (Gottgens et al., 2002); (Pennacchio and Rubin, 2001)). I used a combination of correlation analysis and experimental evidence from the literature to focus in on two of the four methods tested: GERP and PhastCons (placental). GERP calculates conservation at the nucleotide level, whereas PhastCons calculates a conservation score on a region-by-region basis, using a sliding window approach. By combining two contrasting approaches I hoped to identify conserved regulatory elements with even greater accuracy, the premise being that variants scoring highly in both approaches are more likely to be real than variants predicted to be conserved by only one method (Cooper and Shendure, 2011).

A caveat in this assumption is that we know from our correlation analysis that GERP and PhastCons placental only correlate with a Spearman's coefficient value of 0.18772. That is to say, they are only weakly correlated. This begs the question, how much is the method benefiting from using a combined conservation score based on both of these

approaches? Would it in fact perform better using only one or other of these two? In the next chapter, I will address this issue and look at the performance of each method individually.

2.4.3 Model design

The hybrid model I chose to test in this chapter combines ranking and weighting. Model performance was tested using a variety of datasets and ROC curve and AUC statistics.

2.4.3.1 Conclusions on model performance

My model performed very well on the HBB dataset, with an AUC of 0.998 and to a similar level on the HBB data spiked into the SBF2 locus (AUC of 0.993). One reason why the model may identify these HBB variants with almost perfect sensitivity and specificity, even with the change of genomic context, could be related to the type of variants within this dataset. Thalassemia is a highly penetrant, Mendelian disorder, which could be expected to be caused by variants with a high effect size and a highly deleterious effect on phenotype. These variants could be hypothesised to either affect annotation features to a greater extent than weaker variants (i.e. regulatory variants with no link to disease, or less penetrant, low effect, complex trait variants) and therefore could be more easy to identify by studying changes to annotation features.

In contrast, the RAVEN variants are experimentally verified regulatory variants. As such, these variants may be harder to distinguish from background variants, thus accounting for the lower AUCs obtained in comparison to the HBB variant analyses. In addition, this could explain the decreased ability of the method to prioritise RAVEN variants over background when spiked into the SBF2 dataset, as this dataset is likely to contain functional variants that are competing with the RAVEN variants and have more distinct feature annotations than the RAVEN variants.

2.4.3.2 Caveats of feature and model selection method

Although I was able to show that my model was able to prioritise the regulatory variants from the HBB and RAVEN datasets with high specificity and sensitivity, therefore showing its application to real data, questions can be raised as to the method used to select both the features and the model for the prioritisation method. Features were chosen on a relatively ad hoc basis, without formal comparison of the relationship between different features. Similarly, the scores (instead of ranks) assigned to the features MAF, Position, chromosome region score and repetitive element score, were chosen on how important each sub-class is relative to the others (for instance, promoter variants were considered more important than intergenic variants and scored appropriately). This was not an uninformed process as it was strongly influenced by evidence from the literature and general opinion in the scientific community. Nevertheless, this approach could have benefited from more formal model training. This will be the subject of the next chapter.

2.4.4 Test datasets

The development of algorithms to identify functional regulatory SNPs has been impeded by the lack of data on regulatory SNPs (Torkamani and Schork, 2008). The better the data, the more accurate the assessment; good benchmarking data can be used across multiple different methods, allowing a fair comparison between methods. However, very few sources of verified regulatory variants exist, and those that are available are mostly limited in number and are not available to download in bulk format.

Time and effort are required to construct useful datasets, but the information we can get from these data can be limited and biased. Finding a dataset that is applicable across the genome and across variant classes is, therefore, very difficult.

This lack of a gold standard dataset is a major issue hindering the construction and testing of prioritisation methods. To overcome this, I assembled my own repertoire of datasets, consisting of two true positive datasets (HBB and RAVEN) and four background sets (HBB, RAVEN, SBF2 and ENCODE spiking background sets). Although each has its own caveats, they each have fundamental facets of a hypothetical

gold standard dataset: the HBB dataset contains SNPs from the human haemoglobin beta gene (coding and non-coding), the true positive SNPs being disease causing for beta thalassemia; the RAVEN dataset contains non-coding SNPs, all within 10kb of genes with mouse homologs, the true positive SNPs consisting of experimentally verified regulatory SNPs; the SBF2 background set provides a large number of unrelated, low MAF variants (all with a MAF <5%); and the ENCODE background datasets provide an opportunity to test model performance across a range of genomic contexts. The individual aspects of a gold standard dataset presented by this collection of data, make these datasets, when used in combination, a good substitute for a single benchmarking dataset.

2.4.4.1 Pros and cons of the HBB dataset

Using these data I was able to test how well the model performs on different classes of regulatory variants under different genomic contexts. When the pipeline was run on these data it successfully ranked the known SNPs above the background (control) SNPs, as illustrated by the ROC curves in Figures 2.4, 2.5, 2.6 and 2.7 and the AUCs (ranging from 0.797 to 0.998). This showed our model was able to correctly prioritise known disease or functional variants over background variants with un-known function.

The main caveat of the HBB dataset that it is not an unbiased dataset, as the SNPs all map to a very small region of the genome. Because of this, the inheritance of the disease SNPs and control SNPs may be under some level of linkage and not independent. It should also be noted that the "disease" variants are all associated with a Mendelian disorder, ß thalassemia, which is a very specific disease model and most likely has a very different genetic architecture to complex diseases. As I wanted to design a model that can be used for a variety of genomic projects, this dataset only provides limited information and needs to be supported by additional data.

2.4.4.2 Pros and cons of the RAVEN dataset

The RAVEN dataset is roughly 2.5 times larger than the HBB non-coding dataset, thereby increasing the power of my analysis. In addition to the experimentally validated true positive variants, Andersen et al (2008) also compiled a dataset of ~4,000 background variants matched to the true positive variants to within 10kb upstream of the transcription start site of human genes with available human-mouse orthologs. The advantage of this RAVEN background dataset over the HBB background variants is that the variants come from a range of locations across the genome. By combining know regulatory variants with unrelated background variants we remove the evolutionary context of the known variants and locus effects. This was the source of inspiration to extend the analysis further by developing even larger, unrelated background sets, which the known variants could be spiked into, thus providing more information on how well my method can prioritise known regulatory variants.

The RAVEN dataset provides has attributes that the HBB dataset lacks; however, this dataset suffers from its own limitations. A disadvantage of the RAVEN dataset is that all the control non-coding SNPs have a MAF greater then 5%. Therefore, by ranking SNPs on MAF, I have biased the prioritisation against the common background variants and towards the much less frequent regulatory SNPs.

2.4.4.3 Pros and cons of the SBF2 4p background dataset

As both the HBB and RAVEN datasets on their own are limited, I performed an additional spiking analysis, whereby both true positive sets were spiked into a background dataset consisting of the SBF2 linkage-region dataset. As this is both an unrelated locus and made up of SNPs with a MAF <5%, it overcomes the drawbacks of testing each dataset in its native background by removing the positional (and possible linkage) bias between disease SNPs and control SNPs; focusing on regulatory SNPs; and providing a background SNP dataset consisting entirely of SNPs with a MAF <5%.

A drawback to this dataset is that it is likely to contain true functional SNPs that are categorised as control SNPs. These SNPs could potentially out perform the spiked-in regulatory SNPs, lowering the perceived performance of the methods tested on this dataset. This background spiking set describes a single genomic locus and there is no way of telling if it is predictive or representative of the entire genome or if it contains some structural or regional bias I have not accounted for.

2.4.4.4 Advantages of the ENCODE background dataset

Using the 44 regions that make up the ENCODE pilot project as spiking sets has many advantages, one of which is it allowed me to assess how well any prioritisation model can differentiate true regulatory variants from control variants over a range of genomic contexts. A second advantage is the size of the ENCODE dataset: as of June 2012, over 170,000 SNPs from the 1000 Genomes EUR (release 72) dataset were annotated within the 30Mb across the 44 ENCODE pilot regions, potentially including a variety of SNP classes including coding variants and non-coding variants, some of which could be regulatory variants (promoters, long range enhancers; repressors, insulators, etc).

As 14 of these regions were selected because they contain very well studied genes (with known biological and disease functions/roles), it can be assumed that these regions will contain annotated and characterised disease causing protein-coding and regulatory variants. All the regions would be expected to contain functional (non-pathogenic) regulatory variants in addition to non-functional background variants. The ENCODE pilot region dataset is therefore a stringent background dataset, as the protein-coding variants and regulatory variants present will compete against the spiked-in known regulatory variants, affecting the model's ability to prioritise the known variants above the background variants.

By incorporating different genomic loci with a range of non-exonic conservation scores and gene densities, I was able to challenge the model in two additional ways:

1. If the background variants are highly conserved, this would challenge the model's ability to discriminate variants based on conservation. In this scenario,

would the model be able to prioritise the spiked-in variants over the background variants via the other features?

2. As gene-dense regions contain large numbers of background variants that are either exonic, splice or promoter variants, the power of the position score will be affected. Can the model still correctly prioritise the known functional variants above background variants when there are a large number of background variants competing for the position score?

This is a very novel approach. By using these regions as an additional spiking control set I hoped to get a fuller understanding of the performance of my method under varying genomic backgrounds.

2.4.5 Conclusions from the ENCODE spiking analysis

The spiking of HBB and RAVEN variants into the ENCODE pilot project regions can tell a lot about both the performance and functionality of the model. In particular, sub-setting the HBB and RAVEN variants into different categories allowed me to compare how well the model distinguishes different variant classes and to postulate what features might be affecting its performance.

In particular I would like to focus attention on three particular comparisons: the HBB non-coding true positive variants vs. the RAVEN true positive variants; the HBB non-functional variants set as true positives vs. the RAVEN regulatory dataset; and the Full RAVEN dataset vs. the rare RAVEN dataset.

2.4.5.1 Disease causing vs. regulatory SNPs (HBB TP non-coding vs. Full RAVEN regulatory)

Comparing the performance of the model on the HBB non-coding SNPs spiked into the ENCODE background datasets vs. the RAVEN non-coding regulatory variants, we see very similar results to the earlier analyses (HBB and RAVEN in their own background sets and the HBB and RAVEN spiked into SBF2), with AUCs of 0.983 and 0.745 on the

HBB and RAVEN data respectively. This is further validation that my method is performing consistently across different datasets.

2.4.5.2 Functional vs. non-functional SNPs (HBB All False vs. Full RAVEN set)

I next questioned how well the pipeline would perform if I used background SNPs as my true positive SNPs, which are presumably non functional. By comparing the performance of the model on the HBB background variants (all False) vs. the RAVEN variants (full), treating the HBB background variants as true positives, we can assess the models ability to discriminate true positives from false positives. The HBB false positive dataset was compared against the RAVEN variant set as they contain similar numbers of variants (141 vs. 95), with a comparable average MAF of roughly 0.27. This removes any bias caused by number of variants or MAF. Therefore, the difference in AUC (HBB FP: 0.4155 vs. RAVEN: 0.745) can be attributed to the models ability to discriminate between functional non-coding variants and non-functional non-coding variants. It is interesting to note the model actually prioritised the HBB false positive variants worse than you would expect by chance. This could be due to the fact that the ENCODE background datasets all contain real, experimentally verified disease variants (particularly around medically important genes) as well as potentially many regulatory variants we know nothing about, thus a large number of false negatives which the model is correctly prioritising as true positives above the false positive HBB variants.

2.4.5.3 Rare vs. common (Full RAVEN vs. less than 5% RAVEN sets)

As discussed in 2.2.2.1, the AUCs from Table 2.4 show that the model is prioritising the HBB and RAVEN known variants consistently well across different background datasets. However, the question of why the model is prioritising the HBB variants better than the RAVEN variants remains unanswered. My final question on these data was, therefore, how much impact is MAF having on the ability to prioritise functional variants over non-functional variants? I have already shown that when MAF is removed (by correcting the average MAF across variants) the method can still correctly distinguish to a very high degree the functional from non-functional variants. However, how much

would the performance of the method change on the RAVEN data if only rare variants were considered?

To address this, I generated a smaller RAVEN dataset consisting of only those variants that have a MAF of <5%. The average MAF of this dataset was 0.009, much more comparable to the HBB true positive non-coding dataset. Interestingly, when the model was used to prioritise these variants against the ENCODE background datasets the AUC (0.925) was more similar to the HBB non-coding set (0.983) than the full RAVEN dataset (0.745), indicating that the HBB variants are most likely being prioritised better than the RAVEN variants because they are rare, rather than because of some underlying difference in their architecture.

2.4.5.4 Implications

We can therefore conclude from these analyses that true functional variants are prioritised better by my model than non-functional control variants; control variants are prioritised worse than expected by chance; and this effect is not a by-product of weighting of allele frequencies (RAVEN and HBB control av. AF ~0.27). Spiking various subsets of the HBB and RAVEN SNPs into the ENCODE pilot regions has also highlighted the ability of my method to identify true functional variants over background variants.

2.4.6 The difference between implementing in Perl vs. R

For the sake of transparency, simplicity, portability, "market penetration" and adaptability, I re-implemented my prototype Perl codes in R as a series of R functions. Working in R has many advantages over Perl for genomic data. Firstly, R is a software environment as well as a programming language, structured around data frames and capable of dealing with large amounts of data, which can be read in and analysed without the need to compile and run any code. This allows complex operations to be performed on large datasets with speed and efficiency. R is also specifically designed for statistical and data-mining analyses, and has many built-in tools to perform with ease operations that would be quite complicated to do in Perl. As the ranking and scoring of SNPs in my model are all very simple to do in R, it made sense to re-implement in this

environment. The resulting code is shorter and simpler in R, providing a level of transparency beyond anything that can be achieved in Perl. Being part of the R environment, my code can also be linked with other functions and packages, such as ROCR, reducing the time wasted moving from one statistical environment to another. Lastly, having my code written as a series of R functions, there is the potential to reformat the code for it to be wrapped into an R package. This would further simplify the code into a single command, while still providing flexibility and allowing easy modification.

2.4.7 Things to improve

2.4.7.1 Datasets

The HBB and RAVEN true positive datasets are good examples of different classes of regulatory variants (one a Mendelian disease set, the other a purely regulatory set), however the background variants they are compared against have their disadvantages: the HBB background set is a single locus – could have some unique feature we know nothing about; the RAVEN background set contains SNPs that all have a MAF > 0.5%, so are immediately biased against for MAF score; the SBF2 background dataset provides a large number of rare (MAF <0.05%) variants, but this is a locus linked to a specific disorder (BD) and so is likely to contain real causal variants for BD that are competing with our known functional variants. In addition, the SBF2 dataset is concentrated around a single chromosome locus, which may suffer from some unique chromosomal structural organisation I have not taken into account. I therefore needed an improved spiking set that would allow me to study the ability of my method to correctly prioritise know regulatory variants over background variants from a variety of different genomic contexts. For this reason, I developed the ENCODE background dataset.

An additional problem faced when analysing and predicting the functionality of non-coding variants is the breadth of classes of regulatory elements. This is often not taken into account; regulatory variants are all tarred with the same brush under the general descriptor “non-coding variants” or at most “regulatory variants”, when in fact regulatory elements can be divided into subclasses such as non-protein-coding RNAs, promoters, enhancers, repressors and insulators. Similarly, variants can be subdivided

by their biological effect: whether they lead to a "strong" effect (highly penetrant, mendelian diseases); account for a small amount of variance for a complex disorder; or have a functional role, affecting gene expression, but not manifesting as a disease phenotype.

The list of classes that we can use to discriminate different regulatory variants is small but by no means exhaustive: improvements in our understanding of the regulatory architecture of the genome will no doubt add additional classes to this list as we identify more types of regulatory elements.
I therefore need a more comprehensive true positive variant dataset, consisting of large numbers of verified regulatory variants. This will be dealt with in the following chapter.

2.4.7.2 Feature weightings and feature scores

The method by which features and weightings were chosen, though based on logical assumptions, was arbitrary, unsystematic and potentially biased. This therefore needs to be addressed. The first step to correct this imbalance is to convert the various scores used in the model to ranks (so all features are ranked). Once this has been done, I should also re-assess the weightings of the ranks used in the model in a systematic, comprehensive manner.

2.4.8 Conclusions

I have developed a prioritisation method that makes use of a range of functional annotation data to rank SNPs on the likelihood of having a functional effect. This method makes use of a model framework that combines aspects of both a scoring model and a ranking model. Specific features, scores and weightings have been chosen based on multiple pieces of information and tested on a variety of datasets to gauge the model's performance. The performance evaluation showed this method was able to prioritise regulatory variants above background variants with high specificity and sensitivity (AUCs ranging from 0.721 to 0.989). In addition, I have established a spiking strategy to evaluate tool performance. This in itself is a novel approach with great potential. However, in this context the spiking strategy could potentially lead to over-fitting and an over estimation of model performance. Improved performance evaluation

methodology and additional testing is required on more diverse datasets to provide a better picture on how the method is performing and how I can improve it.

The method used for the feature selection and assessment was adhoc and arbitrary. A more systematic, reproducible, unbiased evaluation method is required with:

1. Better training and testing data.
2. More statistically rigorous performance evaluation.
3. A more inclusive feature set.

I will address these three points in the next chapter.

Chapter 3: Model testing using cross-validation and development of an R-package

3.1 Introduction

3.1.1 Summary of Chapter 2

At the start of my PhD, there was a need in the field of genomics for a method to prioritise SNPs (particularly non-coding SNPs) on the basis of putative functionality, making use of the wealth of genomic and epigenomic annotation data available from projects such as the ENCODE project (Consortium, 2012), Functional Annotation of the Mammalian Genome (FANTOM5) project (Consortium et al., 2014) and the 1000 Genomes project (Genomes Project et al., 2010). I chose to develop a method to fill this gap. Crucially, this method would support the interpretation of data from a range of genomics projects including variants from genome-wide association studies (GWASs) and next generation sequencing (NGS) projects. The steps I undertook to construct and test an initial model are described in Chapter 2.

Evaluation of this preliminary model on test data showed that it was able to prioritise positive variants (known disease or regulatory variants) over background variants with high specificity and sensitivity. However, the model building approach I used was unsystematic and the model relied on adhoc thresholds for some of the annotation measures. In addition, the model assessment made use of limited datasets and therefore ran the risk of being over fitted to the test data. It was important, therefore, to develop a more formalised approach, combining: i) systematic model selection; ii) a more rigorous statistical framework for stringent evaluation of model performance and prevention of over fitting; and iii) larger datasets, to increase power. In this chapter I describe the steps taken to address these issues.

3.1.2 Systematic model training and validation

Prediction methods are used across diverse fields ranging from the biosciences to insurance, marketing, meteorology and beyond. More pertinently, in the fields of epidemiology and clinical diagnostics, complex biological data are used to predict

phenotypes and diagnosis. For example, biomarkers are used to predict disease status ((Liu and Albert, 2014); (Rowe et al., 2013)) and microarray data are used to forecast clinical cancer outcomes (Dupuy and Simon, 2007). Similarly, predictive methods are being used across a range of genomic data-mining projects attempting to link genes and proteins to pathogenicity. Although the data used in each of these scenarios differs, the methodologies and data requirements are similar. Two factors are critical to successfully build and evaluate any predictive model. The first is high quality benchmarking data; the second is appropriate training and evaluation methodology (Vihinen, 2012). Both of these factors will be discussed in the following sections.

3.1.3 Defining benchmarking datasets

Development and assessment of predictive methods is often hindered by the use of small, often private, datasets and limited performance evaluation measures (Vihinen, 2012). Such model evaluation is neither comprehensive nor generalisable. A model's performance assessed under these conditions cannot be fairly compared against predictive methods trained on other data. A better approach is to use established benchmarking data in a systematic, impartial analysis. This ensures that the evaluation of performance is consistent across all methods.

Vihinen (2012) suggests that a dataset should meet a minimum set of criteria before it can be considered a definitive benchmarking dataset. It should be: relevant; representative; non-redundant; scalable; simple; reusable; consist of experimentally determined classes; and should contain equal numbers of positive and negative variants. In addition, a benchmarking dataset should be large enough to provide sufficient statistical power (Vihinen, 2012). This demanding set of conditions is difficult to meet in any single dataset. In particular, finding a dataset that contains large numbers of functionally validated positive and negative variants is challenging, verging on impossible. In the Methods and Results sections of this chapter (3.2 and 3.3 respectively), I will describe how I constructed a benchmarking dataset by combining two independent datasets: experimentally verified positive variants from the Human Gene Mutation Database (HGMD) and the ENCODE pilot project background spiking variant dataset introduced in Chapter 2.

3.1.3.1 HGMD variants

The HGMD database is a large scale, comprehensive archive of germline mutations that are implicated or are associated with human disease (Stenson et al., 2003). It contains over 141,000 mutations, including SNPs, indels and rearrangements, which are available via two databases: a public version freely accessible for all registered users from academic institutions and non-profit organisations; and a subscription version, HGMD Professional, available through the purchase of a license. A disadvantage of the public version is it is several years out-of-date in comparison to the professional version. It also cannot be batch-downloaded and the variant annotations are more limited than the professional version. For these reasons, I chose to focus on the professional version and obtained the appropriate license.

The HGMD professional database is subdivided into multiple annotation tables including: MUTATION (single base-pair substitutions; missense/nonsense); DELETION (deletions of 20 bp or less); INSERTION (insertions of 20 bp or less); INDEL (indels of 20 bp or less); DELINS (a combined table for data on deletions, insertions and indels); GROSDEL (for large deletions); GROSINS (for large insertions); COMPLEX (for complex rearrangements); AMPLET (for repeat variations); and PROM (variants causing regulatory abnormalities). I focused on the PROM table, which contains non-protein-coding variants with reported phenotypic impacts. These can be further categorised into the following variant subclasses:

DM and DM? (Disease-causing Mutations): These variants have been reported in the literature to be pathogenic mutations. The 'DM?' subclass are variants where there is doubt regarding the degree of pathogenicity. Diseases represented by the DM group of mutations include Parkinson's disease, glaucoma, cystic fibrosis, aplastic anaemia, Hirschsprung's disease, Cowden's disease, beta thalassaemia, Wilson's disease, retinoblastoma, retinitis pigmentosa and haemophilia.

DP (Disease-associated Polymorphisms): These variants have been reported to be significantly associated with disease; however, are not supported by experimental evidence of functionality.

DFP (Disease-associated Polymorphisms with additional supporting evidence of Functionality): Like the DP class of variants, these variants have been reported to be significantly associated with disease; the difference being they are supported by experimental evidence to be directly functional. Variants in this class are associated with diseases and disorders such as type 2 diabetes, asthma, LDL-cholesterol levels, hypertension, schizophrenia, coronary heart disease, myocardial infarction, rheumatoid arthritis, increased risk of lung cancer, neonatal respiratory distress syndrome, Crohn's disease, polycystic ovary syndrome, macular degeneration, Alzheimer's disease and Graves's disease.

FP (in vitro/laboratory or in vivo Functional Polymorphisms): These variants have been reported to have a functional consequence, but have yet to been associated with a disease phenotype.

FTV: Polymorphic, or rare nonsense, or frame shift variants that have been predicted to alter the gene product (i.e. to result in the production of a truncated product), but as of yet with no reported disease association.

The organisation of the HGMD Professional database into these classes and subclasses allows specific subsets of SNPs to be easily extracted.

3.1.3.2 ENCODE variants

In addition to the HGMD functional variants, a control dataset was required to assess the model's ability to distinguish positive (functional) variants from negative (non-functional) variants. An ideal control dataset would match the positive set in size, for

accurate model assessment (Vihinen, 2012). To my knowledge, no such large, experimentally verified non-functional dataset existed. I therefore chose to use the ENCODE pilot project spiking dataset I developed in Chapter 2 for the evaluation of the preliminary model's performance. This dataset contains variants from the 1000 Genomes European population, restricted to within the boundaries of the 44 ENCODE pilot project regions.

3.1.4 Training methodology

The quality of a predictor depends largely on how the model training has been performed. The most common mistake, as observed by Smialowski et al. (2010), made during model building is the incorrect partitioning of data into training and test datasets. It is important to ensure the training and test data are kept separate, as leakage between these datasets can lead to over fitting and an over optimistic estimate of model performance. Furthermore, construction and evaluation of predictive models require the use of well-established validation methods such as cross-validation, which assesses both a model's performance and its ability to generalise to independent data (Smialowski et al., 2010).

3.1.4.1 Cross-validation

Cross-validation is a statistical method used to assess the performance of a model and to predict how well it will perform on novel data. The cross-validation protocol involves splitting data into multiple training and validation sets and first training the model on the training dataset and then testing it on the validation dataset. Different forms of cross-validation are characterised by different methods of data splitting; for example, leave-one-out cross-validation (LOOCV), repeated random sub-sampling, and *k*-fold cross-validation ((Arlot, 2010); (Hastie, 2009)). In *k*-fold cross-validation the dataset is split into *k* equal sub-samples. For each round of cross-validation, 1 to *k*, one of the *k* sub-samples is used as the validation data, while the other *k*-1 sub-samples are combined into a training dataset. This is repeated until all *k* sub-samples have been used as the validation dataset (see Figure 3.1). *k* can be represented by any positive integer; however, it is most commonly set to ten, as ten-fold cross-validation is a generally

accepted compromise between computational complexity (running time and required CPU's) and statistical power.

Cross-validation can be used for model selection (estimating the performance of different models in order to choose the best one) and model assessment (having chosen a final model, estimating its generalisation error on new data). To perform both model selection and model assessment, the data is best partitioned into three parts (tripartite division): training, validation and test datasets. However, when an analysis is 'data poor', only a limited number of samples being available for training, validation and testing, the power of the cross-validation can be maximised by drawing the training and validation data from a single, larger dataset that is partitioned into folds (for instance, ten folds for ten-fold cross-validation). During each round of ten-fold cross validation, one fold is held out as the validation dataset and the other nine combined into the training dataset (Figure 3.1). Performance error is calculated as the difference in performance on the training and validation datasets across each fold. The best model chosen from the cross-validation is then tested on the hold out test dataset. The generalisation error is calculated as the difference in model performance on the test dataset versus the model training and validation performance. Crucially, the data used for training and validation cannot be used for the final model testing as this would lead to an over estimation of model performance. Similarly, once a model has been run on the test data, it cannot be tweaked or modified (Hastie, 2009) or rerun on the test data.

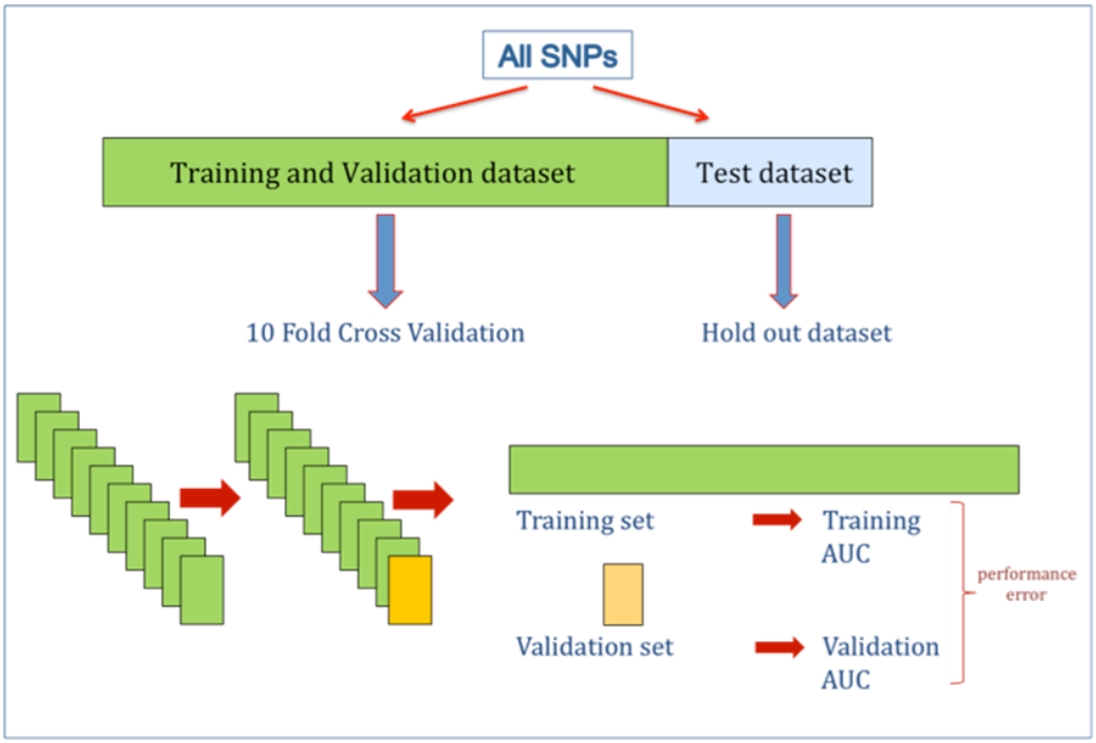


Figure 3.1 Cartoon outlining 10-fold cross-validation using a tripartite data split. The full dataset is divided into two sections: training and validation dataset and a hold out test dataset. The training and validation dataset is partitioned into ten folds. During each round of cross-validation, one fold is set as the validation set and the other nine folds are joined together into the training fold. Each weighting model is run on the training and validation sets in each of the ten folds and the training and validation AUCs for each fold are calculated and recorded. Performance error is calculated as the difference between the training and validation AUCs across all ten folds. The best model from cross-validation (the highest AUC with lowest performance error) is then selected and run on the hold out test dataset. The difference between the test AUC and the average training/validation AUC for that weighting model is used to calculate the generalisation error.

3.1.5 Development of an R package

R, which was created by Ross Ihaka and Robert Gentleman (Ihaka, 1996), is both a free programming language and an environment for statistical computing and graphics (http://www.r-project.org/about.html). The advantages of this system are extensive. R is a flexible, modular language, which allows for effective and simple data handling. As the R language is simple and intuitive, it is easy to learn and is therefore used extensively for the analysis of genomic and epigenomic data. R is highly extendable and,

through the efforts of the Comprehensive R Archive Network (CRAN) and other, newer, distributors and repositories (such as Bioconductor), it is constantly evolving through the inclusion of new packages. Importantly, new R packages have to pass a peer-review process before being accepted by CRAN and Bioconductor, providing confidence in their design and functionality. The modularity of R means that R packages and functions can be used individually or combined into larger R codes. In addition, R allows data to be stored as a range of objects, each of which can be manipulated in different ways. Lastly, R compiles and runs on a range of platforms and systems including UNIX, Linux, Mac OS and Windows, making it universally accessible.

I decided to restructure my list of R functions into an R package. An R package has the following advantages over a series of R functions: it is simpler and faster to run, as the number of commands required to achieve the same output is reduced; and it can be submitted to an R repository, making it easier to distribute and more widely accessible.

Once the decision was made to restructure my R code and functions into an R package, it was necessary to select a package name. I chose to call the R package "SNP Ranking by Function R package" (SuRFR). For the rest of this chapter and this thesis I will refer to the prioritisation R package I have developed as SuRFR

3.1.6 Summary of chapter aims

The aim of this chapter was to improve the SNP prioritisation method described in Chapter 2. This was achieved by i) redesigning the model framework to make it more reproducible; ii) updating the annotation data; iii) expanding the test datasets to increase the power of the analysis; iv) formalising the model testing to prevent over-fitting; and v) restructuring the R code into an R package.

3.2 Materials and Methods

3.2.1 Simplified model framework

The model framework was restructured around the concept of a straight forward rank-of-ranks. For each annotation category, SNPs were ranked from least likely to be functional through to most likely; the ranks across all of the annotation categories were combined using a weighting model to generate an aggregate rank (the rank-of-ranks). Equation 1 describes this model framework:

$$R = rank_i\ (\sum (r_{ij} . w_j))$$

Equation 1.

r_{ij} is the rank of the *i*th variant in the *j*th annotation category, and w_j is the weight for the *j*th annotation category (Ryan et al., 2014).

A central aspect of this method is the weighting term (w_j) a vector of multipliers (one multiplier for each annotation category), which quantifies the importance attributed to each annotation category in the prioritisation of putative functional variants. I developed three different weighting models for SuRFR, for three different categories of regulatory variants: a model designed to be generally applicable to any analysis ("ALL"); a model designed specifically for the prioritisation of rare, highly penetrant disease variants ("DM"); and a model designed for complex trait variants ("DFP").

3.2.2 New annotation data sources

The annotation data classes and sources used in SuRFR are summarised in Table 3.1 and detailed in the following paragraphs:

MAF: I used an updated minor allele frequency (MAF) table for the 1000 Genomes EUR population (release 72). For this annotation, SNPs with the lowest MAF (i.e. rarest SNPs) were ranked highest.

Annotation	Details	Source	Download date
Minor Allele Frequency (MAF)	1000 Genomes phase-1 EUR population	ftp://ftp.ensembl.org/pub/release-72/variation/vcf/homo_sapiens/	Jun-14
RS numbers	1000 Genomes phase-1 EUR population	ftp://ftp.ensembl.org/pub/release-72/variation/vcf/homo_sapiens/	Jun-14
Genomic Evolutionary Rate Profiling (GERP) scores	UCSC hg19 release, hg19.GERP.bed	MySQL --user=genome --host=genome-mysql.cse.ucsc.edu -A -D hg19 -P 3306	Jul-13
Chromatin states	UCSC hg19 release, wgEncodeBroadHmm* x nine cell lines	MySQL --user=genome --host=genome-mysql.cse.ucsc.edu -A -D hg19 -P 3306	Apr-13
DNase hypersensitivity sites	UCSC hg19 release, wgEncodeRegDnaseClustered	MySQL --user=genome --host=genome-mysql.cse.ucsc.edu -A -D hg19 -P 3306	Mar-14
DNase footprints	ENCODE DNase footprints	ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/footprints	Mar-14
Transcription Factor Binding Sites	wgEncodeRegTfbsClusteredInputsV3	MySQL --user=genome --host=genome-mysql.cse.ucsc.edu -A -D hg19 -P 3306	Mar-14
Gene names, exons, introns, splice sites	UCSC hg19 release, knownGene	MySQL --user=genome --host=genome-mysql.cse.ucsc.edu -A -D hg19 -P 3306	Apr-13
CpG islands	UCSC hg19 release, cpgIslandExt	MySQL --user=genome --host=genome-mysql.cse.ucsc.edu -A -D hg19 -P 3306	Apr-13
FANTOM5 CAGE data	FANTOM5 robust transcription start sites	http://fantom.gsc.riken.jp/5/datafiles/phase1.0/	Mar-14
Transcribed Enhancers	Transcribed Enhancer Atlas, permissive enhancer set	http://enhancer.binf.ku.dk/Pre-defined_tracks.html	Mar-14

Table 3.1 This table describes the annotations used in the R package, SuRFR, as well as the sources they were obtained from and the dates they were downloaded (Ryan et al., 2014).

DNase HS: SNPs were ranked on normalised peak score, taking the maximum signal strength across any cell line.

DNase footprints: SNPs were ranked on the number of cell-lines where DNase footprints were observed.

Position: The position score used for the model in Chapter 2 was modified to include annotation data for gene names, exons, introns, splice sites, CpG islands, and CpG shores. Data from the FANTOM5 project, characterising novel transcription start sites (TSSs), was used to annotate previously undocumented promoters (defined as being 1000 bp upstream of FANTOM5 TSSs) and regions 10kb upstream of transcripts. The rank orders of this annotation feature were redefined based on evidence from the literature (see section 3.3.3.3); the new rank order of position categories can be seen in Table 3.2.

Transcribed Enhancers: I collated CAGE-defined transcribed enhancers, identified using data from the FANTOM5 project (Andersson et al., 2014), into a new feature annotation dataset. SNPs were ranked by a binary classification, based on whether or not they overlapped a CAGE-defined transcribed enhancer.

Position	Rank
exon, splice site	5
promoter	4
10 kb upstream and downstream of genes	3
CpG islands and CpG shores	2
intron	1
intergenic	0

Table 3.2 Updated rank orders of position categories. This data is based on enrichment data presented by Hindorff et al. (Hindorff et al., 2009), and Schork et al. (Schork et al., 2013).

Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

cell	Tier	Description	Lineage	Tissue	Karyotype	Sex	Documents	Vendor ID	Term ID	*Label*
GM12878	1	B-lymphocyte, lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasion, Epstein-Barr Virus	mesoderm	blood	normal	F	ENCODE	Coriell GM12878	BTO:0002062	GM12878
H1-hESC	1	embryonic stem cells	inner cell mass	embryonic stem cell	normal	M	ENCODE	WiCell Research Institute WA01	CL:0000007	H1-hESC
HepG2	2	hepatocellular carcinoma	endoderm	liver	cancer	M	ENCODE	ATCC HB-8065	BTO:0000599	HepG2
HMEC	3	mammary epithelial cells	ectoderm	breast	normal	U	Bernstein Crawford Stam	Lonza CC-2551	BTO:0002178	HMEC
HSMM	3	skeletal muscle myoblasts	mesoderm	muscle	normal	U	Bernstein Crawford Stam	Refer to the protocol documents for differing sources CC-2580	BTO:0000256 (non-specific)	HSMM
HUVEC	2	umbilical vein endothelial cells	mesoderm	blood vessel	normal	U	ENCODE	Lonza CC-2517	BTO:0001949	HUVEC
K562	1	leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC	mesoderm	blood	cancer	F	ENCODE	ATCC CCL-243	BTO:0000664	K562
NHEK	3	epidermal keratinocytes	ectoderm	skin	normal	U	ENCODE	Lonza CC-2501	BTO:0000375	NHEK
NHLF	3	lung fibroblasts	endoderm	lung	normal	U	Bernstein Stam	Lonza CC-2512	BTO:0000161 (non-specific)	NHLF

Total = 9

Table 3.3 Table from the UCSC Genome Browser showing the nine cell lines used as the source of the experimental data produced by Ernst et al. used to define chromatin states.

Transcription Factor Binding Sites: I identified an updated transcription factor binding site (TFBS) table from the UCSC Genome Browser. SNPs were ranked on the highest peak signal for any of the transcription factors across all of the cell lines.

Conservation: SuRFR's conservation score was based on GERP rejection substitution (RS) scores. SNPs were ranked from highest to lowest RS; prior to ranking, all negative RS scores were converted to zero.

Chromatin States: The chromatin state data across nine cell lines (Table 3.3) presented by Ernst et al. (2011) was reassessed using multivariable regression. The new chromatin state rankings can be seen in Table 3.4.

Chromatin state classes	Rank
Promoter	10
Strong Enhancer	9
Weak Enhancer	8
Repressed	7
Insulator	6
Repetitive/CNV	5
Transcription Transition	4
Transcription Elongation	3
Weak Transcription	2
Heterochromatin	1

Table 3.4 Rankings of each of the 10 chromatin state classes (best rank: 10, worst rank: 1) defined by the regression analysis described in Results section: Multivariable regression. Each chromatin class is colour coded to reflect the individual chromatin states they represent (shown in full in Table 3.6.)

3.2.3 Construction of test datasets

3.2.3.1 HGMD variants

The HGMD Professional data is available via MySQL. I accessed the database via the command line, using the command format:

```
mysql –A –h host –u username –p password
```

The data I required for each variant from the HGMD mySQL database included: *chromosome, position, ID* (HGMD Accession number)*, Ref base, Alt base, disease, gene, tag* and *dbSNP id*.

This information was not available from any single table; therefore I used JOIN to pull out the appropriate data from multiple tables into a single output. I did this in two steps: i) SNPs present in dbSNP; and ii) SNPs not in dbSNP. I also only extracted variants of the subclasses DM, DFP and FP. The command formats for each step are:

1. SNPs in dbSNP:

Select "chrom", "pos", "dbSNP_id", "ref", "alt", "tag", "id", "gene", "disease" *from hgmd_vcf INNER JOIN prom ON (id = acc_num) LEFT JOIN dbSNP ON (id = hgmd_acc) where hgmd_acc IS NULL and (tag = "DM" or tag = "DFP" or tag = "FP");*

2. SNPs not in dbSNP:

Select "chrom", "pos", "dbSNP_id", "ref", "alt", "tag", "id", "gene", "disease" *from hgmd_vcf INNER JOIN prom ON (id = acc_num) INNER JOIN dbSNP ON (id = hgmd_acc) where (tag = "DM" or tag = "DFP" or tag = "FP");*

These commands provided me with the columns I needed for a total of 1,959 SNPs (1,332 in dbSNP, 627 not in dbSNP). After removing 62 duplicates, this dataset contained 1,897 variants.

A subset of these SNPs overlapped with my RAVEN variants (70 variants). I therefore removed these variants from the HGMD dataset before using it. This left 1,827 variants in the positive dataset (644 DM variants, 686 DFP variants and 497 FP variants). This dataset will be referred to as the 'ALL' dataset from this point onward.

The ALL SNPs were partitioned into a training/validation set (1,440 variants) and a hold out test dataset (387 variants). I constructed two additional training/validation/test datasets by further subdividing the HGMD dataset by variant subclass. These two datasets consisted of i) DM variants only (the DM dataset) and ii) DFP variants only (the DFP dataset). The DM and DFP datasets were split into tripartite subsets as for the ALL HGMD dataset; the DM training/validation set containing 512 SNPs; and the DFP training/validation set containing 534 SNPs.

3.2.3.2 ENCODE variants

The 44 ENCODE pilot project regions contain between them 170,892 variants from the 1000 Genomes EUR population (see Chapter 2 for more details). These variants were divided into two sets: a training/validation set, equal in size to the positive SNP set (i.e. 1,440 SNPs for the ALL dataset; 512 for the DM dataset; and 534 for the DFP dataset) and a background hold out test dataset (169,452 SNPs). All of the SNPs present in the training/validation datasets were excluded from the test dataset.

3.2.3.3 HBB dataset

The HBB variants described in Chapter 2 were used as an additional positive spiking set. The HBB non-coding variants were compared to the HGMD dataset and any SNPs present in both datasets were removed from the HBB dataset. This left me with a HBB non-coding dataset of 27 variants.

3.2.3.4 RAVEN dataset

The RAVEN variants from Chapter 2 were also used to assess the ability of the new models to prioritise regulatory variants. All 95 RAVEN variants were used.

3.2.4 Multivariable regression

Multivariable regression was used for three separate tasks:

i) To correlate the 15 chromatin states with the training data to identify the most informative chromatin state ranking to include in the model.

ii) To compare the predictive power of the old versions of the annotation features (used in Chapter 2) versus the new versions (see Methods 3.2.2).

iii) To guide the parameter boundaries for the grid search algorithm used for parameter optimisation.

All three tasks were performed using the R function *glm()*. The following commands were used to prepare the data for the *glm* function:

Loading the test datasets:

```
HGMD_training_1440 <- read.table("HGMD_1440_training.18.12.13.txt", header=T, sep = "\t", stringsAsFactors = FALSE, na = "NA")

Cross_Val_1440_Null <- read.table("cross_val_encode_1440_Training.18.12.13.txt", header=T, sep = "\t", stringsAsFactors = FALSE, na = "NA")
```

Merging the positive and negative variants into one dataset:

```
HGMD_All_T_F <- merge(HGMD_training_1440, Cross_Val_1440_Null, all= TRUE)
```

Setting the binary classifier (functional/non-functional):

```
TrPos_File <- "4.12.13_hgmd_prom.final.bed"

TrPos<-read.table(TrPos_File,header=T, sep = "\t", stringsAsFactors = FALSE, na = ".")

positives <- TrPos$Pos

HGMD_All_T_F$Score <- 0

HGMD_All_T_F$Score[HGMD_All_T_F$Pos %in% positives] = 1
```

3.2.4.1 Regression of the chromatin states on the full training/validation dataset

Multivariable regression was performed on the chromatin states for all nine of the Ernst cell lines: GM12878, H1-hESC, K562, HepG2, HUVEC, HMEC, HSMM, NHEK, and NHLF (Table 3.3).

```
HGMD_ALL_E_Gm12878 <- glm (y~ HGMD_All_T_F$wgEncodeBroadHmmGm12878HMM+0, family = binomial(link = "logit"))

summary(HGMD_ALL_E_Gm12878)

HGMD_ALL_E_Gm12878_summary <- summary(HGMD_ALL_E_Gm12878)$coef

write.table(HGMD_ALL_E_Gm12878_summary, file = "Multivariable_regression_HGMD_ALL_Ernst_Gm12878.12.3.14.txt", append = FALSE, quote = TRUE, sep = "\t",eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE, qmethod = c("escape", "double"),fileEncoding = "")
```

3.2.4.2 Regression of the new and old versions of the annotation features on the full training/validation dataset

Regression was performed on the normalised ranks of each individual feature using a command such as this one used for MAF:

```
HGMD_ALL_MAF <- glm (y~ HGMD_All_T_F$MAF.rank_normalised, family = binomial(link = "logit"))

summary(HGMD_ALL_MAF)
```

3.2.4.3 Regression of feature annotations on the full training/validation dataset to guide parameter boundaries for grid search algorithm

A multivariable regression analysis was performed on the combined feature set to aid the choice of upper and lower weighting limits for parameter optimisation, using the following commands:

```
# Position + MAF + DNAse f + Cons + DNase c + Ernst + enhancers + TFBSs:

y <- HGMD_All_T_F$Score

HGMD_all_new_rank_tfbs <- glm (y~ HGMD_All_T_F$F_Position.rank + HGMD_All_T_F$MAF.rank_normalised + HGMD_All_T_F$DNase.foot.av.rank + HGMD_All_T_F$Conservation.rank + HGMD_All_T_F$E.DNase.av.rank + HGMD_All_T_F$Ernst.Av.new.rank + HGMD_All_T_F$Enhancers.rank + HGMD_All_T_F$TFBSs.rank, family = binomial(link = "logit"))

summary(HGMD_all_new_rank_tfbs)

HGMD_all_new_rank_tfbs_summary <- summary(HGMD_all_new_rank_tfbs)$coef

write.table(HGMD_all_new_rank_tfbs_summary, file = "Multivariable_regression_HGMD_all_new_rank_TFBSs_summary.5.5.14.txt", append = FALSE, quote = TRUE, sep = "\t",eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE, qmethod = c("escape", "double"),fileEncoding = "")
```

3.2.5 Ten-fold cross-validation

The known functional and pathogenic variants from the HGMD database (ALL dataset) were combined with the background ENCODE variants into a single training/validation dataset of 2,880 SNPs and a test dataset consisting of 169,839 SNPs.

The training/validation dataset was further randomly subdivided into ten folds for cross-validation. Pseudo-code for the R code that was implemented for parameter optimisation and ten-fold cross-validation can be seen in Figure 3.2. Parameter optimisation was

performed using a modified grid search algorithm. This method incorporated multivariable regression on the full training/validation dataset to guide the parameter boundaries of the grid search algorithm. This was similarly performed for the DM and DFP datasets.

Parameter weightings were permuted using brute force permutation of all possible positive integer parameter values. In total over the three datasets, almost half a million permutations of weighting models were assessed using ten-fold cross-validation (n = 450,000).

Performance was measured using ROC curves and AUCs using the R package ROCR (Sing et al., 2005). The objective parameter optimised for weighting parameter selection was maximum AUC, with a threshold acceptable performance error of <0.005 (calculated as the difference between the mean training and validation AUCs: ΔAUC). Three models were developed from this analysis, one for each dataset: 'ALL', 'DM' and 'DFP'. For each of these three datasets, the best model was applied to the hold out test dataset (similarly divided by variant class into ALL, DM and DFP test datasets). Generalisation errors were calculated as the difference between the test AUC and the mean training/validation AUC for that weighting model.

```
permute weighting models to be tested
    foreach weighting model do:
        foreach of the ten cross-validation folds do:
            hold-out one fold (validation set)
            merge remaining folds (training set)
            run weighting model on training set
            run weighting model on validation set
            calculate training and validation AUCs
        end
        calculate the average training and validation
         AUCs across all folds
    end
    calculate performance error for each weighting model
end
determine the optimal weighting model
run the final model on the hold-out test dataset
calculate generalisation error
```

Figure 3.2 Pseudo code for parameter optimisation and ten-fold cross-validation

3.2.6 Building the R code into an R package

3.2.6.1 package-skeleton

To build a new R package I performed the following actions in R:

1. I cleared workspace so as to have a clean R session:

```
rm(list = ls())
```

2. I loaded each of the package functions and data objects one by one.

3. To build the package, I ran the command:

```
package.skeleton ("package_name")
```

4. I edited the package files as follows:

   i) I filled in the DESCRIPTION file and manual pages *(~/package/man)*

   ii) I edited the NAMESPACE file to contain look-up information for functions and objects within the package.

   iii) I wrote a user manual explaining how each part of the R package works, containing real working examples (see Appendix B).

5. Lastly, I built, installed and checked the package using the commands:

```
R CMD build package_name

R CMD install package_name.0.99.tar.gz

R CMD check package_name.0.99.tar.gz
```

3.2.6.2 Sweave vignette

I used R studio to write the sweave vignette for my R package (see Appendix C).

3.3 Results

3.3.1 Construction of training, validation and test datasets

I selected functional non-coding variants with experimentally verified phenotypic impacts from the HGMD PROM database of regulatory variants implicated in disease. This data was then divided into three datasets: DM (known disease causing SNPs: 644 SNPs); DFP (disease-associated variants with functional evidence: 686 SNPs); and ALL (all DM, DFP and FP HGMD PROM SNPs: 1,827 SNPs). For each of these three datasets, an equal number of background variants was obtained by randomly sampling the 1000 Genomes EUR variants located within the ENCODE pilot project regions. Each dataset was divided into a training/validation dataset (ALL: 1,440 known functional variants and 1,440 background variants; DM: 512 known and 512 background variants; and DFP: 534 known and 534 background variants) and a hold out test dataset (387, 132, 152 known variants (ALL, DM and DFP respectively); and 169,452 background variants).

3.3.2 Changes to the feature annotations included in SuRFR

The ENCODE project and other genomics projects are not static data sources, but are constantly being improved (due to technological advances and updated protocols) and expanded to contain new and extended data (e.g. additional cell lines). It was, therefore, important to continue checking these resources regularly to keep abreast of new developments and update the annotation data used by SuRFR. Several updates of features used by SuRFR, as well as some additional annotation features, came to my attention during my second year. This chapter describes the evaluation of the impact of these features on model performance. In addition, this chapter outlines the testing and optimisation of the prioritisation model using cross-validation and reports the performance of SuRFR on a variety of independent datasets.

Some of the features described in Chapter 2 were not incorporated in the model in the most objective or systematic manner. In particular, the chromatin states from Ernst et al. (2011) were integrated into the model without taking into account differences in predictive power for the different chromatin states. For example, the promoter, enhancer,

insulator and transcription chromatin state classes and subclasses were all treated equally, whereas the literature suggests that certain chromatin states are more likely to overlap some regulatory elements more often than others (Ernst et al, 2011). In addition, the rank orders of the Position ranking had been chosen using out-dated data. Furthermore, the conservation score had not been tested systematically to measure the contribution of GERP and PhastCons individually. I, therefore, also re-evaluated the impact of these features on model performance.

3.3.2.1 Updated annotation sources

Minor Allele Frequency (MAF): A new release of the 1000 Genomes MAFs, based on data from 2,504 individuals (compared to 1,092 in the last release), became available. This data was used to rank SNPs, on the basis of MAF, from most rare to most common.

DNase HS clusters: This data contains information on DNase HSs assayed across 125 cell lines, a large increase on the previous version, which was based on 74 cell lines. I performed multivariable regression on the full training and validation dataset, comparing these data and the DNase HS data from the original model (Chapter 2). Table 3.5 shows that the updated DNase HS data has a higher ß coefficient than the old data, therefore incorporation of this data would better enable SuRFR to discriminate between functional and background variants.

3.3.2.2 New annotation data sources

DNase Footprints: Genomic DNase I footprinting data demarcate sequence-specific transcription factors binding sites within regulatory regions, at nucleotide resolution. This data, collected as part of the ENCODE project, consists of high confidence DNase I footprints from 41 cell types (45.1 million footprints in total) (Neph et al., 2012). By combining this data, in addition to data on DNase HS clusters, I anticipated improving the ability of SuRFR to better prioritise regulatory variants by identifying those that overlap DNA elements bound by regulatory factors. Regression of this data on the full

training and validation dataset showed that the DNase footprints dataset is an informative annotation for differentiating between regulatory and background variants in this dataset (with a β coefficient of 2.62; see Table 3.5).

FANTOM5 CAGE data: The FANTOM5 consortium published new data early in 2014, comprehensively mapping TSS and their promoters across 975 human samples (573 primary cells, 152 tissues and 250 cell lines)(Consortium et al., 2014). I hypothesised that inclusion of this data would lead to more accurate promoter identification, thereby improving the accuracy of my position ranking.

Transcribed Enhancers: A by-product of the FANTOM5 project was the identification of CAGE defined transcribed enhancers (Andersson et al., 2014). These were shown to be more accurate predictors of real enhancers than ENCODE data. I therefore tested this feature's ability to predict variant functionality.

TFBSs: I included the wgEncodeRegTfbsClusteredV3 dataset from the UCSC Genome Browser in the parameter optimisation. The highest peak signal for any transcription factor (TF) across all cell lines was used to rank SNPs.

3.3.2.3 Optimisation of the remaining annotation features

In addition to updating the feature annotation data for MAF and DNase HS and incorporating several new features into the model, I also re-evaluated how the remaining features (position, chromatin states and conservation) contributed to the performance of SuRFR.

Position rank:

Hindorff et al. (2009) and Schork at al. (2013) suggested that disease associated variants are more likely to occur in particular position categories, such as enhancer elements and promoters, more often than others ((Hindorff et al., 2009); (Schork et

al., 2013)). Using the genomic enrichment results for disease variants from both these sources, I re-ordered the ranking of the position categories (see Table 3.2). I further modified this annotation feature by incorporating the FANTOM5 TSS data.

Old data	β coeff	p-value	New data	β coeff	p-value
Ernst	2.3776	<2e-16	New Ernst	4.458	<2e-16
DNase-c V1	2.2516	<2e-16	DNase_c V2	2.406	<2e-16
			DNase_F	2.6197	<2e-16
Position	7.888	<2e-16	New Position	11.6197	<2e-16

Table 3.5 Comparison of the old versus new feature annotations for the Chromatin states (Ernst), DNase HS data (DNase HS clusters: DNase_c; and DNase footprints: DNase_F). Regression was performed on the normalised ranks of each annotation feature, allowing the β coefficients to be directly compared.

Chromatin states:

Multivariable logistic regression on the full training/validation dataset was used to assess the relationship between each of the 15 chromatin states and variant class; the ß coefficients indicating the relative correlation of each annotation category to the classifier (i.e. positive or background variant). Table 3.6 shows the average ß coefficients for each chromatin state across the nine cell lines GM12878; H1-hESC; K562; HepG2; HUVEC; HMEC; HSMM; NHEK; and NHLF.

The average β coefficients showed pronounced grouping of "like" categories (promoter with promoter, weak enhancer with weak enhancer, etc.) of chromatin states with similar β coefficients. Using this information I collapsed these similar categories into 10 classes of chromatin states. The two 'Repetitive/CNV' categories had high standard error rates and noise in the data. This class was positioned in the middle of the ranking, between classes that correlated positively with the data classes and those that had a negative

correlation. Using these data I defined the rank order for the chromatin state classes, shown in Table 3.4.

Chromatin State	Average β coefficient across nine cell lines	Std. Error	Pr(>|z|)
15_Repetitive/CNV	13.399	322.129	0.966
1_Active_Promoter	3.551	0.353	0.000
2_Weak_Promoter	2.127	0.338	0.000
3_Poised_Promoter	2.565	0.558	0.006
6_Weak_Enhancer	1.099	0.278	0.008
4_Strong_Enhancer	1.500	0.389	0.025
12_Repressed	0.404	0.142	0.029
9_Txn_Transition	-0.051	0.401	0.729
8_Insulator	0.030	0.412	0.585
10_Txn_Elongation	-0.363	0.191	0.140
7_Weak_Enhancer	-0.072	0.244	0.541
11_Weak_Txn	-0.532	0.110	0.000
13_Heterochrom/lo	-0.659	0.057	0.000
5_Strong_Enhancer	0.010	0.410	0.280
14_Repetitive/CNV	-4.811	246.088	0.867

Table 3.6 Multivariable regression ß coefficients (column 2), standard error rates (column 3) and p value (column 4) for each of the 15 chromatin states averaged across nine cell lines.

Conservation:

I next compared the relative contribution of each of the two conservation methods, GERP and PhastCons, and their combined contribution, to the model's ability to prioritise functional over non-functional variants. I did this by performing multivariable regression on the full HGMD/ENCODE training/validation dataset, to see how well each predictor could differentiate SNPs on the binary classifier (functional/non-functional). Table 3.7 shows the ß coefficient for the combined conservation rank (GERP + PhastCons) and Table 3.8 shows the ß coefficients for each tool individually. The combined conservation score does not correlate well with the classification of functional versus non-functional ($p>0.5$). However, GERP is positively correlated with the classifier, with a ß coefficient of 0.5709 (p-value < 2.78e-15), while PhastCons has a strong negative correlation with the classifier, with a ß coefficient of -1.6903 (p value < 2.38e-9). This indicated that using GERP on its own would have more power than combining GERP and PhastCons together, and that PhastCons is a poor predictor of single SNP function.

	Estimate	Std. Error	z value	Pr(>|z|)
(intercept)	-0.05663	0.04784	-1.184	0.2365
HGMD_ALL$Conservation.rank	0.20567	0.10895	1.888	0.0591

Table 3.7 Multivariable regression output for the combined conservation rank (GERP + PhastCons) on the full training/validation dataset.

	Estimate	Std. Error	z value	Pr(>|z|)	
(intercept)	-6.16035	0.27235	-22.619	< 2e-16	***
HGMD_ALL$GERP	0.57093	0.07227	7.900	2.78E-15	***
HGMD_ALL$PhastCons	-1.69035	0.28320	-5.969	2.39E-09	***

Table 3.8 Multivariable regression ß coefficients on the full training/validation for the two conservation methods individually: GERP and PhastCons.

3.3.2.4 Multivariable regression to select parameter boundaries

Multivariable regression on the training/validation dataset was used to define the upper and lower parameter boundaries for the modified grid search algorithm used for the parameter optimisation step. This was performed for each of the three training/validation datasets: ALL, DM and DFP (Tables 3.9, 3.10 and 3.11 respectively). Using these ß coefficients I chose positive, whole integer parameter ranges for each of the annotation categories for the parameter optimisation of SuRFR. The parameter boundaries used for each of the three datasets are shown in Table 3.12.

ALL training/validation dataset	β coefficient	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.7960	0.2198	-21.8243	0.0000
HGMD_ALL$F_Position.rank	6.3699	0.2763	23.0509	0.0000
HGMD_ALL$DAF.rank_normalised	-1.3615	0.1732	-7.8610	0.0000
HGMD_ALL$DNase.foot.av.rank	0.7172	0.2431	2.9504	0.0032
HGMD_ALL$Conservation.rank	0.6396	0.1648	3.8804	0.0001
HGMD_ALL$E.DNase.av.rank	-0.0496	0.2066	-0.2402	0.8102
HGMD_ALL$Ernst.Av.new.rank	1.9758	0.2798	7.0606	0.0000
HGMD_ALL$Enhancers.rank	-0.1061	0.6572	-0.1614	0.8718
HGMD_ALL$TFBSs.rank	1.1599	0.2031	5.7115	0.0000

Table 3.9 Multivariable regression ß coefficients, standard errors, z values and p-values for the ALL training/validation data.

DM training/validation dataset	β coefficient	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.8927	1.2509	-9.5074	0.0000
HGMD_DM$F_Position.rank	7.5632	0.7236	10.4527	0.0000
HGMD_DM$DAF.rank_normalised	5.0458	0.6299	8.0101	0.0000
HGMD_DM$DNase.foot.av.rank	0.1967	0.6003	0.3277	0.7431
HGMD_DM$Conservation.rank	2.4449	0.4928	4.9612	0.0000
HGMD_DM$E.DNase.av.rank	-0.3372	0.5647	-0.5971	0.5504
HGMD_DM$Ernst.Av.new.rank	2.7075	0.7680	3.5253	0.0004
HGMD_DM$Enhancers.rank	-12.5078	738.4351	-0.0169	0.9865
HGMD_DM$TFBSs.rank	2.0049	0.5556	3.6083	0.0003

Table 3.10 Multivariable regression ß coefficients, standard errors, z values and p-values for the DM training/validation data.

DFP training/validation dataset	β coefficient	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9183	0.3479	-8.3891	0.0000
HGMD_DFP$F_Position.rank	5.6870	0.4372	13.0080	0.0000
HGMD_DFP$DAF.rank_normalised	-3.5053	0.3330	-10.5260	0.0000
HGMD_DFP$DNase.foot.av.rank	1.1763	0.3951	2.9772	0.0029
HGMD_DFP$Conservation.rank	0.0569	0.2530	0.2249	0.8220
HGMD_DFP$E.DNase.av.rank	0.2229	0.3306	0.6743	0.5001
HGMD_DFP$Ernst.Av.new.rank	1.6903	0.4512	3.7462	0.0002
HGMD_DFP$Enhancers.rank	1.4698	1.3626	1.0787	0.2807
HGMD_DFP$TFBSs.rank	0.5706	0.3302	1.7279	0.0840

Table 3.11 Multivariable regression ß coefficients, standard errors, z values and p-values for the DFP training/validation data.

Model	MAF	Conservation	Chromatin States	DNase HS	Position	DNase Footprints	Enhancers	TFBSs
ALL	0-2	0-3	0-8	0-1	0-16	0-3	0-1	0-5
DM	0-13	0-7	0-8	0-1	0-18	0-2	0-1	0-6
DFP	0-1	0-1	0-6	0-1	0-15	0-6	0-6	0-3

Table 3.12 The upper and lower boundaries of the weighting parameters chosen to be tested using the grid search algorithm. Column one describes the three models (ALL, DM and DFP) and each subsequent column shows the range of integer parameters used for the model parameter optimisation.

3.3.3 Ten-fold cross-validation

The ALL, DM and DFP training/validation sets were further partitioned into ten equal folds for ten-fold cross-validation, ensuring no overlap existed between any of the training/validation datasets and the hold out test datasets. I performed weighting model parameter optimisation and ten-fold cross-validation on each of these three datasets and assessed the performance and generalisability of SuRFR using ROC curves and AUC statistics.

3.3.3.1 Training and validation (AUCs, errors, specificity and sensitivity)

The optimum weighting model for each dataset was chosen based on the highest average training/validation AUC with a performance error of less than 0.005. The AUCs for the top 1% of weighting models were very similar, differing by less than 0.003 (ΔAUC ALL: 0.0026; ΔAUC DM: 0.0021; ΔAUC DFP: 0.0011), suggesting a smooth parameter space with few fine-grained local optima. Performance errors for each model (ALL, DM and DFP) were calculated as the difference between the average training and validation AUCs. The AUCs and error rates for each model are shown in Table 3.13. Each model

performed well on the training/validation data, with AUCs ranging from 0.908 to 0.976 and performance errors of less than 0.004, indicating that each model can successfully prioritise functional over background variants with high specificity and sensitivity.

Model	Training AUC	Validation AUC	TEST AUC	Performance error	Generalisation error
ALL	0.944	0.944	0.909	0.000	0.035
DM	0.976	0.976	0.956	0.000	0.020
DFP	0.912	0.908	0.897	0.004	0.013

Table 3.13 Average training, validation and test AUCs for the three SuRFR models run on the cross-validation datasets.

3.3.3.2 Hold out test dataset

The top weighting models for each of the three data classes ALL, DM and DFP, were next run on the hold out test dataset to establish SuRFR's generalisation error. These data are shown in Table 3.13 and Figure 3.3. Again, each of the three models performed with high specificity and sensitivity, producing AUCs of 0.897 to 0.956 and generalisation errors less than 0.035. This suggests that all models are likely to perform equally well on novel data.

3.3.4 Characterisation of regulatory variant classes

The best weighting models for each of the three variant classes are shown in Table 3.14. This data shows that each of the three variant classes is best prioritised by a different combination of genomic annotations. The most informative annotation category across all three variant classes was position (SNP position relative to genes). MAF was a very useful annotation for the prioritisation of DM variants over background, but was not at all useful for prioritising the ALL or DFP classes of regulatory variants. In contrast,

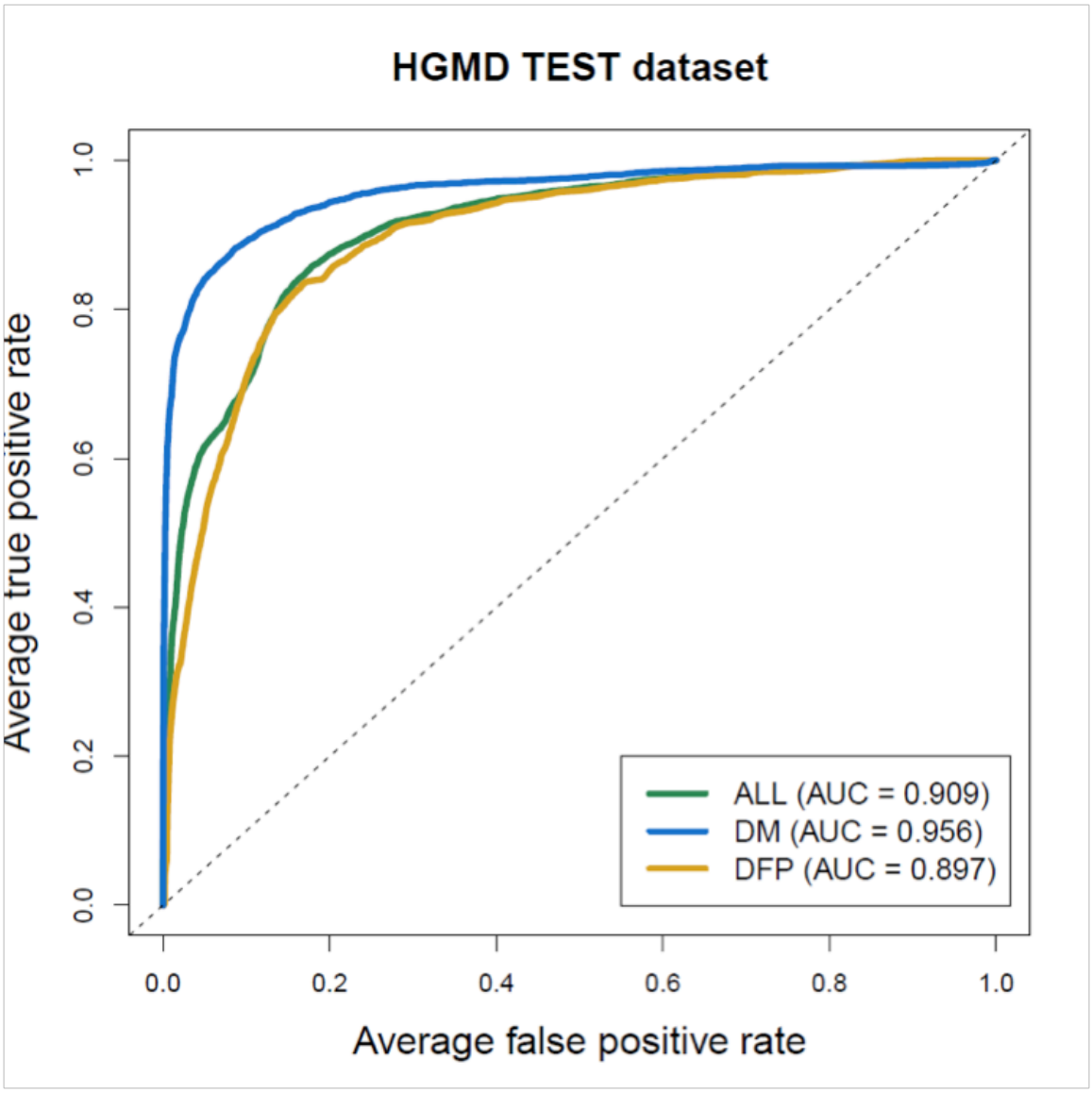


Figure 3.3 ROC curves and AUCs for the three SuRFR models (ALL: green; DM: blue; and DFP: gold) run on the hold-out test dataset. Y-axis represents the average true positive rate; the x-axis represents the average false positive rate and the grey dotted line represents random chance.

conservation was a relatively uninformative annotation: it provided only a minor contribution to SuRFR's ability to prioritise the DM variants, played an even smaller role in the prioritisation of the ALL variants and had no role in the prioritisation of the DFP variants.

The redefined chromatin states had a variable impact on the ability of SuRFR to distinguish functional from non-functional variants, most effectively prioritising the DM variants, closely followed by DFP and being least effective at prioritising the ALL category of variant. In contrast, the TFBS annotation consistently added to the correct ranking of true variants in all three variant classes.

Multivariable regression suggested that DNase HS and DNase footprints are highly correlated features, which may suggest that they provide similar input to the prioritisation of known regulatory variants (Table 3.5). However, when these two features were incorporated in the same model, the DNase footprints were more highly correlated with correct prioritisation of known variants than the DNase HS clusters (Tables 3.9, 3.10 and 3.11). This was reflected in the subsequent weightings assigned to the two annotation categories.

Model	MAF	Conservation	Chromatin States	DNase HS	Position	DNase Footprints	Enhancers	TFBSs
ALL	0	1	1	0	8	0	1	3
DM	12	2	6	1	15	1	0	5
DFP	0	0	3	1	15	3	5	2

Table 3.14 Parameter weightings for best performing weighting model for each variant class from the ten-fold cross-validation analysis. The first column lists the three weighting models (ALL, DM and DFP). Each subsequent column represents a different annotation class. The values represent the weightings of each annotation class defined in each weighting model.

3.3.5 Additional test datasets: HBB and RAVEN

As a further test of the generalisability of SuRFR, I tested the three models on the HBB and RAVEN datasets presented in Chapter 2. The HBB (27 non-coding SNPs not present in the HGMD dataset) and RAVEN (95 regulatory variants not in the HGMD dataset) true positive SNPs were spiked into the 44 ENCODE pilot regions (minus the training/validation SNPs). Figure 3.4 shows the ROC curves and AUCs for these two analyses.

All three models prioritised the non-coding HBB variants with very high specificity and sensitivity; the DM model performed the best, with an AUC of 0.989, followed by the ALL (0.981) and DFP models (0.956). More variation existed in the ability of SuRFR to prioritise the RAVEN variants over the background ENCODE variants; the ALL and DFP classes generated AUCs of 0.921 and 0.937 respectively, while the DM model achieved an AUC of 0.797.

3.3.6 Background variants as known functional variants

The RAVEN background variant dataset contains 3,856 variants all located within 10kb of genes conserved between mice and humans (Andersen et al., 2008). As a negative control I performed a bootstrapping analysis, running SuRFR on 100 randomly sampled subsets of the RAVEN background variants against the remaining background variants. Each subset contained 95 variants, each of which was defined as a "known" (positive) variant; the remaining 3,761 background variants being classed as background (control) variants. The average AUC calculated across the 100 bootstrapping sets was 0.50 (Figure 3.5), indicating that the background variants were not prioritised any better than would be expected by chance. In contrast, the 95 "real" true positive RAVEN variants spiked into the same background dataset produced AUCs of 0.83, 0.845 and 0.842 for the ALL, DM and DFP models respectively. This demonstrates that SuRFR is capable of prioritising functional variants better than non-functional variants.

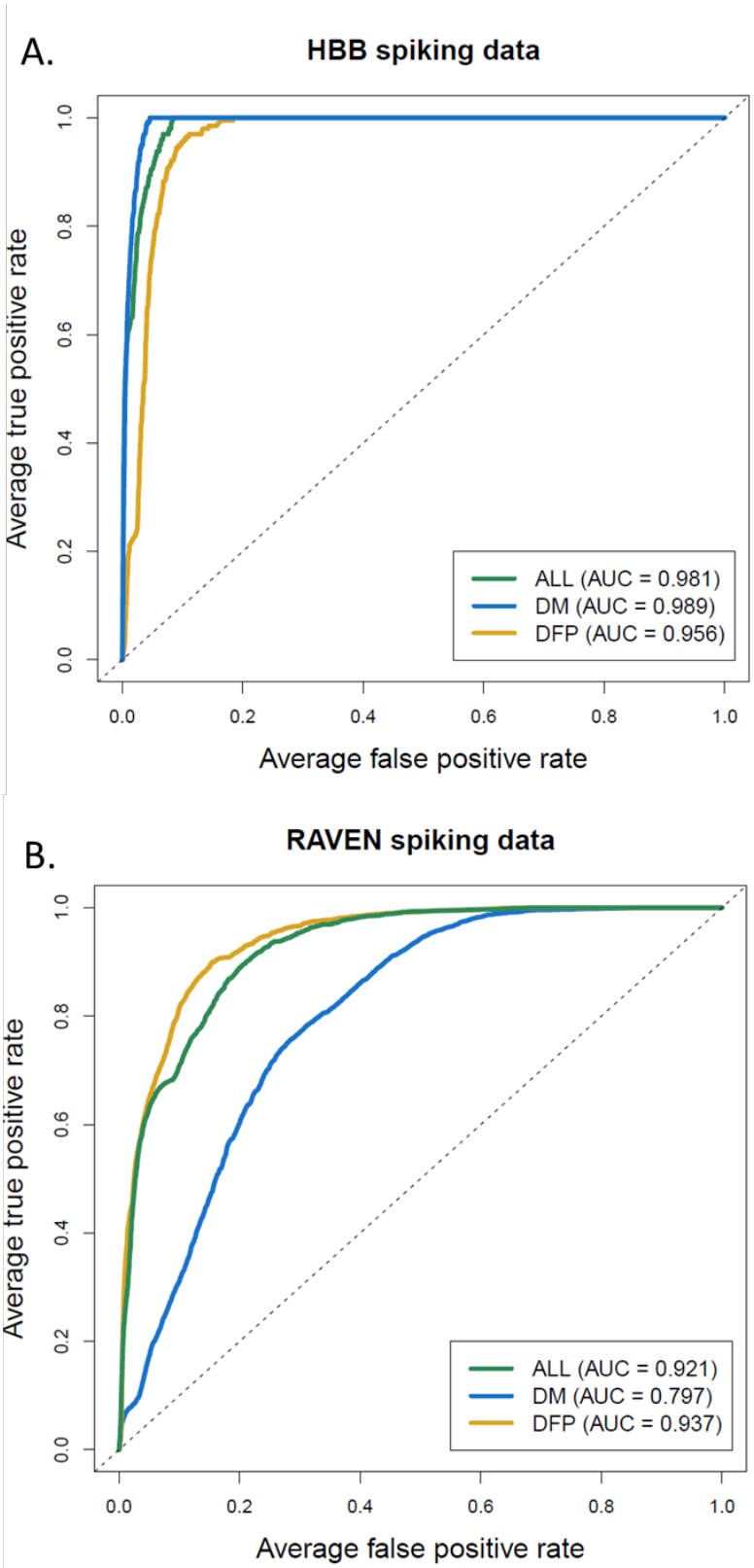


Figure 3.4 Mean ROC curves (y-axis: True positive rate; x-axis: False positive rate) and AUCs for the three SuRFR models (ALL (green) DM (blue) and DFP (gold)) run on: a) HBB non-coding pathogenic and b) RAVEN non-coding regulatory datasets spiked into the ENCODE pilot project background dataset. The dotted grey line indicates random chance.

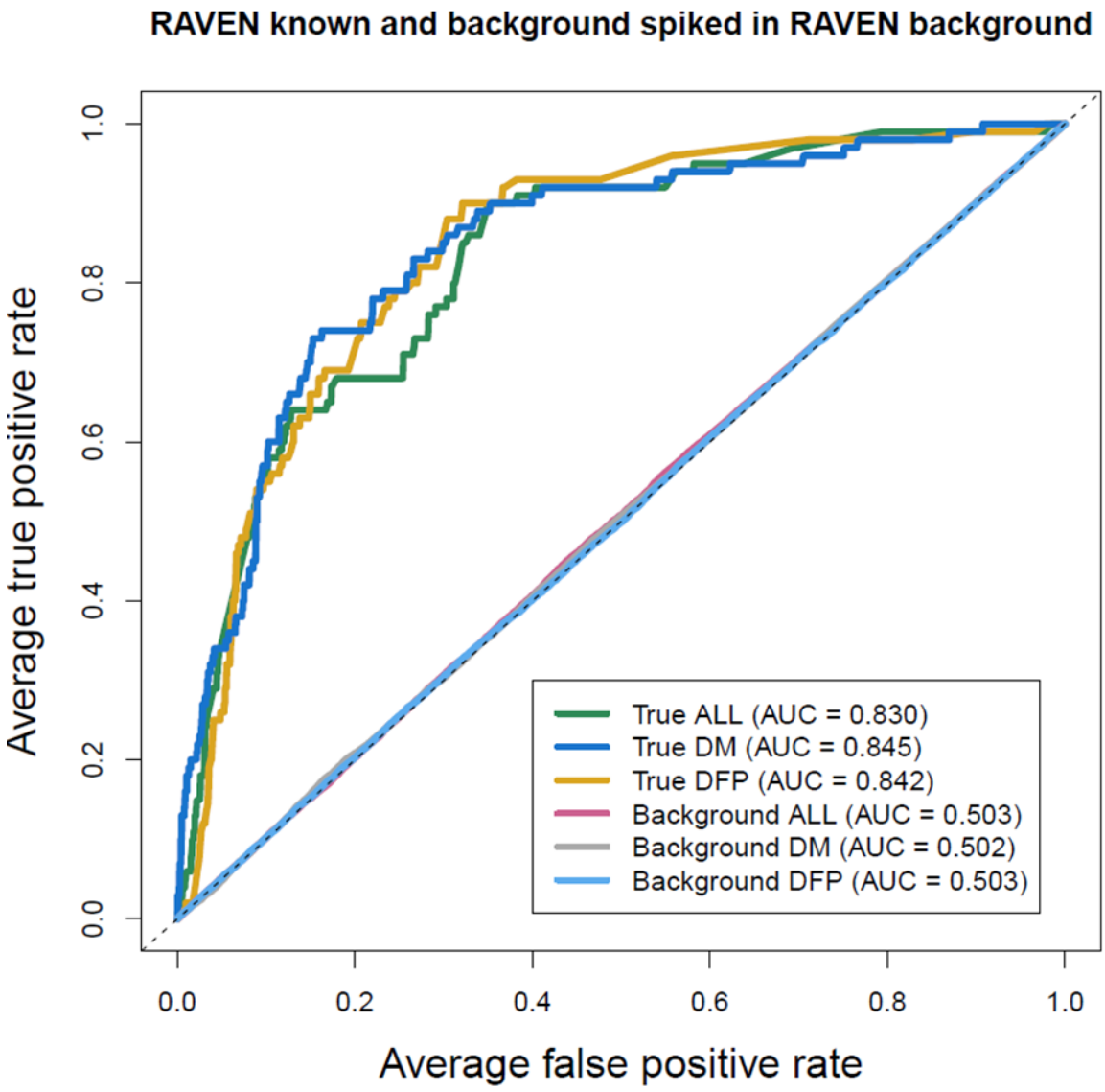


Figure 3.5 ROC curves and AUCs for the three SuRFR models (ALL, DM and DFP) run on i) 100 background datasets classed as functional and ii) the true functional variants run against the background dataset. These results show that SuRFR does not rank the background variants any better than expected by chance, supporting earlier results that showed SuRFR can prioritise functional over background variants.

3.3.7 R package details

Using the R function package.skeleton, I converted my R code into an R package, SuRFR. The SuRFR R package is available from: *http://www.cgem.ed.ac.uk/resources/*

In addition, I wrote a user manual and a sweave vignette for this R package. The user manual can be found in Appendix B while the sweave vignette can be found in Appendix C.

The data presented in this chapter (and Chapter 4) were published by Genome Medicine in October 2014 (Ryan et al., 2014) (see Appendix D).

3.4 Summary and Discussion

3.4.1 Summary of Results

The aim of this chapter was to improve the model I developed in Chapter 2. This was achieved by implementing both a modified model framework and a more structured model assessment protocol. The new model framework was based on a rank-of-ranks, removing the need for arbitrary thresholds and treating each annotation category equally, thereby removing any bias and providing consistency across the annotations. The improved model assessment protocol combined a modified grid search algorithm and ten-fold cross validation. This protocol made use of a benchmarking dataset consisting of regulatory variants from HGMD and background variants from the 1000 Genomes EUR population located within the ENCODE pilot project regions. Performance was measured using ROC curves and AUCs. Three models were developed from this analysis: the ALL, DM and DFP models. The results of the cross-validation showed that each model was able to prioritise their corresponding class of regulatory variants above the background variants with high sensitivity and specificity (AUCs between 0.897 and 0.976: see Table 3.12 and Figure 3.4) and low performance and generalisation errors (Table 3.12). These results suggest that SuRFR does not suffer from over-fitting and is likely to perform equally well on novel data.

3.4.2 Changes to feature annotation data

Projects such as ENCODE are continuing to provide the scientific community with genomic annotation data, from TFBSs, to RNA assays and a range of DNA and histone modifications, across an ever increasing number of cell lines. Genomic annotation data is therefore not static but constantly being updated and expanded. As such it was important to update and expand the annotations used by SuRFR to prioritise putative functional variants.

Table 3.1 lists the annotation features used in this modified version of SuRFR. These features can be divided into three classes: i) annotation features from the original model for which new releases have become available (making use of larger numbers of cell lines and modified experimental design); ii) new annotation features (features that had

not been released at the time of the initial build of SuRFR); and iii) annotations from the original model of SuRFR but incorporated differently (optimised integration of related classes).

By including updated versions of the annotations MAF and DNase HS clusters, I expected to improve the accuracy of my method, as the updated annotation data should be more accurate. Similarly, I hoped to improve accuracy by including additional features that I hypothesised would improve the prioritisation of functional variants over background variants. The new annotations I chose to include were DNase footprints, FANTOM5 CAGE defined promoters, FANTOM CAGE defined transcribed enhancers and TFBSs. Lastly, by changing the way both the position rank and chromatin state rank were calculated, using a more formalised approach (multivariable regression), I intended to optimise the information content of these features.

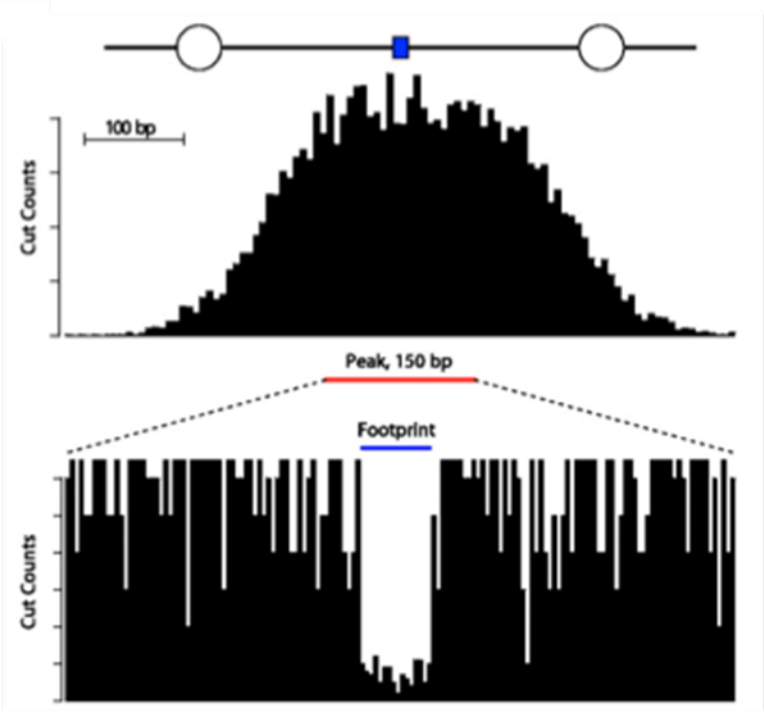


Figure 3.6 DNase HS versus DNase footprint. Figure from (Vernot et al., 2012).

DNase HS data (clusters and footprints): While DNase HSs indicate regions of open chromatin, DNase footprints more specifically reveal single protein-binding events (Madrigal and Krajewski, 2012), as illustrated by Figure 3.6. I hoped, by combining DNase footprinting data with DNase HS data, to improve the specificity of SuRFR, as regions where the two features overlap are potentially more likely to mark true binding events than either feature on its own (Figure 3.6).

I hypothesised two scenarios for these data: i) that there would be a combinatorial effect, each feature aiding the other and adding specificity to the prioritisation of regulatory variants; or ii) that the footprints, being more specific elements, demarking sequences of regulatory factor occupancy on a nucleotide level (in comparison to DNase HS clusters, which mark regions of open chromatin), would provide greater accuracy than the DNase HS clusters and would therefore remove the need for the DNase HS cluster feature. The results from the cross validation analysis surprised me. Although scenario i) appears to be in effect for the DM class of regulatory variants and scenario ii) is true for the DFP class, neither explained why the ALL class of variants required neither DNase HS cluster data or the DNase footprinting data. An explanation for this could be that the ALL model placed greater emphasis on the TFBS annotation data than any other feature, bar position. As the TFBS annotation and DNase HS features all provide information on the likelihood of a variant overlapping a protein binding domain, there is a certain amount of redundancy between these features, thus explaining the lack of DNase HS features in the ALL model. In summary, although these two features provide largely overlapping data, the different variant models required different weightings of these features: for the DM class of variants, each of these features contributes equally to the correct prioritisation of functional variants; while the DFP model relies more heavily on the DNase footprint data; and the ALL model required neither feature.

Transcribed enhancers: the FANTOM5 project produced an atlas of active, transcribed, enhancer regions. These regions were defined by bidirectional CAGE tags, assayed across a range of samples, including 432 primary cells, 135 tissues, and 241 cell lines (Andersson et al., 2014). Using in vitro enhancer assays in HeLa cells, Andersson

et al. (2014) were able to show that bidirectional capped RNAs were a more accurate signature of active enhancers than enhancers predicted by DNase HSs or 'strong enhancer' chromatin states. This database of over 43,000 enhancer candidates was therefore as good a candidate annotation to test as the DNase HSs and the chromatin states. The cross validation analysis supported the inclusion of this feature in both the ALL and, particularly, the DFP model. This was in contrast to the results of the multivariable regression analysis (Tables 3.9, 3.10 and 3.11), which were inconclusive. However, I suggest that this issue derives from the fact that both the enhancer dataset and training/validation dataset are small. I propose this issue is one of data acquisition rather than a lack of correlation between this feature and regulatory effect. Larger numbers of known true positives are required to improve this analysis. In particular, there is an acquisition bias in most known regulatory variant datasets towards variants proximal to genes, specifically within promoter regions. This bias further reduces the likelihood of training data containing sufficient numbers of enhancer variants for us to expect a high correlation between regulatory variants and enhancer features. This feature is therefore an important one to retain in the SuRFR models, allowing us to detect more enhancer variants and reduce the bias away from promoter variants in any future validated regulatory datasets.

TFBSs: TFBSs tend to be short (4-10 bp) DNA sequences that occur repeatedly across the genome. TFBSs are important components of the human regulatory network and changes to these binding sites can affect the ability of transcription factors to bind to them, thus having an effect on function and potentially leading to a disease phenotype. However, only a fraction of predicted sites are real, active regions of transcription factor occupancy that play a role in gene regulation (Cuellar-Partida et al., 2012). Predictive methods, therefore, that use pattern recognition to identify putative novel TFBSs tend to identify a large number of false positives, and so are inherently error prone. In contrast, experimental methods such as chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) have been used to identify regions of (true) transcription factors occupancy genome-wide. The ENCODE consortia performed an integrative analysis on 161 transcription factors across 91 cell types to comprehensively map the human regulatory network ((Wang et al., 2012); (Gerstein et al., 2012)). These data are

catalogued and annotated by the Factorbook repository (Wang et al., 2013) and have been used to generate the annotation dataset wgEncodeRegTfbsClusteredV3. Based on this knowledge, I hypothesised that this dataset of ChIP-seq identified TFBSs would be an important feature to integrate into my prioritisation model. As expected, this annotation played an important role in prioritising functional regulatory variants over background variants, and this role was not variant class specific. In fact, after the position feature, this was the second most universally informative feature (tied with chromatin states). This result reinforces the rule that, wherever possible, experimental data should be used over predicted data, which is inherently error prone.

Chromatin states: In Chapter 2, I split the chromatin states into a binary classification (1/0), either active states with the potential to affect regulation, or inactive states (such as heterochromatin). These scores were summed across the nine cell lines included in the analysis, with possible scores ranging from 0 - 9. However, this did not take into account differences in predictive power between the chromatin states within each class. This also assumed that a SNP overlapping an informative feature in nine cell lines was nine times as informative as a SNP overlapping an informative feature in a single cell line. This is very unlikely to be the case, particularly as Ernst et al. (2011), showed that regulatory regions vary in activity levels across cell types and enhancers show very high tissue specificity (Ernst et al., 2011).

Subsequently, I used multivariable regression on the chromatin states to more accurately determine the correlation between the rankings of the chromatin states and the correct prioritisation of causal variants over background variants. This analysis highlighted a marked grouping of ‘like’ chromatin states (regulatory element classes: promoters, weak enhancers, strong enhancers, repetitive sequences, etc). This supported results from Ernst et al., showing that ‘like’ chromatin states correlated across cell lines: genomic regions annotated as enhancer states (strong or weak) and promoter states (active, weak and poised) in one cell line often remained the same class of state (enhancer or promoter) across the other cell lines (Ernst et al., 2011). I therefore used these results to group the chromatin states into ‘classes’ and ranked the SNPs on these classes (Table 3.3). The

updated chromatin class ranks showed a marked improvement in their correlation with regulatory status in the multivariable regression, with β coefficients increasing from 2.38 for the original ranks to 4.46 for the new rank order. In addition, the new chromatin state feature was the second most universally informative feature (tied with the TFBS data) in the cross-validation analysis, further supporting the data of Ernst and Kellis.

Position: Hindorff (2009) and Schork (2013) independently demonstrated enrichment of disease-associated variants in specific genomic locations more often than others. I, therefore, re-ranked the position categories to reflect the results of their analyses. Similarly, data from the FANTOM5 project provided more accurate data on TSSs across the genome, allowing more accurate mapping of SNPs to promoter and 10 kb upstream regions. These changes improved the accuracy of the position rank, as shown by the multivariable regression analysis performed on the old and new position ranks (ß coefficients changing from 7.89 to 11.62). Although position score had always been the most effective annotation in the prioritisation of functional regulatory variants over background, I anticipated that these changes would also improve the specificity and sensitivity of my model. This was indeed the case, as shown by the results of the three models run on the HBB and RAVEN datasets, both of which produced higher AUCs than the original analyses on these datasets (See section 3.4.5 for more details).

Conservation: Multivariable regression was performed on the two conservation methods, GERP and PhastCons, and their cumulative conservation score. This allowed me to assess the ability of each tool, individually and combined, to predict the functionality of the training/validation SNPs. Surprisingly, I found that the low correlation of the conservation score used in the original model (from Chapter 2) was due to a negative correlation of the PhastCons data to the SNP functional classification, masking the positive correlation of GERP to this classification. Both GERP and PhastCons are recommended in the literature as useful tools for nucleotide sequence based prediction and have been shown to perform to a similar extent (Pollard et al., 2010). This information is difficult to reconcile to the results of the regression analysis in section 3.3.2.3. While this incongruity is difficult to explain, the GERP regression result

appears to reflect previous reports of GERP's performance while the PhastCons does not. This might reflect the difference in resolution of the measures used by both tools (GERP provides a nucleotide level measure, while PhastCons provides a multi-nucleotide region measure). This could also suggest that the PhastCons result is an anomaly in the data. Additional information is needed to clarify this issue. Pending further investigation, I therefore chose to leave out the PhastCons annotation from all future analyses.

Annotation weightings: The SNP rankings for each of the annotation categories were combined into a cumulative rank-of-ranks. Rather than arbitrarily weighting each annotation parameter against the others, I used a model training and assessment protocol to identify the most informative combination of annotation weightings, optimising their relative contribution to the final ranking of SNPs. Using cross-validation and a benchmarking dataset of non-coding disease and regulatory variants and background variants of unknown function provided me with the statistical framework needed for rigorous model assessment as well as an estimation of how well SuRFR would perform on novel data.

3.4.3 Conclusions from cross-validation

The grid search algorithm is an exhaustive search of a manually selected subset of a defined parameter space. This method is commonly used for hyperparameter optimisation, model selection, and to prevent over-fitting. In this context, a hyperparameter is defined as a parameter of a prior distribution, in this case, the weighting of an annotation ranking. Two requirements of the grid search algorithm are: i) the user must manually define the search space; and ii) it must be guided by cross-validation (Hsu, 2010). This method is designed to maximise generalisation by exhaustively searching for the optimum hyperparameter set (in this case, the best combination of annotation weightings) across the parameter space.

There are many practical advantage of the grid search algorithm including: i) speed, as it is highly parallelisable (as the parameter evaluations are independent of each other); ii) simple set-up, as the grid search space can be constructed by brute force permutation; iii) flexibility, as the parameter boundaries can be changed; and most importantly, iv) ease of integration into the widely used cross-validation analysis framework. This method is therefore well suited to the task of model selection and assessment.

A disadvantage of this method is the computational cost of an exhaustive search, which can be outperformed by randomly chosen subsets of the parameter space (Bergstra, 2012). To improve the performance of this method, I adapted it in two ways. Firstly, I performed multivariable regression on the full training/validation set to guide the weighting parameter boundaries, thereby reducing the grid search space. Secondly, I restricted the parameter values to positive integers, further reducing the number of weighting models to be permuted. This reduced the computational intensity of the analysis (the final number of permutations permuted being just under half a million; n = 450,000). Furthermore, the performance of the modified grid search algorithm could be assessed by analysing the distribution of the AUCs produced during cross-validation. These data enabled me to evaluate how well the grid search algorithm worked in comparison to other machine learning approaches.

The AUCs of the top 1% of weightings models (ranked on maximum AUC with a performance error < 0.005) were closely clustered, suggesting the models represented by the group all scored very similarly, arguing for smooth parameter space with few fine-grained local optima. This suggested that the boundaries of the grid search algorithm were well chosen and the most informative subset of the parameter space was interrogated by this analysis. The low performance errors and generalisation errors from the ten-fold cross-validation provided additional evidence of the efficiency and success of this analysis, suggesting that SuRFR is able to prioritise real, functional variants over background variants and it will work equally well on novel data.

Studies have shown that no single machine learning algorithm outperforms all other methods on all data types and the most important factor affecting the performance and reliability of any machine learning algorithm is the training data used ((Tan and Gilbert, 2003); (Vanneschi et al., 2011); (Caruana, 2006)). I am therefore confident that the approach I have used is as effective as any other approach and this is in large-part due to the rigorous benchmarking data I have used.

3.4.4 Implications from characterisation of different regulatory variant classes

The ten-fold cross-validation and subsequent model testing using the hold out test dataset showed that the three classes of functional variants (ALL, DM and DFP) were each best prioritised by different combinations of annotation weightings. Whether this is because different classes of variants are caused by combinatorial changes to genomic features, or because these different variant classes lead to specific combinatorial patterns of genomic features (i.e. cause or effect), cannot be explained by this data alone. However, some of these patterns intuitively make sense. For instance, the DM class of variants were best prioritised by parameter models that included a strong weighting for MAF (rare SNPs ranked higher than common SNPs). This class of variant tends to give rise to rare, high penetrance, Mendelian disorders, with severe phenotypes. It is therefore not surprising that this class are enriched for rare variants and that MAF is a good feature to differentiate them from background variants. Interestingly, these DM variants were also consistently ranked higher than the background variants for a large range of annotation weighting models, suggesting that these variants are associated with changes across many functional annotation categories and are thus identifiable by a range of annotation weighting models.

In contrast, the DFP variants (GWAS significant SNPs with functional evidence) were more difficult to identify, with only a very specific subset of weighting models prioritising them over the background variants. This dataset consists of common SNPs with small effect sizes, likely to result in subtler changes to function (than the DM variants), which, as a result, could be more difficult to detect. This could explain why such a specific-combination of annotation weightings is required to correctly prioritise

them above the background variants. As these SNPs are from association studies, they are also likely to be common variants associated with lower penetrance, complex traits. It is unsurprising, therefore, that the DFP model does not find the prioritisation of rare variants to be a useful predictor.

Across all three variant classes, position was found to be the most informative annotation feature. This is in keeping with the literature, where it has been shown that the influence of a regulatory site on expression drops off almost linearly with distance from the TSS in a 10 kb range (Manolio et al., 2009) and that disease variants are enriched in certain genomic positions, such as coding and promoter regions, over intronic and intergenic regions (Schork et al., 2013).

The ranking of chromatin states (Table 3.3) was chosen based on multivariable regression on the full training and validation dataset, the promoter and enhancer chromatin state classes ranking higher than the other chromatin states. After the position feature, this was the second most informative annotation across all three variant classes. This is in keeping with the literature where it has been shown, for example, that disease variants are over-represented in strong enhancers (Ernst et al., 2011).

The next most informative feature across the three classes was TFBSs. This is not surprising, as changes to TFBSs may alter the binding ability of transcription factors, thereby having an impacting on function and regulation.

Non-coding disease associated variants are enriched in DNase HS and thus putative regulatory sites (Maurano et al., 2012). DNase HS clusters and DNase footprints are highly correlated and provide overlapping information; DNase HSs mark regions of open chromatin while the DNase footprints mark regions of transcription factor occupancy within these broader regions. Despite this, using both features in the same weighting model provides more information than using either feature on its own. This study showed both DNase HS clusters and DNase footprints to be informative markers

of functionality, though neither feature was weighted as strongly as I would have expected. An explanation for this is that DNase HSs and DNase footprints co-localise with many other features including enhancers, TFBSs and promoter regions, and their effectiveness is therefore masked by the inclusion of these other features.

The remaining features had more variant-class-specific roles, being informative in the prioritisation of one class but not necessarily the others (as shown for MAF above). For instance, the transcribed enhancer class of annotation does not correlate with the DM variants and is only modestly informative for prioritising the ALL class above background. In contrast, the transcribed enhancers are highly informative for prioritising the DFP variant class. It is difficult to draw any conclusive hypothesis from this result, as the transcribed enhancer dataset is very limited (roughly 40,000 enhancers across the entire genome) and the p-values from the multivariable regression (Tables 3.8, 3.9, 3.10) were non-significant, indicating there is a lot of variability in the data. More data is therefore needed to validate this result.

Historically, many of the tools used for discriminating functional from non-functional variants made use of evolution as a measure of deleteriousness (Cooper and Shendure, 2011). Phylogenetic and constraint based approaches are designed on the premise that genomic sequence elements that are conserved across species, or in excess of neutral expectation, are likely to have important functions. Therefore, when variation is identified within one of these highly conserved elements, it is predicted to have an impact on function, potentially leading to a disease phenotype. In contrast to this view from the literature, this study suggests conservation is not a particularly informative annotation, playing a minor role in the prioritisation of DM variants, an even smaller contribution to the prioritisation of ALL variants and not contributing at all to the discrimination of DFP variants above background. This could be due to redundancy amongst the annotations (other annotations masking the true information content of this feature), or it could be highlighting the fact that these features are not as enriched in conserved regions as previously assumed. Indeed, some studies have shown that conservation is in fact a poor predictor of regulatory function (Ritchie et al., 2014) and

there is extensive regulatory gain and loss between lineages, indicating that regulatory element positions fluctuate across evolution (Meader et al., 2010).

These data suggest that many annotation categories are correlated and specific subsets of annotations are required to best discriminate the different functional variant classes from the background variants.

3.4.5 Generalisability: performance on HBB and RAVEN

As a further test of the generalisability of the three SuRFR models, I ran them on two additional datasets: the HBB non-coding dataset and the RAVEN dataset, both spiked into the 44 ENCODE regions. All three models performed extremely well on the HBB dataset, with average AUCs ranging from 0.95 to 0.989; the DM model performing the best. The DM model's performance on these data is very similar to its performance on the cross-validation hold out test dataset, where it achieved an AUC of 0.956. This was not due to leakage between datasets, as SNPs present in the HGMD dataset were removed from the HBB dataset prior to testing. This result is unsurprising, as the HBB non-coding dataset contains variants that are very similar to the DM class of HGMD variants (disease mutations for a high penetrance Mendelian disease (beta thalassaemia)). The performance of the DM model is also comparable (AUCs: 0.989 vs 0.983), to the performance of the old model on the same data (Chapter 2, Table 2.4). This result was quite surprising, considering the old model was designed in an unsystematic, ad-hoc manner, and showed that my general premise in Chapter 2, though unjustified, was still good.

More variation could be seen in the performances of the three models on the RAVEN dataset; the ALL and DFP models performing roughly equally well (with AUCs of 0.921 and 0.937 respectively) and the DM model performing with a much lower AUC of 0.797. These results were not unexpected as the RAVEN dataset contains variants that are known to be regulatory, without necessarily a disease phenotype. Therefore, these variants are most similar to the DFP class of variants. As such, I would not expect the DM model to prioritise them as well as the ALL or DFP models. All three models

perform much better than the old model (Chapter 2 Table 2.4, AUC: 0.745) on this dataset.

The comparison of SuRFR's performance on the bootstrapping RAVEN background dataset provided a negative control, complementing the earlier tests of SuRFR's generalisability. These results showed that the positive (functional) variant datasets are not being ranked above background variants due to some artefact in the data, but are instead being truly ranked on their putative functionality.

Taken together, these results suggest that the old version of the prioritisation approach was more similar to the DM model of SuRFR and that the parameter optimisation procedure and cross-validation did improve the performance and generalisability of my prioritisation method. Not only have I made my analysis more robust, I have improved the accuracy and performance of SuRFR during the process.

3.4.6 Benefits of R package and Bioconductor

Implementation of SuRFR as an R package has many advantages, including speed, ease of use and increased market penetration. Integration of SuRFR into the widely used R environment provides flexibility, modularity, adaptability, ease of installation and updates. This facilitates the incorporation of additional modules, functions and annotations in the future and allows it to be combined with other R packages.

I have constructed the SuRFR R package in a way that allows the user to modify the features and parameters to suit their own requirements by specifying a custom model instead of the ALL, DM or DFP models. In addition, the MAF function makes use of a Gamma distribution to allow the optimal MAF range to be modified to suit each analysis. This is particularly useful for the analysis of GWAS data, which, generally consist of common variants, do not benefit from the default MAF setting (which prioritises unique and rare variants over common variants). Figure 3.7 shows three

examples of MAF settings: 3.7.A shows the prioritisation of unique variants above all others; B represents a scenario where SNPs with a MAF of 5% are prioritised highest; and C represents the prioritisation of variants optimised around a 20% MAF.

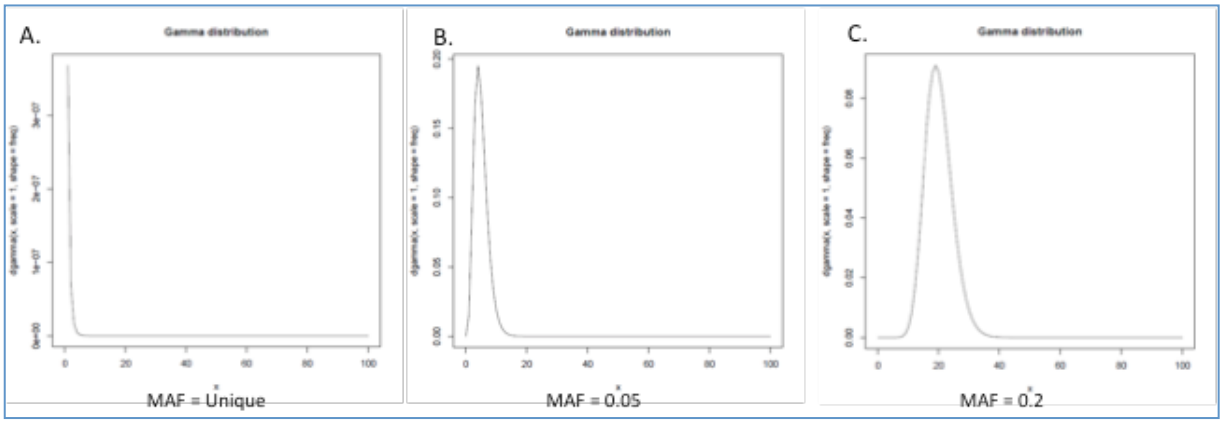


Figure 3.7. Example gamma distributions for three optimum MAFs: A. 0% (unique); B. 5%; and C. 20%. SNPs that are ranked based on their positions on the curve.

3.5 Conclusions

Robust performance assessment requires good benchmarking data and good performance evaluation methodology. I have used both to build upon the work initiated in Chapter 2 to develop a new variant prioritisation R package. I have shown that the final models and weightings chosen from the model assessment and parameter optimisation were able to prioritise known functional variants very well (with high AUCs) and also generalised well to novel data. This analysis also provided interesting biological insights into the functional annotations that correlate with different regulatory variant classes. SuRFR has many advantages over other methods. However, to confirm it is better than other comparable approaches I must do a formal comparative analysis. This will be the topic of Chapter 4.

Chapter 4: Comparison against competing approaches

4.1 Introduction

4.1.1 Review of SuRFR evaluation and performance

In Chapter 3 I described how I designed and tested a new R package, SuRFR, which prioritises genomic variants on the basis of functional annotation. This system makes use of data from multiple annotation categories, ranking SNPs in each category and using a weighting model to combine the individual ranks into a rank-of-ranks. Three weighting models were trained and validated using ten-fold cross-validation: a general model broadly applicable to any genomics analysis (ALL); a model designed for the prioritisation of rare disease variants (DM); and a model for the prioritisation of complex disease variants (DFP). The performance of each of these models has been assessed using a hold out test dataset and two additional, unrelated datasets (the HBB and the RAVEN datasets). All three models were shown to perform with high specificity, sensitivity, and generalisability on the data classes for which they were designed, suggesting that SuRFR will accurately prioritise putative functional variants for further investigation. However, the usefulness of this method cannot be fully established until it has been compared against other related tools.

During the first half of my PhD, no sufficiently comparable approach existed (see Chapter 2 for a summary of the tools that were available during that time); however, from late 2013 onwards, several new methods were published: GWAS3D (May, 2013); FunSeq (October, 2013); CADD (February, 2014); and GWAVA (February, 2014). In this chapter I will describe each of these methods; discuss their pros and cons; and question how well they perform against SuRFR in a comparative analysis.

4.1.2 Update on SNP prioritisation approaches

4.1.2.1 GWAS3D

GWAS3D is a method designed for the interpretation of genomic variants from GWAS studies, but it can also be used for the prioritisation of regulatory variants independent of

GWAS signals (Li et al., 2013). This method was developed as a web-based tool, implemented with a Perl-based web framework, 'Catalyst', using a MySQL database to store the annotation data. The workflow for GWAS3D is shown in Figure 4.1.

Given a set of GWAS data, GWAS3D performs a preliminary filter on the data to filter out less significant SNPs, removing query SNPs that fall above a user defined p-value cut-off. If the input data is not presented in VCF format (and so lacks reference and alternative alleles for each SNP), any SNPs that do not map to HapMap or 1000 Genomes are also filtered out. GWAS3D next identifies all SNPs in linkage disequilibrium (LD) (based on a user-defined LD standard) with each of the lead (query) SNPs. These SNPs are then annotated for a range of features (Figure 4.1, blue coloured block entitled 'GWAS3D Signals Mapping') including distal interactions, active histone marks, conservation, and user-defined data. Any SNP overlapping at least one signal is brought on to the next stage of the pipeline, while any SNP that does not overlap any signal is removed. Next, the binding affinity significance of each SNP for each of the transcription factor (TF) motifs from the ENCODE project is measured using position weight matrices of the transcription factor binding site (TFBS) motifs. The log-odds (LOD) of probabilities of binding for each of these motifs is then compared against the null distribution of binding affinity difference (calculated by permuting each ENCODE motif on all 52 million SNPs in dbSNP), and the p-value of each LOD calculated.

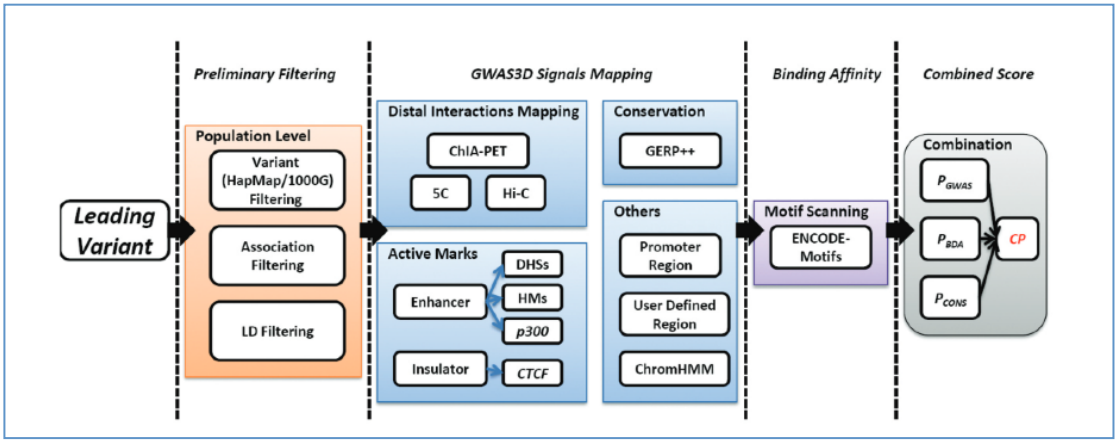


Figure 4.1. This diagram outlines the GWAS3D workflow, and has been taken from Li et al., 2013. See the description of the pipeline (Section 4.1.2.1) for full details (Li et al., 2013).

GWAS3D uses three measurements (GWAS, binding affinity and conservation) and their associated p-values to perform Fisher's combined probability test to calculate a combined p-value (CP) for each variant. The most significant p-value for any SNP in LD with the original lead variant is taken as the CP for that lead variant. All of the original input variants are then ranked on their CP values.

In contrast to the analysis of GWAS data, when a variant list without association data is used, the data cannot be filtered on association significance, nor can a final GWAS measurement be calculated and included in the CP value. In addition, during the LD filtering step of GWAS3Ds protocol, some of the input SNPs are likely to be replaced by alternative SNPs within the same LD block (as the query SNP may not be the lead SNP for that LD block). Therefore, although in theory this method can be used to assess non-GWAS data, in practice, this method is not well suited to such data. For this reason, I chose not to compare SuRFR against GWAS3D.

4.1.2.2 FunSeq

Function based prioritisation of sequence variants (FunSeq) is a variant prioritisation workflow developed to prioritise candidate non-coding cancer drivers (somatic mutations) based on patterns of selection, but which can also be used for personal genomics (germ-line mutations) (Khurana et al., 2013). Figure 4.2 describes how FunSeq scores variants on their predicted deleterious impact. The input SNPs are filtered at each level of the prioritisation workflow; only those SNPs that meet a level's criteria being retained. As a SNP passes each level, it achieves a higher score; scores range from 0 (no levels passed) to 6 (six levels passed).

Khurana et al. (2013) used population-variation data across 1,092 individuals from the 1000 Genomes (Phase 1) project to identify signatures of purifying selection. Using the full range of polymorphisms (SNPs, indels and structural variations) from these individuals, they studied patterns of purifying selection in different functional categories, defined by data from ENCODE. In particular, they looked at non-coding regions. The non-coding regions were first divided into broad categories based on their overlap with

functional data from ENCODE (such as TFBSs, DNase hypersensitive (DNase HS) regions, enhancers and non-coding RNAs). These broad categories were then further subdivided into 677 high-resolution categories (for instance, into different families of TFs). These categories were analysed to see if any were enriched for rare variants under very strong selection. In this way they identified 102 categories (of the 677) that showed statistically significant selective constraints, and specific genomic regions where variants are more likely to have strong phenotypic impact. Using these data, they defined the regions that contained a high fraction of rare variants (covering ~0.02% and ~0.4% of the genome) as "sensitive" and "ultra-sensitive" regions. Within these regions they found ~40 and ~400 fold enrichment respectively of disease-causing mutations from HGMD, therefore providing independent validation that these sensitive and ultra-sensitive regions are functionally important.

The authors next examined somatic variants (cancer variants) and found that 99% of somatic variants occur in non-coding regions, including TFBSs, non-coding RNAs and pseudogenes. Analysis of somatic variants from tumour and normal tissue from the same individual showed an enrichment for missense (~5x), loss-of-function (~14x), sensitive (~1.2x) and ultrasensitive (~2x) variants. Khurana et al. showed that somatic cancer variants are enriched for functionally deleterious mutations and somatic variants in the non-coding elements under strongest selection are the most likely to be cancer drivers.

Although the authors recommend that FunSeq would be best used for tumour genomics, they also suggest that it can be used for the identification of potentially deleterious variants in personal genomics. In this latter capacity, FunSeq is a comparable method to SuRFR. I therefore chose to include it in my tool comparison analysis.

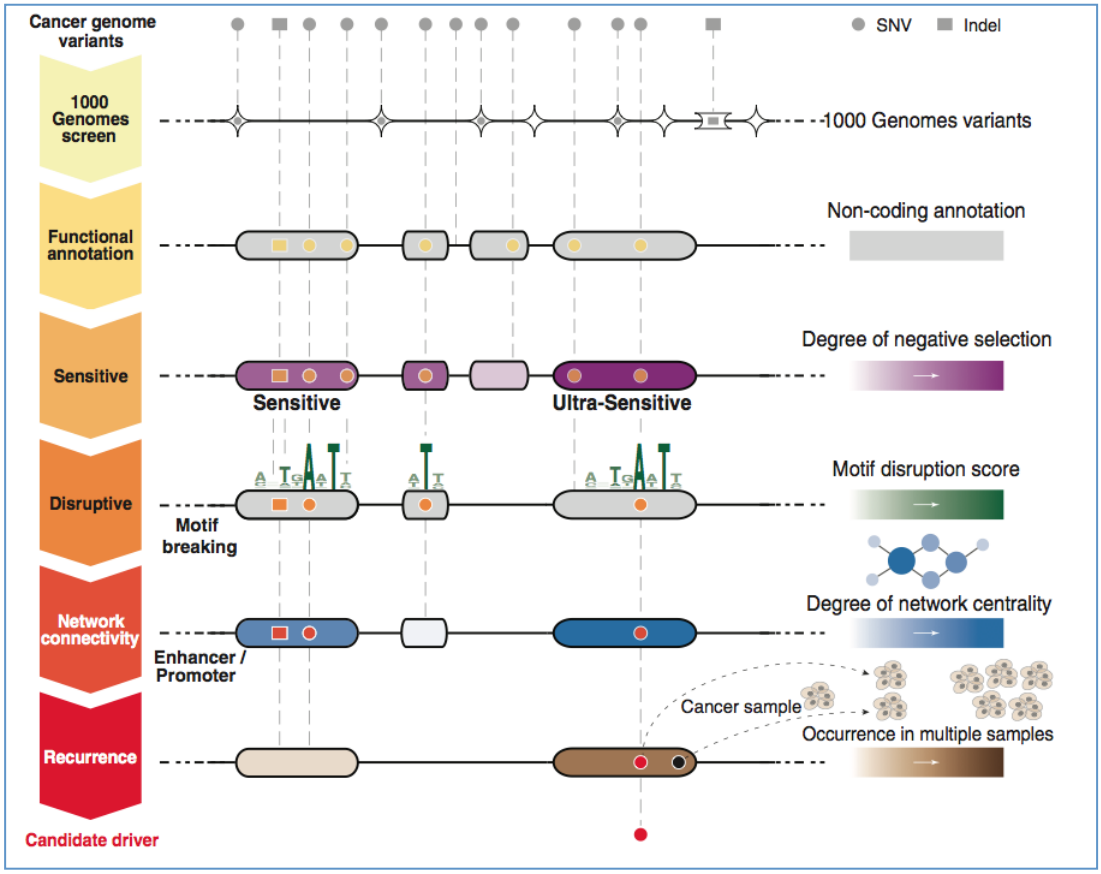


Figure 4.2. This figure presents a graphical overview of the FunSeq workflow, showing the filtering of SNPs to identify candidate non-coding cancer drivers based on patterns of selection. In the first step, the somatic variants are filtered to exclude 1000 Genomes polymorphisms. In the second step, only variants which overlap at least one of the non-coding annotations are retained. In step 3, variants that are located in "sensitive" regions are retained. In step 4, variants are prioritised on whether they disrupt a transcription-factor binding motif, while in step 5, variants are filtered based on whether they reside near the centre of a biological network. Lastly, variants are prioritised based on whether they are located in a region that contains mutations found in other (or multiple) cancer samples. This figure is taken from (Khurana et al., 2013).

4.1.2.3 CADD

Combined Annotation-Dependent Depletion (CADD) is a framework for integrating diverse genome annotation data, designed to score all possible SNPs and indels on deleteriousness (Kircher et al., 2014). This method was trained on a combination of observed and simulated variation. The observed data consisted of 14.9 million SNPs across the human genome with a derived allele frequency (DAF) ≥ 95% (1000 Genomes project), and, as such, are fixed or almost fixed in the population. In contrast, the simulated data consisted of 14.9 simulated *de novo* mutations derived using a custom empirical model of sequence evolution (motivated by parameters of the General Time Reversible (GTR) model (Tavaré, 1986)). The authors claimed an advantage of this training data was that it did not rely on catalogues of known pathogenic variants and therefore was not affected by the acquisition bias from which such data collections suffer.

CADD was built on the premise that selective constraint can be used as a measure of deleteriousness. Linear models were used to correlate 63 genomic annotation features with the observed and simulated datasets. This analysis showed that nearly all of the annotations could be used to discriminate observed from simulated variants. The strongest individual annotation metrics were found to be the conservation features. Using features derived from these 63 genomic annotations, Kircher et al. trained a Support Vector Machine (SVM) with a linear Kernel. From this, ten models (each independently trained on observed variants and different subsets of simulated variants) were developed. Spearman's rank correlations showed that these ten models were highly correlated (rho > 0.99). These ten models were averaged into a single model, which was used to score all (8.6 billion) possible SNVs in the genome (each position being a potential SNP or indel location (Figure 4.3)). The scoring system developed was called the C score. To simplify the C scores, Kircher et al. computed scaled C scores, which represent a variant's rank compared to the previously computed C scores for the 8.6 billion possible variants in the genome. Scaled C scores range in value from 0 – 99, higher scores suggesting greater deleteriousness than lower scores. Figure 4.3 shows that disease variants have on average higher scaled C scores than non-disease variants (see Table 4.3.c: "Olfactory") or random background variants (see Table 4.3.c: "Other")

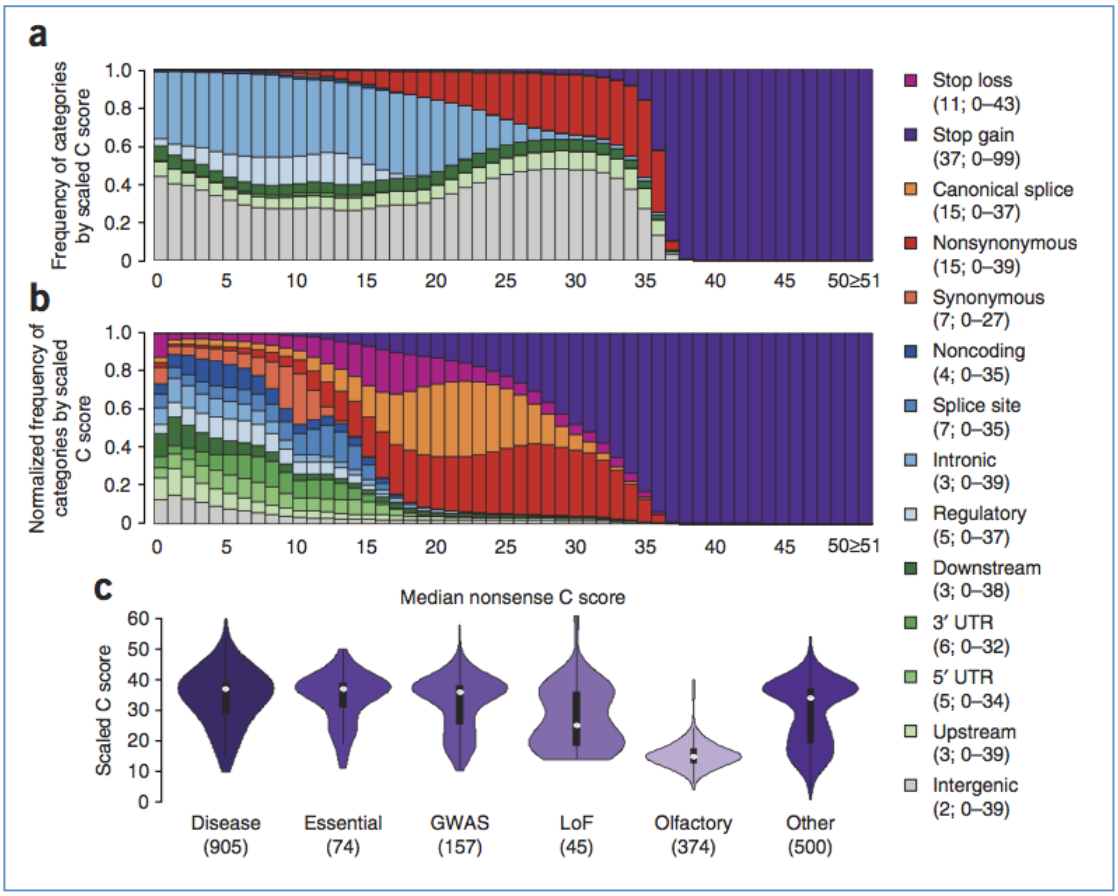


Figure 4.3 Relationships of the CADD scaled C scores (ranging from 0 to 99) to genome-wide variant consequence categories: a) ratio of variants within each variant consequence category for each C score bin (0-1; 1-2; 2-3; … 50-51; ≥51); b) ratio of variants within each variant consequence category, normalised by the number of SNPs with in each category, for each C score bin; the legend for each variant category includes, in brackets, the median and range of scaled C scores for that category; c) violin plots showing the median C scores of potential nonsense variants for 6 classes of genes (genes with at least 5 known pathogenic variants (Disease); genes predicted to be essential (Essential); Genes from GWAS studies harbouring significantly associated variants (GWAS); genes recorded by the 1000 Genomes project as harbouring at least two loss-of-function mutations (LoF); genes encoding olfactory receptor proteins (Olfactory) and a random selection of 500 genes (Other)), showing disease and functional nonsense variants are more likely to have higher C scores than non-disease (Olfactory) or random background nonsense variants (Other). Taken from Kircher et al. (2014) (Kircher et al., 2014).

The authors used this method to prioritise pathogenic and benign variants from the ClinVar database of clinical variation, and showed that CADD could prioritise the pathogenic SNPs above the benign SNPs better than the missense and conservation metrics SIFT, PolyPhen, GERP, PhastCons and PhyloP (Kircher et al., 2014). I therefore considered this method to be an appropriate comparative tool to test against SuRFR.

4.1.2.4 GWAVA

Genome-wide Annotation of Variants (GWAVA) was developed by Ritchie et al. (2014) to prioritise non-coding variants on the likelihood of functionality and, therefore, pathogenicity (Ritchie et al., 2014). This method combines multiple annotations to identify variants that are likely to be functional. These annotations include: regulatory features (such as DNase HS, TFBSs and RNA polymerase binding); genic context (position of variants relative to genomic features such as exons, introns, distance to the nearest TSS, etc.); human variation; conservation; and sequence context (such as G+C content, CpGs and repetitive elements). A modified random forest algorithm was used to train a classifier that integrates these individual annotations into a single metric to discriminate functional variants from background. The classifier was trained on data consisting of 1,614 known disease-implicated, regulatory variants from HGMD (downloaded from Ensembl), and three different background datasets. These background datasets consisted of randomly selected variants from the 1000 Genomes project (with minor allele frequencies ≥ 1%): i) 161,400 variants from across the genome; ii) 16,140 variants matched (to the HGMD regulatory variants) for distance to the nearest TSS; and iii) all variants within a 1 kb window of each of the HGMD variants (5,027 variants). From these three training datasets, three distinct classifiers were developed. Model training and validation was performed using ten-fold cross-validation and performance was measured using ROC curves and AUCs, which showed that the relative performance of the three models improved as the background variant datasets became less stringently matched to the known HGMD variants (Figure 4.4). Independent validation, using pathogenic variants from the ClinVar clinical variant database against non-pathogenic ClinVar variants and 1000 genomes background variants matched by distance to the nearest TSS, showed GWAVA was successfully capable of prioritising pathogenic variants above background variants.

Due to the similarities in the feature sets and model training methods used by both GWAVA and SuRFR, I considered this the most similar method to SuRFR.

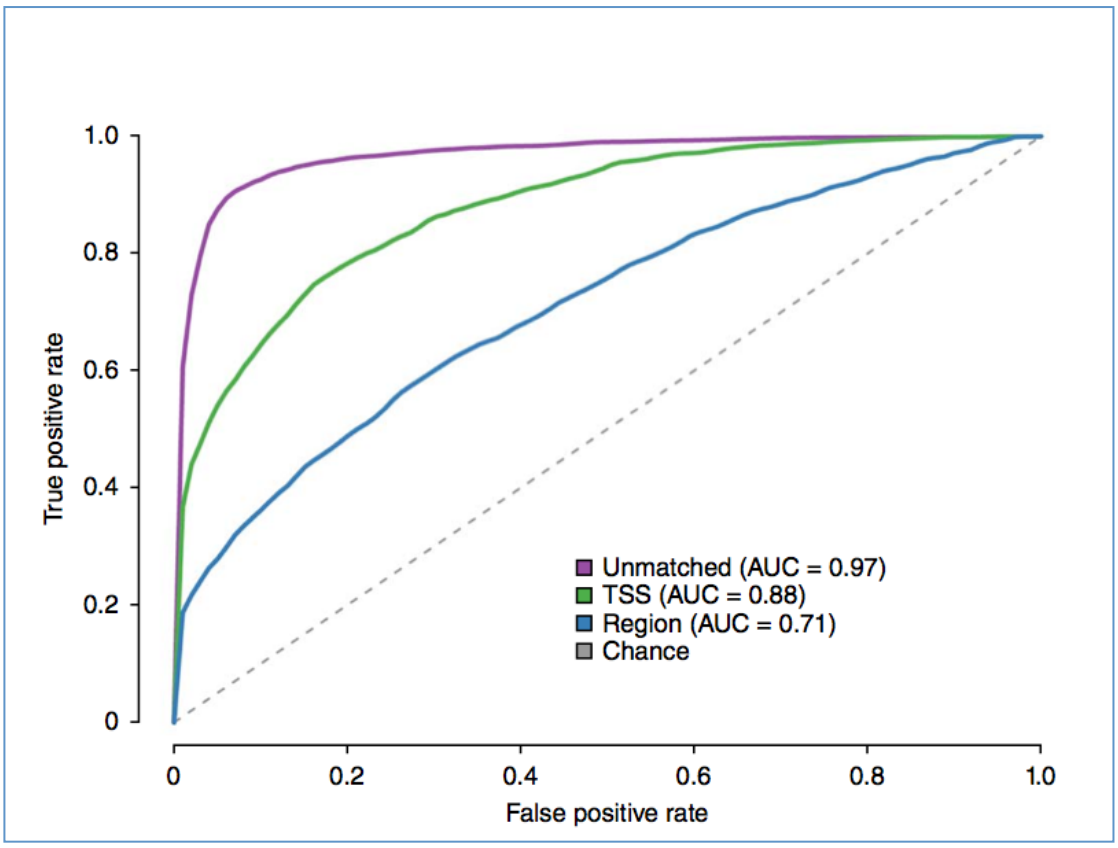


Figure 4.4 Mean ROC curves and AUCs from the ten-fold cross-validation experiments of GWAVA on the three training datasets. Taken from Ritchie et al., 2014 (Ritchie et al., 2014).

4.1.3 Additional datasets

As GWAVA was trained on HGMD data that contained variants that overlap variants from my hold out test dataset, the HBB non-coding dataset and the RAVEN dataset, GWAVA's performance on these three datasets would be inflated due to over-fitting. This meant that none of these datasets could be used for the comparison of SuRFR, CADD, FunSeq and GWAVA. Therefore, an additional, unbiased, dataset was required. The most obvious data to use were the two ClinVar datasets used by Ritchie et al. (2014)

to evaluate GWAVA's performance on novel data, as they had not been used to train any of these four models.

4.1.3.1 ClinVar

The ClinVar database is a public archive of medically important variants and phenotypes, officially launched in April 2013 (Landrum et al., 2014). This resource is funded and curated by the US National Institute of Health (NIH). An advantage of this database is that all submissions are categorised both by data source (whether from clinical tests, literature review or research) and review status (the extent of variant verification: single submission; multiple submissions; or reviewed by an expert panel). In addition, many of these variants have been functionally validated. Using these and other filters, users can select subsets of data that meet specific, user-defined criteria. As this resource is still quite new, it is more limited in size than other resources, such as HGMD. However, it also contains variants that are not yet present in HGMD, making it an excellent independent data source.

4.1.3.2 1000 Genomes variants matched for distance to the TSS

An interesting aspect of the datasets generated by Ritchie et al. (2014) for the training and validation of GWAVA is that the background datasets were matched to the positive variants by distance to the nearest TSS. This, importantly, allowed them to assess the performance of their method, excluding the effect of position. This allowed them to correct for the acquisition bias that exists in databases of regulatory variants (which tend to contain more variants proximal to the TSS than more distal variants). Using a similar method, I generated my own matched background dataset for the RAVEN regulatory dataset, consisting of variants from the 1000 Genomes European (EUR) dataset.

4.1.3.3 Complex trait related datasets

Projects that start with an association signal or a linkage region followed by sequencing or fine mapping of the region, and end with an experimentally, functionally validated regulatory disease variant form an important class of test dataset. It was important to

identify studies like these, to allow me to test the performance of each of the four prioritisation-methods in a less synthetic manner. The following sections describe three complex trait analyses that fit this criterion, which I then used to compare the performances of the prioritisation methods.

SORT1:

Musunuru et al. (2010) investigated a locus on chromosome 1p13 known to be strongly associated with low-density lipoprotein cholesterol levels (LDL-C) and cardiovascular disease (Musunuru et al., 2010). Fine-mapping in the genomic region responsible for LDL-C association, using SNPs genotyped from ~20,000 individuals of European descent (Kathiresan et al., 2009) identified 22 variants. Of these, the six SNPs with the highest association were clustered in a 6.1 kb non-coding region. Luciferase assays and electrophoretic shift assays (EMSA) demonstrated that one of the six, rs12740374, creates a binding site for the transcription factor C/EBP and alters liver-specific expression of the *SORT1* gene.

EGR2:

A good candidate for systemic lupus erythematosus susceptibility (SLE) is the early growth response 2 gene (*EGR2*). Myouzen et al. (2010) performed a case–control association study for SLE, of the 80kb region around the *EGR2* gene (Myouzen et al., 2010). This study identified a single non-coding SNP with a significant p-value. Functional characterisation (EMSA) of the SNPs in complete LD (R^2 = 1.0) with this tagging SNP showed that two SNPs had allelic differences in binding ability. Moreover, luciferase assays performed on these two SNPs showed that one (rs1412554) increased expression by 1.2 fold while the second, (rs1509957) repressed transcriptional activity.

TCF7L2:

In a search for variants associated with type-2 diabetes (T2D), Gaulton et al. (2010) identified a GWAS significant SNP (rs7903146) at the *TCF7L2* locus (Gaulton et al., 2010). This variant and five others in high LD with it were investigated using luciferase assays. Allelic differences in enhancer activity were observed for the tagging SNP, rs7903146.

4.1.4 Summary of chapter aims

The aim of this chapter was to compare the performance of SuRFR against three similar prioritisation approaches, to show if SuRFR is a useful addition to the field.

4.2 Methods

4.2.1 Running GWAVA

GWAVA can be used either as a web based tool that provides pre-computed scores for all known human variants, or run locally as a python script. I chose to use the downloadable command-line version as some of the variants in the test datasets are not included in the 1000 Genomes and so could not be annotated by the online version.

The GWAVA software requires the following python libraries (and their dependencies) to operate:

- *numpy (1.7.0)*
- *scikit-learn (0.14.1)*
- *scipy (0.11.0)*
- *pybedtools (0.6.4)*
- *pandas (0.12.0)*
- *tabix (0.2.5)*

Stewart Morris (SM) installed GWAVA and its environment variables on the server Wheeljack. GWAVA operates via a two-step process: first building a variant annotation table; and second, using the annotation table to compute the classifier scores for each variant.

Step 1: SNP annotation

The input data format is a four-column bed file containing the tab-delimited columns: chromosome; start coordinate; end coordinate; and a unique identification number. For example:

chr1 123455 123456 rs123

Variants in the input were first sorted using the command format:

```
sort -k1,1 -k2,2n ClinVAR_path_non_path.bed -o ClinVAR_path_non_path.sorted.bed
```

The sorted data was run through the annotation script using the command format '*python annotation_script sorted_variant_file annotated_variant_file'*:

```
python gwava_annotate.py ClinVAR_path_non_path.sorted.bed annotated_ClinVAR_path_non_path.sorted.bed.csv
```

Step 2: SNP classification

The annotated variant file was then run through the classifier script using the command format *'python GWAVA_classifier model_type annotated_variants scored_variants'*, model types being 'unmatched', 'tss', or 'region'. E.g.

```
python gwava.py tss annotated_variants.csv variant_scores.bed
```

4.2.2 Running FunSeq

FunSeq is a PERL- and Linux/UNIX-based tool that is available either as a web tool or a downloadable command line program. The command line version of FunSeq requires the files listed in Figure 4.5 and has the following dependencies:

- *Bedtools*

- *Tabix*

- *VAT (snpMapper Module)*

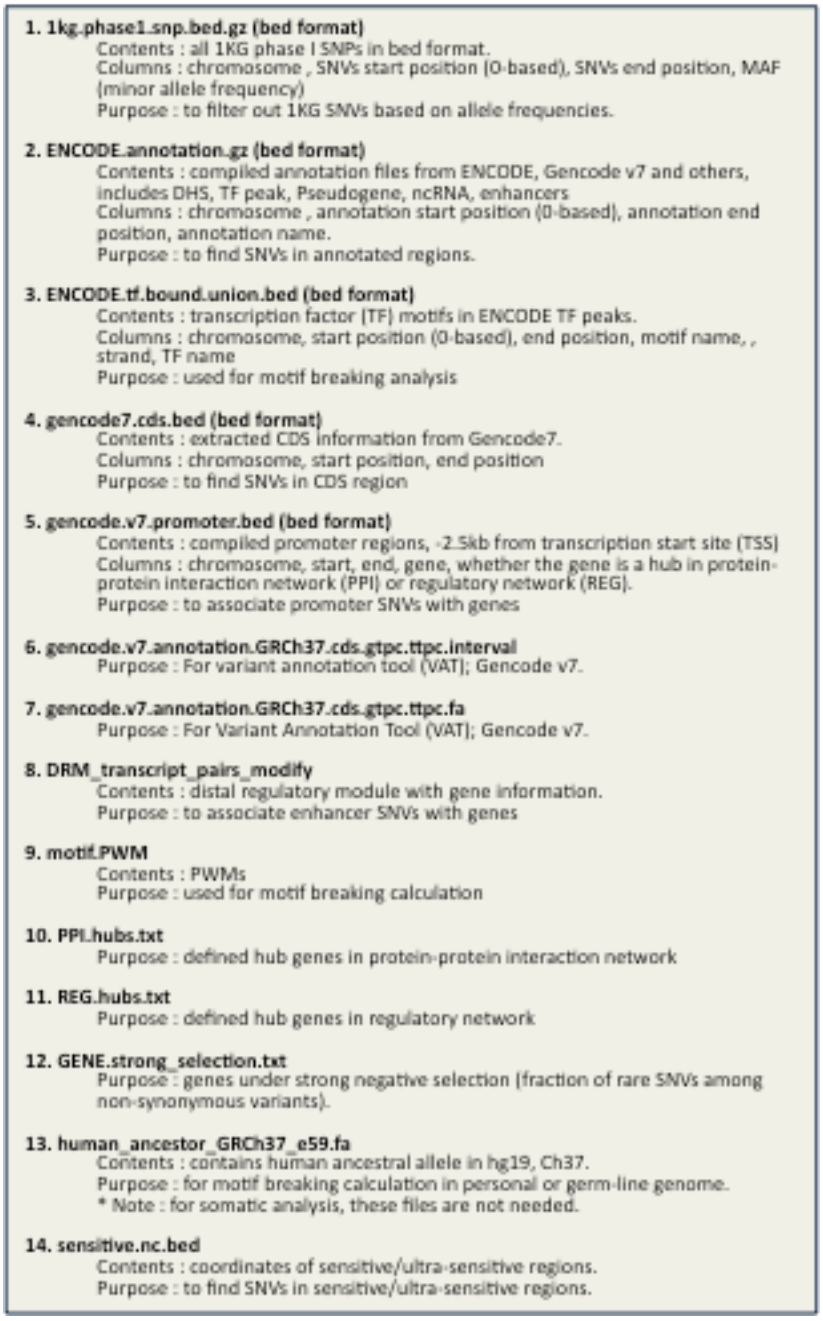
- *Perl 5 or higher*

- *Perl package Parallel::ForkManager*

SM installed FunSeq and its dependencies on the server Wheeljack. The input for FunSeq is a bed file containing the following tab-delimited columns: chromosome; start coordinate; end coordinate; reference allele; and alternative allele. The general usage commands for FunSeq are shown in Figure 4.6

For my analysis, FunSeq was run using the *–m 2* option (germline mutation).

I ran a test comparing different MAF thresholds (0-1 in 0.1 steps) and found '0.1' to be the best compromise between specificity and sensitivity. I therefore set the MAF threshold to*: –maf 0.1*.



1. 1kg.phase1.snp.bed.gz (bed format)
Contents : all 1KG phase I SNPs in bed format.
Columns : chromosome , SNVs start position (0-based), SNVs end position, MAF (minor allele frequency)
Purpose : to filter out 1KG SNVs based on allele frequencies.

2. ENCODE.annotation.gz (bed format)
Contents : compiled annotation files from ENCODE, Gencode v7 and others, includes DHS, TF peak, Pseudogene, ncRNA, enhancers
Columns : chromosome , annotation start position (0-based), annotation end position, annotation name.
Purpose : to find SNVs in annotated regions.

3. ENCODE.tf.bound.union.bed (bed format)
Contents : transcription factor (TF) motifs in ENCODE TF peaks.
Columns : chromosome, start position (0-based), end position, motif name, , strand, TF name
Purpose : used for motif breaking analysis

4. gencode7.cds.bed (bed format)
Contents : extracted CDS information from Gencode7.
Columns : chromosome, start position, end position
Purpose : to find SNVs in CDS region

5. gencode.v7.promoter.bed (bed format)
Contents : compiled promoter regions, -2.5kb from transcription start site (TSS)
Columns : chromosome, start, end, gene, whether the gene is a hub in protein-protein interaction network (PPI) or regulatory network (REG).
Purpose : to associate promoter SNVs with genes

6. gencode.v7.annotation.GRCh37.cds.gtpc.ttpc.interval
Purpose : For variant annotation tool (VAT); Gencode v7.

7. gencode.v7.annotation.GRCh37.cds.gtpc.ttpc.fa
Purpose : For Variant Annotation Tool (VAT); Gencode v7.

8. DRM_transcript_pairs_modify
Contents : distal regulatory module with gene information.
Purpose : to associate enhancer SNVs with genes

9. motif.PWM
Contents : PWMs
Purpose : used for motif breaking calculation

10. PPI.hubs.txt
Purpose : defined hub genes in protein-protein interaction network

11. REG.hubs.txt
Purpose : defined hub genes in regulatory network

12. GENE.strong_selection.txt
Purpose : genes under strong negative selection (fraction of rare SNVs among non-synonymous variants).

13. human_ancestor_GRCh37_e59.fa
Contents : contains human ancestral allele in hg19, Ch37.
Purpose : for motif breaking calculation in personal or germ-line genome.
* Note : for somatic analysis, these files are not needed.

14. sensitive.nc.bed
Contents : coordinates of sensitive/ultra-sensitive regions.
Purpose : to find SNVs in sensitive/ultra-sensitive regions.

Figure 4.5 Required data files for FunSeq, taken from the Funseq manual web page: *http://info.gersteinlab.org/FunSeq*

```
Usage :      ./funseq -f file -maf maf -m <1/2> -inf <bed/vcf> -outf <bed/vcf> -nc
Options :    -f user input SNVs file
             -maf Minor Allele Frequency (MAF) threshold to filter 1KG phase1 SNVs (value 0 ~ 1)
             -m 1 - somatic Genome; 2 - germline or personal Genome
             -inf input format - BED or VCF
             -outf output format - BED or VCF
             -nc [Optional] Only do non-coding analysis.
```

Figure 4.6 Usage commands for FunSeq. Taken from the Funseq manual web page: *http://info.gersteinlab.org/FunSeq*

4.2.3 Running CADD

CADD is available as a web tool. The input data must be in the form of the first five rows of a VCF file without a header row (e.g. chromosome; coordinate (+1); ID; reference allele; and alterative allele).

4.2.4 ClinVar datasets

I made use of multiple datasets from ClinVar in this analysis: a pathogenic dataset, a non-pathogenic dataset and a non-coding pathogenic dataset. The pathogenic dataset and non-pathogenic dataset were both downloaded from the GWAVA support website (ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/annotated/ accessed March 2014). These two datasets consisted of a true positive set of 194 pathogenic variants and a background set of 150 non-pathogenic variants. However, the pathogenic dataset contained a large number of mitochondrial variants. Removing the mitochondrial variants reduced this dataset to 128 pathogenic variants. An additional 58 non-exonic, non-coding SNPs were obtained directly from the ClinVar database.

4.2.5 1000 Genomes background variants

A dataset of 19,400 1000 Genomes background variants matched (by distance to the nearest TSS) to the pathogenic ClinVar variants was downloaded from the GWAVA support website. SM wrote a Perl script to allow us to generate additional background

datasets consisting of 1000 Genomes EUR variants matched by distance to the nearest TSS for each positive variant dataset. The function of this script was to bin all the variants from the 1000 Genomes EUR database into bins of different distances to the nearest TSS. Using this program he was able to randomly sample the 1000 Genomes database for SNPs matched to a list of distances to the nearest TSS, pulling out as many SNPs at each distance as required for each analysis. SM used this method to construct a matched background set of variants for the ClinVar non-coding dataset. This dataset contains 5,800 variants. He repeated this for the RAVEN regulatory dataset, producing a matched background dataset of 9,500 variants.

4.2.6 Complex trait related datasets

The SNPs from both the *SORT1* analysis and *TCF7L2* analysis were used by Ritchie et al. to test the performance of GWAVA and so were available from the GWAVA support website. I constructed annotation tables for both of these SNP sets; the *SORT1* annotation table contained 22 variants and the *TCF7L2* annotation table containing six SNPs.

I ran the lead SNP (rs10761670) from the *EGR2* analysis performed by Myouzen et al. (2010) through the online tool SNAP (https://www.broadinstitute.org/mpg/snap/ldsearch.php accessed May 2014) to identify all the SNPs proxy to this tagging SNP ($R^2 \geq 1.0$). This program returned a list of 35 proxy SNPs (including the tagging SNP). I constructed an annotation table for these 35 SNPs, using the 1000 Genomes Asian (ASN) population to define MAFs. In addition, I also generated a SNP dataset for the 80kb region surrounding the *EGR2* gene. This larger dataset contained all the SNPs from this region present in the 1000 Genomes ASN database (n = 237).

4.3 Results

I compared SuRFR's ability to prioritise known pathogenic variants against three additional variant prioritisation approaches: FunSeq, CADD and GWAVA. Independent data that had not been used for the training of any of these methods was required to compare their performances. This restricted the data sources at my disposal, as the tools were trained on different datasets. Despite this, I was able to identify data not used to train any of these tools (the ClinVar datasets) and used this to compare all of the methods against each other.

4.3.1 Performance of SuRFR versus GWAVA, CADD and FunSeq

4.3.1.1 ClinVar

To compare the performances of SuRFR, GWAVA, FunSeq and CADD, I used an independent dataset of clinical variants from the ClinVar archive of disease variants (Landrum et al., 2014) (see Section 4.1). This dataset consisted of 128 pathogenic variants, extracted from the ClinVar archive by Ritchie et al. (2014) to test the generalisability of GWAVA. I had modified this dataset by removing all mitochondrial variants (reducing the number from 194 to 128 variants). The reasons for this were two-fold: firstly, SuRFR has been trained on nuclear (and not mitochondrial) variants and therefore it cannot be assumed that SuRFR can correctly prioritise functional mitochondrial variants; and secondly, SuRFR relies heavily on genomic annotations that pertain exclusively to nuclear, and not mitochondrial, variants (in particular, histone modifications). As none of the other datasets used in this analysis contained mitochondrial variants, this task did not need to be repeated.

The 128 (nuclear) pathogenic variants were compared against two background datasets: a background dataset of 150 "non-pathogenic" variants (also from the ClinVar archive) and 19,400 variants identified as part of the 1000 Genomes project, distributed across the genome and matched with the pathogenic variants for distance to the nearest TSS. As for the pathogenic variants, these background datasets were selected by Ritchie et al. for their analysis of GWAVA's generalisability.

Closer examination of the pathogenic ClinVar dataset showed that it contained several synonymous, non-synonymous and UTR exonic variants. I therefore also extracted an additional pathogenic dataset directly from the ClinVar archive, consisting of purely non-exonic, non-coding variants (58 non-exonic, non-coding, clinical variants). For this second pathogenic dataset, I generated a background dataset matched by distance to the nearest TSS, 100 times the size of the pathogenic dataset (100 background SNPs matched to each pathogenic SNP).

None of these datasets had been used to train SuRFR, GWAVA, CADD or Funseq, allowing these data to be used for rigorous comparison of tool performance. For the parameters used for each of these tools, see Section 4.2.

Pathogenic ClinVar variants:

I ran SuRFR, GWAVA, CADD and FunSeq on the 128 pathogenic variants in combination with i) the 150 non-pathogenic variants test dataset and ii) the 19,400 matched 1000 Genomes variants. On these data, SuRFR was able to discriminate the pathogenic variants above background with AUCs of 0.80 and 0.85 respectively. On the same data, AUCs of 0.71 and 0.80 were achieved by GWAVA, 0.76 and 0.83 by CADD and 0.54 and 0.48 by FunSeq (Figure 4.8 A & B). These results show that SuRFR outperforms all the other methods on these data. FunSeq's performance on both of these datasets was roughly what you would expect by chance. Based on this result, I chose not to include FunSeq in any of the downstream analyses.

Non-coding versus matched 1000 Genomes background variants:

In contrast, when the performance of SuRFR, GWAVA and CADD on the non-exonic, non-coding pathogenic dataset was compared, all three methods performed at a very similar level, with CADD just outperforming SuRFR (Figure 4.9). The AUCs measured in this analysis were 0.671 (SuRFR), 0.629 (GWAVA) and 0.692 (CADD), all much lower than for the other pathogenic ClinVar dataset (Figure 4.8 A & B).

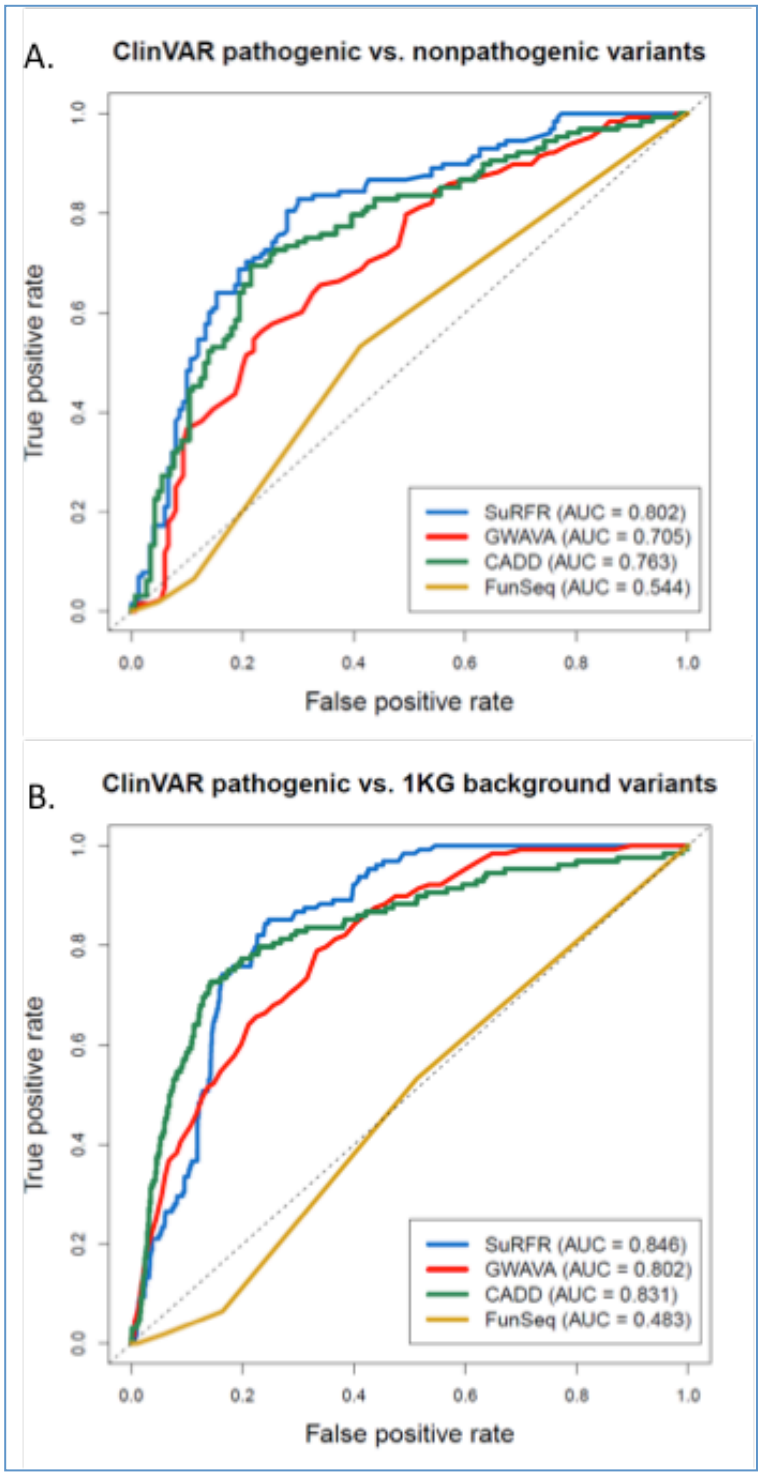


Figure 4.8. Comparison of SuRFR, GWAVA, CADD and FunSeq on A. ClinVar pathogenic vs non-pathogenic variants and B. ClinVar pathogenic vs 19,400 matched 1000 Genomes variants. This plot shows the performance if these four methods via ROC curves (true positive rate on the y-axis, versus false positive rate on the x-axis) and AUCs against the performance expected by

chance (grey dotted line). SuRFR (blue line) outperforms all three models, GWAVA (red line), CADD (green line) and FunSeq (gold line), on both of these datasets.

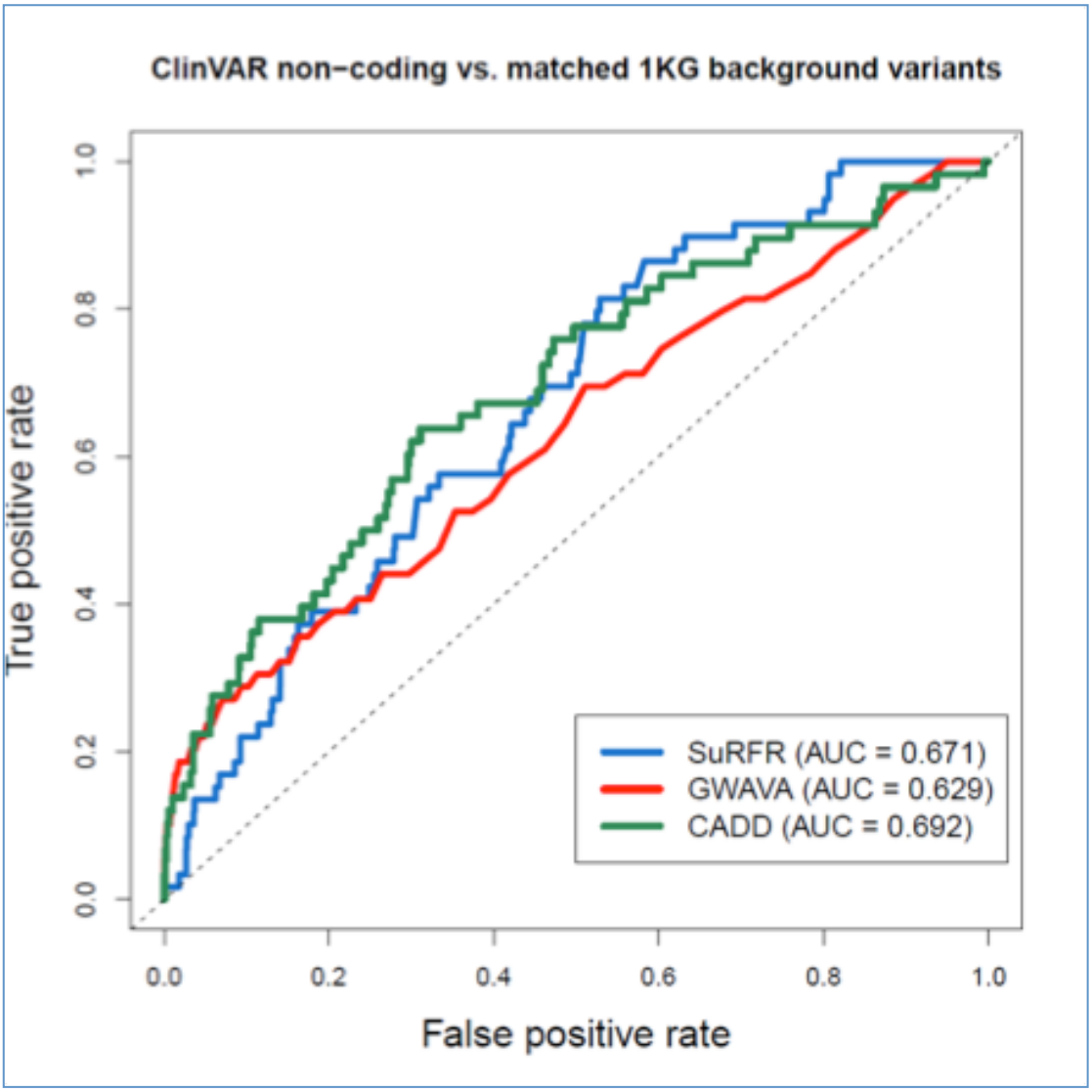


Figure 4.9 ROC curves and AUCs for SuRFR (blue), GWAVA (red) and CADD (green) run on the ClinVar non-exonic, non-coding pathogenic variants versus 5,800 matched 1000 Genomes variants.

4.3.1.2 HBB coding

Of the four tools I have compared, GWAVA is the most similar tool to SuRFR, making use of largely overlapping annotation data to prioritise non-coding variants on predicted functionality. To test and compare the ability of these two methods to prioritise pathogenic coding variants, I ran both tools on a test dataset consisting of coding disease variants for the disease ß thalassemia, located within the HBB gene. Both of these methods were extremely successful at prioritising the coding pathogenic variants above background 1000 Genomes variants, with AUCs of 0.996 and 0.975 for SuRFR (DM model) and GWAVA (TSS model) respectively (Figure 4.10).

4.3.1.3 RAVEN versus background matched by distance to the nearest TSS

As the RAVEN regulatory variants were used as part of the training and validation data for the development of GWAVA, this data could not be used to fairly compare GWAVA's performance against the other tools (as it would give an inflated estimate of GWAVA's performance). However, this data could be used to compare the performance of SuRFR and CADD. I generated a new background variant dataset consisting of 100 matched variants for every RAVEN variant. This background dataset contained 9,500 variants matched for distance to the nearest TSS. SuRFR did not perform as well on this dataset as it had done on the original RAVEN dataset (where the control SNPs were not matched by distance to the nearest TSS), achieving an AUC of 0.702 compared to the previous AUC of 0.94 (both for the DFP model). However, despite this large decrease in performance, SuRFR still performs better than CADD, which achieved an AUC of 0.608 on this data (Figure 4.11).

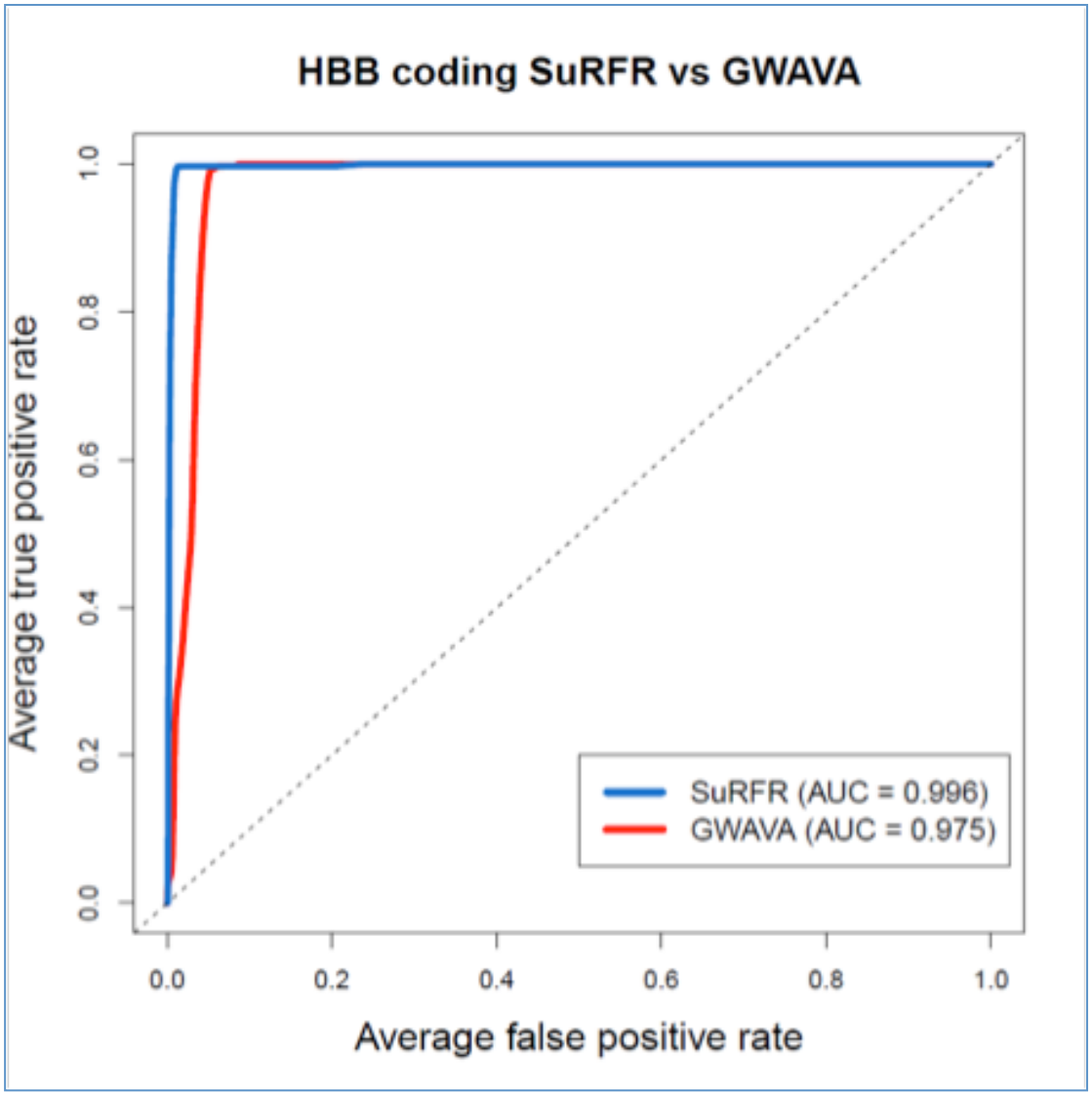


Figure 4.10 ROC curve and AUCs showing the performance of SuRFR (blue line) and GWAVA (red line) on the HBB coding variant dataset, against performance expected by chance (grey dotted line).

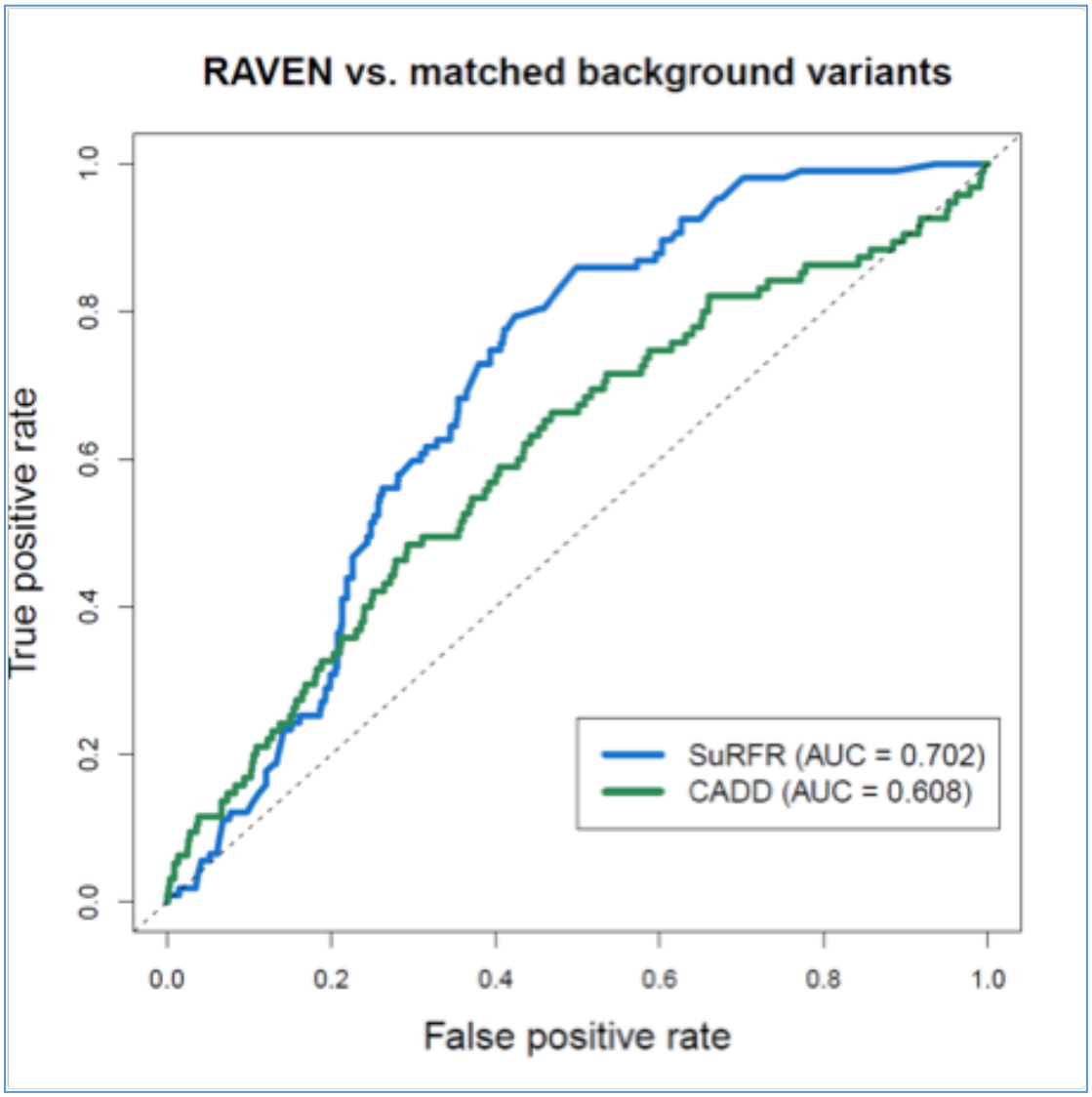


Figure 4.11 ROC curves and AUCs for SuRFR (blue line) and CADD (green line) run on the RAVEN regulatory variants versus a matched control set.

4.3.1.4 Complex trait datasets

I next compared the ability of SuRFR, GWAVA and CADD to prioritise known, functional, pathogenic variants identified by others during the study of complex traits. In each of these three examples, the projects started with an association signal or a candidate gene region. These candidate loci were then analysed in more detail. The outcome of each study was the identification of variants that altered gene expression or protein-DNA interactions, through the use of reporter assays or electrophoretic shift assays (EMSAs).

SORT1:

Musurunu et al. (2010) investigated a region of chromosome 1p13 associated with LDL-C levels. Subsequent fine mapping of this region, genotyping roughly 20,000 individuals of European descent, identified 22 SNPs within the minimal genomic region associated with LDL-C in these individuals. Of these 22 SNPs, the six with the highest association were clustered in a 6.1kb region. Functional investigation of these SNPs confirmed that one of the six, rs12740374, creates a binding site for the transcription factor C/EBP, and functional assays indicated it altered expression of the *SORT1* gene. I ran the 22 SNPs in high LD (R^2 = 1.0) with the tagging SNP (rs10761670) through SuRFR, GWAVA and CADD. SuRFR prioritised the functionally validated SNP (rs12740374) first out of 22, while GWAVA ranked it sixth and CADD ranked it twentieth (Table 4.1).

EGR2:

The chr10q21 candidate locus for SLE is roughly 80kb in size and contains 237 variants from the 1KG Asian population database with a MAF >0.10 (as of June 2014). Assessment of all 237 by GWAVA and SuRFR failed to rank the functionally validated variant, rs1509957, in the top 10% of prioritised SNPs. However, restricting the dataset to the 35 proxy SNPs in high LD (R^2 >= 1.0) with the most significantly associated SNP for SLE (Myouzen et al., 2010), the functional variant ranked very highly for both SuRFR and GWAVA (first and second respectively). In contrast, CADD ranked this variant eighteenth (Table 4.1).

TCF7L2:

The last complex trait example I used to compare the performances of SuRFR, GWAVA and CADD was the *TCF7L2* locus associated with T2D. This locus contains six variants that were functionally assessed, of which one was shown to significantly increase enhancer activity (rs7903146). Prioritisation of these six variants using SuRFR, GWAVA and CADD showed that all three tools ranked rs7903146 second out of six.

	Total number of variants	SuRFR ranking of functional variant	GWAVA ranking of functional variant	CADD ranking of functional variant
SORT1	22	1st out of 22	6th out of 22	20th out of 22
EGR2	35	1st out of 35	2nd out of 35	18th out of 35
TCF7L2	6	2nd out of 6	2nd out of 6	2nd out of 6

Table 4.1 Ranking of functionally validated variants versus background from three complex trait studies.

In conclusion, SuRFR was best able to prioritise known, functionally verified complex trait variants above background variants better than both GWAVA and CADD. In addition, this analysis showed that although CADD may be as good a method for identifying some classes of functional variants, it is not as good as either SuRFR or GWAVA at prioritising validated functional variants for complex traits.

4.4 Summary and Discussion

4.4.1 Pros and Cons of GWAVA, CADD and FunSeq

I compared SuRFR's ability to prioritise pathogenic variants against three tools (GWAVA, FunSeq and CADD) which were also designed to prioritise non-coding variants and all of which were published near the end of my PhD. Although they are all designed to perform the same task, they were developed based on different frameworks: GWAVA being written in Python and making use of a modified random forest algorithm (Ritchie et al., 2014); CADD providing a single measure (C score), based on the integration of a range of annotations, pre-computed for the entire genome (Kircher et al., 2014); and FunSeq scoring variants on the pattern of functional annotations they overlap (Khurana et al., 2013).

4.4.1.1 GWAVA

Of these three, the method that most closely resembles SuRFR is GWAVA. There are two aspects of GWAVA that make it very similar to SuRFR: the feature set incorporated into its prioritisation model; and the training methodology implemented during its development. Like SuRFR, GWAVA was designed to make use of the annotation features conservation, histone marker data, and allele frequency, amongst others. Although the black box nature of GWAVA prevents us from making a proper comparison of the features used by GWAVA versus SuRFR, or the weighting of the features within GWAVA's model, several conclusions can be drawn based on the data presented in the GWAVA paper. While not identical, the functional annotations used by these two methods are similar enough to explain why both methods perform to a similar level on several datasets: the HBB coding dataset (Figure 4.11) and the *EGR2* and *TCF7L2* complex trait datasets (Table 4.1). However, in this chapter I have also shown that SuRFR performs better than GWAVA on several other datasets, including the three ClinVar datasets (Figure 4.9 A. & B. and Figure 4.10), as well as the *SORT1* complex trait dataset. This could potentially either be due to the differences in the functional annotation used by each method (for instance, GWAVA includes GC content and SuRFR uses FANTOM CAGE data and transcribed enhancer data) or due to the way

each method combines the annotation data into a unified model (ranking versus random forest approach).

Another possible explanation for the difference in performance between the two approaches could be training methods used for each tool. GWAVA was trained and validated using a bipartite split of the training data (into training and validation folds for cross-validation). In contrast, SuRFR was trained using a tripartite data split: training, validation and hold out test datasets. Tripartite splitting of the training data allows both performance and generalisation errors to be calculated, quantifying the level of over-fitting to the training data and providing a measure of the generalisability of the method. Bipartite splitting, in contrast, does not allow the amount of over-fitting to be assessed. GWAVA could, therefore, be over-fitted to the training data, explaining why it performs worse on independent datasets.

A final explanation for the difference in performance between SuRFR and GWAVA could be the differences in training data. Although both methods make use of data from HGMD, the data used to train GWAVA is an older release of HGMD, from 2012. This data was not obtained directly from HGMD, but instead from Ensembl, and was less well annotated than the same data obtained directly from HGMD. In contrast, to train SuRFR I used a more recent, professional release of HGMD (2014), which was well annotated and sub-categorised into variant classes (DM, DFP, etc. see Chapter 3 for more details). Using this additional classification data, I was able to filter the PROM table to remove any ambiguous SNPs (removing SNPs that were not verified as disease mutations (the DM? variants) and removing SNPs that had been identified from GWAS but had not been followed up with functional validation (the DP variants)). By only using SNPs that had a proven disease or functional role, I was able to train SuRFR on a much cleaner dataset than the full HGMD regulatory dataset, potentially improving the signal to noise ratio in my data and fitting my weighting model more accurately to disease (and functional) variants.

4.4.1.2 CADD

CADD was developed using a support vector machine learning approach, trained on 14.7 million high-frequency human-derived alleles and an equal number of simulated variants (Kircher et al., 2014). This framework combines a range of 63 annotations (Supplementary Tables 1 and 2 (Kircher et al., 2014)) into a single measure, known as a C score, for each variant. This measure can be viewed as an estimate of deleteriousness and can be used to rank variants on their deleterious effect.

In my comparative analysis of CADD and SuRFR's performances, SuRFR either out-matched or performed as well as CADD on all of my test datasets. This may be because CADD has been trained to differentiate high-frequency alleles from simulated variants of equal frequencies, whereas the datasets I have used contain variants with a range of allele frequencies. This may also be due to biases in the data used to train CADD. For instance, the authors claim the best features were conservation metrics; however, the training data contains positive variants that are fixed or almost fixed in the human population (DAF > 95%), whereas the simulated variants are de novo variants (therefore unique variants) that are likely to overlap many unconstrained regions of the genome, biasing towards conservation metrics.

4.4.1.3 FunSeq

FunSeq's model framework is built on the enrichment of rare variants in particular annotation categories to estimate levels of purifying selection. Although this approach can be used for personal genomics, the authors make a point of highlighting that it is most useful and effective for cancer genomics. This is particularly apparent from my analysis of its performance on the ClinVar datasets, where it was not able to distinguish pathogenic variants from background variants any better than would be expected by chance (Figure 4.8). FunSeq's poor performance on these data is likely to be explained by the manner in which it filters SNPs. In its default mode (for the identification of cancer drivers), FunSeq filters out all variants that occur in the 1000 genomes project, as any variant that is not unique is unlikely to be a cancer driver. For the analysis of germline mutations, FunSeq instead allows the user to define a cut-off MAF; all SNPs

with MAFs above this threshold being discarded. When no threshold was used (setting the MAF cut-off to 1.0), FunSeq could not distinguish the background variants from pathogenic variants, leading to a very high false positive rate. I went on to test a range of potential MAF cut-offs (0.1 to 1.0 in 0.1 steps) and found the best cut-off (producing the best AUC) to be a MAF of 0.1. Using this cut-off, however, many variants within my test datasets were discarded (as they had MAF's greater than 0.1), increasing the false negative rate. In conclusion, although using a MAF cut-off of 0.1 maximised the specificity and sensitivity of this tool, both measures were still very low.

4.4.2 Importance of the ClinVar dataset

As each method was trained on different datasets, it was crucial to find a dataset that had not been used to train any of the models under comparison, as this would lead to an unrealistically exaggerated estimate of that method's performance (due to over-fitting). The ClinVar database, which was not used to train any of the methods, was therefore essential to the fair assessment of each model's performance. Similarly, creating multiple test datasets from this data source allowed me to draw different conclusions from the analysis.

4.4.2.1 Pathogenic versus non-pathogenic

Combining the pathogenic variants with the non-pathogenic variants from ClinVar allowed me to assess how well each method could prioritise known-functional variants against a background set of truly null variants. This was a very rare opportunity, as very few datasets of functionally assessed non-functional, non-disease causing variants exist. However, these variants might yet prove to be functional as they could have a functional role we don't yet have a test for.

The results of this analysis (Figure 4.9A.) showed that SuRFR can prioritise functional versus non-functional variants at least as well as (and marginally better than) CADD and GWAVA and all three greatly outperform FunSeq, which does not perform much better than random chance.

4.4.2.2 Pathogenic versus matched Thousand Genomes background variants

In Chapter 3 I showed that a SNP's position relative to genes (Position annotation) is the most important feature for prioritising functional variants over background variants. A background variant dataset matched to the functional variants with respect to position to the nearest TSS removes positional bias. The prioritisation of variants in such a dataset would be based on their rankings and scores for other features. As SuRFR relies heavily on positional information, such a background dataset would therefore put SuRFR at a disadvantage. In contrast, this type of background dataset is most advantageous to GWAVA, as the TSS model was trained on variants matched by distance to the nearest TSS, and therefore relies more on the other features in its model. This type of data neither positively nor negatively affected CADD's performance, as it was trained on genome-wide data, unlimited by position relative to genes

Although this dataset was in many ways more stringent than the first ClinVar dataset, the results are very similar: SuRFR performs at least as well as GWAVA and CADD and all three tools greatly outperform FunSeq. As FunSeq performed so poorly on both of these datasets I excluded it from all future comparisons.

4.4.2.3 Non-coding, matched background

As discussed in Section 4.4.2.2, background datasets matched by distance to the nearest TSS are specifically designed to be the toughest dataset for SuRFR, as they remove the advantage of the position score (SuRFR's best performing feature). Combining this style of background data with a positive dataset consisting of only non-coding (disease causing / functional) variants creates the most stringent dataset of all, as all three methods prioritise exonic and coding variants above all other variants. Therefore, this dataset truly tests how well each method can prioritise non-exonic, non-coding functional variants.

Although this dataset handicaps SuRFR compared to the other two methods, SuRFR still performs as well as CADD and GWAVA (Figure 4.10), and this result is likely to be a

lower estimate of SuRFR's performance. It is highly unlikely that any real-world scenario would exist where the causal variant (within a region of interest) is matched, by exactly the same distance to the nearest TSS, by such a large number of control variants. Therefore, in any real-world scenario we would expect the position score to have more of an effect and improve the ranking of the causal variant versus background. This observation is justified by the performance of SuRFR on the data from the three complex trait studies outlined in section 4.3.1.4.

4.4.3 Ability to prioritise coding variants: HBB coding

Although GWAVA and SuRFR were designed for the prioritisation of non-coding variants, I chose to test how well these two methods perform when prioritising coding variants in additional to non-coding variants. This analysis verified the fact that while both methods have been trained on non-coding datasets, they can both correctly prioritise coding pathogenic variants over background variants (Figure 4.11). This information is useful, as it is an important advantage over other methods to be able to prioritise coding and non-coding variants simultaneously. Currently one drawback to SuRFR is that is not capable of distinguishing between different classes of coding variants (3'UTR, 5'UTR, synonymous, and non-synonymous substitutions), as it does not make use of annotations that could prioritise these variant sub-classes.

4.4.4 Matching by distance to TSS

Rerunning SuRFR on the RAVEN functional variants and a new background dataset, matched by distance to the nearest TSS, allowed me to once again assess the lower limit of SuRFR's performance, as well as to determine how well CADD performs on the same data. Although SuRFR is at a greater disadvantage on this data than CADD, it still outperforms it (Figure 4.12, AUCs of 0.702 and 0.608 for SuRFR and CADD respectively). This suggests that SuRFR is better suited to the prioritisation of functional non-coding variants (not necessarily with a role in disease) than CADD. The difference in SuRFR's performance on this data versus the first RAVEN dataset (Figure 3.5.) once

again highlights the importance of the Position weighting to SuRFR's ability to prioritise functional variants.

4.4.5 Complex trait datasets

Developing datasets composed of large numbers of real, experimentally verified, functional and/or disease-causing variants is crucial to establishing the performance of a prioritisation approach. Such datasets provide the power required to predict with a reasonable amount of certainty how well a predictive approach will perform on novel data. However, such datasets are also highly synthetic and not representative of the numbers and types of variants likely to be present in a "real world" analysis. For instance, these datasets are enriched for large numbers of pathogenic variants that would not normally be seen in a single disease analysis. In addition, the genomic background that these variants would normally be found in has been altered, loosing the genomic context, local structural information, LD, and the array of allele frequencies, replacing them with random variants. The genomic background for disease variants could have both epistatic and polygenic effects, which would not be seen in curated databases of disease variants.

Experimental approaches to identifying the causal variant(s) for a disease or complex trait often start with either an association signal or linkage data from a family analysis. This information allows investigators to focus on a specific region of interest. Follow-up analysis such as fine-mapping, sequencing, genotyping etc., can then be used to prioritise a subset of variants for functional investigation (using luciferase assays, EMSA shift assays etc.), which can lead to the discovery of variant(s) that alter gene expression and can result in a pathogenic phenotype.

4.4.5.1 Positive attributes

The three complex trait studies used in section 4.3.1.4 provided an opportunity to test how well the three tools SuRFR, GWAVA and CADD perform on single phenotype/single locus datasets. These three different examples of complex trait analyses each represent a different type of study: i) each of these datasets represents a different complex traits (LDL_C) and diseases (SLE and diabetes); ii) the causal variants

have different functional roles (the LDL-C variant creates a binding site for the transcription factor C/EBP while the diabetes variant affects enhancer activity); and iii) the regions under investigation range in size from a 6 kb to 80 kb.

4.4.5.2 Negative attributes

These are only three examples representing the search for causal variants associated with complex disease; I have no examples of Mendelian diseases. This means I can test the performance of the ALL and DFP models of SuRFR, but do not have appropriate data to test the DM model. In addition, the regions under investigation are all less than 100 kb; therefore, the number of SNPs included in each study is limited. When the full region around the SLE locus was investigated, the number of SNPs increased from 35 to 237. When this larger SNP set was run through GWAVA and SuRFR, neither method prioritised the causal variant in the top 10% (GWAVA ranked this variant 162nd and SuRFR ranked it 118th out of 237). This suggests that this region therefore contained many variants, which, in addition to the causal variant, have functional roles identifiable from the annotation data (active enhancers, regulatory variants etc.) but not associated with SLE.

Caution should therefore be taken when searching for candidate causal variants, as many variants in the genome will also have regulatory functions and will be identifiable by the functional annotation data used to prioritise disease variants. This highlights the importance of reducing the list of candidate variants to be prioritised using any available *a priori* information, such as linkage and association signals.

4.4.6 Conclusions

In this chapter I have compared the performance of SuRFR, GWAVA, CADD and FunSeq on a variety of test datasets. I have shown that SuRFR performs at least as well as its nearest competitors, and in some instances out performs them.

The differences in performance between these four methods are likely to be due to a combination of differences in model design as well as differences in the training data used. Furthermore, differences in performance for some of these methods can be

accounted for by the fact that they are designed with a different main purpose (FunSeq is designed for the cancer genomics; CADD is a measure of deleteriousness rather than functionality).

In addition to SuRFR performing better on these data than the other three methods, SuRFR also has the advantage of being implemented as an R package and being part of the R environment. This is an advantage over CADD, which is a web-based method, as it does not limit the number of variants that can be analysed at one time (and is also better than the downloadable version of CADD which requires a large amount of free memory: 79 Gb). This is also an advantage over GWAVA, which is written in Python, as R is a statistical framework, allowing downstream analysis without exporting to another software. At every point during the running of the R package, users can understand the extent to which the various annotations contribute to the variant rankings, allowing construction of hypotheses based on the data obtained.

The most important advantage SuRFR has over these other methods is its flexibility, allowing the user to change the weighting vector used to suit their own hypotheses and also allowing additional annotation sources to be included in its framework. For these reasons, SuRFR is an excellent addition to ranks of variant prioritisation methods and I am confident it will hold its own against its competitors. Plans for future developments will be discussed in Chapter 6 (Discussion).

Chapter 5: Application of SuRFR to the study of psychiatric illness

5.1 Introduction

There has been considerable success in the search for genetic determinants of Mendelian diseases. In contrast, it has proven very difficult to identify the genetic variants contributing to complex diseases and disorders. Factors that might contribute to this are the differences in genetic architecture and the greater environmental contribution to complex disease (see Chapter 1: section 1.4). The genetics of psychiatric illness have proven particularly difficult to unravel; it is only within the last two years that the first major successes in the field of psychiatric genomics have been achieved, with the Psychiatric Genomics Consortium (PGC) identifying 108 loci associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics, 2014). Similarly, in 2014 a genome-wide association study (GWAS) combining PGC bipolar disorder samples (Psychiatric, 2011) and samples derived from the MooDS (Systematic Investigation of the Molecular Causes of Major Mood Disorders and Schizophrenia) consortium identified five loci that showed genome wide significant association with bipolar disorder (Muhleisen et al., 2014). Three of these loci had been previously reported ((Ferreira et al., 2008); (Psychiatric, 2011); (Chen et al., 2013)) and two were novel loci. The smaller number of loci identified in the bipolar disorder GWAS compared to the schizophrenia GWAS is likely to reflect, in part, the difference in sample sizes of these two GWASs (Chapter 1: Section 1.5).

Although these successes are important and pave the way for similar discoveries for other psychiatric illnesses, GWASs are only the first step in identifying variants, genes and pathways that contribute to illness. Further investigation of these findings, such as functional investigation of candidate variants (Muhleisen et al., 2014) and pathway analysis (Nurnberger et al., 2014) are needed to link association data to the underlying biology (Shinozaki and Potash, 2014).

Some disorders are unlikely ever to be aided by GWAS, mainly due to an insufficient number of affected individuals (as limited numbers means insufficient power for true

associations to reach genome-wide significance). In such cases, methods such as whole genome sequencing (WGS), family studies and linkage analysis, will be important in identifying risk variants and elucidating candidate genes. On the other hand, genomics projects (GWAS, WGS, etc.), which function on a genome-wide scale, are implicating an ever-increasing number of variants. The experimental assays currently available to characterise putative susceptibility variants are too costly and time consuming to perform on large numbers of SNPs. Bioinformatic prioritisation of candidate variants is an essential aid to the analysis of these data, as it allows experimental follow-up to be focused on those SNPs with the highest likelihood of being functional based on the currently available evidence. In this chapter, I will describe the application of SuRFR to two family-based psychiatric illness analyses, investigating the genetics of bipolar disorder and major depressive disorder.

5.1.1 Bipolar disorder

Bipolar disorder (BD) is a major psychiatric condition with a lifetime prevalence of 1% (Mülheisen et al., 2014). It is characterised by an episodic, recurrent change in mood that ranges from severe depression to elation (mania) (Craddock and Sklar, 2009). The World Health Organisation (WHO) has classified BD as one of the top ten leading causes of global disease burden for the 15-44 year old age group (Muhleisen et al., 2014). Family, twin and adoption studies have shown that there is a strong heritable component to this disorder, with studies suggesting between 60-85% of risk variance being attributable to genetic factors ((Smoller and Finn, 2003); (Nothen et al., 2010)).

The aetiology of BD is complex: multiple genetic and environmental factors contribute to disease risk ((Lichtenstein et al., 2009); (Shinozaki and Potash, 2014)). Although, traditionally, BD and other psychiatric disorders were considered to be clinically distinct, there is growing evidence for shared phenotypes across many psychiatric disorders, in particular BD, schizophrenia, schizoaffective disorder and major depressive disorder ((Lichtenstein et al., 2009); (Shinozaki and Potash, 2014)). This overlap in clinical features may reflect overlapping genetic causes ((Barnett and Smoller, 2009); (Cross-Disorder Group of the Psychiatric Genomics et al., 2013)). Similarly, it is recognised that several clinical subtypes of BD exist, including bipolar disorder type I, bipolar disorder type II, and bipolar type schizoaffective disorder (Craddock and Sklar,

2009). Taken together, these factors make studying the genetics of BD complicated. Despite the large amount of effort that has gone into linkage, sequencing and association studies of BD, there are no confirmed, functionally validated susceptibility variants for BD. In addition, the handful of loci that have been implicated in GWAS as being significantly associated with BD only explain a small proportion of the heritability. A number of next generation sequencing (NGS) projects have been undertaken to identify rare variants that contribute to disease susceptibility ((Chen et al., 2013); (Georgi et al., 2014);(Ament et al., 2015)); however, as with GWAS, additional validation and functional investigation of these data are required to confirm the role of these variants in disease aetiology. Table 5.1 outlines some of the largest BD studies that have been performed over the last decade (GWAS with ~1,000 or more cases; sequencing of over 200 individuals).

There are a number of strategies that together should result in the discovery of susceptibility variants for BD. These include analysis of larger GWAS cohorts and whole genome sequencing. A complementary strategy is to reduce heterogeneity. This can be achieved by sub-typing diagnoses ((Lee et al., 2011); (Greenwood et al., 2012)), stratifying illness by co-morbid conditions (Kerner et al., 2013) and/or using family studies (Georgi et al., 2014).

Study	Details of project design	Results
(Baum et al., 2008)	GWAS of BD in two independent case-controls sets of European ancestry, with 461 cases and 563 controls; and 772 cases and 876 controls respectively. Roughly 550,000 SNPs within genes were genotyped in these samples.	No variant reached genome-wide significance in the individual studies; however the combined study returned one significantly associated variant ($p= 1.5 \times 10^{-8}$), located in an intron of the gene *DGKH*. The authors considered this a good candidate gene, as it plays a role in the lithium sensitive phosphatidyl inositol pathway.
(Wellcome Trust Case Control, 2007)	GWAS of seven common diseases, including BD, in a British case-control set. The BD GWAS was performed on 1,868 cases and 3,000 controls, using the Affymetrix GeneChip 500K Mapping Array Set.	No genome-wide significant association was observed for BD. However, *KCNC2*, *GABRB1*, *GRM7* and *SYN3* all showed association at $P< 5 \times 10^{-7}$.
(Sklar et al., 2008)	GWAS of samples from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD) study (1,461 cases and 2,008 controls).	No variant reached genome-wide significance; however, the strongest association was seen for a variant in *MYO5B* ($P=1.66 \times 10^{-7}$). In addition, comparison of top associated SNPs from this study and the WTCCC showed that there was a concordance of signals for SNPs in the gene *CACNA1C*.
(Ferreira et al., 2008)	GWAS of BD in a dataset referred to as the ED-DUB-STEP2 dataset, consisting of cases and controls from the University of Edinburgh, Trinity College Dublin and the STEP-BD study. In total, this dataset consisted of 4,387 cases and 6,209 controls.	This analysis identified two strongly associated regions. The first, rs10994336 ($P= 9.1 \times 10^{-9}$) was in the gene *ANK3* and the second, rs1006737 ($P = 7.1 \times 10^{-8}$), was in the previously reported gene *CACNA1C*.
(Ollila et al., 2009)	GWAS replication study. The authors took the strongest associated SNPs from two GWAS studies (WTCCC, 2007; and Baum et al, 2008) and genotyped these 26 variants in a Finnish BD family cohort (723 individuals from 180 families).	Confirmed six associations: *DFNB31* (rs10982256), *SORCS2* (rs4411993, rs7683874, rs10937823), *SCL39A3* (rs4806874), and *DGKH* (rs9315885).
(Smith et al., 2009)	The authors conducted two GWAS, one on samples of European ancestry (EA: 1,001 cases and 1,033 controls) and one of African ancestry (AA: 345 cases and 670 controls).	No signal reached genome-wide significance; however the top two variants from each study (EA and AA) were: an intergenic region of Xq27 and *NAP1* (EA); and *DPY19L3* and *NTRK2* (AA).
(Wang et al., 2010)	Performed a genome-wide meta-analysis of two cohorts of combined BD and SCZ cases (653, 1172 cases and 1034, 1379 controls respectively).	Identified five loci associated with both BD and SCZ: *NAP5*, chr6q15 (near *GABRR1*), *CNTNAP2*, chr9q33.1 (near *ASTN2*) and *NALCN*.
(Howrigan et al., 2011)	Used prior findings from genome-wide linkage analysis to re-analyse the GWAS data from the STEP-BD. Using this linkage data, they implemented a weighted FDR approach.	No SNPs reached genome-wide significance.
(Psychiatric, 2011)	GWAS of 7,481 BD cases and 9,250 controls and a replication cohort of 4,496 BD cases and 42,422 controls.	Confirmed the previously identified association for CACNA1C and identified a new association with an intronic variant in ODZ4.

Study	Details of project design	Results
(Smith et al., 2011)	GWAS of BD on samples from the Bipolar Genome Study (BiGS), consisting of 2,191 cases and 1,434 controls.	No variants reached genome-wide significance; however, the authors noted that the variant rs2367911, near the gene *CACNA2D1* had suggestive association ($P=5.9 \times 10^{-6}$). This gene is related to the previously reported gene *CACNA1C*.
(Cichon et al., 2011)	GWAS of BD in a German cohort (MooDS). Discovery cohort consisted of 682 cases and 1,300 controls; total number of samples between the discovery cohort and two replication cohorts were 6,030 cases and 31,749 controls.	This study identified rs1064395 as a risk factor for BD. This variant is located in the 3'UTR of the gene *NCAN*, encoding an extracellular matrix glycoprotein, thought to play a role in migration and cell adhesion.
(Lee et al., 2011)	GWAS of a sub-set of BD (bipolar I) in a Han Chinese cohort. The discovery cohort consisted of 1,000 cases and 1,000 controls; the replication cohort contained 409 cases and 1,000 controls.	Although no genome-wide associated regions were identified, several suggestive loci were found ($P< 10^{-6}$): *SP8, ST8SIA2, KCTD12* and *CACNB2*.
(Greenwood et al., 2012)	The authors attempted to counteract the genetic heterogeneity of BD by using temperament as a quantitative trait to define subtypes of BD. Genotyping was performed on 1,263 BD cases and 1,434 controls.	Using five subscales of temperament (hyperthymic, dysthymic, cyclothymic, irritable and anxious), the authors identified three significantly associated regions: chr1 (*INTS7* gene), chr12 (*MDM1*) and chr22 (*FBLN1*).
(Kerner et al., 2013)	Exome sequencing study of individuals from a family with BD and co-morbid anxiety spectrum disorders. This study compared the exomes of three affected sisters against one unaffected brother and 200 population controls.	Exome sequencing identified very rare, heterozygous variants in eight brain expressed genes: *IQUB, JMJD1C, GADD45A, GOLGB1, PLSCR5, VRK2, MESDC2* and *FGGY*. Predicted by at least one functional predictive algorithm (out of three: SIFT, PolyPhen and Mutation Taster) to be potentially protein damaging.
(Chen et al., 2013)	Meta-analysis of a cohort consisting of cases and controls of European and Asian ancestry. This analysis combines two GWAS; the first (phase I) based on 6,658 cases and 8,187 controls, the second (phase II) tested in a sample of 484 cases and 1,823 controls. Together, this combined dataset comprised ~17,000 samples.	The discovery phase of this study identified one genome-wide significant result near *TRANK1* (rs9834970, $P= 2.4 \times 10^{-11}$). In addition, there was suggestive evidence of association for a variant near *ANK3* ($P < 10^{-6}$). These associations were replicated in the phase II data.
(Muhleisen et al., 2014)	In the largest BD GWAS to date, the authors genotyped 2.3 million SNPs in a cohort of 24,025 patients and controls. These samples were a combination of samples from the BGC-BD and MooDS consortia (9,747 cases and 14,278 controls).	This analysis identified five genome-wide significantly associated loci: *ANK3, ODZ4, TRANK1, DCY2* and an intergenic region on 6q16.1. Three of these regions had previously been reported, while the other two were novel.
(Xu et al., 2014)	GWAS performed on Canadian and UK population cohorts, consisting of 950 BD cases and 950 controls.	No genome-wide significant results. However, this study identified several suggestive associations with variants in regions previously implicated in other GWAS studies (including *SYNE1* on chr6q25, *PPP2R2C* on chr4p16.1, *ZNF659* on chr3p24.3, *CNTNAP5* on chr2q14.5 and *CDH13* on chr16q23.3).

Study	Details of project design	Results
(Georgi et al., 2014)	WGS analysis of a genetic isolate (large old order Amish pedigree) with BD. Of the 497 individuals in this family, the genomes of 50 (consisting of 18 parent-child trios) were sequenced and a further 388 family members were genotyped.	Using a combination of linkage, association and WGS analysis, this study identified five nominally significant linkage regions: chr2p25.3-p25.1, chr4p16.3, chr7q21.11-q31.33, chr16p13.3-13.12 and chr18p11.22-q13.1. Analysis of the variants located under these peaks identified several amish-specific putative damaging exonic missense variants; however, no evidence strongly implicated any one locus or a common pathway.
(Nurnberger et al., 2014)	Meta-analysis of four published GWAS to identify biological pathways that contribute to BD. 966 genes with two or more variants associated with BD ($P < 0.5$) in three of four GWAS were included in this analysis.	17 pathways were implicated in this analysis, of which 6 were associated with BD in both the initial and replication samples. These pathways included: hormonal regulation, calcium channels, second messenger systems and glutamate signalling.
(Ament et al., 2015)	Whole genome sequencing of 200 individuals from 41 families multiply affected with BD. This study focused on 3,087 genes with i. evidence of association from GWAS or ii. with know synaptic functions. Targeted sequencing of a subset of these candidate genes (26) was performed in an additional 3,014 cases and 1,717 controls.	The aim of this study was to identify uncommon and rare variants that might influence risk for bipolar disorder. This analysis focused on genes with *a priori* functional or GWAS association evidence. BD pedigrees were shown to have an increased burden of rare variants in genes and pathways that regulate neuronal excitability, particularly in ion channel genes. Most of the risk variants identified were non-coding variants predicted to have regulatory functions, suggesting an important role for the regulation of gene expression in BD.

Table 5.1. Summary of large-scale analyses of BD performed over the last decade.

5.1.2 Major depressive disorder

Major depressive disorder (MDD), also known as major depression and unipolar depression, is a debilitating psychiatric disorder, characterised by a persistent depressive mood, loss of interest or pleasure in normally enjoyable activities and changes to sleep and appetite (Verbeek et al., 2013). MDD is one of the most common psychiatric disorders, with a lifetime prevalence of ~15% (Kessler et al., 2005). In 1996 the World Health Organisation predicted MDD would be the second leading cause of disability worldwide by 2020 (after ischemic heart disease) (Murray and Lopez, 1996). Almost twenty years later, this prediction is on track; MDD is the third leading cause of disability in Europe and in the US is reported as being the greatest cause of disability of any biomedical disease (Flint and Kendler, 2014).

Family studies have shown there is a genetic component to MDD, with heritability estimated to be 0.37 (95% confidence intervals 31-42%)(Sullivan et al., 2000). Both early onset and recurrence of depression are associated with higher familial aggregation ((Wray et al., 2012); (Sullivan et al., 2000); (Kendler et al., 2005)). This disorder is twice as common in women as men ((Wray et al., 2012); (Wilhelm et al., 2003)).

To date, ten GWASs have been published (see Table 5.2). Only two of these have returned genome-wide significant associations ((Kohli et al., 2011); (consortium, 2015)). The first, by Kohli et al. (2011) identified a ~450kb region associated with MDD. Two tagging SNPs within this region were found to be associated with altered expression of *SLC6A15*. The second, by the CONVERGE consortium (2015), attempted to reduce genetic heterogeneity by focusing on women with recurrent MDD, of Han Chinese ancestry (all four grandparents were Han Chinese). This study used low-coverage whole genome sequencing of 5,303 cases and 5,337 controls (also Han Chinese women). After quality control, the SNP set for GWAS consisted of 6,242,619 SNPs. This study identified, and later independently replicated, two genome-wide significantly associated regions for MDD. The first variant, rs12415800 ($P = 2.53 \times 10^{-10}$), is located near the *SIRT1* gene on chromosome 10; the second variant, rs126244970 ($P = 6.45 \times 10^{-12}$), is also on chromosome 10, in an intron of the *LHPP* gene.

Study	Details of project design	Results
(Sullivan et al., 2009)	GWAS of 435,291 SNPs genotyped in 1,738 MDD cases and 1,802 controls from a Dutch cohort.	No SNP reached genome-wide significance. However, of the top 200 ranked SNPs, 11 localised to a 167 kb region, which overlaps the gene *PCLO*. The protein encoded by this gene is known to be involved in neurotransmission.
(Lewis et al., 2010)	GWAS of 471,747 SNPs genotyped in a UK cohort of 1,636 MDD cases and 1,594 controls.	No genome-wide significant results were identified in this study. A SNP in *BICC1* achieved suggestive evidence of association ($P < 10^{-6}$), but this finding has not been replicated.
(Muglia et al., 2010)	The authors performed GWAS on two independent European cohorts: first 1,022 cases of MDD and 1000 controls (genotyped using the Illumina 550 platform); and second 492 MDD cases and 1052 controls (genotyped using the	Neither of the two separate GWASs, nor the meta-analysis, identified any genome-wide significant associations.

	Affymetrix 5.0 platform). These independent datasets were also studied together in a meta-analysis.	
(Rietschel et al., 2010)	GWAS of 604 patients with MDD and 1,364 controls from a German cohort.	No SNPs reached genome-wide significance. Two SNPs showed nominally significant association, one of which is located in a putative regulatory element for *HOMER1*. Evidence from animal studies and human imaging studies support the hypothesis that *HOMER1* may play a role in the aetiology of MDD through a dysregulation of cognitive and motivational processes.
(Shyn et al., 2011)	GWAS of 1,221 cases and 1,636 controls from a US cohort. The authors also conducted a meta-analysis of three European-ancestry GWAS datasets totalling 3,957 cases and 3,428 controls.	This study failed to identify any variants that reached genome-wide significance. The strongest evidence for association in this analysis was observed for three intronic SNPs in *SP4, ATP6V1B2* and *GRM7*. Prior biological evidence suggested GRM7 to be a strong candidate gene for MDD. However, this has yet to be replicated in any other GWAS study.
(Shi et al., 2011)	GWAS on a US cohort consisting of 1,020 MDD cases and 1,636 controls.	No genome-wide significant results were identified in this study. The strongest evidence of association was observed on chr18q22.1.
(Kohli et al., 2011)	Discovery set consisted of 353 MDD cases and 366 controls from a clinic in Munich Germany. The replication set consisted of 3,738 cases and 10,635 controls from six independent cohorts of German, Dutch, UK and African American origin.	This study identified a single genome-wide significant association with a variant on chr12q21.31. This variant appears to be part of a haplotype containing seven additional common variants in LD with the tagging SNP, covering a region of ~450kb. This region is a gene desert, the closest gene to which is *SLC6A15* (a further 287kb distal to the associated region). Gene expression showed that two of the common variants in the associated region altered the expression of *SLC6A15* in the hippocampus.
(Wray et al., 2012)	GWAS of the MDD2000+ cohort, consisting of 2,431 cases and 3,673 controls. In addition, the authors performed a meta-analysis including two additional datasets (totalling 5,763 cases and 6,901 controls).	No SNPs in either the MDD2000+ study nor in the meta-analysis reached genome-wide significance.
(Major Depressive Disorder Working Group of the Psychiatric et al., 2013)	This is the largest GWAS for MDD to date, consisting of 9,240 cases and 9,519 controls of recent European ancestry. In addition to a large replication cohort (6,783 cases and 50,695 controls), this study stratified cases by phenotypes including sex, recurrence and age of onset amongst others.	No SNP reached genome-wide significance.
(consortium, 2015)	This study focused on a cohort of Han Chinese women, 5,303 with recurrent MDD and 5,337 controls. Low-coverage whole-genome sequencing was used to genotype the cohort; 6,242,629 SNPs were used for the GWAS.	Two genome-wide significant loci were identified: rs12415800 ($P = 2.53 \times 10^{-10}$), located near the *SIRT1* gene on chromosome 10; and rs126244970 ($P = 6.45 \times 10^{-12}$), in an intron of the LHPP gene. Neither of these significantly associated variants were replicated in a comparison against the PGC MDD GWAS data (2013).

Table 5.2 Summary of the ten MDD GWAS studies performed to date.

So far, candidate gene approaches have analysed almost 200 genes in the search for genes and pathways that might function in the aetiology of MDD. So far, these studies have had limited success, and many groups working on the same gene report contrary and conflicting findings. A meta-analysis of 26 genes yielded significant results for 7 genes (Flint and Kendler, 2014). However, the mean effect size across these studies was shown to be 1.35, and the variants tested were shown to be common. These two facts together suggest that these associations, if real, should have been identified by at least one of the ten MDD GWASs published to date. As they have not, it is possible that these are false positive findings (Flint and Kendler, 2014). However, increased genetic heterogeneity of GWAS samples might have masked the association with a subtype of illness. Similarly, rare variants or variants with a lower effect size, which would also be missed by GWAS, may still contribute to MDD (Flint and Kendler, 2014).

Although GWASs have yet to identify replicated variants associated with MDD, we have still learnt something from these analyses, as the lack of results provide clues to the genetic architecture of MDD:

1. Large numbers of common variants of small effect sizes (odds ratios of less than 1.2) could account for a large portion of the genetics of MDD. If this is the case, increased sample sizes will be needed to identify these variants, as increasing the number of cases of MDD used for GWAS will improve the power to detect common variants of small effect sizes (<1.2) (Levinson et al., 2014).

2. The limited success of GWAS could also point to the role of individually rare variants with higher effect sizes in causing complex traits such as MDD (McCellan and King, 2010). Although each of these variants might only occur in a small subset of cases, collectively they could contribute a significant portion of the genetics of MDD. These rare variants are unlikely to be indentified by current GWAS approaches (Flint and Kendler, 2014).

3. The environmental component for MDD is quite substantial. Sullivan et al. (2000) showed that variance in liability to MD is mostly due to individual specific environmental effects (95% confidence interval 58%-67%)(Sullivan et al., 2000). Focusing analyses on cases of MDD with more homogenous

environmental backgrounds could be a way of reducing heterogeneity in the data; for instance, studying women with perinatal and post-partum MDD (Wray et al., 2012).

4. Genetic heterogeneity, coupled with phenotypic homogeneity, could reduce the power of association studies. In their 2014 paper, Flint and Kendler (Flint and Kendler, 2014) described the following scenario: if two unrelated pathways lead to MDD and 50 variants contribute to disease aetiology through one pathway and another 50 contribute through the second pathway (both sub-types of MDD presenting with the same phenotypes), the power to detect either pathway is reduced by half. They suggested that without prior knowledge of these two pathways, the results of such an analysis would be difficult to decipher.

Sub-setting cases based on additional phenotypes (for instance, by sex, co-morbid psychiatric traits, severity of symptoms, combination of symptoms, biomarkers (e.g., MRI data), early-onset, or other co-morbid illnesses) could provide more homogenous datasets. Similarly, family studies of psychiatric illness provide a level of genetic homogeneity, as cases share genetic factors contributing to disease susceptibility.

5.1.3 Collaborative efforts with Cold Spring Harbour Laboratories

Our group has been working in collaboration with Prof. Dick McCombie's group at Cold Spring Harbour Laboratories (CSHLs) to investigate the genetic causes of several psychiatric illnesses using family studies. I have worked on two of these projects as part of my PhD project: the Scottish BD family project; and a second Scottish family, presenting with both MDD and idiopathic oedema.

5.1.4 The Scottish bipolar family project

5.1.4.1 Background

In 1996 a linkage analysis was performed on twelve Scottish mood disorder families (Blackwood et al., 1996). One of these families, from now on referred to as "SBF2" (Figure 5.1), with multiple cases of BD and MDD, generated a significant two-point LOD score (LOD >= 3.3; (Lander and Kruglyak, 1995)) of 4.1 (at recombination fraction $\theta = 0$) with the marker D4S394 on chromosome 4p, under the narrow diagnostic model (BD cases only; MDD coded as 'unknown'). Little change was seen in the linkage to this marker when the broad model (BD and MDD) was used (LOD 3.95, $\theta = 0$), showing evidence of linkage to BD and MDD in this region (Blackwood et al., 1996). Re-analysis of this family, using an extended pedigree and additional microsatellite markers, increased the LOD score to 4.41 at marker D4S394 (under the narrow diagnostic model) (Le Hellard et al., 2007). Additional investigation of this region using a robust variant components analysis method (Visscher et al., 1999), showed very strong evidence for a quantitative trait locus in this region affecting both bipolar disorder and MDD, achieving a maximum LOD of 5.9 and explaining roughly 25% of variance for these traits in this pedigree. This region has failed to be identified as significantly associated to BD by GWAS; however, a variant at the centromeric end of this linkage region (rs215411) was significantly associated with SCZ in a recent GWAS (Schizophrenia Working Group of the Psychiatric Genomics, 2014). In addition, there is tentative association evidence suggesting chr4p15-16 is a putative locus for susceptibility for BD ((Christoforou et al., 2007); (Baum et al., 2008); (Ollila et al., 2009)), as well as tentative linkage evidence (Georgi et al., 2014). These, together with the linkage analysis, suggest that the region may contain a rare genetic risk variant for psychiatric illness, segregating in this family under a dominant model with reduced penetrance

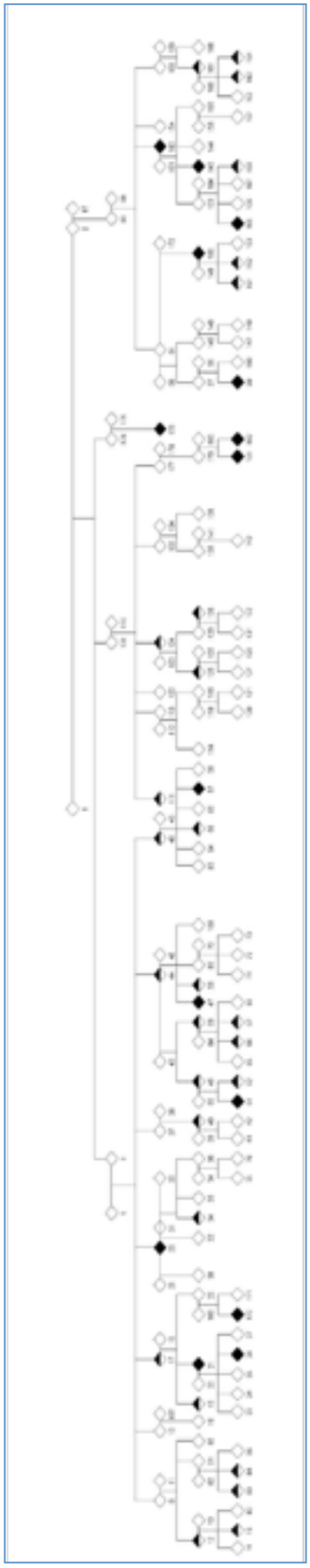


Figure 5.1. Diagram of the current pedigree of SBF2, originally examined as part of a study on bipolar families performed by Douglas Blackwood et al. (Blackwood, 1996). In this diagram, the affection status of individuals is represented by various proportions of black and white diamonds: individuals with BD are represented by fully black diamonds; individuals with unipolar disorder (rMDD) are depicted by half black/half white diamonds. The five individuals included in the whole genome sequencing analysis (IDs 17, 21, 29, 33 and 39) are all marked with red asterisks.

5.1.4.2 Collaboration with CSHLs on the SBF2 project

Based on the linkage analyses (which showed strong evidence of linkage to this region under a dominant model ((Blackwood et al., 1996); (Le Hellard et al., 2007))) and the variance components analysis (which suggested the association is unlikely to be due to a polygenic component (Visscher et al., 1999)), we hypothesised that the genetic risk variant for BD and MDD in SBF2 would be a rare variant, segregating with incomplete penetrance, located within the linkage region on chromosome 4.

Our collaborator at CSHLs, Prof. Dick McCombie, set out to generate sequencing data from the 4p locus in this family. At the time this project was initiated (2009), several sequencing options were available. Whole genome sequencing was chosen as the most effective method to achieve sufficient coverage at lowest cost.

CSHLs sequenced the whole genomes of five individuals from SBF2, three affected-carriers of the disease-linked haplotype on chromosome 4p15-16 (ID 17 (MDD), ID 21 (BD) and ID 29 (BD)) and two unaffected, married in individuals (ID 33 and ID 39). Samples were sequenced on the Illumina GAIIx platform. The mean sequencing depth across the five individuals ranged from 36x to 58x. These data were processed using the Illumina pipeline v1.5/v1.6 for base calling. Sequence alignment was performed using BWA (Li and Durbin, 2009), and the Genome Analysis ToolKit (GATK)(McKenna et al., 2010) was used to analyse the sequencing data. SNPs were filtered using the following GATK filtering thresholds:

- Filter out SNPs which have a phred-scaled Qscore >= 30
- Filter out SNPs within clusters (3 SNPs within 10bp of each other)
- Filter out SNPs which have <10X coverage
- Include SNPs which are found in chrM and chrRandom
- Include SNPs in repeat regions

This analysis focused on the SBF2 disease-linked haplotype, defined as an approximately 20Mb region on chromosome 4p (Le Hellard et al., 2007). CSHLs transferred files containing the SNPs from the linkage region as VCF files. The downstream analysis of this data is described in more detail in section 5.3.1.

Prior to my PhD project, the Evan's group had performed an initial investigation of the SNP data from the disease-linked chromosome, focusing on the coding SNPs. No unique, putatively functional, non-synonymous or missense SNPs were identified in the cases, suggesting that the causal variant might be a non-coding variant. It was therefore important to be able to functionally interpret and compare non-coding variants (regulatory variants in untranslated regions (UTRs), introns, promoters and intergenic regions; splice variants, etc). I therefore aimed to use SuRFR to prioritise these variants on the likelihood of functionality.

5.1.5 Co-morbid major depressive disorder and idiopathic oedema

5.1.5.1 Idiopathic Oedema

Idiopathic oedema (IO) is also known as cyclical oedema, periodic oedema, the fluid-retention syndrome, and, less formally, unexplained swelling (Denning et al., 1990). This disorder is characterised by intermittent swelling symptoms, often occurring at two or more sites simultaneously, including the face, hands, fingers, feet, breasts, abdomen and limbs. These symptoms also include fluid retention and an increase in body weight from the morning to the evening (diurnal weight variation). The amount of weight change considered clinically significant is still under debate. Thorn's operational criterion (Thorn, 1968), suggests a diurnal weight variation exceeding 1.4 Kg to be diagnostic for IO. This is still commonly used for the diagnosis of this disorder; however, a study of 'normal' fluid retention versus that experienced by women with IO did not find this diagnostic measure capable of discriminating between cases and controls (Denning et al., 1990). This is supported by the most recent assessment of the community prevalence of swelling symptoms (Dunnigan et al., 2004), which reported a median self-recorded daily weight gain for patients with severe IO to be 0.89 Kg. This study also found that the severity of discomfort experienced by individuals with IO is disproportionate to the amount of swelling observed by clinicians, and changes in swelling are often more obvious to those suffering from IO and their close relatives (Dunnigan et al., 2004).

In recent years, it has become more commonly accepted that symptoms of IO often form a clinical triad consisting of swelling symptoms, functional-autonomic symptoms (irritable bowel syndrome, urge frequency, and incontinence of micturition, vasomotor symptoms with pallor, faintness and syncope) and affective disturbances (anxiety and life-event stress)(Dunnigan and Pelosi, 1993). IO is also associated with obesity, diabetes and hypothyroidism (Pelosi et al., 1986). Many patients also suffer from psychological symptoms including depression; patients with IO were shown to be significantly more likely to have MDD than a cohort of female hospital outpatients (Pelosi et al., 1986). Despite increased knowledge of the symptoms and clinical manifestation, IO remains a poorly understood condition.

5.1.5.2 F224: a family multiply affected my idiopathic oedema

This condition mostly affects women, although a few cases have been reported in men (Hoffman et al., 1998). Until 1993 this disorder had not been seen in adolescents or children. At that time a study published by Dunnigan and Pelosi (1993) reported 18 cases of pre-pubescent idiopathic oedema, 15 girls and 3 boys, from 13 families. Apart from the usual triad of symptoms (swelling, autonomic and affective disturbances), all 18 children in the study by Dunnigan and Pelosi appeared healthy; laboratory tests excluded allergic, obstructive, cardio-vascular, and hypoproteinaemic causes of oedema. Treatment with drugs, such as chlorpropamide and spironolactone, ephedrine, captopril and bromocriptine produced no consistent improvement. However, all but one of the eighteen children showed a marked improvement in symptoms on administration of a carbohydrate-limited diet (120-140g carbohydrates per day). Relapses in diet were associated with a return of IO symptoms, which were also brought on and exacerbated by stressful life events (Dunnigan and Pelosi, 1993).

Five of these children were related through their mothers, four sisters, all of whom also suffered from IO, as did their mother (see pedigree for F224, Figure 5.2). These five children showed symptoms of IO as early as three months of age. The early onset of the disorder in these individuals and the family history of IO suggest this to be a case of

early-onset, familial IO. This family will be identified as F224 for the remainder of this thesis.

Figure 5.2 Pedigree of family F224, multiply affected by idiopathic oedema (IO). Individuals affected with IO are drawn in purple, unaffected individuals are coloured grey. The five affected offspring (47,48,49,50 and 51) were all included in the study of IO by Dunnigan and Pelosi, 1993.

5.1.5.3 Linkage analysis of four families with IO and MDD

In 2008, an extended pedigree of F224 was reported with multiple cases of IO and MDD (Anderson et al., 2008). Of the 28 affected individuals in this family, 18 had both MDD and IO, 7 were affected with MDD only and the remaining 3 suffered from IO only. F224, along with three additional, smaller families (F225, F226 and F364), were used in a genome-wide linkage analysis to identify regions of the genome associated with both MDD and IO. The primary aim of this analysis was to use the co-occurrence of IO and MDD in these cases to delineate a sub-phenotype, to identify regions of the genome harbouring causative variants for MDD.

Two disease definitions were used in this analysis: a narrow (individuals affected with MDD) and a broad (individuals with MDD and IO) definition. Parametric linkage and non-parametric multipoint variance component analysis were performed using 371 microsatellite markers in the four families. Four parametric linkage models were run, autosomal dominant and autosomal recessive models (with corresponding disease allele frequencies of 0.012 and 0.300 respectively) being fitted to both the narrow and broad disease definitions. Table 5.3 shows the markers with the maximum LOD scores for each model for both F224 (F224 LOD) and the cross-family linkage analysis (Total LOD).

	Narrow Dominant	Broad Dominant	Narrow Recessive	Broad Recessive
F224 LOD:	0.91	0.92	0.85	?
Total LOD:	1.73	1.55	2.2	1.2
Marker:	D14S275	D14S275	D8S260	D7S516
Chromosome:	Chr14q	Chr14q	Chr8q	Chr7q

Table 5.3. Markers with the highest LOD scores from the marker specific analysis performed by Anderson et al. (2008) on the four families with co-morbid IO and MDD.

Although this analysis failed to identify any regions of the genome significantly linked to MDD and IO (LOD >= 3.3; (Lander and Kruglyak, 1995)), several regions were identified with suggestive evidence of linkage (LOD >= 1.9; Lander and Kruglyak, 1995):

Chr8q: the marker D8S260 achieved the highest marker specific LOD across all four families. This LOD was with the narrow recessive model, suggesting this locus to be linked to MDD and not IO.

Chr14q: the marker D14S275 achieved the second highest LOD (1.73) with the narrow dominant model, as well as the highest LOD (1.55) for the broad dominant model. Both of these models are based on autosomal dominant form of inheritance, implicating a single, high penetrance variant in this region. This marker also had the highest family specific (F224) LOD (0.92) for the broad dominant model, suggesting a risk variant for both IO and MDD in this region.

Chr7q: the best result for the broad recessive model was a LOD of 1.2 with marker D7S516 on chr7q.

The linkage results for the two disease definitions used (narrow (MDD only) and broad (MDD and IO)), suggest that although there might be a common locus contributing to both disorders, MDD and IO might also be caused by variants on different loci. Specifically, chr8q is the most likely locus for a susceptibility variant for MDD, while chr7p and chr14q may contribute to both disorders.

5.1.5.4 F224 as a family case study for idiopathic oedema and depression

The relationship between these two conditions in this family is not clear, but intriguing. What is clear is that there is a high burden of IO in this family and a large number of individuals suffering from depression, many in the form of MDD. As this family appears to suffer from an early onset, familial form of IO, with more severe swelling symptoms and functional-autonomic symptoms (Dunnigan and Pelosi, 1993), these individuals might also suffer from more severe affective disturbances than more common forms of IO, which could be diagnostically similar to MDD. Therefore, the psychiatric symptoms used to diagnose MDD in these co-morbid individuals might be part of the depressive symptoms known to contribute to IO aetiology. In addition, the cases of MDD without IO in this family might be sporadic, or the IO symptoms may be very minor in nature. Based on a lifetime prevalence of ~15% for MDD (Kessler et al., 2005), the number of

expected cases in a family of 61 individuals would be ~ 9 cases; the number of cases of MDD without IO in this family (7) falls within this expected range.

MDD and IO in this family might be two separate phenotypes with a common (or overlapping) cause, or MDD in individuals with IO might be a component part of the (severe) IO phenotype. Either way, this indicates the presence of a variant or variants that can predispose to MDD. The IO and MDD phenotype in this family can therefore be used to select a phenotypic subtype of MDD in a genetically homogeneous background. It can be hypothesised that under such circumstances of reduced phenotypic and genetic heterogeneity, it might be easier to identify a functional variant than looking for common variants within a genetically and phenotypically diverse group of individuals. Furthermore, any genetic risk factors for depression identified from this study might be generalisable in the population.

5.1.5.5 Potential mechanisms of action for IO and MDD

Although the biological mechanisms underlying IO and MDD are not yet known, several theories have been put forward based on the phenotypes and aetiology of IO:

Immune response:

In a study of four cases of adult IO, serum levels of cytokines were shown to be abnormal; an increased serum concentration of SIL-2R was observed and TNF-alpha, IFN-gamma and IL-2 were found to be transiently elevated (Hoffman et al., 1993). The authors suggested that in these individuals, the formation of oedema could be a consequence of activation of T cells, resulting in the production of cytokines, and a cytokine induced alteration of the function of endothelial cells. However, they could not postulate the source of the stimulus leading to T cell activation. There is also a potential role for an activated immune system, and by extension autoimmunity, in the pathogenesis of psychiatric illnesses ((Maes et al., 2008); (Miller, 2010); (Davison, 2012)).

Abnormal neurotransmitter function:

Abnormal neurotransmitter function could cause both the functional-autonomic symptoms of IO (via the autonomic nervous system) and the affective disturbances associated with both IO and MDD (Dunnigan and Pelosi, 1993).

Insulin:

IO shares common symptoms with diabetic oedema. The high co-occurrence of a diabetic family history, obesity and weight gain with IO suggests a potential link between IO and an insulin-mediated abnormality of carbohydrate metabolism (Dunnigan and Pelosi, 1993). This is supported by the positive effect of a carbohydrate-restricted diet. Similarly, this link suggests a potential role for insulin in the aetiology of this disorder. More specifically, as insulin levels have been shown to be normal in many incidences of IO, there might be a variation in the function of insulin receptors in individuals with this condition (Dunnigan and Pelosi, 1993). In addition, acute, sub-acute and chronic diabetic oedema, are clinically similar to IO, suggesting a common cause (Dunnigan and Pelosi, 1993). This could suggest a pathogenic role of insulin, or a variation in the function of insulin receptors.

Furthermore, there is a very strong link between diabetes and MDD (Vancampfort et al., 2015b), BD (Vancampfort et al., 2015a) and schizophrenia (Foley et al., 2015).
Insulin has also been reported to multiply effect the transport of water and electrolytes (Dunnigan and Pelosi, 1993) including but not limited to stimulating the sympathetic nervous system (Landsberg and Young, 1985) and by regulating membrane ion transport, modifying calcium exchange and thereby modulating arteriolar tone (Blaustein, 1977).

Ion channels:

Changes to the structure and function of ion channels could be responsible both for the fluid retention symptoms of IO (as described by the potential effect of insulin in the previous section) and depression (ion channels being consistently implicated in psychiatric illness)((Ament et al., 2015); (Schizophrenia Working Group of the Psychiatric Genomics, 2014)).

Vasculature leakage:

There is some evidence for the role of capillary leakage, leading to an increased diffusion of fluid into the extra vascular space, in the aetiology of IO (Dunnigan et al., 2004). It has also been shown that when an individual suffers from emotional stress, neurological vasodilator pathways are activated, leading to increased blood flow to the skin and muscles (Greenfield, 1966). Variation in vasculature could provide a link between the fluid retention symptoms of IO and stress. Similarly, it has been hypothesised that increased permeability in the blood-brain barrier can lead to psychiatric illness through an inflammation response ((Maes et al., 2008); (Shalev et al., 2009)). Vascular degeneration in the brain, leading to changes in the blood brain barrier and impaired amyloid beta-peptide clearance, has also been implicated in the pathogenesis of Alzheimer's disease (Bell and Zlokovic, Acta Neuropathol, 2009 (Bell and Zlokovic, 2009)).

Stress and environmental factors:

The impact of stress on the symptoms of both IO and MDD adds to the evidence that there is a strong environmental component to both of these disorders. Genes and variants that interact with environmental factors could play important roles in the pathogenicity of both IO and MDD.

5.1.5.6 Collaboration with CSHLs on the F224 project

We have also collaborated with Prof. Dick McCombie's CSHLs group to study the co-occurrence of MDD with IO in F224 (Figure 5.3). Five individuals from this family were sequenced; a parent-offspring triad (IDs 25, 26 and 50 from the pedigree shown in Figure 5.3), the child being one of the children from the original early-onset IO analysis described in section 5.1.4, and two additional affected individuals, a mother and daughter pair (IDs 31 and 53). All four affected individuals suffer from both MDD and IO. The whole genomes of these five individuals were sequenced with an average depth of 31-40X, 90% of the genome being covered by a read depth of at least 20X. The

GATK quality control thresholds described in section 5.1.4.2 were also used on these data.

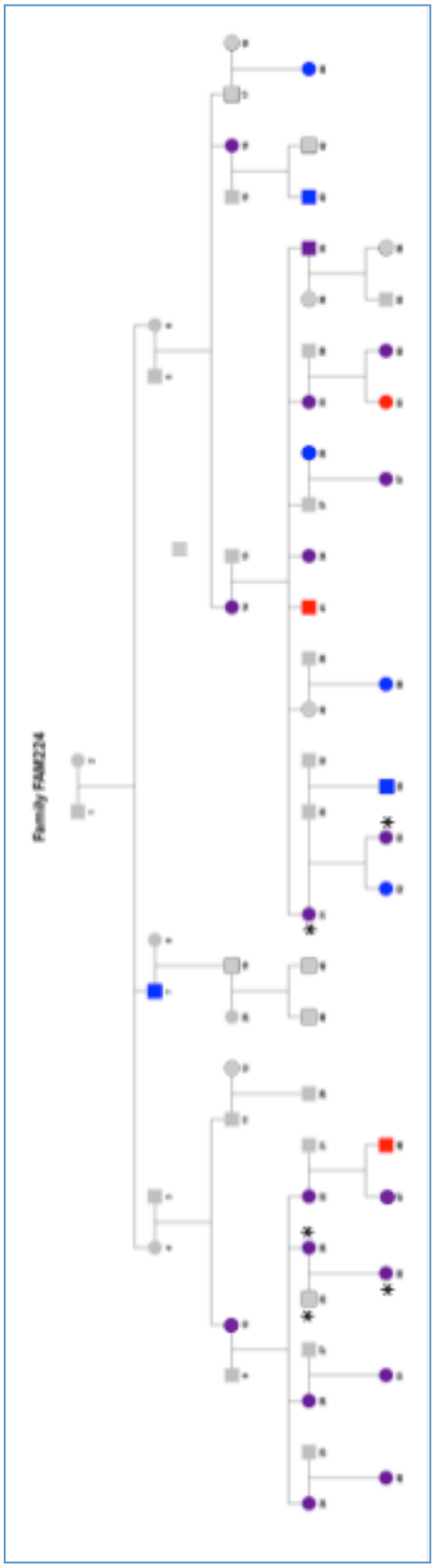


Figure 5.3 Family 224 pedigree showing all 61 individuals (28 affected and 33 unaffected). Unaffected individuals are coloured in grey; individuals with MDD are coloured in blue, individuals with IO are coloured in red; and individuals with both MDD and IO are coloured in purple. DNA from five individuals from this pedigree was sequenced (*).

After quality filtering each genome was represented by roughly ~3.4 million SNPs. These data were sent to us by CSHLs as VCF files. Stewart Morris (SM) identified all heterozygous SNPs common to the affected individuals but not present in the control individual. This filtered dataset contained 142,374 SNPs.

The aim of this project was to perform prioritisation analysis of the variants identified. A plausible model for the genetic risk variant in this family would be a rare variant of moderate effect, lying within one of the linkage regions. This is however by no means the only potential model; risk variant(s) of any minor allele frequency could exist anywhere in the genome, functioning on their own or in combination, affecting gene-gene interactions. I therefore chose a two pronged approach for this analysis: i) focusing on the three genomic loci with suggestive evidence of linkage to MDD and IO and ii) a whole genome analysis.

5.1.6 Summary of chapter aims

The aim of this chapter was to use SuRFR to prioritise variants from two projects studying psychiatric illness in families: bipolar disorder and MDD in family SBF2 and depression and MDD in the F224 family, co-occurring with idiopathic oedema.

5.2 Methods

5.2.1 SuRFR Annotation:

Using the coordinates of the Family SBF2 and F224 variant datasets (based on the GRCh37/hg19 assembly of the human genome), I annotated these variants using the annotation function, "Denovo_anno_table", from the R package SuRFR, using the following command format:

```
ann_table <- Denovo_anno_table(file, pop="EUR", threads=1)
```

I used this annotation data to rank the variants on the basis of predicted function. I also used these annotation tables to identify all of the exonic variants in the SBF2 and F224 datasets, which I extracted and saved as VCF files.

5.2.2 SuRFR prioritisation:

I ran the annotation data through SuRFR's prioritisation function, "SuRFR_analysis_tab_file", to prioritise the full variant sets for SBF2 and F224 (using SuRFR's DM model):

```
SBF2_DM <- SuRFR_analysis_tab_file(filename, DM, unique)

F224_DM <- SuRFR_analysis_tab_file(filename, DM, unique)
```

5.2.3 Ensembl's Variant Effect Predictor:

I ran the exonic datasets for SBF2 and F224 through the online version of Ensembl's Variant EffectPredictor (http://www.ensembl.org/Homo_sapiens/Tools/VEP, accessed 15/6/15) (McLaren et al., Bioinformatics, 2010). For this, I used the following options:

Assembly: GRCh37.p13
Species: Human (Homo sapiens)
Data format: Ensembl default:
chr end end disease allele/ref allele (e.g., 1 818046 818046 T/C)

5.3 Results

5.3.1 Whole genome sequencing study of SBF2

As previously described, linkage analysis performed on SBF2 identified a region of chromosome 4 that co-segregated with disease status, making this a good candidate region to search for a putative susceptibility variant for BD and MDD. The population frequency of a variant with a large effect size would be expected to be very low, as BD is rare (~1%) and no variants on chromosome 4p reach genome-wide significance in GWAS. I therefore hypothesised that the disease predisposing variant for BD and MDD in SBF2 to be either a rare or unique variant.

Our collaborators at CSHLs provided us with whole genome sequencing data for five members of SBF2, three affected (two with BD and one with MDD) and two unaffected married-in individuals. SM processed the GATK SNP file and extracted SNP position, reference base and alternative allele. He then identified the SNPs that were unique to the disease-linked chromosome (within the five individuals) by identifying those SNPs that were heterozygous in all the cases and homozygous in the two controls. The script written by SM also allowed for sequence failure of one affected and one unaffected individual at any position that would otherwise qualify as a disease-chromosome-linked SNP. In this way SM extracted all the SNPs present on the disease-linked haplotype (defined as chr4: 6,534,951 – 26,495,592).

Once the list of variants on the disease-linked chromosome was generated, SM identified the variant frequencies using data from the thousand genomes project (1KG) and HapMap. Only SNPs with a Minor Allele Frequency (MAF) below 5% in the 1KG European population were included in the next step as the goal was to search for a dominant rare variant (one that is present on the disease-linked chromosome of family SBF2 and only found rarely, if at all, in public databases). This dataset consisted of 739 variants.

I annotated the 739 variants using the annotation function of my R package SuRFR. Table 5.4 shows the distribution of these variants across the genomic position categories

used by SuRFR. This table shows that the majority of variants in this dataset are intergenic (i.e. at least 10kb away from a gene). Only nine variants were found to be located in exons or splice sites.

Score	*Genomic Feature*	*Number of variants*
0	intergenic	467
1	intron	190
2	CpG islands and CpG shores	1
3	10 kb upstream and downstream of genes	67
4	promoter	5
5	exon, splice site	9

Table 5.4 Location of variants relative to genomic features. Column 1 contains the scores used by SuRFR; column 2 describes the type of position category associated with that score; and column 3 shows the number of variants mapping to each genomic feature.

5.3.1.1 SuRFR analysis

I ranked these variants on the basis of their estimated functional potential using the "DM" model of SuRFR. Table 5.5 shows the results for the top 37 variants (top 5%) from this analysis (full dataset can be found in Appendix E). Of these 37 variants, 2 are exonic (position score 5) and two are located in putative promoter regions (1kb upstream of the TSS). In addition, 29 of these variants have not been observed in the 1KG EUR dataset, suggesting they might be unique to this family. 12 variants overlap high DNase HS peak signals (>500) and 10 overlap high TFBS peak signals. 13 variants overlap DNase footprints in at least one cell line. The SNPs in this list are within or nearby several interesting candidate genes, which are either associated with or functional candidates for BD (see Section 5.4.2).

Rank of Ranks	Gene	Position	GERP Raw	GERP Normalised Rank	DNase HS Raw	DNase HS Normalised Rank	DNase HS Footprints Raw	DNase HS Footprints Normalised Rank	Chromatin State Score	Chromatin State Normalised Rank	Position Score	Position Normalised Rank	MAF Raw	MAF Normalised Rank	Enhancers Raw	Enhancers Normalised Rank	TFBSs Raw	TFBSs Normalised Rank	DM Score
1	AK91889	chr4:13889985-13889986	1.710	0.923	1000	1.000	1	0.965	9	0.981	1	0.889	unique	1.000	0	0.001	781	1.0	39.924
2	LOC285548	chr4:13549912-13549913	0.000	0.001	1000	1.000	2	0.973	10	1.000	4	0.988	unique	1.000	0	0.001	1000	1.0	39.793
3	C1QTNF7	chr4:15422449-15422450	0.000	0.001	875	0.969	16	1.000	9	0.981	3	0.981	unique	1.000	0	0.001	397	1.0	39.337
4	C1QTNF7	chr4:15422447-15422448	0.000	0.001	875	0.969	11	0.996	9	0.981	3	0.981	unique	1.000	0	0.001	397	1.0	39.333
5	AFAP1	chr4:7865716-7865717	0.000	0.001	990	0.976	12	0.997	9	0.981	1	0.889	unique	1.000	0	0.001	254	0.9	37.839
6	FAM200B	chr4:15682717-15682718	0.000	0.001	0	0.001	0	0.001	10	1.000	4	0.988	unique	1.000	0	0.001	381	0.9	37.552
7	CNO	chr4:6727701-6727702	0.000	0.001	0	0.001	0	0.001	8	0.919	3	0.981	unique	1.000	0	0.001	456	1.0	37.004
8	MRFAP1	chr4:6642640-6642641	3.630	0.992	1000	1.000	1	0.965	10	1.000	5	1.000	0.01	0.552	0	0.001	1000	1.0	36.574
9	KCNIP4	chr4:21598318-21598319	0.000	0.001	312	0.917	0	0.001	8	0.919	1	0.889	unique	1.000	0	0.001	246	0.9	36.405
10	ABLIM2	chr4:8046394-8046395	0.086	0.609	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	541	1.0	36.210
11	GPR78	chr4:8610943-8610944	1.140	0.848	0	0.001	0	0.001	2	0.601	1	0.889	unique	1.000	0	0.001	223	0.9	35.254
12	AK091889	chr4:13934995-13934996	0.225	0.697	1000	1.000	8	0.993	8	0.919	3	0.981	0.01	0.552	0	0.001	560	1.0	35.045
18	SORCS2	chr4:7726334-7726335	0.000	0.001	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	183	0.9	34.751
18	SORCS2	chr4:7726338-7726339	0.000	0.001	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	183	0.9	34.751
18	SORCS2	chr4:7726422-7726423	0.000	0.001	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	183	0.9	34.751
18	SORCS2	chr4:7726459-7726460	0.000	0.001	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	183	0.9	34.751
18	SORCS2	chr4:7726492-7726493	0.000	0.001	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	183	0.9	34.751
18	SORCS2	chr4:7726510-7726511	0.000	0.001	0	0.001	0	0.001	7	0.811	1	0.889	unique	1.000	0	0.001	183	0.9	34.751
19	MGC4836	chr4:14608854-14608855	0.225	0.697	0	0.001	0	0.001	1	0.449	1	0.889	unique	1.000	0	0.001	217	0.9	34.008
20	KCNIP4	chr4:20848541-20848542	1.190	0.855	0	0.001	1	0.965	9	0.981	1	0.889	unique	1.000	0	0.001	0	0.0	33.905
23	GPR78	chr4:8610779-8610780	0.000	0.001	0	0.001	0	0.001	6	0.645	1	0.889	unique	1.000	0	0.001	223	0.9	33.828
23	GPR78	chr4:8610877-8610878	0.000	0.001	0	0.001	0	0.001	6	0.645	1	0.889	unique	1.000	0	0.001	223	0.9	33.828
23	GPR78	chr4:8610880-8610881	0.000	0.001	0	0.001	0	0.001	6	0.645	1	0.889	unique	1.000	0	0.001	223	0.9	33.828
24	SLIT2	chr4:20531943-20531944	0.235	0.721	0	0.001	0	0.001	8	0.919	3	0.981	unique	1.000	0	0.001	0	0.0	33.681
25	LOC285548	chr4:13549106-13549107	0.000	0.001	567	0.951	0	0.001	10	1.000	5	1.000	0.01	0.552	0	0.001	1000	1.0	33.581
26	SH3TC1	chr4:8201279-8201280	0.468	0.758	1000	1.000	0	0.001	10	1.000	1	0.889	0.01	0.552	0	0.001	1000	1.0	33.478
27	DRD5	chr4:9786380-9786381	1.900	0.936	0	0.001	0	0.001	7	0.811	3	0.981	unique	1.000	0	0.001	0	0.0	33.461
28	LOC285547,NKX3-2	chr4:13536454-13536455	0.000	0.001	345	0.926	0	0.001	10	1.000	3	0.981	0.01	0.552	0	0.001	1000	1.0	33.271
29	MGC4836	chr4:14721034-14721035	0.000	0.001	740	0.959	1	0.965	9	0.981	1	0.889	0.0013	1.000	0	0.001	0	0.0	33.156
30	C1QTNF7	chr4:15335504-15335505	1.680	0.917	766	0.962	4	0.985	8	0.919	3	0.981	0.03	0.326	0	0.001	575	1.0	32.748
31	C1QTNF7	chr4:15335509-15335510	1.570	0.905	766	0.962	4	0.985	8	0.919	3	0.981	0.03	0.326	0	0.001	575	1.0	32.724
33	LOC441009	chr4:14905419-14905420	0.000	0.001	0	0.001	1	0.965	7	0.811	3	0.981	unique	1.000	0	0.001	0	0.0	32.555
33	LOC441009	chr4:14905421-14905422	0.000	0.001	0	0.001	1	0.965	7	0.811	3	0.981	unique	1.000	0	0.001	0	0.0	32.555
34	SORCS2	chr4:7401204-7401205	0.688	0.792	0	0.001	0	0.001	8	0.919	1	0.889	unique	1.000	0	0.001	0	0.0	32.441
35	CLNK	chr4:10576246-10576247	0.361	0.742	0	0.001	0	0.001	8	0.919	1	0.889	unique	1.000	0	0.001	0	0.0	32.341
37	ABLIM2	chr4:8076567-8076568	0.000	0.001	0	0.001	0	0.001	8	0.919	3	0.981	unique	1.000	0	0.001	0	0.0	32.241
37	CC2D2A	chr4:15531017-15531018	0.000	0.001	0	0.001	0	0.001	8	0.919	3	0.981	unique	1.000	0	0.001	0	0.0	32.241

Table 5.5. Top 5% (37) ranked variants from the output of SuRFR DM model. This table shows the rank of ranks for each variant in column 1, the name of the gene associated with the variant (the gene closest to the variant) in column 2, and the coordinates of the SNP in column 3. Columns 4-19 show the raw feature data (white columns) and normalised ranks (blue columns) of each variant for each of the main feature categories (GERP, DNase HS clusters, DNase HS footprints, chromatin states, position, MAF (unique variants are those not seen in the 1KG EUR dataset, Enhancers and TFBSs). The final column shows the combined score used to rank the variants, based on the weighted ranks from each feature category. See Chapter 3 for details on the weighting models used for each version of SuRFR (i.e. the ALL, DM and DFP models).

5.3.1.2 VEP analysis of exonic and splice variants

I next analysed all the exonic and splice variants within the full set of 739 SNPs. As SuRFR has not been designed to discriminate between different types of exonic variants, I used Ensembl's Variant Effect Predictor (VEP) to analyse these variants. VEP predicts the effect of a variant on gene transcripts (McLaren et al., 2010).

The VEP output comes in two forms: i) a summary panel and pie chart (Figures 5.4 and 5.6) giving a brief overview of the VEP job (including the number of variants included, whether they overlap genes or regulatory features (transcription factor binding sites), and the proportion of types of consequences VEP predicts the variants to cause, based on the total number of transcripts affected) and ii) a table, detailing the effect of each variant on each transcript. This table contains details on the location of the genomic locations of the variants, any genes they overlap, whether they are existing variations (known from dbSNP or 1000 Genomes), what their predicted consequence is and the predicted impact of this consequence (Modifier, Low, Moderate, or High. See the description on the VEP website: http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences), and whether SIFT and PolyPhen predict them to have a deleterious effect on protein structure and function.

Figure 5.4 and Table 5.6 summarise these data. All nine variants are found in variant databases such as the 1000 Genomes project, dbSNP and the HapMap project. Two variants are coding and one lies within a splice site. Only one variant was shown to be a non-synonymous substitution, changing a glutamic acid residue (E) to a lysine residue (K) in the protein CC2D2A. VEP predicted this variant to have a moderate impact on protein function. In addition, SIFT and PolyPhen respectively predicted this variant to be deleterious and possibly damaging. The other coding variant, in MRFAP1, is a synonymous substitution, predicted by VEP to have a low likelihood of pathogenic consequence and by SIFT and PolyPhen to be benign. Variant rs3733510 overlaps a splice site; however, VEP did not predict it to have a pathogenic consequence. The remaining six exonic variants were found to overlap 5' and 3'UTRs. Re-sequencing of a

subset of the 739 variants by J.C. Yao et al. at CSHLs validated five of the nine exonic variants. The remaining four were not tested.

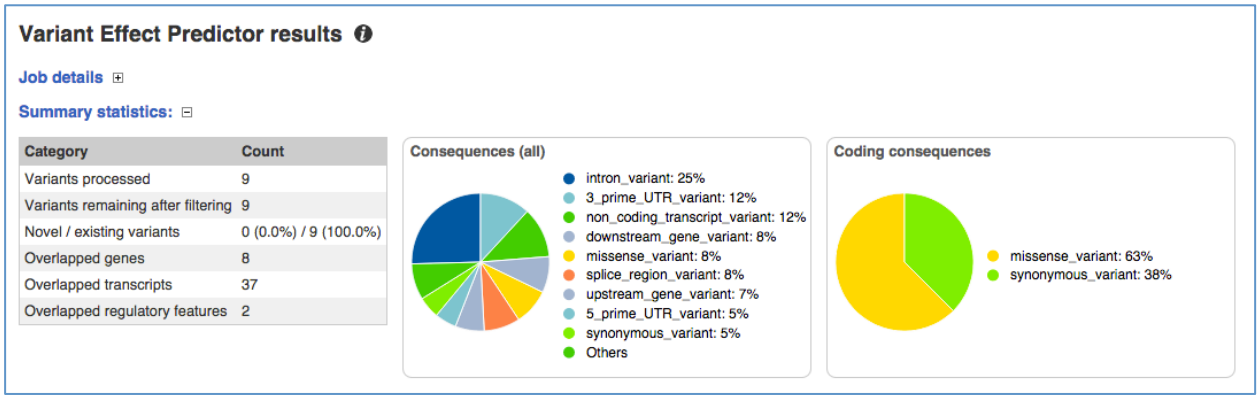


Figure 5.4 Summary statistics from VEP for the nine exonic / splice site variants from the SBF2 disease-linked haplotype. The summary table reports the number of variants included in the VEP job, how many of these were known variants, and whether any of these overlapped genes, transcripts or regulatory features (TFBSs). The two pie charts summarise the proportion of consequences for all of the transcripts these variants overlap; the pie chart on the left reported all consequences; the one on the right showing the consequences to the protein coding part of the transcripts.

Gene	Position (Hg19)	Disease allele	Reference allele	RS number	MAF	Type	Consequence	Validated by resequencing
MRFAP1	chr4:6642640-6642641	T	C	rs115295191	0.01	synonymous exonic	Low	yes
SORCS2	chr4:7742238-7742239	T	G	rs181197155	0.01	3'UTR exonic	Modifier	yes
SORCS2	chr4:7744215-7744216	C	G	rs115102964	0.01	3'UTR exonic	Modifier	not tested
AFAP1-AS1	chr4:7761604-7761605	T	C	rs141056486	0.01	3'UTR exonic	Modifier	yes
LOC285548	chr4:13549106-13549107	T	G	rs144248652	0.01	5'UTR, promoter	Modifier	not tested
CC2D2A	chr4:15474865-15474866	A	G	rs183968785	0.01	5'UTR, intronic	Modifier	yes
CC2D2A	chr4:15534867-15534868	G	A	rs144439937	0.01	missense exonic, E/K	Moderate	yes
SLIT2	chr4:20611628-20611629	A	G	rs3733510	0.04	splice	Low	not tested
GBA3	chr4:22820675-22820676	T	T	rs5015834	unique	3'UTR exonic	Modifier	not tested

Table 5.6. Summary of VEP results for the eight exonic and single splice site variant in the SBF2 dataset. This table shows the gene the variant overlaps (column 1), the genomic position of the variant (column 2), the disease and reference alleles (columns 3 and 4 respectively), the rs number associated with that genomic position (column 5), the minor allele frequency (MAF) of the variants (column 6), the main consequence type VEP predicts the variant to have across all transcripts (column 7) and the predicted impact of this consequence (Low, Modifier, Moderate, High. Column 8). In addition, I have included a description of whether this variant has been validated.

5.3.1.3 Investigation of variants in LD with the schizophrenia GWAS variant rs215411

A recent GWAS identified a variant within the chr4p15 locus as being significantly associated with schizophrenia. This variant is located in an intergenic region (the nearest gene is ~35kb away), suggesting it might function as an enhancer or other long range regulatory element. I identified the linkage disequilibrium block (D' = 0.8) around this variant using HaploView (HapMap V2, Release 24, population CEU, solid spine of LD)((Barrett et al., 2005)). This region spans the chr4 region from 23,323,427 to 23,446,949 bp. The SBF2 sequencing data (MAF <0.5) contains six variants that lie within this region (Figure 5.5). Any of these six variants could be hypothesised to function as long-range enhancers. The highest ranking of these SNPs ranked 366th out of the SBF2 SuRFR ranking data. Excluding all variants with a position score > 0 (i.e. only ranking intergenic variants), reduced this dataset to 467 intergenic variants. The highest ranking variant from the rs215411 linkage region ranked 94th out of the intergenic variants.

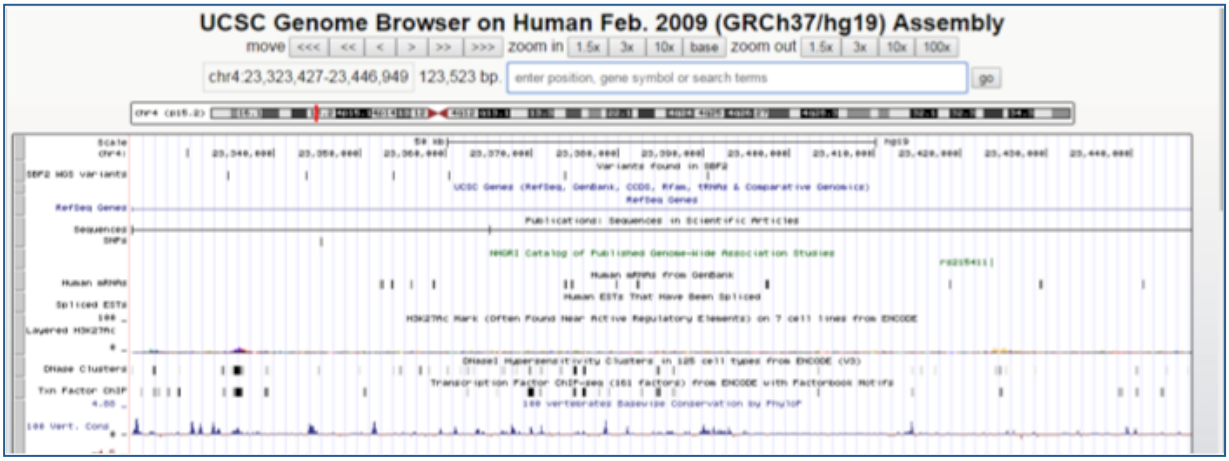


Figure 5.5. Screen shot of the UCSC genome browser showing the linkage block (D' = 0.8) around the SZC GWAS variant rs215411. The top track, "SBF2 WGS variants", is a custom track, showing the locations of the six SBF2 rare variants (MAF <0.5) that are located in this region. The GWAS variants rs215411 is coloured green and can be seen to the right hand side of the region.

5.3.2 Whole genome sequencing study of F224

Our collaborators at CSHLs generated WGS data for a family with rMDD and co-occurring IO (F224). From these data, SM extracted the heterozygous variants that were present in the four cases and absent from the married-in control. I annotated the 142,374 SNPs in this dataset using the annotation function of SuRFR. I did not use a MAF cut-off for this dataset as it was unclear what minor allele frequency to expect for the risk variant. I used the DM model to prioritise the rare variants over common variants as I hypothesised the susceptibility variant(s) to be of moderate to high effect size (best identified by the DM model), potentially uncommon or rare in the general population.

5.3.2.1 SuRFR analysis

I ranked all 142,374 variants using the DM model of SuRFR. The top 30 variants from this analysis are shown in Table 5.7, along with the nearest gene and their MAFs. In addition, I extracted subsets of this ranked data for each of the three loci implicated by linkage analysis: chr7p (chr7: 1-59,000,000), chr8q (chr8: 47,000,000-147,000,000) and chr14q (chr14: 19,000,000-68,000,000). The top 30 variants from each of these regions are shown in Table 5.9 A, B and C, along with the nearest gene and their MAF. Only one of the top 30 whole genome variants lies in a linkage region: the SNP at chr8: 82754557 ranks 7th. The highest ranked variant from the chr7q locus ranks 243rd and the best variant from chr14q ranks 56th.

DM Rank	Pos	Gene	MAF
1	chr1:17053828-17053829	TRNA_Gly	0
2	chr3:49131667-49131668	QRICH1	0
3	chr10:127585303-127585304	FANK1	0
4	chr10:127585239-127585240	FANK1	0
5	chr10:127584197-127584198	DHX32	0
6	chr15:68549703-68549704	CLN6	0
7	chr8:82754557-82754558	SNX16	0
8	chr3:75721827-75721828	LOC401074	0
9	chr1:16946433-16946434	CROCCP2	0
10	chr13:45496246-45496247	TRNA,TRNA_Glu	0
11	chr17:7742123-7742124	KDM6B	0
12	chr1:202159258-202159259	LGR6	0
13	chr7:128114467-128114468	METTL2B	0
14	chr1:17053860-17053861	ESPNP	0
15	chr20:52557957-52557958	BCAS1	0
16	chr1:16946437-16946438	CROCCP2	0
17	chr19:41724686-41724687	AXL	0
18	chr4:190942598-190942599	FRG2,LOC100288255	0
19	chr1:17230278-17230279	CROCC	0
20	chr6:160141872-160141873	WTAP	0
21	chr1:17223097-17223098	CROCC	0
22	chr1:17055001-17055002	ESPNP	0
23	chr15:52973078-52973079	FAM214A	0
24	chr1:205242768-205242769	TMCC2	0.00396
25	chr1:17230835-17230836	CROCC	0
26	chr17:8085508-8085509	MIR3676,MIR4521,TMEM107,TRNA_Ile,TRNA_Trp	0
27	chr1:17275053-17275054	CROCC	0
29	chr9:66458422-66458423	CR627148	0
29	chr9:66458427-66458428	CR627148	0
30	chr3:75708181-75708182	FRG2C	0

Table 5.7. The top 30 ranked variants from the full F224 dataset (142,374 SNPs). Column 1: the DM rank of each variant; column 2: the coordinates of the variants; column 3: the gene associated with that variant; column 4: the MAF of the variant based on the 1KG EUR dataset.

A.

DM Rank	Position	Gene	MAF
243	chr7:5971722-5971723	RSPH10B,RSPH10B2	0
311	chr7:8192283-8192284	ICA1	0
525	chr7:5971813-5971814	RSPH10B,RSPH10B2	0
603	chr7:1210069-1210070	AK090593	0
624	chr7:30953049-30953050	AQP1	0.07124
675	chr7:37393176-37393177	ELMO1	0.15699
707	chr7:33914840-33914841	AK025321	0.12797
764	chr7:44885664-44885665	H2AFV	0
891	chr7:8192408-8192409	ICA1	0
968	chr7:1923717-1923718	MAD1L1	0
983	chr7:37400026-37400027	ELMO1	0
1127	chr7:6765592-6765593	PMS2CL	0.15831
1145	chr7:27153280-27153281	HOXA3	0.11082
1189	chr7:946442-946443	ADAP1	0
1247	chr7:6764939-6764940	PMS2CL	0.15831
1270	chr7:4681720-4681721	FOXK1	0.13061
1275	chr7:2354351-2354352	SNX8	0.27968
1337	chr7:57927687-57927688	.	0
1337	chr7:57927757-57927758	.	0
1360	chr7:44297487-44297488	CAMK2B	0.06992
1393	chr7:2353164-2353165	SNX8	0
1492	chr7:50544965-50544966	DDC	0
1600	chr7:28682603-28682604	CREB5	0
1712	chr7:18826241-18826242	HDAC9	0
1719	chr7:2550887-2550888	LFNG	0.29815
1781	chr7:38217682-38217683	STARD3NL	0.2124
1815	chr7:25992509-25992510	MIR148A	0.32718
1876	chr7:33941736-33941737	BMPER	0.13456
1882	chr7:4691059-4691060	FOXK1	0.13193
1908	chr7:1927571-1927572	MAD1L1	0

B.

DM Rank	Position	Gene	MAF
7	chr8:82754557-82754558	SNX16	0
94	chr8:82754690-82754691	SNX16	0
106	chr8:82754569-82754570	SNX16	0
106	chr8:82754575-82754576	SNX16	0
246	chr8:99437569-99437570	KCNS2	0
286	chr8:145913057-145913058	ARHGAP39	0
378	chr8:52731133-52731134	PCMTD1	0
383	chr8:144614395-144614396	ZC3H3	0
422	chr8:52731139-52731140	PCMTD1	0
449	chr8:72752977-72752978	LOC100132891	0.12005
465	chr8:72752974-72752975	LOC100132891	0.12005
484	chr8:118992955-118992956	EXT1	0
608	chr8:72754341-72754342	MSC	0.17678
617	chr8:69244017-69244018	C8orf34	0
739	chr8:145002831-145002832	PLEC	0
844	chr8:145180833-145180834	KIAA1875	0
862	chr8:97657216-97657217	CPQ	0.18997
957	chr8:103135687-103135688	NCALD	0
962	chr8:141807239-141807240	PTK2	0
1215	chr8:72752881-72752882	LOC100132891	0.12401
1302	chr8:145180717-145180718	KIAA1875	0
1369	chr8:72753496-72753497	LOC100132891	0.19789
1379	chr8:72754747-72754748	MSC	0.17678
1421	chr8:86056652-86056653	LRRCC1	0
1601	chr8:143407916-143407917	TSNARE1	0
1662	chr8:52365517-52365518	PXDNL	0
1942	chr8:72757931-72757932	LOC100132891	0.11609
1942	chr8:72758068-72758069	LOC100132891	0.11609
1945	chr8:101968938-101968939	YWHAZ	0.20185
2149	chr8:134195284-134195285	WISP1	0.25462

C.

DM Rank	Position	Gene	MAF
56	chr14:24436047-24436048	DHRS4,DHRS4L2	0
291	chr14:38679813-38679814	SSTR1	0
346	chr14:24895009-24895010	KHNYN	0.05937
580	chr14:24435995-24435996	DHRS4,DHRS4L2	0
593	chr14:23526973-23526974	CDH24	0.17282
621	chr14:23527114-23527115	CDH24	0.17414
821	chr14:62025935-62025936	PRKCH	0
1444	chr14:23525877-23525878	CDH24	0.07784
1501	chr14:56045798-56045799	KTN1-AS1	0.26253
1605	chr14:31352862-31352863	COCH,LOC100506071	0.02902
1645	chr14:31629097-31629098	HECTD1	0.02111
1738	chr14:31527746-31527747	AP4S1	0
1741	chr14:55913274-55913275	TBPL2	0
2042	chr14:25081523-25081524	GZMH	0.09235
2351	chr14:35452125-35452126	SRP54	0.281
2460	chr14:19614777-19614778	P77SP	0.07652
2668	chr14:25361408-25361409	STXBP6	0
2671	chr14:24456191-24456192	DHRS4,DHRS4L1,DHRS4L2	0.22559
2890	chr14:32244009-32244010	NUBPL	0.06069
2900	chr14:23023288-23023289	AV852A1N1T,TCR-alpha, TCRA, TRA,TRA@, TRAC, TRD, hADV36S1	0
3158	chr14:55191681-55191682	SAMD4A	0.13588
3248	chr14:55879684-55879685	FBXO34	0.36675
3280	chr14:55892037-55892038	FBXO34,TBPL2	0.17282
3323	chr14:32075135-32075136	NUBPL	0
3323	chr14:32075140-32075141	NUBPL	0
3577	chr14:24455725-24455726	DHRS4,DHRS4L1,DHRS4L2	0
4022	chr14:61698768-61698769	PRKCH	0.09763
4270	chr14:38686629-38686630	SSTR1	0.23219
4333	chr14:32296678-32296679	NUBPL	0.01451
4415	chr14:57274518-57274519	OTX2	0.3496

Table 5.8. The top 30 DM ranked variants from the three F224 linkage regions: chr7q (A), chr8q (B) and chr14q (C). For each table, column 1 shows the rankings of these SNPs against the full F224 dataset, column 2 shows their position (coordinates in Hg19 format), column 3 contains the gene associated with the variant (if blank, this variant is intergenic) and column 4 contains the variant MAFs.

5.3.2.2 Analysis of exonic variants using VEP

Using the annotation data from SuRFR, I was able to identify all of the variants in the WGS dataset that overlapped exons. I extracted this subset of 2,912 variants from the full dataset and ran it through the Variant Effect Predictor (VEP) to identify any variants with deleterious effects (missense, stop loss, stop gain and frameshift mutations). The summary statistics from VEP on this dataset can be seen in Figure 5.6. Of these 2,912 variants, 180 are novel (not seen in the 1000 Genomes database or dbSNP).

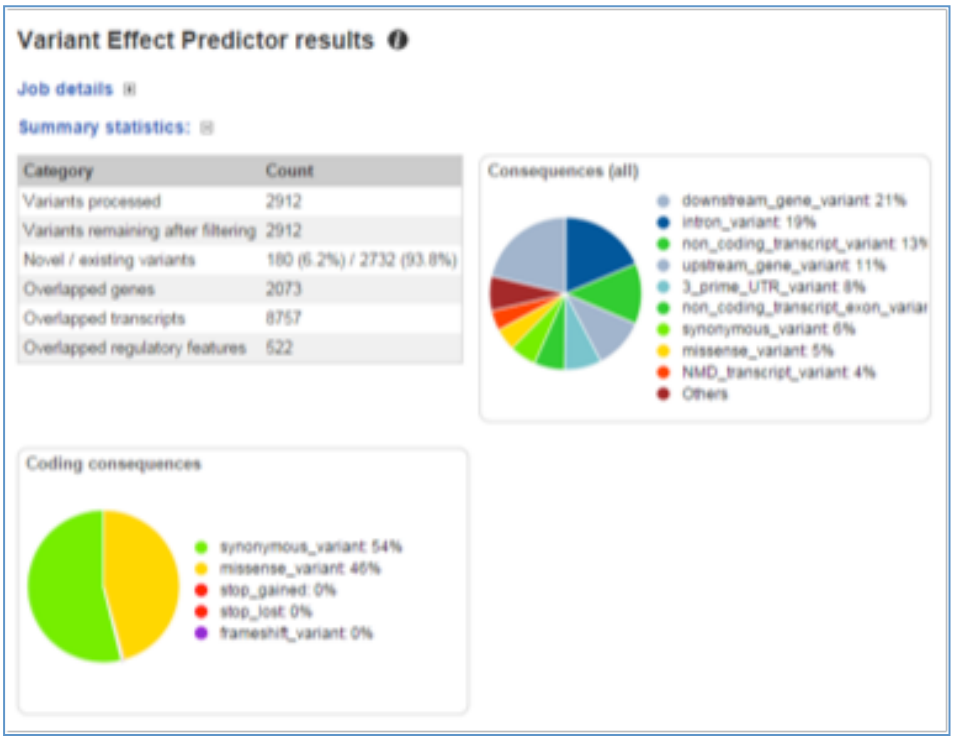


Figure 5.6. Summary statistics from VEP for the full exonic F224 dataset.

	WGS	7q	8q	14q
Total variants	142,374	3,817	4,692	1,775
exonic	2,912	163	78	81
missense	477	27	8	11
stop lost	3	0	0	0
stop gain	4	0	0	0
frameshift	2	0	0	0

Table 5.9 Breakdown of the types of deleterious exonic variants present in the full F224 dataset (across the whole genome (WGS)), and within each of the three linkage regions (chr7q, chr8q and chr14q).

Of these 2,912 variants, 477 were missense mutations, 3 were stop lost mutations, 4 were stop gain mutations and 2 were frameshift mutations (Table 5.9). None of these were located in the linkage regions. All nine of the stop-loss, stop-gain and frameshift mutations were predicted by VEP to have a high impact on protein structure (Table 5.10). Six of these variants are not found in the 1000 Genomes database or other human variation databases, suggesting they are unique to F224.

The other 2,903 variants were predicted to have either a moderate or low risk of pathogenicity. However, 141 of these variants were predicted by either SIFT or PolyPhen (or both) to be potentially deleterious, (See Appendix F for a summary of the VEP output table for these 141 variants). Of these 141 variants, 46 had a MAF <= 0.05, of which 44 had a MAF <= 0.01. Only one of the 141 variants (rs351855) was annotated by VEP as having clinical significance (highlighted in the table in Appendix F).

Using the annotation data generated by SuRFR I extracted the subset of exonic variants within the three linkage regions (chr7p, chr8q and chr14q)(Figure 5.7, Table 5.9). All 322 variants were missense variants or synonymous variants. 39 of these missense variants were predicted by VEP to have a moderate impact on protein structure and function (Table 5.11), of which 11 were predicted by SIFT and/or PolyPhen to be potentially damaging (highlighted in pink).

#Uploaded_variation	Location	Allele	Consequence	IMPACT	SYMBOL	EUR_MAF	CLIN_SIG
10_99376152_A/G	10:99376152-99376152	G	stop_lost	HIGH	MORN4	G:0.1978	-
11_55036812_T/C	11:55036812-55036812	C	stop_lost	HIGH	TRIM48	-	-
12_6948468_T/C	12:6948468-6948468	C	stop_lost	HIGH	LEPREL2	-	-
11_55861650_C/G	11:55861650-55861650	G	stop_gained	HIGH	OR8I2	G:0.0944	-
11_56310356_A/T	11:56310356-56310356	T	stop_gained	HIGH	OR5M11	T:0.0974	-
3_75786916_C/A	3:75786916-75786916	A	stop_gained	HIGH	ZNF717	-	-
3_75787127_G/T	3:75787127-75787127	T	stop_gained	HIGH	ZNF717	-	-
11_7022531_A/CT	11:7022530-7022531	CT	frameshift_variant	HIGH	ZNF214	-	-
9_14720357_C/TA	9:14720356-14720357	TA	frameshift_variant	HIGH	CER1	-	-

Table 5.10 Summary data from VEP for the nine exonic variants predicted to have high impact on protein structure and function (IMPACT column). Six of these variants have not been seen in the 1000 Genomes EUR database, suggesting they are unique to family F224.

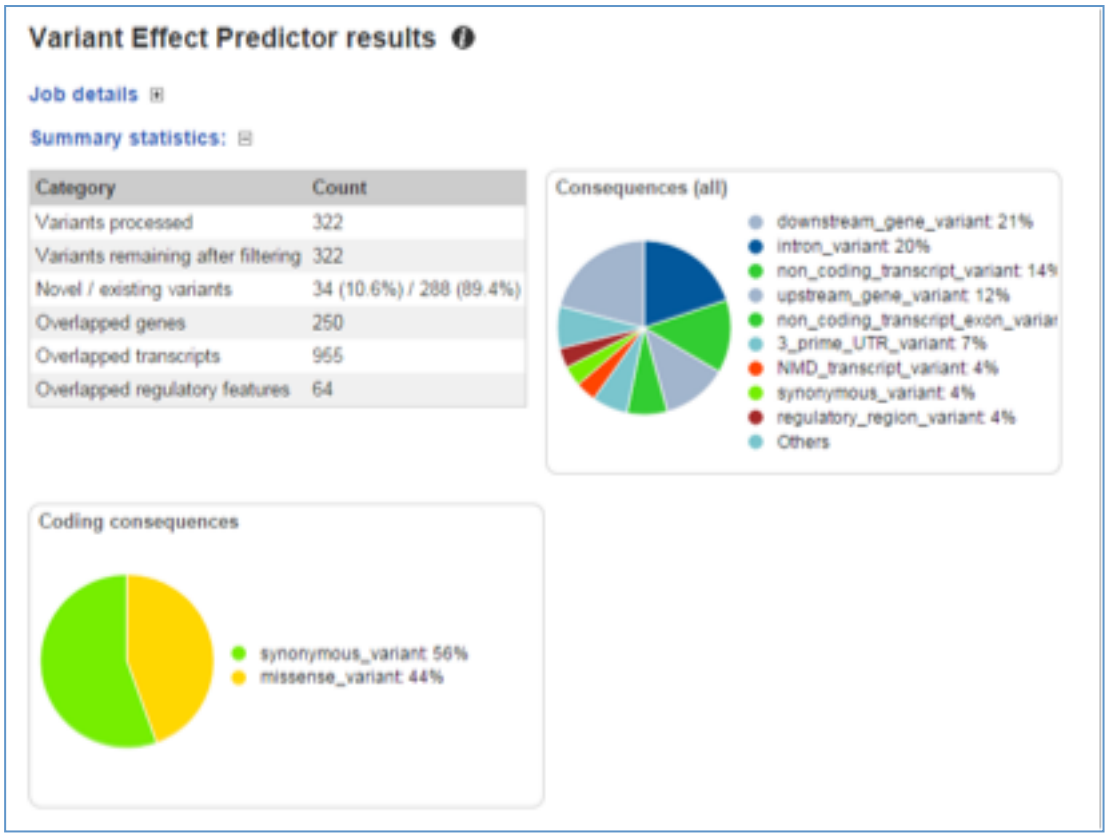


Figure 5.7 Summary statistics from VEP for the subset of exonic variants that overlap the three linkage regions (chr7q, chr8q and chr14q). All the protein coding variants were identified by VEP as being either synonymous substitutions or missense variants.

#Uploaded_variation	Location	Allele	IMPACT	SYMBOL	Amino_acids	Existing_variation	SIFT	PolyPhen	EUR_MAF
7_1586662_T/C	7:1586662-1586662	C	MODERATE	TMEM184A	S/G	rs3779607,COSM3762544	tolerated(0.5)	benign(0)	C:0.6213
7_1733102_G/T	7:1733102-1733102	T	MODERATE	AC074389.6	L/M	rs4720922	-	possibly_damaging(0.82)	T:0.2744
7_34009946_C/T	7:34009946-34009946	T	MODERATE	BMPER	A/V	rs10265207,COSM3762645	-	unknown(0)	T:0.4682
7_64439613_T/G	7:64439613-64439613	G	MODERATE	ZNF117	K/N	rs3807068,COSM3762802,COSM1091258	tolerated(0.09)	probably_damaging(0.981)	G:0.4016
7_64439701_C/T	7:64439701-64439701	T	MODERATE	ZNF117	C/Y	rs3807069,COSM3762803	deleterious(0.04)	probably_damaging(0.999)	T:0.4016
7_64452830_C/T	7:64452830-64452830	T	MODERATE	ERV3-1	C/Y	rs34639489	deleterious(0.01)	probably_damaging(0.999)	T:0.5606
7_76069811_T/C	7:76069811-76069811	C	MODERATE	ZP3	S/P	rs2906999	tolerated(0.27)	benign(0)	C:0.5209
7_82784456_A/G	7:82784456-82784456	G	MODERATE	PCLO	S/P	rs6972461	-	benign(0.032)	G:0.2704
7_100551314_C/G	7:100551314-100551314	G	MODERATE	MUC3A	A/G	rs79779043,COSM4161463,COSM4161462	-	unknown(0)	T:0
7_100551322_C/T	7:100551322-100551322	T	MODERATE	MUC3A	P/S	rs76260977,COSM4161465,COSM4161464	-	unknown(0)	-
7_100551331_G/T	7:100551331-100551331	T	MODERATE	MUC3A	V/L	rs77979106,COSM4161467,COSM4161466	-	unknown(0)	-
7_100551332_T/C	7:100551332-100551332	C	MODERATE	MUC3A	V/A	rs77022449,COSM4161469,COSM4161468	-	unknown(0)	-
7_100551334_A/T	7:100551334-100551334	T	MODERATE	MUC3A	T/S	rs76311216,COSM4161471,COSM4161470	-	unknown(0)	-
7_100551370_A/T	7:100551370-100551370	T	MODERATE	MUC3A	S/C	COSM4161473,COSM4161472	-	unknown(0)	-
7_100551371_G/C	7:100551371-100551371	C	MODERATE	MUC3A	S/T	rs75463292,COSM4161475,COSM4161474	-	unknown(0)	-
7_100551397_G/T	7:100551397-100551397	T	MODERATE	MUC3A	V/L	rs79748732,COSM1622097,COSM1622098	-	unknown(0)	-
7_100551400_G/A	7:100551400-100551400	A	MODERATE	MUC3A	G/S	COSM4161485,COSM4161484	-	unknown(0)	-
7_102389697_C/T	7:102389697-102389697	T	MODERATE	FAM185A	L/F	rs141352868,COSM3762293,COSM3762292	deleterious_low_confidence(0)	benign(0.001)	T:0.0815
7_128119514_G/A	7:128119514-128119514	A	MODERATE	METTL2B	E/K	rs1065267,COSM3762362	tolerated(0.39)	benign(0.014)	-
7_139167934_T/G	7:139167934-139167934	G	MODERATE	KLRG2	K/T	rs1860150,COSM1622418	tolerated_low_confidence(1)	benign(0)	G:0.5239
7_140158851_C/G	7:140158851-140158851	G	MODERATE	MKRN1	V/L	rs2272095	deleterious(0.02)	benign(0.303)	G:0.2893
7_142028388_A/T	7:142028388-142028388	T	MODERATE	TRBV6-1	Q/L	-	-	-	-
7_142143800_A/G	7:142143800-142143800	G	MODERATE	TRBV6-7	Y/H	-	-	-	-
7_142180647_G/T	7:142180647-142180647	T	MODERATE	TRBV6-5	A/D	-	-	-	-
7_142326607_C/T	7:142326607-142326607	T	MODERATE	TRBV19	L/F	-	-	probably_damaging(0.998)	-
7_142334648_A/T	7:142334648-142334648	T	MODERATE	TRBV20-1	R/W	-	deleterious(0.04)	benign(0.112)	-
7_142378633_G/A	7:142378633-142378633	A	MODERATE	TRBV25-1	V/M	rs17243	deleterious(0.02)	possibly_damaging(0.715)	A:0.5646
8_2046700_C/T	8:2046700-2046700	T	MODERATE	MYOM2	T/M	rs2294066	-	benign(0.149)	T:0.166
8_6873603_T/G	8:6873603-6873603	G	MODERATE	DEFA3	D/A	rs145076681,COSM1132709	tolerated(0.77)	benign(0)	-
8_11189591_C/T	8:11189591-11189591	T	MODERATE	SLC35G5	R/W	COSM3718698	tolerated(0.17)	benign(0)	-
8_30999280_G/T	8:30999280-30999280	T	MODERATE	WRN	L/F	rs1801195,COSM3763319,CM004850	tolerated(0.89)	benign(0)	T:0.4404
8_52733050_T/A	8:52733050-52733050	A	MODERATE	PCMTD1	N/I	rs12335014,COSM3982549	tolerated(0.21)	benign(0.263)	A:0.6292
8_68421768_G/C	8:68421768-68421768	C	MODERATE	CPA6	S/C	rs17853192	-	probably_damaging(0.927)	C:0.0905
8_110539186_G/A	8:110539186-110539186	A	MODERATE	PKHD1L1	V/I	rs1783174	tolerated_low_confidence(0.62)	benign(0.009)	A:0.1402
8_142489410_A/G	8:142489410-142489410	G	MODERATE	MROH5	L/S	rs2748418,COSM4162706,COSM4162705	tolerated(0.07)	benign(0.283)	G:0.3867
14_21967916_A/G	14:21967916-21967916	G	MODERATE	METTL3	S/P	rs1139130,COSM3999223	-	benign(0)	G:0.5
14_74041748_A/G	14:74041748-74041748	G	MODERATE	ACOT2	H/R	rs149033118,COSM4148229	-	benign(0.008)	-
14_102815042_C/T	14:102815042-102815042	T	MODERATE	CINP	R/H	rs7011	tolerated(0.18)	benign(0.024)	T:0.2654
14_102901201_A/G	14:102901201-102901201	G	MODERATE	TECPR2	I/V	rs10149146,COSM3999163	tolerated_low_confidence(1)	benign(0)	G:0.2982
14_106208082_G/T	14:106208082-106208082	T	MODERATE	IGHG1	L/M	-	tolerated(0.33)	possibly_damaging(0.459)	-
14_106208086_A/C	14:106208086-106208086	C	MODERATE	IGHG1	D/E	-	tolerated(1)	benign(0.003)	-
14_106209119_T/C	14:106209119-106209119	C	MODERATE	IGHG1	K/R	-	tolerated(0.5)	benign(0.004)	-
14_106235611_A/T	14:106235611-106235611	T	MODERATE	IGHG3	F/Y	-	tolerated(1)	benign(0.001)	-
14_106235767_C/T	14:106235767-106235767	T	MODERATE	IGHG3	S/N	-	tolerated(1)	benign(0.001)	-
14_106236128_T/A	14:106236128-106236128	A	MODERATE	IGHG3	Y/F	-	tolerated(1)	benign(0)	-
14_106236143_G/A	14:106236143-106236143	A	MODERATE	IGHG3	P/L	-	tolerated(0.12)	benign(0.041)	-

Table 5.11. Summary of VEP output for the 39 exonic, missense variants predicted to have a moderate impact on protein structure. Variants highlighted in pink are predicted by SIFT and/or PolyPhen to be deleterious.

5.4 Discussion

5.4.1 Summary

SuRFR was designed to aid in the analysis of variants from a range of genomics projects, including WGS data. To that effect, I have used SuRFR to prioritise variants from two WGS projects: the Scottish BD family (SBF2) project and the F224 project, a family with major depressive disorder and idiopathic oedema. The goal was to prioritise the putative functional variants for further investigation. This analysis focused on single nucleotide polymorphisms, as SuRFR has not been trained to prioritise indels or CNVs.

I have trained three different SuRFR models, each designed for specific analysis types, for prioritising causal variants for diseases with different genetic architectures. I previously showed that the ALL model is capable of prioritising known disease variants from a range of disease architectures above background variants. However, when the mode of inheritance is known to be Mendelian-like (rare variants of large effect), using the DM model improves the likelihood of correctly prioritising the causal variants above background variants. It is therefore important to correctly identify the disease model of the variants under investigation. For both SBF2 and F224 I hypothesised the susceptibility variants to be variants of large effect with medium to high penetrance, that are either rarely seen in the general population, or are unique to these families. I therefore used the DM model of SuRFR for both of these analyses. However, as the linkage results for F224 were less informative that for SBF2, I did not pre-filter the F224 dataset on MAF, allowing more common variants of high effect to also be evaluated by SuRFR.

5.4.2 SBF2

5.4.2.1 Candidate genes:

Amongst the top ranking SuRFR output variants from the ~20Mb disease-linked haplotype, there were several that were within or nearby interesting candidate genes:

C1QTNF7:

Conduct disorder (CD) is one of the most prevalent childhood psychiatric disorders, characterised by aggressive behaviour, persistent rule breaking, and associated with alcohol problems (Dick et al., 2011). The C1q and tumour necrosis factor-related protein 7, *C1QTNF7*, is an extracellular protein of unknown function. Two variants in this gene showed genome-wide significant associations ($P<5 \times 10^{-8}$) with CD in a study of 872 cases of CD and 3,091 controls (Dick et al., 2011). As CD is a psychiatric condition that can co-morbidly occur in youth with BD (Joshi and Wilens, 2009) and there is evidence of genetic overlap between psychiatric disorders (see Chapter 1: 1.4), this gene is an interesting candidate for BD.

The top two ranking *C1QTNF7* variants (ranking 3rd and 4th overall) are unique variants, not seen in the 1000 genomes database. These two SNPs are located one base pair apart and overlap both a DNase HS cluster and three TFBSs, including c-FOS (Figure 5.7). c-FOS is known to be expressed in the brain (Herrera and Robertson, 1996) and to play an important role in regulation of synaptic plasticity (Cohen and Greenberg, 2008). Variants associated with both schizophrenia risk and protection, have been identified in the *c-FOS* gene (Boyajyan et al., 2015). $GABA_B$ receptors, which have been implicated in psychiatric disorders (de Bartolomeis and Tomasetti, 2012), have also been shown to be linked to stress-induced c-FOS activation in the hippocampus (O'Leary et al., 2014).

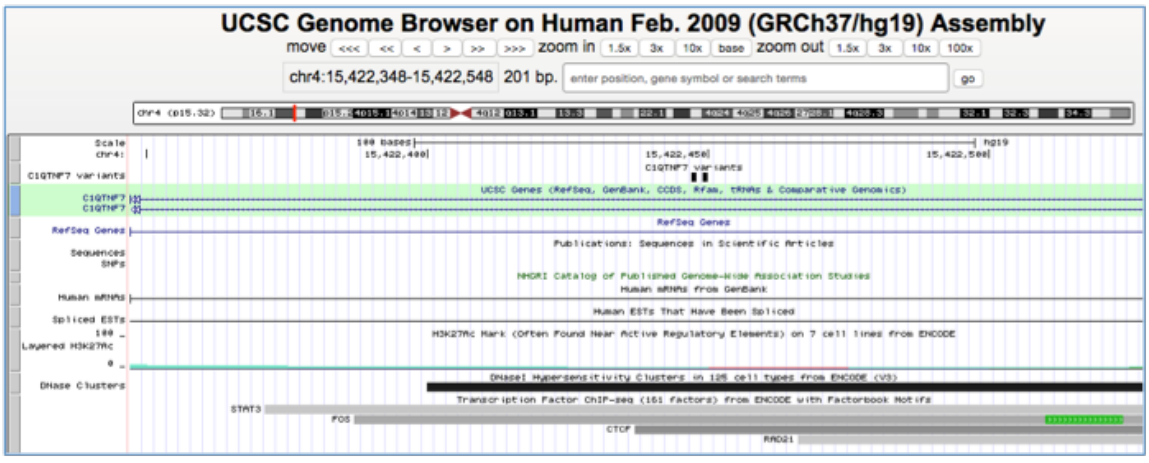


Figure 5.7 Image from the UCSC genome browser, showing the location of the two intronic C1QTNF7 variants and their overlap with a binding site for the brain-expressed transcription factor c-FOS.

Two additional *C1QTNF7* variants are in the top 5% table (Table 5.4). These variants, both with MAFs of 0.03 in the 1000 genomes EUR dataset, are located 3 base pairs apart from each other ~7kb upstream of the *C1QTNF7* transcription start site. These two variants also overlap a DNase HS cluster and several TFBSs (POLR2A, BCL11A, PAX5, YY1 and EP300).

KCNIP4:

The protein encoded by this gene is a member of the family of voltage-gated (Kv) channel-interacting proteins (KCNIPs). Members of this protein family are small calcium binding proteins and interaction partners of the voltage-gated potassium channel subunit Kv4 family (Weissflog et al., 2013). A candidate gene based association study (594 adult attention-deficit/hyperactivity disorder (ADHD) cases, 630 BD cases and 974 controls) showed *KCNIP4* to be associated with both ADHD (best p = 0.0079) and BD (best p = 0.0043) (Weissflog et al., 2013). However, the power of this study was limited due to the small number of samples.

Voltage-gated calcium channels and their binding partners have been implicated consistently in GWAS of psychiatric illness (Lee et al., 2012). In particular, the potassium voltage gated protein KCNC2, along with its interaction partner ANK3 (Judy et al., 2013)(which helps to regulate the localisation of voltage-gated ion channels (Garrido et al., 2003)) have both been implicated in multiple GWASs of BD ((Wellcome Trust Case Control, 2007); (Ferreira et al., 2008); (Muhleisen et al., 2014)). These calcium-binding proteins, including KCNIP4, are therefore excellent biological candidates for conferring risk of BD and depression. Together, these results suggest *KCNIP4* may play a role in conferring risk for BD and psychiatric illness. Two intronic variants for *KCNIP4* are present in the top 5% of ranked variants (9th and 20th respectively).

SORCS2:

SORCS2 is a member of the sortilin family of mammalian type-I transmembrane receptors containing a Vsp10p domain ((Hermey, 2009); (Willnow et al., 2008)). The sortilins are fundamental for development and maintenance of neuronal synaptic

properties, signalling characteristics, morphology and growth ((Jansen et al., 2007); (Hermey, 2009); (Lane et al., 2012)). Five members of the sortilins are found in vertebrates (SORL1, SORT1, SORCS1, SORCS2 and SORCS), all of which are expressed in the brain (Willnow et al., 2008). There is growing evidence that this family of receptors are potential neuronal disease genes, *SORL1* and *SORCS1* being implicated in Alzheimer disease (Reitz et al., 2013).

SORCS2 is a neuronal receptor involved in protein trafficking. This gene was implicated by a GWAS (best p value = 0.000014) ((Wellcome Trust Case Control, 2007); (Baum et al., 2008); (Ollila et al., 2009)) and an association study comparing 576 schizophrenia patients and 506 BD patients with 607 controls from the Scottish population (p = 0.0003)(Christoforou et al., 2007). However, despite these early suggestive findings, associations with this gene have failed to reach genome-wide significance in the largest BD GWAS to date by Mülheisen et al. (2014) (Table 5.1). This does not mean that rare variants within this gene are not associated with BD. Nor does it preclude the possibility of Scottish specific variation playing a role in the aetiology of BD in SBF2, as has been suggested for variants in *Neuregulin* and *DISC1* ((Walker et al., 2010); (Hennah et al., 2009)).

The top 5% of ranked variants from family SBF2 included 8 non-coding variants around this gene and within its introns. In addition, two variants within the 3'UTR of *SORCS2* were found. VEP classifies all 3' and 5' UTR variants as "modifiers", implying they may function through a regulatory function affecting gene expression, or mRNA stability or localisation (Duan et al., Hum Mol Genetics, 2003).

DRD5:

This gene encodes a dopamine receptor. There exists a lot of evidence implicating dopamine receptors in the aetiology of psychiatric conditions, in particular schizophrenia ((Brisch et al., 2014); (Hoenicka et al., 2007)). Most recently, the 2014 PGC schizophrenia GWAS identified the locus containing the *DRD2* gene as a genome-wide significantly associated region with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics, 2014). However, the mechanisms by which dopamine receptors might contribute to psychiatric disorders are still being investigated ((Laruelle, 2014);

(Grace, 2012); (Lodge and Grace, 2011)). We do know that neurotransmitter systems are closely linked, with interactions occurring between dopamine, glutamate (de Bartolomeis and Tomasetti, 2012) and serotonin (de Bartolomeis et al., 2013). In addition, all effective antipsychotic drugs (both classical and current drugs) for schizophrenia function through mechanisms that include dopamine and related neurochemical pathways (Brisch et al., 2014). A variant 5' to the DRD5 gene ranked 27th in the DM rankings. This variant appears to be unique to family SBF2.

SLIT2:

The SLIT2 protein acts as a molecular guidance cue in neuronal migration. The Drosophila homolog of this protein was shown to be a neuronally expressed protein that plays a role in axon guidance (Itoh et al., 1998). Although this gene has not been shown to be associated with any of the major psychiatric illnesses, this gene was identified in an association with anger in suicide attempts (Sokolowski et al., 2010). This, along with its function as an axon guidance molecule, important for neuronal wiring, makes it an interesting candidate gene for BD.

The 24th best ranking variant from the family SBF2 was a unique, intronic *SLIT2* variant, located 7 base pairs from an intron-exon boundary and predicted by VEP to be located in a splice site, though not to affect splicing.

CC2D2A:

This gene encodes a coil-coil domain protein. Mutations in *CC2D2A* have known to cause Joubert syndrome and Meckel syndrome, two forms of ciliopathies, which encompass a range of symptoms including mental retardation (Bachmann-Gagescu et al., 2012). This protein interacts with CEP290 (Gorden et al., 2008), which in turn has been shown to interact with DISC1 (Millar et al., 2003). *CEP290* has been implicated as an autism candidate gene (Cukier et al., 2014), while *DISC1* (disrupted in schizophrenia 1) has been implicated in several psychiatric illnesses ((Bradshaw and Porteous, 2012); (Brandon and Sawa, 2011); (Millar et al., 2001)), suggesting CC2D2A may be part of a larger network of proteins involved in psychiatric illness.

This gene harbours the only non-synonymous substitution (rs144439937) in the set of rare (MAF <0.05) variants from the SBF2 disease-linked chromosome. This variation results in a lysine residue (K) being replaced by a glutamate (E) at amino acid position 507. VEP reported this variant to be moderately likely to have a negative consequence to protein function. This variant changes a basic amino acid (K) to an acidic amino acid (E), which might alter the electrostatics of the surface of the protein, affecting how it interacts with other proteins. The population frequency of rs144439937 (MAF 0.008) meant it was excluded from the initial analysis of coding variants described in section 5.1.4.2 (as it was not unique to family SBF2). However, it is possible that this variant confers risk of illness via an interaction with other variants.

Intergenic variants within the rs215411 linkage region:

Six variants from SBF2 were found to be in LD with a genome-wide significant variant for schizophrenia, rs215411. On its own, the rank of the highest ranking variant (94th out of the 467 intergenic SBF2 variants) may not be sufficient to include this SNP in any follow-on experimental analysis; however, the independent evidence from the schizophrenia GWAS, suggesting a susceptibility variant for psychiatric illness may be located in this region, adds additional weight.

5.4.2.2 Potential future work on the SBF2 project

Experimental follow up:

SuRFR is a predictive method, designed to be an aid to genomics projects, prioritising variants from most likely to be functional (and therefore the best starting point for follow-up analysis), to least likely to be functional. Experimental analysis is needed to verify the functionality of high-ranking variants. Methods to functionally validate variants in the lab include luciferase assays and EMSA shift assays or other related approaches, which can be used to show changes in gene expression or in the binding ability of transcription factors. Several reviews discussing the various options available to experimentally establish the functional consequence of regulatory variants have been recently published ((MacArthur et al., 2014); (Li et al., 2015); (Knight, 2014)).

Validation of sequencing:

No sequencing platform has been shown to be 100% accurate (Lam et al., 2012). Based on cross-platform comparative analysis and our own sequencing validation experiments, it appears that rare or unique variants have the lowest validation rates and are most likely contain erroneous variants (Lam et al., 2012). Several courses of action are available to deal with this: i) The GATK SNP filtering steps chould be re-evaluated (more stringent thresholds should be used); ii) All high-ranking SNPs chould be re-sequenced in these individuals to show they are real variants.

Confirm association of causal variants with illness in additional cases/controls:

Candidate SNPs identified in the affected individuals should be followed up in additional family members to confirm that these variants segregate with illness. High-ranking common SNPs chould also be genotyped across a large number of individuals from the Scottish population (including other individuals with BP) to see if they are associated with BP on a population level. This is particularly relevant for the K507E variant in CC2D2A.

Consider additional disease model hypothesis:

I have hypothesised a dominant disease model, and have predicted the chr4p locus identified through linkage analysis to harbour a highly penetrant, rare susceptibility variant. However, the LOD generated by the linkage analysis (LOD = 4.1) does not negate the possibility of risk variants occurring elsewhere in the genome. Therefore, there could be other interpretations for genetic risk of BD in this family. For instance, the chr4p locus might harbour several variants that interact and together lead to illness. Similarly there might be a number of variants across the genome that could each individually have a small effect but together concert a polygenic risk.

5.4.2.3 Active projects on the Family SBF2 data

Our collaborators at CSHLs are sequencing the genomes of additional individuals from family SBF2. Another collaborator, Prof. Andrew McIntosh and his group, have genotyped over 50 individuals from family SBF2 and are calculating a polygenic risk

score for the family members. In addition, Andrew McIntosh's group is re-assessing the linkage in SBF2 using additional individuals and genotyping data. These data should be available in the next few months.

5.4.3 F224

Early onset cases of any disorder tend to be more severe than late-onset versions and are more likely to have a strong genetic component ((Gogtay et al., 2011); (Agopian et al., 2012); (Childs and Scriver, 1986)). Familial, highly penetrant disorders are therefore good candidates for gene identification. The related individuals with IO identified by Dunnigan and Pelosi, shown to also be affected by depressive symptoms, frequently diagnosed as MDD, are therefore potentially useful in the discovery of the genetic factors linked to IO and MDD.

The disease model for MDD and IO in F224 is less clear than for BD and MDD in family SBF2. The linkage analysis, although suggestive, did not point to any region of the genome with sufficient significance to focus the search on any one locus. This could be due to: heterogeneity across the four families; a multi-locus interaction model; a more polygenic model; and/or a large environmental component to illness. I used SuRFR to rank all 142,374 variants on the basis of predicted function. I then used VEP to identify proteins predicted to have a deleterious impact on protein structure and function.

5.4.3.1 Candidate genes from the SuRFR ranking analysis

Due to the lack of clarity from the linkage data, I have analysed the variants identified by whole genome sequencing in four datasets: the whole genome data; the chr7q linkage region; the chr8q linkage region; and the chr14q linkage region. The following section summarises the best candidate genes from these datasets.

QRICH1:

Little is known about the glutamate rich protein 1 (*QRICH1*) gene or its encoded protein. However, this protein is predicted to contain a caspase activation recruitment domain

(CARD) (from UniprotKB, 27th July 2015). The CARD family of proteins play an important role in regulating apoptosis, inflammation signalling and NF-kB signalling (Kao et al., 2015).

A non-coding variant in the promoter region of *QRICH1* ranked second out of the whole-genome SuRFR ranking data. This variant overlaps a range of strong regulatory features including the chromatin state active promoter, DNase HS in 125 cell lines; 37 TFBSs; and has a high conservation score. Due to the potential role of inflammation in the aetiology of IO and psychiatric illness, this is an interesting gene for further investigation.

ICA1:

The islet cell auto antigen (ICA1) gene has been associated with type I diabetes and plays a role in glucose regulation (Arvan et al., 2012). *ICA1* has also been implicated as an autoimmune gene (Johar et al., 2015). The link between IO and diabetic oedema implicates genes such as ICA1. This gene would therefore be an interesting gene to investigate further.

Two intronic *ICA1* variants were in the top 30 chr7q linkage region variants, ranking 2nd and 9th (311th and 891st out of the whole-genome data).

AQP1:

Aquaporin 1 plays a critical role in water transport across the peritoneal membrane, which forms the lining of the abdominal cavity, containing the blood vessels, lymph vessels and nerves (Morelle and Devuyst, 2015). Increased expression of AQP1 in peritoneal capillaries leads to increased water permeability (Devuyst and Ni, 2006).

An intronic variant of this gene ranked 5th of the variants in the chr7q linkage region and 624th out of the whole genome data. This variant is an excellent candidate for the capillary leakage phenotype seen in IO, which might also confer risk to psychiatric illness via inflammatory response due to abnormal blood-brain barrier communication ((Maes et al., 2008); (Shalev et al., 2009)). In addition, Aquaporin 4 (*AQP4*) is known to

be expressed in the brain (Amiry-Moghaddam et al., 2004) and has been proposed to function in concert with the potassium channel Kir4.1 (Nagelhus et al., 2004), suggesting a functional link between aquaporins, and potassium gated ion channels and a novel pathway involved in the aetiology of psychiatric illness.

CAMK2B:

Increased expression of the Calcium/calmodulin-dependent protein kinase II B (CAMK2B) protein in the frontal cortex has been shown in patients with schizophrenia and depression (Novak et al., 2006). Furthermore, increased expression of this protein was reported in the prefrontal cortex of suicide victims (Choi et al., 2011). This gene is therefore a good candidate for both the psychological symptoms of IO and MDD.

An intronic *CAMK2B* variant ranked 20th out of the chr7 linkage region variants (1360th overall). This variant overlaps the binding sites of over 20 transcription factors (by ChIP-seq data) and DNase HS in 110 cell lines.

DDC:

The Dopa Decarboxylase (*DDC*) gene is also known as Aromatic L-Amino Acid Decarboxylase (AADC). AADC is a key component of the serotonin and dopamine synthesis pathways ((Deneris and Wyler, 2012); (Cenci, 2014)). This protein is also expressed in blood vessel associated cells (Bertler et al., 1966). Mutations of this gene (homozygous and compound heterozygous mutations) lead to AADC deficiency, which negatively affects neurotransmitter metabolism, which in turn leads to a deficiency of both serotonin and dopamine (Brun et al., 2010)) clinically characterised this disorder as consisting of vegetative symptoms, oculogyric crises (a prolonged involuntary upward eye movement), dysteria (uncontrollable repetitive muscle movements) and severe neurological dysfunction which usually begins in infancy or early childhood. Other symptoms reported by Swoboda et al. (2003) include emotional lability and irritability, as well as gastrointestinal problems such as reflux disease, constipation and diarrhoea (Swoboda et al., 2003).

Some of these symptoms appear related to IO (motional lability, irratibility, gastrointestinal problems), although the overall aetiology is much more extreme. A regulatory variant could be hypothesised to lead to a less severe disorder than AADC, such as the intronic DDC variant which ranks 22nd out of the chr7q variants and 1492th out of the whole-genome ranked data.

KCNS2:

This gene encodes a potassium voltage gated channel. As described earlier in this discussion (5.4.2.1), potassium voltage gated channels are widely expressed in both the central and peripheral nervous system, mediate neuronal excitability ((Yellen, 2002); (McKeown et al., 2008)), and have been implicated via multiple lines of evidence to play a role in the pathology of psychiatric illness (Brisch et al., 2014).

A variant less than 1.5kb upstream of the *KCNS2* gene ranked 5th out of the chr8 linkage region variants and 246th out of the whole-genome ranked data.

NCALD:

A variant in the *NCALD* gene on chromosome 8 has been reported to be a risk variant for coeliac disease (Monten et al., 2015). In the same report, the authors suggest a link between *NCALD*, coeliac disease and nutrient signalling.

An intronic *NCALD* variant ranked 18th in the chr8q linkage region and 957th out of the whole genome ranking. This variant overlaps binding sites for nine transcription factors (UCSC genome browser, ENCODE ChIP-Seq track, accessed 27th July 2015): POLR2A, ATF2, FOXM1, EZH2, WRNIP1, STAT1, RELA, CHD1 and IKZF1. Of these, ATF2 responds to stress-related stimuli and plays a role in inflammation (Yu et al., 2014). Both IO and this transcription factor has been linked with obesity ((Pelosi et al., 1986); (Miyata et al., 2013)). In addition, the most successful treatment of IO is a reduced carbohydrate diet, which would include a reduction in gluten containing foods. Taken together these data suggest a potential overlap in function for NCALD in celiac disease and IO and a link between IO, obesity, ATF2 and NCALD.

SSTR1:

This gene encodes the somatostatin receptor 1, which Egerod et al. (2015) showed plays a role in somatostatin secretion in gastric somatostatin cells (Egerod et al., 2015). They also showed that this action is regulated by a combination of hormones, neurotransmitters, neuropeptides and metabolites (Egerod et al., 2015). A closely related protein, SSTR2, has also been implicated in the pathogenicity of Alzeihmer's disease (Adori et al., 2015).

Two variants for this gene were included in the top ranking variants for the chr14 linkage region. The first, which ranked 2nd (291st in the whole-genome data) lies within the 3' UTR of *SSTR1*; the second, ranking 28th (4270th in the whole-genome data) is located downstream of this gene.

5.4.3.2 Analysis of coding variants using VEP:

As with the SBF2 data, I also focused on the variants that overlapped exons. Because SuRFR cannot discriminate between different classes of exonic variants (synonymous, non-synonymous, UTR, etc), I used VEP to search this list of variants for ones that potentially have a deleterious effect on protein structure and function.

Nine variants were predicted by VEP to have a high impact, located in the genes: *MORN4, ZNF214, TRIM48, OR812, OR5M11, LEPREL2, ZNF717* and *CER1*. Of these, the most interesting candidate genes are *CER1* and *MORN4*. The *CER1* gene encodes a cytokine that may play a role in anterior neural induction and somite formation during embryogenesis (Uniprot, 27th July, 2015). *MORN4* has been shown through Drosophila and Mouse models to have a role in axon degeneration (Bhattacharya et al., 2012). Little is known in the literature about *ZNF717* (a pseudogene), *ZNF214* (a zinc finger protein), *OR812* and *OR5M11* (two olfactory receptor genes), *TRIM48* (a RING finger protein) or *LEPREL2* (a collagen prolyl hydroxylase).

Only one variant from this dataset was reported by VEP to have a clinical significance (although only predicted by VEP to have a moderate impact). This variant, rs351855 is a missense variant of the Fibroblast Growth Factor Receptor 4 (*FGFR4*) gene. This SNP is

associated with susceptibility to ischemic stroke ((Zhang et al., 2012a); (Yin et al., 2014)) and has been shown to modulate the association of a variant in the Klotho Beta (KLB) protein (expressed in the digestive system, regulates bile acid production and associated with diarrhoea (Camilleri et al., 2014)) with colonic transit in irritable bowl syndrome with diarrhoea (IBS) (Wong et al., 2011). According to UniProt (http://www.uniprot.org/uniprot/P22455#section_comments, accessed 12th August 2015), GO terms associated with *FGFR4* include cell migration, signalling pathways (including insulin receptor signalling and nerotrophin TRK receptor signalling pathway), and glucose homeostasis (see Figure 5.8 for the full list of GO terms reported by UniProt). Many of these terms support a potential biological link between this gene and IO and MDD.

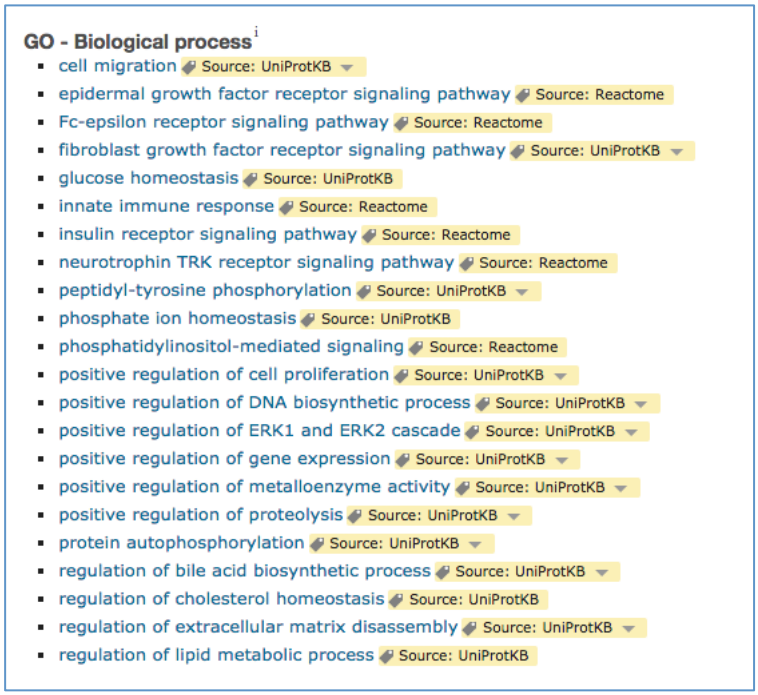


Figure 5.8 Screen shot from the UniProt webpage for the FGFR4 protein, showing the gene ontology terms associated with FGFR4.

5.4.3.3 Summary of implicated mechanisms

The strongest hypothesis for IO in the literature is vascular leakage in the capillary bed. The capillary bed and vasculature system work in tandem with the lymphatic system, both being integral parts of vasculature structure, regulating fluid release and uptake from the blood into the interstitial fluid and back into the blood (http://anatomyandphysiologyi.com/lymphatic-system/ accessed July 2015). Furthermore there is evidence in the literature of shared swelling symptoms between IO and lymph vessel diseases (characterised by lymphedema, swelling of the limbs due to a build up of lymph fluid in soft tissue), such as Elephantiasis ((Babu and Nutman, 2014). See also the WHO report on elephantiasis: http://www.who.int/mediacentre/factsheets/fs102/en/).
The dysregulation of either part of this vascular mechanism could explain the swelling symptoms observed in IO. In addition, the neurovasculature is an important component of the brain; defects in the mechanism of the blood-brain barrier and lymphatic vasculature have been implicated in neurological disorders ((Shalev et al., 2009); (Bell and Zlokovic, 2009)). The lymphatic system is also an important element of the immune system (Liao and von der Weid, 2015), which has recently been implicated as a common pathway for schizophrenia, BD and MDD (Network and Pathway Analysis Subgroup of Psychiatric Genomics, 2015). Taken together, these results suggest there might exist a common mechanism contributing to the range of phenotypes (swelling symptoms, functional-autonomic and affective disturbances) associated with IO and depression in F224. This hypothesis is supported in this analysis by the high-ranking variant identified in AQP1, which suggests a mechanism combining vasculature, lymphatic system and a potential immune/inflammatory response in the aetiology of IO and MDD.

There could also be a neurological cause for both IO and MDD in this family. The evidence that stress worsens both the swelling and affective symptoms of IO also supports the theory of a neurological component. Among some of the highest-ranking variants from SuRFR were variants involved in neuronal processes. These included the potassium channel protein KCNS2, the calcium/calmodulin dependent protein CAMK2B and the serotonin/dopamine synthesis protein DDC (which, being expressed in the blood-brain barrier, points back to the first hypothesis).

Interestingly, although drugs had very little effect, a change in diet from a normal diet to low carbohydrate diet has been shown to have greatest effect on IO swelling symptoms (Dunnigan and Pelosi, 1993). In cases with successful treatment, symptoms only returned when the diet restrictions were not followed or stressful life events occurred. This suggests that both diet and stress are important factors in the pathophysiology of IO, potentially implicating a carbohydrate metabolism/ insulin related cause to this disorder. The variant in the *NCALD* gene, which has been linked to coeliac disease, as well as a SNP in *ICA1*, which plays a role in glucose metabolism, are good candidates to follow-up this particular theory.

This analysis has highlighted several mechanisms that might play a role in the aetiology of IO and depression. It should be noted that none of these three mechanisms are mutually exclusive and some of the genes impacted may be involved in more than one pathway/mechanism.

5.4.3.4 Future work on the Idiopathic oedema data

One major drawback to my analysis of the IO data was its size, as the raw WGS data contained over 12 million variants, a number of which are likely to be sequencing errors. Before further analysis is done on these data, variants should be filtered as follows:

1. As with the SBF2 data, the GATK quality control step could be repeated using more stringent thresholds to improve the signal to noise ratio and increase our chances of identifying the true causal variants.

2. WGS of additional individuals from F224 would allow additional variants not present in all cases to be excluded, further reducing the number of SNPs to be analysed.

3. The linkage analysis performed by Anderson et al. (2008) included three other families. Analysing sequencing data from these families could allow us to filter variants further. Across these families there could be genes and/or pathways that harbour different variants, all contributing to disease risk.

Once a filtered set of SNPs has been produced and re-analysed (using the methods described in this chapter), candidate variants should be validated by Sanger sequencing. Many experimental methods are available to functionally characterise these variants. Some examples include: i) obtain gene expression data in appropriate cell lines and test whether these variants are eQTLs; ii) perform luciferase assays and EMSA shift assays to test if variants are regulatory variants and if they directly alter TFBSs; iii) use genome editing techniques such as CRISPR to recreate these variants in cell lines and use these cultured cells to test if variants alter the stability or localisation of mRNA or proteins.

5.5 Conclusion

I have used SuRFR to prioritise putative causal variants associated with two different psychiatric illnesses in two different projects. These two prioritised lists of variants will be used to guide the selection of variants for experimental and genetic investigation as part of a major international collaboration.

I have also discussed several drawbacks to the project design, and ways of improving the quality of the data to be analysed using SuRFR. I showed in Chapter 4 that the more refined the variant data, the better SuRFR performs. These points will be discussed further in Chapter 6.

Chapter 6: Discussion

6.1 Summary of thesis

Whole genome and whole exome sequencing methods have generated large amounts of data on human genetic variation. As these methods become more affordable they are likely to become routine tools in the investigation of the genetic basis of human disease (Wang et al., 2015). As "Big Data" continues to get bigger, we need tools to identify the signal of true pathogenic variation over the background noise generated by benign variation. This task is particularly challenging for non-coding variants, as our knowledge of what defines functional and pathogenic non-coding variation is limited. In this chapter I will summarise the aims of my PhD and my progress towards achieving them. I will also discuss limitations of this project and suggest directions this project could be taken in the future.

6.1.1 Aim 1

The first aim of my PhD was to develop a bioinformatics tool to prioritise variants on the basis of their putative functional and pathogenic roles. I have addressed this aim in Chapter 2 and Chapter 3, where I outlined the development and testing of my method, SuRFR, an R package for the ranked prioritisation of candidate causal variants. The modular design and tuneable parameterisation of SuRFR allows for simple and efficient incorporation of publically available data and prior biological knowledge into the ranking scheme. In Chapter 2, I introduced: i) the annotations used to prioritise known functional and pathogenic variants over background variants; ii) the training datasets I constructed for this analysis and iii) the principles behind the initial model I implemented. In Chapter 3, I expanded on the topics introduced in Chapter 2 and presented the formalised model training protocol used to develop SuRFR.

SuRFR produces rank orderings of variants for each of a wide diversity of functional genomic measures and annotations. These include: minor allele frequency (MAF); position of SNPs relative to genic elements (exons, introns, promoters, etc.); DNase hypersensitivity sites (DNase HS); chromatin states; transcription factor binding sites

(TFBSs); enhancers; and conservation. These individual ranks are then combined into a final rank using a weighting system parameterised and tested through ten-fold cross-validation. Central to the success of the parameterisation and testing of this approach is the quality of the training data. Known regulatory variants were obtained from the Human Gene Mutation Database (HGMD), while background variants were obtained by randomly sampling SNPs from the 1000 Genomes project located within the ENCODE pilot regions. Known and background variants were randomly assigned to training/validation sets and a hold out test set. Performance was measured using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUCs) statistics. Performance and generalisation errors were calculated to estimate the generalisability of the method and to predict its performance on novel data.

In Chapter 3 I showed that the AUCs from the optimum combination of weightings run on the hold out test dataset were very high (0.90-0.95), indicating that the method works well to prioritise known regulatory variants over background variants on an independent dataset. In addition, the performance and generalisation errors were low (0.004-0.030), indicating the likelihood of the pipeline performing equally well on novel data.

This analysis has provided insight into the extent to which different classes of functional annotation are most useful for the identification of known regulatory variants. I have shown that known regulatory variants tend to overlap some functional categories more than others: the most important factor for identifying a true variant across all regulatory classes of regulatory variant is position relative to genes (upstream proximity to the transcription start site (TSS) strongly affects the likelihood of a SNP being functional). However, this could reflect the acquisition bias that exists in all databases of known pathogenic variants, which are enriched for variants proximal to genes. Additional training data, for variants with a known disease role located in all genomic regions, is needed to rectify this imbalance. Unfortunately, such data is not currently available in the numbers needed to provide sufficient power.

6.1.2 Aim 2

The second aim of my PhD was to perform a comparative analysis between SuRFR and other prioritisation tools. In Chapter 4 I compared the performance of SuRFR against three related tools which were all published near the end of my PhD: GWAVA, CADD and FunSeq. In this chapter I showed that SuRFR performs equally well when applied to novel data (data that had not been used to train any of these methods). These data included variants from the ClinVAR database of clinical variation, a novel non-coding dataset (the RAVEN dataset), and coding variants from the HbVar database. In addition, when GWAVA, CADD and SuRFR were run on three datasets consisting of SNPs identified from the investigation complex traits ((Musunuru et al., 2010); (Myouzen et al., 2010); (Gaulton et al., 2010)), SuRFR outperformed both GWAVA and CADD.
Whilst SuRFR performs as well as these other methods, it also has several additional advantages in its design and implementation. These are detailed below:

6.1.2.1 Integration

Being an R package, SuRFR is a component of the R environment and can be used in combination with other R packages without the need for additional data formatting. R is becoming an increasingly important tool in genomics, mainly due to the advances and improvements being made to the R software project Bioconductor. The aim of this project is to provide a comprehensive suite of tools for the analysis of high throughput genomics data (Huber et al., 2015). The R packages provided and maintained by Bioconductor are individually useful, but collectively have even greater merit as they allow the analysis and interpretation of genomics data in a unified framework. In addition, any other R packages (either private or from other repositories such as CRAN) can be used in conjunction with those curated by Bioconductor.

6.1.2.2 Modularity

SuRFR has been constructed in such a way as to allow the user to incorporate additional data in the future. One example would be to include expression data generated in a specific cell line to add greater discriminatory power to the ranked list of variants.

6.1.2.3 Flexibility

Although I have trained three models (ALL, DM and DFP), SuRFR can also be run using custom models defined by the user, based on the weightings they feel most appropriate to their data. For instance, MAF can not only be up-weighted or down-weighted according to the user's preferences, but also the optimal MAF (in the default mode, set to unique, "0") can be specified by the user (for example, if the best associated SNP from a GWAS has a MAF of 0.3, the user may wish to set the optimum MAF to 0.3).

6.1.3 Aim 3

The third aim of my PhD was to apply SuRFR to the study of psychiatric illness. In Chapter 5, I analysed whole genome sequencing data from a large Scottish family with bipolar disorder (SBF2) and a second family with major depressive disorder (MDD) and idiopathic oedema (IO)(F224). Using SuRFR to prioritise these data I have highlighted several plausible candidate genes and variants for follow-up analysis. These include variants in pathways previously implicated in psychiatric illness including calcium channels and synaptic proteins, as well as high ranking variants from novel genes and pathways. One of the advantages of SuRFR is that the features that contribute to the rankings are easy to identify and investigate further. For instance, two high ranking variants from the SBF2 analysis (ranking 3rd and 4th overall), which lie within an intron of C1QTNF7 (itself a candidate psychiatric gene – see section 5.4.2 for further information), appear to have ranked highly because they were: unique to the family; had high DNase HS scores; overlapped DNase footprints in a large number of cell lines (16 and 11 respectively); had high chromatin state scores (Chromatin state score of 9, suggesting strong enhancers); and overlapped several TFBSs (including a binding site for c-Fos , a brain expressed transcription factor known to play a role in synaptic plasticity (Cohen and Greenberg, 2008)). These data together support the potential pathogenicity of these variants whilst also suggesting a potential mechanism of action.

These two family projects, involving international collaborations, are on-going. More information is currently being gathered on the variation present within these families and the segregation of variation with illness (additional linkage data, inclusion of additional

individuals, etc), which should become available in the near future. This will allow us the opportunity to finalise the subset of high ranking variants to be taken forward for follow-on experimental investigation.

6.2 Project limitations

6.2.1 Acquisition bias of training data

Methods that rely on catalogues of know pathogenic variants, including GWAVA and SuRFR, are limited by the number of variants within these catalogues. These datasets also suffer from acquisition bias, and tend to over represent variants near to or within genes. This issue particularly affects our ability to identify signals associated with long-range enhancers.

A new computational method was published in August 2015 which attempted to overcome this issue by implementing a sequence-based approach (Lee et al., 2015). This method predicts the functional effects of regulatory variants by training a gapped *k*-mer support vector machine (gkm-SVM) using cell-type specific sequence features of regulatory elements, including DNase HSs and putative TFBSs. The premise for this method is that cell-type specific regulatory elements can be identified using cell-type specific genomic features and that these data can be used to predict the effect of SNPs on these features in their native genomic contexts. The gkm-SVM produces a regulatory sequence vocabulary by generating scores for all unique 10-mer sequences, which it compares against the known regulatory sequences. The difference in gkm-SVM score between the wild-type variant and the SNP, termed deltaSVM, is used to predict how big a functional effect the SNP has. The larger the score (either positive or negative), the greater the SNP effect.

The authors trained the gkm-SVM using DNase HS data from specific cell types to identify genomic sequences that are likely to also have regulatory activity within those specific cell types and therefore predict the likelihood of novel variants affecting regulatory activity, thus identifying DNase quantitative trait loci (dsQTLs) (SNPs that are highly correlated with DNase-seq read depth (Degner et al., 2012)). This method was compared against GWAVA, CADD and GERP (Lee et al., 2015) using a dataset of

known dsQTL SNPs and non-dsQTL SNPs with comparable levels of DNase HS. The gkm-SVM was shown to predict SNPs associated with dsQTLs more accurately than the other methods (AUCs of 0.75, 0.63, 0.69 and 0.56 for gkm-SVM, CADD, GWAVA and GERP respectively).

An additional experiment reported in this paper to support gkm-SVM's performance (Lee et al., 2015) made use of the analysis of *SORT1* by Musurunu et al. (Musunuru et al., 2010). This analysis investigated a region of chromosome 1p13 associated with LDL-C levels and identified a single SNP, rs12740374, as altering the hepatic expression of the *SORT1* gene. I also used the data from this analysis to compare the performances of SuRFR, CADD and GWAVA in Chapter 4 (Section 4.3.1.4). In their study, Lee et al. showed that the deltaSVM for the functional SNP rs12740374 was only higher than for the surrounding SNPs when the gkm-SVM was trained on data from an appropriate cell-type (HepG2). When other cell-types were used (MEL and LNCaP) the gkm-SVM could not prioritise rs12740374 better than the background SNPs. This highlights that the performance of this method is very sensitive to the training data used and if the appropriate cell-type specific data is unavailable, its performance suffers. In contrast, SuRFR identified this variant 1st out of 22 SNPs, without requiring cell-type specific data, suggesting SuRFR may compare favourably with the gkm-SVM method on other data.

6.2.2 Limitations of family-based sequencing projects

There are many advantages to using next generation sequencing (NGS) methods to study variation contributing to the full spectrum of human disease types and associated genetic architectures. One of the greatest advantages of whole genome sequencing is that it detects common and rare variants, both within protein coding sequences and non-coding sequences, in the same assay. In addition, it is not limited to the study of single nucleotide polymorphisms, also being capable of identifying indels, CNVs and translocation events. However, there are several challenges facing the application of NGS.

The first relates to sequencing accuracy and sensitivity. Lam et al. (2012) compared the performances of the two leading sequencing platforms (Illumina and Complete Genomics) (Lam et al., 2012). This study showed that while there was an 88% concordance between platforms (with a sensitivity of 99.34%), there existed platform specific variation (2.7% for Complete Genomics and 9.2% for Illumina). The platform specific variation was also shown to be enriched for novel variants and was shown to have a false positive rate of at least 35%, suggesting many of these variants are likely to be errors. The concordance of indels was even lower between platforms, only 26.5% being common to both Complete Genomics and Illumina. The platform specific indels were also more difficult to validate as they were found to be more likely to overlap repeats, making them difficult to amplify by PCR for Sanger sequencing. These results suggest that care should be taken when using NGS methods to identify putative disease variants, as real variants will be missed and false positives will be included. Lam et al. further suggested that comprehensive variation detection could be better achieved by using at least two platforms. However, they also recognise that this would not always be possible due to the added expense.

The second challenge to be faced is base-calling and sequence alignment, which can also affect the sensitivity and specificity of sequencing data. Failure to accurately align sequence data to a reference genome can lead to large portions of the sequencing data being missed. Similarly, the quality control thresholds used by base-calling algorithms can leading to large error rates: too lax and they can lead to the inclusion of false positives; too severe and true SNPs can be left out as false negatives)(Nielsen et al., 2011). These should also therefore be considered carefully before sequence data is analysed.

While issues concerning the accuracy of sequencing data can affect whether a true variant is identifiable, there are other reasons why a causal variant may not be found in a family study. One reason is that the proposed genetic architecture is wrong, leading to incorrect filtering of pathogenic variants. In the analysis of the SBF2 disease-linked haplotype, I used a MAF threshold of 0.05, so only the uncommon, rare and unique

variants were included in my analysis. However, if my hypothesised disease model is incorrect, I might have missed susceptibility variants.

A second reason why a causal variant may not be identified could be within-family locus heterogeneity, where multiple variants are responsible for disease, but in different branches of the family. Bilineal inheritance would confound traditional methods, such as focusing on variants shared in cases but absent in controls, and linkage analysis. This has been shown by Rehman et al. (2015), who studied the effect of familial locus heterogeneity in a large number of families with various forms of hereditary hearing impairment (Rehman et al., 2015). In this analysis, the authors identified a large difference between the expected maximum LOD (mLOD) (calculated based on a fully penetrant, autosomal recessive marker) and the genome-wide LOD scores, suggesting that multiple loci are contributing to this disease in different parts of the family. To overcome this, Rehman et al. proposed splitting a heterogeneous family into smaller units, calculating new mLODs for each unit, and comparing this to the LODs generated for each unit. In addition, affection status of individuals within each unit can also be modified (cases alternately set to unknown) to compensate for heterogeneity within each family sub-unit, thereby identifying the affected individuals that are likely to segregate the same causal variants (for the full workflow, see Figure 4 from (Rehman et al., 2015)). Rehman et al. identified linkage regions using this approach and performed exome sequencing to confirm segregation of causal variants with the phenotype of interest.

The low LOD scores generated for F224 might indicate that this family is an example of familial heterogeneity. The linkage analysis could therefore be repeated, splitting the family into sub-units and comparing the mLOD for each sub-unit against the actual LODs achieved to identify individuals that segregate the same causal variant. In addition, this analysis could be performed for different phenotypes (idiopathic oedema only, major depressive disorder only and both diagnoses) to identify causal variants that segregate with one diagnoses but not the other.

6.3 Potential improvements to SuRFR

While I have shown SuRFR is capable of prioritising candidate SNPs with high specificity and sensitivity, there are several ways that I could improve SuRFR's functionality and usefulness in the future:

6.3.1 Coding variants

Many methods are available that are capable of differentiating between different classes of coding variants (UTR vs. protein coding; synonymous, non-synonymous, missense, splice variants, etc.), including Ensembl's Variant Effect Predictor (McLaren et al., 2010), SIFT (Ng and Henikoff, 2001), polyPhen (Adzhubei et al., 2010), SNAP (Bromberg and Rost, 2007), FATHMM (Shihab et al., 2013) and PANTHER (Thomas et al., 2003). Any one of these methods can be used separately to predict the deleteriousness of the coding variants identified during SuRFR's annotation step. However, it would be useful to have a unified framework that incorporates all the relevant data to prioritise different classes of coding variants and non-coding variants in a single pipeline. One method to do this would be to include the output of one or several of these protein-coding prediction methods into SuRFR's annotation table. This would allow SuRFR to discriminate between non-synonymous, synonymous, frame-shift, UTR, etc, while still incorporating all the previous genomic features.

Earlier this year, Dong et al. published a comparison of 18 deleterious-scoring methods, including three conservation scores, eleven functional prediction scores and four ensemble methods (combining multiple methods in a single output) (Dong et al., 2015). Using three independent datasets, this study found that the novel ensemble method being presented in this paper outperformed all the other methods. FATHMM was found to be the best performing individual tool, while the next best performing ensemble method (combining SIFT, PolyPhen-2, LRT, MutationTaster and PhyloP scores) was KGGSeq (Li et al., 2012). One of the main conclusions from this analysis was that ensemble methods can perform better than their individual component scores and that ensemble methods that included protein-specific features only perform better than methods that utilise general genomic annotation data (Dong et al., 2015).

This analysis suggests that the best way to incorporate a protein-level deleteriousness measure into SuRFR would be to incorporate an ensemble method into SuRFR's annotation data and train an additional model using protein-coding data from HGMD.

6.3.2 Indels

SNPs represent a large proportion of human variation, but are not the only class of variant that has been implicated in human disease. Indels are also known to have pathogenic roles ((Mullaney et al., 2010)). However, few prioritisation methods are designed to functionally interpret the deleteriousness of indels. One exception is the gkm-SVM method described in section 6.1.2 (Lee et al., 2015). This method allows indels to be analysed by summing the deltaSVM score across all affected nucleotides.

Although SuRFR is not currently trained to analyse indels, gkm-SVM suggests a framework that could be used to modify SuRFR's SNP functionality: by summing scores across all affected base positions. This, however, fails to take indel length into account, biasing the method in favour of longer indels. There is currently insufficient evidence suggesting that longer indels are more likely to be deleterious than short indels.

Deletions could alternatively be scored by summing and averaging the scores across all deleted bases (thus taking into account indel length). However, a deletion that overlaps both a single highly functional variant and many non-functional variants would not be prioritised by this method, as the average signal would be low. Instead, indels could be prioritised based on the highest functional score of any base affected by the indel. Possible approaches to scoring insertions is less obvious. One option would be to combine the scores of the bases either side of the insertion.

As with designing a SNP prioritisation method, the best way to resolve these options would be to use known pathogenic indels and background indels as a model training set. The HGMD database contains a catalogue of pathogenic indel data, while the 1000

genomes project also contains large numbers of indels. These data could be used to train an indel-specific version of SuRFR.

6.3.3 Variant interactions

As discussed in Chapter 1, there is a lot of evidence suggesting that some diseases are caused by multiple interacting variants. When two or more variants affect disease susceptibility, the performance of the predictive method can be increased by allowing for interactions between variants (Krzywinski and Altman, 2014). It would be interesting to look into the possibility of generating a two-point or multi-point version of SuRFR, which could take into account multiple interacting variants. However, the development of such a method would be limited by the availability of appropriate training data, as few validated epistatic interactions are catalogued.

6.3.4 Expression and methylation data

Disease risk variants are known to be enriched in expression quantitative trait loci (eQTLs) and methylation quantitative trait loci (meQTLs) ((Nicolae et al., 2010); (Gamazon et al., 2013); (Richards et al., 2012); (Westra et al., 2013)). See Albert and Kruglyal (2015) for a review of recent human eQTL datasets and the disease/trait studied (Albert and Kruglyak, 2015). Furthermore, there is evidence of cross-talk between DNA and histone methylation, gene expression being controlled by both forms of methylation, both together or independently (Du et al., 2015). Similarly, DNA methylation regions have been shown to overlap promoter regions and to be enriched for disease variants (Ma et al., 2015).

While DNA methylation data may be prioritised by SuRFR (by substituting the coordinates of SNPS in the input file with the coordinates of differentially methylated CpGs), it would be a useful extension to SuRFR's remit if it could be modified to function as a formal add-on to methylation packages to prioritise differentially methylated probes with similar p-values. In addition, databases of eQTL data, such as the Genotype-Tissue Expression (GTEx) project (Consortium, 2013), and meQTLs (Lemire et al., 2015) could be tested as additional prioritisation features.

6.3.5 Increased flexibility

SuRFR has been designed to make use of locally stored annotation data, constructing an annotation table from these data. In addition to R, this process relies on methods including bedtools and vcftools. The resulting annotation table is a simple dataframe, which is then used by the ranking function to create a second output table containing the individual ranks for each feature as well as the overall rank of these ranks, combined using user-specified weighting parameters.

I hope to submit SuRFR to the Bioconductor project for integration into their database of R packages. To this end, I have been in contact with the Bioconductor package development support team to discuss the restructuring that is needed before they can accept SuRFR. Earlier this year, the Bioconductor team published a paper outlining their future plans for genomic analyses projects (Huber et al., 2015). Their main aim for the future is to have a single universal system to store and manipulate genomic data. As part of this, Bioconductor have developed a new R object class called GRanges, a standardised format for storing all data pertaining to genomic coordinates and annotation data. For SuRFR to be accepted by Bioconductor I will have to reformat the data into a GRange object. An advantage of restructuring SuRFR into a GRange object is that all GRanges are compatible with other GRange objects, therefore increasing the flexibility of SuRFR further.

An additional requirement for SuRFR to be accepted is that all of the annotation data used by SuRFR (currently stored locally) must be added to AnnotationHub, a centralised annotation database that is updated and controlled by Bioconductor.

6.3.6 Tissue/cell type specificity

SNPs associated with complex traits have been shown to have tissue dependent effects on gene expression (Fu et al., 2012), suggesting that tissue specificity plays an important role in disease. Similarly, eQTLs are largely context dependent, being active in specific cell types at specific time points ((Nica et al., 2011); (Grundberg et al., 2012)).

Although tissue and cell-type specific annotation data is available for many features (including histone modifications, DNase HS, DNase footprints, enhancers, tissue specific promoters), the development of tissue/ cell-type specific versions of SuRFR is limited by the lack of sufficient cell-type specific and disease specific training data (the data that is available for single traits/diseases being limited in size). Despite this caveat, it could still be useful to incorporate cell-type or tissue specific data into the output of SuRFR. These data can be used in addition to the weighted rank of ranks for users to discriminate between variants. For instance, it would be interesting to identify brain-specific features that overlap variants of interest for psychiatric illness.

6.4 Conclusions

In this thesis I have described the development, testing and application of a novel computational approach for the functional investigation of putatively deleterious variants. This method filled a niche that was not covered by other tools. Since then, other tools have been developed that perform the same role; however, I have also shown that SuRFR compares favourably with these other approaches, confirming its continued relevance (Ryan et al., 2014)(Appendix D).

All prioritisation approaches are stepping-stones on the path to identifying true risk and causal disease variants. As such, their usefulness is in directing future research efforts towards a subset of variants to be followed up further, rather than being the end point of an analysis. In this context, the potential future plans of this project can be divided in two directions: the first, following up the candidate variants identified in Chapter 5; the second, expanding SuRFR's remit to allow the investigation of additional variant types in an improved model. These two directions are equally exciting and present the possibility of furthering our understanding of what constitutes a deleterious variant and how these variants function in disease.

References

ADIE, E. A., ADAMS, R. R., EVANS, K. L., PORTEOUS, D. J. & PICKARD, B. S. 2005. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics,* 6**,** 55.

ADORI, C., GLUCK, L., BARDE, S., YOSHITAKE, T., KOVACS, G. G., MULDER, J., MAGLOCZKY, Z., HAVAS, L., BOLCSKEI, K., MITSIOS, N., UHLEN, M., SZOLCSANYI, J., KEHR, J., RONNBACK, A., SCHWARTZ, T., REHFELD, J. F., HARKANY, T., PALKOVITS, M., SCHULZ, S. & HOKFELT, T. 2015. Critical role of somatostatin receptor 2 in the vulnerability of the central noradrenergic system: new aspects on Alzheimer's disease. *Acta Neuropathol,* 129**,** 541-63.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AGOPIAN, A. J., EASTCOTT, L. M. & MITCHELL, L. E. 2012. Age of onset and effect size in genome-wide association studies. *Birth Defects Res A Clin Mol Teratol,* 94**,** 908-11.

ALBERT, F. W. & KRUGLYAK, L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet,* 16**,** 197-212.

AMBERGER, J. S., BOCCHINI, C. A., SCHIETTECATTE, F., SCOTT, A. F. & HAMOSH, A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res,* 43**,** D789-98.

AMENT, S. A., SZELINGER, S., GLUSMAN, G., ASHWORTH, J., HOU, L., AKULA, N., SHEKHTMAN, T., BADNER, J. A., BRUNKOW, M. E., MAULDIN, D. E., STITTRICH, A. B., ROULEAU, K., DETERA-WADLEIGH, S. D., NURNBERGER, J. I., JR., EDENBERG, H. J., GERSHON, E. S., SCHORK, N., BIPOLAR GENOME, S., PRICE, N. D., GELINAS, R., HOOD, L., CRAIG, D., MCMAHON, F. J., KELSOE, J. R. & ROACH, J. C. 2015. Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proc Natl Acad Sci U S A,* 112**,** 3576-81.

AMIRY-MOGHADDAM, M., FRYDENLUND, D. S. & OTTERSEN, O. P. 2004. Anchoring of aquaporin-4 in brain: molecular mechanisms and implications for the physiology and pathophysiology of water transport. *Neuroscience,* 129**,** 999-1010.

ANDERSEN, M. C., ENGSTROM, P. G., LITHWICK, S., ARENILLAS, D., ERIKSSON, P., LENHARD, B., WASSERMAN, W. W. & ODEBERG, J. 2008. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol,* 4**,** e5.

ANDERSON, C. A., MACLEAN, A., DUNNIGAN, M. G., PELOSI, A. J., MURRAY, V., MCKEE, I., MCDONALD, G., BURT, D. W., MORRICE, D. R., MUIR, W. J., VISSCHER, P. M. & BLACKWOOD, D. H. 2008. A genome-wide linkage study in families with major depression and co-morbid unexplained swelling. *Am J Med Genet B Neuropsychiatr Genet,* 147**,** 356-62.

ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I., HOOF, I., BORNHOLDT, J., BOYD, M., CHEN, Y., ZHAO, X., SCHMIDL, C., SUZUKI, T., NTINI, E., ARNER, E., VALEN, E., LI, K., SCHWARZFISCHER, L., GLATZ, D., RAITHEL, J., LILJE, B., RAPIN, N., BAGGER, F. O., JORGENSEN, M., ANDERSEN, P. R., BERTIN, N., RACKHAM, O., BURROUGHS, A. M., BAILLIE, J. K., ISHIZU, Y., SHIMIZU, Y., FURUHATA, E., MAEDA, S., NEGISHI, Y., MUNGALL, C. J., MEEHAN, T. F., LASSMANN, T., ITOH, M., KAWAJI, H., KONDO, N., KAWAI, J., LENNARTSSON, A., DAUB, C. O., HEUTINK, P., HUME, D. A., JENSEN, T. H., SUZUKI, H., HAYASHIZAKI, Y., MULLER, F., CONSORTIUM, F., FORREST, A. R., CARNINCI, P., REHLI, M. & SANDELIN, A. 2014. An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-61.

ANNEY, R., KLEI, L., PINTO, D., REGAN, R., CONROY, J., MAGALHAES, T. R., CORREIA, C., ABRAHAMS, B. S., SYKES, N., PAGNAMENTA, A. T., ALMEIDA, J., BACCHELLI, E., BAILEY, A. J., BAIRD, G., BATTAGLIA, A., BERNEY, T., BOLSHAKOVA, N., BOLTE, S., BOLTON, P. F., BOURGERON, T., BRENNAN, S., BRIAN, J., CARSON, A. R., CASALLO, G., CASEY, J., CHU, S. H., COCHRANE, L., CORSELLO, C., CRAWFORD, E. L., CROSSETT, A., DAWSON, G., DE JONGE, M., DELORME, R., DRMIC, I., DUKETIS, E., DUQUE, F., ESTES, A., FARRAR, P., FERNANDEZ, B. A., FOLSTEIN, S. E., FOMBONNE, E., FREITAG, C. M., GILBERT, J., GILLBERG, C., GLESSNER, J. T., GOLDBERG, J., GREEN, J., GUTER, S. J., HAKONARSON, H., HERON, E. A., HILL, M., HOLT, R., HOWE, J. L., HUGHES, G., HUS, V., IGLIOZZI, R., KIM, C., KLAUCK, S. M., KOLEVZON, A., KORVATSKA, O., KUSTANOVICH, V., LAJONCHERE, C. M., LAMB, J. A., LASKAWIEC, M., LEBOYER, M., LE COUTEUR, A., LEVENTHAL, B. L., LIONEL, A. C., LIU, X. Q., LORD, C., LOTSPEICH, L., LUND, S. C., MAESTRINI, E., MAHONEY, W., MANTOULAN, C., MARSHALL, C. R., MCCONACHIE, H., MCDOUGLE, C. J., MCGRATH, J., MCMAHON, W. M., MELHEM, N. M., MERIKANGAS, A., MIGITA, O., MINSHEW, N. J., MIRZA, G. K., MUNSON, J., NELSON, S. F., NOAKES, C., NOOR, A., NYGREN, G., OLIVEIRA, G., PAPANIKOLAOU, K., PARR, J. R., PARRINI, B., PATON, T., PICKLES, A., PIVEN, J., POSEY, D. J., POUSTKA, A., POUSTKA, F., et al. 2010. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet,* 19**,** 4072-82.

ARLOT, C. 2010. A survey of cross-validation procedures for model selection. *Statist. Surv,* 4.

ARVAN, P., PIETROPAOLO, M., OSTROV, D. & RHODES, C. J. 2012. Islet autoantigens: structure, function, localization, and regulation. *Cold Spring Harb Perspect Med,* 2.

BABU, S. & NUTMAN, T. B. 2014. Immunology of lymphatic filariasis. *Parasite Immunol,* 36**,** 338-46.

BACHMANN-GAGESCU, R., ISHAK, G. E., DEMPSEY, J. C., ADKINS, J., O'DAY, D., PHELPS, I. G., GUNAY-AYGUN, M., KLINE, A. D., SZCZALUBA, K., MARTORELL, L., ALSWAID, A., ALRASHEED, S., PAI, S., IZATT, L., RONAN, A., PARISI, M. A., MEFFORD, H., GLASS, I. & DOHERTY, D. 2012. Genotype-phenotype correlation in CC2D2A-related Joubert syndrome reveals an association with ventriculomegaly and seizures. *J Med Genet,* 49**,** 126-37.

BAKER, L. A. 2014. Do our "big data" in genetic analysis need to get bigger? *Psychophysiology,* 51**,** 1321-2.

BARENBOIM, M. & MANKE, T. 2013. ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. *Bioinformatics,* 29**,** 2197-8.

BARNETT, J. H. & SMOLLER, J. W. 2009. The genetics of bipolar disorder. *Neuroscience,* 164**,** 331-43.

BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-5.

BASSETT, A. S. & CHOW, E. W. 2008. Schizophrenia and 22q11.2 deletion syndrome. *Curr Psychiatry Rep,* 10**,** 148-57.

BAUER, D. E., KAMRAN, S. C., LESSARD, S., XU, J., FUJIWARA, Y., LIN, C., SHAO, Z., CANVER, M. C., SMITH, E. C., PINELLO, L., SABO, P. J., VIERSTRA, J., VOIT, R. A., YUAN, G. C., PORTEUS, M. H., STAMATOYANNOPOULOS, J. A., LETTRE, G. & ORKIN, S. H. 2013. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science,* 342**,** 253-7.

BAUM, A. E., AKULA, N., CABANERO, M., CARDONA, I., CORONA, W., KLEMENS, B., SCHULZE, T. G., CICHON, S., RIETSCHEL, M., NOTHEN, M. M., GEORGI, A., SCHUMACHER, J., SCHWARZ, M., ABOU JAMRA, R., HOFELS, S., PROPPING, P., SATAGOPAN, J., DETERA-WADLEIGH, S. D., HARDY, J. & MCMAHON, F. J. 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry,* 13**,** 197-207.

BELL, R. D. & ZLOKOVIC, B. V. 2009. Neurovascular mechanisms and blood-brain barrier disorder in Alzheimer's disease. *Acta Neuropathol,* 118**,** 103-13.

BERGSTRA, B. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research,* 13**,** 281-305.

BERNDT, S. I., GUSTAFSSON, S., MAGI, R., GANNA, A., WHEELER, E., FEITOSA, M. F., JUSTICE, A. E., MONDA, K. L., CROTEAU-CHONKA, D. C., DAY, F. R., ESKO, T., FALL, T., FERREIRA, T., GENTILINI, D., JACKSON, A. U., LUAN, J., RANDALL, J. C., VEDANTAM, S., WILLER, C. J., WINKLER, T. W., WOOD, A. R., WORKALEMAHU, T., HU, Y. J., LEE, S. H., LIANG, L., LIN, D. Y., MIN, J. L., NEALE, B. M., THORLEIFSSON, G., YANG, J., ALBRECHT, E., AMIN, N., BRAGG-GRESHAM, J. L., CADBY, G., DEN HEIJER, M., EKLUND, N., FISCHER, K., GOEL, A., HOTTENGA, J. J., HUFFMAN, J. E., JARICK, I., JOHANSSON, A., JOHNSON, T., KANONI, S., KLEBER, M. E., KONIG, I. R., KRISTIANSSON, K., KUTALIK, Z., LAMINA, C., LECOEUR, C., LI, G., MANGINO, M., MCARDLE, W. L., MEDINA-GOMEZ, C., MULLER-NURASYID, M., NGWA, J. S., NOLTE, I. M., PATERNOSTER, L., PECHLIVANIS, S., PEROLA, M., PETERS, M. J., PREUSS, M., ROSE, L. M., SHI, J., SHUNGIN, D., SMITH, A. V., STRAWBRIDGE, R. J., SURAKKA, I., TEUMER, A., TRIP, M. D., TYRER, J., VAN VLIET-OSTAPTCHOUK, J. V., VANDENPUT, L., WAITE, L. L., ZHAO, J. H., ABSHER, D., ASSELBERGS, F. W., ATALAY, M., ATTWOOD, A. P., BALMFORTH, A. J., BASART, H., BEILBY, J., BONNYCASTLE, L. L., BRAMBILLA, P., BRUINENBERG, M., CAMPBELL, H., CHASMAN, D. I., CHINES, P. S., COLLINS, F. S., CONNELL, J. M., COOKSON, W. O., DE FAIRE, U., DE VEGT, F., DEI, M., DIMITRIOU, M., EDKINS, S., ESTRADA, K., EVANS, D. M., FARRALL, M., FERRARIO, M. M., et al. 2013. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet,* 45**,** 501-12.

BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L., ECKER, J. R., FARNHAM, P. J., HIRST, M., LANDER, E. S., MIKKELSEN, T. S. & THOMSON, J. A. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol,* 28**,** 1045-8.

BERTLER, A., FALCK, B., OWMAN, C. & ROSENGRENN, E. 1966. The localization of monoaminergic blood-brain barrier mechanisms. *Pharmacol Rev,* 18**,** 369-85.

BERTRAM, L., MCQUEEN, M. B., MULLIN, K., BLACKER, D. & TANZI, R. E. 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet,* 39**,** 17-23.

BERTRAM, L. & TANZI, R. E. 2008. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nat Rev Neurosci,* 9**,** 768-78.

BETANCUR, C. 2011. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res,* 1380**,** 42-77.

BHATTACHARYA, M. R., GERDTS, J., NAYLOR, S. A., ROYSE, E. X., EBSTEIN, S. Y., SASAKI, Y., MILBRANDT, J. & DIANTONIO, A. 2012. A model of toxic neuropathy in Drosophila reveals a role for MORN4 in promoting axonal degeneration. *J Neurosci,* 32**,** 5054-61.

BLACKWOOD, D. H., HE, L., MORRIS, S. W., MCLEAN, A., WHITTON, C., THOMSON, M., WALKER, M. T., WOODBURN, K., SHARP, C. M., WRIGHT, A. F., SHIBASAKI, Y., ST CLAIR, D. M., PORTEOUS, D. J. & MUIR, W. J. 1996. A locus for bipolar affective disorder on chromosome 4p. *Nat Genet,* 12**,** 427-30.

BLAUSTEIN, M. P. 1977. Sodium ions, calcium ions, blood pressure regulation, and hypertension: a reassessment and a hypothesis. *Am J Physiol,* 232**,** C165-73.

BLENNOW, K., DE LEON, M. J. & ZETTERBERG, H. 2006. Alzheimer's disease. *Lancet,* 368**,** 387-403.

BOFFELLI, D., MCAULIFFE, J., OVCHARENKO, D., LEWIS, K. D., OVCHARENKO, I., PACHTER, L. & RUBIN, E. M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science,* 299**,** 1391-4.

BOYAJYAN, A., ZAKHARYAN, R., ATSHEMYAN, S., CHAVUSHYAN, A. & MKRTCHYAN, G. 2015. Schizophrenia-associated Risk and Protective Variants of c-Fos Encoding Gene. *Recent Adv DNA Gene Seq*.

BOYCOTT, K. M., VANSTONE, M. R., BULMAN, D. E. & MACKENZIE, A. E. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet,* 14**,** 681-91.

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., CHERRY, J. M. & SNYDER, M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res,* 22**,** 1790-7.

BRADSHAW, N. J. & PORTEOUS, D. J. 2012. DISC1-binding proteins in neural development, signalling and schizophrenia. *Neuropharmacology,* 62**,** 1230-41.

BRANDON, N. J. & SAWA, A. 2011. Linking neurodevelopmental and synaptic theories of mental illness through DISC1. *Nat Rev Neurosci,* 12**,** 707-22.

BREST, P., LAPAQUETTE, P., SOUIDI, M., LEBRIGAND, K., CESARO, A., VOURET-CRAVIARI, V., MARI, B., BARBRY, P., MOSNIER, J. F., HEBUTERNE, X., HAREL-BELLAN, A., MOGRABI, B., DARFEUILLE-MICHAUD, A. & HOFMAN, P. 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet,* 43**,** 242-5.

BRISCH, R., SANIOTIS, A., WOLF, R., BIELAU, H., BERNSTEIN, H. G., STEINER, J., BOGERTS, B., BRAUN, K., JANKOWSKI, Z., KUMARATILAKE, J., HENNEBERG, M. & GOS, T. 2014. The role of

dopamine in schizophrenia from a neurobiological and evolutionary perspective: old fashioned, but still in vogue. *Front Psychiatry,* 5**,** 47.

BROMBERG, Y. & ROST, B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res,* 35**,** 3823-35.

BRUGGER, S. M., MERRILL, A. E., TORRES-VAZQUEZ, J., WU, N., TING, M. C., CHO, J. Y., DOBIAS, S. L., YI, S. E., LYONS, K., BELL, J. R., ARORA, K., WARRIOR, R. & MAXSON, R. 2004. A phylogenetically conserved cis-regulatory module in the Msx2 promoter is sufficient for BMP-dependent transcription in murine and Drosophila embryos. *Development,* 131**,** 5153-65.

BRUN, L., NGU, L. H., KENG, W. T., CH'NG, G. S., CHOY, Y. S., HWU, W. L., LEE, W. T., WILLEMSEN, M. A., VERBEEK, M. M., WASSENBERG, T., REGAL, L., ORCESI, S., TONDUTI, D., ACCORSI, P., TESTARD, H., ABDENUR, J. E., TAY, S., ALLEN, G. F., HEALES, S., KERN, I., KATO, M., BURLINA, A., MANEGOLD, C., HOFFMANN, G. F. & BLAU, N. 2010. Clinical and biochemical features of aromatic L-amino acid decarboxylase deficiency. *Neurology,* 75**,** 64-71.

CALABRIA, A., MOSCA, E., VITI, F., MERELLI, I. & MILANESI, L. 2010. SNPRanker: a tool for identification and scoring of SNPs associated to target genes. *J Integr Bioinform,* 7.

CAMILLERI, M., BUSCIGLIO, I., ACOSTA, A., SHIN, A., CARLSON, P., BURTON, D., RYKS, M., RHOTEN, D., LAMSAM, J., LUEKE, A., DONATO, L. J. & ZINSMEISTER, A. R. 2014. Effect of increased bile acid synthesis or fecal excretion in irritable bowel syndrome-diarrhea. *Am J Gastroenterol,* 109**,** 1621-30.

CARBONELL, J., ALLOZA, E., ARCE, P., BORREGO, S., SANTOYO, J., RUIZ-FERRER, M., MEDINA, I., JIMENEZ-ALMAZAN, J., MENDEZ-VIDAL, C., GONZALEZ-DEL POZO, M., VELA, A., BHATTACHARYA, S. S., ANTINOLO, G. & DOPAZO, J. 2012. A map of human microRNA variation uncovers unexpectedly high levels of variability. *Genome Med,* 4**,** 62.

CARDNO, A. G. & OWEN, M. J. 2014. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. *Schizophr Bull,* 40**,** 504-15.

CARLBORG, O. & HALEY, C. S. 2004. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet,* 5**,** 618-25.

CARUANA, N.-M. 2006. An empirical comparison of supervised learning algorithms. *ICML '06 Proceedings of the 23rd international conference on Machine learning***,** 161-168.

CENCI, M. A. 2014. Presynaptic Mechanisms of l-DOPA-Induced Dyskinesia: The Findings, the Debate, and the Therapeutic Implications. *Front Neurol,* 5**,** 242.

CHEN, D. T., JIANG, X., AKULA, N., SHUGART, Y. Y., WENDLAND, J. R., STEELE, C. J., KASSEM, L., PARK, J. H., CHATTERJEE, N., JAMAIN,

S., CHENG, A., LEBOYER, M., MUGLIA, P., SCHULZE, T. G., CICHON, S., NOTHEN, M. M., RIETSCHEL, M., BIGS, MCMAHON, F. J., FARMER, A., MCGUFFIN, P., CRAIG, I., LEWIS, C., HOSANG, G., COHEN-WOODS, S., VINCENT, J. B., KENNEDY, J. L. & STRAUSS, J. 2013. Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol Psychiatry,* 18**,** 195-205.

CHILDS, B. & SCRIVER, C. R. 1986. Age at onset and causes of disease. *Perspect Biol Med,* 29**,** 437-60.

CHOI, K., LE, T., XING, G., JOHNSON, L. R. & URSANO, R. J. 2011. Analysis of kinase gene expression in the frontal cortex of suicide victims: implications of fear and stress. *Front Behav Neurosci,* 5**,** 46.

CHRISTOFOROU, A., LE HELLARD, S., THOMSON, P. A., MORRIS, S. W., TENESA, A., PICKARD, B. S., WRAY, N. R., MUIR, W. J., BLACKWOOD, D. H., PORTEOUS, D. J. & EVANS, K. L. 2007. Association analysis of the chromosome 4p15-p16 candidate region for bipolar disorder and schizophrenia. *Mol Psychiatry,* 12**,** 1011-25.

CHUN, S. & FAY, J. C. 2009. Identification of deleterious mutations within three human genomes. *Genome Res,* 19**,** 1553-61.

CICHON, S., MUHLEISEN, T. W., DEGENHARDT, F. A., MATTHEISEN, M., MIRO, X., STROHMAIER, J., STEFFENS, M., MEESTERS, C., HERMS, S., WEINGARTEN, M., PRIEBE, L., HAENISCH, B., ALEXANDER, M., VOLLMER, J., BREUER, R., SCHMAL, C., TESSMANN, P., MOEBUS, S., WICHMANN, H. E., SCHREIBER, S., MULLER-MYHSOK, B., LUCAE, S., JAMAIN, S., LEBOYER, M., BELLIVIER, F., ETAIN, B., HENRY, C., KAHN, J. P., HEATH, S., BIPOLAR DISORDER GENOME STUDY, C., HAMSHERE, M., O'DONOVAN, M. C., OWEN, M. J., CRADDOCK, N., SCHWARZ, M., VEDDER, H., KAMMERER-CIERNIOCH, J., REIF, A., SASSE, J., BAUER, M., HAUTZINGER, M., WRIGHT, A., MITCHELL, P. B., SCHOFIELD, P. R., MONTGOMERY, G. W., MEDLAND, S. E., GORDON, S. D., MARTIN, N. G., GUSTAFSSON, O., ANDREASSEN, O., DJUROVIC, S., SIGURDSSON, E., STEINBERG, S., STEFANSSON, H., STEFANSSON, K., KAPUR-POJSKIC, L., ORUC, L., RIVAS, F., MAYORAL, F., CHUCHALIN, A., BABADJANOVA, G., TIGANOV, A. S., PANTELEJEVA, G., ABRAMOVA, L. I., GRIGOROIU-SERBANESCU, M., DIACONU, C. C., CZERSKI, P. M., HAUSER, J., ZIMMER, A., LATHROP, M., SCHULZE, T. G., WIENKER, T. F., SCHUMACHER, J., MAIER, W., PROPPING, P., RIETSCHEL, M. & NOTHEN, M. M. 2011. Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am J Hum Genet,* 88**,** 372-81.

COETZEE, S. G., RHIE, S. K., BERMAN, B. P., COETZEE, G. A. & NOUSHMEHR, H. 2012. FunciSNP: an R/bioconductor tool integrating

functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res,* 40**,** e139.

COHEN, S. & GREENBERG, M. E. 2008. Communication between the synapse and the nucleus in neuronal development, plasticity, and disease. *Annu Rev Cell Dev Biol,* 24**,** 183-209.

COMBARROS, O., CORTINA-BORJA, M., SMITH, A. D. & LEHMANN, D. J. 2009. Epistasis in sporadic Alzheimer's disease. *Neurobiol Aging,* 30**,** 1333-49.

CONDE, L., VAQUERIZAS, J. M., DOPAZO, H., ARBIZA, L., REUMERS, J., ROUSSEAU, F., SCHYMKOWITZ, J. & DOPAZO, J. 2006. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res,* 34**,** W621-5.

CONSORTIUM, C. 2015. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature,* 523**,** 588-91.

CONSORTIUM, E. P. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol,* 9**,** e1001046.

The ENCODE Project Consortium, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

CONSORTIUM, F., THE, R. P., CLST, FORREST, A. R., KAWAJI, H., REHLI, M., BAILLIE, J. K., DE HOON, M. J., HABERLE, V., LASSMANN, T., KULAKOVSKIY, I. V., LIZIO, M., ITOH, M., ANDERSSON, R., MUNGALL, C. J., MEEHAN, T. F., SCHMEIER, S., BERTIN, N., JORGENSEN, M., DIMONT, E., ARNER, E., SCHMIDL, C., SCHAEFER, U., MEDVEDEVA, Y. A., PLESSY, C., VITEZIC, M., SEVERIN, J., SEMPLE, C., ISHIZU, Y., YOUNG, R. S., FRANCESCATTO, M., ALAM, I., ALBANESE, D., ALTSCHULER, G. M., ARAKAWA, T., ARCHER, J. A., ARNER, P., BABINA, M., RENNIE, S., BALWIERZ, P. J., BECKHOUSE, A. G., PRADHAN-BHATT, S., BLAKE, J. A., BLUMENTHAL, A., BODEGA, B., BONETTI, A., BRIGGS, J., BROMBACHER, F., BURROUGHS, A. M., CALIFANO, A., CANNISTRACI, C. V., CARBAJO, D., CHEN, Y., CHIERICI, M., CIANI, Y., CLEVERS, H. C., DALLA, E., DAVIS, C. A., DETMAR, M., DIEHL, A. D., DOHI, T., DRABLOS, F., EDGE, A. S., EDINGER, M., EKWALL, K., ENDOH, M., ENOMOTO, H., FAGIOLINI, M., FAIRBAIRN, L., FANG, H., FARACH-CARSON, M. C., FAULKNER, G. J., FAVOROV, A. V., FISHER, M. E., FRITH, M. C., FUJITA, R., FUKUDA, S., FURLANELLO, C., FURINO, M., FURUSAWA, J., GEIJTENBEEK, T. B., GIBSON, A. P., GINGERAS, T., GOLDOWITZ, D., GOUGH, J., GUHL, S., GULER, R., GUSTINCICH, S., HA, T. J., HAMAGUCHI, M., HARA, M., HARBERS, M., HARSHBARGER, J., HASEGAWA, A., HASEGAWA, Y., HASHIMOTO, T., HERLYN, M., HITCHENS, K. J., HO SUI, S. J., HOFMANN, O. M., et al. 2014. A promoter-level mammalian expression atlas. *Nature,* 507**,** 462-70.

CONSORTIUM, G. T. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet,* 45**,** 580-5.

COOPER, G. M. & SHENDURE, J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet,* 12**,** 628-40.

COOPER, G. M., STONE, E. A., ASIMENOS, G., PROGRAM, N. C. S., GREEN, E. D., BATZOGLOU, S. & SIDOW, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res,* 15**,** 901-13.

COTTON, R. G., AUERBACH, A. D., BECKMANN, J. S., BLUMENFELD, O. O., BROOKES, A. J., BROWN, A. F., CARRERA, P., COX, D. W., GOTTLIEB, B., GREENBLATT, M. S., HILBERT, P., LEHVASLAIHO, H., LIANG, P., MARSH, S., NEBERT, D. W., POVEY, S., ROSSETTI, S., SCRIVER, C. R., SUMMAR, M., TOLAN, D. R., VERMA, I. C., VIHINEN, M. & DEN DUNNEN, J. T. 2008. Recommendations for locus-specific databases and their curation. *Hum Mutat,* 29**,** 2-5.

CRADDOCK, N. & SKLAR, P. 2009. Genetics of bipolar disorder: successful start to a long journey. *Trends Genet,* 25**,** 99-105.

CROSS-DISORDER GROUP OF THE PSYCHIATRIC GENOMICS, C. 2013. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet,* 381**,** 1371-9.

CROSS-DISORDER GROUP OF THE PSYCHIATRIC GENOMICS, C., LEE, S. H., RIPKE, S., NEALE, B. M., FARAONE, S. V., PURCELL, S. M., PERLIS, R. H., MOWRY, B. J., THAPAR, A., GODDARD, M. E., WITTE, J. S., ABSHER, D., AGARTZ, I., AKIL, H., AMIN, F., ANDREASSEN, O. A., ANJORIN, A., ANNEY, R., ANTTILA, V., ARKING, D. E., ASHERSON, P., AZEVEDO, M. H., BACKLUND, L., BADNER, J. A., BAILEY, A. J., BANASCHEWSKI, T., BARCHAS, J. D., BARNES, M. R., BARRETT, T. B., BASS, N., BATTAGLIA, A., BAUER, M., BAYES, M., BELLIVIER, F., BERGEN, S. E., BERRETTINI, W., BETANCUR, C., BETTECKEN, T., BIEDERMAN, J., BINDER, E. B., BLACK, D. W., BLACKWOOD, D. H., BLOSS, C. S., BOEHNKE, M., BOOMSMA, D. I., BREEN, G., BREUER, R., BRUGGEMAN, R., CORMICAN, P., BUCCOLA, N. G., BUITELAAR, J. K., BUNNEY, W. E., BUXBAUM, J. D., BYERLEY, W. F., BYRNE, E. M., CAESAR, S., CAHN, W., CANTOR, R. M., CASAS, M., CHAKRAVARTI, A., CHAMBERT, K., CHOUDHURY, K., CICHON, S., CLONINGER, C. R., COLLIER, D. A., COOK, E. H., COON, H., CORMAND, B., CORVIN, A., CORYELL, W. H., CRAIG, D. W., CRAIG, I. W., CROSBIE, J., CUCCARO, M. L., CURTIS, D., CZAMARA, D., DATTA, S., DAWSON, G., DAY, R., DE GEUS, E. J., DEGENHARDT, F., DJUROVIC, S., DONOHOE, G. J., DOYLE, A. E., DUAN, J., DUDBRIDGE, F., DUKETIS, E., EBSTEIN, R. P., EDENBERG, H. J., ELIA, J., ENNIS, S., ETAIN, B., FANOUS, A., FARMER, A. E., FERRIER, I. N., FLICKINGER, M., FOMBONNE, E., FOROUD, T., FRANK, J., FRANKE, B., et al. 2013. Genetic relationship

between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet,* 45**,** 984-94.

CUELLAR-PARTIDA, G., BUSKE, F. A., MCLEAY, R. C., WHITINGTON, T., NOBLE, W. S. & BAILEY, T. L. 2012. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics,* 28**,** 56-62.

CUKIER, H. N., DUEKER, N. D., SLIFER, S. H., LEE, J. M., WHITEHEAD, P. L., LALANNE, E., LEYVA, N., KONIDARI, I., GENTRY, R. C., HULME, W. F., BOOVEN, D. V., MAYO, V., HOFMANN, N. K., SCHMIDT, M. A., MARTIN, E. R., HAINES, J. L., CUCCARO, M. L., GILBERT, J. R. & PERICAK-VANCE, M. A. 2014. Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Mol Autism,* 5**,** 1.

CUNNINGHAM, F., AMODE, M. R., BARRELL, D., BEAL, K., BILLIS, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FITZGERALD, S., GIL, L., GIRON, C. G., GORDON, L., HOURLIER, T., HUNT, S. E., JANACEK, S. H., JOHNSON, N., JUETTEMANN, T., KAHARI, A. K., KEENAN, S., MARTIN, F. J., MAUREL, T., MCLAREN, W., MURPHY, D. N., NAG, R., OVERDUIN, B., PARKER, A., PATRICIO, M., PERRY, E., PIGNATELLI, M., RIAT, H. S., SHEPPARD, D., TAYLOR, K., THORMANN, A., VULLO, A., WILDER, S. P., ZADISSA, A., AKEN, B. L., BIRNEY, E., HARROW, J., KINSELLA, R., MUFFATO, M., RUFFIER, M., SEARLE, S. M., SPUDICH, G., TREVANION, S. J., YATES, A., ZERBINO, D. R. & FLICEK, P. 2015. Ensembl 2015. *Nucleic Acids Res,* 43**,** D662-9.

DAVISON, K. 2012. Autoimmunity in psychiatry. *Br J Psychiatry,* 200**,** 353-5.

DAVYDOV, E. V., GOODE, D. L., SIROTA, M., COOPER, G. M., SIDOW, A. & BATZOGLOU, S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol,* 6**,** e1001025.

DAYEM ULLAH, A. Z., LEMOINE, N. R. & CHELALA, C. 2013. A practical guide for the functional annotation of genetic variations using SNPnexus. *Brief Bioinform,* 14**,** 437-47.

DE BARTOLOMEIS, A., BUONAGURO, E. F. & IASEVOLI, F. 2013. Serotonin-glutamate and serotonin-dopamine reciprocal interactions as putative molecular targets for novel antipsychotic treatments: from receptor heterodimers to postsynaptic scaffolding and effector proteins. *Psychopharmacology (Berl),* 225**,** 1-19.

DE BARTOLOMEIS, A. & TOMASETTI, C. 2012. Calcium-dependent networks in dopamine-glutamate interaction: the role of postsynaptic scaffolding proteins. *Mol Neurobiol,* 46**,** 275-96.

DE GOBBI, M., VIPRAKASIT, V., HUGHES, J. R., FISHER, C., BUCKLE, V. J., AYYUB, H., GIBBONS, R. J., VERNIMMEN, D., YOSHINAGA, Y., DE JONG, P., CHENG, J. F., RUBIN, E. M., WOOD, W. G., BOWDEN, D. &

HIGGS, D. R. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science,* 312**,** 1215-7.

DE RUBEIS, S., HE, X., GOLDBERG, A. P., POULTNEY, C. S., SAMOCHA, K., CICEK, A. E., KOU, Y., LIU, L., FROMER, M., WALKER, S., SINGH, T., KLEI, L., KOSMICKI, J., SHIH-CHEN, F., ALEKSIC, B., BISCALDI, M., BOLTON, P. F., BROWNFELD, J. M., CAI, J., CAMPBELL, N. G., CARRACEDO, A., CHAHROUR, M. H., CHIOCCHETTI, A. G., COON, H., CRAWFORD, E. L., CURRAN, S. R., DAWSON, G., DUKETIS, E., FERNANDEZ, B. A., GALLAGHER, L., GELLER, E., GUTER, S. J., HILL, R. S., IONITA-LAZA, J., JIMENZ GONZALEZ, P., KILPINEN, H., KLAUCK, S. M., KOLEVZON, A., LEE, I., LEI, I., LEI, J., LEHTIMAKI, T., LIN, C. F., MA'AYAN, A., MARSHALL, C. R., MCINNES, A. L., NEALE, B., OWEN, M. J., OZAKI, N., PARELLADA, M., PARR, J. R., PURCELL, S., PUURA, K., RAJAGOPALAN, D., REHNSTROM, K., REICHENBERG, A., SABO, A., SACHSE, M., SANDERS, S. J., SCHAFER, C., SCHULTE-RUTHER, M., SKUSE, D., STEVENS, C., SZATMARI, P., TAMMIMIES, K., VALLADARES, O., VORAN, A., LI-SAN, W., WEISS, L. A., WILLSEY, A. J., YU, T. W., YUEN, R. K., STUDY, D. D. D., HOMOZYGOSITY MAPPING COLLABORATIVE FOR, A., CONSORTIUM, U. K., COOK, E. H., FREITAG, C. M., GILL, M., HULTMAN, C. M., LEHNER, T., PALOTIE, A., SCHELLENBERG, G. D., SKLAR, P., STATE, M. W., SUTCLIFFE, J. S., WALSH, C. A., SCHERER, S. W., ZWICK, M. E., BARETT, J. C., CUTLER, D. J., ROEDER, K., DEVLIN, B., DALY, M. J. & BUXBAUM, J. D. 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature,* 515**,** 209-15.

DECIPHERING DEVELOPMENTAL DISORDERS, S. 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature,* 519**,** 223-8.

DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J. B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E., STEPHENS, M., GILAD, Y. & PRITCHARD, J. K. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature,* 482**,** 390-4.

DENERIS, E. S. & WYLER, S. C. 2012. Serotonergic transcriptional networks and potential importance to mental health. *Nat Neurosci,* 15**,** 519-27.

DENNING, D. W., DUNNIGAN, M. G., TILLMAN, J., DAVIS, J. A. & FORREST, C. A. 1990. The relationship between 'normal' fluid retention in women and idiopathic oedema. *Postgrad Med J,* 66**,** 363-6.

DEVUYST, O. & NI, J. 2006. Aquaporin-1 in the peritoneal membrane: Implications for water transport across capillaries and peritoneal dialysis. *Biochim Biophys Acta,* 1758**,** 1078-84.

DICK, D. M., ALIEV, F., KRUEGER, R. F., EDWARDS, A., AGRAWAL, A., LYNSKEY, M., LIN, P., SCHUCKIT, M., HESSELBROCK, V.,

NURNBERGER, J., JR., ALMASY, L., PORJESZ, B., EDENBERG, H. J., BUCHOLZ, K., KRAMER, J., KUPERMAN, S. & BIERUT, L. 2011. Genome-wide association study of conduct disorder symptomatology. *Mol Psychiatry,* 16**,** 800-8.

DONG, C., WEI, P., JIAN, X., GIBBS, R., BOERWINKLE, E., WANG, K. & LIU, X. 2015. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet,* 24**,** 2125-37.

DRUMM, M. L., ZIADY, A. G. & DAVIS, P. B. 2012. Genetic variation and clinical heterogeneity in cystic fibrosis. *Annu Rev Pathol,* 7**,** 267-82.

DU, J., JOHNSON, L. M., JACOBSEN, S. E. & PATEL, D. J. 2015. DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol,* 16**,** 519-32.

DUDBRIDGE, F. 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genet,* 9**,** e1003348.

DUNNIGAN, M. G., HENDERSON, J. B., HOLE, D. & PELOSI, A. J. 2004. Unexplained swelling symptoms in women (idiopathic oedema) comprise one component of a common polysymptomatic syndrome. *QJM,* 97**,** 755-64.

DUNNIGAN, M. G. & PELOSI, A. J. 1993. Familial idiopathic oedema in prepubertal children: a new syndrome. *Q J Med,* 86**,** 301-13.

DUPUY, A. & SIMON, R. M. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst,* 99**,** 147-57.

EATON, W. W., MARTINS, S. S., NESTADT, G., BIENVENU, O. J., CLARKE, D. & ALEXANDRE, P. 2008. The burden of mental disorders. *Epidemiol Rev,* 30**,** 1-14.

EGEROD, K. L., ENGELSTOFT, M. S., LUND, M. L., GRUNDDAL, K. V., ZHAO, M., BARIR-JENSEN, D., NYGAARD, E. B., PETERSEN, N., HOLST, J. J. & SCHWARTZ, T. W. 2015. Transcriptional and Functional Characterization of the G Protein-Coupled Receptor Repertoire of Gastric Somatostatin Cells. *Endocrinology***,** EN20151388.

ELGAR, G. & VAVOURI, T. 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet,* 24**,** 344-52.

ERNST, J. & KELLIS, M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol,* 28**,** 817-25.

ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M., KU, M., DURHAM, T., KELLIS, M. & BERNSTEIN, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

FERREIRA, M. A., O'DONOVAN, M. C., MENG, Y. A., JONES, I. R., RUDERFER, D. M., JONES, L., FAN, J., KIROV, G., PERLIS, R. H., GREEN, E. K., SMOLLER, J. W., GROZEVA, D., STONE, J., NIKOLOV, I., CHAMBERT, K., HAMSHERE, M. L., NIMGAONKAR, V. L., MOSKVINA, V., THASE, M. E., CAESAR, S., SACHS, G. S., FRANKLIN, J., GORDON-SMITH, K., ARDLIE, K. G., GABRIEL, S. B., FRASER, C., BLUMENSTIEL, B., DEFELICE, M., BREEN, G., GILL, M., MORRIS, D. W., ELKIN, A., MUIR, W. J., MCGHEE, K. A., WILLIAMSON, R., MACINTYRE, D. J., MACLEAN, A. W., ST, C. D., ROBINSON, M., VAN BECK, M., PEREIRA, A. C., KANDASWAMY, R., MCQUILLIN, A., COLLIER, D. A., BASS, N. J., YOUNG, A. H., LAWRENCE, J., FERRIER, I. N., ANJORIN, A., FARMER, A., CURTIS, D., SCOLNICK, E. M., MCGUFFIN, P., DALY, M. J., CORVIN, A. P., HOLMANS, P. A., BLACKWOOD, D. H., GURLING, H. M., OWEN, M. J., PURCELL, S. M., SKLAR, P., CRADDOCK, N. & WELLCOME TRUST CASE CONTROL, C. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet,* 40**,** 1056-8.

FLINT, J. & KENDLER, K. S. 2014. The genetics of major depression. *Neuron,* 81**,** 484-503.

FOLEY, D. L., MACKINNON, A., MORGAN, V. A., WATTS, G. F., CASTLE, D. J., WATERREUS, A. & GALLETLY, C. A. 2015. Common familial risk factors for schizophrenia and diabetes mellitus. *Aust N Z J Psychiatry*.

FROMER, M., POCKLINGTON, A. J., KAVANAGH, D. H., WILLIAMS, H. J., DWYER, S., GORMLEY, P., GEORGIEVA, L., REES, E., PALTA, P., RUDERFER, D. M., CARRERA, N., HUMPHREYS, I., JOHNSON, J. S., ROUSSOS, P., BARKER, D. D., BANKS, E., MILANOVA, V., GRANT, S. G., HANNON, E., ROSE, S. A., CHAMBERT, K., MAHAJAN, M., SCOLNICK, E. M., MORAN, J. L., KIROV, G., PALOTIE, A., MCCARROLL, S. A., HOLMANS, P., SKLAR, P., OWEN, M. J., PURCELL, S. M. & O'DONOVAN, M. C. 2014. De novo mutations in schizophrenia implicate synaptic networks. *Nature,* 506**,** 179-84.

FU, J., WOLFS, M. G., DEELEN, P., WESTRA, H. J., FEHRMANN, R. S., TE MEERMAN, G. J., BUURMAN, W. A., RENSEN, S. S., GROEN, H. J., WEERSMA, R. K., VAN DEN BERG, L. H., VELDINK, J., OPHOFF, R. A., SNIEDER, H., VAN HEEL, D., JANSEN, R. C., HOFKER, M. H., WIJMENGA, C. & FRANKE, L. 2012. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet,* 8**,** e1002431.

FU, W., O'CONNOR, T. D. & AKEY, J. M. 2013. Genetic architecture of quantitative traits and complex diseases. *Curr Opin Genet Dev,* 23**,** 678-83.

GAMAZON, E. R., BADNER, J. A., CHENG, L., ZHANG, C., ZHANG, D., COX, N. J., GERSHON, E. S., KELSOE, J. R., GREENWOOD, T. A., NIEVERGELT, C. M., CHEN, C., MCKINNEY, R., SHILLING, P. D., SCHORK, N. J., SMITH, E. N., BLOSS, C. S., NURNBERGER, J. I.,

EDENBERG, H. J., FOROUD, T., KOLLER, D. L., SCHEFTNER, W. A., CORYELL, W., RICE, J., LAWSON, W. B., NWULIA, E. A., HIPOLITO, M., BYERLEY, W., MCMAHON, F. J., SCHULZE, T. G., BERRETTINI, W. H., POTASH, J. B., ZANDI, P. P., MAHON, P. B., MCINNIS, M. G., ZOLLNER, S., ZHANG, P., CRAIG, D. W., SZELINGER, S., BARRETT, T. B. & LIU, C. 2013. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry,* 18**,** 340-6.

GARRIDO, J. J., FERNANDES, F., MOUSSIF, A., FACHE, M. P., GIRAUD, P. & DARGENT, B. 2003. Dynamic compartmentalization of the voltage-gated sodium channels in axons. *Biol Cell,* 95**,** 437-45.

GAULTON, K. J., NAMMO, T., PASQUALI, L., SIMON, J. M., GIRESI, P. G., FOGARTY, M. P., PANHUIS, T. M., MIECZKOWSKI, P., SECCHI, A., BOSCO, D., BERNEY, T., MONTANYA, E., MOHLKE, K. L., LIEB, J. D. & FERRER, J. 2010. A map of open chromatin in human pancreatic islets. *Nat Genet,* 42**,** 255-9.

GENOMES PROJECT, C., ABECASIS, G. R., ALTSHULER, D., AUTON, A., BROOKS, L. D., DURBIN, R. M., GIBBS, R. A., HURLES, M. E. & MCVEAN, G. A. 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

GENOMES PROJECT, C., ABECASIS, G. R., AUTON, A., BROOKS, L. D., DEPRISTO, M. A., DURBIN, R. M., HANDSAKER, R. E., KANG, H. M., MARTH, G. T. & MCVEAN, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

GEORGI, B., CRAIG, D., KEMBER, R. L., LIU, W., LINDQUIST, I., NASSER, S., BROWN, C., EGELAND, J. A., PAUL, S. M. & BUCAN, M. 2014. Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genet,* 10**,** e1004229.

GEORGIEVA, L., REES, E., MORAN, J. L., CHAMBERT, K. D., MILANOVA, V., CRADDOCK, N., PURCELL, S., SKLAR, P., MCCARROLL, S., HOLMANS, P., O'DONOVAN, M. C., OWEN, M. J. & KIROV, G. 2014. De novo CNVs in bipolar affective disorder and schizophrenia. *Hum Mol Genet,* 23**,** 6677-83.

GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K. K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R., MIN, R., ALVES, P., ABYZOV, A., ADDLEMAN, N., BHARDWAJ, N., BOYLE, A. P., CAYTING, P., CHAROS, A., CHEN, D. Z., CHENG, Y., CLARKE, D., EASTMAN, C., EUSKIRCHEN, G., FRIETZE, S., FU, Y., GERTZ, J., GRUBERT, F., HARMANCI, A., JAIN, P., KASOWSKI, M., LACROUTE, P., LENG, J., LIAN, J., MONAHAN, H., O'GEEN, H., OUYANG, Z., PARTRIDGE, E. C., PATACSIL, D., PAULI, F., RAHA, D., RAMIREZ, L., REDDY, T. E., REED, B., SHI, M., SLIFER, T., WANG, J., WU, L., YANG, X., YIP, K. Y., ZILBERMAN-SCHAPIRA, G., BATZOGLOU, S., SIDOW, A., FARNHAM, P. J., MYERS, R. M.,

WEISSMAN, S. M. & SNYDER, M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature,* 489**,** 91-100.

GHANEM, N., JARINOVA, O., AMORES, A., LONG, Q., HATCH, G., PARK, B. K., RUBENSTEIN, J. L. & EKKER, M. 2003. Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res,* 13**,** 533-43.

GIARDINE, B., VAN BAAL, S., KAIMAKIS, P., RIEMER, C., MILLER, W., SAMARA, M., KOLLIA, P., ANAGNOU, N. P., CHUI, D. H., WAJCMAN, H., HARDISON, R. C. & PATRINOS, G. P. 2007. HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Hum Mutat,* 28**,** 206.

GIORDANO, S., AMATO, F., ELCE, A., MONTI, M., IANNONE, C., PUCCI, P., SEIA, M., ANGIONI, A., ZARRILLI, F., CASTALDO, G. & TOMAIUOLO, R. 2013. Molecular and functional analysis of the large 5' promoter region of CFTR gene revealed pathogenic mutations in CF and CFTR-related disorders. *J Mol Diagn,* 15**,** 331-40.

GOGTAY, N., VYAS, N. S., TESTA, R., WOOD, S. J. & PANTELIS, C. 2011. Age of onset of schizophrenia: perspectives from structural neuroimaging studies. *Schizophr Bull,* 37**,** 504-13.

GOODSWEN, S. J., GONDRO, C., WATSON-HAIGH, N. S. & KADARMIDEEN, H. N. 2010. FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. *BMC Bioinformatics,* 11**,** 311.

GORDEN, N. T., ARTS, H. H., PARISI, M. A., COENE, K. L., LETTEBOER, S. J., VAN BEERSUM, S. E., MANS, D. A., HIKIDA, A., ECKERT, M., KNUTZEN, D., ALSWAID, A. F., OZYUREK, H., DIBOOGLU, S., OTTO, E. A., LIU, Y., DAVIS, E. E., HUTTER, C. M., BAMMLER, T. K., FARIN, F. M., DORSCHNER, M., TOPCU, M., ZACKAI, E. H., ROSENTHAL, P., OWENS, K. N., KATSANIS, N., VINCENT, J. B., HILDEBRANDT, F., RUBEL, E. W., RAIBLE, D. W., KNOERS, N. V., CHANCE, P. F., ROEPMAN, R., MOENS, C. B., GLASS, I. A. & DOHERTY, D. 2008. CC2D2A is mutated in Joubert syndrome and interacts with the ciliopathy-associated basal body protein CEP290. *Am J Hum Genet,* 83**,** 559-71.

GOTTGENS, B., BARTON, L. M., CHAPMAN, M. A., SINCLAIR, A. M., KNUDSEN, B., GRAFHAM, D., GILBERT, J. G., ROGERS, J., BENTLEY, D. R. & GREEN, A. R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. *Genome Res,* 12**,** 749-59.

GRACE, A. A. 2012. Dopamine system dysregulation by the hippocampus: implications for the pathophysiology and treatment of schizophrenia. *Neuropharmacology,* 62**,** 1342-8.

GRATTEN, J., VISSCHER, P. M., MOWRY, B. J. & WRAY, N. R. 2013. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet,* 45**,** 234-8.

GREENFIELD, A. D. 1966. Survey of the evidence for active neurogenic vasodilatation in man. *Fed Proc,* 25**,** 1607-10.

GREENWOOD, T. A., AKISKAL, H. S., AKISKAL, K. K., BIPOLAR GENOME, S. & KELSOE, J. R. 2012. Genome-wide association study of temperament in bipolar disorder reveals significant associations with three novel Loci. *Biol Psychiatry,* 72**,** 303-10.

GRUNDBERG, E., SMALL, K. S., HEDMAN, A. K., NICA, A. C., BUIL, A., KEILDSON, S., BELL, J. T., YANG, T. P., MEDURI, E., BARRETT, A., NISBETT, J., SEKOWSKA, M., WILK, A., SHIN, S. Y., GLASS, D., TRAVERS, M., MIN, J. L., RING, S., HO, K., THORLEIFSSON, G., KONG, A., THORSTEINDOTTIR, U., AINALI, C., DIMAS, A. S., HASSANALI, N., INGLE, C., KNOWLES, D., KRESTYANINOVA, M., LOWE, C. E., DI MEGLIO, P., MONTGOMERY, S. B., PARTS, L., POTTER, S., SURDULESCU, G., TSAPROUNI, L., TSOKA, S., BATAILLE, V., DURBIN, R., NESTLE, F. O., O'RAHILLY, S., SORANZO, N., LINDGREN, C. M., ZONDERVAN, K. T., AHMADI, K. R., SCHADT, E. E., STEFANSSON, K., SMITH, G. D., MCCARTHY, M. I., DELOUKAS, P., DERMITZAKIS, E. T., SPECTOR, T. D. & MULTIPLE TISSUE HUMAN EXPRESSION RESOURCE, C. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet,* 44**,** 1084-9.

GUO, L., DU, Y., CHANG, S., ZHANG, K. & WANG, J. 2014. rSNPBase: a database for curated regulatory SNPs. *Nucleic Acids Res,* 42**,** D1033-9.

GUSELLA, J. F., WEXLER, N. S., CONNEALLY, P. M., NAYLOR, S. L., ANDERSON, M. A., TANZI, R. E., WATKINS, P. C., OTTINA, K., WALLACE, M. R., SAKAGUCHI, A. Y. & ET AL. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature,* 306**,** 234-8.

HAGER, G. 2009. Footprints by deep sequencing. *Nat Methods,* 6**,** 254-5.

HARDY, J. & SINGLETON, A. 2009. Genomewide association studies and human disease. *N Engl J Med,* 360**,** 1759-68.

HAROLD, D., ABRAHAM, R., HOLLINGWORTH, P., SIMS, R., GERRISH, A., HAMSHERE, M. L., PAHWA, J. S., MOSKVINA, V., DOWZELL, K., WILLIAMS, A., JONES, N., THOMAS, C., STRETTON, A., MORGAN, A. R., LOVESTONE, S., POWELL, J., PROITSI, P., LUPTON, M. K., BRAYNE, C., RUBINSZTEIN, D. C., GILL, M., LAWLOR, B., LYNCH, A., MORGAN, K., BROWN, K. S., PASSMORE, P. A., CRAIG, D., MCGUINNESS, B., TODD, S., HOLMES, C., MANN, D., SMITH, A. D., LOVE, S., KEHOE, P. G., HARDY, J., MEAD, S., FOX, N., ROSSOR, M., COLLINGE, J., MAIER, W., JESSEN, F., SCHURMANN, B., HEUN, R., VAN DEN BUSSCHE, H., HEUSER, I., KORNHUBER, J., WILTFANG, J.,

DICHGANS, M., FROLICH, L., HAMPEL, H., HULL, M., RUJESCU, D., GOATE, A. M., KAUWE, J. S., CRUCHAGA, C., NOWOTNY, P., MORRIS, J. C., MAYO, K., SLEEGERS, K., BETTENS, K., ENGELBORGHS, S., DE DEYN, P. P., VAN BROECKHOVEN, C., LIVINGSTON, G., BASS, N. J., GURLING, H., MCQUILLIN, A., GWILLIAM, R., DELOUKAS, P., AL-CHALABI, A., SHAW, C. E., TSOLAKI, M., SINGLETON, A. B., GUERREIRO, R., MUHLEISEN, T. W., NOTHEN, M. M., MOEBUS, S., JOCKEL, K. H., KLOPP, N., WICHMANN, H. E., CARRASQUILLO, M. M., PANKRATZ, V. S., YOUNKIN, S. G., HOLMANS, P. A., O'DONOVAN, M., OWEN, M. J. & WILLIAMS, J. 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet,* 41**,** 1088-93.

HASTIE, T., FRIEDMAN 2009. *The Elements of Statistical Learning: Data Mining, Inference and prediction*, Springer.

HENN, B. M., BOTIGUE, L. R., BUSTAMANTE, C. D., CLARK, A. G. & GRAVEL, S. 2015. Estimating the mutation load in human genomes. *Nat Rev Genet,* 16**,** 333-43.

HENNAH, W., THOMSON, P., MCQUILLIN, A., BASS, N., LOUKOLA, A., ANJORIN, A., BLACKWOOD, D., CURTIS, D., DEARY, I. J., HARRIS, S. E., ISOMETSA, E. T., LAWRENCE, J., LONNQVIST, J., MUIR, W., PALOTIE, A., PARTONEN, T., PAUNIO, T., PYLKKO, E., ROBINSON, M., SORONEN, P., SUOMINEN, K., SUVISAARI, J., THIRUMALAI, S., ST CLAIR, D., GURLING, H., PELTONEN, L. & PORTEOUS, D. 2009. DISC1 association, heterogeneity and interplay in schizophrenia and bipolar disorder. *Mol Psychiatry,* 14**,** 865-73.

HERMEY, G. 2009. The Vps10p-domain receptor family. *Cell Mol Life Sci,* 66**,** 2677-89.

HERRERA, D. G. & ROBERTSON, H. A. 1996. Activation of c-fos in the brain. *Prog Neurobiol,* 50**,** 83-107.

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., FIELDS, S. & STAMATOYANNOPOULOS, J. A. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods,* 6**,** 283-9.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. & MANOLIO, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

HOENICKA, J., ARAGUES, M., PONCE, G., RODRIGUEZ-JIMENEZ, R., JIMENEZ-ARRIERO, M. A. & PALOMO, T. 2007. From dopaminergic genes to psychiatric disorders. *Neurotox Res,* 11**,** 61-72.

HOLLINGWORTH, P., HAROLD, D., SIMS, R., GERRISH, A., LAMBERT, J. C., CARRASQUILLO, M. M., ABRAHAM, R., HAMSHERE, M. L., PAHWA, J. S., MOSKVINA, V., DOWZELL, K., JONES, N., STRETTON, A., THOMAS, C., RICHARDS, A., IVANOV, D., WIDDOWSON, C., CHAPMAN, J., LOVESTONE, S., POWELL, J., PROITSI, P., LUPTON, M. K., BRAYNE, C., RUBINSZTEIN, D. C., GILL, M., LAWLOR, B., LYNCH, A., BROWN, K. S., PASSMORE, P. A., CRAIG, D., MCGUINNESS, B., TODD, S., HOLMES, C., MANN, D., SMITH, A. D., BEAUMONT, H., WARDEN, D., WILCOCK, G., LOVE, S., KEHOE, P. G., HOOPER, N. M., VARDY, E. R., HARDY, J., MEAD, S., FOX, N. C., ROSSOR, M., COLLINGE, J., MAIER, W., JESSEN, F., RUTHER, E., SCHURMANN, B., HEUN, R., KOLSCH, H., VAN DEN BUSSCHE, H., HEUSER, I., KORNHUBER, J., WILTFANG, J., DICHGANS, M., FROLICH, L., HAMPEL, H., GALLACHER, J., HULL, M., RUJESCU, D., GIEGLING, I., GOATE, A. M., KAUWE, J. S., CRUCHAGA, C., NOWOTNY, P., MORRIS, J. C., MAYO, K., SLEEGERS, K., BETTENS, K., ENGELBORGHS, S., DE DEYN, P. P., VAN BROECKHOVEN, C., LIVINGSTON, G., BASS, N. J., GURLING, H., MCQUILLIN, A., GWILLIAM, R., DELOUKAS, P., AL-CHALABI, A., SHAW, C. E., TSOLAKI, M., SINGLETON, A. B., GUERREIRO, R., MUHLEISEN, T. W., NOTHEN, M. M., MOEBUS, S., JOCKEL, K. H., KLOPP, N., WICHMANN, H. E., PANKRATZ, V. S., SANDO, S. B., AASLY, J. O., BARCIKOWSKA, M., WSZOLEK, Z. K., DICKSON, D. W., GRAFF-RADFORD, N. R., PETERSEN, R. C., et al. 2011. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet,* 43**,** 429-35.

HOWRIGAN, D. P., LAIRD, N. M., SMOLLER, J. W., DEVLIN, B. & MCQUEEN, M. B. 2011. Using linkage information to weight a genome-wide association of bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet,* 156B**,** 462-71.

HSU, C., LIN. 2010. *A Practical Guide to Support Vector Classification* [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [Accessed March 2014.

HUBER, W., CAREY, V. J., GENTLEMAN, R., ANDERS, S., CARLSON, M., CARVALHO, B. S., BRAVO, H. C., DAVIS, S., GATTO, L., GIRKE, T., GOTTARDO, R., HAHNE, F., HANSEN, K. D., IRIZARRY, R. A., LAWRENCE, M., LOVE, M. I., MACDONALD, J., OBENCHAIN, V., OLES, A. K., PAGES, H., REYES, A., SHANNON, P., SMYTH, G. K., TENENBAUM, D., WALDRON, L. & MORGAN, M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods,* 12**,** 115-21.

IHAKA, G. 1996.

R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics,* 5**,** 299-314.

INTERNATIONAL SCHIZOPHRENIA, C. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature,* 455**,** 237-41.

INTERNATIONAL SCHIZOPHRENIA, C., PURCELL, S. M., WRAY, N. R., STONE, J. L., VISSCHER, P. M., O'DONOVAN, M. C., SULLIVAN, P. F. & SKLAR, P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature,* 460**,** 748-52.

IOSSIFOV, I., O'ROAK, B. J., SANDERS, S. J., RONEMUS, M., KRUMM, N., LEVY, D., STESSMAN, H. A., WITHERSPOON, K. T., VIVES, L., PATTERSON, K. E., SMITH, J. D., PAEPER, B., NICKERSON, D. A., DEA, J., DONG, S., GONZALEZ, L. E., MANDELL, J. D., MANE, S. M., MURTHA, M. T., SULLIVAN, C. A., WALKER, M. F., WAQAR, Z., WEI, L., WILLSEY, A. J., YAMROM, B., LEE, Y. H., GRABOWSKA, E., DALKIC, E., WANG, Z., MARKS, S., ANDREWS, P., LEOTTA, A., KENDALL, J., HAKKER, I., ROSENBAUM, J., MA, B., RODGERS, L., TROGE, J., NARZISI, G., YOON, S., SCHATZ, M. C., YE, K., MCCOMBIE, W. R., SHENDURE, J., EICHLER, E. E., STATE, M. W. & WIGLER, M. 2014. The contribution of de novo coding mutations to autism spectrum disorder. *Nature,* 515**,** 216-21.

ITOH, A., MIYABAYASHI, T., OHNO, M. & SAKANO, S. 1998. Cloning and expressions of three mammalian homologues of Drosophila slit suggest possible roles for Slit in the formation and maintenance of the nervous system. *Brain Res Mol Brain Res,* 62**,** 175-86.

JANSEN, P., GIEHL, K., NYENGAARD, J. R., TENG, K., LIOUBINSKI, O., SJOEGAARD, S. S., BREIDERHOFF, T., GOTTHARDT, M., LIN, F., EILERS, A., PETERSEN, C. M., LEWIN, G. R., HEMPSTEAD, B. L., WILLNOW, T. E. & NYKJAER, A. 2007. Roles for the pro-neurotrophin receptor sortilin in neuronal development, aging and brain injury. *Nat Neurosci,* 10**,** 1449-57.

JOHAR, A. S., MASTRONARDI, C., ROJAS-VILLARRAGA, A., PATEL, H. R., CHUAH, A., PENG, K., HIGGINS, A., MILBURN, P., PALMER, S., SILVA-LARA, M. F., VELEZ, J. I., ANDREWS, D., FIELD, M., HUTTLEY, G., GOODNOW, C., ANAYA, J. M. & ARCOS-BURGOS, M. 2015. Novel and rare functional genomic variants in multiple autoimmune syndrome and Sjogren's syndrome. *J Transl Med,* 13**,** 173.

JONSSON, T., STEFANSSON, H., STEINBERG, S., JONSDOTTIR, I., JONSSON, P. V., SNAEDAL, J., BJORNSSON, S., HUTTENLOCHER, J., LEVEY, A. I., LAH, J. J., RUJESCU, D., HAMPEL, H., GIEGLING, I., ANDREASSEN, O. A., ENGEDAL, K., ULSTEIN, I., DJUROVIC, S., IBRAHIM-VERBAAS, C., HOFMAN, A., IKRAM, M. A., VAN DUIJN, C. M., THORSTEINSDOTTIR, U., KONG, A. & STEFANSSON, K. 2013. Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med,* 368**,** 107-16.

JOSHI, G. & WILENS, T. 2009. Comorbidity in pediatric bipolar disorder. *Child Adolesc Psychiatr Clin N Am,* 18**,** 291-319, vii-viii.

JUDY, J. T., SEIFUDDIN, F., PIROOZNIA, M., MAHON, P. B., BIPOLAR GENOME STUDY, C., JANCIC, D., GOES, F. S., SCHULZE, T., CICHON, S., NOETHEN, M., RIETSCHEL, M., DEPAULO, J. R., JR., POTASH, J. B. & ZANDI, P. P. 2013. Converging Evidence for Epistasis between ANK3 and Potassium Channel Gene KCNQ2 in Bipolar Disorder. *Front Genet,* 4**,** 87.

KAO, W. P., YANG, C. Y., SU, T. W., WANG, Y. T., LO, Y. C. & LIN, S. C. 2015. The versatile roles of CARDs in regulating apoptosis, inflammation, and NF-kappaB signaling. *Apoptosis,* 20**,** 174-95.

KARAYIORGOU, M., MORRIS, M. A., MORROW, B., SHPRINTZEN, R. J., GOLDBERG, R., BORROW, J., GOS, A., NESTADT, G., WOLYNIEC, P. S., LASSETER, V. K. & ET AL. 1995. Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc Natl Acad Sci U S A,* 92**,** 7612-6.

KAROLCHIK, D., HINRICHS, A. S. & KENT, W. J. 2011. The UCSC Genome Browser. *Curr Protoc Hum Genet,* Chapter 18**,** Unit18 6.

KATHIRESAN, S., WILLER, C. J., PELOSO, G. M., DEMISSIE, S., MUSUNURU, K., SCHADT, E. E., KAPLAN, L., BENNETT, D., LI, Y., TANAKA, T., VOIGHT, B. F., BONNYCASTLE, L. L., JACKSON, A. U., CRAWFORD, G., SURTI, A., GUIDUCCI, C., BURTT, N. P., PARISH, S., CLARKE, R., ZELENIKA, D., KUBALANZA, K. A., MORKEN, M. A., SCOTT, L. J., STRINGHAM, H. M., GALAN, P., SWIFT, A. J., KUUSISTO, J., BERGMAN, R. N., SUNDVALL, J., LAAKSO, M., FERRUCCI, L., SCHEET, P., SANNA, S., UDA, M., YANG, Q., LUNETTA, K. L., DUPUIS, J., DE BAKKER, P. I., O'DONNELL, C. J., CHAMBERS, J. C., KOONER, J. S., HERCBERG, S., MENETON, P., LAKATTA, E. G., SCUTERI, A., SCHLESSINGER, D., TUOMILEHTO, J., COLLINS, F. S., GROOP, L., ALTSHULER, D., COLLINS, R., LATHROP, G. M., MELANDER, O., SALOMAA, V., PELTONEN, L., ORHO-MELANDER, M., ORDOVAS, J. M., BOEHNKE, M., ABECASIS, G. R., MOHLKE, K. L. & CUPPLES, L. A. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet,* 41**,** 56-65.

KAVANAGH, D. H., TANSEY, K. E., O'DONOVAN, M. C. & OWEN, M. J. 2015. Schizophrenia genetics: emerging themes for a complex disorder. *Mol Psychiatry,* 20**,** 72-6.

KEINAN, A. & CLARK, A. G. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science,* 336**,** 740-3.

KELLIS, M., WOLD, B., SNYDER, M. P., BERNSTEIN, B. E., KUNDAJE, A., MARINOV, G. K., WARD, L. D., BIRNEY, E., CRAWFORD, G. E., DEKKER, J., DUNHAM, I., ELNITSKI, L. L., FARNHAM, P. J., FEINGOLD, E. A., GERSTEIN, M., GIDDINGS, M. C., GILBERT, D. M., GINGERAS, T. R., GREEN, E. D., GUIGO, R., HUBBARD, T., KENT, J., LIEB, J. D., MYERS, R. M., PAZIN, M. J., REN, B., STAMATOYANNOPOULOS, J. A., WENG, Z., WHITE, K. P. &

HARDISON, R. C. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A,* 111**,** 6131-8.

KENDLER, K. S., GATZ, M., GARDNER, C. O. & PEDERSEN, N. L. 2005. Age at onset and familial risk for major depression in a Swedish national twin sample. *Psychol Med,* 35**,** 1573-9.

KERNER, B., RAO, A. R., CHRISTENSEN, B., DANDEKAR, S., YOURSHAW, M. & NELSON, S. F. 2013. Rare Genomic Variants Link Bipolar Disorder with Anxiety Disorders to CREB-Regulated Intracellular Signaling Pathways. *Front Psychiatry,* 4**,** 154.

KESSLER, R. C., BERGLUND, P., DEMLER, O., JIN, R., MERIKANGAS, K. R. & WALTERS, E. E. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry,* 62**,** 593-602.

KHURANA, E., FU, Y., COLONNA, V., MU, X. J., KANG, H. M., LAPPALAINEN, T., SBONER, A., LOCHOVSKY, L., CHEN, J., HARMANCI, A., DAS, J., ABYZOV, A., BALASUBRAMANIAN, S., BEAL, K., CHAKRAVARTY, D., CHALLIS, D., CHEN, Y., CLARKE, D., CLARKE, L., CUNNINGHAM, F., EVANI, U. S., FLICEK, P., FRAGOZA, R., GARRISON, E., GIBBS, R., GUMUS, Z. H., HERRERO, J., KITABAYASHI, N., KONG, Y., LAGE, K., LILUASHVILI, V., LIPKIN, S. M., MACARTHUR, D. G., MARTH, G., MUZNY, D., PERS, T. H., RITCHIE, G. R., ROSENFELD, J. A., SISU, C., WEI, X., WILSON, M., XUE, Y., YU, F., GENOMES PROJECT, C., DERMITZAKIS, E. T., YU, H., RUBIN, M. A., TYLER-SMITH, C. & GERSTEIN, M. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science,* 342**,** 1235587.

KING, D. C., TAYLOR, J., ELNITSKI, L., CHIAROMONTE, F., MILLER, W. & HARDISON, R. C. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res,* 15**,** 1051-60.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet,* 46**,** 310-5.

KNIGHT, J. C. 2014. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Med,* 6**,** 92.

KOHLI, M. A., LUCAE, S., SAEMANN, P. G., SCHMIDT, M. V., DEMIRKAN, A., HEK, K., CZAMARA, D., ALEXANDER, M., SALYAKINA, D., RIPKE, S., HOEHN, D., SPECHT, M., MENKE, A., HENNINGS, J., HECK, A., WOLF, C., ISING, M., SCHREIBER, S., CZISCH, M., MULLER, M. B., UHR, M., BETTECKEN, T., BECKER, A., SCHRAMM, J., RIETSCHEL, M., MAIER, W., BRADLEY, B., RESSLER, K. J., NOTHEN, M. M., CICHON, S., CRAIG, I. W., BREEN, G., LEWIS, C. M., HOFMAN, A., TIEMEIER, H., VAN DUIJN, C. M., HOLSBOER, F.,

MULLER-MYHSOK, B. & BINDER, E. B. 2011. The neuronal transporter gene SLC6A15 confers risk to major depression. *Neuron,* 70**,** 252-65.

KOLBE, D., TAYLOR, J., ELNITSKI, L., ESWARA, P., LI, J., MILLER, W., HARDISON, R. & CHIAROMONTE, F. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res,* 14**,** 700-7.

KRZYWINSKI, M. & ALTMAN, N. 2014. Points of significance: two-factor designs. *Nat Methods,* 11**,** 1187-8.

KUHN, R. M., HAUSSLER, D. & KENT, W. J. 2013. The UCSC genome browser and associated tools. *Brief Bioinform,* 14**,** 144-61.

KUMAR, V., WESTRA, H. J., KARJALAINEN, J., ZHERNAKOVA, D. V., ESKO, T., HRDLICKOVA, B., ALMEIDA, R., ZHERNAKOVA, A., REINMAA, E., VOSA, U., HOFKER, M. H., FEHRMANN, R. S., FU, J., WITHOFF, S., METSPALU, A., FRANKE, L. & WIJMENGA, C. 2013. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet,* 9**,** e1003201.

LAM, H. Y., CLARK, M. J., CHEN, R., CHEN, R., NATSOULIS, G., O'HUALLACHAIN, M., DEWEY, F. E., HABEGGER, L., ASHLEY, E. A., GERSTEIN, M. B., BUTTE, A. J., JI, H. P. & SNYDER, M. 2012. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol,* 30**,** 78-82.

LAMBERT, J. C., HEATH, S., EVEN, G., CAMPION, D., SLEEGERS, K., HILTUNEN, M., COMBARROS, O., ZELENIKA, D., BULLIDO, M. J., TAVERNIER, B., LETENNEUR, L., BETTENS, K., BERR, C., PASQUIER, F., FIEVET, N., BARBERGER-GATEAU, P., ENGELBORGHS, S., DE DEYN, P., MATEO, I., FRANCK, A., HELISALMI, S., PORCELLINI, E., HANON, O., EUROPEAN ALZHEIMER'S DISEASE INITIATIVE, I., DE PANCORBO, M. M., LENDON, C., DUFOUIL, C., JAILLARD, C., LEVEILLARD, T., ALVAREZ, V., BOSCO, P., MANCUSO, M., PANZA, F., NACMIAS, B., BOSSU, P., PICCARDI, P., ANNONI, G., SERIPA, D., GALIMBERTI, D., HANNEQUIN, D., LICASTRO, F., SOININEN, H., RITCHIE, K., BLANCHE, H., DARTIGUES, J. F., TZOURIO, C., GUT, I., VAN BROECKHOVEN, C., ALPEROVITCH, A., LATHROP, M. & AMOUYEL, P. 2009. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet,* 41**,** 1094-9.

LANDER, E. & KRUGLYAK, L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet,* 11**,** 241-7.

LANDRUM, M. J., LEE, J. M., RILEY, G. R., JANG, W., RUBINSTEIN, W. S., CHURCH, D. M. & MAGLOTT, D. R. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res,* 42**,** D980-5.

LANDSBERG, L. & YOUNG, J. B. 1985. Insulin-mediated glucose metabolism in the relationship between dietary intake and sympathetic nervous system activity. *Int J Obes,* 9 Suppl 2**,** 63-8.

LANE, R. F., ST GEORGE-HYSLOP, P., HEMPSTEAD, B. L., SMALL, S. A., STRITTMATTER, S. M. & GANDY, S. 2012. Vps10 family proteins and the retromer complex in aging-related neurodegeneration and diabetes. *J Neurosci,* 32**,** 14080-6.

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., FERREIRA, T., WOOD, A. R., WEYANT, R. J., SEGRE, A. V., SPELIOTES, E. K., WHEELER, E., SORANZO, N., PARK, J. H., YANG, J., GUDBJARTSSON, D., HEARD-COSTA, N. L., RANDALL, J. C., QI, L., VERNON SMITH, A., MAGI, R., PASTINEN, T., LIANG, L., HEID, I. M., LUAN, J., THORLEIFSSON, G., WINKLER, T. W., GODDARD, M. E., SIN LO, K., PALMER, C., WORKALEMAHU, T., AULCHENKO, Y. S., JOHANSSON, A., ZILLIKENS, M. C., FEITOSA, M. F., ESKO, T., JOHNSON, T., KETKAR, S., KRAFT, P., MANGINO, M., PROKOPENKO, I., ABSHER, D., ALBRECHT, E., ERNST, F., GLAZER, N. L., HAYWARD, C., HOTTENGA, J. J., JACOBS, K. B., KNOWLES, J. W., KUTALIK, Z., MONDA, K. L., POLASEK, O., PREUSS, M., RAYNER, N. W., ROBERTSON, N. R., STEINTHORSDOTTIR, V., TYRER, J. P., VOIGHT, B. F., WIKLUND, F., XU, J., ZHAO, J. H., NYHOLT, D. R., PELLIKKA, N., PEROLA, M., PERRY, J. R., SURAKKA, I., TAMMESOO, M. L., ALTMAIER, E. L., AMIN, N., ASPELUND, T., BHANGALE, T., BOUCHER, G., CHASMAN, D. I., CHEN, C., COIN, L., COOPER, M. N., DIXON, A. L., GIBSON, Q., GRUNDBERG, E., HAO, K., JUHANI JUNTTILA, M., KAPLAN, L. M., KETTUNEN, J., KONIG, I. R., KWAN, T., LAWRENCE, R. W., LEVINSON, D. F., LORENTZON, M., MCKNIGHT, B., MORRIS, A. P., MULLER, M., SUH NGWA, J., PURCELL, S., RAFELT, S., SALEM, R. M., SALVI, E., et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* 467**,** 832-8.

LARUELLE, M. 2014. Schizophrenia: from dopaminergic to glutamatergic interventions. *Curr Opin Pharmacol,* 14**,** 97-102.

LE HELLARD, S., LEE, A. J., UNDERWOOD, S., THOMSON, P. A., MORRIS, S. W., TORRANCE, H. S., ANDERSON, S. M., ADAMS, R. R., NAVARRO, P., CHRISTOFOROU, A., HOULIHAN, L. M., DETERA-WADLEIGH, S., OWEN, M. J., ASHERSON, P., MUIR, W. J., BLACKWOOD, D. H., WRAY, N. R., PORTEOUS, D. J. & EVANS, K. L. 2007. Haplotype analysis and a novel allele-sharing method refines a chromosome 4p locus linked to bipolar affective disorder. *Biol Psychiatry,* 61**,** 797-805.

LEE, D., GORKIN, D. U., BAKER, M., STROBER, B. J., ASONI, A. L., MCCALLION, A. S. & BEER, M. A. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet,* 47**,** 955-61.

LEE, J. A. & LUPSKI, J. R. 2006. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron,* 52**,** 103-21.

LEE, M. T., CHEN, C. H., LEE, C. S., CHEN, C. C., CHONG, M. Y., OUYANG, W. C., CHIU, N. Y., CHUO, L. J., CHEN, C. Y., TAN, H. K., LANE, H. Y., CHANG, T. J., LIN, C. H., JOU, S. H., HOU, Y. M., FENG, J., LAI, T. J., TUNG, C. L., CHEN, T. J., CHANG, C. J., LUNG, F. W., CHEN, C. K., SHIAH, I. S., LIU, C. Y., TENG, P. R., CHEN, K. H., SHEN, L. J., CHENG, C. S., CHANG, T. P., LI, C. F., CHOU, C. H., CHEN, C. Y., WANG, K. H., FANN, C. S., WU, J. Y., CHEN, Y. T. & CHENG, A. T. 2011. Genome-wide association study of bipolar I disorder in the Han Chinese population. *Mol Psychiatry,* 16**,** 548-56.

LEE, P. H. & SHATKAY, H. 2008. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res,* 36**,** D820-4.

LEE, S. H., DECANDIA, T. R., RIPKE, S., YANG, J., SCHIZOPHRENIA PSYCHIATRIC GENOME-WIDE ASSOCIATION STUDY, C., INTERNATIONAL SCHIZOPHRENIA, C., MOLECULAR GENETICS OF SCHIZOPHRENIA, C., SULLIVAN, P. F., GODDARD, M. E., KELLER, M. C., VISSCHER, P. M. & WRAY, N. R. 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet,* 44**,** 247-50.

LEMIRE, M., ZAIDI, S. H., BAN, M., GE, B., AISSI, D., GERMAIN, M., KASSAM, I., WANG, M., ZANKE, B. W., GAGNON, F., MORANGE, P. E., TREGOUET, D. A., WELLS, P. S., SAWCER, S., GALLINGER, S., PASTINEN, T. & HUDSON, T. J. 2015. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun,* 6**,** 6326.

LEVINSON, D. F., MOSTAFAVI, S., MILANESCHI, Y., RIVERA, M., RIPKE, S., WRAY, N. R. & SULLIVAN, P. F. 2014. Genetic studies of major depressive disorder: why are there no genome-wide association study findings and what can we do about it? *Biol Psychiatry,* 76**,** 510-2.

LEVY, D., RONEMUS, M., YAMROM, B., LEE, Y. H., LEOTTA, A., KENDALL, J., MARKS, S., LAKSHMI, B., PAI, D., YE, K., BUJA, A., KRIEGER, A., YOON, S., TROGE, J., RODGERS, L., IOSSIFOV, I. & WIGLER, M. 2011. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron,* 70**,** 886-97.

LEWIS, C. M., NG, M. Y., BUTLER, A. W., COHEN-WOODS, S., UHER, R., PIRLO, K., WEALE, M. E., SCHOSSER, A., PAREDES, U. M., RIVERA, M., CRADDOCK, N., OWEN, M. J., JONES, L., JONES, I., KORSZUN, A., AITCHISON, K. J., SHI, J., QUINN, J. P., MACKENZIE, A., VOLLENWEIDER, P., WAEBER, G., HEATH, S., LATHROP, M., MUGLIA, P., BARNES, M. R., WHITTAKER, J. C., TOZZI, F., HOLSBOER, F., PREISIG, M., FARMER, A. E., BREEN, G., CRAIG, I. W. & MCGUFFIN, P. 2010. Genome-wide association study of major recurrent depression in the U.K. population. *Am J Psychiatry,* 167**,** 949-57.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, M. J., WANG, L. Y., XIA, Z., SHAM, P. C. & WANG, J. 2013. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res,* 41**,** W150-8.

LI, M. J., YAN, B., SHAM, P. C. & WANG, J. 2015. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief Bioinform,* 16**,** 393-412.

LI, M. X., GUI, H. S., KWAN, J. S., BAO, S. Y. & SHAM, P. C. 2012. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res,* 40**,** e53.

LI, X. & MONTGOMERY, S. B. 2013. Detection and impact of rare regulatory variants in human disease. *Front Genet,* 4**,** 67.

LIAO, S. & VON DER WEID, P. Y. 2015. Lymphatic system: an active pathway for immune protection. *Semin Cell Dev Biol,* 38**,** 83-9.

LICHTENSTEIN, P., YIP, B. H., BJORK, C., PAWITAN, Y., CANNON, T. D., SULLIVAN, P. F. & HULTMAN, C. M. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet,* 373**,** 234-9.

LIPPERT, C., LISTGARTEN, J., DAVIDSON, R. I., BAXTER, S., POON, H., KADIE, C. M. & HECKERMAN, D. 2013. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Sci Rep,* 3**,** 1099.

LIU, D. & ALBERT, P. S. 2014. Combination of longitudinal biomarkers in predicting binary events. *Biostatistics,* 15**,** 706-18.

LIU, L., SABO, A., NEALE, B. M., NAGASWAMY, U., STEVENS, C., LIM, E., BODEA, C. A., MUZNY, D., REID, J. G., BANKS, E., COON, H., DEPRISTO, M., DINH, H., FENNEL, T., FLANNICK, J., GABRIEL, S., GARIMELLA, K., GROSS, S., HAWES, A., LEWIS, L., MAKAROV, V., MAGUIRE, J., NEWSHAM, I., POPLIN, R., RIPKE, S., SHAKIR, K., SAMOCHA, K. E., WU, Y., BOERWINKLE, E., BUXBAUM, J. D., COOK, E. H., JR., DEVLIN, B., SCHELLENBERG, G. D., SUTCLIFFE, J. S., DALY, M. J., GIBBS, R. A. & ROEDER, K. 2013. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet,* 9**,** e1003443.

LIU, Y., XU, H., CHEN, S., CHEN, X., ZHANG, Z., ZHU, Z., QIN, X., HU, L., ZHU, J., ZHAO, G. P. & KONG, X. 2011. Genome-wide interaction-based association analysis identified multiple new susceptibility Loci for common diseases. *PLoS Genet,* 7**,** e1001338.

LODGE, D. J. & GRACE, A. A. 2011. Hippocampal dysregulation of dopamine system function and the pathophysiology of schizophrenia. *Trends Pharmacol Sci,* 32**,** 507-13.

MA, M., RU, Y., CHUANG, L. S., HSU, N. Y., SHI, L. S., HAKENBERG, J., CHENG, W. Y., UZILOV, A., DING, W., GLICKSBERG, B. S. & CHEN, R. 2015. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics,* 16 Suppl 8**,** S3.

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., BARRETT, J. C., BIESECKER, L. G., CONRAD, D. F., COOPER, G. M., COX, N. J., DALY, M. J., GERSTEIN, M. B., GOLDSTEIN, D. B., HIRSCHHORN, J. N., LEAL, S. M., PENNACCHIO, L. A., STAMATOYANNOPOULOS, J. A., SUNYAEV, S. R., VALLE, D., VOIGHT, B. F., WINCKLER, W. & GUNTER, C. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature,* 508**,** 469-76.

MACISAAC, K. D., LO, K. A., GORDON, W., MOTOLA, S., MAZOR, T. & FRAENKEL, E. 2010. A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLoS Comput Biol,* 6**,** e1000773.

MADRIGAL, P. & KRAJEWSKI, P. 2012. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front Genet,* 3**,** 230.

MAES, M., KUBERA, M. & LEUNIS, J. C. 2008. The gut-brain barrier in major depression: intestinal mucosal dysfunction with an increased translocation of LPS from gram negative enterobacteria (leaky gut) plays a role in the inflammatory pathophysiology of depression. *Neuro Endocrinol Lett,* 29**,** 117-24.

MAIER, R., MOSER, G., CHEN, G. B., RIPKE, S., CROSS-DISORDER WORKING GROUP OF THE PSYCHIATRIC GENOMICS, C., CORYELL, W., POTASH, J. B., SCHEFTNER, W. A., SHI, J., WEISSMAN, M. M., HULTMAN, C. M., LANDEN, M., LEVINSON, D. F., KENDLER, K. S., SMOLLER, J. W., WRAY, N. R. & LEE, S. H. 2015. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet,* 96**,** 283-94.

MAJOR DEPRESSIVE DISORDER WORKING GROUP OF THE PSYCHIATRIC, G. C., RIPKE, S., WRAY, N. R., LEWIS, C. M., HAMILTON, S. P., WEISSMAN, M. M., BREEN, G., BYRNE, E. M., BLACKWOOD, D. H., BOOMSMA, D. I., CICHON, S., HEATH, A. C., HOLSBOER, F., LUCAE, S., MADDEN, P. A., MARTIN, N. G., MCGUFFIN, P., MUGLIA, P., NOETHEN, M. M., PENNINX, B. P., PERGADIA, M. L., POTASH, J. B., RIETSCHEL, M., LIN, D., MULLER-MYHSOK, B., SHI, J., STEINBERG, S., GRABE, H. J., LICHTENSTEIN,

P., MAGNUSSON, P., PERLIS, R. H., PREISIG, M., SMOLLER, J. W., STEFANSSON, K., UHER, R., KUTALIK, Z., TANSEY, K. E., TEUMER, A., VIKTORIN, A., BARNES, M. R., BETTECKEN, T., BINDER, E. B., BREUER, R., CASTRO, V. M., CHURCHILL, S. E., CORYELL, W. H., CRADDOCK, N., CRAIG, I. W., CZAMARA, D., DE GEUS, E. J., DEGENHARDT, F., FARMER, A. E., FAVA, M., FRANK, J., GAINER, V. S., GALLAGHER, P. J., GORDON, S. D., GORYACHEV, S., GROSS, M., GUIPPONI, M., HENDERS, A. K., HERMS, S., HICKIE, I. B., HOEFELS, S., HOOGENDIJK, W., HOTTENGA, J. J., IOSIFESCU, D. V., ISING, M., JONES, I., JONES, L., JUNG-YING, T., KNOWLES, J. A., KOHANE, I. S., KOHLI, M. A., KORSZUN, A., LANDEN, M., LAWSON, W. B., LEWIS, G., MACINTYRE, D., MAIER, W., MATTHEISEN, M., MCGRATH, P. J., MCINTOSH, A., MCLEAN, A., MIDDELDORP, C. M., MIDDLETON, L., MONTGOMERY, G. M., MURPHY, S. N., NAUCK, M., NOLEN, W. A., NYHOLT, D. R., O'DONOVAN, M., OSKARSSON, H., PEDERSEN, N., SCHEFTNER, W. A., SCHULZ, A., SCHULZE, T. G., SHYN, S. I., SIGURDSSON, E., SLAGER, S. L., et al. 2013. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry,* 18**,** 497-511.

MAKAROV, V., O'GRADY, T., CAI, G., LIHM, J., BUXBAUM, J. D. & YOON, S. 2012. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics,* 28**,** 724-5.

MALHOTRA, D., MCCARTHY, S., MICHAELSON, J. J., VACIC, V., BURDICK, K. E., YOON, S., CICHON, S., CORVIN, A., GARY, S., GERSHON, E. S., GILL, M., KARAYIORGOU, M., KELSOE, J. R., KRASTOSHEVSKY, O., KRAUSE, V., LEIBENLUFT, E., LEVY, D. L., MAKAROV, V., BHANDARI, A., MALHOTRA, A. K., MCMAHON, F. J., NOTHEN, M. M., POTASH, J. B., RIETSCHEL, M., SCHULZE, T. G. & SEBAT, J. 2011. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron,* 72**,** 951-63.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MARSHALL, C. R., NOOR, A., VINCENT, J. B., LIONEL, A. C., FEUK, L., SKAUG, J., SHAGO, M., MOESSNER, R., PINTO, D., REN, Y., THIRUVAHINDRAPDURAM, B., FIEBIG, A., SCHREIBER, S., FRIEDMAN, J., KETELAARS, C. E., VOS, Y. J., FICICIOGLU, C., KIRKPATRICK, S., NICOLSON, R., SLOMAN, L., SUMMERS, A., GIBBONS, C. A., TEEBI, A., CHITAYAT, D., WEKSBERG, R., THOMPSON, A., VARDY, C., CROSBIE, V., LUSCOMBE, S., BAATJES,

R., ZWAIGENBAUM, L., ROBERTS, W., FERNANDEZ, B., SZATMARI, P. & SCHERER, S. W. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet,* 82**,** 477-88.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H., BRODY, J., SHAFER, A., NERI, F., LEE, K., KUTYAVIN, T., STEHLING-SUN, S., JOHNSON, A. K., CANFIELD, T. K., GISTE, E., DIEGEL, M., BATES, D., HANSEN, R. S., NEPH, S., SABO, P. J., HEIMFELD, S., RAUBITSCHEK, A., ZIEGLER, S., COTSAPAS, C., SOTOODEHNIA, N., GLASS, I., SUNYAEV, S. R., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-5.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.

MCCLELLAN, J. & KING, M. C. 2010a. Genetic heterogeneity in human disease. *Cell,* 141**,** 210-7.

MCCLELLAN, J. & KING, M. C. 2010b. Genomic analysis of mental illness: a changing landscape. *JAMA,* 303**,** 2523-4.

MCCLELLAN, J. M., SUSSER, E. & KING, M. C. 2007. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry,* 190**,** 194-9.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

MCKEOWN, L., SWANTON, L., ROBINSON, P. & JONES, O. T. 2008. Surface expression and distribution of voltage-gated potassium channels in neurons (Review). *Mol Membr Biol,* 25**,** 332-43.

MCLAREN, W., PRITCHARD, B., RIOS, D., CHEN, Y., FLICEK, P. & CUNNINGHAM, F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26**,** 2069-70.

MEADER, S., PONTING, C. P. & LUNTER, G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res,* 20**,** 1335-43.

MERELLI, I., CALABRIA, A., COZZI, P., VITI, F., MOSCA, E. & MILANESI, L. 2013. SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. *BMC Bioinformatics,* 14 Suppl 1**,** S9.

MEYNERT, A. M., ANSARI, M., FITZPATRICK, D. R. & TAYLOR, M. S. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics,* 15**,** 247.

MILLAR, J. K., CHRISTIE, S., ANDERSON, S., LAWSON, D., HSIAO-WEI LOH, D., DEVON, R. S., ARVEILER, B., MUIR, W. J., BLACKWOOD, D. H. & PORTEOUS, D. J. 2001. Genomic structure and localisation within a linkage hotspot of Disrupted In Schizophrenia 1, a gene disrupted by a translocation segregating with schizophrenia. *Mol Psychiatry,* 6**,** 173-8.

MILLAR, J. K., CHRISTIE, S. & PORTEOUS, D. J. 2003. Yeast two-hybrid screens implicate DISC1 in brain development and function. *Biochem Biophys Res Commun,* 311**,** 1019-25.

MILLAR, J. K., WILSON-ANNAN, J. C., ANDERSON, S., CHRISTIE, S., TAYLOR, M. S., SEMPLE, C. A., DEVON, R. S., ST CLAIR, D. M., MUIR, W. J., BLACKWOOD, D. H. & PORTEOUS, D. J. 2000. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet,* 9**,** 1415-23.

MILLER, A. H. 2010. Depression and immunity: a role for T cells? *Brain Behav Immun,* 24**,** 1-8.

MIYATA, Y., FUKUHARA, A., OTSUKI, M. & SHIMOMURA, I. 2013. Expression of activating transcription factor 2 in inflammatory macrophages in obese adipose tissue. *Obesity (Silver Spring),* 21**,** 731-6.

MONTEN, C., GUDJONSDOTTIR, A. H., BROWALDH, L., ARNELL, H., NILSSON, S., AGARDH, D. & NALUAI, A. T. 2015. Genes involved in muscle contractility and nutrient signaling pathways within celiac disease risk loci show differential mRNA expression. *BMC Med Genet,* 16**,** 44.

MORELLE, J. & DEVUYST, O. 2015. Water and solute transport across the peritoneal membrane. *Curr Opin Nephrol Hypertens,* 24**,** 434-43.

MORTIMER, S. A., KIDWELL, M. A. & DOUDNA, J. A. 2014. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet,* 15**,** 469-79.

MUGLIA, P., TOZZI, F., GALWEY, N. W., FRANCKS, C., UPMANYU, R., KONG, X. Q., ANTONIADES, A., DOMENICI, E., PERRY, J., ROTHEN, S., VANDELEUR, C. L., MOOSER, V., WAEBER, G., VOLLENWEIDER, P., PREISIG, M., LUCAE, S., MULLER-MYHSOK, B., HOLSBOER, F., MIDDLETON, L. T. & ROSES, A. D. 2010. Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. *Mol Psychiatry,* 15**,** 589-601.

MUHLEISEN, T. W., LEBER, M., SCHULZE, T. G., STROHMAIER, J., DEGENHARDT, F., TREUTLEIN, J., MATTHEISEN, M., FORSTNER, A. J., SCHUMACHER, J., BREUER, R., MEIER, S., HERMS, S., HOFFMANN, P., LACOUR, A., WITT, S. H., REIF, A., MULLER-MYHSOK, B., LUCAE, S., MAIER, W., SCHWARZ, M., VEDDER, H., KAMMERER-CIERNIOCH, J., PFENNIG, A., BAUER, M.,

HAUTZINGER, M., MOEBUS, S., PRIEBE, L., CZERSKI, P. M., HAUSER, J., LISSOWSKA, J., SZESZENIA-DABROWSKA, N., BRENNAN, P., MCKAY, J. D., WRIGHT, A., MITCHELL, P. B., FULLERTON, J. M., SCHOFIELD, P. R., MONTGOMERY, G. W., MEDLAND, S. E., GORDON, S. D., MARTIN, N. G., KRASNOW, V., CHUCHALIN, A., BABADJANOVA, G., PANTELEJEVA, G., ABRAMOVA, L. I., TIGANOV, A. S., POLONIKOV, A., KHUSNUTDINOVA, E., ALDA, M., GROF, P., ROULEAU, G. A., TURECKI, G., LAPRISE, C., RIVAS, F., MAYORAL, F., KOGEVINAS, M., GRIGOROIU-SERBANESCU, M., PROPPING, P., BECKER, T., RIETSCHEL, M., NOTHEN, M. M. & CICHON, S. 2014. Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat Commun,* 5**,** 3339.

MULLANEY, J. M., MILLS, R. E., PITTARD, W. S. & DEVINE, S. E. 2010. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet,* 19**,** R131-6.

MURRAY, C. J. & LOPEZ, A. D. 1996. Evidence-based health policy--lessons from the Global Burden of Disease Study. *Science,* 274**,** 740-3.

MUSUNURU, K., STRONG, A., FRANK-KAMENETSKY, M., LEE, N. E., AHFELDT, T., SACHS, K. V., LI, X., LI, H., KUPERWASSER, N., RUDA, V. M., PIRRUCCELLO, J. P., MUCHMORE, B., PROKUNINA-OLSSON, L., HALL, J. L., SCHADT, E. E., MORALES, C. R., LUND-KATZ, S., PHILLIPS, M. C., WONG, J., CANTLEY, W., RACIE, T., EJEBE, K. G., ORHO-MELANDER, M., MELANDER, O., KOTELIANSKY, V., FITZGERALD, K., KRAUSS, R. M., COWAN, C. A., KATHIRESAN, S. & RADER, D. J. 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature,* 466**,** 714-9.

MYOUZEN, K., KOCHI, Y., SHIMANE, K., FUJIO, K., OKAMURA, T., OKADA, Y., SUZUKI, A., ATSUMI, T., ITO, S., TAKADA, K., MIMORI, A., IKEGAWA, S., YAMADA, R., NAKAMURA, Y. & YAMAMOTO, K. 2010. Regulatory polymorphisms in EGR2 are associated with susceptibility to systemic lupus erythematosus. *Hum Mol Genet,* 19**,** 2313-20.

NAGELHUS, E. A., MATHIISEN, T. M. & OTTERSEN, O. P. 2004. Aquaporin-4 in the central nervous system: cellular and subcellular distribution and coexpression with KIR4.1. *Neuroscience,* 129**,** 905-13.

NAJ, A. C., JUN, G., BEECHAM, G. W., WANG, L. S., VARDARAJAN, B. N., BUROS, J., GALLINS, P. J., BUXBAUM, J. D., JARVIK, G. P., CRANE, P. K., LARSON, E. B., BIRD, T. D., BOEVE, B. F., GRAFF-RADFORD, N. R., DE JAGER, P. L., EVANS, D., SCHNEIDER, J. A., CARRASQUILLO, M. M., ERTEKIN-TANER, N., YOUNKIN, S. G., CRUCHAGA, C., KAUWE, J. S., NOWOTNY, P., KRAMER, P., HARDY, J., HUENTELMAN, M. J., MYERS, A. J., BARMADA, M. M., DEMIRCI, F. Y., BALDWIN, C. T., GREEN, R. C., ROGAEVA, E., ST GEORGE-HYSLOP, P., ARNOLD, S. E., BARBER, R., BEACH, T., BIGIO, E. H.,

BOWEN, J. D., BOXER, A., BURKE, J. R., CAIRNS, N. J., CARLSON, C. S., CARNEY, R. M., CARROLL, S. L., CHUI, H. C., CLARK, D. G., CORNEVEAUX, J., COTMAN, C. W., CUMMINGS, J. L., DECARLI, C., DEKOSKY, S. T., DIAZ-ARRASTIA, R., DICK, M., DICKSON, D. W., ELLIS, W. G., FABER, K. M., FALLON, K. B., FARLOW, M. R., FERRIS, S., FROSCH, M. P., GALASKO, D. R., GANGULI, M., GEARING, M., GESCHWIND, D. H., GHETTI, B., GILBERT, J. R., GILMAN, S., GIORDANI, B., GLASS, J. D., GROWDON, J. H., HAMILTON, R. L., HARRELL, L. E., HEAD, E., HONIG, L. S., HULETTE, C. M., HYMAN, B. T., JICHA, G. A., JIN, L. W., JOHNSON, N., KARLAWISH, J., KARYDAS, A., KAYE, J. A., KIM, R., KOO, E. H., KOWALL, N. W., LAH, J. J., LEVEY, A. I., LIEBERMAN, A. P., LOPEZ, O. L., MACK, W. J., MARSON, D. C., MARTINIUK, F., MASH, D. C., MASLIAH, E., MCCORMICK, W. C., MCCURRY, S. M., MCDAVID, A. N., MCKEE, A. C., MESULAM, M., MILLER, B. L., et al. 2011. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet,* 43**,** 436-41.

NEED, A. C., MCEVOY, J. P., GENNARELLI, M., HEINZEN, E. L., GE, D., MAIA, J. M., SHIANNA, K. V., HE, M., CIRULLI, E. T., GUMBS, C. E., ZHAO, Q., CAMPBELL, C. R., HONG, L., ROSENQUIST, P., PUTKONEN, A., HALLIKAINEN, T., REPO-TIIHONEN, E., TIIHONEN, J., LEVY, D. L., MELTZER, H. Y. & GOLDSTEIN, D. B. 2012. Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet,* 91**,** 303-12.

NEPH, S., VIERSTRA, J., STERGACHIS, A. B., REYNOLDS, A. P., HAUGEN, E., VERNOT, B., THURMAN, R. E., JOHN, S., SANDSTROM, R., JOHNSON, A. K., MAURANO, M. T., HUMBERT, R., RYNES, E., WANG, H., VONG, S., LEE, K., BATES, D., DIEGEL, M., ROACH, V., DUNN, D., NERI, J., SCHAFER, A., HANSEN, R. S., KUTYAVIN, T., GISTE, E., WEAVER, M., CANFIELD, T., SABO, P., ZHANG, M., BALASUNDARAM, G., BYRON, R., MACCOSS, M. J., AKEY, J. M., BENDER, M. A., GROUDINE, M., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature,* 489**,** 83-90.

NETWORK & PATHWAY ANALYSIS SUBGROUP OF PSYCHIATRIC GENOMICS, C. 2015. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat Neurosci,* 18**,** 199-209.

NG, P. C. & HENIKOFF, S. 2001. Predicting deleterious amino acid substitutions. *Genome Res,* 11**,** 863-74.

NICA, A. C., PARTS, L., GLASS, D., NISBET, J., BARRETT, A., SEKOWSKA, M., TRAVERS, M., POTTER, S., GRUNDBERG, E., SMALL, K., HEDMAN, A. K., BATAILLE, V., TZENOVA BELL, J., SURDULESCU,

G., DIMAS, A. S., INGLE, C., NESTLE, F. O., DI MEGLIO, P., MIN, J. L., WILK, A., HAMMOND, C. J., HASSANALI, N., YANG, T. P., MONTGOMERY, S. B., O'RAHILLY, S., LINDGREN, C. M., ZONDERVAN, K. T., SORANZO, N., BARROSO, I., DURBIN, R., AHMADI, K., DELOUKAS, P., MCCARTHY, M. I., DERMITZAKIS, E. T., SPECTOR, T. D. & MU, T. C. 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet,* 7**,** e1002003.

NICOLAE, D. L., GAMAZON, E., ZHANG, W., DUAN, S., DOLAN, M. E. & COX, N. J. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet,* 6**,** e1000888.

NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet,* 12**,** 443-51.

NOTHEN, M. M., NIERATSCHKER, V., CICHON, S. & RIETSCHEL, M. 2010. New findings in the genetics of major psychoses. *Dialogues Clin Neurosci,* 12**,** 85-93.

NOVAK, G., SEEMAN, P. & TALLERICO, T. 2006. Increased expression of calcium/calmodulin-dependent protein kinase IIbeta in frontal cortex in schizophrenia and depression. *Synapse,* 59**,** 61-8.

NURNBERGER, J. I., JR., KOLLER, D. L., JUNG, J., EDENBERG, H. J., FOROUD, T., GUELLA, I., VAWTER, M. P., KELSOE, J. R. & PSYCHIATRIC GENOMICS CONSORTIUM BIPOLAR, G. 2014. Identification of pathways for bipolar disorder: a meta-analysis. *JAMA Psychiatry,* 71**,** 657-64.

O'LEARY, O. F., FELICE, D., GALIMBERTI, S., SAVIGNAC, H. M., BRAVO, J. A., CROWLEY, T., EL YACOUBI, M., VAUGEOIS, J. M., GASSMANN, M., BETTLER, B., DINAN, T. G. & CRYAN, J. F. 2014. GABAB(1) receptor subunit isoforms differentially regulate stress resilience. *Proc Natl Acad Sci U S A,* 111**,** 15232-7.

OLLILA, H. M., SORONEN, P., SILANDER, K., PALO, O. M., KIESEPPA, T., KAUNISTO, M. A., LONNQVIST, J., PELTONEN, L., PARTONEN, T. & PAUNIO, T. 2009. Findings from bipolar disorder genome-wide association studies replicate in a Finnish bipolar family-cohort. *Mol Psychiatry,* 14**,** 351-3.

OTT, J., WANG, J. & LEAL, S. M. 2015. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet,* 16**,** 275-84.

PELOSI, A. J., SYKES, R. A., LOUGH, J. R., MUIR, W. J. & DUNNIGAN, M. G. 1986. A psychiatric study of idiopathic oedema. *Lancet,* 2**,** 999-1002.

PENNACCHIO, L. A. & RUBIN, E. M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet,* 2**,** 100-9.

PEREZ-PALMA, E., BUSTOS, B. I., VILLAMAN, C. F., ALARCON, M. A., AVILA, M. E., UGARTE, G. D., REYES, A. E., OPAZO, C., DE FERRARI, G. V., ALZHEIMER'S DISEASE NEUROIMAGING, I. & GROUP, N.-L. N. F. S. 2014. Overrepresentation of glutamate signaling in Alzheimer's disease: network-based pathway enrichment using meta-analysis of genome-wide association studies. *PLoS One,* 9**,** e95413.

PETROVSKI, S., WANG, Q., HEINZEN, E. L., ALLEN, A. S. & GOLDSTEIN, D. B. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet,* 9**,** e1003709.

PHILLIPS, P. C. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet,* 9**,** 855-67.

POLLARD, K. S., HUBISZ, M. J., ROSENBLOOM, K. R. & SIEPEL, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res,* 20**,** 110-21.

POMERANTZ, M. M., AHMADIYEH, N., JIA, L., HERMAN, P., VERZI, M. P., DODDAPANENI, H., BECKWITH, C. A., CHAN, J. A., HILLS, A., DAVIS, M., YAO, K., KEHOE, S. M., LENZ, H. J., HAIMAN, C. A., YAN, C., HENDERSON, B. E., FRENKEL, B., BARRETINA, J., BASS, A., TABERNERO, J., BASELGA, J., REGAN, M. M., MANAK, J. R., SHIVDASANI, R., COETZEE, G. A. & FREEDMAN, M. L. 2009. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet,* 41**,** 882-4.

PRABHU, S. & PE'ER, I. 2012. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res,* 22**,** 2230-40.

PSYCHIATRIC, G. C. B. D. W. G. 2011. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet,* 43**,** 977-83.

PURCELL, S. M., MORAN, J. L., FROMER, M., RUDERFER, D., SOLOVIEFF, N., ROUSSOS, P., O'DUSHLAINE, C., CHAMBERT, K., BERGEN, S. E., KAHLER, A., DUNCAN, L., STAHL, E., GENOVESE, G., FERNANDEZ, E., COLLINS, M. O., KOMIYAMA, N. H., CHOUDHARY, J. S., MAGNUSSON, P. K., BANKS, E., SHAKIR, K., GARIMELLA, K., FENNELL, T., DEPRISTO, M., GRANT, S. G., HAGGARTY, S. J., GABRIEL, S., SCOLNICK, E. M., LANDER, E. S., HULTMAN, C. M., SULLIVAN, P. F., MCCARROLL, S. A. & SKLAR, P. 2014. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature,* 506**,** 185-90.

QU, H. & FANG, X. 2013. A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics,* 11**,** 135-41.

REHMAN, A. U., SANTOS-CORTEZ, R. L., DRUMMOND, M. C., SHAHZAD, M., LEE, K., MORELL, R. J., ANSAR, M., JAN, A., WANG, X., AZIZ, A., RIAZUDDIN, S., SMITH, J. D., WANG, G. T., AHMED, Z. M., GUL, K., SHEARER, A. E., SMITH, R. J., SHENDURE, J., BAMSHAD, M. J.,

NICKERSON, D. A., UNIVERSITY OF WASHINGTON CENTER FOR MENDELIAN, G., HINNANT, J., KHAN, S. N., FISHER, R. A., AHMAD, W., FRIDERICI, K. H., RIAZUDDIN, S., FRIEDMAN, T. B., WILCH, E. S. & LEAL, S. M. 2015. Challenges and solutions for gene identification in the presence of familial locus heterogeneity. *Eur J Hum Genet,* 23**,** 1207-15.

REITZ, C., TOSTO, G., VARDARAJAN, B., ROGAEVA, E., GHANI, M., ROGERS, R. S., CONRAD, C., HAINES, J. L., PERICAK-VANCE, M. A., FALLIN, M. D., FOROUD, T., FARRER, L. A., SCHELLENBERG, G. D., GEORGE-HYSLOP, P. S., MAYEUX, R. & ALZHEIMER'S DISEASE GENETICS, C. 2013. Independent and epistatic effects of variants in VPS10-d receptors on Alzheimer disease risk and processing of the amyloid precursor protein (APP). *Transl Psychiatry,* 3**,** e256.

REPLICATION, D. I. G., META-ANALYSIS, C., ASIAN GENETIC EPIDEMIOLOGY NETWORK TYPE 2 DIABETES, C., SOUTH ASIAN TYPE 2 DIABETES, C., MEXICAN AMERICAN TYPE 2 DIABETES, C., TYPE 2 DIABETES GENETIC EXPLORATION BY NEX-GENERATION SEQUENCING IN MUYLTI-ETHNIC SAMPLES, C., MAHAJAN, A., GO, M. J., ZHANG, W., BELOW, J. E., GAULTON, K. J., FERREIRA, T., HORIKOSHI, M., JOHNSON, A. D., NG, M. C., PROKOPENKO, I., SALEHEEN, D., WANG, X., ZEGGINI, E., ABECASIS, G. R., ADAIR, L. S., ALMGREN, P., ATALAY, M., AUNG, T., BALDASSARRE, D., BALKAU, B., BAO, Y., BARNETT, A. H., BARROSO, I., BASIT, A., BEEN, L. F., BEILBY, J., BELL, G. I., BENEDIKTSSON, R., BERGMAN, R. N., BOEHM, B. O., BOERWINKLE, E., BONNYCASTLE, L. L., BURTT, N., CAI, Q., CAMPBELL, H., CAREY, J., CAUCHI, S., CAULFIELD, M., CHAN, J. C., CHANG, L. C., CHANG, T. J., CHANG, Y. C., CHARPENTIER, G., CHEN, C. H., CHEN, H., CHEN, Y. T., CHIA, K. S., CHIDAMBARAM, M., CHINES, P. S., CHO, N. H., CHO, Y. M., CHUANG, L. M., COLLINS, F. S., CORNELIS, M. C., COUPER, D. J., CRENSHAW, A. T., VAN DAM, R. M., DANESH, J., DAS, D., DE FAIRE, U., DEDOUSSIS, G., DELOUKAS, P., DIMAS, A. S., DINA, C., DONEY, A. S., DONNELLY, P. J., DORKHAN, M., VAN DUIJN, C., DUPUIS, J., EDKINS, S., ELLIOTT, P., EMILSSON, V., ERBEL, R., ERIKSSON, J. G., ESCOBEDO, J., ESKO, T., EURY, E., FLOREZ, J. C., FONTANILLAS, P., FOROUHI, N. G., FORSEN, T., FOX, C., FRASER, R. M., FRAYLING, T. M., FROGUEL, P., FROSSARD, P., GAO, Y., GERTOW, K., GIEGER, C., GIGANTE, B., GRALLERT, H., GRANT, G. B., GRROP, L. C., GROVES, C. J., et al. 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet,* 46**,** 234-44.

RHINN, H., FUJITA, R., QIANG, L., CHENG, R., LEE, J. H. & ABELIOVICH, A. 2013. Integrative genomics identifies APOE epsilon4 effectors in Alzheimer's disease. *Nature,* 500**,** 45-50.

RICHARDS, A. L., JONES, L., MOSKVINA, V., KIROV, G., GEJMAN, P. V., LEVINSON, D. F., SANDERS, A. R., MOLECULAR GENETICS OF

SCHIZOPHRENIA, C., INTERNATIONAL SCHIZOPHRENIA, C., PURCELL, S., VISSCHER, P. M., CRADDOCK, N., OWEN, M. J., HOLMANS, P. & O'DONOVAN, M. C. 2012. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol Psychiatry,* 17**,** 193-201.

RIETSCHEL, M., MATTHEISEN, M., FRANK, J., TREUTLEIN, J., DEGENHARDT, F., BREUER, R., STEFFENS, M., MIER, D., ESSLINGER, C., WALTER, H., KIRSCH, P., ERK, S., SCHNELL, K., HERMS, S., WICHMANN, H. E., SCHREIBER, S., JOCKEL, K. H., STROHMAIER, J., ROESKE, D., HAENISCH, B., GROSS, M., HOEFELS, S., LUCAE, S., BINDER, E. B., WIENKER, T. F., SCHULZE, T. G., SCHMAL, C., ZIMMER, A., JURAEVA, D., BRORS, B., BETTECKEN, T., MEYER-LINDENBERG, A., MULLER-MYHSOK, B., MAIER, W., NOTHEN, M. M. & CICHON, S. 2010. Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. *Biol Psychiatry,* 68**,** 578-85.

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., KIM, Y., LEE, S. H., MAGNUSSON, P. K., SANCHEZ, N., STAHL, E. A., WILLIAMS, S., WRAY, N. R., XIA, K., BETTELLA, F., BORGLUM, A. D., BULIK-SULLIVAN, B. K., CORMICAN, P., CRADDOCK, N., DE LEEUW, C., DURMISHI, N., GILL, M., GOLIMBET, V., HAMSHERE, M. L., HOLMANS, P., HOUGAARD, D. M., KENDLER, K. S., LIN, K., MORRIS, D. W., MORS, O., MORTENSEN, P. B., NEALE, B. M., O'NEILL, F. A., OWEN, M. J., MILOVANCEVIC, M. P., POSTHUMA, D., POWELL, J., RICHARDS, A. L., RILEY, B. P., RUDERFER, D., RUJESCU, D., SIGURDSSON, E., SILAGADZE, T., SMIT, A. B., STEFANSSON, H., STEINBERG, S., SUVISAARI, J., TOSATO, S., VERHAGE, M., WALTERS, J. T., MULTICENTER GENETIC STUDIES OF SCHIZOPHRENIA, C., LEVINSON, D. F., GEJMAN, P. V., KENDLER, K. S., LAURENT, C., MOWRY, B. J., O'DONOVAN, M. C., OWEN, M. J., PULVER, A. E., RILEY, B. P., SCHWAB, S. G., WILDENAUER, D. B., DUDBRIDGE, F., HOLMANS, P., SHI, J., ALBUS, M., ALEXANDER, M., CAMPION, D., COHEN, D., DIKEOS, D., DUAN, J., EICHHAMMER, P., GODARD, S., HANSEN, M., LERER, F. B., LIANG, K. Y., MAIER, W., MALLET, J., NERTNEY, D. A., NESTADT, G., NORTON, N., O'NEILL, F. A., PAPADIMITRIOU, G. N., RIBBLE, R., SANDERS, A. R., SILVERMAN, J. M., WALSH, D., WILLIAMS, N. M., WORMLEY, B., PSYCHOSIS ENDOPHENOTYPES INTERNATIONAL, C., ARRANZ, M. J., BAKKER, S., BENDER, S., BRAMON, E., COLLIER, D., CRESPO-FACORRO, B., et al. 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet,* 45**,** 1150-9.

RITCHIE, G. R., DUNHAM, I., ZEGGINI, E. & FLICEK, P. 2014. Functional annotation of noncoding sequence variants. *Nat Methods,* 11**,** 294-6.

RONEMUS, M., IOSSIFOV, I., LEVY, D. & WIGLER, M. 2014. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet,* 15**,** 133-41.

ROWE, C. C., BOURGEAT, P., ELLIS, K. A., BROWN, B., LIM, Y. Y., MULLIGAN, R., JONES, G., MARUFF, P., WOODWARD, M., PRICE, R., ROBINS, P., TOCHON-DANGUY, H., O'KEEFE, G., PIKE, K. E., YATES, P., SZOEKE, C., SALVADO, O., MACAULAY, S. L., O'MEARA, T., HEAD, R., COBIAC, L., SAVAGE, G., MARTINS, R., MASTERS, C. L., AMES, D. & VILLEMAGNE, V. L. 2013. Predicting Alzheimer disease with beta-amyloid imaging: results from the Australian imaging, biomarkers, and lifestyle study of ageing. *Ann Neurol,* 74**,** 905-13.

RYAN, N. M., MORRIS, S. W., PORTEOUS, D. J., TAYLOR, M. S. & EVANS, K. L. 2014. SuRFing the genomics wave: an R package for prioritising SNPs by functionality. *Genome Med,* 6**,** 79.

SABO, P. J., HAWRYLYCZ, M., WALLACE, J. C., HUMBERT, R., YU, M., SHAFER, A., KAWAMOTO, J., HALL, R., MACK, J., DORSCHNER, M. O., MCARTHUR, M. & STAMATOYANNOPOULOS, J. A. 2004. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A,* 101**,** 16837-42.

SACCONE, S. F., BOLZE, R., THOMAS, P., QUAN, J., MEHTA, G., DEELMAN, E., TISCHFIELD, J. A. & RICE, J. P. 2010. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res,* 38**,** W201-9.

SAMUELS, M. E. & ROULEAU, G. A. 2011. The case for locus-specific databases. *Nat Rev Genet,* 12**,** 378-9.

SANDERS, S. J., ERCAN-SENCICEK, A. G., HUS, V., LUO, R., MURTHA, M. T., MORENO-DE-LUCA, D., CHU, S. H., MOREAU, M. P., GUPTA, A. R., THOMSON, S. A., MASON, C. E., BILGUVAR, K., CELESTINO-SOPER, P. B., CHOI, M., CRAWFORD, E. L., DAVIS, L., WRIGHT, N. R., DHODAPKAR, R. M., DICOLA, M., DILULLO, N. M., FERNANDEZ, T. V., FIELDING-SINGH, V., FISHMAN, D. O., FRAHM, S., GARAGALOYAN, R., GOH, G. S., KAMMELA, S., KLEI, L., LOWE, J. K., LUND, S. C., MCGREW, A. D., MEYER, K. A., MOFFAT, W. J., MURDOCH, J. D., O'ROAK, B. J., OBER, G. T., POTTENGER, R. S., RAUBESON, M. J., SONG, Y., WANG, Q., YASPAN, B. L., YU, T. W., YURKIEWICZ, I. R., BEAUDET, A. L., CANTOR, R. M., CURLAND, M., GRICE, D. E., GUNEL, M., LIFTON, R. P., MANE, S. M., MARTIN, D. M., SHAW, C. A., SHELDON, M., TISCHFIELD, J. A., WALSH, C. A., MORROW, E. M., LEDBETTER, D. H., FOMBONNE, E., LORD, C., MARTIN, C. L., BROOKS, A. I., SUTCLIFFE, J. S., COOK, E. H., JR., GESCHWIND, D., ROEDER, K., DEVLIN, B. & STATE, M. W. 2011. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron,* 70**,** 863-85.

SCHIZOPHRENIA PSYCHIATRIC GENOME-WIDE ASSOCIATION STUDY, C. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet,* 43**,** 969-76.

SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS, C. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature,* 511**,** 421-7.

SCHODEL, J., BARDELLA, C., SCIESIELSKI, L. K., BROWN, J. M., PUGH, C. W., BUCKLE, V., TOMLINSON, I. P., RATCLIFFE, P. J. & MOLE, D. R. 2012. Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nat Genet,* 44**,** 420-5, S1-2.

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., TOBACCO, GENETICS, C., BIPOLAR DISORDER PSYCHIATRIC GENOMICS, C., SCHIZOPHRENIA PSYCHIATRIC GENOMICS, C., SCHORK, N. J., ANDREASSEN, O. A. & DALE, A. M. 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.

SCHULZE, T. G., AKULA, N., BREUER, R., STEELE, J., NALLS, M. A., SINGLETON, A. B., DEGENHARDT, F. A., NOTHEN, M. M., CICHON, S., RIETSCHEL, M., BIPOLAR GENOME, S. & MCMAHON, F. J. 2014. Molecular genetic overlap in bipolar disorder, schizophrenia, and major depressive disorder. *World J Biol Psychiatry,* 15**,** 200-8.

SEBAT, J., LAKSHMI, B., MALHOTRA, D., TROGE, J., LESE-MARTIN, C., WALSH, T., YAMROM, B., YOON, S., KRASNITZ, A., KENDALL, J., LEOTTA, A., PAI, D., ZHANG, R., LEE, Y. H., HICKS, J., SPENCE, S. J., LEE, A. T., PUURA, K., LEHTIMAKI, T., LEDBETTER, D., GREGERSEN, P. K., BREGMAN, J., SUTCLIFFE, J. S., JOBANPUTRA, V., CHUNG, W., WARBURTON, D., KING, M. C., SKUSE, D., GESCHWIND, D. H., GILLIAM, T. C., YE, K. & WIGLER, M. 2007. Strong association of de novo copy number mutations with autism. *Science,* 316**,** 445-9.

SERRETTI, A. & FABBRI, C. 2013. Shared genetics among major psychiatric disorders. *Lancet,* 381**,** 1339-41.

SHALEV, H., SERLIN, Y. & FRIEDMAN, A. 2009. Breaching the blood-brain barrier as a gate to psychiatric disorder. *Cardiovasc Psychiatry Neurol,* 2009**,** 278531.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SHI, J., POTASH, J. B., KNOWLES, J. A., WEISSMAN, M. M., CORYELL, W., SCHEFTNER, W. A., LAWSON, W. B., DEPAULO, J. R., JR., GEJMAN, P. V., SANDERS, A. R., JOHNSON, J. K., ADAMS, P., CHAUDHURY, S.,

JANCIC, D., EVGRAFOV, O., ZVINYATSKOVSKIY, A., ERTMAN, N., GLADIS, M., NEIMANAS, K., GOODELL, M., HALE, N., NEY, N., VERMA, R., MIREL, D., HOLMANS, P. & LEVINSON, D. F. 2011. Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatry,* 16**,** 193-201.

SHIHAB, H. A., GOUGH, J., COOPER, D. N., STENSON, P. D., BARKER, G. L., EDWARDS, K. J., DAY, I. N. & GAUNT, T. R. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat,* 34**,** 57-65.

SHINOZAKI, G. & POTASH, J. B. 2014. New developments in the genetics of bipolar disorder. *Curr Psychiatry Rep,* 16**,** 493.

SHYN, S. I., SHI, J., KRAFT, J. B., POTASH, J. B., KNOWLES, J. A., WEISSMAN, M. M., GARRIOCK, H. A., YOKOYAMA, J. S., MCGRATH, P. J., PETERS, E. J., SCHEFTNER, W. A., CORYELL, W., LAWSON, W. B., JANCIC, D., GEJMAN, P. V., SANDERS, A. R., HOLMANS, P., SLAGER, S. L., LEVINSON, D. F. & HAMILTON, S. P. 2011. Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry,* 16**,** 202-15.

SIEPEL, A., BEJERANO, G., PEDERSEN, J. S., HINRICHS, A. S., HOU, M., ROSENBLOOM, K., CLAWSON, H., SPIETH, J., HILLIER, L. W., RICHARDS, S., WEINSTOCK, G. M., WILSON, R. K., GIBBS, R. A., KENT, W. J., MILLER, W. & HAUSSLER, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res,* 15**,** 1034-50.

SIFRIM, A., POPOVIC, D., TRANCHEVENT, L. C., ARDESHIRDAVANI, A., SAKAI, R., KONINGS, P., VERMEESCH, J. R., AERTS, J., DE MOOR, B. & MOREAU, Y. 2013. eXtasy: variant prioritization by genomic data fusion. *Nat Methods,* 10**,** 1083-4.

SILVENTOINEN, K., SAMMALISTO, S., PEROLA, M., BOOMSMA, D. I., CORNES, B. K., DAVIS, C., DUNKEL, L., DE LANGE, M., HARRIS, J. R., HJELMBORG, J. V., LUCIANO, M., MARTIN, N. G., MORTENSEN, J., NISTICO, L., PEDERSEN, N. L., SKYTTHE, A., SPECTOR, T. D., STAZI, M. A., WILLEMSEN, G. & KAPRIO, J. 2003. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res,* 6**,** 399-408.

SING, T., SANDER, O., BEERENWINKEL, N. & LENGAUER, T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics,* 21**,** 3940-1.

SKLAR, P., SMOLLER, J. W., FAN, J., FERREIRA, M. A., PERLIS, R. H., CHAMBERT, K., NIMGAONKAR, V. L., MCQUEEN, M. B., FARAONE, S. V., KIRBY, A., DE BAKKER, P. I., OGDIE, M. N., THASE, M. E., SACHS, G. S., TODD-BROWN, K., GABRIEL, S. B., SOUGNEZ, C., GATES, C., BLUMENSTIEL, B., DEFELICE, M., ARDLIE, K. G.,

FRANKLIN, J., MUIR, W. J., MCGHEE, K. A., MACINTYRE, D. J., MCLEAN, A., VANBECK, M., MCQUILLIN, A., BASS, N. J., ROBINSON, M., LAWRENCE, J., ANJORIN, A., CURTIS, D., SCOLNICK, E. M., DALY, M. J., BLACKWOOD, D. H., GURLING, H. M. & PURCELL, S. M. 2008. Whole-genome association study of bipolar disorder. *Mol Psychiatry,* 13**,** 558-69.

SMIALOWSKI, P., FRISHMAN, D. & KRAMER, S. 2010. Pitfalls of supervised feature selection. *Bioinformatics,* 26**,** 440-3.

SMITH, E. N., BLOSS, C. S., BADNER, J. A., BARRETT, T., BELMONTE, P. L., BERRETTINI, W., BYERLEY, W., CORYELL, W., CRAIG, D., EDENBERG, H. J., ESKIN, E., FOROUD, T., GERSHON, E., GREENWOOD, T. A., HIPOLITO, M., KOLLER, D. L., LAWSON, W. B., LIU, C., LOHOFF, F., MCINNIS, M. G., MCMAHON, F. J., MIREL, D. B., MURRAY, S. S., NIEVERGELT, C., NURNBERGER, J., NWULIA, E. A., PASCHALL, J., POTASH, J. B., RICE, J., SCHULZE, T. G., SCHEFTNER, W., PANGANIBAN, C., ZAITLEN, N., ZANDI, P. P., ZOLLNER, S., SCHORK, N. J. & KELSOE, J. R. 2009. Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol Psychiatry,* 14**,** 755-63.

SMITH, E. N., KOLLER, D. L., PANGANIBAN, C., SZELINGER, S., ZHANG, P., BADNER, J. A., BARRETT, T. B., BERRETTINI, W. H., BLOSS, C. S., BYERLEY, W., CORYELL, W., EDENBERG, H. J., FOROUD, T., GERSHON, E. S., GREENWOOD, T. A., GUO, Y., HIPOLITO, M., KEATING, B. J., LAWSON, W. B., LIU, C., MAHON, P. B., MCINNIS, M. G., MCMAHON, F. J., MCKINNEY, R., MURRAY, S. S., NIEVERGELT, C. M., NURNBERGER, J. I., JR., NWULIA, E. A., POTASH, J. B., RICE, J., SCHULZE, T. G., SCHEFTNER, W. A., SHILLING, P. D., ZANDI, P. P., ZOLLNER, S., CRAIG, D. W., SCHORK, N. J. & KELSOE, J. R. 2011. Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. *PLoS Genet,* 7**,** e1002134.

SMOLLER, J. W. & FINN, C. T. 2003. Family, twin, and adoption studies of bipolar disorder. *Am J Med Genet C Semin Med Genet,* 123C**,** 48-58.

SOKOLOWSKI, M., WASSERMAN, J. & WASSERMAN, D. 2010. Association of polymorphisms in the SLIT2 axonal guidance gene with anger in suicide attempters. *Mol Psychiatry,* 15**,** 10-1.

ST CLAIR, D., BLACKWOOD, D., MUIR, W., CAROTHERS, A., WALKER, M., SPOWART, G., GOSDEN, C. & EVANS, H. J. 1990. Association within a family of a balanced autosomal translocation with major mental illness. *Lancet,* 336**,** 13-6.

STEFANSSON, H., RUJESCU, D., CICHON, S., PIETILAINEN, O. P., INGASON, A., STEINBERG, S., FOSSDAL, R., SIGURDSSON, E., SIGMUNDSSON, T., BUIZER-VOSKAMP, J. E., HANSEN, T., JAKOBSEN, K. D., MUGLIA, P., FRANCKS, C., MATTHEWS, P. M., GYLFASON, A., HALLDORSSON, B. V., GUDBJARTSSON, D., THORGEIRSSON, T. E.,

SIGURDSSON, A., JONASDOTTIR, A., JONASDOTTIR, A., BJORNSSON, A., MATTIASDOTTIR, S., BLONDAL, T., HARALDSSON, M., MAGNUSDOTTIR, B. B., GIEGLING, I., MOLLER, H. J., HARTMANN, A., SHIANNA, K. V., GE, D., NEED, A. C., CROMBIE, C., FRASER, G., WALKER, N., LONNQVIST, J., SUVISAARI, J., TUULIO-HENRIKSSON, A., PAUNIO, T., TOULOPOULOU, T., BRAMON, E., DI FORTI, M., MURRAY, R., RUGGERI, M., VASSOS, E., TOSATO, S., WALSHE, M., LI, T., VASILESCU, C., MUHLEISEN, T. W., WANG, A. G., ULLUM, H., DJUROVIC, S., MELLE, I., OLESEN, J., KIEMENEY, L. A., FRANKE, B., GROUP, SABATTI, C., FREIMER, N. B., GULCHER, J. R., THORSTEINSDOTTIR, U., KONG, A., ANDREASSEN, O. A., OPHOFF, R. A., GEORGI, A., RIETSCHEL, M., WERGE, T., PETURSSON, H., GOLDSTEIN, D. B., NOTHEN, M. M., PELTONEN, L., COLLIER, D. A., ST CLAIR, D. & STEFANSSON, K. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature,* 455**,** 232-6.

STENSON, P. D., BALL, E. V., MORT, M., PHILLIPS, A. D., SHIEL, J. A., THOMAS, N. S., ABEYSINGHE, S., KRAWCZAK, M. & COOPER, D. N. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat,* 21**,** 577-81.

STRITTMATTER, W. J., SAUNDERS, A. M., SCHMECHEL, D., PERICAK-VANCE, M., ENGHILD, J., SALVESEN, G. S. & ROSES, A. D. 1993. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A,* 90**,** 1977-81.

SULLIVAN, P. F. 2010. The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron,* 68**,** 182-6.

SULLIVAN, P. F., DALY, M. J. & O'DONOVAN, M. 2012. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet,* 13**,** 537-51.

SULLIVAN, P. F., DE GEUS, E. J., WILLEMSEN, G., JAMES, M. R., SMIT, J. H., ZANDBELT, T., AROLT, V., BAUNE, B. T., BLACKWOOD, D., CICHON, S., COVENTRY, W. L., DOMSCHKE, K., FARMER, A., FAVA, M., GORDON, S. D., HE, Q., HEATH, A. C., HEUTINK, P., HOLSBOER, F., HOOGENDIJK, W. J., HOTTENGA, J. J., HU, Y., KOHLI, M., LIN, D., LUCAE, S., MACINTYRE, D. J., MAIER, W., MCGHEE, K. A., MCGUFFIN, P., MONTGOMERY, G. W., MUIR, W. J., NOLEN, W. A., NOTHEN, M. M., PERLIS, R. H., PIRLO, K., POSTHUMA, D., RIETSCHEL, M., RIZZU, P., SCHOSSER, A., SMIT, A. B., SMOLLER, J. W., TZENG, J. Y., VAN DYCK, R., VERHAGE, M., ZITMAN, F. G., MARTIN, N. G., WRAY, N. R., BOOMSMA, D. I. & PENNINX, B. W. 2009. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol Psychiatry,* 14**,** 359-75.

SULLIVAN, P. F., NEALE, M. C. & KENDLER, K. S. 2000. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry,* 157**,** 1552-62.

SWOBODA, K. J., SAUL, J. P., MCKENNA, C. E., SPELLER, N. B. & HYLAND, K. 2003. Aromatic L-amino acid decarboxylase deficiency: overview of clinical features and outcomes. *Ann Neurol,* 54 Suppl 6**,** S49-55.

TAN, A. C. & GILBERT, D. 2003. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics,* 2**,** S75-83.

TAVARÉ 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences,* 17**,** 57-86.

THOMAS, P. D., CAMPBELL, M. J., KEJARIWAL, A., MI, H., KARLAK, B., DAVERMAN, R., DIEMER, K., MURUGANUJAN, A. & NARECHANIA, A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res,* 13**,** 2129-41.

THORN, G. W. 1968. Approach to the patient with "idiopathic edema" or "periodic swelling". *JAMA,* 206**,** 333-8.

THURMAN, R. E., RYNES, E., HUMBERT, R., VIERSTRA, J., MAURANO, M. T., HAUGEN, E., SHEFFIELD, N. C., STERGACHIS, A. B., WANG, H., VERNOT, B., GARG, K., JOHN, S., SANDSTROM, R., BATES, D., BOATMAN, L., CANFIELD, T. K., DIEGEL, M., DUNN, D., EBERSOL, A. K., FRUM, T., GISTE, E., JOHNSON, A. K., JOHNSON, E. M., KUTYAVIN, T., LAJOIE, B., LEE, B. K., LEE, K., LONDON, D., LOTAKIS, D., NEPH, S., NERI, F., NGUYEN, E. D., QU, H., REYNOLDS, A. P., ROACH, V., SAFI, A., SANCHEZ, M. E., SANYAL, A., SHAFER, A., SIMON, J. M., SONG, L., VONG, S., WEAVER, M., YAN, Y., ZHANG, Z., ZHANG, Z., LENHARD, B., TEWARI, M., DORSCHNER, M. O., HANSEN, R. S., NAVAS, P. A., STAMATOYANNOPOULOS, G., IYER, V. R., LIEB, J. D., SUNYAEV, S. R., AKEY, J. M., SABO, P. J., KAUL, R., FUREY, T. S., DEKKER, J., CRAWFORD, G. E. & STAMATOYANNOPOULOS, J. A. 2012. The accessible chromatin landscape of the human genome. *Nature,* 489**,** 75-82.

TORKAMANI, A. & SCHORK, N. J. 2008. Predicting functional regulatory polymorphisms. *Bioinformatics,* 24**,** 1787-92.

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

UHER, R. 2014. Gene-environment interactions in severe mental illness. *Front Psychiatry,* 5**,** 48.

VACIC, V., MCCARTHY, S., MALHOTRA, D., MURRAY, F., CHOU, H. H., PEOPLES, A., MAKAROV, V., YOON, S., BHANDARI, A., COROMINAS, R., IAKOUCHEVA, L. M., KRASTOSHEVSKY, O.,

KRAUSE, V., LARACH-WALTERS, V., WELSH, D. K., CRAIG, D., KELSOE, J. R., GERSHON, E. S., LEAL, S. M., DELL AQUILA, M., MORRIS, D. W., GILL, M., CORVIN, A., INSEL, P. A., MCCLELLAN, J., KING, M. C., KARAYIORGOU, M., LEVY, D. L., DELISI, L. E. & SEBAT, J. 2011. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature,* 471**,** 499-503.

VANCAMPFORT, D., MITCHELL, A. J., DE HERT, M., SIENAERT, P., PROBST, M., BUYS, R. & STUBBS, B. 2015a. Prevalence and predictors of type 2 diabetes mellitus in people with bipolar disorder: a systematic review and meta-analysis. *J Clin Psychiatry*.

VANCAMPFORT, D., MITCHELL, A. J., DE HERT, M., SIENAERT, P., PROBST, M., BUYS, R. & STUBBS, B. 2015b. Type 2 Diabetes in Patients with Major Depressive Disorder: A Meta-Analysis of Prevalence Estimates and Predictors. *Depress Anxiety*.

VANNESCHI, L., FARINACCIO, A., MAURI, G., ANTONIOTTI, M., PROVERO, P. & GIACOBINI, M. 2011. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Min,* 4**,** 12.

VERBEEK, E. C., BEVOVA, M. R., BOCHDANOVITS, Z., RIZZU, P., BAKKER, I. M., UITHUISJE, T., DE GEUS, E. J., SMIT, J. H., PENNINX, B. W., BOOMSMA, D. I., HOOGENDIJK, W. J. & HEUTINK, P. 2013. Resequencing three candidate genes for major depressive disorder in a Dutch cohort. *PLoS One,* 8**,** e79921.

VERNOT, B., STERGACHIS, A. B., MAURANO, M. T., VIERSTRA, J., NEPH, S., THURMAN, R. E., STAMATOYANNOPOULOS, J. A. & AKEY, J. M. 2012. Personal and population genomics of human regulatory variation. *Genome Res,* 22**,** 1689-97.

VIHINEN, M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics,* 13 Suppl 4**,** S2.

VISSCHER, P. M., GODDARD, M. E., DERKS, E. M. & WRAY, N. R. 2012. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol Psychiatry,* 17**,** 474-85.

VISSCHER, P. M., HALEY, C. S., HEATH, S. C., MUIR, W. J. & BLACKWOOD, D. H. 1999. Detecting QTLs for uni- and bipolar disorder using a variance component method. *Psychiatr Genet,* 9**,** 75-84.

VOIGHT, B. F., SCOTT, L. J., STEINTHORSDOTTIR, V., MORRIS, A. P., DINA, C., WELCH, R. P., ZEGGINI, E., HUTH, C., AULCHENKO, Y. S., THORLEIFSSON, G., MCCULLOCH, L. J., FERREIRA, T., GRALLERT, H., AMIN, N., WU, G., WILLER, C. J., RAYCHAUDHURI, S., MCCARROLL, S. A., LANGENBERG, C., HOFMANN, O. M., DUPUIS, J., QI, L., SEGRE, A. V., VAN HOEK, M., NAVARRO, P., ARDLIE, K., BALKAU, B., BENEDIKTSSON, R., BENNETT, A. J., BLAGIEVA, R., BOERWINKLE, E., BONNYCASTLE, L. L., BENGTSSON BOSTROM,

K., BRAVENBOER, B., BUMPSTEAD, S., BURTT, N. P., CHARPENTIER, G., CHINES, P. S., CORNELIS, M., COUPER, D. J., CRAWFORD, G., DONEY, A. S., ELLIOTT, K. S., ELLIOTT, A. L., ERDOS, M. R., FOX, C. S., FRANKLIN, C. S., GANSER, M., GIEGER, C., GRARUP, N., GREEN, T., GRIFFIN, S., GROVES, C. J., GUIDUCCI, C., HADJADJ, S., HASSANALI, N., HERDER, C., ISOMAA, B., JACKSON, A. U., JOHNSON, P. R., JORGENSEN, T., KAO, W. H., KLOPP, N., KONG, A., KRAFT, P., KUUSISTO, J., LAURITZEN, T., LI, M., LIEVERSE, A., LINDGREN, C. M., LYSSENKO, V., MARRE, M., MEITINGER, T., MIDTHJELL, K., MORKEN, M. A., NARISU, N., NILSSON, P., OWEN, K. R., PAYNE, F., PERRY, J. R., PETERSEN, A. K., PLATOU, C., PROENCA, C., PROKOPENKO, I., RATHMANN, W., RAYNER, N. W., ROBERTSON, N. R., ROCHELEAU, G., RODEN, M., SAMPSON, M. J., SAXENA, R., SHIELDS, B. M., SHRADER, P., SIGURDSSON, G., SPARSO, T., STRASSBURGER, K., STRINGHAM, H. M., SUN, Q., SWIFT, A. J., THORAND, B., et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet,* 42**,** 579-89.

WALKER, R. M., CHRISTOFOROU, A., THOMSON, P. A., MCGHEE, K. A., MACLEAN, A., MUHLEISEN, T. W., STROHMAIER, J., NIERATSCHKER, V., NOTHEN, M. M., RIETSCHEL, M., CICHON, S., MORRIS, S. W., JILANI, O., STCLAIR, D., BLACKWOOD, D. H., MUIR, W. J., PORTEOUS, D. J. & EVANS, K. L. 2010. Association analysis of Neuregulin 1 candidate regions in schizophrenia and bipolar disorder. *Neurosci Lett,* 478**,** 9-13.

WAN, X., YANG, C., YANG, Q., XUE, H., TANG, N. L. & YU, W. 2010. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics,* 26**,** 30-7.

WANG, J., ZHUANG, J., IYER, S., LIN, X., WHITFIELD, T. W., GREVEN, M. C., PIERCE, B. G., DONG, X., KUNDAJE, A., CHENG, Y., RANDO, O. J., BIRNEY, E., MYERS, R. M., NOBLE, W. S., SNYDER, M. & WENG, Z. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res,* 22**,** 1798-812.

WANG, J., ZHUANG, J., IYER, S., LIN, X. Y., GREVEN, M. C., KIM, B. H., MOORE, J., PIERCE, B. G., DONG, X., VIRGIL, D., BIRNEY, E., HUNG, J. H. & WENG, Z. 2013. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res,* 41**,** D171-6.

WANG, K., ZHANG, H., MA, D., BUCAN, M., GLESSNER, J. T., ABRAHAMS, B. S., SALYAKINA, D., IMIELINSKI, M., BRADFIELD, J. P., SLEIMAN, P. M., KIM, C. E., HOU, C., FRACKELTON, E., CHIAVACCI, R., TAKAHASHI, N., SAKURAI, T., RAPPAPORT, E., LAJONCHERE, C. M., MUNSON, J., ESTES, A., KORVATSKA, O., PIVEN, J., SONNENBLICK, L. I., ALVAREZ RETUERTO, A. I., HERMAN, E. I.,

DONG, H., HUTMAN, T., SIGMAN, M., OZONOFF, S., KLIN, A., OWLEY, T., SWEENEY, J. A., BRUNE, C. W., CANTOR, R. M., BERNIER, R., GILBERT, J. R., CUCCARO, M. L., MCMAHON, W. M., MILLER, J., STATE, M. W., WASSINK, T. H., COON, H., LEVY, S. E., SCHULTZ, R. T., NURNBERGER, J. I., HAINES, J. L., SUTCLIFFE, J. S., COOK, E. H., MINSHEW, N. J., BUXBAUM, J. D., DAWSON, G., GRANT, S. F., GESCHWIND, D. H., PERICAK-VANCE, M. A., SCHELLENBERG, G. D. & HAKONARSON, H. 2009. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature,* 459**,** 528-33.

WANG, K. S., LIU, X. F. & ARAGAM, N. 2010. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophr Res,* 124**,** 192-9.

WANG, Q., LU, Q. & ZHAO, H. 2015. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front Genet,* 6**,** 149.

WARD, A. J. & COOPER, T. A. 2010. The pathobiology of splicing. *J Pathol,* 220**,** 152-63.

WARD, L. D. & 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**,** D930-4.

WEI, W. H., HEMANI, G. & HALEY, C. S. 2014. Detecting epistasis in human complex traits. *Nat Rev Genet,* 15**,** 722-33.

WEISSFLOG, L., SCHOLZ, C. J., JACOB, C. P., NGUYEN, T. T., ZAMZOW, K., GROSS-LESCH, S., RENNER, T. J., ROMANOS, M., RUJESCU, D., WALITZA, S., KNEITZ, S., LESCH, K. P. & REIF, A. 2013. KCNIP4 as a candidate gene for personality disorders and adult ADHD. *Eur Neuropsychopharmacol,* 23**,** 436-47.

WELLCOME TRUST CASE CONTROL, C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

WELTER, D., MACARTHUR, J., MORALES, J., BURDETT, T., HALL, P., JUNKINS, H., KLEMM, A., FLICEK, P., MANOLIO, T., HINDORFF, L. & PARKINSON, H. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res,* 42**,** D1001-6.

WESTRA, H. J., PETERS, M. J., ESKO, T., YAGHOOTKAR, H., SCHURMANN, C., KETTUNEN, J., CHRISTIANSEN, M. W., FAIRFAX, B. P., SCHRAMM, K., POWELL, J. E., ZHERNAKOVA, A., ZHERNAKOVA, D. V., VELDINK, J. H., VAN DEN BERG, L. H., KARJALAINEN, J., WITHOFF, S., UITTERLINDEN, A. G., HOFMAN, A., RIVADENEIRA, F., T HOEN, P. A., REINMAA, E., FISCHER, K., NELIS, M., MILANI, L., MELZER, D., FERRUCCI, L., SINGLETON, A. B., HERNANDEZ, D. G., NALLS, M. A., HOMUTH, G., NAUCK, M., RADKE, D., VOLKER, U.,

PEROLA, M., SALOMAA, V., BRODY, J., SUCHY-DICEY, A., GHARIB, S. A., ENQUOBAHRIE, D. A., LUMLEY, T., MONTGOMERY, G. W., MAKINO, S., PROKISCH, H., HERDER, C., RODEN, M., GRALLERT, H., MEITINGER, T., STRAUCH, K., LI, Y., JANSEN, R. C., VISSCHER, P. M., KNIGHT, J. C., PSATY, B. M., RIPATTI, S., TEUMER, A., FRAYLING, T. M., METSPALU, A., VAN MEURS, J. B. & FRANKE, L. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet,* 45**,** 1238-43.

WHO. 2015. *http://www.who.int/mental_health/management/en/* [Online]. 2015].

WILHELM, K., MITCHELL, P., SLADE, T., BROWNHILL, S. & ANDREWS, G. 2003. Prevalence and correlates of DSM-IV major depression in an Australian national survey. *J Affect Disord,* 75**,** 155-62.

WILLNOW, T. E., PETERSEN, C. M. & NYKJAER, A. 2008. VPS10P-domain receptors - regulators of neuronal viability and function. *Nat Rev Neurosci,* 9**,** 899-909.

WILSON, H. S. & SKODOL, A. 1994. Special report: DSM-IV: overview and examination of major changes. *Arch Psychiatr Nurs,* 8**,** 340-7.

WINGENDER, E. 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform,* 9**,** 326-32.

WISTE, A., ROBINSON, E. B., MILANESCHI, Y., MEIER, S., RIPKE, S., CLEMENTS, C. C., FITZMAURICE, G. M., RIETSCHEL, M., PENNINX, B. W., SMOLLER, J. W. & PERLIS, R. H. 2014. Bipolar polygenic loading and bipolar spectrum features in major depressive disorder. *Bipolar Disord,* 16**,** 608-16.

WONG, B. S., CAMILLERI, M., CARLSON, P. J., GUICCIARDI, M. E., BURTON, D., MCKINZIE, S., RAO, A. S., ZINSMEISTER, A. R. & GORES, G. J. 2011. A Klothobeta variant mediates protein stability and associates with colon transit in irritable bowel syndrome with diarrhea. *Gastroenterology,* 140**,** 1934-42.

WRAY, N. R., GODDARD, M. E. & VISSCHER, P. M. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res,* 17**,** 1520-8.

WRAY, N. R. & GOTTESMAN, II 2012. Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Front Genet,* 3**,** 118.

WRAY, N. R., PERGADIA, M. L., BLACKWOOD, D. H., PENNINX, B. W., GORDON, S. D., NYHOLT, D. R., RIPKE, S., MACINTYRE, D. J., MCGHEE, K. A., MACLEAN, A. W., SMIT, J. H., HOTTENGA, J. J., WILLEMSEN, G., MIDDELDORP, C. M., DE GEUS, E. J., LEWIS, C. M., MCGUFFIN, P., HICKIE, I. B., VAN DEN OORD, E. J., LIU, J. Z., MACGREGOR, S., MCEVOY, B. P., BYRNE, E. M., MEDLAND, S. E.,

STATHAM, D. J., HENDERS, A. K., HEATH, A. C., MONTGOMERY, G. W., MARTIN, N. G., BOOMSMA, D. I., MADDEN, P. A. & SULLIVAN, P. F. 2012. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol Psychiatry,* 17**,** 36-48.

WRAY, N. R., YANG, J., HAYES, B. J., PRICE, A. L., GODDARD, M. E. & VISSCHER, P. M. 2013. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet,* 14**,** 507-15.

XIA, K., GUO, H., HU, Z., XUN, G., ZUO, L., PENG, Y., WANG, K., HE, Y., XIONG, Z., SUN, L., PAN, Q., LONG, Z., ZOU, X., LI, X., LI, W., XU, X., LU, L., LIU, Y., HU, Y., TIAN, D., LONG, L., OU, J., LIU, Y., LI, X., ZHANG, L., PAN, Y., CHEN, J., PENG, H., LIU, Q., LUO, X., SU, W., WU, L., LIANG, D., DAI, H., YAN, X., FENG, Y., TANG, B., LI, J., MIEDZYBRODZKA, Z., XIA, J., ZHANG, Z., LUO, X., ZHANG, X., ST CLAIR, D., ZHAO, J. & ZHANG, F. 2014. Common genetic variants on 1p13.2 associate with risk of autism. *Mol Psychiatry,* 19**,** 1212-9.

XU, B., ROOS, J. L., LEVY, S., VAN RENSBURG, E. J., GOGOS, J. A. & KARAYIORGOU, M. 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet,* 40**,** 880-5.

XU, H., GREGORY, S. G., HAUSER, E. R., STENGER, J. E., PERICAK-VANCE, M. A., VANCE, J. M., ZUCHNER, S. & HAUSER, M. A. 2005. SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics,* 21**,** 4181-6.

XU, W., COHEN-WOODS, S., CHEN, Q., NOOR, A., KNIGHT, J., HOSANG, G., PARIKH, S. V., DE LUCA, V., TOZZI, F., MUGLIA, P., FORTE, J., MCQUILLIN, A., HU, P., GURLING, H. M., KENNEDY, J. L., MCGUFFIN, P., FARMER, A., STRAUSS, J. & VINCENT, J. B. 2014. Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. *BMC Med Genet,* 15**,** 2.

YANDELL, M., HUFF, C., HU, H., SINGLETON, M., MOORE, B., XING, J., JORDE, L. B. & REESE, M. G. 2011. A probabilistic disease-gene finder for personal genomes. *Genome Res,* 21**,** 1529-42.

YELLEN, G. 2002. The voltage-gated potassium channels and their relatives. *Nature,* 419**,** 35-42.

YIN, C., LI, S., ZHAO, W., GUO, Y., ZHANG, Y. & FENG, J. 2014. The role of fibroblast growth factor receptor 4 polymorphisms in the susceptibility and clinical features of ischemic stroke. *J Clin Neurosci,* 21**,** 246-9.

YU, T., LI, Y. J., BIAN, A. H., ZUO, H. B., ZHU, T. W., JI, S. X., KONG, F., YIN DE, Q., WANG, C. B., WANG, Z. F., WANG, H. Q., YANG, Y., YOO, B. C. & CHO, J. Y. 2014. The regulatory role of activating transcription factor 2 in inflammation. *Mediators Inflamm,* 2014**,** 950472.

ZHANG, H. F., ZHAO, K. J., YANG, P. F., FANG, Y. B., ZHANG, Y. H., LIU, J. M. & HUANG, Q. H. 2012a. Association between fibroblast growth factor receptor 4 Gly388Arg polymorphism and ischaemic stroke. *J Int Med Res,* 40**,** 1708-14.

ZHANG, X., COWPER-SAL LARI, R., BAILEY, S. D., MOORE, J. H. & LUPIEN, M. 2012b. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res,* 22**,** 1437-46.

ZOLLNER, S. & PRITCHARD, J. K. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet,* 80**,** 605-15.