Investigation of new strategies for identifying causal mechanisms in schizophrenia taking bioinformatics approaches beyond genome-wide association studies

Author:
John HUBERT

Supervisor:
Prof Valentina ESCOTT-PRICE

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

MRC Centre for Neuropsychiatric Genetics and Genomics
School of Medicine

April 6, 2023

Declaration of Authorship

I, John HUBERT, declare that this thesis titled, “Investigation of new strategies for identifying causal mechanisms in schizophrenia taking bioinformatics approaches beyond genome-wide association studies” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: *John Hubert*

Date: 28/08/2022

"Slartibartfast: Perhaps I'm old and tired, but I think that the chances of finding out what's actually going on are so absurdly remote that the only thing to do is to say, 'Hang the sense of it,' and keep yourself busy. I'd much rather be happy than right any day.

Arthur Dent: And are you?

Slartibartfast: Ah, no. Well, that's where it all falls down, of course..."

Hitchhikers Guide to the Galaxy (2005) *Based on the book by Douglas Adams*

Acknowledgements

I would like to first acknowledge my supervisor, Valentina Escott-Price who kept me sane while I tried to do the impossible and gave me all the opportunities to grow as a researcher.

I would like to thank my second and third supervisors Michael O'Donovan and James Walters who gave me much needed support into the clinical side of schizophrenia, while I confused them with computational mumbo-jumbo.

Thank you as well to Ric Anney, who, after my year out was able to give me much needed help in getting this thesis submitted.

I acknowledge the generous funding from Cardiff University school of medicine, without which, I would have never been able to undertake this PhD.

I acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

To the strongest people I know, my mother Veronica, my sister Rachael and my brother Alex, there are not enough words in the English language to describe how thankful I am that I have you in my life.

Thank you to Will, one of my closest friends and confidants when times were tough.

And finally to all the friends I made in Cardiff. Every smile and compassionate action made a difference.

Contents

List of Figures

List of Tables

CARDIFF UNIVERSITY

Abstract

Investigation of new strategies for identifying causal mechanisms in schizophrenia taking bioinformatics approaches beyond genome-wide association studies

by John HUBERT

The goal of the research presented here is to provide support towards finding true precision medicine for patients with schizophrenia. In ideal circumstances, and as hinted towards already within oncology (Le Tourneau et al., 2015), this would be to obtain quantitative information from the patient, and use this information to identify an effective treatment option. As the heritability of the genetic liability towards schizophrenia is around 80%, genetic studies are an ideal base to build individualised treatment options.

GWA studies have successfully discovered over 150 loci associated with schizophrenia and have confirmed that the condition is polygenic, i.e. each risk variant in an individual's genome has a small effect size, but there are a large number of these variants which contribute to the pathogenesis of schizophrenia. The explanation of how these variants contribute to the biological processes that cause schizophrenia is however, unknown.

PRSs provide a metric to measure the genetic liability to any individual disorder and capture a large component of the genetic risk towards schizophrenia. In addition, studies using schizophrenia PRS have found genetic overlap with other disorders. If the PRS was designed to focus on specific genes/pathways, it could give a clearer insight into the biological mechanisms that cause schizophrenia.

However, there are many genetic, statistical and computation problems when using multiple PRS across multiple traits. Therefore, I present SurPRSe, a bioinformatics workflow to produce robust gene-set specific PRS that can be compared across multiple traits. I use this method to investigate the relationship between schizophrenia, subcortical brain volume sizes and cognition.

List of Abbreviations

ASD	**A**utism **S**pectrum **D**isorder
ADHD	**A**ttention **D**eficit **H**yperactivity **D**isorder
AS	**A**nkylosing **S**pondylitis
AUC	**A**rea **U**nder the **C**urve
ASL	**A**rterial **S**pin **L**abeled
ARC	neuronal **A**ctivity-**R**egulated **C**ytoskeleton associated protein
BASH	**B**ourne **A**gain **SH**ell
BMI	**B**ody **M**ass **I**ndex
BGG	**B**est **G**uess **G**enotype
BP	**B**ase-**P**air
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
BOLD fMRI	**B**lood **O**xygenation **L**evel **D**ependent **f**unctional **M**agnetic **R**esonance **I**maging
BDNF	**B**rain-**D**erived **N**eurotrophic **F**actor
CNV	**C**opy **N**umber **V**ariant
CoMPaSS	**Co**gnition (in) **M**ood, **P**sychosis (and) **S**chizophrenia **S**tudy
CD	**C**rohn's **D**isease
CNS	**C**entral **N**ervous **S**ystem
CPU	**C**entral **P**rocessing **U**nit
COMT	**C**atechol-**O**-**M**ethyl **T**ransferase
DSM	**D**iagnostic (and) **S**tatistical **M**anual (of Mental Disorders)
dbSNP	**S**ingle **N**ucleotide **P**olymorphism **d**ata**b**ase

DNA	**D**eoxyribo**N**ucleic **A**cid
dztc	**D**i**Z**ygotic twin correlation
EA	**E**ducational **A**ttainment
eQTL	**e**xpression **Q**uantitative **T**rait **L**oci
ExAC	**Ex**ome **A**ggregation **C**onsortium
FWER	**F**amily-**W**ise **E**rror **R**ate
FDR	**F**alse **D**iscovery **R**ate
FMRP	**F**ragile X **M**ental **R**etardation **P**rotein
FA	**F**ractional **A**nsiotropy
GWA	**G**enome **W**ide **A**ssociation
GRC	**G**enome **R**eference **C**onsortium
g	The **G**eneral factor of intelligence
GB	**G**iga**B**ytes
GO	**G**ene**O**ntology
GMT	**G**ene **M**atrix **T**ranspose
GSMR	**G**eneralised **S**ummary-data-based **M**endelian **R**andomisation
HPC	**H**igh **P**erformance **C**omputing
HRC	**H**aplotype **R**eference **C**onsortium reference panel
IQ	**I**ntelligence **Q**uotient
ICD	**I**nternational (Statistical) **C**lassification (of) **D**iseases (and Related Health Problems)
INFO	**INFO**rmation score
KB	**K**ilo-**B**ase
LD	**L**inkage **D**isequilibrium
LDSC	**LD** **S**core **R**egression
MATRICS	**M**easurement **A**nd **T**reatment **R**esearch (to) **I**mprove **C**ognition (in) **S**chizophrenia
miRNA	**M**icro **RNA**

MAF	**M**inor **A**llele **F**requency
MHC	**M**ajor **H**istocompatibility **C**omplex
MGS	**M**olecular **G**enetics of **S**chizophrenia
MCCB	MATRICS **C**ognitive **C**onsensus **B**attery
MR	**M**endelian **R**andomisation
MRS	**M**agnetic **R**esonance **S**pectroscopy
mztc	**M**ono**Z**ygotic twin correlation
MRI	**M**agnetic **R**esonance **I**maging
DTI	**D**iffusion **T**ensor **I**maging
MDD	**M**ajor **D**epressive **Disorder**
NCBI	**N**ational **C**enter (for) **B**iotechnology **I**nformation
NMDA	**N**-**M**ethyl-**D**-**A**spartate
OR	**O**dds **R**atio
PGC	**P**sychiatric **G**enomics **C**onsortium
PRS	**P**olygenic **R**isk **S**core
Pt	**P** value **t**hreshold
PRSAVE	**P**olygenic **R**isk **S**core **A**nalysis **V**iewing **E**nvironment
PCa	**a**ggressive **P**rostate **C**ancer
PPSR	**P**seudo **P**rofile **S**core **R**egression
PCA	**P**rinciple **C**omponent **A**nalysis
pLI	**p**robability of being **L**oss of Function **Intolerant**
PMRS	**P**roton **M**agnetic **R**esonance **S**pectroscopy
PET	**P**ositron **E**mission **T**omography
rsID	**r**eference **s**np (cluster) **ID**
ssID	**s**ubmitted **s**np **ID**
SNV	**S**ingle **N**ucleotide **V**ariant
SNP	**S**ingle **N**ucleotide **P**olymorphism

SCAN	**S**chedules (for) **C**linical **A**ssessment (in) **N**europsychiatry
SurPRSe	**Su**pe**r**computing (with) **P**olygenic **R**isk **S**core **e**valuation
SCZ	**SC**hi**Z**ophrenia
SE	**S**tandard **E**rror
SPECT	**S**ingle **P**hoton **E**mission **C**omputed **T**omography
UCSC	**U**niversity (of) **C**alifornia **S**anta **C**ruz
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
QC	**Q**uality **C**ontrol
VCFS	**V**elo-**c**ardio-**f**acial **S**yndrome
WGCNA	**W**eighted **G**ene **C**o-expression **N**etwork **A**nalysis

1 Introduction

The underlying pathophysiology of schizophrenia is challenging to identify. A large amount of progress has been made in identifying the genetic causes of schizophrenia, but there is still a large gap in linking these causal mutations to biological pathways.

A method already exists which can quantify the genetic risk of an individual with schizophrenia (an individualised genetic score), and a number of biological pathways associated with schizophrenia have already been found. Could incorporating these biological pathways into the individuals' genetic score improve our understanding of the genetic architecture of schizophrenia?

The diagnosis of schizophrenia is dependent on psychiatric history and the examination of the patients' mental status (Keshavan et al., 2020). Imaging procedures are currently used to exclude cases where there is secondary psychosis from, for example, substance abuse or medical illnesses. However, there is evidence that this method may not be cost effective as observed in Lubman et al (2002) , where scans of half of chronic schizophrenia patients and 77% of scans of first episode psychosis were classified as not clinically significant. However, 4 out of 340 scans did observe previously unsuspecting pathology, and 50% of scans observed abnormal pathology. An argument can be made that any procedure which captures abnormal schizophrenia pathology is cost-effective, but that does not argue against researching methods which may identify the schizophrenia patients which were classified as clinically normal in Lubman et al. (Lubman et al., 2002).

A 'rule in' approach where biological biomarkers are used to confirm a diagnosis of schizophrenia may be more reliable, as observed in many other medical conditions (for example, bacterial infections in the lung diagnosed with a combination of chest X-rays and sputum microscopy.). However, schizophrenia does not have any biological biomarkers for use in diagnosis. Neuroimaging is a strong candidate to develop biomarkers in schizophrenia. Imaging procedures can capture phenotypic variations in molecular and cellular disease targets, and is versatile with respect to the measurement of multiple pathophysiological mechanisms, including brain structural integrity deficits, functional dysconnectivity, and systems of altered neurotransmitters (Kraguljac et al., 2021).

Attempting to use genetic variants another approach to identifying a biomarker. For example, the use of a large database of individual's brain volumes and genotypes (ENIGMA) found that some genetic mutations were associated with larger hippocampal volume (Stein et al., 2012). The same association analyses can be attempted for brain volumes and common schizophrenia mutations.

Another strong candidate for a biomarker in schizophrenia may be found in researching the genetics of cognition within schizophrenia patients. Cognitive deficits are central to schizophrenia and meet all the requirements of being an endophenotype of the disorder, aka a trait that is a quantitative, heritable, trait-related deficit assessed within a laboratory environment (Green and Harvey, 2014; Braff, 1993).

1.1 Summary of Introduction

Within this chapter, I will describe:

- How the history of the genetics of schizophrenia began with confusing and contrasting viewpoints, but then lead towards renewed optimism with the rise of genetic technologies including **G**enome **W**ide **A**ssociation (GWA) studies.

- The use of GWA studies to disentangle the common risk of schizophrenia and how this has lead towards the formation of individualised genetic profiles or "**P**olygenic **R**isk **S**cores (PRSs)".

- The ability to incorporate gene-set analysis within the PRS and how it will be applied to subcortical brain volumes and cognition within schizophrenia patients.

1.2 The Genetics of Schizophrenia

1.2.1 Schizophrenia before human molecular genetics

Schizophrenia is an overarching term used to describe the collection of symptoms observed in psychiatric patients. This is evident by the change in classification of "Schizophrenia" in **D**iagnostic (and) **S**tatistical **M**anual (of Mental Disorders) (DSM)-IV to "Schizophrenia spectrum" within DSM-V (Glasheen et al., 2016). It is a deeply complex disorder which constitutes *positive* symptoms including hallucinations and delusions, and *negative* symptoms including diminished emotional or cognitive functions (Tandon et al., 2013). Due to the intrinsic difficulty of the object of schizophrenia itself and problems with the standard of technology, the history of the genetics of schizophrenia is complex and confusing (Henriksen, Nordgaard, and Jansson, 2017).

In a paper with Jung, Bleuler described Dementia praecox (later differentiated and first coined as ′schizophrenia′) as being monogenic following the monogenic transmission discoveries provided by Mendel (Bleuler and Jung, 1908; Bleuler, 1911; Henriksen, Nordgaard, and Jansson, 2017). The monogenic transmission was quickly disproved due to the inability for the theory to fit empirical data observed at the time, but the concept of a singular gene being at least dominantly causative for schizophrenia was postulated until at least 1989 (Holzman, 1989).

Before the advent of the Human Genome Project, genetic inferences into schizophrenia were proposed using pedigree analysis (Henriksen, Nordgaard, and Jansson, 2017). The first goal was to provide evidence for a genetic link for schizophrenia, and to disprove the various psychoanalytical hypotheses of schizophrenia causation. For example, from the 1950′s to the 1970′s it was common to find the negative stereotype of a "schizophrenogenic mother" in psychiatric literature, whereby the mother′s upbringing of the child induced the development of schizophrenia in the child′s later life-course (Seeman, 2016).

Twin studies from the 1960′s showed empirical evidence that there was at least a genetic component to schizophrenia, and a review of all twin studies estimates the heritability to schizophrenia to be 81% (Fischer, 1973; Sullivan, Kendler, and Neale, 2003).

To add context the the phrase "heritability for the genetic liability to schizophrenia" mentioned in the previous paragraph, we can model the variance in any phenotype (including schizophrenia) as the sum of it environmental variance and the genetic variance (Lee et al., 2011). Heritability is described in (Sullivan, Kendler, and Neale, 2003) to be narrow-sense heritability, in which only the variance in the additive genetic components are estimated (Lee et al., 2011). Broad-sense heritability includes all genetic effects including, for example, epistatic and dominant genetic effects (Lee et al., 2011). However, an assumption for heritability is that the trait is normally distributed, otherwise termed the disease liability scale. This works well for continuous traits, but in terms of binary traits including schizophrenia (at least a case-controlled design for GWA studies), this assumption does not hold. We also observe ascertainment bias as the number of cases within the study design is usually higher than the trait prevalence within the general population. So the observed heritability is converted to the liability scale and corrected for population prevalence and the number of cases.

1.2.2 Advent of linkage and candidate gene studies

By the 1990's the neuropsychiatric field placed a significant portion of their focus onto the genetic mechanisms of schizophrenia. At a similar time, the Human Genome Project provided the ability to begin to directly examine DNA as a way to assess the genetic risk towards schizophrenia in individuals (Lander et al., 2001). The initial forays into DNA-based methods was 'linkage analysis', which looked at extended families and/or sibling pairs in an attempt to associate regions of the genome (not individual variants) with schizophrenia (Henriksen, Nordgaard, and Jansson, 2017). Despite the limitation of only analysing at most, a small group of individuals, linkage analysis takes advantage of the fact that genetic markers which are located very close on the genome, tend to be inherited together as observed during meiosis. Therefore, variants surrounding the schizophrenia risk loci will be 'linked' to the loci. While linkage analysis produced results which struggled to be replicated, meta-analysis of linkage studies indicated that the risk of schizophrenia was found in many different chromosomal regions (Ng et al., 2009) and the effect sizes of the alleles within these regions towards schizophrenia appeared to be very low. Linkage analysis hinted towards the polygenic nature of the genetic liability towards schizophrenia and subsequently provided evidence that linkage analysis lacked power when analysing schizophrenia.

In an attempt to counteract the small effect sizes of susceptibility alleles towards schizophrenia, the field examined explored the 'candidate gene' approach, where specific genes were tested for their correlation to schizophrenia within a case-control study design (Henriksen, Nordgaard, and Jansson, 2017). Despite over a 1000 genes having been tested, very few have been shown to be reliably associated to schizophrenia (e.g. DISC1, NRG1 and COMT) but even these have been disputed due to a lack of replication and statistical power (Gejman, Sanders, and Kendler, 2011).

For example, in one of the earliest genetic association studies, **S**ingle **N**ucleotide **P**olymorphisms (SNPs) from 14 candidate genes (RGS4, DISC1, DTNBP1, STX7, TAAR6, PPP3CC, NRG1, DRD2, HTR2A, DAOA, AKT1, CHRNA7, COMT, and ARVCF) were tested for an association with schizophrenia but no experiment-wide or gene-wide significance was found (Sanders et al., 2008). This signified that these specific genes were unlikely to account for a substantial portion of schizophrenia risk. However, these genes may still account for a tiny portion of schizophrenia risk. For example, DISC1 mutant animal models display behavioural, neurostructural and neurochemical phenotypes related to schizophrenia, and is implicated in affecting dopamine signalling pathways (Dahoun et al., 2017). However, with respect to animal models, the relevance of the schizophrenia phenotypes observed in mice for human pathology has been debated.

1.2.3 Rare genetic variation in schizophrenia

While the linkage and candidate gene studies were proving to be unconvincing, there was a growing interest in the potential association of a recurrent deletion in chromosomal band 22q11.2 to schizophrenia (Avramopoulos, 2018). The 22q deletion causes **V**elo-**c**ardio-**f**acial **S**yndrome (VCFS) and it was noted that at least 10% of these patients were also reported to have some form of psychiatric disorder (Chow, Bassett, and Weksberg, 1994). Further follow up studies have found that carriers of this deletion increased the risk of obtaining schizophrenia by a factor of approximately 68, and is today known as one of the most common **C**opy **N**umber **V**ariants (CNVs) associated with schizophrenia (Rees et al., 2014b; Marshall et al., 2017). Chromosomal band 22q11.21 was the first CNV to be discussed as being a risk factor for schizophrenia. Further, the identification of smaller CNVs was made possible by genotyping through microarray's and the ongoing human genome project (Avramopoulos, 2018).

As these schizophrenia factors could be robustly replicated, a segment of schizophrenia genetics research diverged into the identification and study of rare and de-novo risk variants (Kirov et al., 2009). These experiments are carefully designed using, for example, parent-patient trios (the genome of an affected patient and both of their parents) as the sample, and/or exome sequencing to identify more CNVs associated to schizophrenia, while cutting down on expenditure (Avramopoulos, 2018). In rare variation, schizophrenia research examined de-novo mutation, rare CNV and rare **S**ingle **N**ucleotide **V**ariant (SNV) (defined as point mutations with a frequency less than 1%).

De novo variants

I have previously stated that schizophrenia was estimated to be around 81% heritable. This suggests that much of the risk for schizophrenia is inherited. However, there may be alleles that contribute to schizophrenia liability that are not inherited, i.e. mutations which are newly arising. Initial evidence for this came from the observation that increased paternal age at conception was associated with schizophrenia risk (McGrath et al., 2014). Increased paternal age is correlated with de novo mutations. Molecular evidence occurred at a similar time to the discovery of CNVs associated with schizophrenia. First, the rate of mutation for de novo CNVs was significantly elevated in schizophrenia (5%) versus controls (2%) (Rees, O'Donovan, and Owen, 2015). Additionally, the median size of de novo CNVs > 100kb was larger in schizophrenia cases compared with controls (Rees, O'Donovan, and Owen, 2015). After calculation of the selection rate, it is hypothesised that schizophrenia-associated de novo CNVs are purged from the population in less than five generations.

One of the first investigations into the genes and pathways that are disrupted by de novo mutations was by Kirov et al (2012). Proteomic and gene ontology data were used to define the gene-sets to test for association with de novo CNVs. First, a gene was considered to be a 'hit' if the CNV overlapped with the gene according to **B**ase-**P**air (BP). After controlling for biases related to CNV gene-set analysis and partitioning overlapping gene-sets, the association analysis was performed and significance was assessed by one-sided test of an excess of genes hit in the gene set by case CNVs (Kirov et al., 2012). It was found that genes disrupted by de novo CNVs are enriched for genes in the post-synaptic-density proteome and that this association is driven by genes encoding **N**-**M**ethyl-**D**-**A**spartate (NMDA) receptor and neuronal **A**ctivity-**R**egulated **C**ytoskeleton associated protein (ARC) complexes, both of which are involved in synaptic plasticity (Kirov et al., 2012).

Further work by Fromer et. al. (Fromer et al., 2014) implicated disrupted synaptic plasticity through an **F**ragile X **M**ental **R**etardation **P**rotein (FMRP) targets gene-set. Brain expressed genes repressed by FMRP were previously shown to be enriched for de novo mutations in **A**utism **S**pectrum **D**isorder (ASD) and contain multiple genes within the NMDA and ARC complexes. Fromer et al (2014) showed that genes repressed by FMRP were also enriched for de novo variants in schizophrenia.

It was also one of the first studies to show that nonsynonymous de novo mutations were enriched for inherited risk alleles as well, implicating overlap between de novo variant risk and rare inherited variant risk (Fromer et al., 2014). This was further validated in Singh et. al. (Singh et al., 2022) who identified ultra-rare variants (combines de novo, SNV's and protein truncating mutations) in 10 genes that conferred a substantial risk for schizophrenia. The annotated functions of these genes are diverse and include ion transport (CACNA1G, GRIN2A, and GRIA3), neuronal migration and growth (TRIO), transcriptional regulation (SP4, RB1CC1, and SETD1A), nuclear transport (XPO7), and ubiquitin ligation (CUL1, HERC1) (Singh et al., 2022).

CNVs and SNVs

The majority of CNVs increase the risk to schizophrenia substantially, with **O**dds **R**atios (ORs) between 2-60. It is howewver, difficult to ascertain biological insights from these mutations as multiple genes and regulatory elements are disrupted by a singluar CNV. However single gene disruptor CNVs at NRXN1, VIPR2 and PAK7 have been associated with schizophrenia, but only NRXN1 survives multiple testing against all genes in the genome. NRXN1 encodes an adhesion molecule involved in linking presynaptic and postsynaptic neurons. In 2015, a study by Pocklington et al. (Pocklington et al., 2015), supported the previous associations to schizophrenia above, and also reported first genomic evidence of CNVs disrupting genes associated with GABAergic signalling in schizophrenia. Since 2015, an association of CNVs have been implicated with Dystrophin and its binding partners (Marshall et al., 2017), and genes expressed during the consolidation, retrieval or extinction of associative memories (Clifton et al., 2017). There is evidence to support overlap of CNVs and SNVs where in Purcell et al (2014), exome sequencing was used where an increased burden of SNVs were observed in a set of 2546 genes, selected for having a higher probability for being associated with schizophrenia. Enrichments were found in genes affiliated with NMDA receptor and ARC complexes, and FMRP targets. Further support was found in Singh et al (2022) where as well as de novo mutations, SNVs conferred a substantial risk to genes involved in NMDA receptor complexes.

Summary

While progress is being made on the rare variants associated with schizophrenia, their contribution to the overall risk of the disorder is modest. The problem still existed on how to identify the variants which are lowly penetrant and later estimated to confer over one third of the genetic risk towards schizophrenia (Purcell et al., 2009).

In 1996, Risch and Merikangas (1996) proposed that genome wide association studies, rather than genome wide linkage studies, would provide a much more powerful statistical model to test the association of risk variants with modest effects towards complex human diseases. The scientific community started to use GWA studies on the trait of schizophrenia and have not looked back since.

1.3 Schizophrenia GWAS

GWA studies test for associations between multiple common genetic risk variants and/or loci and a trait of interest. The most common study design (but by no means the only design) is a case/control approach, whereby a select number of individuals who have the trait of interest, and a control group of randomly selected individuals are both genotyped. If any individual allellic variant is found more frequently in the cases over the controls, then these variants may indicate a genetic association with the trait of interest. As the design requires the simultaneous statistical testing for potentially millions of variants, the risk of Type I errors is commonly minimised by using a stringent threshold for statistical significance, usually $p < 5 \times 10^{-08}$. This essentially means that there is at least a 5 in a 100,000,000 chance that SNPs found below this threshold have been incorrectly associated with schizophrenia.

Despite evidence of the practicality of GWA studies in 1996, the first schizophrenia GWA was not published until 2007 (Risch and Merikangas, 1996; Lencz et al., 2007). Initially the delay was due to the cost of genotyping, which was quickly overcome through the development of microarray and chip technology (which enabled the ability to scan between 200,000 ~ 2,000,000 SNPs genome-wide), and the development of genetic sequencing projects including Hapmap and the 1000 genomes project (Visscher et al., 2017; Henriksen, Nordgaard, and Jansson, 2017; Frazer et al., 2007; Auton et al., 2015).

1.3.1 The first schizophrenia GWA studies

It was evident after the first GWA studies on schizophrenia that there that was another issue: sample size. In the first schizophrenia GWA study by Lencz *et al.* (2007), the primary case/control analysis only used 71 cases and 31 controls and provided one significantly associated SNP with schizophrenia located near the CSF2RA gene (rs4129148; see Figure 1.1). This was not replicated in future schizophrenia GWA studies.

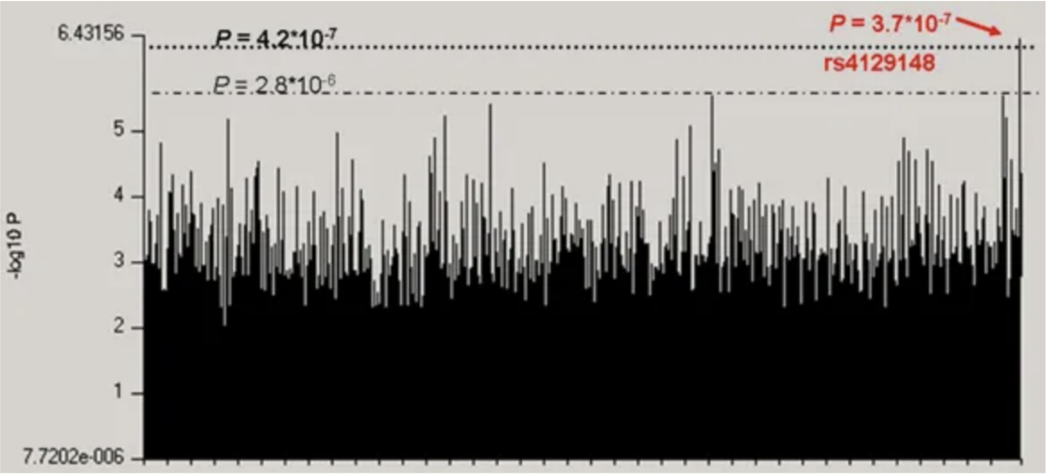


FIGURE 1.1: **Manhattan plot from Lencz et al., 2007**. The genome-wide significant SNP is declared by the red arrow and text.

In the following year, a GWA study of 479 cases and 2,937 controls yielded 12 loci, the most notable being loci surrounding the gene ZNF804A, but rs4129148 was not included and no SNP conferred an OR more than 1.5 (O'Donovan et al., 2008). Not only was it evident that the effect conferred by each SNP was less than expected, but the GWA studies were, at the time, unknowingly hampered by population stratification and an inability to detect **L**inkage **D**isequilibrium (LD) structure. If the common genetic risk of schizophrenia was hypothesised to confer at least half of the risk towards schizophrenia, then it was clear that either schizophrenia GWA studies were not adequately designed to capture this risk in the general population, or the power of the studies needed to increase.

A defining paper in 2009 from the International Schizophrenia Consortium (Purcell et al., 2009) validated the association found for ZNF804A, and provided multiple novel associations within the **M**ajor **H**istocompatibility **C**omplex (MHC) region on the human genome (See Figure 1.2). These associations were found by combining samples with other consortia including the **M**olecular **G**enetics of **S**chizophrenia (MGS) and SGENE consortia. This indicated that GWA studies could capture schizophrenia risk, but the power of the studies needed to dramatically increase, potentially beyond any practical, cost-effective approaches from individual groups.

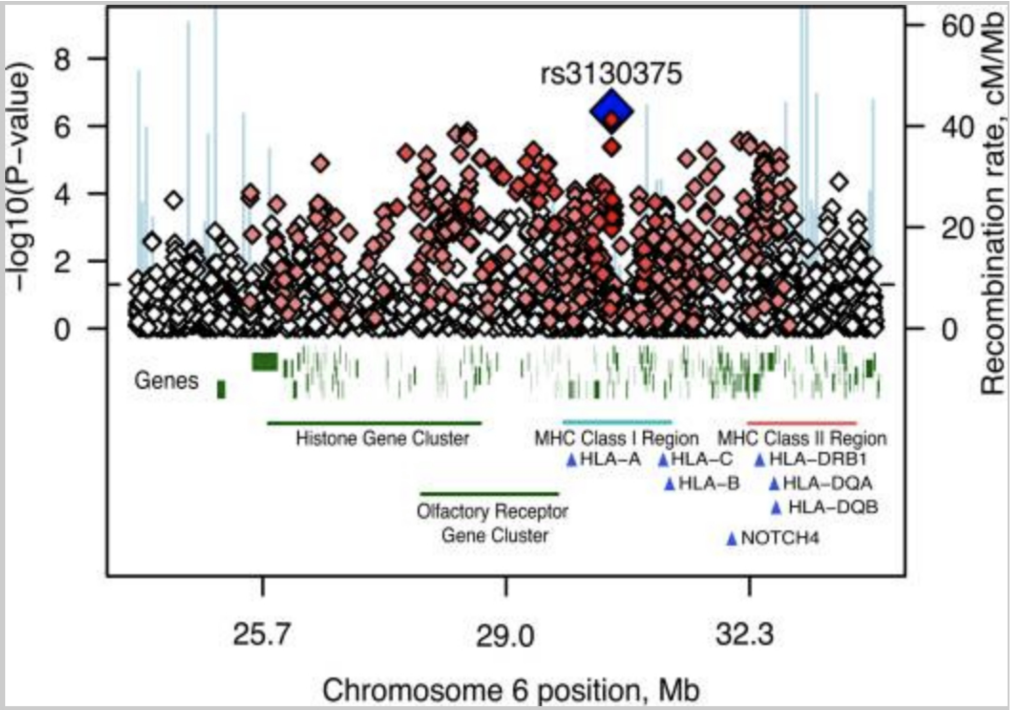


FIGURE 1.2: **Manhattan plot of the MHC region from Purcell et al., 2009**. Recombination rate is signified by the light blue bar graph.

The objective for schizophrenia GWA studies became clear, large consortia would have to be created in order to obtain the samples required to identify robust individual variants significantly associated to schizophrenia.

1.3.2 The creation of the Psychiatric Genomics Consortium (PGC)

The PGC was created in 2007 to investigate the genetics of complex psychiatric disorders, and its schizophrenia group currently has over 400 investigators from 40 different countries. The first GWA study from the PGC was published in 2011 (Ripke et al., 2011), and found seven loci associated to schizophrenia, five of which were novel (See Figure 1.3). In total (including replication), 36 studies including 17,836 cases and 33,859 controls were collected together and these samples have been used in all future PGC schizophrenia GWA studies since. In addition, the GWA study identified loci (rs4765905, rs10994359, rs2239547) that were susceptible to both schizophrenia and bipolar disorder, a finding which initialised a research direction into the combined genetics of various neuropsychiatric traits. The main finding of the GWA was that further samples were still needed in order to provide more than a handful of risk loci associated with schizophrenia, and indeed any significant findings from biological pathway analysis. Further information on PGC1 can be found in Chapter 2.

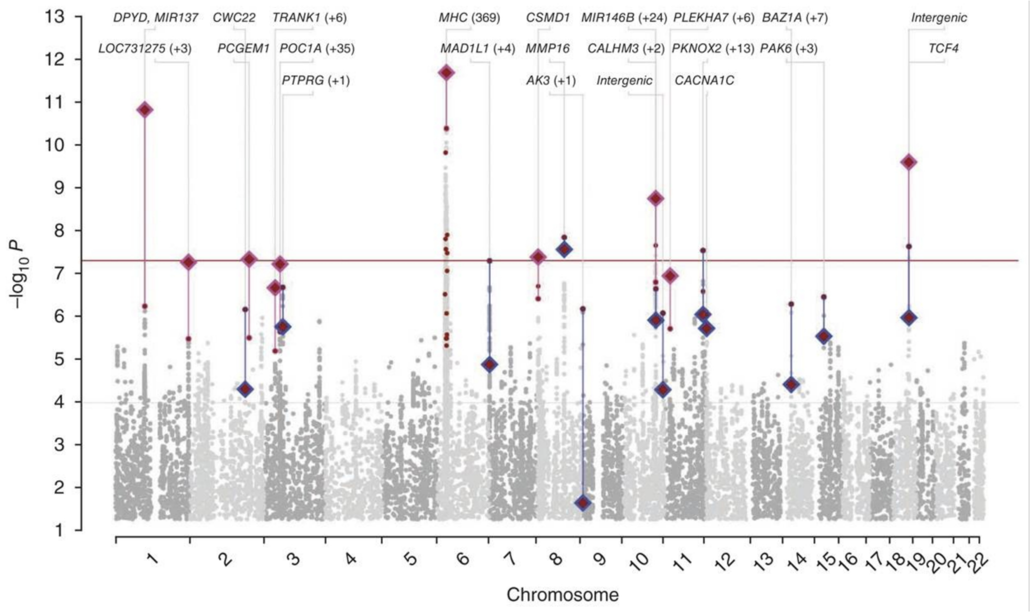


FIGURE 1.3: **Manhattan plot from Ripke et al., 2011**

In 2014, the PGC published their second GWA, PGC2 which identified 108 loci, including 83 novel associations (Ripke et al., 2014; See Figure 1.4 and for more information see Chapter 2). Of particular note within this GWA study was the observation that while all 108 loci could be credible linked to 108 SNPs, only 10 could be credibly associated to any non-synonymous exonic polymorphism (Ripke et al., 2014). This has led to further examinations of schizophrenia risk loci being associated to **e**xpression **Q**uantitative **T**rait **L**ocis (eQTLs)s and epigenetic markers (O'Brien et al., 2018; Wockner et al., 2015).

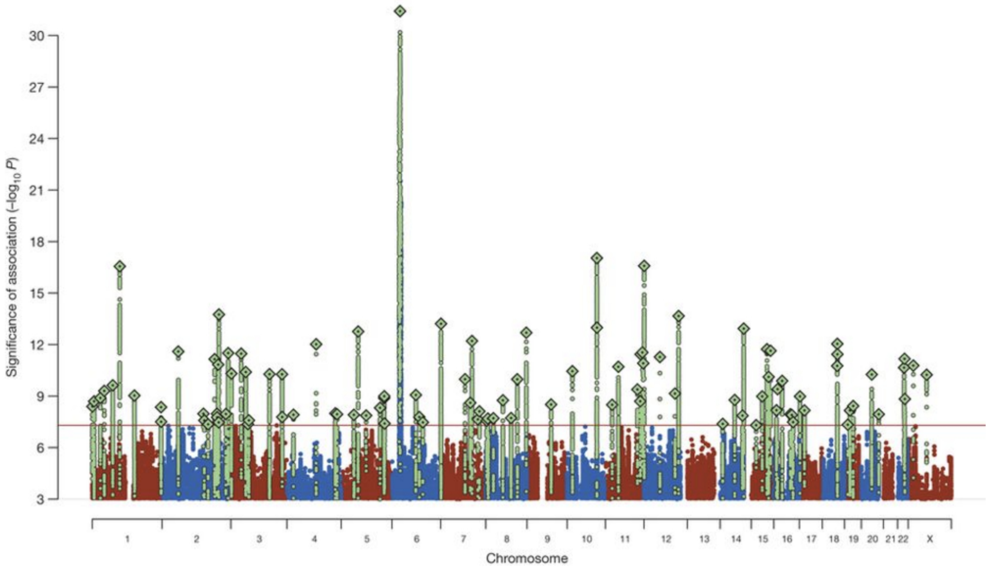


FIGURE 1.4: **Manhattan plot from Ripke et al., 2014**

The latest and largest schizophrenia GWA study at the time of writing, was published in 2018, which took advantage of samples obtained from the PGC and patient records of individuals within the UK who were prescribed clozapine (for more information see Chapter 2; Pardiñas et al., 2018). They found 145 genome-wide significant loci associated with schizophrenia and identified the first biological gene-sets associated to schizophrenia (Pardiñas et al., 2018; See Figure 1.5).

GWA studies have discovered a considerable number of schizophrenia risk loci with small individual effects. However, there were potentially large numbers of risk variants which did not surpass genome-wide significance, but may collectively contribute to schizophrenia risk. The PRS method was developed in order to examine the polygenic component of a particular disease or disorder.

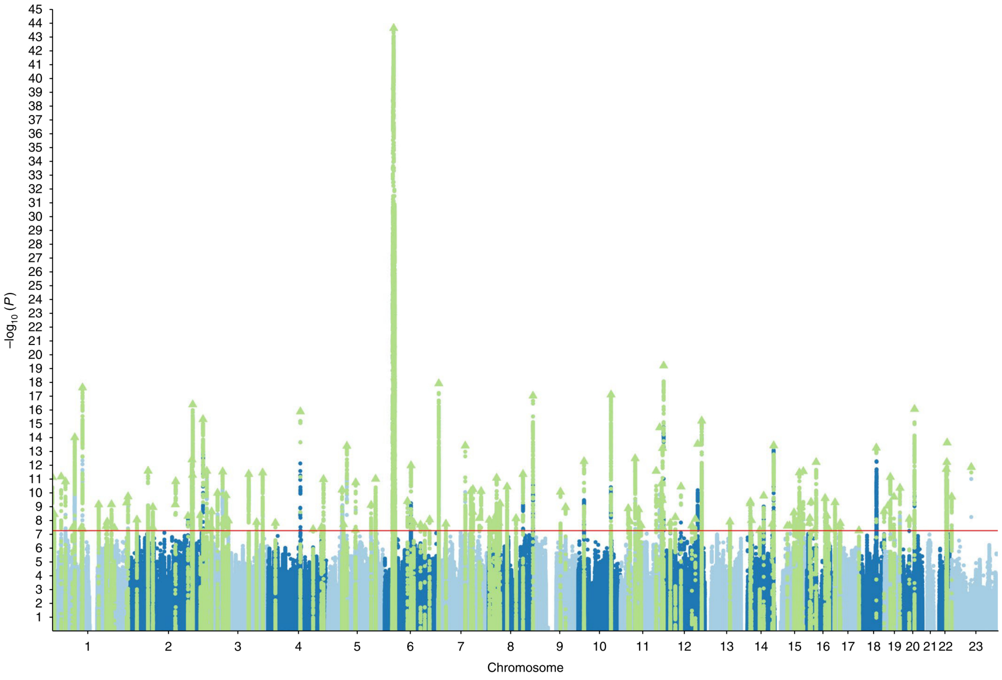


FIGURE 1.5: **Manhattan plot from Pardiñas et al., 2018**

1.3.3 Does rare and common variation in schizophrenia overlap?

One of the first studies to examine rare and common liability for schizophrenia was Purcell et al. (2014). They took 2,546 genes hypothesised to be enriched in mutations associated with schizophrenia, genome-wide CNV studies, GWA studies, and exome sequencing of de novo mutations and found that cases had a higher rate of rare disruptive mutations versus controls (Purcell et al., 2014). However, in a case only analysis of samples from PGC1, CNVs, SNVs and GWA studies were uncorrelated.

Recent evidence suggests a negative correlation between schizophrenia-associated CNV carrier status and the common risk variant burden. In Tansey et al. (2016) they aimed to decipher between two models of schizophrenia with respect to rare variant variation. One proposal was the extreme heterogeneity model, where it was proposed that schizophrenia is a collection of disparate set of distinct disorders among which a specific mutation would share a small, homogeneous sub-group. As schizophrenia is polygenic, it can only apply to those cases with highly penetrant mutations like CNVs. If the extreme heterogeneity hypothesis extends from alleles to pathophysiology, overlap should not be observed for these additional risk factors between carriers of different CNVs, or between carriers of CNVs and people with schizophrenia who do not carry CNVs. The other model is schizophrenia being a polygenic disorder where the disorder is the result of an accumulation of risk factors sufficient to surpass a threshold of disease liability. To examine these hypotheses, they evaluated whether individuals with a diagnosis of schizophrenia who carry a schizophrenia-associated CNV also share a common risk allele burden with those who have schizophrenia without a schizophrenia-associated CNV (Tansey et al., 2016).

First, it was found that both the schizophrenia PRS with CNV cases and the schizophrenia PRS without CNV cases could significantly differentiate from controls (where a significantly higher p-value was found with the schizophrenia PRS with cases ($P = 1.43 * 10^{287}$ versus without CNV $P = 2.25 * 10^{17}$) (Tansey et al., 2016). It was also found that within schizophrenia cases, common risk contributed to patients with a high OR CNV or a low OR CNV and, at most **P** value **t**hreshold (Pt) schizophrenia cases with a high OR CNV had a lower PRS for schizophrenia compared with a) cases without a known schizophrenia-associated CNV, and b) with cases with a lower OR CNV.

Bergen et al (2019) expanded on this analysis by assessing the relationship between three classes of CNVs (samples with CNVs associated with schizophrenia; CNVs that span over 500kb; total CNV burden) and their PRS. Mean PRS between study subjects with and without rare CNVs were compared. Logistic regression modelled the joint effects of PRS and CNVs on schizophrenia liability. Samples with schizophrenia-associated CNVs had a lower PRS in proportion of the effect size of the CNV (Bergen et al., 2019). For example, the strongest associated schizophrenia CNV, the 22q11.2 deletion, required little added effect from the PRS to reach a diagnosis of schizophrenia observed in the sample (Bergen et al., 2019). Large deletions and increased CNV burden were also associated with lower polygenic risk in schizophrenia case (Bergen et al., 2019).

with respect to de novo mutations in schizophrenia, Rees et al (2020) examined the relationship between de novo variant mutations and common risk using the polygenic transmission disequilibrium test (Weiner et al., 2017). Probands (the first individual with a suspected diagnosis of schizophrenia) carrying candidate schizophrenia-related de novo variants had a significantly lower mean polygenic transmission disequilibrium test than that of probands who did not carry one of these de novo variants (Rees et al., 2020). The overtransmission of common risk alleles from parents is about seven times as great to non-carriers than to carriers of candidate schizophrenia-related de novo variants (Rees et al., 2020).

1.3.4 Pathway analysis in schizophrenia GWAS

In 2015, Pocklington et al. (2015) derived 134 gene-sets relevant to the functioning and development of the nervous system based on earlier case-control CNV studies showing that case CNVs were enriched for synaptic and neurodevelopmental genes (Glessner et al., 2010; Walsh et al., 2008). They found that the gene-sets were enriched for CNVs in schizophrenia (Pocklington et al., 2015).

Given the evidence above indicating an overlap of rare and common variant variation in schizophrenia, Pardinas et al (2018) performed a gene-set analysis of these 134 gene-sets in the CLOZUK meta-analysis case/control GWA study using MAGMA (Leeuw et al., 2015). After multiple testing correction and stepwise conditional analysis, six gene-sets were found to be significantly associated with schizophrenia (Targets of FMRP, Abnormal behavior, 5-HT2C receptor complex, Abnormal nervous system electrophysiology, Voltage-gated calcium channel complexes, Abnormal long-term potentiation). In addition, recent studies also identified that mutation intolerant genes were enriched for schizophrenia CNVs, and a MAGMA gene-set analysis of a loss of function gene-set (n = 3,230) found that this gene set was also enriched for common variant variation as well (Pardiñas et al., 2018).

In Schijven et al (2018) further gene-set analysis was performed using MAGMA (Leeuw et al., 2015) on the MsigDB gene ontology database which found enrichment of common variants in synaptic plasticity and neuron differentiation gene sets (Leeuw et al., 2015; Liberzon et al., 2011). In support of these findings, they also performed gene set analysis using MAGMA, MAGENTA and INRICH (Leeuw et al., 2015; Segrè et al., 2010; Lee et al., 2012) on synaptic signalling pathways in KEGG (Kanehisa and Goto, 2000), and found further enrichment in dopaminergic and cholinergic synapses (2018).

Protein-protein interaction analysis (analysing whether the top unique genes in the gene sets show more direct/indirect interaction with each other and with other proteins than expected by chance) used on 22 unique genes from the KEGG dopaminergic, cholinergic or long-term potentiation pathways found more direct interactions with each other and more indirect interactions to other proteins than expected by chance (2018). However, this was not the case for direct interactions with other proteins.

Weighted **G**ene **C**o-expression **N**etwork **A**nalysis (WGCNA) was performed on PGC2 and found 12 gene co-expression modules with sizes ranging from 40 to 1813 genes (Radulescu et al., 2020). Briefly, WGCNA is a data-mining method which uses gene expression data to define a biological network between the genes based on the pairwise correlation (or co-expression) of the genes. Selected modules from the WGCNA network were tested for association with schizophrenia PRS, diagnosis, and genes containing GWA study significant loci within PGC2. One module was found to be associated to all three variables and contained genes involved in synaptic signaling and neuroplasticity (Radulescu et al., 2020).

1.4 Polygenic risk score definition

There are many aspects to a PRS, this section will be structured in the following format:

- Broad definition of a PRS
- The units/values that make up a PRS (the risk variants or SNPs)
- Which genetic model it describes
- How it accounts for LD
- How it accounts for genetic phenomena within a population over time
- Limitations of the method

1.4.1 Broad Definition

A standard PRS combines the effect sizes of all SNPs across the genome into one risk score for each individual (See Figure 1.6). The genetic burden of a particular disorder can then be assessed, and the ability of the risk score to predict the disease status of any one individual can be determined (Purcell et al., 2009).

$$\boxed{PRS_i = \frac{1}{m}\sum_{k}^{m} \beta_k \cdot n^{(i)}_{SNP_k}}$$

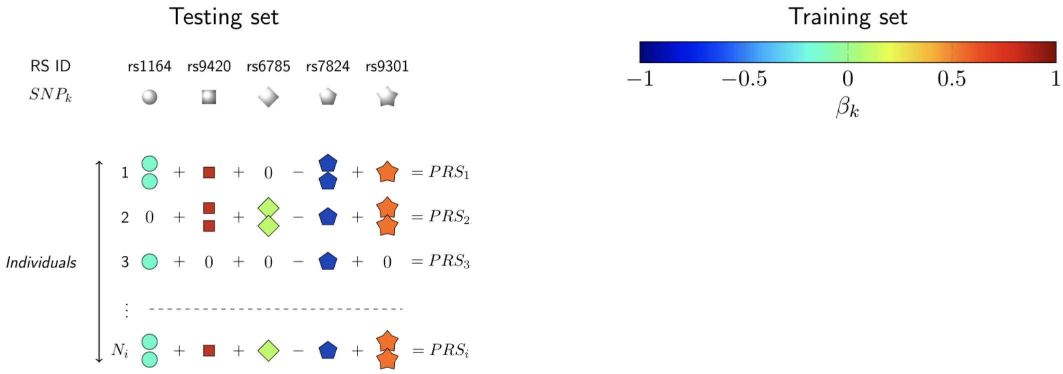


FIGURE 1.6: **Calculation of PRS per individual.** Effect sizes for each SNP k (β_k) are extracted from the training data set, usually limited by a p value threshold. The polygenic score for each individual i (PRS_i) is then calculated in the testing data set. For any individual at any selected SNP, the number of risk alleles the individual has is determined (0,1 or 2) and the dot product is calculated between all SNPs and their corresponding effect size. The polygenic score is the weighted sum of the individual's risk alleles. $\frac{1}{m}$ signifies the weighting of each score required to account for missing genotypes; in PLINK v1.90 the missing genotypes are estimated using 2 × Minor Allele Frequency (MAF). Each risk allele is represented by a different shape and the effect sizes are represented by a colour corresponding to the training set colour bar.

The PRS is in essence, the weighted sum of the individual's risk alleles. The terminology 'risk' within the definition of 'polygenic risk score' can be adjusted depending on which biological trait the PRS is defining. For example, it is logical to refer to the weighted sum of schizophrenia alleles as a polygenic risk score because the trait produces symptoms which impede normal biological functioning. However, if the PRS was describing the trait of height, it makes more biological sense to label this as a 'polygenic predisposition score', because differences in height in a population do not impede normal biological functioning in humans. To reduce confusion for the reader, all polygenic scores will be referred to with the short-hand PRS.

1.4.2 SNPs in a Polygenic Risk Score

At any locus on an individual's genome (their SNP), the individual may carry between 0, 1 or 2 risk alleles. The determination of whether the individual carries a risk allele is inferred by using an independent cohort (the training set as described in Figure 1.6). Briefly, at any singular locus, there will be a major allele and a minor allele within the cohort. The major allele is the allele which is found to be the most frequent in that cohort (for example: over 50% for a bi-allelic locus). The declaration of a major allele is binary, the size of the cohort and the magnitude of the percentage difference between the major allele and other alleles has no bearing of the allele being declared as a major allele within the cohort.

Conversely, the minor allele is the allele at that specific locus that is the second most common within the cohort and is usually the frequency of this allele that is used to calculate the **M**inor **A**llele **F**requency (MAF) at each SNP within the PRS. The frequency of the major allele can be inferred from the MAF if the site is bi-allelic.

A 'risk allele' is determined when, at that particular locus, a GWA study has been performed on the cohort for a particular trait and it has been found that the allele is statistically significantly associated with that trait. The calculation to determine significance is usually the log of the OR (quantification of the strength of association between group A (e.g. cases of schizophrenia) over group B (e.g. control cohort)). In the case where the trait is continuous, the calculation to determine significance is usually a beta coefficient extracted from the fit of a predefined statistical model suiting the trait in question. The log is applied to OR to temper the difference between the relative difference in probability between the two groups and the resulting odds ratio produced. For example, if looking at a range of schizophrenia risk alleles and comparing across only their odds ratios, the magnitude in the difference of odds ratios between each risk allele may be significantly larger than the magnitude of the actual differences in effect sizes. This could encourage an over-emphasis of schizophrenia risk alleles at the extreme ends of the distribution of effect sizes. By taking the log OR, these magnitudes of the differences between odds ratios are transformed to be symmetrical to the magnitudes of differences to their respective effect sizes.

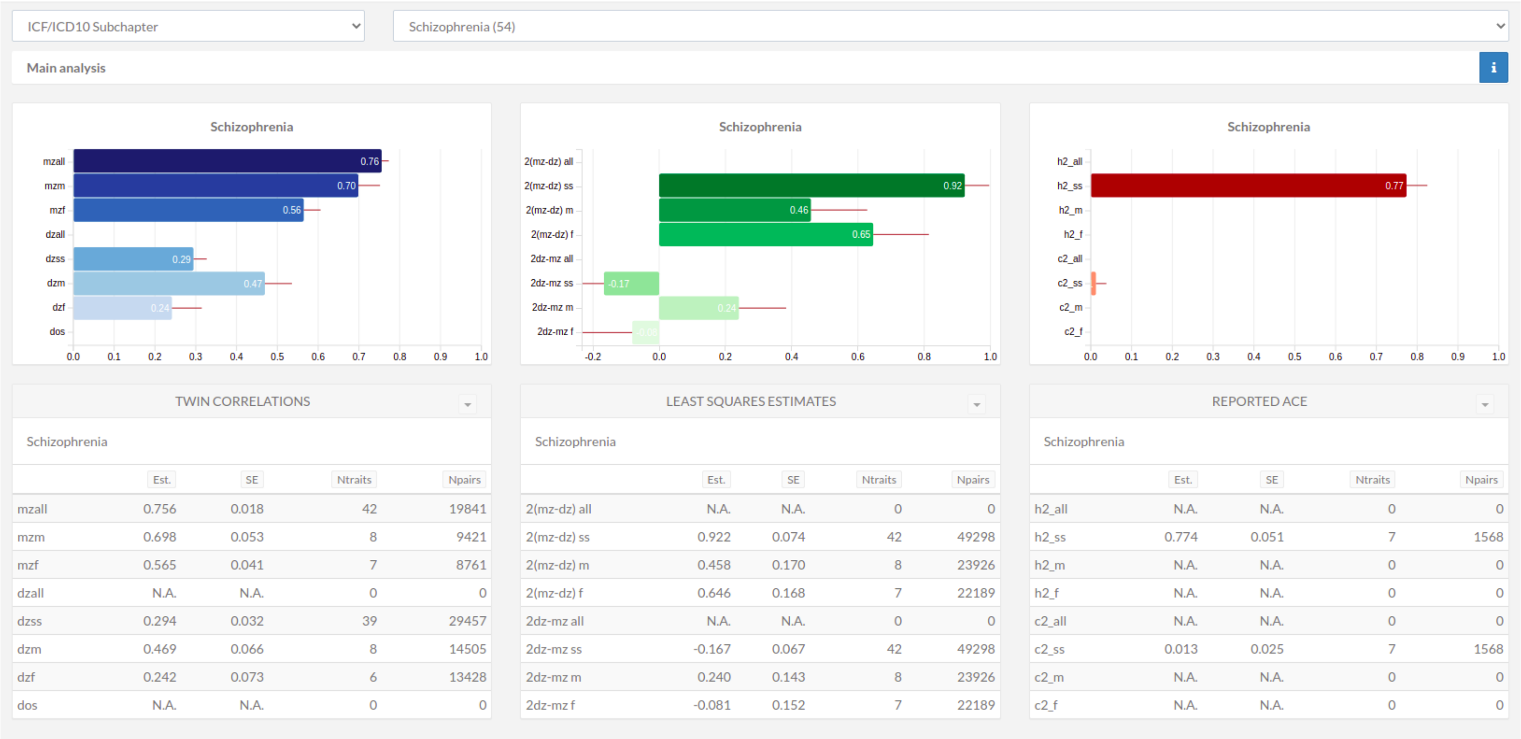
If the allele in question has a log OR more than one and is significant, the allele is likely to be the minor allele because there will be selective pressure against that allele. The situation in which the major allele may be the risk allele at the specified locus is when the odds ratio is less than one combined with a statistically significant p-value and so the minor allele is protective against that trait. However, it is usually preferably interpreted that the minor allele is protective rather than the major allele is a risk allele.

The sum of the number of risk alleles weighted by the log OR or effect sizes of these risk alleles is a singular score per individual that represents the genetic loading for that particular trait or disease (Lewis and Vassos, 2020). This statistical model of an individual's genetic predisposition for a trait assumes that the trait in question has an additive genetic architecture and that the individuals risk alleles are independent from one another.

1.4.3 Genetic model of schizophrenia

Polderman et al (2015a) supports the theory that the genetic architecture of schizophrenia is additive. They examined the heritability of thousands of traits over 50 years of twin studies. Within the trait of schizophrenia, Polderman et al. (2015a) collated the correlation metrics provided for both monozygotic twins and dizygotic twins across 54 studies. All these studies had to provide a correlation metric that was either intraclass, Pearson, polychoric or tetrachoric correlations. If the twin correlations did not exist, then they estimated the correlation based on least-squares or maximum-likelihood methods. In the twin study design method, there are two extreme hypotheses that you can draw based on the correlations for monozygotic and dizygotic twin pairs. If the correlations within the monozygotic and dizygotic twin populations are the same, in which case the phenotypes observed in each twin are entirely caused by non-genetic factors. If the ratio of correlations in the monozygotic twin population compared to the dizygotic twin population is 2:1, this indicates that the phenotypes observed in each twin is solely caused by additive genetic factors. If the ratio is in between these two metrics, it is inferred that there are shared genetic and environmental factors influencing the trait.

For the majority of traits, it was found that in 84% of cases, the correlation of the monozygotic twins was higher than the dizygotic twins correlation metric. In the case of schizophrenia, there was not enough information to provide an all encompassing metric, but for same sex pairs, the metric of 2rDZ - rMZ was -0.17, and the correlations for both male and female monozygotic twins was higher than that of the correlations for the dizygotic twins (See Figure 1.7).



TWIN CORRELATIONS

Schizophrenia

	Est.	SE	Ntraits	Npairs
mzall	0.756	0.018	42	19841
mzm	0.698	0.053	8	9421
mzf	0.565	0.041	7	8761
dzall	N.A.	N.A.	0	0
dzss	0.294	0.032	39	29457
dzm	0.469	0.066	8	14505
dzf	0.242	0.073	6	13428
dos	N.A.	N.A.	0	0

LEAST SQUARES ESTIMATES

Schizophrenia

	Est.	SE	Ntraits	Npairs
2(mz-dz) all	N.A.	N.A.	0	0
2(mz-dz) ss	0.922	0.074	42	49298
2(mz-dz) m	0.458	0.170	8	23926
2(mz-dz) f	0.646	0.168	7	22189
2dz-mz all	N.A.	N.A.	0	0
2dz-mz ss	-0.167	0.067	42	49298
2dz-mz m	0.240	0.143	8	23926
2dz-mz f	-0.081	0.152	7	22189

REPORTED ACE

Schizophrenia

	Est.	SE	Ntraits	Npairs
h2_all	N.A.	N.A.	0	0
h2_ss	0.774	0.051	7	1568
h2_m	N.A.	N.A.	0	0
h2_f	N.A.	N.A.	0	0
c2_all	N.A.	N.A.	0	0
c2_ss	0.013	0.025	7	1568
c2_m	N.A.	N.A.	0	0
c2_f	N.A.	N.A.	0	0

FIGURE 1.7: **Summary of Twin study papers on the trait of Schizophrenia (A)** Twin correlations estimates for schizophrenia. Estimates were correlation coefficients derived from the DerSimonian-Laird random effect meta-analytical approach. (Shulze, 2004). mzall = mztc for males and females, mzm = mztc for males, mzf = mztc for females, dzall = dztc for males and females, dzss = dztc for twin pairs of the same sex, dzm = dztc for males, dzf = dztc for females, dos = dztc for opposite sex twin pairs. **(B)** Least squares estimate for the relationship between monozygotic and dizygotic twin pairs. 2(mz-dz) is equivalent to subtracting the dztc from the mztc and multiplying by 2. 2dz-mz is equivalent to subtracting the mztc from the dztc multiplied by two. These estimates provide an estimation of heritability (2(mz-dz)) and the shared environment (2dz-mz) directly from the twin correlations. **(C)** ACE model (A = Additive genetic variance, C = environmental factors, E = measurement error) estimates for schizophrenia. H2 = Heritability, C2 = estimate of the shared environment.

1.4.4 Accounting for Linkage Disequilibrium

For the PRS to be accurate, there is an assumption that each SNP is independent from one another, and must therefore take into account the genetic phenomena known as LD. Briefly, LD is the correlation structure between SNPs or the non-random association between SNPs at separate loci in a population (Visscher et al., 2012). It occurs due to the process of many genetic phenomena including genetic drift, mutation rates, population structure and genetic recombination which all occur over several generations within a population. The important information to note is that if two SNPs are found to be in LD with each other, then they are no longer independent and the model of a PRS is no longer accurate. The issue of accounting for LD is a contentious one within PRS calculations, with some methods choosing to model the linkage dis-equilibrium across all SNPs (LDpred method reference), while other methods (PLINK v1.90 and PRSice2), choosing to remove the SNPs that show LD signal within a specified range of values.

A large number of the SNPs will likely be removed from the input data sets, and there may be small correlations between the remaining SNPs.

1.4.5 Population genetics of schizophrenia

A note that particularly within the trait of schizophrenia, the additive evolutionary model is vastly overly-simplified and there is still some confusion as to why schizophrenia alleles persist in the population, despite evidence of high heritability and low fecundity within schizophrenia patients (Power et al., 2013). In evolutionary theory, this presents a paradox as schizophrenia risk alleles should have been eliminated through the process of negative selection. Alternative evolutionary theories indicate an influence of positive selection (Fujito et al., 2018), balancing selection (Sato and Kawata, 2018), and background selection (Pardiñas et al., 2018).

A few studies suggest the theory that schizophrenia alleles have, at some point in the history of human evolution, provided some form of benefit for human survival and are therefore maintained through positive selection. For example, the gene sequence 1.3kb upstream of the ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2 (ST8SIA2) gene contains three SNPs associated with schizophrenia (Fujito et al., 2018). ST8SIA2 encodes a sialtransferase that is responsible for the production of polysialic acid (PSA), which has many important functions within the brain including cell-cell communication and function of ion channels. Because the functional biological applications of ST8SIA2 are currently evidenced to be pervasive throughout the brain and important, schizophrenia risk alleles could be selected for, as biological function out-weights the risk of obtaining schizophrenia phenotypes. However, in this particular study, the display of positive selection assumes a functional link between these three schizophrenia associated SNPs and the biological outcome, of which the only evidence supplied is decreased social motivation and increased aggressive behavior in two mouse models (Kröcher et al., 2015; Calandreau et al., 2010).

1.4.6 Limitations of the PRS method

As described in Figure 1.6, a PRS requires the use of two independent data sets, the testing set which must contain genotype level data (See Chapter 2 for more information) and the training set which is usually a published GWA study. A limitation of this approach is that genotype level data is rarely publicly available due to patient confidentiality. Additionally, another limitation surrounds shared samples between data sets. A PRS requires no overlapping samples between testing and training data sets, but samples can be hard to identify in the training set as the structure of the data set does not include sample information. Sample identification is usually inferred by the paper which was published alongside the data. As observed in Chapter 2, the cohorts used to create each data set can become increasingly complicated, especially if multiple data sets are resourced from the same consortium or group.

1.5 Polygenic risk score Application

PRS can be applied over several different methodologies, traits and populations. Below I will describe the:

- Early PRS application
- Advantages of PRS application
- Uses within a clinical setting
  - Prediction of a clinical outcome

  - Prediction across ethnic groups
  - Prediction of schizophrenia
- Relationship of schizophrenia PRS with other traits

1.5.1 Early application of PRS

At the same time as realising that GWA studies required more power to capture the common genetic component of schizophrenia risk in the study from the International Schizophrenia Consortium (Purcell et al., 2009), another line of thought was whether there was a polygenic component for schizophrenia as first described by Gottesman and Shields (1967). If this was the case, then even variants which did not reach genome-wide significance could contribute useful information on the common genetic disposition towards schizophrenia. In this instance, Purcell et al. (2009) first selected variants from within the schizophrenia GWA at various predefined significance thresholds from here-on referred to as a Pt. They then used these 'score alleles' to generate aggregate risk scores in an independent target data set. The alleles are referred to as 'score alleles' because of the inability to differentiate between the true risk alleles from within the schizophrenia GWA study from variants unassociated to schizophrenia (Purcell et al., 2009). The polygenic component was shown to be highly associated with schizophrenia ($p = 1.9 \times 10^{-19}$) (Purcell et al., 2009). In PGC2, another PRS was created and confirmed that the PRS was associated to schizophrenia, was able to predict case/control status and was estimated to explain around 7% of the liability variation in schizophrenia (Ripke et al., 2014). Over time, the method behind the PRS became more solidified into the procedure as described in Figure 1.6.

1.5.2 Advantages of PRS application

The main advantage of using PRSs is it's ability to produce a single risk score per individual. The score can be used as a variable to answer scientific hypotheses relevant to the trait from which the PRS was derived. In addition, a PRS uses two datasets in it's creation, one dataset contains the list of SNPs per individual (testing dataset) and the other dataset contains the effect sizes for each SNP (training dataset). Individuals can therefore be stratified into clinical groups using the effect sizes as a quantitative measurement for the boundary between groupings. In the case of schizophrenia, the PRS can be assessed for its ability to differentiate between patients and controls within the testing data set, using the effect sizes from the training set (Dudbridge, 2013). PRS indexes common variant liability to a given disorder and therefore, cross-disorder analysis is viable. If, for example, an individual was interested in investigating the genetic risk for schizophrenia in patients diagnosed with bipolar disorder, a PRS would be a useful analysis method (Lee et al., 2013; Ward et al., 2017).

1.5.3 PRS in a clinical setting

PRS also has potential in a clinical setting, as it is an early quantitative measurement that can be recorded before disease onset has occurred. However, a PRS measures the risk of developing a disease; A PRS does not display a binary decision on whether the individual will obtain a disease (Sugrue and Desikan, 2019). This has implications in it's applicability within a clinical setting. For example, information which informs the risk of disease is based on GWA studies which measures the prevalence of the disease, aka the number of cases of who have the disease over a specified point in time.

Prediction of a clinical outcome

However, in a clinical setting, a PRS would attempt to predict the incidence of the disease, aka the number of new cases of the disease (Sugrue and Desikan, 2019; Noordzij et al., 2010). The PRS may therefore be inaccurate if the incidence in the disease changes as this may indicate that the prevalence data on which the PRS is derived, no longer represents the proportion of the disease in the population. The evaluation on how well a PRS predicts the clinical outcome of any disease is under debate. Many studies use diagnostic predictive measures including area under the curve (Escott-Price et al., 2019), and positive predictive value and negative predictive value (Li et al., 2021). These measures evaluate the predictive performance of PRS to discriminate between groups with disease vs groups without disease. Other methods are in development including, for example, a method applied in Seibert et al. (2018) which uses a modified form of a PRS to assess the risk of obtaining **a**ggressive **P**rostate **C**ancer (PCa). Diagnosis of PCa has a high false positive rate and screening procedures often cannot separate aggressive PCa patients from patients with indolent disease. Therefore, the aim would be to reduce unnecessary screening of these false positives patients, while still identifying patients who are at a high risk of obtaining PCa.

Seibert et al. (2018) used a method from Desikan et al. (2017), which identifies the SNPs that are associated with a high risk of the disease, and then applies Cox proportional hazard models (usually used in survival analysis) to reduce the previous list of SNPs to a list of SNPs which is associated with a lower survival rate for patients with the disease.

Briefly, survival analysis is an all-encompassing term for a range of statistical tests where the outcome variable is the amount of time until an event occurs. In respect to PCa, most likely the events would be the time of death.

While Seibert et al. (2018) displayed that a modified PRS does predict age at onset of PCa but, one of the major limitations is statistical power, especially due to the fact that the development dataset was a collation of several studies of varied design.

The most promising uses of PRS with risk prediction of complex disorder are similar to the example above, where the PRS has been used as an addition to other clinical risk factors. Using PRS alone does not usually provide enough predictive power over other clinical risk factors. For example, the combination of lifestyle, biochemistry, clinical, and historical risk factors produced an **A**rea **U**nder the **C**urve (AUC) of 82% when predicting the 10 year risk of cardiovascular disease (Assmann, Cullen, and Schulte, 2002). In most instances, a PRS alone does not provide more predictive power, but in some circumstances, a PRS can provide more predictive power than previously used predictive measurements. For example, in **A**nkylosing **S**pondylitis (AS), a PRS was shown to have a higher AUC than genetic testing at a single locus (HLA-B27; AUC = 0.869), measurements of acute phase reactants (AUC = 0.7) and MRI imaging of sacroiliac joints (AUC = 0.885) (Li et al., 2021). However, one of the limitations of (2021) was that the predictive power differed between different ethnic populations.

Prediction across ethnic groups

The application of PRS across different ethnic groups is a challenge when using it within a clinical setting. Unfortunately, most GWA studies which provide the training set for the PRS, come from a European ancestry (Lewis and Vassos, 2020). Therefore, the predictive accuracy of the PRS outside of a European population may be attenuated due to tagging SNPs (only genotype a single SNP representing a region of SNPs located close to the genotyped SNP), differences in patterns of LD, and potential differential genetic drift between the ethnicities which may bias the PRS. However, in the case for schizophrenia, a large-scale GWA study for East Asian populations has been conducted and compared to a European cohort (Lam et al., 2019a).

Prediction of schizophrenia

A schizophrenia PRS has been shown to display moderate predictive ability (AUC = 0.61), but this signal is too low for clinical utility. However, a schizophrenia PRS could provide extra information in stratifying patients with schizophrenia. For example, in a study by Vassos et al. (2017) examining first-episode psychosis patients, they found that patients who went on to develop schizophrenia (n = 86) had a higher PRS than those who went on to develop another form of psychosis (n = 65; Nagelkerke's R2 of 9%). While the predictive ability is still low, and there was quite a low sample size in this experiment, the advantage of this experimental design is that it a) only requires the genotyping of individuals presenting with psychosis, and b) the clinical outcome is not a binary treat/not treat (Lewis and Vassos, 2020).

1.5.4 Prediction of PRS in complex traits

Prediction methods

Before analysing prediction across multiple traits, it is prudent to explain the metrics which confers how well the PRS predicts any singular trait. Most research-based PRS prediction uses population genetics over a singular individual.

When assessing either case/control status or a continuous trait, sometimes the variance (R_2) is used (Lewis and Vassos, 2020). For a continuous variable, the R^2 from a linear regression is reported and it captures the proportion of variance which is explained by the PRS. For case/control status, a logistic regression is used as the outcome is binary. To get an estimate of how much variance a PRS captures for a binary outcome, the Nagelkerke R^2 is reported as it is the most comparable to the R^2 in a linear regression. In some cases (because the prevalence of cases and controls in the sample may not match the prevalence of cases and controls in the population) the Nagelkerke R^2 is converted to the liability scale as defined in Lee et. al. (2012).

The predictive ability of a PRS can be defined in terms of AUC which takes a value of 0.5 to 1 (Lewis and Vassos, 2020). This gives an overall summary of the predictive ability of the model where the closer the value is to 1, the better the predictive ability of the PRS (and other risk factors (eg age and sex) if included in the model). For example, when predicting case/control status, it is the probability that a randomly selected case will have a higher PRS than a randomly selected control.

Odds ratios are frequently used if analysing how well PRS can predict differences in risk between sub-groups in the population (Lewis and Vassos, 2020). The predictive ability can be defined as the proportion of the population which has a k-fold increased odds (eg $k = 2, 3, 4...$) compared to the disease risk in the population. The PRS can also be split into deciles, quartiles and/or quantiles and the predictive ability of the PRS can be defined as the odds ratio of disease for an individual in the top (eg) decile compared to individuals in another section of the distribution e.g.(0-90%, 0-10% or 30-60%).

Traits outside the brain

Height has been an interest in genetic prediction as it can be used as a model for complex trait prediction. Between 2014 and 2018, SNP heritability on it's own could explain between 17-19% of the variability in height, but in 2017, Lippert et al (2017) displayed that selected SNPs could explain 53% of the variability in height, as sex was included in the model and the population sample was diverse in ancestry (You et al., 2021). The best prediction so far, used a PRS which on it's own provided an R_2 value of 0.73. If including sex, parental height and PRSs of both parents and proband, the R^2 increased to 0.82 (You et al., 2021).

The best prediction of Type II diabetes was achieved by Liu et al. (2021), with a top AUC of 0.901 [95% CI: (0.790, 0.800)]. When the prediction was solely based on the PRS, the AUC was 0.749. Covariates including, but not limited to, age, sex, the first 10 population principal components, **B**ody **M**ass **I**ndex (BMI), diastolic blood pressure, glucose level, and cholesterol level accounted for the difference in these two AUC metrics showing that a combination of genetic and associated environmental factors could be used as an accurate predictor of Type II diabetes.

Another study by Khera et al. (Khera et al., 2018) compared the AUCs across the PRS of five different traits including Type II diabetes (AUC = 0.734), Coronary artery disease (AUC = 0.813), Atrial fibrillation (AUC = 0.782), inflammatory bowel disease (AUC = 0.648) and Breast cancer (AUC = 0.695). The idea was to test whether PRS conferred comparable or better power to separate out individuals with a high risk of the trait in a population, over rare monogenic mutations. For all five cases this was true as PRS identified 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively. For coronary artery disease, the prevalence shown by the PRS is 20-fold higher than carrier frequency of familial hypercholesterolemia (rare monogenic) mutations conferring comparable risk in previous studies (Khera et al., 2018).

Neuropsychiatric traits

Using the largest GWA study at the time of writing, a PRS was created to predict the prevalence of **M**ajor **D**epressive **Disorder** (MDD) in a European/North American population (Wray et al., 2018). The PRS explained 1.9% of the variance in liability and the odds ratio of MDD liability for the tenth decile versus the first was 2.4 (Wray et al., 2018).

In a similar pioneering study for ASD, the IPsych ASD GWA study (see Chapter 2) was split into five separate sets of testing and training samples to test the prediction of the ASD PRS (Grove et al., 2019). The observed Nagelkerke's R^2 explained by the PRS was 2.45%. Grove et al (2019) then improved the predictive accuracy (Nakelkerke's $R^2 = 3.77\%$) by including PRS from other traits that were correlated with ASD. These included but were not limited to: schizophrenia, depressive symptoms, **A**ttention **D**eficit **H**yperactivity **D**isorder (ADHD), MDD, extraversion, agreeableness and childhood intelligence (Grove et al., 2019). The choice of traits was determined by selecting the traits which containing the highest Nagelkerke's R^2 when the trait was predicting ASD.

A study by Vassos et al. (Vassos et al., 2017) examined whether a PRS could separate case/control status of first-degree psychosis patients, and whether it could stratify which patients had a diagnosis of schizophrenia or a diagnosis of other psychoses. For the testing set, the Biomedical Research Centre (BRC) for Mental Health Genetics and Psychosis (GAP) study was used whereby participants were selected for patients with a diagnosis of first episode of nonorganic psychosis and control participants. The training set for all PRS was PGC2 (see Chapter 2) split between a cohort with European ancestry and African ancestry. The PRS explained 3.6% of the variance in case/control status for first degree psychosis using Nagelkerke's R_2 on the liability scale. When stratified, the PRS explained 9.4% of the variance in European ancestry but was only predictive in African ancestry (1.1%) when more samples were added. When only looking at cases within the samples (aka all patients meeting a diagnosis of non-organic psychosis), the prediction was split to discern between cases with a diagnosis of schizophrenia versus cases who never met the diagnostic critera for schizophrenia. It was found that schizophrenia cases had a higher PRS than those cases with other psychoses (Nagelkerke's $R^2 = 9.2\%$).

In the latest GWA study of Bipolar disorder containing 41,917 bipolar disorder cases and 371,549 controls of European ancestry, it was found that the PRS explained 4% of phenotypic variance in these samples (Mullins et al., 2021).

1.5.5 PRS and the genetic architecture of schizophrenia

Schizophrenia PRS has been found to correlate with the length of hospitalisation for schizophrenia patients within inpatient admissions. Furthermore, at supported housing facilities, and the schizophrenia PRS for chronically ill resident schizophrenia patients was substantially higher than the PRS of resident schizophrenia patients who were less severely ill.

There has been a lot of research into the association of schizophrenia PRS and treatment resistance to antipsychotics, but the evidence has been conflicting. There have been reports that patients with treatment resistant antipsychotic treatment (Clozapine) had a positive association to a high schizophrenia PRS (Frank et al., 2015) and first-episode psychosis patients with treatment resistance tended to have a high schizophrenia PRS (Zhang et al., 2019). However, further studies have found no association of schizophrenia PRS with treatment response, but a possible association to premorbid IQ and earlier age of onset within treatment response patients (Legge et al., 2020; Kowalec et al., 2021).

A schizophrenia PRS has also been associated with lower cognition in multiple studies (Ripke et al., 2014; Pardiñas et al., 2018; Hubbard et al., 2016; Mistry et al., 2018). However, both the direction of effect and the strength of the association changes between different definitions of cognition (eg premorbid IQ, educational attainment) (Richards et al., 2020; Dickinson et al., 2020). Intermediate phenotypes of schizophrenia including neuroimaging measures has been investigated, but there has not been many studies displaying strong evidence of associations between schizophrenia PRS and structural changes within the brain. Cao et al. (2021) did however show significant associations between schizophrenia PRS and lower functional connectivity (aka visual system, default-mode system, frontoparietal system) using network-based statistic analysis.

Schizophrenia PRS has been found to be associated with negative and disorganised symptoms of schizophrenia (e.g. withdrawal, loss of motivation, loss of concentration), albeit with a small reported amount of variance explained, but no association has been found of schizophrenia to its positive symptoms (e.g. hallucinations, delusions) (Legge et al., 2021). Schizophrenia PRS has been associated with psychotic symptoms in first-episode psychosis and bipolar samples (Legge et al., 2021). There is a suggestion that individuals with a high schizophrenia PRS may be suggestive of treatment non-response in MDD. But the association of the PRS to non-response was only nominal (p =0.003).

The genetic architectures between schizophrenia and other disorders has been investigated using PRSs. Individuals with neuropsychiatric disorders including bipolar disorder, schizoaffective disorder and depression have all been found to have an elevated schizophrenia PRS (Mistry et al., 2018; Hamshere et al., 2011; Tesli et al., 2014; Milaneschi et al., 2016). The schizophrenia PRS explained between 1% to 6% of the variation in psychiatric disorders including, but not limited to: depression and bipolar disorder (Mistry et al., 2018).

Both schizophrenia and bipolar PRS were able to distinguish patients with broadly defined psychosis and their unaffected relatives from controls, schizophrenia PRS and Bipolar PRS explained 9% and 2% of the variation in psychosis risk respectively.However, the separation was modest for distinguishing between unaffected relatives (schizophrenia PRS P-value = $1.2 * 10^{-}4$, Bipolar PRS P-value = $2.1 * 10^{-}2$).

The genetic overlap between schizophrenia and multiple other disorders suggests that focusing the schizophrenia PRS on specific genes/pathways may be a viable research approach when examining specific biological effects associated with schizophrenia.

1.5.6 Advantages of PRS application in schizophrenia

Schizophrenia is highly heritable (current estimate is approximately 80%) and the PRS of schizophrenia has repeatedly been shown to be associated with schizophrenia in independent samples and individuals with other neuropsychiatric traits and disorders. This provides a quantitative measurement of any individual's liability to schizophrenia in the form of a single number.

The power of schizophrenia PRSs is substantially higher compared to other neuropsychiatric complex traits and disorders, with the latest schizophrenia dataset at the time of writing containing approximately 40,000 cases and 64,000 controls (Pardiñas et al., 2018). In addition, the schizophrenia PRS is one of the best performing PRSs in terms of phenotypic variance explained (7%) as compared to other neuropsychiatric complex traits (Ripke et al., 2014).

1.5.7 Limitations of PRS application in schizophrenia

Schizophrenia is a broad description of multiple facets and traits which makes it a difficult disorder to dissect whether genetic mutations are causal towards it (Jablensky, 2010). This is evident in the clinical presentation and the illness course of schizophrenia where there is substantial heterogeneity between patients (Jonas et al., 2019).

PRS was found to be the most important predictor of case/control status out of various traits (including educational attainment, sex, parental depression) using permutation feature importance. It was also shown that various machine learning models including LASSO and ridge-penalised logistic regression, support vector machines (SVM), random forests, boosting, neural networks and stacked models did not add substantial value over a logistic regression of case/control status (Bracher-Smith et al., 2022). In schizophrenia, it appears difficult to improve on the PRS method or incorporate factors into the PRS that could capture more variation in the liability to schizophrenia. This limits the progress and identifying the endophenotypes of schizophrenia, and subsequently, hinders the ability to stratify individuals with schizophrenia. For example, some studies have suggested the schizophrenia PRS is sensitive to the positive symptoms of schizophrenia (Allardyce et al., 2017), is associated to the negative symptoms of schizophrenia (Jones et al., 2016) or not associated to either (Derks et al., 2012).

The conflict in associations may also be due to the fact that the schizophrenia PRS only captures a median of 7% of the variance in liability to schizophrenia when the SNP-based heritability is 24% and the estimated total genetic liability is 80% (Fusar-Poli et al., 2022). In addition, the schizophrenia PRS does not capture rare or structural changes which contibute towards the genetic liability including CNVs.

1.6 Deriving Polygenic risk scores

Structure of this section:

- Issues with PRS derivation
  - Sample Overlap (GWA studies)
  - Sample Overlap (genotype and training set)
  - Population genetics
  - Set-based Analyses within PRS
- Types of PRSs
- Performing and comparing PRS across multiple data sets.
- Generating gene-set PRS
- Interpretation of gene-set PRS

1.6.1 Issues with PRS Derivation

There are many systematic problems that must be at least considered before performing PRS analyses. These issues include sample overlap between GWA studies (Lin and Sullivan, 2009; LeBlanc et al., 2018), sample overlap between the genotype data and the training set (aka the GWA study) (Choi, Mak, and O'Reilly, 2020; Wray et al., 2013), population structure (Sul, Martin, and Eskin, 2018; Márquez-Luna et al., 2017) and integration of set-based analyses (Baker et al., 2018).

Sample Overlap

LeBlanc et al. (2018) provided an example of when overlapping samples could produce spurious correlation in a bi-variate analysis between two GWA studies. 100,000 SNPs were simulated with a randomised MAF for two studies. Within these simulated studies, 5,000 subjects were shared out of a total sample size of 12,000 subjects. After performing a GWA analysis on the SNPs the two studies independently, a uniform distribution of p-values were produced. However, (LeBlanc et al., 2018) shows that if you select SNPs from one study (Study 2) based on the observed SNPs in the other study that have significant p-values, the distribution of SNPs in study two is no longer uniform and shows inflated p-values.

Sample overlap within GWA studies therefore affects the genetic signal within the training set of the PRS, if the training set was a meta-analysis of several GWA studies. Whether this affects the accuracy of the PRS itself is debatable and depends on how the GWA meta-analysis was derived. The overlap of samples between GWA studies does not violate the assumption that there must be no sample overlap between the testing and the training set within the PRS, but it may violate the assumption that all SNPs are independent of each other within the PRS. This is because within a GWA study, the association of each SNP to the specified trait is performed one SNP at a time, and therefore any correlation between SNPs is not accounted for (Choi, Mak, and O'Reilly, 2020). However, most PRS analyses include a procedure that accounts for correlated SNPs, which may remove any erroneous SNPs due to sample overlap, but these procedures are designed to account for the genetic phenomena of LD rather than sample overlap. More research within this area would be required to gauge the full affects of GWA sample overlap on the result of the PRS. Regardless, there are techniques which correct for sample overlap between GWA studies and they are briefly described below.

One of the first attempts to address overlapping samples in multiple GWA studies was by Lin and Sullivan (2009), who calculated the correlation between all studies based on the number of overlapping samples in each study and adjusted the resulting genetic effect for each SNP in the final meta-analysis of all the studies combined. However, their solution only applies to multiple case-control GWA studies which used the same methodology (to produce the GWA in their analyses (Lin and Sullivan, 2009). A note that an extension to the proposed approach by Lin and Sullivan (2009) was provided by (Han et al., 2016), which assumes heterogeneity between the effect sizes of the genetic effects (aka a random effects model) within each GWA study. Within this approach the studies are first 'decoupled' before performing the meta-analysis. The correlation between all the studies is reduced to zero, and the variances of the genetic effects within each study is increased to adjust to the manipulation of the data (Han et al., 2016).

Most publicly available data sets do not contain genotype information, which limits the number of research groups able to perform these meta-analyses. Performing meta-analysis with GWA summary statistics is a potential solution. Chen et al. (2017) use a similar equation first derived by Lin and Sullivan (Lin and Sullivan, 2009) to produce a continuous test statistic labelled λ_{meta} that provides an estimate of the overlapping samples within two summary level GWA studies. If λ_{meta} is equivalent to one, the samples are presumed to be drawn from a similar population. If λ_{meta} is less than one, the genetic effect sizes are too similar due to sample overlap, and if λ_{meta} is more than one the two cohorts are too dissimilar, potentially due to differing data analysis protocols or explanations within the genetic architectures of the two cohorts. λ_{meta} therefore flags the studies which overlap, and Chen et al. (2017) have derived another metric **P**seudo **P**rofile **S**core **R**egression (PPSR) which generates a genetic similarity matrix for all samples within the two cohorts. However, this approach is limited by the requirement of an 'analysis hub' which is able to produce multiple randomised PRS for each individual (Chen et al., 2017).

If the hypothesis of the research is to test the association between a PRS and a specific trait, then sample overlap between the testing and training data sets can result in the inflation in the magnitude of the association (Choi, Mak, and O'Reilly, 2020).

For example, assume that a single SNP in the population is not associated with a specific phenotype (the correlation (R) between the SNP and the phenotype is zero).

If a sample of the population is selected for a training sample (a GWA study), the expected R^2 value (the variance between the SNP and the phenotype) from a sample size of N is $1/(N-1)$ or $1/N$ if N is large (Wray et al., 2013). This is slightly inaccurate as the value in the population is 0, but the training sample produces a value slightly above 0. For one SNP, this discrepancy is negligible.

However, most PRSs contain thousands, to hundreds of thousands of independent SNPs (predictors) and therefore a set of these m uncorrelated SNPs would result in explaining m/N of the variation between SNP and phenotype (as in a PRS, the SNPs are summed).

For example, when fitted together in a regression analysis in a training sample of N_t = 100, a set of 10 independent SNPs would, on average, explain 10% (R2 = 0.10) of phenotypic variance in the training sample under the null hypothesis of no true association.

If the sample size within the PRS is small but the number of SNPs used is large, the R^2 for the training set may be high by chance and therefore the variation explained by the predictor may be grossly over-estimated.

If samples overlap between the training and the testing set, then based on the equations described previously, the bias within the reported association being proportional to the number of samples that overlap between the training and the testing data set (Wray et al., 2013; Choi, Mak, and O'Reilly, 2020).

Population Genetics

Population structure affecting the association of a singular SNP to a specific phenotype within GWA studies was modelled by Sul et al. (2018). Briefly, the structure of a GWA study assumes that the SNP being tested is independent from all other SNPs in the individual's genome. However, if the SNP is found to be associated with the trait, a proportion of this signal is likely to be caused by a number of SNPs in the individual's genome. This proportion of signal is not modelled within a GWA study. If the population used within the GWA contains a high proportion of related individuals, the proportion of signal contributing to the association between a singular SNP and a phenotype may be substantial and produce false positive results (Sul, Martin, and Eskin, 2018). The logic described above is relevant for any genetic association study, including PRSs.

Population structure can also affect the PRS if the populations within the training set (the GWA study) diverge. PRSs are substantially better at predicting genetic risk in European populations over other populations because the samples used in most GWA studies which form the training set of the PRS are European. The correlation between the phenotype predicted by the phenotype and the observed phenotype declines the more divergent the GWA is from the population the PRS is being applied to (Martin et al., 2019).

There is a suggestion that the fine-scale population structure within sub-populations may affect the predictive accuracy of PRS. In a study by Sakaue et al. (2020), a combination of linear and non-linear dimension reduction methods (**P**rinciple **C**omponent **A**nalysis (PCA) and **U**niform **M**anifold **A**pproximation and **P**rojection (UMAP); (McInnes, Healy, and Melville, 2020)) discovered subtle genetic differences between the sub-populations of Japan. PRSs differed substantially between the mainland population of Japan and the population of the individuals on the islands surrounding the mainland for several traits including height and BMI (Sakaue et al., 2020).

Set-based Analysis within PRS

Since a PRS is the weighted sum of the individual effects of a group of SNPs, it can be defined as a set-based analysis. Therefore, the SNPs used in the set can be controlled to represent, for example, the entire genome (all SNPs within the dataset), a genic PRS (all SNPs are contained within coding regions on the genome), or selected based on their location within functionally relevant biological pathways (SNPs that are located within genes describing a single biological function, for example, from Gene Ontology (Ashburner et al., 2000)). However, for traits that contain a liability dispersed across the genome a SNP set PRS defining a biological pathway will have a lower liability for the trait, then a SNP set PRS encompassing the entire genome (Baker et al., 2018). The power of the PRS describing a biological pathway will be lower than the genome-wide PRS.

Within this chapter, systemic issues involved in PRS analyses will be tackled from a bioinformatics/computational point of view and will be focused in onto four major points:

- Types of PRSs
- Performing and comparing PRS across multiple data sets.
- Generating gene-set PRS
- Interpretation of gene-set PRS

1.6.2 Types of PRSs

PRS can be calculated using several different methodologies and several pieces of software. A common approach is to measure the correlation structure between all variants, and use this structure to assess the best genome-wide prediction. Vilhjalmsson et al. (2015) created LDpred, which uses a Bayesian approach to this method. The PRS is created by initially calculation the posterior mean effects from the training data set. This is achieved by conditioning on a genetic architecture prior and the LD structure from a reference panel (2015). There are two parameters to the genetic architecture prior; the heritability explained by the phenotypes and the proportion of causal variants (2015).

Novel methodologies which extend the Baysian approach are currently in development (Lloyd-Jones et al., 2019).

Alternatively, the correlation structure between variants can be calculated using statistical or regularisation techniques including LASSO regression (Tibshirani, 1996). Mak et al. (2017) created lassosum (`https://github.com/tshmak/lassosum`) an R package that uses penalised regression to adjust the effect sizes of the PRS.

Another common approach is to initially 'clump or prune' the data to account for the correlation structure between SNPs and then to sum all the SNPs meeting a Pt to gauge the proportion of causal SNPs. This methodology is currently implemented within PRSice (Choi, Mak, and O'Reilly, 2020) and PLINK v1.90 (Chang et al., 2015) and will be used within this thesis. A detailed description of this method is provided below in 1.6.3.

1.6.3 PRSs across multiple data sets

Chapter 1.4 explains that a PRS requires two data sets to produce genetic profiles; A target data set with genotype information for each individual, and a training data set which describes the effect of each mutation on the trait of interest. For this study we have used a training set in the form of a large `.txt` file as displayed in Table 1.1.

TABLE 1.1: Raw training set input file describing the IQ3 GWA study[§]

SNP[†]	UNIQUE_ID[†]	CHR*	POS*	A1*	A2	EAF_HRC	Zscore	stdBeta*	SE[†]	P*	N_analyzed	minINFO[†]	EffectDirection[‡]
rs12184267	1:715265	1	715265	t	c	0.04	0.92	6.88e-03	0.01	0.35	225955	0.81	-???????????????++?
rs12184277	1:715367	1	715367	a	g	0.96	-0.66	-4.91e-03	0.01	0.51	226215	0.81	+???????????????–?
rs12184279	1:717485	1	717485	a	c	0.04	1.05	7.91e-03	0.01	0.29	226224	0.81	-???????????????++?
rs116801199	1:720381	1	720381	t	g	0.04	0.30	2.22e-03	0.01	0.76	226626	0.81	-???????????????++?
rs12565286	1:721290	1	721290	c	g	0.04	0.57	4.17e-03	0.01	0.57	226528	0.81	-???????????????++?

[§] Only first 5 rows included. The data set displayed here is identical to the raw data set that can be downloaded from `https://ctg.cncr.nl/software/summary_statistics`. Full description of the data set is also found at the aforementioned hyperlink

*Columns must be present for PRS analysis

[†] Columns useful for PRS analysis

[‡] Columns potentially problematic to read into various software

In this table, from left to right, the columns labelled SNP (**r**eference **s**np (cluster) **ID** (rsID) for each SNP), and UNIQUE_ID (Chromosome and base pair for each SNP) do not need to be present in order to produce a PRS but are supplemental to a PRS analysis. CHR (chromosome number), POS (base-pair position) and A1 (allele from which the effect size is derived) are required in order to produce a PRS. EAF_HRC (effect allele frequency in the Haplotype Reference Consortium reference panel **H**aplotype **R**eference **C**onsortium reference panel (HRC)) and Zscore (metric from the meta-analysis that produced the GWA study) do not need to be included in order to create a PRS.

stdBeta (the standardised beta of A1) needs to be included, while **S**tandard **E**rror (SE) does not need to be included but can be used as supplemental information within a PRS analysis. For example, if within an experiment, several gene-set PRS with varying numbers of SNPs within each set are created and one gene set produces results disparate from the other PRS sets, Knowing the effect sizes and SEs of these individual SNP may provide insight into why the results are disparate. For example, a set might contain a single SNP with an unusually high effect size that explains most of the signal within the PRS SNP set.

P (the p-value of the SNP) is required to produce a PRS.

The minINFO (INFO score of each SNP) columns doe not need to be present in order to produce a PRS but is supplemental to a PRS analysis. INFO scores are created from a program called IMPUTE2 and present a metric for the quality of the imputation for each SNP. Imputation itself allows for the genotyping of variants used for the GWA study that have not been directly genotyped by analysing the LD structure of the SNPs that have been genotyped (Li et al., 2009). This allows for more power within the GWA study without the extra time and monetary cost of genotyping millions of SNPs.

The EffectDirection column may be problematic to read into various software because each row is almost entirely made up of special characters instead of a number or a character. The '?' symbol within the EffectDirection column indicates a missing value.

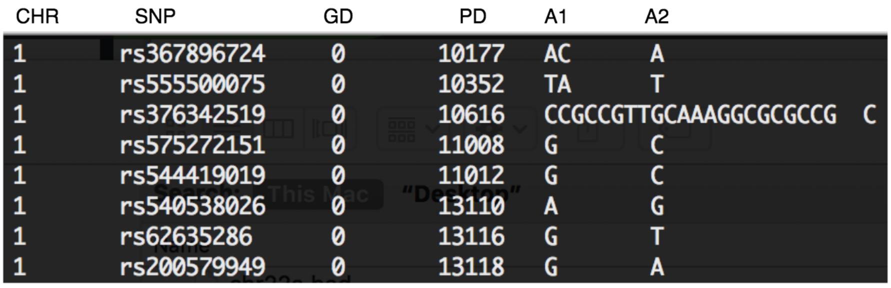
The target data set is usually in either a dosage format or a **B**est **G**uess **G**enotype (BGG) format. Briefly, dosage data is genotype data that has been previously imputed by imputation software (for example IMPUTE2 (Howie, Donnelly, and Marchini, 2009)). As mentioned previously, imputation allows for SNP to be used which have not been directly genotyped. In the output file, this results in a numerical values for each SNP called a dosage. A dosage is the linear transformation of the posterior genotype probabilities. For example, if a SNP is recorded as being ambiguous (A/B within the illumina and/or DBsnp nomenclature (Nelson et al., 2012)), and the genotype probabilities are A/A = 0.1; A/B: 0.4; B/B: 0.5, then the dosage calculation for this particular SNP will be $0 * A/A + 1 * A/B + 2 * B/B = 0.4 + 2 * 0.5 = 1.4$.

While dosage data is robust (as it takes into account all three probabilities per SNP), it may not be suitable for an analysis which relies on a prior assumption of knowing what the most frequent genotype is within a population. In this circumstance, the BGG is calculated, whereby, per SNP, the genotype with the highest probability past a certain threshold is used. However, if the threshold was 0.8, then using the example above, since the highest probability for a genotype is B/B: 0.5, the SNP would be returned as missing. If the combinations were instead: A/A: 0.05; A/B: 0.05, B/B: 0.9, then the genotype B/B would be returned.

Only BGG was used in this thesis and is referenced as the PLINK v1.90 .bed/.bim/.fam format. The 'human readable' format is contained within the .bed and .bim files. These two files explain the data contained within the .bed file. The .bim files contains information about the BGG calculated SNPs located within the samples; information about the samples used to calculated the genotypes is contained within the .fam file. Further information about the files is described in Figure 1.8.

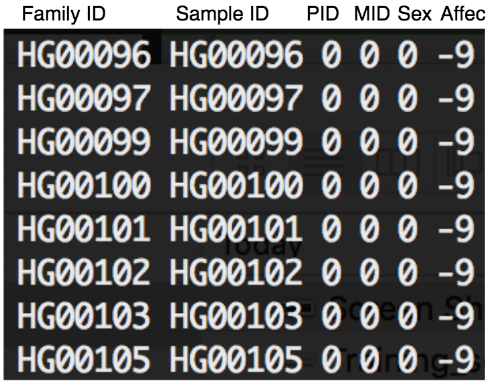
The PLINK v1.90 file format requires a `.bim` file which supplies information about each SNP within the data set. It also required a `.fam` file which provides information on the individuals within the data set. The final file, the `.bed` file, is a binary file which is the only file that is read by the software PLINK and contains all the information provided by the `.fam` and the `.bim` file.

To produce an accurate and reliable PRS, there are seven general analysis steps that must be performed as outlined in Figure 1.9.



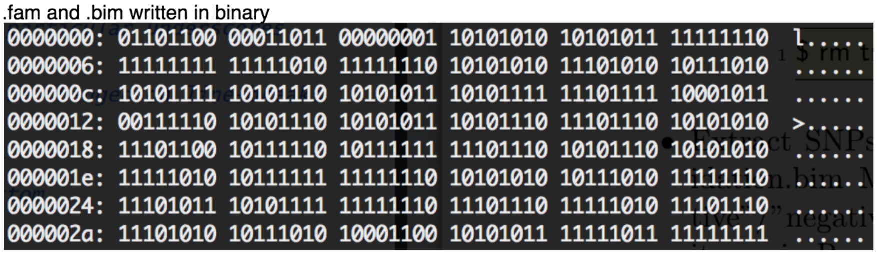
```
CHR        SNP            GD       PD     A1        A2
1       rs367896724       0       10177   AC        A
1       rs555500075       0       10352   TA        T
1       rs376342519       0       10616   CCGCCGTTGCAAAGGCGCGCCG  C
1       rs575272151       0       11008   G         C
1       rs544419019       0       11012   G         C
1       rs540538026       0       13110   A         G
1       rs62635286        0       13116   G         T
1       rs200579949       0       13118   G         A
```

.bim



```
Family ID   Sample ID  PID  MID  Sex  Affection
HG00096 HG00096 0 0 0 -9
HG00097 HG00097 0 0 0 -9
HG00099 HG00099 0 0 0 -9
HG00100 HG00100 0 0 0 -9
HG00101 HG00101 0 0 0 -9
HG00102 HG00102 0 0 0 -9
HG00103 HG00103 0 0 0 -9
HG00105 HG00105 0 0 0 -9
```

.fam



.fam and .bim written in binary

```
0000000: 01101100 00011011 00000001 10101010 10101011 11111110  l.....
0000006: 11111111 11111010 11111110 10101010 11101010 10111010  ......
000000c: 10101111 10101110 10101011 10101111 11101111 10001011  ......
0000012: 00111110 10101110 10101011 10101110 11101110 10101010  >.....
0000018: 11101100 10111110 10111111 11101110 10101110 10101010  ......
000001e: 11111010 10111111 11111110 10101010 10111010 11111110  ......
0000024: 11101011 10101111 11111110 11101110 11111010 11101110  ......
000002a: 11101010 10111010 10001100 10101011 11111011 11111111  ......
```

.bed

FIGURE 1.8: **PLINK v1.90 file format.** PLINK v1.9 (Chang et al., 2015) requires three files as an input; two human readable files indicated here by '`.bim`' and '`.fam`' and a binary file indicated by '`.bed`'. CHR = Chromosome, GD = Genetic distance, PD = Physical position, A1 = Allele 1, A2 = Allele 2. PID = Paternal ID, MID = Maternal ID. Sex can take the values of: 1 = male, 2 = female; other = unknown. Affection can take the values of either 1 = unaffected, 2 = affected; other = unknown. Further information found here: `https://www.cog-genomics.org/plink2/formats`.

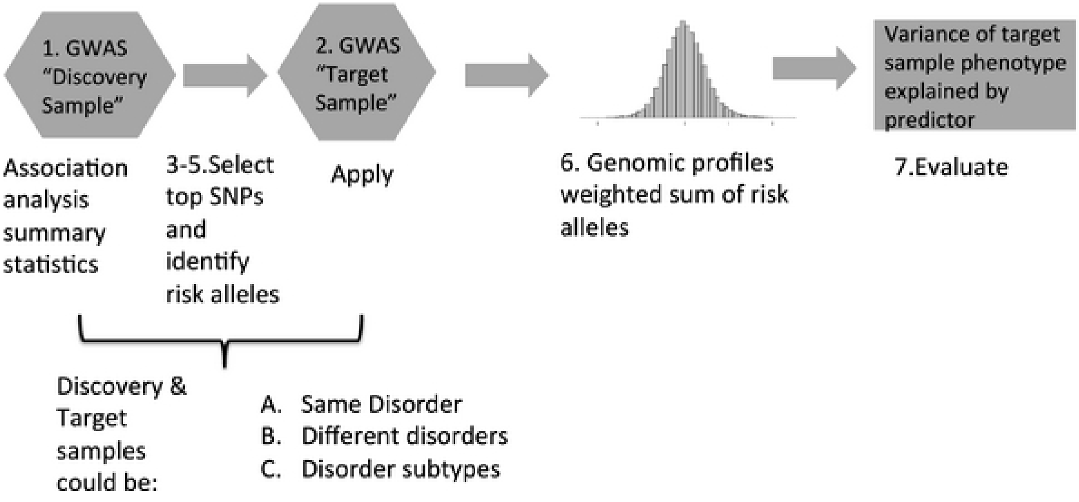

1. GWAS "Discovery Sample"
2. GWAS "Target Sample"
Variance of target sample phenotype explained by predictor
Association analysis summary statistics
3-5.Select top SNPs and identify risk alleles
Apply
6. Genomic profiles weighted sum of risk alleles
7.Evaluate
Discovery & Target samples could be:
A. Same Disorder
B. Different disorders
C. Disorder subtypes

FIGURE 1.9: **PRS analysis steps.** Step 1: Identify a ′discovery′ (or alternatively named ′training′) data set from a GWA study. Step 2: Identify a ′target′ (alternatively named ′testing′) data set. The target data set **must** contain genotype information for each individual. There must not be any sample overlap between training and testing data sets. Step 3: Establish the number of SNPs in common between testing and training data set. Step 4: Apply QC to the training data set. Remove low frequency SNPs, indels (insertion-deletion mutations), low quality variants, and the MHC region. Step 5: Construct a list of SNPs after accounting for LD structure. Step 6: apply a Pt to all SNPs. Step 7: Generate a PRS in the testing data set. Diagram provided by Katherine Tansey.

Steps 1 and 2: Identifying Testing and Training data sets

After identifying the input data sets for PRS analysis, coherence must be achieved between both training set and testing set. Coherence is achieved when, a) the SNP identifiers are identical between both training and testing data set and b) Both data sets have undergone stringent quality control steps.

If a PRS analysis is only performed once, this task is fairly simple if the scripts are manually performed. First, a PLINK v1.90 command is used to remove any un-wanted and/or irregular SNPs. Then a few R commands are used to performed quality control and merge the training and testing sets together on the SNP ID column. Finally, LD is accounted in the combined training and testing set by using another PLINK v1.90 command. By examining Table 1.1 and the ′`.bim`′ file in Figure 1.8 for example, there are multiple columns of data that are almost entirely compatible between both training and target set including ′CHR′ and ′SNP′. Additionally, some columns including ′PD′

and 'POS' contain similar information but have column headings which are slightly different.

Issues arise when standardising multiple different training sets and testing sets. Each PRS analysis should aim to be reproducible. Human error becomes more and more likely to impact the reproducibility of the results as more data sets are processed for each PRS. Examine the differences between Table 1.3 (which describes the raw training data set from PGC1), and Table 1.2 (which is a copy of Table 1.1 displaying the raw IQ3 data set).

TABLE 1.2: Raw training set input file describing the IQ3 GWA study§

SNP†	UNIQUE_ID†	CHR*	POS*	A1*	A2	EAF_HRC	Zscore	stdBeta*	SE†	P*	N_analyzed	minINFO†	EffectDirection‡
rs12184267	1:715265	1	715265	t	c	0.04	0.92	6.88e-03	0.01	0.35	225955	0.81	-???????????????++?
rs12184277	1:715367	1	715367	a	g	0.96	-0.66	-4.91e-03	0.01	0.51	226215	0.81	+???????????????–?
rs12184279	1:717485	1	717485	a	c	0.04	1.05	7.91e-03	0.01	0.29	226224	0.81	-???????????????++?
rs116801199	1:720381	1	720381	t	g	0.04	0.30	2.22e-03	0.01	0.76	226626	0.81	-???????????????++?
rs12565286	1:721290	1	721290	c	g	0.04	0.57	4.17e-03	0.01	0.57	226528	0.81	-???????????????++?

§Only first 5 rows included. The data set displayed here is identical to the raw data set that can be downloaded from `https://ctg.cncr.nl/software/summary_statistics`. Full description of the data set is also found at the aforementioned hyperlink
*Columns must be present for PRS analysis
†Columns useful for PRS analysis
‡Columns potentially problematic to read into various software

TABLE 1.3: Raw training set input file describing the PGC1 GWA study§

snpid†	hg18chr*	bp*	a1*	a2	or*	se†	pval*	info†	ngt	CEUaf
rs3131972	1	742584	AA	G	1.03	0.08	0.76	0.16	0	0.16
rs3131969	1	744045	A	G	1.02	0.08	0.78	0.22	0	0.13
rs3131967	1	744197	T	C	1.02	0.09	0.79	0.21	0	.
rs1048488	1	750775	T	C	0.97	0.08	0.76	0.16	0	0.84
rs12562034	1	758311	A	G	1.00	0.08	0.99	0.19	3	0.09

§Only first 5 rows included. Full description of data set at:
`https://www.med.unc.edu/pgc/results-and-downloads`
*Columns must be present for PRS analysis
†Columns useful for PRS analysis

There are several differences between these two data sets. The heading for the SNP identifier is different between Table 1.2 and Table 1.3, there are two SNP identifier columns in Table 1.2 and one in Table 1.3, almost all columns within Table 1.2 are upper case while almost all columns in 1.3 are lower case, the genome build is included within the heading of stating the chromosomes in Table 1.3 but not included within Table 1.2, and the OR is used as a measurement of effect size in Table 1.3 but the standardised beta coefficient is used within Table 1.2.

If the above process was not automated, a single script per data set would have to be created, because although the processing steps are the same for each data set, the inconsistencies between each training set would need to be accounted for. Every inconsistency would need to be recorded to enable that the script per training set, is reproducible. Additionally, the steps to create a PRS after this standardisation step would have to be repeated for each training data set. These PRS steps are standardised using the PLINK v1.90 software so these analysis steps would be repeated verbatim for each training set.

Automation would remove the requirement to record all the inconsistencies between training sets and remove the requirement to create more than one set of scripts for each PRS. Instead of a script per training set, a program would be used that requires a standardised input for the training set. The inconsistencies between each training set would either be accounted for before the training set is standardised within the program, or if the variations between training sets are limited, can be accounted for within the program itself. The PRS is stringent and reproducible.

There are three options that are available. Either to use pre-existing software that can perform the majority of the PRS analysis, account for all training data sets that will be used within the project, or build a full PRS bioinformatics workflow specific to a variation on a PRS analysis,

Using pre-existing software would save a large amount of time to create all PRSs for the project and does not require a sufficient understanding of computer science. However, the quality of these programs, especially the programs created in academic environments, is an unknown.

In a typical software development project, the quality control of the program is extremely rigorous. In an ideal environment, each software project requires between five to ten individuals (Ahmed, Arshad, and Mahmood, 2018) and contains an infrastructure that requires that the program is separated into simplistic units where each individual unit is tested to ensure it matches several criteria including the redundancy, the unit size, the complexity of the unit and how the unit in question interacts with the other units in the program (Baggen et al., 2011). In an academic environment, the quality of the software aligns with the peer-review process, a piece of software is more likely to be frequently used and trusted by multiple individuals if the piece of software is published. The units of the program are not examined, and the quality is examined via the expected outputs that the software produces. For example, for the PRS software PRSice (Choi and O'Reilly, 2019), less than 5% of the units have been tested and the space between the last two releases was one year and one month `https://github.com/choishingwan/PRSice/releases/tag/2.3.5`.

Using a singular script per data set is acceptable for analyses involving one or two data sets or an analysis that is unlikely to be repeated in the future. However, if multiple data sets are used or if multiple PRSs are required, then a script would have to be created for each PRS. These scripts would need to be accurately documented in order to differentiate between each analyses. Each script would contain analogous code, and the time taken to complete the task would be directly correlated to the number of data sets and the number of PRS required.

A bioinformatics workflow has an initially slower preparation time to produce multiple PRSs, but the analysis is generalised across multiple input data sets and multiple parameters to produce the PRS. The time taken to complete the task is not correlated to the number of data sets or the number of PRS required for the task. The creation of the workflow is not as stringent as software, but because the goal is to speed up analyses rather than generalisation, there is less of an obligation for the time-consuming task of separating the workflow into testable units.

A more complex issue involves the SNP identifiers in both training and target data sets.

Step 3: Determining SNPs in common between Target and Training data sets

For most GWA studies released into the public domain, the identities of each mutation are commonly assigned identifiers known as the rsID from the **S**ingle **N**ucleotide **P**olymorphism **d**ata**b**ase (dbSNP) (Sherry et al., 2001). Within the Bioinformatics community, there is a common mis-conception that these identifiers are unique descriptions for each SNP. Due to the way that genome builds are created (The position of SNPs are defined by a reference genome whereby a selection of individuals are sequenced and the resulting genome fragments are aligned together to create a reference human genome) and the way that SNPs are assigned identifiers, duplications between rsIDs are possible and are unfortunately, fairly common.

Before a SNP is given a rsID, it is a **s**ubmitted **s**np **ID** (ssID) when it is initially submitted to dbSNP. This submission includes the sequence of the SNP and the sequence of the surrounding flanking regions either side of the SNP in question. Through multiple computational algorithms, QC and using **B**asic **L**ocal **A**lignment **S**earch **T**ool (BLAST) (Altschul et al., 1990) or MEGABLAST (Lobo, 2008), the flanking sequence is aligned to the appropriate contig on the appropriate genome build (Camacho et al., 2009; Morgulis et al., 2008). BLAST rapidly aligns and compares an input sequence to a database of known sequences (Lobo, 2008). MEGABLAST performs a similar function but compares the input sequence to sequences with only minor variations and can handle a larger input sequence (Lobo, 2008).

If there are multiple ssIDs at the same location, the identifiers are 'clustered' together and given a rsID. If there are no other ssIDs or rsIDs at that location, the ssID is converted into a rsID (McEntyre and Ostell, 2002).

The SNP FAQ archive (*SNP FAQ Archive [Internet]* 2005) states a number of situations in which there can be the same rsIDs at multiple physical location on the same chromosome. There are three possible reasons for this, the flanking sequence for the ssID was too short when it was submitted, the SNP is in a repetitive region of the chromosome, or there are variations in the SNPs flanking sequence. The problem is not limited to each chromosome either, as a rsID can map to multiple chromosomes if the flanking region of the SNP is in a particularly repetitive region of the genome build.

Depending on the quality and age of the GWA study in question, it is also entirely possible that multiple rsIDs will describe the same physical locations on the genome. In short, SNP discovery took place before the human genome was "built". If multiple rsIDs were accidentally assigned to the same physical location, the dbSNP team merged the identifiers into one rsID in a later genome build (McEntyre and Ostell, 2002; Sherry et al., 2001).

All three data sets referred to in this section (Figure 1.8, Tables 1.1 and 1.3) use rsIDs as the identifiers for each SNP. However, PGC1 uses the University (of) California Santa Cruz (UCSC) hg18 / National Center (for) Biotechnology Information (NCBI) b36 genome build, which means that not all BP positions for each rsID in PGC1 will match the same rsIDs in the other two data sets.

In order to keep to a standardised PRS procedure, I decided that all data sets should be converted to the same genome build before processing. Then, instead of using rsIDs as the unique identifier for each SNP, the chromosome number and the BP position would be combined (eg for Table 1.1, rs12184267 = 1:715265) and any duplicated identifiers would be removed from the analysis.

Step 4: Quality Control

It is standard procedure in most PRS analysis to remove rare SNPs, to remove SNPs of low genotyping quality, to filter variants which deviate from Hardy-Weinberg equilibrium and remove SNPs with a high genotyping missingness. As all of these steps can be achieved with PLINK v1.9 commands, automation of these steps would save considerable time to produce PRS (Chang et al., 2015).

In addition, a more complex but systematic step is the removal of variants with complementary genotypes (for example the variant A:T or the variant G:C). This issue is under-reported. For example, Allele frequencies for a variant within DRD2, a gene linked to schizophrenia, was misrepresented (Sand, 2007). **D**eoxyribo**N**ucleic **A**cid (DNA) is composed of two anti-parallel strands within the human genome. When the DNA is genotyped for each data set, there is ambiguity on which strand was called for each variant. Since most GWA studies do not state the strand in the publicly released data sets, SNPs with alleles 'A:T' or 'C:G' must be removed from both data sets in a PRS as they are complementary bases and it is therefore impossible to tell which strand the variant was called from. To avoid substantial data loss, other alleles (eg A:G or G:A) may be flipped to reach a consensus between the two data sets.

Step 5: Clumping

Clumping removes the genetic signal of LD from the PRS. For multiple genome-wide PRS this is one PLINK v1.9 command repeated for each PRS. Clumping becomes computationally complex for gene-set PRS. See section 1.6.5 for further information.

Step 6: P-value thresholding and Creating genomic profiles

A PRS for any one individual is the summation of their genotypes at a selection of variants, weighted by the effect sizes of the variants on the trait of interest. If the effect sizes are unadjusted, the effect size estimates could be poor with a high standard error. One of the methods to address this issue is to use Pt to filter the selection of SNPs to those with training set P-value below a certain threshold. All excluded SNPs have an effect size estimate set to zero. At a lower Pt, the quality of information for the trait (likelihood the variant is causal) will be substantially higher but the amount of data available will be low. The opposite will be true at a higher Pt.

The best Pt is selected through a process analogous to tuning parameter optimisation, multiple Pt are processed and the Pt which produces the most accurate PRS for the trait of interest is selected. The definition of the 'most accurate' PRS is complex and trait-dependant. Currently, PRSs for many traits are weak proxies for true genetic liability (phenotypic variance explained for many traits is approximately $R^2 < 0.01$). When testing for association, any significant result with a low effect size may be caused by uncorrected confounding effects. In addition, pleiotropy must also be considered when comparing PRS from multiple traits as there is likely to be shared genetic aetiology between the vast majority of phenotypes. For example, a higher predisposition to cognitive performance will, on average, lead to greater educational attainment and higher socio-economic position. A high socio-economic position is associated with the vast majority of diseases, therefore, a genetic component of most diseases will include the genetic aetiology of cognition. With a large enough sample size, associations will be found between the genetics of cognition and the disease (vertical pleiotropy) and between different diseases (horizontal pleiotropy).

In these instances, it may be preferable to obtain and examine a selection of pre-determined Pt as it may inform the strength of the association between the trait(s) in the PRS and inform whether the association was caused by uncorrected confounding genetic effects.

The process of summing genotypes for the creation PRS profiles are simple PLINK v1.9 commands (Chang et al., 2015).

Step 7 (see Figure 1.9) is difficult to automate as in most circumstances, the evaluation of a PRS is performed using logisitic or linear regression which contains covariates and/or a sample reduction stage. Covariates are usually specific for the data set that they were derived from and it may be prudent to select a subset from the available covariates, depending on the scientific hypothesis that is asked. Concordantly, samples may be excluded or included dependent on the scientific hypothesis and/or the inclusion criteria covering an array of what may be tens to hundreds of available phenotypes.

1.6.4 Clumping + thresholding vs other methods

For a whole genome PRS, clumping and Pt selects variants that are only weakly correlated with each other, for use within the PRS. In computational terms, it selects the most significant SNP below the Pt iteratively, computes the correlation between this SNP and all other nearby SNP within a genetic distance of w (also below the Pt), and removes all SNPs that are correlated with the index SNP beyond a pre-selected value r_c^2. The aim of this procedure is to balance between selecting the most predictive SNPs while simultaneously reducing statistical noise. The advantages of this procedure is that it is computationally simple to perform and more computationally efficient past the clumping stage (as less SNPs usually equates to less data to process) (Privé et al., 2019). It also removes SNP that are in LD with each other ad therefore prunes redundant correlated effects from the PRS.

The balance between power and noise is also controlled by the user with the hyper-parameters of w_c, r_c^2 and Pt. While defaults are in place for software including PLINK and PRSice, these parameters will likely change depending on which trait or disease is being analysed. Dependent on the hyper-parameters selected, the clumping procedure may remove independently predictive variants that are correlated with other SNPs (Privé et al., 2019). The power of the PRS is also reduced as the number of SNP is reduced significantly after performing clumping and thresholding.

Other methods including LDpred, lassosum and PRS-CS have been derived which aims to account for the LD while using all the SNPs for the PRS (Vilhjálmsson et al., 2015; Mak et al., 2017; Ge et al., 2019). The simplest method to account for LD is to use linear regression to account for redundant correlated effects, but this results in overfitting when used with a large amount of covariates (aka each SNP) in the model. The data is spread too thinly and results in unstable effect estimates and large standard errors. Another option is to use a baysian approach as used in LDpred, a tool to calculate PRS (Vilhjálmsson et al., 2015). in LDpred the posterior mean effects size (the effect size for each SNP when calculating the PRS) is calculated from the training set by conditioning on a genetic architecture prior and an LD reference panel prior (eg. the 1000 genomes project genotype data) (Vilhjálmsson et al., 2015). The genetic architecture prior is calculated using a heritability estimate from the the training set and the fraction of causal SNPs for the trait (aka the fraction of SNPs with non-zero effect sizes). This allows for better power than clumping and thresholding as all SNP are used. However, LDpred relies on a reference panel for estimation of LD and if the reference panel population structure is too dissimilar from the population structure within the training set, the power of the prediction may decrease (Vilhjálmsson et al., 2015). The same reasoning applies if the population structure is heterogenous across cohorts used within the training set. LDpred also uses a point-normal mixture prior distribution to simulate the genetic architecture prior and this may not be equivalent to a true genetic architecture.

Mak et al (2017) improved on the performance and predictive ability of the PRS using LASSO penalised regression. Penalised regression overcomes the overfitting problem by incorporating a penalty term into the large number of predictors (aka SNPs) (Newcombe et al., 2019). The likelihood of the regression coefficients is modified by the penalty term, with a large penalty leading to the exclusion of many SNPs (Newcombe et al., 2019). The penalty term is optimised depending on the predictive performance in the testing set. Mak et al (2017) disputed the performance of LDpred and claimed that lassosum provided a better predictive power of the PRS, but the main advantage of lassosum is its computational speed. It was faster than LDpred and with 500 participants and 8,000,000 SNPs, lassosum took around 15 min without parallel processing. This is comparable with the computational speed of clumping and thresholding.

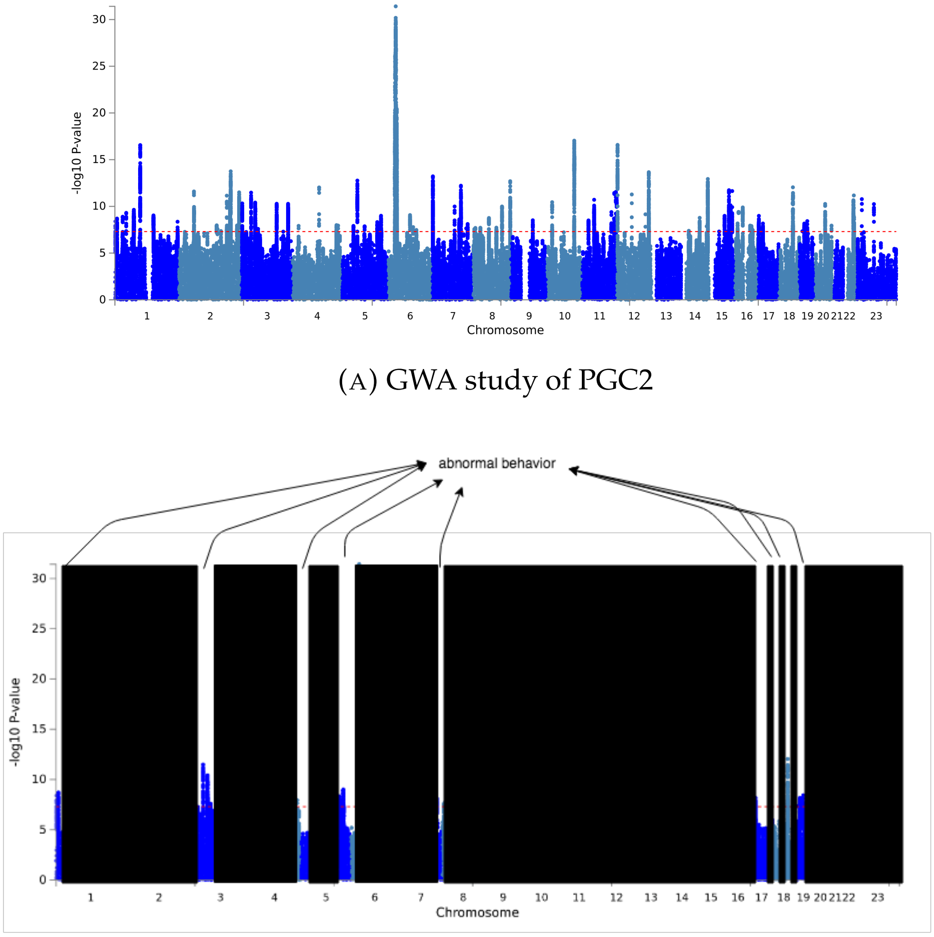
PRS-CS (a bayesian method) attempts to tackle the limitation of heterogeneity between the reference set and the training set and/or heterogeneity within the cohorts of the training set by using a different prior, the SNP effect sizes (Ge et al., 2019). Shrinkage applied to each SNP is adaptive to the strength of its association signal in GWA study (Ge et al., 2019). Despite the improvements listed above, PRS is still limited in its prediction accuracy by the training set sample size, heritability and genetic architecture of the trait being predicted, of which most bayesian PRS methods currently require.

1.6.5 Performing gene-set PRSs

PRSs are a standard methodology to assess the genetic liability to human disorders or phenotypes with a large common genetic risk component. Assessment of liability with PRS can include risk prediction, sample stratification or can be used to dissect the relationships between contrasting subphenotypes (e.g. see Escott-Price et al., 2015; Allardyce et al., 2018; Foley et al., 2017, respectively).

It is of significant interest to incorporate a biological pathway component into the PRS. The aim would be to test whether a set of variants, weighted by the effect of the genetic risk of the trait of interest, is associated at the genome-wide or gene-set specific level (a gene-set PRS). A gene-set PRS would still aim to assess the genetic liability to human disorders or phenotypes, but the advantage is that the gene-set PRS would now confer an extra dimension of describing potentially biological informative features of the trait of interest (Baker et al., 2018). These PRS could then for example, be used to prioritise genes or biological pathways for further functional studies (Baker et al., 2018).

One potential method to create a gene-set PRS is to partition the polygenic risk into SNP sets describing genes or biological pathways which capture a large percentage of the common variation for the trait of interest. A conceptual visualisation of how this is performed in the training set of the PRS is described in Figure 1.10.



(A) GWA study of PGC2

(B) abnormal behavior gene-set PRS

FIGURE 1.10: **Concept of a gene-set PRS.**

Clumping

There are many computational consequences of attempting to perform a gene-set PRS across multiple data sets. When identifying the SNPs in common between testing and training sets, extra steps are required to ensure that accurate and reliable PRSs are produced. The boundaries between gene sets need to be correctly defined and mapped onto the 'consensus' identifiers between the testing and training data sets. Each gene-set must be tested to ensure there is at least two SNPs within the gene-set PRS and supplementary information including the number of SNPs within the gene-set must be logged.

However, most consequences to performing a gene-set PRS over a genome-wide PRS do not occur until pruning or clumping (step 5 in Figure 1.9).

Pruning is an algorithm in PLINK v1.90 that removes the confounding genetic signal due to LD. Each SNP is ordered based on its position in the genome. The first SNP is taken and the correlation between it and the following SNPs is computed (for example the next 50 SNPs). When the correlation coefficient between the first SNP and the other 50 SNPs is observed beyond a set threshold provided by the user, PLINK v1.90 removes one SNP from the correlated pair, keeping the one with the largest MAF. PLINK v1.90 continues on with the next SNP which has not yet been removed.

Clumping is an alternative algorithm in PLINK v1.90 that removes the confounding genetic signal due to LD. The computational procedure of clumping is also performed by the software PLINK v1.90 but the broad procedure is as follows:

1. A 'window' is first defined where variants are included in the clumping procedure if they physically lie within a specified length of the genome.
2. A statistic demonstrating importance (usually a P-value) is used to sort SNPs within the specified window.
3. The first SNP (e.g. the most important/significant SNP) is selected as the index SNP and PLINK v1.90 removes all SNPs in this window that are correlated past a specified threshold with the index SNP.
4. After all relevant SNPs are removed, the next most significant SNP is selected as the index SNP. No index SNP is ever removed from the analysis.

Both methodologies are equally proficient at removing confounding genetic signal. However, in order to obtain a PRS with a higher predictive power, clumping is preferred. In pruning, the SNPs are randomly removed but with clumping, the SNPs with the strongest signal (lowest p-value) are preferentially retained.

In the case of gene-set PRS if the clumping procedure was performed before defining genic boundaries for each gene-set, the accuracy of the gene-set PRS would be affected. Each gene-set only confers the genetic predisposition to each trait with variants located within the physical location of the gene-set and the regulatory regions of the gene-set. If clumping was performed including all variants genome-wide, then the 'window' defined by PLINK v1.90's clumping procedure would include SNPs that are not located with the gene-set. Therefore, some SNPs might be unfairly removed from the analysis if an index SNP lies outside a genic region. The genic regions must be defined before clumping is performed.

Missing SNPs

Figure 1.10 highlights a further computational issue when performing gene-set PRS. Due to the large size of genotype data that now exists (for example CLOZUK (Pardiñas et al., 2018) is 49 **G**iga**B**ytes (GB) in size when compressed), many analyses are split so that each chromosome of input data sets use parallel processing methods (e.g. a supercomputer). Analysis of Genome-wide PRS is simple with this method of processing as it is almost guaranteed that at least one SNP will be on each chromosome. The data set is initially split by chromosome, and each chromosome is processed in parallel. In the case of Figure 1.10b, there are multiple chromosomes where no SNPs are present. Depending on the gene-set, the chromosomes containing no genetic information can change drastically.In addition, if the input gene-set is relatively small (10-20 genes), then when the PRS is limited by the Pt, a low threshold (e.g. p-value of 5e-08) might contain zero SNPs. As of 30/12/2021, there are over 650,000 gene-sets stratified for homo-sapiens within the gene ontology database (see `http://geneontology.org/stats.html`) and six gene-sets significantly associated with schizophrenia. Correcting for missing SNPs for each gene-set individually would be time-consuming. It would be beneficial to optimise this area of gene-set PRS processing.

1.6.6 Interpretation of gene-set PRSs

Genome-wide PRSs are designed to confer information about the identity of individuals with a high risk for a complex trait. Information is however, lost on the individual's genetic profile which may be informative for patient stratification and evaluation of treatment response. A gene-set PRS aims to account for genomic substructure by conferring the risk of a trait within a biological pathway. In order to test whether the biological component of the gene-set PRS has an important effect for the trait in question, the gene-set PRS should confer information about the trait that would not otherwise be visible within a genome-wide PRS. However, as there are up to hundreds of thousands of gene-sets to test against, it would be optimal to compare the gene-set PRS with the genome-wide PRS to ensure that the gene-set PRS confers signal for the trait. A direct comparison of the gene-set PRS and the genome-wide PRS would be unfair as, the intronic regions of the genome are excluded from a gene-set PRS, and due to the significant amount of information conveyed by the genome-wide PRS over the gene-set PRS. A PRS defining only genic regions of the genome (a genic PRS) would be an ideal comparison for any gene-set PRS. This genic PRS would exclude the intronic regions of the genome from the genome-wide PRS.

In addition, the optimal Pt for a gene-set PRS might be entirely different from a genome-wide or a genic PRS. The ideal analysis pipeline for a feasible interpretation of a gene-set PRS would therefore consist of the simultaneous production of a:

- Genome-wide PRS
- Genic PRS
- Gene-set PRS

at a selection of Pt for all PRS.

Ideally, once the PRSs are produced, the input gene sets and Pt would be selected and used to explore how the burden of the genetic risk for the trait varies across the gene-sets.

1.7 Schizophrenia gene-set PRS in Imaging Genetics

There is an increasing number of theories suggesting that abnormalities in early brain development occurring at birth and late development abnormalities around the onset of psychosis, appear in schizophrenia patients (Weinberger, 1987; Kelly et al., 2018). Neuroimaging studies have the potential to examine whether these two models of brain development are causal for schizophrenia.

1.7.1 History of imaging in schizophrenia

Imaging techniques are broadly described into three categories:

- Brain chemistry

- Brain function
- Brain structure

Brain chemistry

Neurotransmitter receptors including 5HT-1/5HT-2, D2/D3, NMDA have been examined using **P**ositron **E**mission **T**omography (PET), **S**ingle **P**hoton **E**mission **C**omputed **T**omography (SPECT), **M**agnetic **R**esonance **S**pectroscopy (MRS) and **P**roton **M**agnetic **R**esonance **S**pectroscopy (PMRS) for their association with schizophrenia (Keshavan et al., 2020). These analyses have produced mixed results. For example, a number of PET studies have found a reduced number of 5HT-1 receptors within the pons and the midbrain, and a reduced number of 5HT-2 receptors in the neocortex (Nikolaus, Müller, and Hautzel, 2016). However, there was no difference in the binding of serotonin transporters of schizophrenia patients over controls. Similarly, for investigations into the glutamate system, there is a suggestion of NMDA receptors in schizophrenia patients from PMRS, but the levels of the neurotransmitter GABA was not significantly different between schizophrenia patients and controls (Schür et al., 2016). Replication was not observed when using PET and SPECT.

The most replicable studies have been on receptors to the neurtransmitter dopamine, where PET has shown direct evidence that D2/D3 receptors are the primary site of action of most antipsychotic drugs (Stone et al., 2009). However, neurotransmitters do not work in isolation and the system of neurotransmitters within the brain will likely change with the progression of schizophrenia.

Brain function

The first neuroimaging studies in schizophrenia used Xenon inhalation, SPECT and PET to measure cerebral blood flow as it is coupled to brain metabolism (Ingvar and Franzén, 1974; O'Connell et al., 1989). Metabolism itself was directly measured using PET. Improvements in neuroscientific methods allowed for the analysis of neural activity and microvascular function in schizophrenia using **B**lood **O**xygenation **L**evel **D**ependent **f**unctional **M**agnetic **R**esonance **I**maging (BOLD fMRI) and **A**rterial **S**pin **L**abeled (ASL) perfusion MRI (Keshavan et al., 2020).

In summary, these techniques have suggested an altered metabolic/hemodynamic activity in the frontal, cingulate, parietal and occipital brain regions of schizophrenia patients, and hyperactivity in the putamen and sensorimotor regions of schizophrenia patients.

Brain structure

There is previous evidence of smaller hippocampal, amygdala, thalamus, nucleus accumbens, and intercranial volumes in schizophrenia patients as compared with controls, as well as larger pallidum and lateral ventricles in schizophrenia patients as compared with controls (Keshavan et al., 2020). There are also reports of widespread cortical thinning and a smaller cortical surface area in schizophrenia patients. There is growing traction that the disconnectivity hypothesis (whereby the disorder involves abnormal or insufficient communication between functional brain regions) first proposed by Friston and Frith (1995), may be the core pathology of schizophrenia. Techniques including DTI have allowed the examination of the white matter microstructure of various cortical and subcortical regions (Keshavan et al., 2020). Computing inter-regional correlations of regional gray matter morphology and complex network analysis computations can aid in the examination of the disconnectivity hypothesis (Wheeler and Voineskos, 2014).

However, the majority of these studies examine patients who have chronic schizophrenia, or patients taking at least one anti-psychotic medication. It would be useful to examine to what extent, genetics is causal to the pathogenesis of schizophrenia in the brain.

1.7.2 ENIGMA consortium

In 2009, A consortium named the Enhancing Neuroimaging Genetics through Meta-Analsis (ENIGMA) was formed with the aim of discovering how the common genetics of humans relate to brain measures derived from Neuroimaging methodology. This goal required large data-sets which could only be achieved through the collaboration of multiple working research groups, up to 50 by the time the first few papers by ENIGMA were published (Stein et al., 2012; Hibar et al., 2015). These papers investigated which common variants affected mean-bilateral hippocampal, total, intercranial brain volumes

(Stein et al., 2012) and subcortical brain volumes (Hibar et al., 2015) through the use of GWA studies and **M**agnetic **R**esonance **I**maging (MRI) scans.

With respect to schizophrenia, the first study was by Van Erp et al (2016) which found that there were differences in subcortical brain volumes between 2028 schizophrenia cases and 2540 controls with smaller hippocampus, amygdala, thalamus, nucleus accumbens and intracranial volumes, and larger pallidum and lateral ventricle volumes. In a follow up study approximately doubling the sample size (4474 cases, 5098 controls), the cortical brain region was analysed and found that schizophrenia patients had a widespread thinner cortex and a smaller surface area of the cortex compared with controls (Erp et al., 2018). ENIGMA developed a **D**iffusion **T**ensor **I**maging (DTI) protocols to analyse microstructure abnormalities, which are undetectable on traditional MRI scans (Thompson et al., 2020; Kochunov et al., 2018). Kelly et al. (2018) investigated the white matter microstructure in schizophrenia cases compared with controls and found that in 20 of the 25 brain regions tested (including but not limited to: anterior limb of internal capsule, corpus callosum, cingulum, fornix and superior corona radiata), the **F**ractional **A**nsiotropy (FA) was significantly lower in schizophrenia patients (1962 in total) as compared to controls (2359 in total). The same was observed for the average FA across all brain regions. Briefly, FA is the measurement of the ansiotropy (property of a material which allows it to change or assume different properties in different directions) of water molecules. Water diffuses freely in all directions in an environment without obstacles. This can be disrupted by cellular bodies, cell membranes and/or macromolecules (including but not limited to axons and dendrites), if diffusion only occurs on the axis of the axon. The degree of ansiotropy, between 0 and 1, can be detected and infer alterations in the axonal diameter, fiber density or myelin structure, aka white matter microstructure.

One of the first ENIGMA studies investigating if common SNPs are associated with brain volumes was by Franke et al (2016), which used a 33,636 cases, 43,008 controls schizophrenia GWA study, but found no association of a PRS to with any brain volume metric for 11,840 subjects. The PRS which captured the most variation was for hippocampal brain volume, but this association was almost entirely driven by a single SNP, rs2268894 (Franke et al., 2016). Following on from this study, Smeland et al. (Smeland and Andreassen, 2018) created a novel approach (condFDR) to test the overlap between schizophrenia and subcortical brain volumes which uses a bayesian **F**alse **D**iscovery

Rate (FDR) approach. the condFDR method leverages overlapping associations in independent GWA study to re-rank test statistics and hypothetically increase the power. However, only a further 5 loci were implicated across the intercranial volume, hippocampus volume and putamin volume. Outside of the subcortex, Lee et al (2016) used GCTA to perform partitioned heritability analysis on 1,750 healthy individuals, and found that SNPs that are associated with schizophrenia explained a significant proportion of heritability in eight brain regions (right temporal pole, left superior frontal, superior temporal, inferior parietal, lateral occipital, and entorhinal cortices, the cuneus and intercranial volume) .

1.7.3 Application of schizophrenia gene-set PRS to subcortical brain imaging Volumes

The first study to examine how the genetics of schizophrenia affects brain morphometry in schizophrenia was a candidate gene study performed on 11 relatives of a pedigree. A SNP located on chromosome 5p14.1–13.1 was found to be associated with ventricular enlargement and frontoparietal atrophy using computed topography (Shihabuddin et al., 1996). In a twin study, it was shown that 4.7% of the genetic variance in schizophrenia was shared with global white matter structure (white matter is found predominantly within subcortical regions; Bohlken et al. (2016)).

Since polygenic risk scores were preferred over candidate gene studies for schizophrenia, no consistent associations between a PRS and any brain region size have been found. Since schizophrenia is complex, and it may be that only certain biological pathways affect certain brain regions, if any consistent association of a gene-set PRS with a subcortical brain region is found, it would progress the field forward.

1.8 Application of Schizophrenia gene-set PRS to Cognitive phenotypes

An alternative to imaging genetics to study brain function and structure is to use cogntitive phenotypes. Numerous patient, family, twin, prospective, and high-risk studies have shown that schizophrenia is associated with deviations in cognition. For example, Lencz et al. (2014) displayed that higher PRSs for schizophrenia are associated with lower general cognitive ability, Ranlund et al. (2018) found that higher PRSs for schizophrenia are associated with poorer spatial visualization skills and Savage et al. (Savage et al., 2018) showed that intelligence has a protective effect on schizophrenia risk.

Promisingly, these cognitive deviations have been observed in unaffected relatives, suggesting a neurobiological risk over environmental effects.

However, while associations between the various cognitive phenotypes is strong, these studies do not address the direction of causation between schizophrenia liability and cognitive phenotypes. By separating out the PRS into biological pathways, the extent to which a schizophrenia PRS is mediated by cognition-related pathways and schizophrenia related pathways can be examined.

2 Datasets used for analyses

The design of a **G**enome **W**ide **A**ssociation (GWA) study is expensive, where a single sample can cost between approximately \$40 to \$200 depending on the genotyping chip that was used (Quick et al., 2020). A full human genome sequence is now approximately \$1000 per sample. Therefore, it is common practice to create a meta-analysis of multiple GWA studies and/or genotype data in order to maximise statistical power. A consequence of this however, is a decrease in distinctive identifiers for data sets and a higher chance of overlapping samples between data sets. To avoid confusion, data sets will be grouped into traits.

Each data set will be in one of two formats; either summary statistics or genotype data. Summary statistics confers information of how each **S**ingle **N**ucleotide **P**olymorphism (SNP) is associated to a defined trait in a population. Genotype data contains an estimate of the genotype at each SNP for any particular individual within the data set.

2.1 Schizophrenia data sets

A total of eight schizophrenia data sets were used within this thesis. Table 2.1 displays concise information about all data sets which is expanded upon in the sections below.

TABLE 2.1: Schizophrenia data sets.

Data set	N† Samples	N† Cases	N† Controls	N† SNPs	Paper	Data set URL
PGC1	20,899	8,832	12,067	849,241	(Ripke et al., 2011)	`https://www.med.unc.edu/pgc/results-and-downloads`
PGC1+Sweden	32,143	13,833	18,310	4,819,154	(Ripke et al., 2013)	`https://www.med.unc.edu/pgc/results-and-downloads`
PGC2*	up to 150,064	up to 36,989	up to 113,075	9,444,231	(Ripke et al., 2014)	`https://www.med.unc.edu/pgc/results-and-downloads`
CardiffCOGS	1,024	1,024	0	9,332,862	(Lynham et al., 2018b)	**N/A (in house)**
CLOZUK	35,302	11,260	24,542	42,561,547	(Pardiñas et al., 2018)	**N/A (in house)**
PGC2noCLOZUK	69,516	29,415	40,101	5,008,739	(Ripke et al., 2014)	**N/A (in house)**
The CLOZUK meta-analysis	105,318	40,675	64,643	8,171,062	(Pardiñas et al., 2018)	`http://walters.psycm.cf.ac.uk/`
SCZminusCOGS	104,294	39,651	64,643	5,550,204	**N/A (unpublished)**	**N/A (in house)**

*Full data set not used in thesis

†N = 'Number of'

The 'Dataset' column refers to the name of the data set, which will be referred to in this thesis hereafter. N Samples refers to the total number of individuals within each data set. The N cases and N controls describes the number of individuals who were cases and controls when the data set contained a binary trait of interest. N SNPs displays the total number of SNPs within each data set. Paper is the reference to the paper in which the data set was derived. The data set URL is the hyperlink to the raw data set.

2.1.1 PGC1 (Summary Statistics)

- Number of Samples = 20,899
    - Cases = 8,832
    - Controls = 12,067
- Total Number of SNPs = 849,241
- Paper URL: `https://www.ncbi.nlm.nih.gov/pubmed/21926974`
- Data set URL: `https://www.med.unc.edu/pgc/results-and-downloads`

PGC1 (Ripke et al., 2011) was amongst the first schizophrenia GWA studies (Purcell et al., 2009; O'Donovan et al., 2008) to define individual schizophrenia loci associated to the trait of schizophrenia. All individuals were of European ancestry and further sample information can be found in the supplementary note of Ripke *et. al.* (2011).

A special note should be made on the fact that this GWA study was originally built using the **U**niversity (of) **C**alifornia **S**anta **C**ruz (UCSC) hg18 / **N**ational **C**enter (for) **B**iotechnology **I**nformation (NCBI) b36 genome build (`https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.12/`. The position of SNPs are defined by a reference genome which had to be created whereby a selection of individuals are sequenced and the resulting genome fragments are aligned together to create a reference human genome. In the case of PGC1, the reference genome used was not the reference genome used for all other data sets(UCSC hg19 / **G**enome **R**eference **C**onsortium (GRC) h37) but can be converted if required.

2.1.2 PGC1+Sweden (Summary Statistics)

- Number of Samples = 32,143
    - Cases = 13,833
    - Controls = 18,310
- Total Number of SNPs = 4,819,154
- Paper URL: `https://www.ncbi.nlm.nih.gov/pubmed/23974872`

- Data set URL: `https://www.med.unc.edu/pgc/results-and-downloads`

PGC1+Sweden (Ripke et al., 2013) expanded on the original PGC1 (Ripke et al., 2011) GWA study by initially producing a GWA study using samples from a Swedish cohort (5,001 cases, 6,243 controls) and then meta-analysed these samples with PGC1. All individuals from PGC1 are within PGC1+Sweden.

2.1.3 PGC2* (Summary Statistics)

- Number of Samples = up to 150,064
  - Cases = 36,989
  - Controls = 113,075
- Total Number of SNPs = 9,444,231
- Paper URL: `https://www.nature.com/articles/nature13595`
- Data set URL: `https://www.med.unc.edu/pgc/results-and-downloads`

PGC2 (Ripke et al., 2014) is a large meta-analysis of 49 ancestry matched case-control samples (34,241 cases and 45,604 controls; 46 European and 3 east Asian ancestry), 3 family-based samples from Europe (1,235 parent affected-offspring trios) and the deCODE cohort (1,513 cases and 66,236 controls; European ancestry). While the summary statistics are available for the full data set, it would not be useful in **P**olygenic **R**isk **S**core (PRS) analysis for this thesis as many samples contained within this data set would overlap with many different genotype data sets required for PRS analysis. Additionally, samples would need to be reduced down due to population stratification and relatedness between individuals in the GWA study (Pardiñas et al., 2018). Population stratification is where non-random mating occurs (and usually due to geographic isolation). If different populations are used within a GWA study, associations between a SNP and schizophrenia may be due to the genetic differences between populations and be unrelated to schizophrenia. All samples from PGC1+Sweden are contained in this GWA study.

2.1.4 CardiffCOGS (Genotype data)

- Number of Samples = 1,024
  - Cases = 1,024
  - Controls = 0
- Total Number of SNPs = 9,332,862
- Paper URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6019354/`
- Data set URL: **N/A (in-house)**

The CardiffCOGS cohort has been increasing in size consistently over a period of approximately 10 years. Its first reported use (N = 571) was in a study by Rees *et al.* (2014a) to examine the contribution of 15 **C**opy **N**umber **V**ariants (CNVs) to schizophrenia-associated loci. All patients from Rees *et al.* (2014a) CardiffCOGS were recruited from community, in-patient and voluntary sector mental health services in the UK. For each individual, A **D**iagnostic (and) **S**tatistical **M**anual (of Mental Disorders) (DSM)-IV criterium based, best-estimate lifetime diagnosis was arrived at using **S**chedules (for) **C**linical **A**ssessment (in) **N**europsychiatry (SCAN) instrument interviews and a review of case-notes on each individual. These DSM-IV criterium based diagnoses and recruitment procedures continued as CardiffCOGS increased in sample size (Rehman, 2011; American Psychiatric Association, 2000).

The next reported use of CardiffCOGS was in Pardiñas *et al.* which separated CardiffCOGS into two waves, CardiffCOGS1 (N = 512) and CardiffCOGS2 (N = 247). The main use of these individuals was to test for validation of treatment-resistant schizophrenia (symptoms persist despite two or more trials of antipsychotic medications of adequate dose and duration (Potkin et al., 2020)) in a much larger treatment-resistant schizophrenia sample, CLOZUK (Pardiñas et al., 2018). CardiffCOGS1 was genotyped by the Broad institute (Massachusetts, USA) using Illumina HumanOmniExpress-12 and OmniExpressExome-8 chips while CardiffCOGS2 genotyping was performed by deCODE in Iceland using Illumina HumanOmniExpress-12 chips (Pardiñas et al., 2018; Rees et al., 2014a).

Lynham *et al.* (2018c) was the latest to use CardiffCOGS in a study which compared the phenotypic differences of cognition across the schizophrenia/bipolar diagnostic spectrum. In this cohort, the data set is referred to as the

Cognition (in) **M**ood, **P**sychosis (and) **S**chizophrenia **S**tudy (CoMPaSS) but includes all previous CardiffCOGS individuals. In total, 824 cases were used within this study including a DSM-IV diagnosis of schizophrenia (n = 558), schizoaffective depressive (n = 112), schizoaffective bipolar (n = 76) or bipolar disorder (n = 78), as well as 103 control participants.

To ascertain a cognitive phenotype within the neuropsychiatric case group, Lynham *et al.* (2018c) used the MATRICS **C**ognitive **C**onsensus **B**attery (MCCB) (Nuechterlein et al., 2008a). In essence, the MCCB measures seven domains of cognition including, but not limited to, speed of processing, verbal learning and social cognition by performing 10 tasks (Lynham et al., 2018c). At each task, the mean and standard deviation of the control group (N = 103) was used to derive a z score for each individual within the neuropsychiatric case group. A composite cognitive score of all tasks was created for each individual in accordance with MCCB procedures.

For simplicity, the data set used in this thesis will be referred to as 'CardiffCOGS'. While 1024 individuals is stated as the size of CardiffCOGS, this figure encompasses the entire cohort to date, without restrictions to any one measurement (e.g. to DSM-IV diagnosis). When restrictions are applied, they will be explicitly mentioned.

2.1.5 CLOZUK (Genotype data)

- Number of Samples = 35,302
  - Cases = 11,260
  - Controls = 24,542
- Total Number of SNPs = 42,561,547
- Paper URL: `http://www.ncbi.nlm.nih.gov/pubmed/29483656`
- Data set URL: **N/A (in-house)**

In the UK, there is a compulsory clozapine blood monitoring system for any individual prescribed the antipsychotic medication clozapine. The medication is licensed for treatment resistant schizophrenia. Pardiñas *et al.* (2018) acquired anonymous aliquots of blood samples regularly collected as part of routine checks for agranulocytosis, a rare adverse effect of taking clozapine. The ascertainment of samples was in line with the UK Human tissue act and followed national research ethics approval.

These samples (defined as CLOZUK1 and CLOZUK2) were then genotyped as described in Pardiñas *et al.* (2018) and made-up a substantial portion of the cases within the CLOZUK data set Table 2.2

TABLE 2.2: CLOZUK Genotype Samples*.

Dataset	Samples in GWAS	Genotyping chip
CLOZUK1*	5,528	OmniExpress
CardiffCOGS1*	512	OmniExpress
CLOZUK2*	4,973	OmniExpress
CardiffCOGS2*	247	OmniExpress
WTCCC2†	4,641	Illumina 1.2M
Cardiff Controls†	1,078	OmniExpress
Generation Scotland†	6,480	OmniExpress
T1DGC†	2,532	HumanHap 550
POBI†	2,516	Illumina 1.2M
TWINSUK†	2,426	Illumina 317/610/660/1M
QIMR†	2,339	Illumina 317/610/660
TEDS†	1,752	OmniExpress
GERAD†	778	Illumina 660

*table adapted from Pardiñas *et al.* (2018). * Schizophrenia cases.
† Control samples

The 'Dataset' column indicates the identifier of a subset of samples within the CLOZUK data set. The 'Samples in GWAS' column indicates the number of samples within that CLOZUK subset. The Genotyping chip column indicates on each array of how each of the CLOZUK sub-samples were sequenced.

CardiffCOGS samples (as described previously) contained a portion of treatment-resistant schizophrenia patients that were used to validate a treatment-resistant schizophrenia diagnosis within the CLOZUK1 and CLOZUK2 samples. Altogether, CardiffCOGS1+2 and CLOZUK 1+2 encompass the cases of CLOZUK.

Control samples were collected from publicly available sources or via collaboration. A note must be made that for the cohorts CLOZUK1, CardiffCOGS1, WTCCC2 and Cardiff Controls, there were a total of 6,040 cases and 5,719 controls overlapping with the data set referred to as PGC2 in this thesis.

2.1.6 PGC2noCLOZUK (Summary Statistics)

- Number of Samples = 69,516
  - Cases = 29,415
  - Controls = 40,101
- Total Number of SNPs = 5,008,739
- Paper URL: `https://www.nature.com/articles/nature13595`
- Data set URL: **N/A (in-house)**

The overlapping samples referred to in CLOZUK were removed from PGC2 to create a new PGC2 summary statistics data set independent from the CLOZUK samples. Duplicate samples were identified by identical ID's and were removed using PLINK.

2.1.7 The CLOZUK meta-analysis (Summary Statistics)

- Number of Samples = 105,318
  - Cases = 40,675
  - Controls = 64,643
- Total Number of SNPs = 8,171,062
- Paper URL: `https://www.ncbi.nlm.nih.gov/pubmed/29483656`
- Data set URL: `http://walters.psycm.cf.ac.uk/`

As CLOZUK and PGC2noCLOZUK were entirely independent from one another, Pardiñas *et al.* (2018) performed a meta-analysis on both data sets to create a new GWA study. METAL was used to perform the meta-analysis using their fixed effect procedure, with weighting derived from standard errors (Willer, Li, and Abecasis, 2010).

2.1.8 SCZminusCOGS (Summary Statistics)

- Number of Samples = 104,294
  - Cases = 39,651
  - Controls = 64,643

- Total Number of SNPs = 5,550,204
- Paper URL: **N/A (unpublished)**
- Data set URL: **N/A (in-house)**

In order to create schizophrenia PRSs in CardiffCOGS with the largest sample size available, the CardiffCOGS samples were removed from the CLOZUK meta-analysis. METAL (Willer, Li, and Abecasis, 2010) was used to meta-analyse two in-house data sets to create an equivalent of a SCZminusCOGS data set. The fixed-effects model was used and all SNPs were limited to an INFO > 0.9 (Willer, Li, and Abecasis, 2010).

2.2 General Intelligence data set

The complex traits genetics lab in Amsterdam have been releasing GWA studies after stages of steadily increasing their sample size for the trait of general intelligence in the population or **I**ntelligence **Q**uotient (IQ), as observed with the **P**sychiatric **G**enomics **C**onsortium (PGC) data sets. The data set that will be analysed here will be the latest IQ GWA study: IQ3 from Savage *et. al* (2018) which explains 4% of the variation in IQ observed within samples.

TABLE 2.3: General Intelligence data sets.

	IQ3
Reference	Savage, J. E. et al. 2018. Nature genetics
Sample Size	269,867
Number of associated SNPs	242 lead SNPs (205 loci)
Gene-sets identified	6*
Paper URL	`https://www.nature.com/articles/s41588-018-0152-6`
Data set URL	`https://ctg.cncr.nl/software/summary_statistics`

*(3 conditional MAGMA):,neurogenesis, central nervous system neuron differentiation, and regulation of synapse structure or activity processes)

The first row contains the identifiers for each public general intelligence data set which will be referred to these identifiers for the rest of the thesis. Information pertaining to the reference where each of the data sets is in the 'Reference' row. The 'Sample size' row indicates the total number of samples within each data set. The number of associated SNPs row contains the total number of significantly associated loci to the trait of general intelligence within IQ3. The Paper and Data set URL gives a hyperlink to the paper referencing the data set and the location of the raw data set.

2.2.1 IQ3 (Summary Statistics)

- Number of Samples = 269,867
  - Cases = **N/A**
  - Controls = **N/A**
- Total Number of SNPs = 9,295,119
- Number of associated SNPs = 242 lead SNPs (205 loci)
- Gene-sets identified: 6 (3 conditional MAGMA:,neurogenesis, central nervous system neuron differentiation, and regulation of synapse structure or activity processes)
- Paper URL: `https://www.nature.com/articles/s41588-018-0152-6`
- Data set URL: `https://ctg.cncr.nl/software/summary_statistics`

*The number of associated SNPs row contains the total number of significantly associated loci to the trait of general intelligence within IQ3.

Large sample sizes are required to obtain effective signals for the IQ. In IQ3, the intelligence measurements of 14 independent cohorts were meta-analysed together (Savage et al., 2018). The assessment of IQ across all 14 cohorts was calculated using neurocognitive tests calculating the fluid domains of cognitive functioning (reasoning/thinking, processing speeds, problem solving in novel situations). Despite differences between cognitive tests performed on the different cohorts, a latent factor described as *The General factor of intelligence (g)* was used to capture the variance in common across the cognitive tasks. The association of each SNP to *g* was used for the IQ3 GWA study. The effect sizes for each SNP were standardised based on the method described in Zhu *et al.* (2016).

2.3 Autism Spectrum Disorder data set

2.3.1 iPsych ASD (Summary Statistics)

- Number of Samples = 46,350
  - Cases = **18,381**
  - Controls = **27,969**
- Total Number of SNPs = 9,112,387
- Paper URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6454898/`

- Data set URL: `https://www.med.unc.edu/pgc/results-and-downloads/`

The iPsych GWA study found five loci associated to Autism Spectrum disorder trait (Grove et al., 2019). In this thesis, the use of the iPSYCH **A**utism **S**pectrum **D**isorder (ASD) dataset was confined to Chapter 3 whereby the objective was to determine the two bioinformatics tools to produce gene-set PRS.

2.4 Early Growth Genetics dataset

2.4.1 Birth length (Summary Statistics)

- Number of Samples = 28 459
  - Cases = **N/A**
  - Controls = **NA**
- Total Number of SNPs = 2,201,971
- Paper URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4447786/`
- Data set URL: `http://egg-consortium.org/birth-length.html`

The Body length GWA was created to investigate the relationship between fetal and infancy length growth and various complex diseases including cardiovascular disease and type 2 diabetes. Seven independent SNPs were found to be associated with birth length as measured using standardised procedures (Valk et al., 2015).

2.5 General population data sets

The creation of trait specific data sets is one method to tackle the research of life-threatening and disabling conditions in humans. An alternative yet complementary approach is to create a sizeable cohort with a compendium of phenotypes. As the phenotype is not limited to any singular trait, the sample size can be substantially higher than the trait specific data sets, allowing for higher power for the PRS analyses. In addition, the data sets are commonly used within the neuropsychiatric field which means that they are ideal when testing the efficacy of gene-set PRS.

2.5.1 European 1000 Genomes (Genotype data)

- Number of Samples = 503

  - Cases = **N/A**
  - Controls = **N/A**

- Total Number of SNPs = 11,915,643
- Paper URL: `https://www.nature.com/articles/nature15393`
- Data set URL: `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`

In phase 3 of the 1000 genomes project, various sequencing data are collected together from 26 populations from around the world (Auton and Abecasis, 2015). For the purposes of this thesis, the phase 3 raw data are converted into the plink file format PLINK REFERENCE and limited to the 503 individuals with a European descent.

2.5.2 UK Biobank (Genotype data)

- Number of Samples = 443,031 (white UK and Irish)
  - Cases = **N/A**
  - Controls = **N/A**
- Total Number of SNPs = 35,884,914
- Paper URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380465/`
- Data set URL: `https://biobank.ctsu.ox.ac.uk/showcase/`

The UK Biobank genotype data contains both genetic and phenotype information for individuals from the UK (Sudlow et al., 2015). Phenotypes recorded include schizophrenia (520 individuals as diagnosed by the International Statistical Classification of Diseases and Related Health Problems or the ICD-10 and death records) and demographic data (**B**ody **M**ass **I**ndex (BMI)). 7,654,308 SNPs remain after standardised **Q**uality **C**ontrol (QC) steps.

2.5.3 Summary

In this Chapter, I have outlined and described all data-sets that were used within this thesis. In total there are eight data sets that are related to schizophrenia, one data set that is related to general intelligence within the population, one data set related to autism and two data sets that are population samples.

3 SurPRSe workflow and PRSAVE shiny app

3.1 Introduction

In order for a **P**olygenic **R**isk **S**core (PRS) to convey the liability of a biological pathway, only the **S**ingle **N**ucleotide **P**olymorphisms (SNPs) within that biological pathway must be used for the respective PRS. There is no standardised software for producing a Gene set PRS. The aim of this chapter is to describe the bioinformatics workflow I created to perform gene-set PRS accurately, and to investigate existing software which attempts to produce gene-set PRS. My bioinformatics workflow was named **Su**pe**r**computing (with) **P**olygenic **R**isk **S**core **e**valuation (SurPRSe) and is run on Cardiff's supercomputer, Hawk. PRSet is the component of PRSice2 (Choi, Mak, and O'Reilly, 2020) which produces gene-set PRS.

As stated in Chapter 1.6, the production and analysis steps for a PRS is quite straightforward, but computational issues arise when attempting to make a gene-set PRS across multiple input data sets and gene-sets. As stated previously, there are three options to produce gene-set PRS, either to use pre-existing software, to use a set of scripts per each **G**enome **W**ide **A**ssociation (GWA) study or to create a bioinformatics workflow. At the time of writing, PRSet is the only pre-existing software that is able to produce gene-set PRS (Choi and O'Reilly, 2019). However, at the time at which these PRS were being produced, the first version of PRSet was only recently made available to the public. The program is written in C++, a programming language of which the only available support was from the developer. As there was little support available for PRSet, this option was discounted.

11 data sets were used within this thesis. Creating a set of scripts per data set would be too time consuming and the lack of structure may affect the accuracy of the PRS, especially when the PRS were compared to one another. Therefore, the creation of SurPRSe was the most favourable option. The bioinformatics workflow was written in R and BASH, two languages which are more accessible than C++. This allowed more opportunity for support and a better understanding from other academics of how SurPRSe works. SurPRSe will be compared to PRSet as an initial measurement of

quality. The quality and the application of the gene-set PRSs will be further assessed using two case studies within the trait of schizophrenia. Both of these case studies will involve both the comparison of multiple gene-set PRS to each other and the comparison to the current standardised method of PRS, which is to use all SNPs available within the sample.

SurPRSe will create the PRSs, while association testing will be visualised in a shiny app named PRSAVE. As shiny apps are interactive, it will enable a quick and easy visualisation of the PRS without the difficulty of parsing through a large output of summary statistics.

In this chapter, my SurPRSe workflow will aim to produce gene-set PRS that targets the 1000 genomes population sample and trains on the iPsych ASD GWAS. Gene-sets will be taken directly from the Gene-ontology resource available on MsigDB (Subramanian et al., 2005; Liberzon et al., 2011). SurPRSe will be evaluated on:

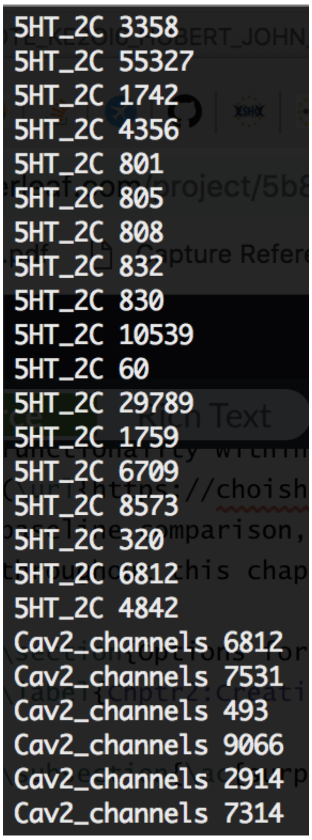
- Analysis set-up
- Speed of processing
- Production of PRS

3.2 Workflow Development - SurPRSe

I created a bioinformatics workflow named SurPRSe aimed at solving the issues in section 1.6. Specifically, SurPRSe will enable the simultaneous production of genome-wide, gene-wide and gene-set PRS profiles in one collated output file.

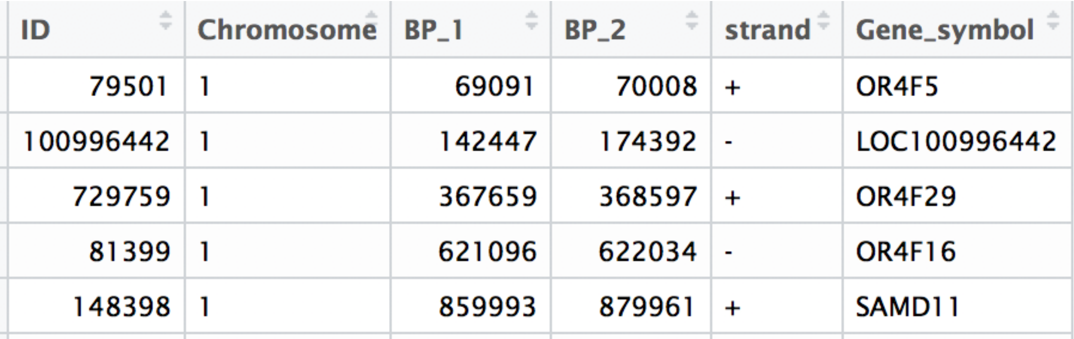
3.2.1 File inputs

The input to SurPRSe is the training data set, the testing data set, a gene set annotation file as described in Figure 3.1 and a gene location file in an **N**ational **C**enter (for) **B**iotechnology **I**nformation (NCBI) format as displayed in Figure 3.2. The gene set location file was downloaded from: `https://ctg.cncr.nl/software/MAGMA/aux_files/NCBI37.3.zip`. The gene set annotation file was curated depending on the trait used. For schizophrenia, the gene-sets were provided in-house in the same format used for Pardinas et al. (2018). All identifiers were entrez-gene identifiers and no identifiers were removed or added from their source files. The file format was converted to the format described in Figure 3.1. The output to the workflow are PRS profiles of gene-set PRSs, a gene-wide PRS and a genome-wide PRS.



```
5HT_2C 3358
5HT_2C 55327
5HT_2C 1742
5HT_2C 4356
5HT_2C 801
5HT_2C 805
5HT_2C 808
5HT_2C 832
5HT_2C 830
5HT_2C 10539
5HT_2C 60
5HT_2C 29789
5HT_2C 1759
5HT_2C 6709
5HT_2C 8573
5HT_2C 320
5HT_2C 6812
5HT_2C 4842
Cav2_channels 6812
Cav2_channels 7531
Cav2_channels 493
Cav2_channels 9066
Cav2_channels 2914
Cav2_channels 7314
```

FIGURE 3.1: **SurPRSe gene-set annotation file input**. The first row is the name of each gene-set repeated to match the number of genes within each gene-set. The second column contain the Entrez gene (NCBI) identifiers. For example, 5HT-2C contains 18 genes where the first gene in the list is the entrez gene ID 3358. This gene is 5-hydroxytryptamine receptor 2C which encodes a seven-transmembrane G-protein-coupled receptor and responds to the neurotransmitter serotonin.



ID	Chromosome	BP_1	BP_2	strand	Gene_symbol
79501	1	69091	70008	+	OR4F5
100996442	1	142447	174392	-	LOC100996442
729759	1	367659	368597	+	OR4F29
81399	1	621096	622034	-	OR4F16
148398	1	859993	879961	+	SAMD11

FIGURE 3.2: **NCBI gene location file input**. ID = NCBI gene identifier, Chromosome = chromosomal location of gene, BP_1 = the start of the coding region of the gene, BP_2 = the end of the coding region of the gene, strand = whether the gene is located on the positive or the negative strand, Gene_symbol = the Gene name identifier.

3.2.2 File outputs

The output of SurPRSe is a table within a text file where each column represents a PRS and each row is an individual. The column names act as the identifier for each PRS (see Table 3.1).

TABLE 3.1: Output file from SurPRSe.

FID	IID	PHENO	extended_geneset_SCORE_GO_CARDIAC-_CHAMBER_DEVELOPMENT_1e-04	extended_geneset_SCORE_GO_CARDIAC-_CHAMBER_DEVELOPMENT_0.01
HG00096	HG00096	-9	0	2.98e-03
HG00097	HG00097	-9	-0.03	1.48e-03
HG00099	HG00099	-9	0	2.51e-03
HG00100	HG00100	-9	0	1.00e-04
HG00101	HG00101	-9	-0.03	-4.02e-03

FID = Family ID, IID = Individual ID, PHENO = Phenotype of individual, GO = Gene Ontology.

The first two columns (FID and IID) represent the unique identifiers for each individual. The PHENO column represents the phenotype of the individual but this column usually contains the value ’-9’ which indicates that the data are missing. biobanking has enabled the release of thousands of phenotypes for each individual’s genotype. This vast amount of information usually requires the phenotypes to be stored in a different file.

Each PRS profile is stored as a column name which promotes text-mining procedures. For example in the column: ’extended_geneset_SCORE_GO_CARDIAC_CHAMBER_DEVELOPMENT_1e-04’ from Table 3.1, each group of words between the underscores can be grouped as: "gene region length"_"common ID"_"Gene-set name"_"significance threshold used". The 'gene region length' describes whether the flanking regions outside the gene (e.g. the promotor) were included (extended) or were not included (normal). The 'common ID' was used to identify this type of column name.

3.2.3 Architecture

In essense, SurPRSe follows the flow diagram depicted in Figure 3.3. At present, SurPRSe is only able to work within a server or supercomputer environment. The locations of both training set and testing set on the server are used as an input. The data are copied across to the working environment (the locations on the server where all input files and output files are).

The training data set is read into SurPRSe. Assuming that the training data set is a suitable standard and passes quality control measures, SurPRSe loads in the testing data set. If the testing data set passes **Q**uality **C**ontrol (QC) measures (see 1.6.3) and if both training data set and testing data set are compatible with each other for PRS analysis (see 1.6.3, SurPRSe is able to run a whole genome, gene-centric and a gene-set PRS automatically.

The output is a table of the various PRS profiles which can be downloaded off of the server when SurPRSe has finished processing. If the user wishes to select gene-sets associated with the input training GWA study, SurPRSe has the functionality to perform gene-set analysis on the training set GWA study using MAGMA (Leeuw et al., 2015).

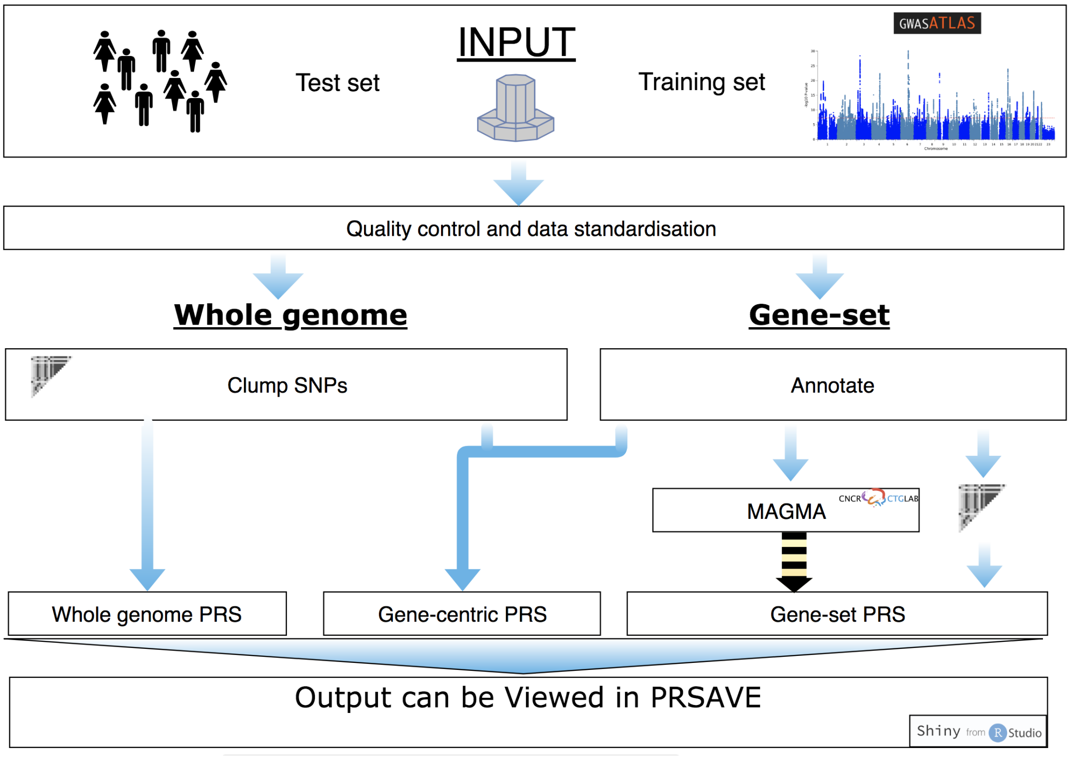


FIGURE 3.3: **SurPRSe Flow Diagram**. A broad overview of the functionality of SurPRSe. Each blue arrow indicates the direction of processing throughout SurPRSe. The yellow and black striped arrow indicates the process can only be performed once MAGMA gene-set analysis takes place. The gene-sets that are output by MAGMA are re-input into SurPRSe.The image of the training set GWA study was taken from GWASATLAS (Tian et al., 2020).

Logistically, SurPRSe is able to produce PRS via a workflow of **B**ourne **A**gain **SH**ell (BASH) and R scripts. The function of each script is provided below.

Module 1: Preparation and configuration

Scripts used:

- run.sh
- config.sh
- run_config.sh
- plink_config.sh

This module defines the parameters for running the SurPRSe workflow.

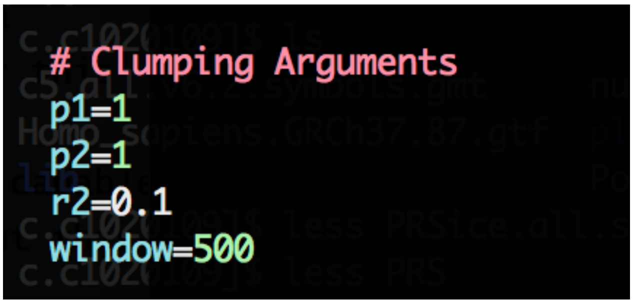
run.sh

`run.sh` controls the desired output of SurPRSe. If the user only wants to produce a genome-wide PRS for example, `run.sh` will activate the relevant scripts required for this to happen. This script is also able to trigger a MAGMA gene-set analysis on the PRS training data set if required but this should be performed independently of any analysis that produces a PRS.

config.sh and run_config.sh

`run.sh` should never be altered by the user and is in fact controlled by `config.sh` and `run_config.sh`, which allows the user to define the parameters that SurPRSe is capable of. These two scipts also ensures the analysis is performed in the correct directories on the Hawk **H**igh **P**erformance **C**omputing (HPC) cluster.

`config.sh` is the only file that is edited by the user. Each variable name should not be changed as it is stored and sourced throughout SurPRSe. The values of each variable can be edited by the user. For example for Figure 3.4:



```
# Clumping Arguments
p1=1
p2=1
r2=0.1
window=500
```

FIGURE 3.4: **SurPRSe arguments for clumping**. Taken from a subsection of `PRS_arguments_script.sh`.

`p1` controls the significance threshold for *index* SNPs. The index SNP is the SNP that represents all other SNPs in its haplotype block. `p2` controls the significance threshold for *clumped* SNPs. `r2` controls the **L**inkage **D**isequilibrium (LD) threshold when applying the clumping procedure and the `window` variable controls the physical distance threshold for clumping in SurPRSe. The `window` variable is measured in **K**ilo-**B**ase (KB). Further information on clumping can be found in Chapter 1.4.

plink_config.sh

This script contains three R scripts which create `.txt` files of some of the arguments specified in `PRS_arguments_script.sh`. Two of the R scripts create a lower threshold and an upper threshold file in a format that PLINK recognises if using their software to define the **P** value **t**hreshold (Pt) of the PRS analysis. For example, if the Pt was 0.05 and the lower bound was 0, PLINK includes all SNPs with P-value from 0 to 0.05, including any SNP with P-value equal to 0.05. The last R script performs a similar action but for the chromosomes which are considered for the analysis.

Module 2: Quality Control

This module handles the pre-processing and QC of the training data set and the testing data set. This includes clumping, INFO score and MAF thresholds for all SNPs and data harmonisation between training set and testing set.

Scripts used:

- training_set_QC.sh
- training_set_QC.R
- testing_set_QC_and_clumping.sh
- MAF_INFO_score_QC.R
- testing_training_harmonisation.R

training_set_QC.sh and training_set_QC.R
`training_set_QC.sh` contains the R script: `training_set_QC.R`, which is a QC check of the training set and also contains commands to split the genotype data into different files dependant on which chromosome the SNPs are located within the data set if required.

testing_set_QC_and_clumping.sh
This script predominantly handles the processing before, during and directly after the clumping procedure (see Chapter 1.4). The training data set and the testing data set are harmonised so that only SNPs within both data sets are considered for the PRS. Clumping then may be performed dependant on whether the user wishes to produce a gene-set PRS. All commands for clumping are handled by PLINK 1.9 (Chang et al., 2015). Further clean-up steps are then performed to remove auxiliary files that are created as a part of the clumping procedure.

MAF_INFO_score_QC.R
Called from within `testing_set_QC_and_clumping.sh`. The training data set is split by the chromosome number and QC'd based on **M**inor **A**llele **F**requency (MAF) and the INFO score. The MAF is the frequency at which the second most common allele

occurs across the individuals within the training set and the INFO score. The INFO score quantifies the likelihood that the SNP was imputed correctly and takes a value between 0 to 1 (where the closer the value is to 1, the higher certainty of imputation).

testing_training_harmonisation.R
Called from within `testing_set_QC_and_clumping.sh`. `testing_training_harmonisation.R` checks for duplicated SNP identifiers and harmonises the SNP identifiers from the training set with the SNP identifiers from the testing set. Each SNP is tested for mis-information due to allele flipping (see Chapter 1.4) and checks are made to see if the previous steps were processed correctly.

Module 3: Create Gene-set PRS

Gene-set PRS are created within this script. Includes all the steps listed above but is performed individually for each gene-set that is provided as an input to SurPRSe.

Scripts used:

- geneset_PRS_analysis.sh
- Assign_SNPs_to_genes_from_geneset.R
- training_set_QC_geneset.R
- geneset_PRS_scoring.R
- geneset_PRS_scoring_plink.sh
- Collate_all_genesets.R

geneset_PRS_analysis.sh
Gene-set PRS are processed here. Initially commands are used to set up a directory structure on the supercomputer hawk to keep the files processed for gene-set PRS separate from that of the genome-wide PRS. The gene-set analysis software MAGMA is used to annotate the SNPs to each gene-set with helper R scripts (Leeuw et al., 2015). Finally, PLINK and further helper R scripts are used to produce gene-set PRS (Chang et al., 2015).

Assign_SNPs_to_genes_from_geneset.R
An R script located within `geneset_PRS_analysis.sh`. This script predominantly acts as a check to see if MAGMA has annotated the SNPs to each gene within each gene-set correctly (Leeuw et al., 2015).

training_set_QC_geneset.sh

An R script called from within `geneset_PRS_analysis.sh`. This script collates the training set for PRS analysis together into one file after it was previously split by chromosome to allow for parallel processing

geneset_PRS_scoring.R
An R script called from within `geneset_PRS_analysis.sh`. For each gene-set, the script creates the `.score` file required for the sotware PLINK 1.9 to create polygenic risk scores and records cases where no SNPs are recorded within the gene-set (eg. a small gene-set at a Pt of 5e-08) (Chang et al., 2015).

geneset_PRS_scoring_plink.sh
A shell script run in parallel within `geneset_PRS_analysis.sh`. Polygenic risk scoring is processed by PLINK 1.9 (Chang et al., 2015).

Collate_all_genesets.R
An R script located within `geneset_PRS_analysis.sh` which converts all gene-set PRS profiles created with `geneset_PRS_scoring_plink.sh` into one tab delimited text file. Each gene-set PRS is defined in its respective column from within the text file.

Module 4: Create Whole Genome PRS

If the user selects the option to produce a genome-wide PRS, the following scripts will be run and a genome-wide PRS produced:

- genomewide_PRS_analysis.sh
- extracting_useful_SNP_information.R
- training_set_QC_genomewide.R
- PRS_scoring_genomewide.R
- PRS_scoring_genomewide_plink.R

genomewide_PRS_analysis.sh
Similar in function as `testing_set_QC_and_clumping.sh`. Is only used if gene-set analysis is not specified within the configuration `config.sh`

extracting_useful_SNP_information.R

A script that records the number of SNPs in both training and testing set during the processing of SurPRSe. The number of SNPs is saved within a tab-delimited text file for the user to check to ensure that the analysis has been performed correctly and provides useful supplemental information.

training_set_QC_genomewide.sh
Similar in function to `training_set_QC_geneset.R` but formats the training set to be processed by PLINK 1.9 (Chang et al., 2015).

training_set_QC_genomewide.sh

Performs Polygenic risk scoring for a genome-wide PRS.

training_set_QC_genomewide_plink.sh

Performs a similar function as `geneset_PRS_scoring.R` but creates a PLINK score file for a whole genome PRS.

Module 5: Auxiliary analyses

In addition to the output of either a gene-set or genome-wide PRS, SurPRSe is able to produce a number of files that supplement a PRS analysis. It is able to run MAGMA (Leeuw et al., 2015) in order to define gene-sets for the gene-set PRS, convert various formats of gene-set annotation files, and provide a gene-wide PRS which encompasses all genes from a background gene annotation file.

Scripts used:

- MAGMA_extract_SNP_list.R
- MAGMA_gene_set_analysis.sh
- SNP_loc_creator_and_gmt_formatter.R
- training_set_QC_genomewide_unclumped.R
- Collate_all_PRS_files_together.R

MAGMA_extract_SNP_list.R

Creates a table of SNPs identifiers between testing and training data set that can be used as an input into MAGMA (Leeuw et al., 2015).

MAGMA_gene_set_analysis.sh

A helper script that performs a MAGMA gene-set analysis on the training data set. Also performs a **F**amily-**W**ise **E**rror **R**ate (FWER) correction by running 100,000 permutations at an alpha of 0.05.

SNP_loc_creator_and_gmt_formatter.R

A common file format for gene-set analysis is the **G**ene **M**atrix **T**ranspose (GMT) format (see Figure A.6a). This R script converts the gene-set input for MAGMA's gene-set analysis into the GMT format.

training_set_QC_genomewide_unclumped.R

Similar in function to `training_set_QC_genomewide.R` but SNPs prior to elimination via the clumping procedure are included.

Collate_all_PRS_files_together.R

Similar in function to `Collate_all_genesets.R` but collates the genome-wide, the gene-centric and the gene-set PRS together into a singular output file.

3.3 App Development: Polygenic Risk Score Analysis Viewing Environment (PRSAVE)

The output of SurPRSe is a multi-row table of PRS which may include gene-set PRS, genome-wide PRS and gene-wide PRS. For use within this thesis, the majority of these PRS were used to test for an association with a phenotype. A linear/logistic regression was the most commonly used statistical test. If just using the table of regression results per PRS, it may be hard to interpret without a visual aid. To aid in this process, a shiny app called PRSAVE was developed. PRSAVE is able to plot the regression results of the genome-wide, the gene-centric and the gene-set PRS in the same set of plots within an internet browser located at: `https://johnhubertjj.shinyapps.io/Viewing_PRS_two_files/`.

3.3.1 File input

The output to SurPRSe is a table of PRS. Regression analysis must then be performed manually.

After statistical analysis with each of the PRS, the output is semi-standardised dependent on the phenotype, the research question and the type of statistical modelling that was used. PRSAVE requires at least the estimate (aka effect size), the standard error of the estimate, the p-value and the r squared value from the statistical model converted into summary statistics for each gene-set acPRS (labelled as ′score′ as ‘acPRS is not a standardised abbreviation within the Neuropsychiatric field; see Table 3.2).

TABLE 3.2: input file for PRSAVE.

.id	score	estimate	SE	tvalue	p	r.squared	lower	upper
Cognition_all_samples	extended_geneset_SCORE_5HT_2C_1e-06	-0.01	0.04	-0.28	0.77	8.82e-05	-0.10	0.08
Cognition_all_samples	extended_geneset_SCORE_5HT_2C_1e-04	-0.04	0.04	-0.94	0.34	9.92e-04	-0.13	0.05
Cognition_all_samples	extended_geneset_SCORE_5HT_2C_0.01	-0.03	0.04	-0.77	0.43	6.76e-04	-0.12	0.05
Cognition_all_samples	extended_geneset_SCORE_5HT_2C_0.05	-0.02	0.04	-0.56	0.57	3.52e-04	-0.11	0.06
Cognition_all_samples	extended_geneset_SCORE_5HT_2C_0.1	-0.02	0.04	-0.63	0.52	4.52e-04	-0.11	0.05

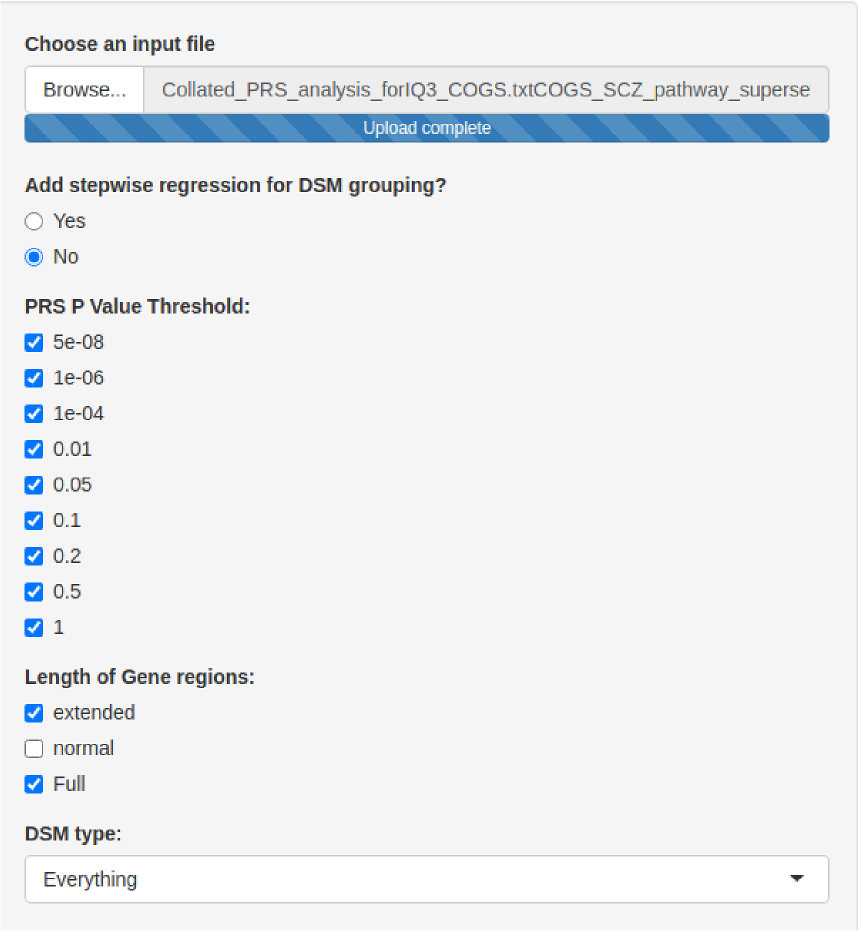
.id = grouping variable, score = PRS, estimate = standardised coefficient of regression, SE = Standard Error, p = p value, lower = lower confidence interval, upper = upper confidence interval

The first column of the input file to PRSAVE is a grouping variable if required. If there is no grouping variable, the same string is repeated for each row. The ′score′ column contains the PRS profile identifier. The remaining columns contain summary statistics for a regression. Of particular note, the estimate is the beta coefficient for the regression. p contains the p-value of the regression. The r.squared contains information describing the variation explained by the PRS within the regression. This is Nagelkerke′s r-squared when the regression is logistic. The ′lower′ and ′upper′ columns contain the 95% confidence intervals for the beta coefficient.

Polygenic **R**isk **S**core **A**nalysis **V**iewing **E**nvironment (PRSAVE) was designed to visualise the output file after logistic or linear regression with a PRS, using the ggplot2 package for R.

3.3.2 Output

On entering the landing page, the user must first upload an input file to visualise (see Figure 3.5.



Choose an input file

Collated_PRS_analysis_forIQ3_COGS.txtCOGS_SCZ_pathway_superse

Add stepwise regression for DSM grouping?

○ Yes
● No

PRS P Value Threshold:

☑ 5e-08
☑ 1e-06
☑ 1e-04
☑ 0.01
☑ 0.05
☑ 0.1
☑ 0.2
☑ 0.5
☑ 1

Length of Gene regions:

☑ extended
□ normal
☑ Full

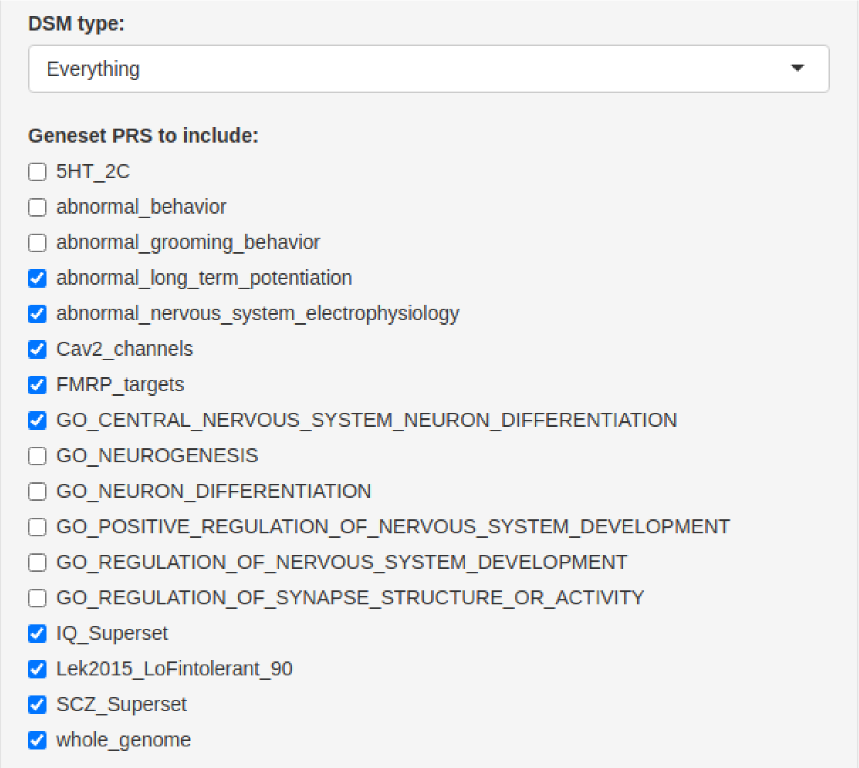
DSM type:

Everything

FIGURE 3.5: **Top half of the sidepanel of PRSAVE**. The top button is where the user uploads the input file required for PRSAVE to populate the page. The second option asks whether stepwise regression should be applied to a grouping variable (in this case. DSM. The third option [provides a series of checkboxes for the P-value thresholds the user wishes to include in the plots to the right hand side of the landing page. The fourth option contains a series of checkboxes of where to define the gene regions for the gene-set PRS. The fifth option allows for the user to select individuals based on a grouping variable (in this case, the DSM)

There is an initial option to provide stepwise regression, but this option is depreciated and will be removed on the next iteration of PRSAVE. The third option for the user to select is the Pt under which the PRS scores were modelled. These selections will populate these PRS results on the right hand side of the page. The fourth option is to select the physical length of the genes that were used within the gene set PRSs. Extended gene regions indicate that as well as the inclusion of SNPs within the gene boundary, the regulatory regions of the gene were included as well (35 kb upstream and 10 kb downstream of the gene boundaries). 'Normal' indicates that only the gene boundaries were used and 'full' includes the genome wide and the genic PRS into the plots to the right of the page as well.

The remainder of the sidepanel (Figure 3.6) is a selection of check-boxes of which PRS to include within the analysis. These are the column headings of the PRS within the input file and are printed verbatim.



DSM type:

Everything

Geneset PRS to include:

- □ 5HT_2C
- □ abnormal_behavior
- □ abnormal_grooming_behavior
- ☑ abnormal_long_term_potentiation
- ☑ abnormal_nervous_system_electrophysiology
- ☑ Cav2_channels
- ☑ FMRP_targets
- ☑ GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION
- □ GO_NEUROGENESIS
- □ GO_NEURON_DIFFERENTIATION
- □ GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT
- □ GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT
- □ GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY
- ☑ IQ_Superset
- ☑ Lek2015_LoFintolerant_90
- ☑ SCZ_Superset
- ☑ whole_genome

FIGURE 3.6: **Selection of PRS to include within the output of PRSAVE**. This selection is directly below Figure 3.5

After all the sidepanel inputs are selected by the user, three plots will populate the page to the right hand side of the sidebar. The topmost plot is the p-value plot (Figure 3.7).

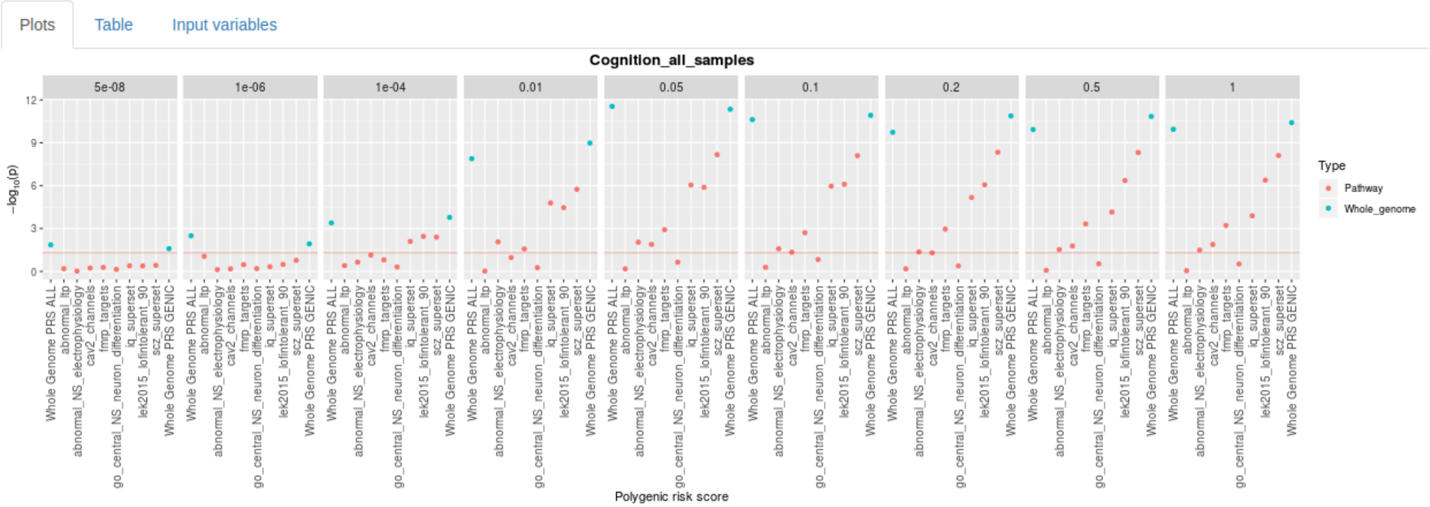


FIGURE 3.7: **P value plot from PRSAVE**. There are n number of faceted plots for each Pt m, where n is the number of selected Pt check boxes in the sidebar and m are the values of these check boxes. The x-axis contains the PRS selected from the sidebar. If the whole genome PRS is selected, the genome-wide PRS will show as the left-most point within each Pt facet, and the genic PRS will show as the right-most point within each Pt facet. The y-axis is the -log10 P value of each PRS which was calculated as the level of association of the PRS with the response variable (in this example above, cognition within schizophrenia). If a point is above the red line, the PRS is considered to be significantly associated with the response variable.

Each plot (the p-value plot, the beta plot and the r-squared plot) on the landing page is faceted by Pt selected by the user in the side panel. They are ordered from the lowest selected value to the highest selected value (in the case of Figure 3.7; 5e-08 and 1). Each PRS that is selected by the user in the side panel is located on the x-axis of every plot. If the 'Whole genome' option within 'Geneset PRS to include' and 'Full' in 'Length of Gene regions' options are selected on the side panel then the genome wide and the genic plots are also included within every plot. The genome-wide will always be the right-most point on the x-axis and the genic PRS will always be the left-most option on the x-axis. The y-axis records the -log10 P value of each PRS to normalise the data and therefore allow a better visualisation of the differences between each PRS. A red line signifying the level of nominal significance is also included to allow the user to clearly see which PRS were found to be associated to the response variable.

Directly below the P value plot is the beta plot as described in 3.8. If any of the error bars for each PRS are non-overlapping with the red line (an effect size of zero), that PRS is deemed to be significantly associated with the response variable.

The final plot on the landing page is the R-squared plot which reports the variance of the response variable explained by the PRS in the form a percentage (Figure 3.9. The direction of the effect size and the p-value of the PRS is also incorporated.

Supplementary information for the plots on the landing page is provided on the 'Table' (Figure 3.10) and the 'Input variables' (Figure 3.11) tabs. The aim of the 'Table' tab is to provide the user with the raw data in order to supplement the plots on the landing page. The 'Input variables' tab contains a page of text that can be recognised by the programming language R, so if the user wishes to alter the plots manually, they are already given example code to do so.

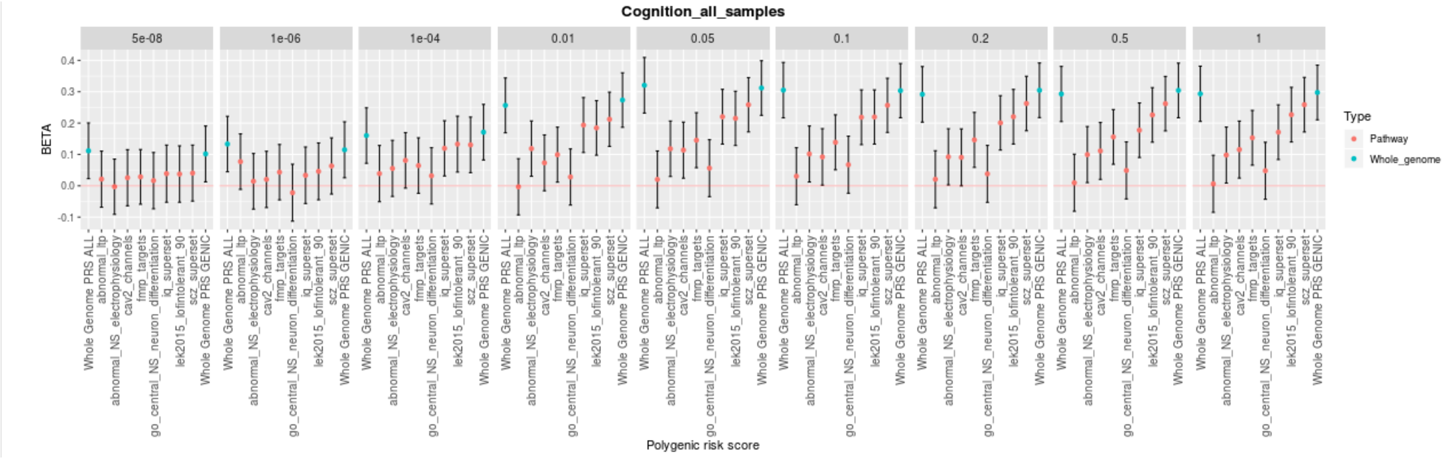


FIGURE 3.8: **Beta plot from PRSAVE**. The x-axis, facets and layout of the plots are equivalent to Figure 3.7. The y-axis is the effect size for each PRS. The error bars for each point signify the 5% Confidence interval for each PRS. The red line signifies an effect size of zero.

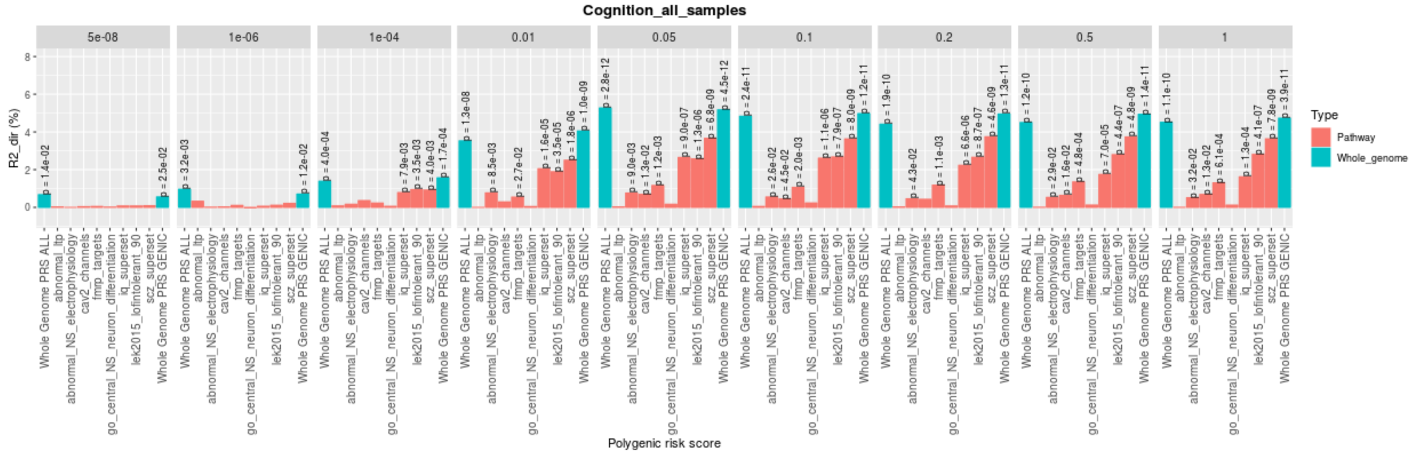
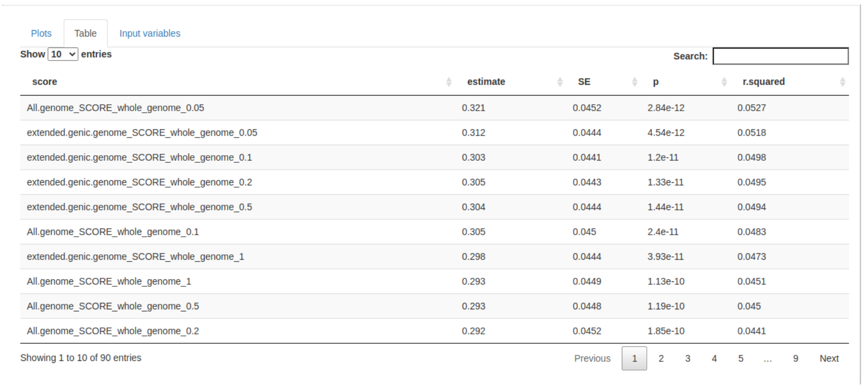


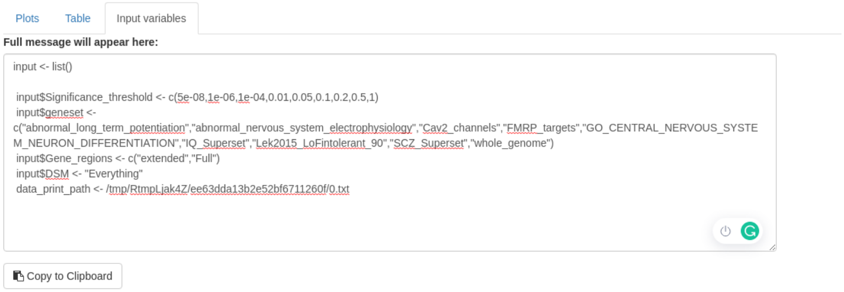
FIGURE 3.9: r^2 **plot from PRSAVE**. The x-axis and facets are equivalent to Figure 3.7. Instead of points, each PRS is represented by a bar. The y-axis describes the r^2 value of the PRS in the form of a percentage. The direction of effect size is also incorporated into the plot (if the bar is below the x-axis it contained a negative effect size and vis-versa.) If an associated of the PRS with the response variables was significant after multiple correction testing, the FDR corrected p-value will be displayed at the top of the relevant bar



score	estimate	SE	p	r.squared
All.genome_SCORE_whole_genome_0.05	0.321	0.0452	2.84e-12	0.0527
extended.genic.genome_SCORE_whole_genome_0.05	0.312	0.0444	4.54e-12	0.0518
extended.genic.genome_SCORE_whole_genome_0.1	0.303	0.0441	1.2e-11	0.0498
extended.genic.genome_SCORE_whole_genome_0.2	0.305	0.0443	1.33e-11	0.0495
extended.genic.genome_SCORE_whole_genome_0.5	0.304	0.0444	1.44e-11	0.0494
All.genome_SCORE_whole_genome_0.1	0.305	0.045	2.4e-11	0.0483
extended.genic.genome_SCORE_whole_genome_1	0.298	0.0444	3.93e-11	0.0473
All.genome_SCORE_whole_genome_1	0.293	0.0449	1.13e-10	0.0451
All.genome_SCORE_whole_genome_0.5	0.293	0.0448	1.19e-10	0.045
All.genome_SCORE_whole_genome_0.2	0.292	0.0452	1.85e-10	0.0441

Showing 1 to 10 of 90 entries

FIGURE 3.10: **Input file table**. The input file used to as an input for PRSAVE, displayed as a data table

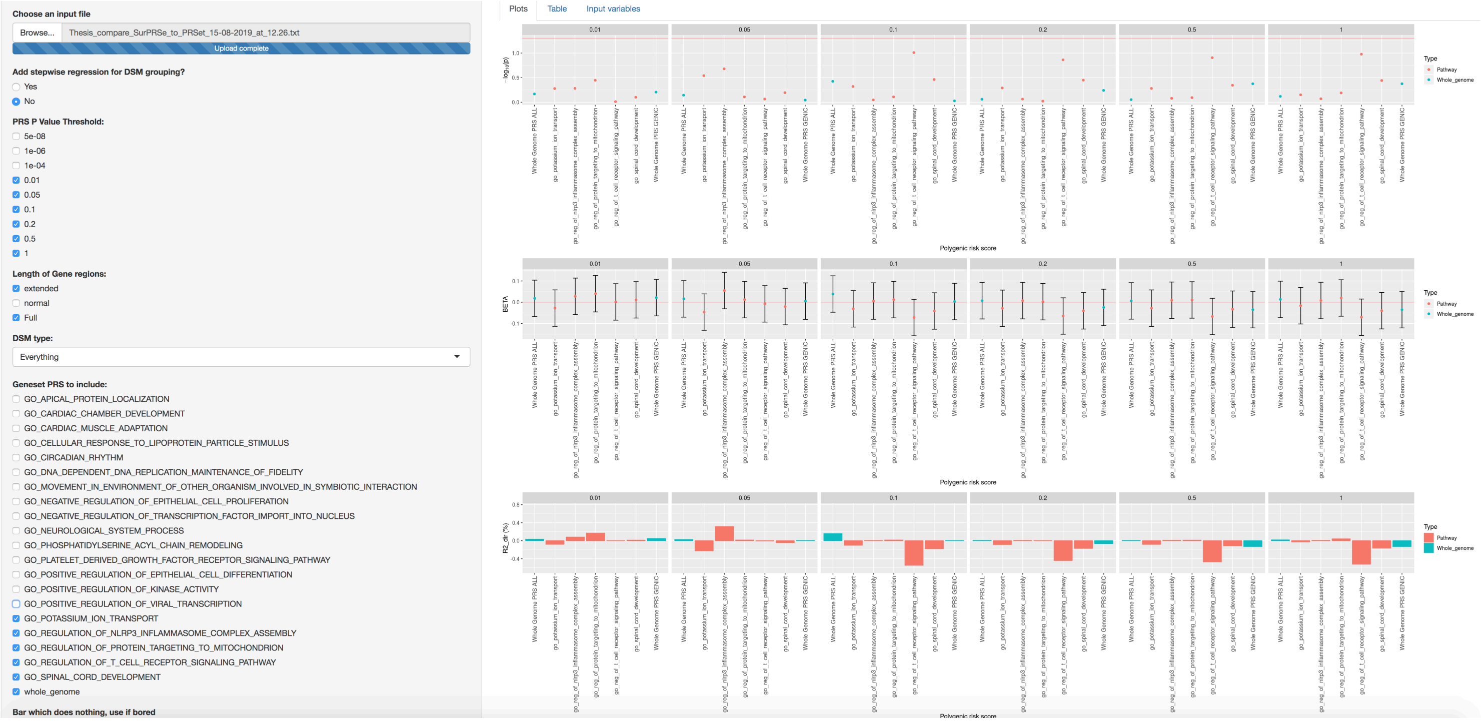


Full message will appear here:

```
input <- list()

 input$Significance_threshold <- c(5e-08,1e-06,1e-04,0.01,0.05,0.1,0.2,0.5,1)
 input$geneset <-
c("abnormal_long_term_potentiation","abnormal_nervous_system_electrophysiology","Cav2_channels","FMRP_targets","GO_CENTRAL_NERVOUS_SYSTE
M_NEURON_DIFFERENTIATION","IQ_Superset","Lek2015_LoFintolerant_90","SCZ_Superset","whole_genome")
 input$Gene_regions <- c("extended","Full")
 input$DSM <- "Everything"
 data_print_path <- /tmp/RtmpLjak4Z/ee63dda13b2e52bf6711260f/0.txt
```

FIGURE 3.11: **Plotting code**. The PRSAVE user inputs to the produce the plots in the form of R code

The full layout of all the individual components of PRSAVE above, is provided in Figure 3.12. Note that in Figure 3.12, a different input file was used and therefore displays different results than displayed in the figure above.



Choose an input file

Browse... Thesis_compare_SurPRSe_to_PRSet_15-08-2019_at_12.26.txt

Upload complete

Add stepwise regression for DSM grouping?

○ Yes
● No

PRS P Value Threshold:

- □ 5e-08
- □ 1e-06
- □ 1e-04
- ☑ 0.01
- ☑ 0.05
- ☑ 0.1
- ☑ 0.2
- ☑ 0.5
- ☑ 1

Length of Gene regions:

- ☑ extended
- □ normal
- ☑ Full

DSM type:

Everything

Geneset PRS to include:

- □ GO_APICAL_PROTEIN_LOCALIZATION
- □ GO_CARDIAC_CHAMBER_DEVELOPMENT
- □ GO_CARDIAC_MUSCLE_ADAPTATION
- □ GO_CELLULAR_RESPONSE_TO_LIPOPROTEIN_PARTICLE_STIMULUS
- □ GO_CIRCADIAN_RHYTHM
- □ GO_DNA_DEPENDENT_DNA_REPLICATION_MAINTENANCE_OF_FIDELITY
- □ GO_MOVEMENT_IN_ENVIRONMENT_OF_OTHER_ORGANISM_INVOLVED_IN_SYMBIOTIC_INTERACTION
- □ GO_NEGATIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION
- □ GO_NEGATIVE_REGULATION_OF_TRANSCRIPTION_FACTOR_IMPORT_INTO_NUCLEUS
- □ GO_NEUROLOGICAL_SYSTEM_PROCESS
- □ GO_PHOSPHATIDYLSERINE_ACYL_CHAIN_REMODELING
- □ GO_PLATELET_DERIVED_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY
- □ GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_DIFFERENTIATION
- □ GO_POSITIVE_REGULATION_OF_KINASE_ACTIVITY
- □ GO_POSITIVE_REGULATION_OF_VIRAL_TRANSCRIPTION
- ☑ GO_POTASSIUM_ION_TRANSPORT
- ☑ GO_REGULATION_OF_NLRP3_INFLAMMASOME_COMPLEX_ASSEMBLY
- ☑ GO_REGULATION_OF_PROTEIN_TARGETING_TO_MITOCHONDRION
- ☑ GO_REGULATION_OF_T_CELL_RECEPTOR_SIGNALING_PATHWAY
- ☑ GO_SPINAL_CORD_DEVELOPMENT
- ☑ whole_genome

Bar which does nothing, use if bored

Plots
Table
Input variables
0.01
0.05
0.1
0.2
0.5
1
−log10(p)
BETA
R2_dir (%)
Whole Genome PRS ALL
go_potassium_ion_transport
go_reg_of_nlrp3_inflammasome_complex_assembly
go_reg_of_protein_targeting_to_mitochondrion
go_reg_of_t_cell_receptor_signaling_pathway
go_spinal_cord_development
Whole Genome PRS GENIC
Polygenic risk score
Type
Pathway
Whole_genome

FIGURE 3.12: **Screenshot of the output from Section 3.5.3 in PRSAVE**. The bar on the left hand-side indicates options to view the polygenic risk score. The user uploads the input file to PRSAVE (See Table 3.2) by clicking on the 'Browse' icon. The user then selects the options from 'PRS P Value Threshold', 'Length of Gene regions' and the 'Geneset PRS to include' to inform which PRS to view. The results are displayed on the right-hand side of the screen. The title of each plot is the Pt and the X-axis for each plot are the PRS. The legend indicates that the red coloured points are gene-set PRS and the blue coloured points are genome-wide or gene-centric PRS. In the top row of plots, the y-axis is the -log10 P value and the red line indicates a nominal P-value of 0.05. The y-axis in the centre row of plots indicates the beta coefficient from the regression. The error bars indicate the Standard error of the Beta coefficient. The bottom row of plots have a y-axis indicating the r-squared value with the direction of the beta incorporated. If the value is below zero then the PRS contained a negative direction of effect on the phenotype in question and vis-versa if the value is positive.

There is no standardised method for creating a bioinformatics workflow (Leipzig, 2016). There is no requirement for a workflow to align to an architecture or to a standardised practice in writing code, and each line of code is not required to be tested for quality. They are therefore prone to human and systematic error. Unfortunately, there are not many solutions and/or the resources available within an academic environment to solve this problem as it requires an engineering solution, a resource not available to most neuropsychiatric departments. However, the probability of human error can be reduced, if the ability to interpret the data is improved, and the complexity of the output from the data is reduced.

Within SurPRSe, the number of gene-sets that are used to create the gene-set PRS is up to the user, and depending on the scope of the research project, can number in the thousands. It would be impractical to check if each gene-set PRS has been processed correctly by hand. PRSAVE aims to counter this issue by adding an interactive component to the output of a gene-set PRS analysis. The user can quickly plot and visualise the gene-set PRS without any coding knowledge, making the results more accessible. Additionally, by selecting the parameters for the inclusion of the gene-set PRS, and the ability to remove unwanted PRS, helps to interpret and visualise the data, especially to external users.

3.4 Testing

3.4.1 SurPRSE

Analysis set-up

SurPRSe was set-up to run multiple gene-set PRS, a single gene-centric PRS and a single genome-wide PRS at eight pre-defined Pt (5e-08, 1e-06, 1e-04, 0.01, 0.05, 0.1, 0.2, 0.5, 1). The input to SurPRSe was the 1000 genomes population sample (Auton and Abecasis, 2015) which acted as the target set and the iPsych GWA study (Grove et al., 2019) which acted as the training set. Both data sets were located in the working directory where the analysis took place. For the gene-set PRS, the gene-sets were defined as 20 **G**ene**O**ntology (GO) ontology sets located within the MSigDB database (located here: `http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C5` and described in Table 3.3). The annotations were downloaded directly from MsigDB in the format of a MsigDB annotation file (see Figure A.6). SurPRSe was tested with an input of 1, 5, 10 then 20 gene-sets in four separate runs. In order to provide a benchmark for SurPRSe, the same analysis was set up using PRSet, a bioinformatics tool that is a component of the software PRSice (Choi and O'Reilly, 2019). PRSice is software designed to automate the majority of the steps required to produce a PRS. This includes QC (clumping, data cleaning and strand flipping of SNPs, Pt, and finally the calculation and evaluation of the PRS. The first version of the software was released in 2015 (Euesden, Lewis, and O'Reilly, 2015), and a second version was released in 2019 (Choi and O'Reilly, 2019) which improved the speed and the efficiency of version one. PRSet is a new component of PRSice which allows for the production of gene-set PRS.

TABLE 3.3: Gene sets used in comparison of SurPRSe to PRSet.

Gene set	Membership*	Number of Genes
Positive regulation of Viral Transcription	A,B,C,D	39
Cardiac Chamber Development	B,C,D	144
DNA dependent DNA replication maintenance of fidelity	B,C,D	24
Circadian rhythm	B,C,D	137
Phosphatidylserine acyl chain remodeling	B,C,D	17
Spinal cord development	C,D	106
Platelet derived growth factor receptor signaling pathway	C,D	34
Cellular response to lipoprotein particle stimulus	C,D	13
Regulation of NLRP3 inflammasome complex assembly	C,D	11
Positive regulation of epithelial cell differentiation	C,D	57
Positive regulation of kinase activity	D	482
Negative regulation of transcription factor import into nucleus	D	39
Potassium ion transport	D	154
Regulation of T cell receptor signaling pathway	D	29
Cardiac muscle adaptation	D	11
Negative regulation of epithelial cell proliferation	D	116
Movement in environment of other organism involved in symbiotic interaction	D	87
Regulation of protein targeting to mitochondrion	D	98
Apical protein localization	D	12
Neurological system process	D	1242

*The Gene-sets used in the four separate runs of PRSet and SurPRSe. Letters indicate which gene-set was used when testing different numbers of gene-sets. A = 1, B = 5, C = 10, D = 20.

In Table 3.3, the first column indicates the gene-set taken from MSigDB (Subramanian et al., 2005; Liberzon et al., 2011). The 'Membership' column indicates which gene-set was used for each run in SurPRSe. The last column indicates the number of genes that are contained within each gene-set.

3.4.2 Processing Speed

The speed of SurPRSe was measured using the command-line tool 'time' wrapped around the SurPRSe workflow within a hawk job submission script (Supplementary

Figure A.1). As with the analysis set-up, PRSet will be used as a benchmark and wrapped in its own submission script (See Supplementary Figure A.2).

The time command produces three output metrics: Real, User and Sys. The Real metric describes the wall clock time from when the job is submitted to the Hawk supercluster, to when the job finished processing on the Hawk supercluster.

The User metric measures the amount of **C**entral **P**rocessing **U**nit (CPU) time spent in user-mode code. Simply put, it is the time that code which can be accessed by the user is being processed by the Hawk processors. This excludes any time where for example, a file is being written into one of the directories on Hawk or anything inside the kernel (code which automatically allocates memory depending on the needs of the job and cannot be accessed by the user). The User metric may exceed the Real metric in cases where more than one CPU is used as the User metric records the time across all processes, across all CPUs.

The sys metric records the time spent in the kernel while the code is processing. This is code that cannot be accessed by the user and can involve, for example, the allocation of memory or accessing a network.

The analysis set-up is the same as described in section 3.4.1. The Real, User and sys metrics were compared when the input to each bioinformatics tool was 1, 5, 10 and then 20 gene-sets respectively.

3.4.3 Production of PRS

The number of SNPs before clumping, the number of mismatched SNPs and the total number of SNPs used for polygenic risk scoring was recorded for each analysis run. The genome-wide, the gene-centric and the 20 gene-set PRS produced by both SurPRSe and PRSet were compared by calculating the pearson correlation coefficient between each relevant PRS.

3.4.4 Data Visualisation

SurPRSe has no functionality to visualise data. Instead, the statistical analysis performed with the PRS is processed using a semi-automated R script. The output of this script is input into the shiny app PRSAVE (see section 3.3). PRSAVE accepts standardised summary statistics as an input which can be created from the output of both SurPRSe. The -log10 P-value, the beta coefficient of the statistical test and the r-squared value are all plotted on PRSAVE. Interactive decile and centile analysis of the PRS is also possible.

A randomly generated set of normally distributed continuous values for the individuals of the 1000 genomes data set (using mean = 0 and sd = 1) was created using the R command Rnorm. This provided a set of values to test for association with the PRS produced by SurPRSe and enabled an in depth look at the data visualisation capabilities of PRSAVE.

The 1000 genomes random phenotype was tested for association with the genome-wide, the gene-centric and the 20 gene-set PRS produced by SurPRSe using a separate R script. The results were saved as the input file to PRSAVE. A screen-shot of the plots displaying the -log10 P-value, the beta coefficient of the linear regression and the r-squared value in PRSAVE was produced.

3.5 Results

3.5.1 Analysis set up

SurPRSe contains a wiki page which allows an easy set-up on the Hawk supercomputer (`https://github.com/johnhubertjj/SurPRSe/wiki`). PRSet similarly contains an in-depth guide on how to set up an analysis (`http://www.prsice.info/prset_detail/`) and is easily suited towards a supercomputer environment despite no explicit instructions on set up.

3.5.2 Processing Speed

SurPRSe and PRSet both had comparative Sys times up to 10 gene sets (see Table 3.4). PRSet was more efficient than SurPRSe for any number of gene-sets for the User times. PRSet was more efficient across all metrics with 20 gene-sets. The SurPRSe processes were run with 22 CPUs simultaneously on Hawk, while PRSet used one CPU.

TABLE 3.4: Processing times for the runthrough of PRSet and SurPRSe with an increasing number of gene-sets used as an input.

Number of Gene-sets		**SurPRSe**	**PRSet**
1	Real	15m 24.14s	27m 18.12s
	User	55m 9.60s	8m 30.234s
	Sys	10m 9.56s	11m 41.10s
5	Real	20m 30.58s	27m 45.56s
	User	57m 43.15s	8m 35.85s
	Sys	11m 8.59s	11m 35.83s
10	Real	35m 1.18s	30m 55.79s
	User	58m 16.26s	8m 34.88s
	Sys	11m 25.44s	11m 49.50s
20	Real	41m 11.29s	28m 46.50s
	User	66m 38.67s	8m 42.89s
	Sys	13m 1.56s	11m 40.44s

The first column of Table 3.4 displays the number of gene-sets used for each run of the bioinformatics tool. The second column displays the respective times for both SurPRSe and PRSet to finish a run.

As more gene-sets were input into each tool the User metric steadily increased for both SurPRSe and PRSet. The rate at which time increased appeared to be higher for SurPRSe. The Real metric did not substantially change for PRSet across all gene-set inputs. In comparison, SurPRSe appeared to show a linear increase in time for the Real metric.

3.5.3 Production of PRS

There are approximately 40,000 extra SNPs included within the PRSet analysis (See table 3.5). In order to explain the discrepancy between these results, the 'Apical Protein Localisation' gene-set PRS and the 'Cardiac Chamber Development' gene-set PRS were produced again with PRSet, but the '–print-snp' argument was included to directly compare the SNPs included within each analysis with SurPRSe (See table 3.8).

TABLE 3.5: Number of SNPs at each stage of PRS production.

Stage of Analysis	Number of SNPs in SurPRSe	Number of SNPs in PRSet
iPSYCH total*	9,112,386	9,112,386
1000genomes total*	11,915,643	11,915,643
After merge and QC	3,615,347	7,320,806
After Clumping	71,572	114,179

*total refers to the number of SNPs that were available from the data set provided in the original publications (Grove et al., 2019; Auton and Abecasis, 2015)

SurPRSe produced a PRS profile for each gene set PRS at each Pt. NA values were produced if the gene set contained no SNPs at the specified Pt. PRSet produced an output file, but the number of PRS profiles at each gene-set for each Pt did not match the number of input PRSs profiles. No error was recorded in the log file. Therefore, to ensure that the PRSet results are accurate, the PRSet was re-run at a Pt of 0.2, 0.5 and 1 as these thresholds produced a PRS profile across all gene set PRSs. Only gene set PRS at these Pt were compared with SurPRSe.

The genome-wide PRS between SurPRSe and PRSet showed a perfect correlation at Pt = 5e-08 (See Figure 3.13). As more SNPs were included into the PRS, the correlation between the bioinformatics tools was reduced to ≈ 0.85. This correlation measure was maintained until all the SNPs were included in the model (Pt = 1).

Using the methodology described in Section 3.4.1, there was almost no correlation (≈ 0) between the gene-centric PRS (see Figure 3.14) and every gene-set PRS at each Pt (See Figures A.3, A.4 and A.5).

Comparison of genome-wide PRS

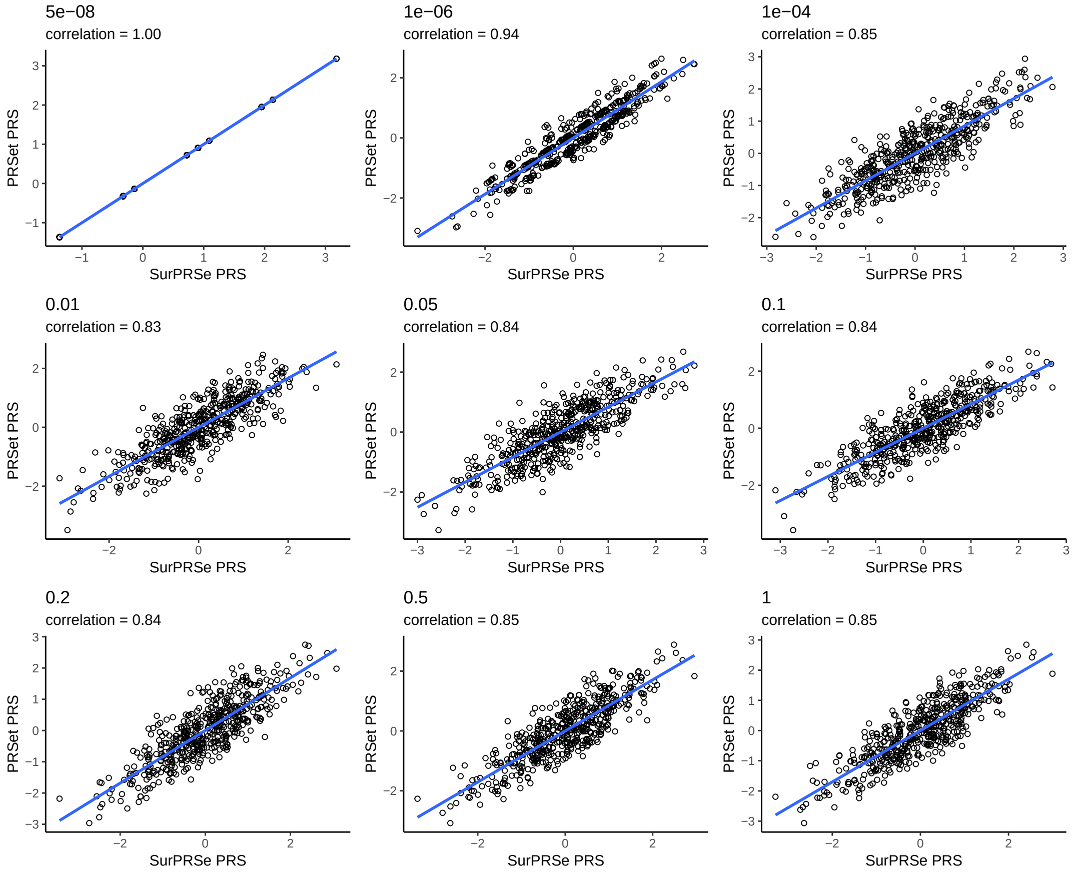


FIGURE 3.13: **Genome-wide comparison of PRS between SurPRSe and PRSet.** Titles of each plot declare the Pt at which the PRS were produced. The pearson correlation coefficient is at the top left of each plot. Each point signifies a PRS of an individual from the 1000 genomes project. The reduction in data within the 5e-08 and the 1e-06 plots is due to the fact that only select individuals had any SNP below these Pt. The PRS were standardised by subtracting the mean from the individual score and dividing by the standard deviation. PRSice version 2.2.6 was used to produce the polygenic risk scores.

Comparison of gene-centric PRS

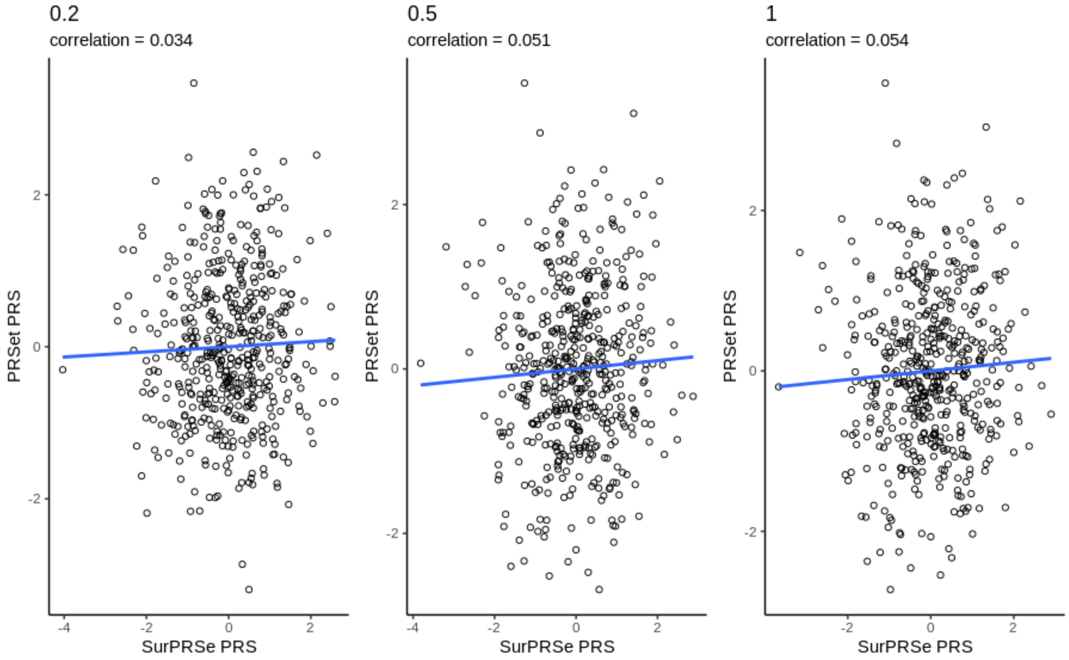
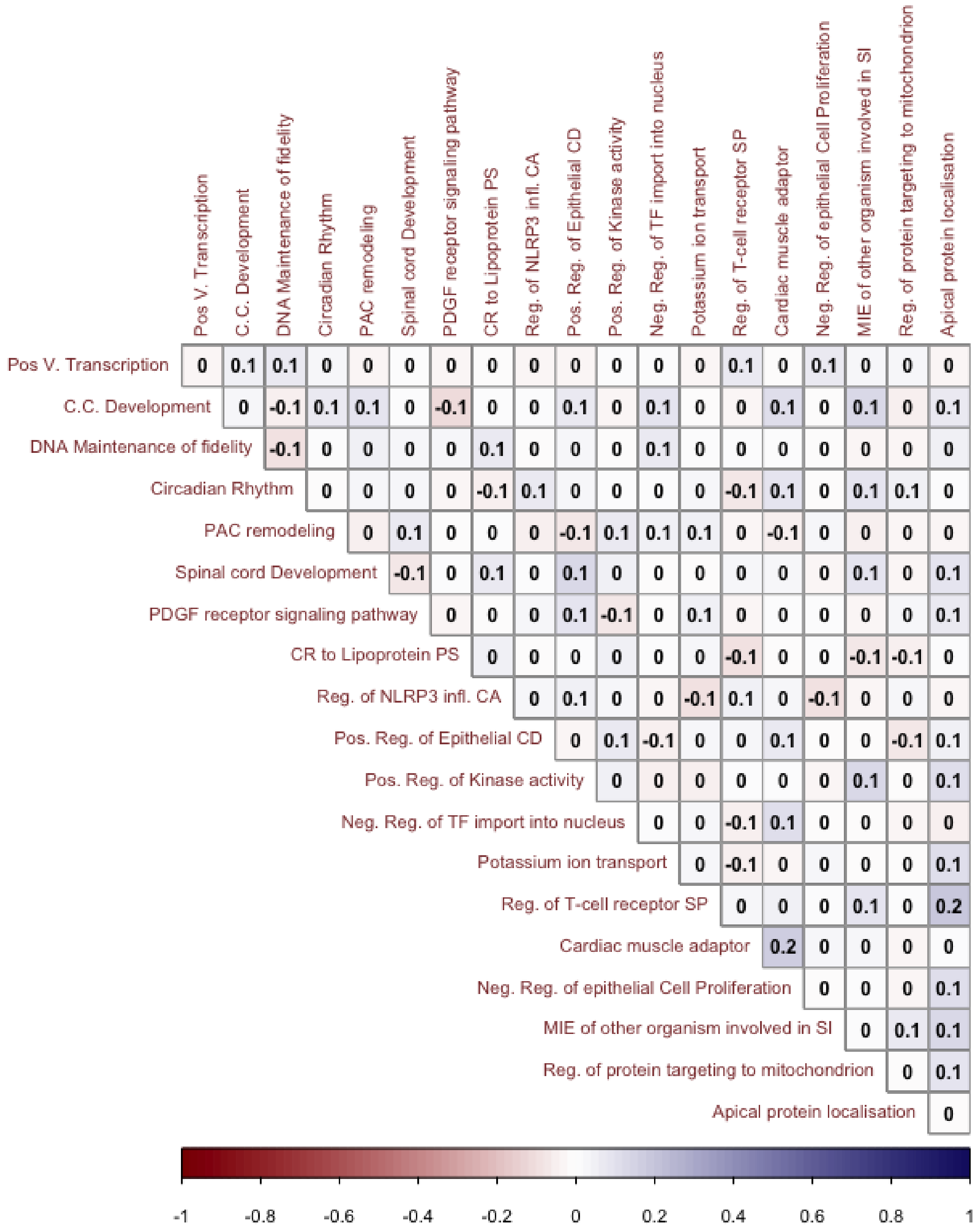


FIGURE 3.14: **Gene-centric comparison of PRS between SurPRSe and PRSet.** Titles of each plot declare the Pt at which the PRS were produced. The Pearson correlation coefficient is provided as the subtitle to each plot. Each point signifies a PRS of an individual from the 1000 genomes project. The PRS were standardised by subtracting the mean from the individual score and dividing by the standard deviation. Each PRS was defined within the genic regions described by the annotation file for each bioinformatics tool.

Comparison of gene-set PRS

No PRSet gene-set PRS or SurPRSe gene-set PRS correlated with each other using the methods described within each of the respective bioinformatics tools inputs. This was the same whether the Pt = 0.2 (figure A.4), the Pt = 0.5 (Supplementary figure A.5), or Pt = 1 (Supplementary figure A.3).



	Pos V. Transcription	C.C. Development	DNA Maintenance of fidelity	Circadian Rhythm	PAC remodeling	Spinal cord Development	PDGF receptor signaling pathway	CR to Lipoprotein PS	Reg. of NLRP3 infl. CA	Pos. Reg. of Epithelial CD	Pos. Reg. of Kinase activity	Neg. Reg. of TF import into nucleus	Potassium ion transport	Reg. of T-cell receptor SP	Cardiac muscle adaptor	Neg. Reg. of epithelial Cell Proliferation	MIE of other organism involved in SI	Reg. of protein targeting to mitochondrion	Apical protein localisation
Pos V. Transcription	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0	0	0
C.C. Development		0	-0.1	0.1	0.1	0	-0.1	0	0	0.1	0	0.1	0	0	0.1	0	0.1	0	0.1
DNA Maintenance of fidelity			-0.1	0	0	0	0	0.1	0	0	0	0.1	0	0	0	0	0	0	0
Circadian Rhythm				0	0	0	0	-0.1	0.1	0	0	0	0	-0.1	0.1	0	0.1	0.1	0
PAC remodeling					0	0.1	0	0	0	-0.1	0.1	0.1	0.1	0	-0.1	0	0	0	0
Spinal cord Development						-0.1	0	0.1	0	0.1	0	0	0	0	0	0	0.1	0	0.1
PDGF receptor signaling pathway							0	0	0	0.1	-0.1	0	0.1	0	0	0	0	0	0.1
CR to Lipoprotein PS								0	0	0	0	0	0	-0.1	0	0	-0.1	-0.1	0
Reg. of NLRP3 infl. CA									0	0.1	0	0	-0.1	0.1	0	-0.1	0	0	0
Pos. Reg. of Epithelial CD										0	0.1	-0.1	0	0	0.1	0	0	-0.1	0.1
Pos. Reg. of Kinase activity											0	0	0	0	0	0	0.1	0	0.1
Neg. Reg. of TF import into nucleus												0	0	-0.1	0.1	0	0	0	0
Potassium ion transport													0	-0.1	0	0	0	0	0.1
Reg. of T-cell receptor SP														0	0	0	0.1	0	0.2
Cardiac muscle adaptor															0.2	0	0	0	0
Neg. Reg. of epithelial Cell Proliferation																0	0	0	0.1
MIE of other organism involved in SI																	0	0.1	0.1
Reg. of protein targeting to mitochondrion																		0	0.1
Apical protein localisation																			0

FIGURE 3.15: **Gene set wide comparison of PRS between SurPRSe and PRSet at Pt = 0.2.** Each box signifies the correlation coefficient between the gene set PRS produced by SurPRSe and PRSet at a Pt of 0.2. The y-axis labels signify the identifiers of the gene set PRS from PRSet and the x-axis signifies PRS from SurPRSe. The diagonal describes direct comparisons of gene-set PRS. The legend and colour of the boxes indicates the correlation coefficient where red is negative and blue is positive. The numerical correlation coefficient is also displayed within each box.

In order to differentiate between an error in the definitions of the gene-sets provided to both SurPRSe and PRSet and an error in the processing of the PRS scores themselves, the gene-set analysis was re-performed using identical sources for gene-sets and their chromosomal locations. In PRSet the definition of each gene-set was manually curated as a separate file using the `.bed` file format (See Appendix; Figure A.7).

PRSet has two methods of inputting a gene-set input, the 'two files' method where one file is obtained from MsigDB (definitions of gene-sets and gene identifiers from ensemble; see Figure 3.6) and a second file is an ensembl GTF file to define locations (Figure 3.7), or a singular .bed file (See Appendix; Figure A.7) at the time of writing. SurPRSe uses a similar method to the PRSet 'two files' method, but at the time of writing, the gene locations file and the gmt file is obtained from NCBI rather than Ensembl.There may be inconsistencies between the two sources for the definitions of the gene-set inputs. For example for the ENSA gene, in Ensembl, the gene is located between **B**ase-**P**airs (BPs) 150,600,851 and 150,629,612 on the reverse strand of chromosome one. However, in NCBI, the gene is stated to be located between BPs 50621246 and 150629612. Therefore, by defining the gene sets manually as a .bed format using the NCBI gene-sets as a reference, there is parity between the gene-set input into PRSet and the gene-set input into SurPRSe.

Two gene-sets at three Pt (0.2,0.5,1) were manually set up to have the exact same gene locations for PRSet as the input annotation file used for SurPRSe. After running each workflow to produce the two sets of gene-set PRS, the PRS were compared using the Pearson correlation coefficient. For consistency, the gene sets are described as "Apical Protein Localisation" and "Cardiac Chamber Development" but these gene-sets have been altered from their original sources in order to maintain consistency between the inputs to SurPRSe and PRSet (See Figure 3.16). This is because gene identifiers are not 1:1 between different resources. In ensembl and NCBI for the same gene-set, some genes may be excluded because an identifier was not allocated to that gene. Therefore, only the genes in common between both ensembl and NCBI were used as an input for SurPRSe and PRSet.

TABLE 3.6: MSigDB File Format

Set	Gene1	Gene2
Apical Protein Localisation	ENSG0000023612	ENSG00000237957
Cardiac Chamber Development	ENSG00000288937	ENSG00000288824

Set describes the name of the gene set. Each Gene is ordered along the same row for the same gene set. Each gene ID is an ensembl gene ID.

TABLE 3.7: GTF File Format

Seqname	feature	start	end	strand
1	gene	11869	14409	+
1	transcript	11869	14409	+

Seqname = Name of chromosome or scaffold, source = data source, feature = feature type id (e.g. gene, Variation, transcript), start = start position of feature with numbering starting at 1, stop = end position of feature with numbering starting at 1, score = float value, strand = either positive (+) or negative (-), frame = either 0, 1 or 2. the number indicates at what position in the feature is the first base of a codon. attribute = semi-colon separated list of information about the feature.

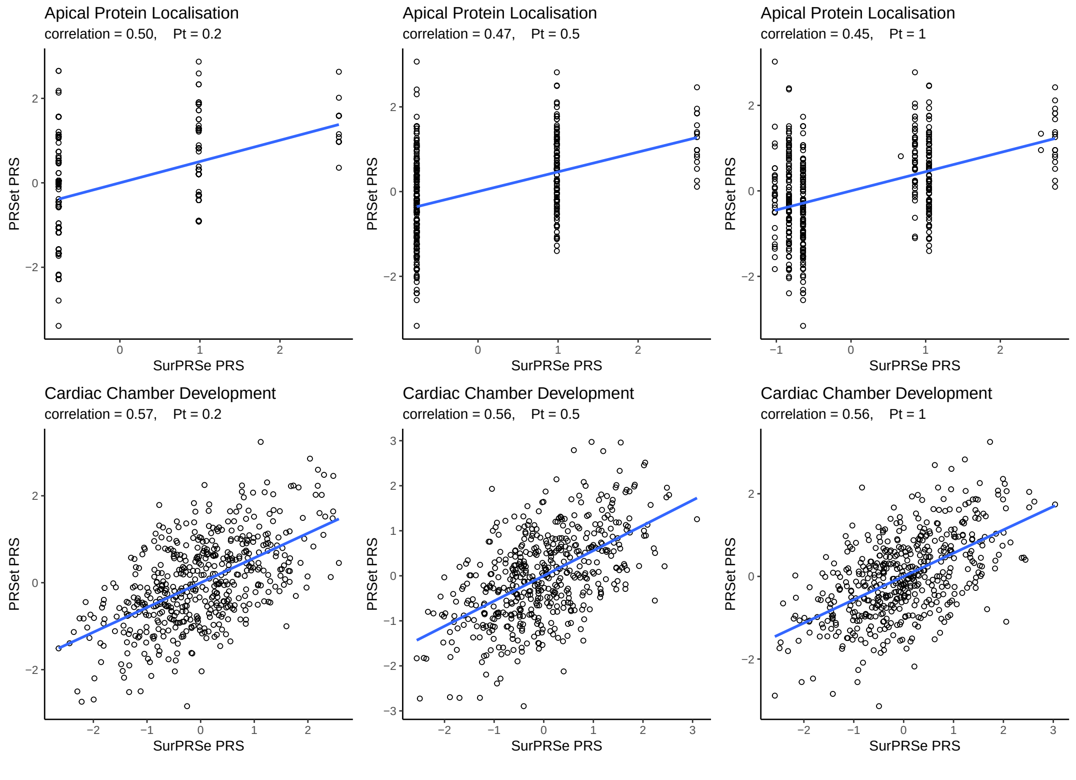


FIGURE 3.16: **Comparison of an altered 'Apical Protein Localisation' gene-set PRS and an altered 'Cardiac Chamber Development' gene-set PRS between SurPRSe and PRset**. 'Apical Protein Localisation' contained 5 genes and 'Cardiac Chamber Development' contained 108 genes. The Pearson correlation coefficient followed by the Pt is at the top left of each plot. Each point signifies a PRS of an individual from the 1000 genomes project. All 503 individuals are included within each plot. The PRS were standardised by subtracting the mean from the individual score and dividing by the standard deviation.

When the gene-set annotations are directly comparable, SurPRSe and PRSet produce PRS which are correlated between 40% and 60%. When the gene-set was small (5 genes) SurPRSe PRS appeared to cluster into three separate groups whereas the spread of the PRSet PRS was larger. To test whether the correlations observed were not due to random chance, the correlation analysis was performed again, but the SurPRSe 'Apical Protein Localisation' gene-set PRS was tested for correlation with the PRSet 'Cardiac Chamber Development' gene-set PRS. As these gene sets do not contain the same genes, There should be zero correlation between these two PRS. This would support the observation that the correlation was genuine between SurPRSe and PRSet when the input gene set was the same.

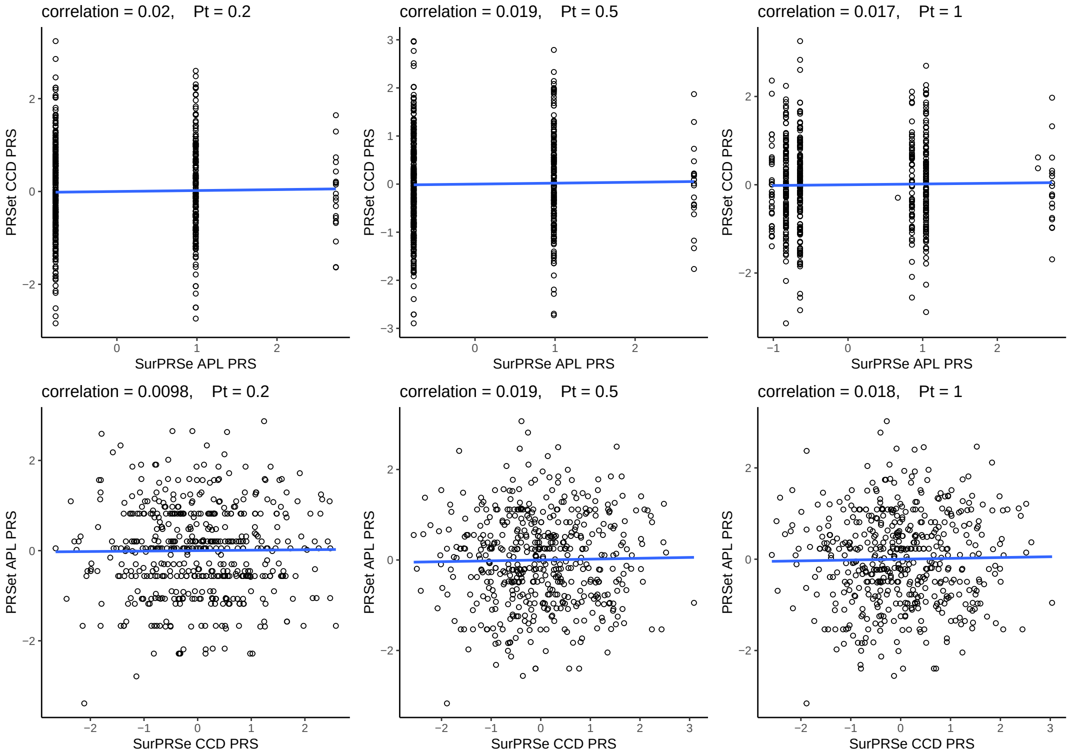


FIGURE 3.17: **Comparison of an altered 'Apical Protein Localisation' gene-set PRS from SurPRSe and PRSet vs. an altered 'Cardiac Chamber Development' gene-set PRS from PRset and SurPRSe respectively.** APL = 'Apical Protein Localisation' which contained 5 genes. 'CCD' = 'Cardiac Chamber Development' which contained 108 genes. The Pearson correlation coefficient followed by the Pt is provided as the title to each plot. Each point signifies a PRS of an individual from the 1000 genomes project. The PRS were standardised by subtracting the mean from the individual score and dividing by the standard deviation.

As displayed in Figures 3.17 and 3.18, there was no correlation between the 'Cardiac

Chamber Development' PRS and the 'Apical Protein Localisation' PRS at any Pt. This includes when the PRS was generated solely by PRSet or SurPRSe or if the PRS was compared across the different software tool platforms.

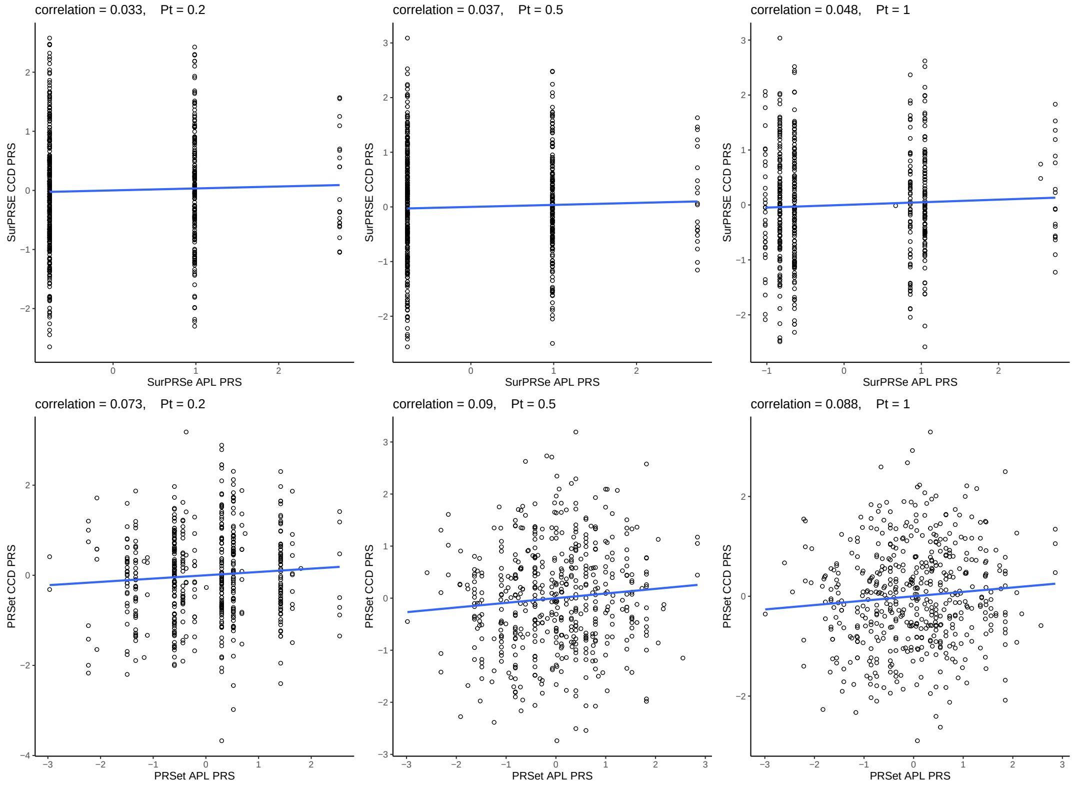


FIGURE 3.18: **Comparison of an altered 'Apical Protein Localisation' gene-set PRS vs. an altered 'Cardiac Chamber Development' gene-set PRS from SurPRSe, then from PRSet respectively.** APL = 'Apical Protein Localisation' which contained 5 genes. 'CCD' = 'Cardiac Chamber Development' which contained 108 genes. The Pearson correlation coefficient followed by the Pt is provided as the title to each plot. Each point signifies a PRS of an individual from the 1000 genomes project. The PRS were standardised by subtracting the mean from the individual score and dividing by the standard deviation.

TABLE 3.8: SNPs contributing to the Apical Protein Localisation gene-set PRS of SurPRSe and PRSet at Pt = 0.2

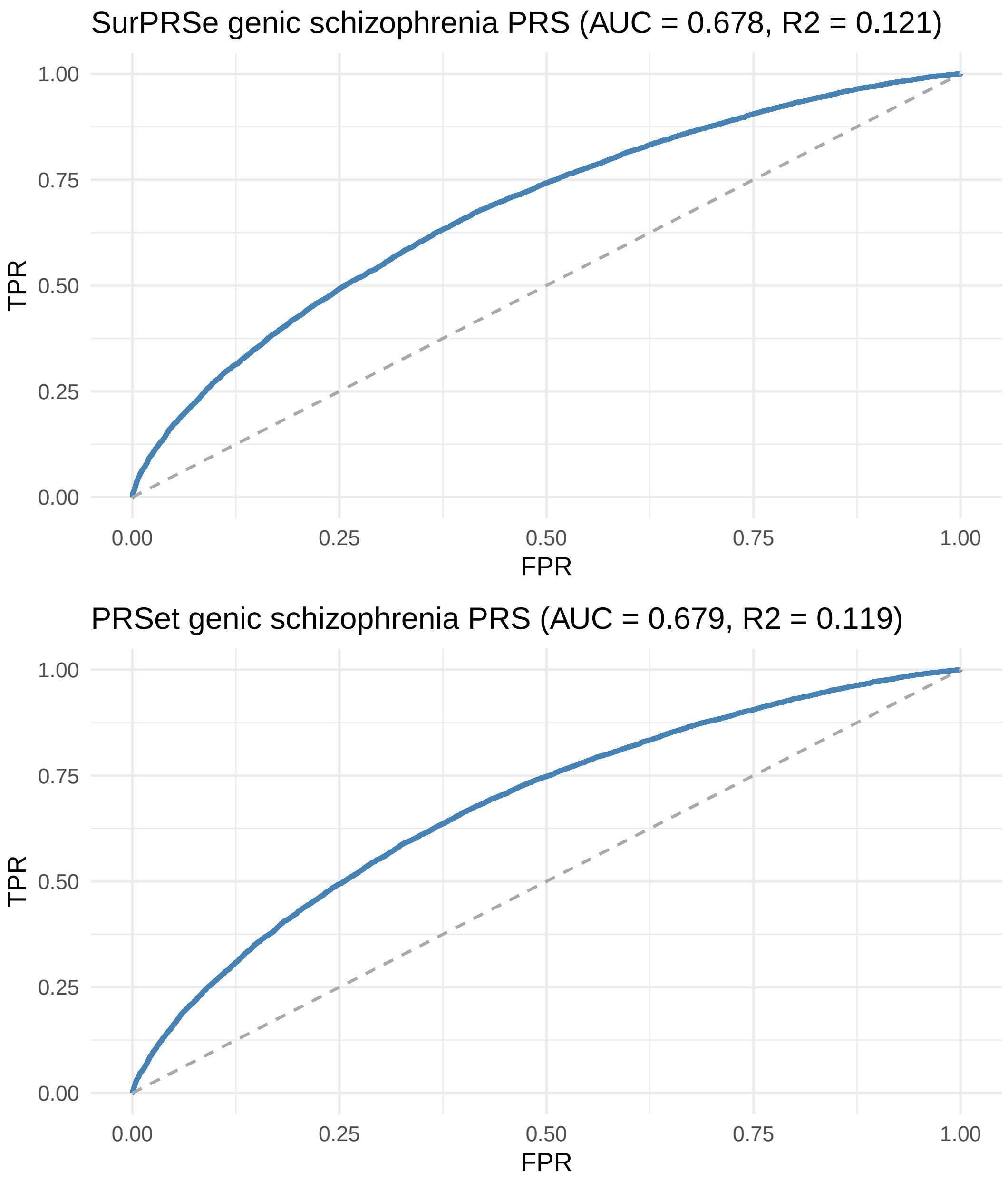
Combined SNP identifiers	CHR_S_test	GD	BP_S_test	A1_S_test	A2_S_test	CHR_P	SNP_P	BP_P	Pvalue_P	Geneset_P	CHR_S_train	BP_S_train	A1_S_train	A2_S_train	INFO_S	BETA_S	SE_S	P_S_train
22:25596029	22	0	25596029	G	A	22	rs4239914	25596029	0.012210	Y	22	25596029	G	A	0.974	-0.0379025	0.0151	0.012210
22:25602708	NA	NA	NA	NA	NA	22	rs2252878	25602708	0.212700	Y	NA	NA	NA	NA	NA	NA	NA	NA
2:209009916	2	0	209009916	T	C	2	rs17538374	209009916	0.741800	Y	2	209009916	T	C	0.991	-0.0053040	0.0162	0.741800
2:209010777	NA	NA	NA	NA	NA	2	rs3214759	209010777	0.122000	Y	2	209010777	G	GC	0.971	-0.0310986	0.0201	0.122000
2:209025923	2	0	209025923	T	G	2	rs2441351	209025923	0.004972	Y	2	209025923	T	G	0.989	0.0473033	0.0169	0.004972
2:209027569	NA	NA	NA	NA	NA	2	rs200926418	209027569	0.496400	Y	NA	NA	NA	NA	NA	NA	NA	NA

Combined SNP identifiers = SNP IDs converted to chromosome:base-pair format within each software to allow easy comparison between SurPRSe and PRSet. CHR_S_test = the chromosome number as it appears within the 1000 genomes data set taken from SurPRSe. GD = Gene distance (the physical distance of the SNP from the nearest gene). BP_S_test = The base-pair location as described within the 1000 genomes data set taken from SurPRSe. A1_S_test = A1 from the 1000 genomes data set as defined by SurPRSe. A2_S_test = A2 from the 1000 genomes data set as defined by SurPRSe. CHR_P = chromosome number from the SNP-list output produced from PRSet. SNP_P = SNP rsID's taken directly from the SNP-list output produced from PRSet. BP_P = The base-pair location as described within SurPRSe. Pvalue_P = P-value of each SNP output from PRSet. Geneset_P = A binary variable (Y = Yes, N = No) describing whether the SNP was included within the PRS of PRSet. CHR_S_train = chromosome position recorded in the iPSYCH GWA study as output from SurPRSe. BP_S_train = Base-pair position recorded in the iPSYCH GWA study as output from SurPRSe. A1_S_train = A1 from the iPSYCH GWA study as defined by SurPRSe. A2_S_train = A2 from the iPSYCH GWA study as defined by SurPRSe. INFO_S = INFO score provided within the SurPRSe output. BETA_S = effect size of each SNP provided within the SurPRSe output. SE_S = Standard error of each SNP provided within the SurPRSe output. P_S_train = P-value of each SNP provided within the SurPRSe output. NA's indicate where data was missing.

While the polygenic risk scores are at least partly correlated between SurPRSe and PRSet, further analysis was required to explain the stratification of the Apical Protein Localisation PRS observed in SurPRSe (See figures 3.18 and 3.17) as compared to the continuous data format of the PRSet PRS. A possible explanation is a different set of variants selected for in SurPRSe as compared to PRSet.

While some discrepancies were explained with the analysis above, it is necessary to check that SurPRSe is producing accurate PRS using predictive outcomes that are known to be true. To examine the differences in correlation between genome-wide and genic PRS for SurPRSe and PRSet, a schizophrenia PRS was created using each tool and the accuracy of each PRS to predict case/control status in schizophrenia patients was examined. CLOZUK was used as the testing data set (11,260 cases, 24,542 controls) and PGC2noCLOZUK was used as the training set (29,415 cases, 40,101 controls). To record the predictive accuracy, both AUC and the nagelkerke R2 were used. With these sample sizes, we would expect the nagelkerke R2 to be between 0.1 and 0.2 and the AUC to be above 0.6 (Ripke et al., 2014).

The SurPRSe schizophrenia genome-wide PRS had an AUC of 0.707 when predicting case/control status in CLOZUK. The nagelkerke R2 from a logistic regression model performed with this PRS, produced a value of 0.159. In comparison, the PRSet schizophrenia genome-wide PRS had an AUC of 0.695 and and Nagelkerke R2 of 0.14. Both SurPRSe and PRSet had comparative prediction to previous schizophrenia PRS with a similar sample size (Ripke et al., 2014). The genic PRS predictive accuracy is shown in figure 3.19. There is a slight loss in predictive accuracy (SurPRSe AUC = 0.678, R2 = 0.121, PRSet AUC = 0.679, R2 = 0.119) but is also still comparative to previous schizophrenia PRS (Ripke et al., 2014).


SurPRSe genic schizophrenia PRS (AUC = 0.678, R2 = 0.121)

PRSet genic schizophrenia PRS (AUC = 0.679, R2 = 0.119)

FIGURE 3.19: Area under the curve plots for the genic schizophrenia PRS. The SurPRSe PRS is the top plot, the PRSet PRS is the bottom plot. TPR = True Positive Rate, FPR = False Positive Rate. The AUC and Nagelkerke r-squared values are within the brackets of each plot title.

To ensure account for any bias, a negative control was produced where the PGC2noCLOZUK training set was replaced with a bodylength GWA study; a trait with no association with schizophrenia; to produce a body length PRS using both SurPRSe and PRSet. The AUC and nagelkerke R2 was taken for each trait, where I would expect these two PRS to have an AUC of 0.5 and an R2 value close to zero.

Both the genic and genome-wide body length PRS produced by SurPRSe and PRSet did not show any predictive ability for schizophrenia case/control status in CLOZUK. The SurPRSe genome-wide PRS had an AUC of 0.503 and an R2 of 6.16e-05. The PRSet genome-wide PRS had an AUC of 0.505 and an R2 of 1.34e-04. Figure 3.20 shows the predictive accuracy of the genic body length PRS for SurPRSe (AUC = 0.503, R2 = 1.41e-05) and PRSet (AUC = 0.504, R2 = 4.58e-05).

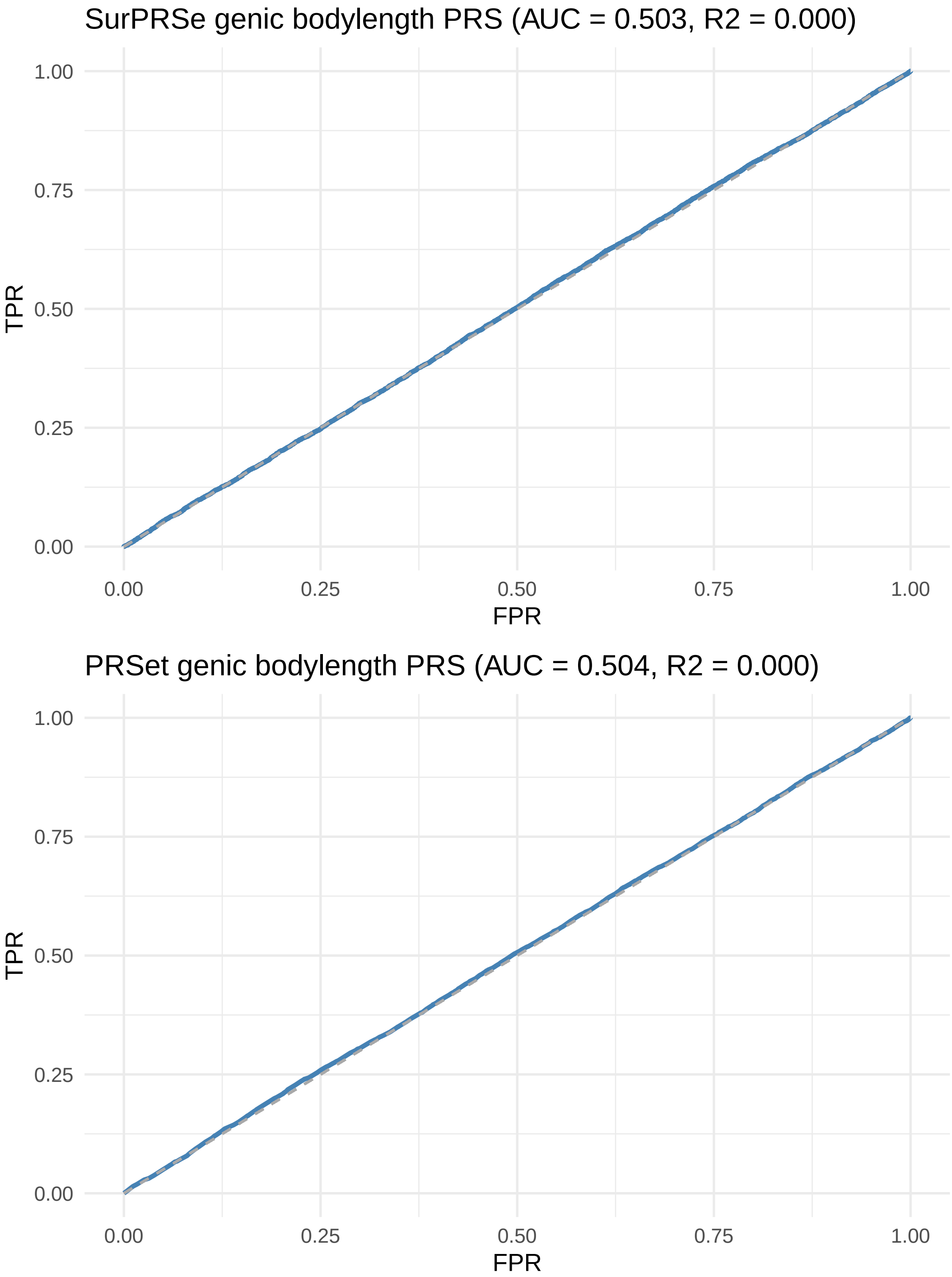


FIGURE 3.20: Area under the curve plots for the genic body length PRS. The SurPRSe PRS is the top plot, the PRSet PRS is the bottom plot. TPR = True Positive Rate, FPR = False Positive Rate. The AUC and Nagelkerke r-squared values are within the brackets of each plot title.

Finally, to test whether clumping was the cause of the discrepencies between the gene-set PRS of SurPRSe and PRSet, a PRS was created for the CCD gene-set where the only SNPs that were included were those that were within the CCD gene-set and have already underwent the clumping procedure as defined within PLINK (p1 = 1, p2 = 2, R2 = 0.1). The correlation between the SurPRSe PRS and the PRSet PRS was recorded and can be observed in figure 3.21.

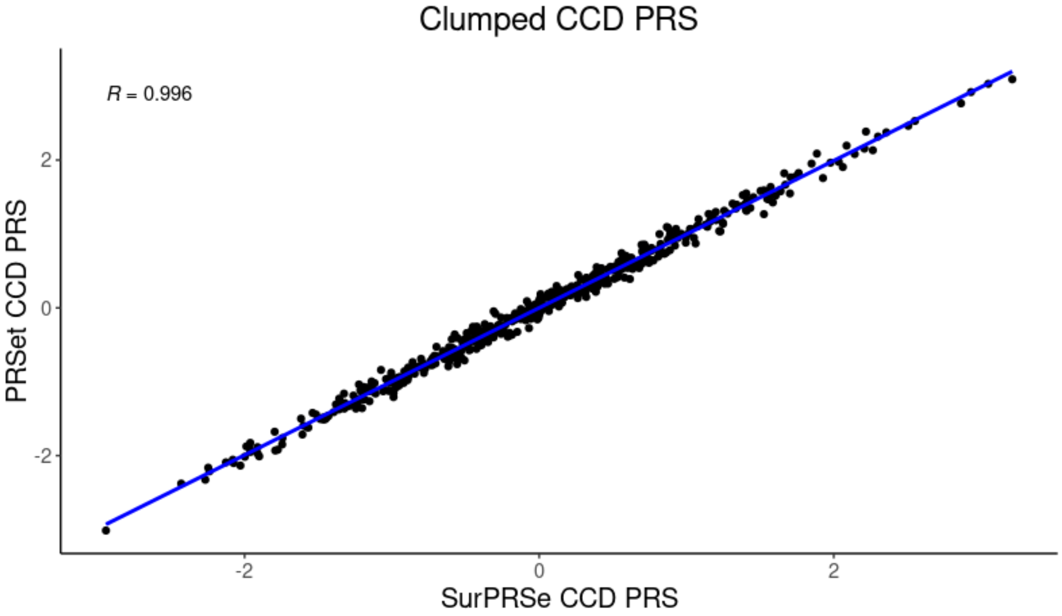


FIGURE 3.21: Comparison of the ckumped 'Cardiac Chamber Development' gene set PRS created by SurPRSe and PRSet. 'CCD' = 'Cardiac Chamber Development'. Within the 108 genes of the CCD, SNPs were excluded using PLINKs clumping procedure. The PRS were standardised by subtracting the mean from the individual score and dividing by the standard deviation.

The correlation between both the SurPRSe PRS and the PRSet PRS was almost 1. This indicates that the discrepencies between the two tools is likely to lie within the different clumping procedures and/or in their QC procedures. With respect to QC, it may be that PRSet prefers power over noise in the production of their gene-set PRS due to the increased number of SNP observed as compared to SurPRSe in all PRS (the genic, genome-wide and gene-set), but very similar predictive ability in the genic and genomewide PRS despite more available SNPs.

With respect to clumping, although theoretically there may not be a noticeable difference between PRSet and SurPRSe, there may be a computational difference. SurPRSe selects the SNPs within the gene-set PRS specified, and then applies PLINKs clumping procedure on these selection of SNPs. This is performed separately for each gene-set. Since PLINK uses a sliding window to identify index SNPs, clumping is not performed across genes or gene-sets. PRSet uses a "capture the flag" system, which, instead of removing SNPs that are not within genic regions, it assigns a binary flag (either 1 or 0) to state whether any one SNP is located within each gene-set. This speeds up the clumping procedure as each SNP is submitted only once for all input gene-sets, while in SurPRSe, the processing time and computational resources required is correlated to the number of gene-sets that are input. As no SNPs are removed in the PRSet procedure, it may be that some SNP are erroneously flagged to be clumped, but further investigation into the accuracy of PRSet is beyond the remit of this thesis.

Reasons for SNP discrepancies:

22:25602708: INFO score of 0.883, PRSet did not remove despite declaring a threshold of 0.9 when running PRSet.

2:209010777: Had an A2 recorded as 'GC'. As multiple bases were detected within the A2 column, it was removed from SurPRSe.

2:209027569: INFO score of 0.893, PRSet did not remove despite despite declaring a threshold of 0.9 when running PRSet.

All other SNPs were included within both SurPRSe and PRSet. In total, three SNPs contributed to the SurPRSe PRS (22:25596029, 2:209009916,2:209025923) while six contributed to the PRSet PRS.

The SNP discrepancies between SurPRSe and PRSet within the Cardiac Chamber Development gene-set PRS was recorded. In total there were 674 SNPs. 223 SNPs were found to be in common between SurPRSe and PRSet.

200 SNPs were excluded/included due to differences in the clumping procedure (for the full list, see Supplementary Table A.1. For example, rs56261301 is in LD with rs11250569. rs56261301 was included in both SurPRSe and PRSet until the clumping procedure. rs56261301 was included post-clumping within SurPRSe but not in PRSet. It has a p-value of 0.48 and is the head of a small clump containing three SNPs.

One SNP, rs28620303, contained an A1 (the effect allele of rs28620303) and an A2 (The non-effect allele of rs28620303) of base C and base G respectively. This should have been removed as it is an ambiguous SNP. It is not possible to pair-up the alleles with complementary base-pairs (eg A/T or C/G SNPs) across both the target and testing data sets because, if the genotyping chips and/or the chromosome strand of the SNP is unknown, it is impossible to discern if the SNP is referring to the same allele or not. Allele frequencies could be used to infer which alleles are on the same strand, but in the circumstances where SNPs have an MAF close to 50% or when the testing and target data are from different populations, it may not be an accurate estimate. This SNP was removed from SurPRSe but retained in PRSet.

PRSet does not have a P1 and a P2 clumping argument, PLINK (software which makes up a component of SurPRSe does. P1 defines the p-value threshold for index SNPs, of which the 'clumps' of SNPs within a certain physical distance are tested for LD with this index SNP. P2 indicates the p-value threshold for the clumped SNPs, so that any SNP within the 'clump' that is above a specified p-value, is removed. This may explain the remaining 44 SNPs which were included within PRSet but excluded within SurPRSe.

3.6 Discussion

I have shown that there are large differences between SurPRSe and PRSet with the production of gene-set PRS. Whole genome PRS between both pieces of software appears to produce similar results and on further investigation, it appears that the reason for this discrepancy is SNP selection for both pieces of software. SurPRSe is more stringent than PRSet.

Wiki pages for both pieces of software exist to enable to user to produce at least one example gene-set PRS before performing their own analyses. PRSet was found to be faster than SurPRSe. A novel shiny application was created called PRSAVE, which was designed to visualise gene-set PRS.

It was expected that PRSet was more scalable and efficient than SurPRSe. PRSet is written predominantly in C++, a language which uses memory and processors much more efficiently than R and BASH, the languages of SurPRSe. However, one advantage of SurPRSe is that R is a much simpler language to understand within the neuropsychiatric field as the programming language is predominantly used. Issues that arise in producing a gene-set PRS can therefore be understood and fixed faster.

There is a strong correlation between the genome-wide PRS of SurPRSe and PRSet. This indicates that despite being written in different computational languages and likely to have different architectures, both bioinformatics tools follows a similar methodology. There was an absence of correlation between the gene-centric and the 20 GO gene-set PRS. The input gene ID's for the 'Apical Protein Localisation' and the 'Cardiac Chamber Development' gene-set was then adjusted for SurPRSe and PRSet, so that the input was identical. The correlation between SurPRSe and PRSet PRS outputs then increased from approximately 0, to approximately 0.5. This indicates that there is an issue in the derivation of the SNPs within the gene-sets before the PRS is produced.

PRSet uses Ensembl ID's and SurPRSe uses NCBI ID's. NCBI and Ensembl ids have been shown to contain different definitions for the chromosomal locations of genes. Therefore there is an increased likelihood for differing pools of SNPs used for the gene-set PRS produced by SurPRSe and PRSet. Gene-set PRS are only found to be correlated if the chromosomal locations are identical between SurPRSe and PRSet. The absence of correlation between the 'Cardiac Chamber Development' PRS and the 'Apical Protein Localisation' PRS using an identical chromosomal location input supports this finding.

The 'Apical Protein Localisation' PRS between SurPRSe and PRSet also display that SurPRSe appears to be more conservative when assigning a PRS to each individual. The increased variation between PRS in the PRSet over SurPRSe indicates that PRSet is using more SNPs than SurPRSe to calculate a PRS despite similar QC procedures used for each bioinformatics tool. Further investigation into the source code for each tool would be required in order to investigate the differences between the PRS but there was no remit within this project to do it here.

PRSAVE is a shiny application which will allow scientists to visualise gene-set PRS results without requiring knowledge of any programming language. With the use of a conventional PRS analysis, the ratio of PRS per trait is 1:1, and most tests for association and/or prediction of the PRS with the trait can be summarised within a single plot. For gene-set PRS, the ratio of PRS per trait is n:1, where n is the number of gene-sets tested. With the increase in data, the ability to interpret the results of any association and/or prediction of every PRS with the trait, within the same plot becomes increasingly difficult. Applications like PRSAVE, allows for the user to instantly change the parameters for the plot to reduce/increase complexity, with the aim of increasing the ability to interpret the results. If gene-set PRS analysis is to be used within either a clinical or research environment, applications like PRSAVE become increasingly beneficial to promote understanding across scientific disciplines and save valuable time perfecting the usage of gene-set PRS.

4 Application of SurPRSe to subcortical brain volumes in schizophrenia

Introduction

Schizophrenia is highly heritable and contains a common genetic component that explains one third of the total risk to the devastating psychiatric disorder (Ripke et al., 2014). Schizophrenia has been confirmed to have a polygenic nature. In the latest **G**enome **W**ide **A**ssociation (GWA) study by Pardiñas et al. (2018), 145 independent risk loci were found to be associated to schizophrenia. Our understanding of the genetic factors related to schizophrenia have increased exponentially in recent years, but despite this, there has been little increase in our understanding of the neurobiology of schizophrenia.

It has long been assumed that the impact of the genetic risk to schizophrenia affects the individuals' brain anatomy and function. The identification of common risk alleles associated to schizophrenia should therefore open up new approaches to explore the neuroanatomical basis of schizophrenia. Despite this, the results from the neuroimaging genetics field attempting to connect the common allele risk for schizophrenia to anatomical structures in the brain have been unconvincing.

Subcortical brain volumes have been shown to have a heritability estimate of 44% to 88% depending on the brain region analysed (Braber et al., 2013; Satizabal et al., 2017). It has previously been shown that there are differences in subcortical brain volumes between healthy controls and schizophrenia patients (Erp et al., 2016; Okada et al., 2016). Recent large studies focusing on **P**olygenic **R**isk **S**cores (PRSs) have shown no association of the PRS to the volume of subcortical brain regions (Franke et al., 2016; Reus et al., 2017). The lack of association of common schizophrenia risk to subcortical brain region volumes indicates that the differences in subcortical brain structure observed in clinical samples may be consequences, rather than intermediate phenotypes of the disorder (Caseras et al., 2015). However, recent work by Warland *et al.* (2019) has suggested that there is an association between the rare genetics of schizophrenia with subcortical volumes, by finding associations between the schizophrenia **C**opy **N**umber **V**ariant (CNV) carriers and the sizes of the right thalamus, right hippocampus and the right accumbens for Biobank participants with no history of severe neuropsychiatric disorders (n = 9,112).

Due to the polygenic nature of the liability in schizophrenia, previous analyses have increased the number of samples to identify more variants associated with schizophrenia. Success has been seen between PGC1 (5,001 cases, 6,243 controls) and PGC2 (36,989 cases and 113,075 controls) where the variation of common genetic risk explained in case-control status increased from approximately 5-7% to around 15-20% respectively.

In this study, I aim to apply the PRS bioinformatics workflow **Su**pe**r**computing (with) **P**olygenic **R**isk **S**core **e**valuation (SurPRSe) in order to test whether the absence of genetic correlations between schizophrenia and subcortical brain volumes were due to the lack of power using previous schizophrenia training samples. I increase the quality of information for schizophrenia using three Psychiatric Genomics Consortium (PGC) genome-wide association (GWA) studies (PGC1 (Ripke et al., 2011), PGC1+Sweden (Ripke et al., 2013), PGC2 (Ripke et al., 2014)) and the largest schizophrenia GWA study at the time of writing (CLOZUK meta analysis) (Pardiñas et al., 2018) as the training data sets. Subcortical brain volume sizes were obtained from UK Biobank (Collins, 2012), and it was also used as the testing set to maximise the power. I do not expect there to be any significant associations of any genome-wide schizophrenia PRS with any subcortical brain volume. However, if the sample size is the main limiting factor in discovering a genetic relationship between schizophrenia PRS and subcortical brain volumes, the variation explained by the genome-wide schizophrenia PRS should increase as the sample size contained within each GWA increase as well.

There is evidence to suggest that schizophrenia patients display brain morphological abnormalities, but there is heterogeneity in the sizes of the same brain regions across different studies (Erp et al., 2016; Haijma et al., 2013). No genetic correlation between schizophrenia and any subcortical brain region has been found (Erp et al., 2016). In addition, any reports of an association of schizophrenia PRS to subcortical brain regions have been contradicted by another study (Merwe et al., 2019). It is possible, that the absence of association is caused by the heterogeneity in the effect of schizophrenia alleles across the brain. It may be that only a subset of schizophrenia risk alleles correspond to subcortical brain volume sizes and/or the brain volumes may not relate to the same subset of schizophrenia alleles (Grama et al., 2020). Recent studies have attempted to identify 'core gene-sets' that make a larger contribution to SZ risk as compared to the rest of the genome (Rammos et al., 2019), or to limit the PRS to genes up/down regulated by MIR137 (Hill et al., 2014) in order to predict brain anatomy (Cosgrove et al., 2018) and functional connectivity (Liu et al., 2020). I aim to use SurPRSe to investigate whether there is an association of gene-set PRS and a genomic PRS (A PRS where all **S**ingle **N**ucleotide **P**olymorphisms (SNPs) must be within genomic regions) which are enriched for common schizophrenia variants, with subcortical brain volumes within healthy controls (Pardiñas et al., 2018).

4.1 Materials and Methods

4.1.1 Samples

UK Biobank

Genotype data curation from UK Biobank has been previously described by Hagenaars et al., 2016a and within Chapter 2. Genotyping was used using two differemnt arrays. The Affymetrix UK BiLEVE Axiom array (807,411 probes) on an initial 50,000 participants, and the Affymetrix UK Biobank Axiom® array (820,967 probes) for the remaining participants. The two arrays are extremely similar (with over 95% common content), but there was a mixture of arrays used for the samples of which imaging data was available.

Imaging data from the first release of UK Biobank consisted of an initial 4,446 subjects (2,342 males/2,104 females; mean age ± s.d. = 55.52 ± 7.62 years; range = 40-70 years) (Reus et al., 2017). In 2015, UK Biobank extended the number of individuals for brain scanning to 100,000 individuals by 2023. At the time of this analysis, the imaging data for a further 13,706 individuals had been released to make a total of 18,152 individuals.

Images of the brain for each individual was taken using a Siemens Skyra 3T running VD13A SP4 (as of October 2015), with a standard Siemens 32-channel RF receive head coil (`https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/bmri_V4_23092014.pdf`). The images were T1 weighted. T1 weighted imaging is a structural brain imaging technique aimed to produce high-resolution depiction of brain anatomy. There is a strong contrast between white and grey brain matter, reflecting the differences in the interaction of water with the surrounding tissues. It is primarily related to the calculation of brain structure volumes.

T1-weighted subcortical brain region images were separated into the left and right hemispheres (Thalamus, Caudate Nucleus, Putamen, Pallidum, Hippocampus, Amygdala, the Nucleus Accumbens and the Lateral Ventricles). The volumes for these hemisphere were calculated using FreeSurfer v5.3 (`https://surfer.nmr.mgh.harvard.edu`). Bilateral subcortical volumes were obtained by averaging the volume of left and right subcortical structures. Participants were excluded if no genetic information was provided alongside volumetric phenotypes and vice-versa.

Schizophrenia GWA Studies

PGC1 (Ripke et al., 2011) and PGC1+sweden (Ripke et al., 2013) GWA study summary statistics were used as training sets in the polygenic risk score analysis and can be downloaded from the PGC website (`http://www.med.unc.edu/pgc/downloads/`). PGC2noCLOZUK referred to here is the re-analysis of the latest PGC dataset sample (Ripke et al., 2014) with the CLOZUK schizophrenia GWAS samples (6,040 cases and 5,719 controls) removed as described previously (Pardiñas et al., 2018). The final PGC2noCLOZUK GWAS summary statistics dataset (29,415 cases and 40,101 controls) was used as a training set in the polygenic risk score analysis. The CLOZUK meta-analysis (40,675 cases, 64,643 controls) is the combined meta-analysis of the CLOZUK dataset (11,260 cases, 24,542 controls) and PGC2noCLOZUK with duplicate PGC2 and CLOZUK samples removed respectively (Pardiñas et al., 2018). Further information about all the data sets described above can be found within Chapter 2.

4.1.2 Gene sets

Seven gene-sets were used for the gene-set PRS. The Mouse Genome Informatics database (Blake et al., 2003) accounts for three gene-sets; abnormal behaviour (MP:0004924; 717,522 SNPs spanning over 2037 genes included in the PRS), abnormal long term potentiation (MP:0002207; 68,686 SNPs from 157 genes included in the PRS) and abnormal nervous system electrophysiology (MP:0002272; 106,641 SNPs included in 213 genes considered in the PRS). These three sets relate to behavioural and neurophysiological correlates of learning. The other four sets composed of: targets of the fragile X mental retardation protein (FMRP targets; 403,723 SNPs from 839 genes considered for the PRS (Darnell et al., 2011)), the 5-HT2C receptor complex (5HT-2C channels; 4435 SNPs included in 18 genes considered in the PRS (Bécamel et al., 2002)), the voltage-gated calcium channel complexes (CaV2 channels;107,987 SNPs from 207 genes included in the PRS (Swantje et al., 2010)) and loss of function intolerant genes as defined by the Exome Aggregation Consortium using their gene-level constraint metric (pLI $>= 0.9$) (LoF intolerant; 1,152,144 SNPs from 3,191 genes included in the PRS (Lek et al., 2016a)). Altogether, these gene-sets together account for 39% of the SNP-based heritability, a substantial amount for the regions these sets cover across the genome (Pardiñas et al., 2018). Further information about the sources of these gene-sets can be found in (Pardiñas et al., 2018).

4.1.3 Polygenic risk scores

All polygenic risk scores were calculated using SurPRSe. For all analyses, single nucleotide polymorphisms (SNPs) with a low minor allele frequency (MAF < 0.1), low quality (INFO < 0.9 or SE > 5 if INFO score not available), ambiguous alleles or residing in the extended MHC region (MHC = chromosome 6: 24MB – 34MB) were removed.

Before calculating the PRS, SNPs were pruned to account for linkage disequilibrium, removing SNPs within 500kb (–clump-kb) and r2 > 0.1 (–clump-r2) of another associated SNP above a specified association/significance threshold. Different **P** value **t**hreshold (Pt) for the calculation of each PRS were used dependent on which analysis was performed. No PRS was calculated outside of the Pt values of Pt $<$ 1e-06, 1e-04, 0.01, 0.05, 0.1, 0.2, 0.5 or 1.

A Pt of 0.05 was used to calculate the PRS for the analysis which examined the four waves of schizophrenia GWA studies. From the latest schizophrenia GWA study, it was found that a schizophrenia PRS at a Pt of 0.05 captures the most variation in case control status (Ripke et al., 2014). Further PRS at a Pt of 1e-06 and 0.5 can be found in Appendix B, section B.1.

Within the gene-set PRS analysis, all eight thresholds (Pt $<$ 1e-06, 1e-04, 0.01, 0.05, 0.1, 0.2, 0.5, 1) were included when the PRS was tested for an association with the left+right subcortical brain region size. This was because the optimal Pt for each gene-set PRS is unknown *a priori*. The Pt of 0.05 was included as this threshold has been shown to account for the majority of the variation explained for schizophrenia (Ripke et al., 2014). For both left and right hemispheres of each region within the gene-set PRS analysis, three Pt were used (Pt $<$ 1e-06, 0.05, 1).

All polygenic risk scores in UK Biobank samples were corrected for eight population covariates (PC1-8) and the genotyping array and subsequently standardised (mean of zero and a standard deviation of 1) before testing against the relevant volumetric phenotype (Smith et al., 2016).

Whole genome polygenic risk scores

The subcortical brain volumes of the UK Biobank samples were tested for association with schizophrenia PRS using four case/control training datasets. Four whole genome polygenic risk scores were produced at Pt = 0.05, one for each iteration of the schizophrenia GWA studies. Linear regressions were used to analyse whether the PRS were associated with subcortical brain regions. In each regression, sex, age and intercranial volume were included as covariates. No selection of SNPs in terms of physical location was used when comparing each schizophrenia trained polygenic risk score. The standardised regression coefficient from each linear regression and its confidence interval was compared across all four schizophrenia trained PRS for each subcortical brain region hemisphere for the selected p-value significance thresholds for the SNPs in the training set as previously mentioned.

Gene-set polygenic risk scores

For each gene-set, SNPs were limited to the gene boundaries within the set and a polygenic risk score was conducted in the same manner as for the aforementioned polygenic risk score method training only on the CLOZUK meta-analysis. The standardised score (mean of 0 and a standard deviation of 1) was then tested for association with the subcortical brain region volumes. A genic-wide PRS was created whereby the whole genome polygenic risk score was limited to the SNPs within genic boundaries as described by NCBI build 37.2 obtained from the annotation software MAGMA (`https://ctg.cncr.nl/software/magma`; Accessed 10/11/2017; (Leeuw et al., 2015)). The r-squared for the whole genome polygenic risk scores and the genic-wide polygenic risk scores were compared to the gene-set polygenic risk scores (observed which r-squared value was higher) to assess whether the gene-set PRS captured more variation than the genome-wide and the genic-wide PRS.

All PRS p-values were corrected for multiple comparisons (specified as FDR p in text) using the Benjamini-Hochberg **F**alse **D**iscovery **R**ate (FDR) procedure (Benjamini and Hochberg, 1995) at an α of 0.05. FDR was applied separately twice; once to the analysis which involved the waves of the different GWA studies and again to the analysis involving the gene-set PRS. For the analysis involving the waves of the different GWA studies, the FDR was applied to all p-values, across all four GWA studies simultaneously.

4.2 Results

4.2.1 Whole genome polygenic risk scores

I examined the relationship between schizophrenia genetic risk and subcortical brain volumes in healthy individuals. We investigated the effect of increasing the training data set sample size and the testing data set sample size, thereby increasing the power of the analysis.

There was a nominally significant negative association found between the CLOZUK meta-analysis genome-wide PRS and the caudate nucleus at a Pt of 0.5 (left+right and left hemispheres) and 1 (left+right, left and right hemispheres).

A nominally significant negative association was also found between the CLOZUK meta-analysis genome-wide PRS and the pallidum (left+right) at a Pt of 0.5 and 1.

No result passed multiple testing correction when all brain regions, training sets and Pt thresholds were included.

There were no significant associations observed between schizophrenia genetic risk and all other subcortical brain volumes. This observation was consistent irrespective of the training set that was used to inform the PRS. All standardised coefficients were within 0.1 units from the null and no observable trends in effect sizes were observed within subcortical brain volumes, across hemispheres or across p value significance thresholds (For Pt at 0.05: Figure 4.1, For Pt across all three thresholds (1e-06, 0.05, 0.5) see B.1 and B.2.

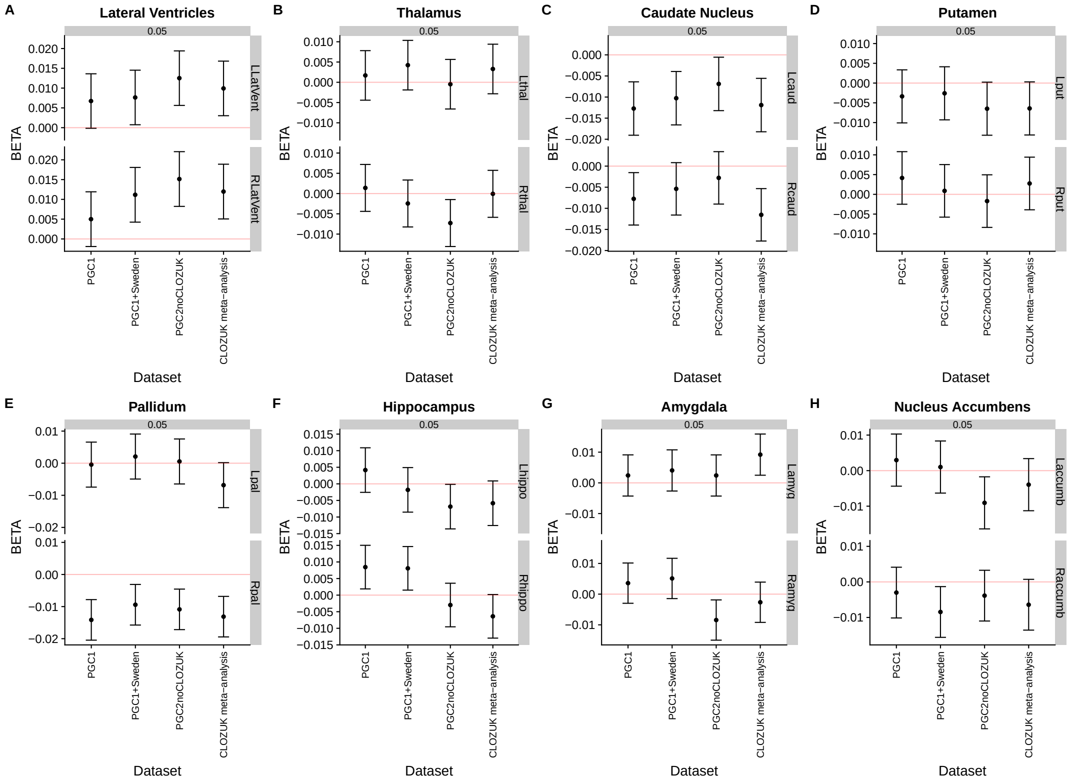


FIGURE 4.1: **Comparisons of associations of polygenic risk scores to subcortical brain volumes in UK Biobank samples**. **A-H** Titles of each plot indicate the subcortical brain region. All PRS have been calculated at a Pt of 0.05, as indicated at the top of the figure. The top section of each plot is the left hemisphere of the specified brain region (indicated by the prefix 'L') and the bottom section is the right hemisphere of the specified brain region (indicated by the prefix 'R'). The x-axis indicates the training data-set used for the PRS. The GWA studies increase in power from left to right. The BETA title on the x-axis indicates the standardised coefficient of the linear regression. The red line indicates a null value across each plot. Error bars displayed here are standard error bars.

There were nominally significant results within the left and right Caudate Nucleus and the right Pallidum when the CLOZUK meta-analysis was used as the training set, but these did not pass multiple testing correction when all brain regions, training sets and Pt thresholds were included.

When the left and right structures were averaged, all standardised coefficients were within 0.1 units from the null and no observable trends in association were observed across p value significance thresholds for any brain region (Figure 4.2).

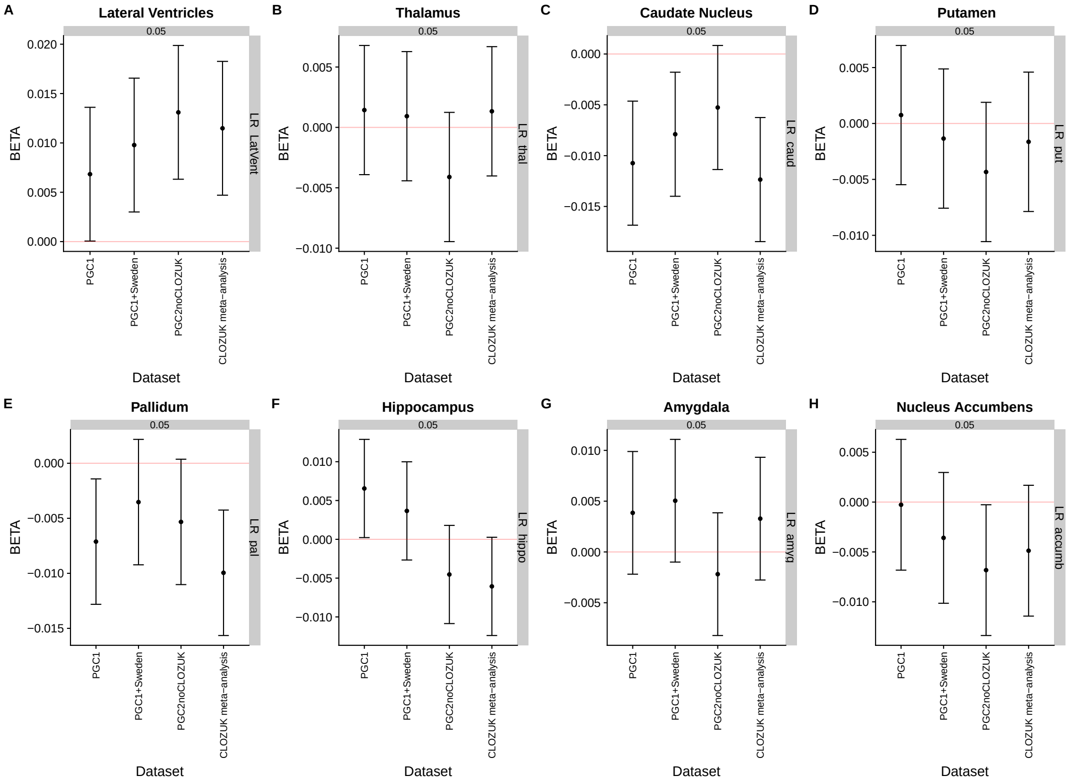


FIGURE 4.2: **Comparisons of associations of polygenic risk scores to subcortical brain volumes in UK Biobank samples**. **A-H** Titles of each plot indicate the subcortical brain region. All PRS have been calculated at a Pt of 0.05, as indicated at the top of each plot. The 'LR' notation on the right side of the y-axis indicates that the brain region has been averaged between the left and right hemisphere. The x-axis indicates the training dataset used for the PRS. At each Pt the GWA studies increase in power from left to right. The BETA title on the x-axis indicates the standardised coefficient of the linear regression. The red line indicates a Null value across each plot. Error bars displayed here are standard error bars.

4.2.2 Gene-set polygenic risk scores

I investigated whether there were any significant associations of the genome-wide and genic PRS with any subcortical brain region volume. Then, I investigated whether using previous significantly associated schizophrenia gene-sets to create PRSs displayed an association between common schizophrenia genetic risk and subcortical brain volumes.

As observed in the previous analysis, the genome-wide PRS was negatively significantly associated with the Left Caudate Nucleus, the right caudate nucleus, the combined left+right caudate nucleus and the combined left+right pallidum brain volume at a Pt of 1 (Figure 4.4 and Figure 4.5). A negative association for the Left+right pallidum and the left caudate nucleus and the left+right caudate nucleus was also found at a Pt of 0.5. At the same Pt, negative associations of the genic-wide PRS was also found with the right caudate nucleus and the left+right caudate nucleus.

Secondly, I investigated whether any gene-set PRS captured more variation in the subcortical brain region size over and above the genome-wide PRS.

There was a significant association found for the left hippocampus with the 'abnormal ltp' gene-set PRS at a Pt of 1e-06 and for the left caudate nucleus with the 'abnormal nervous system electrophysiology' gene-set PRS at a Pt of 1e-06. No significant associations for either gene-set PRS was found in the combined or opposite hemisphere, or across other Pt. No other gene-set PRS was significantly associated with any other subcortical brain region at any Pt.

Thirdly, I investigated whether any gene-set PRS captured variation in the subcortical brain volume above and beyond the genome-wide and/or the genic-wide PRS.

Within the hippocampus, the 'abnormal ltp' gene-set captured more variation than both the genic-wide and the genome-wide PRS. The same result was seen in both right and left hemispheres of the hippocampus (Figure 4.3). 'Abnormal ltp' passed FDR correction at a Pt $<$ 1e-06 within the left hemisphere of the hippocampus. The signal for 'abnormal ltp' was maintained across all Pt for the hippocampus.

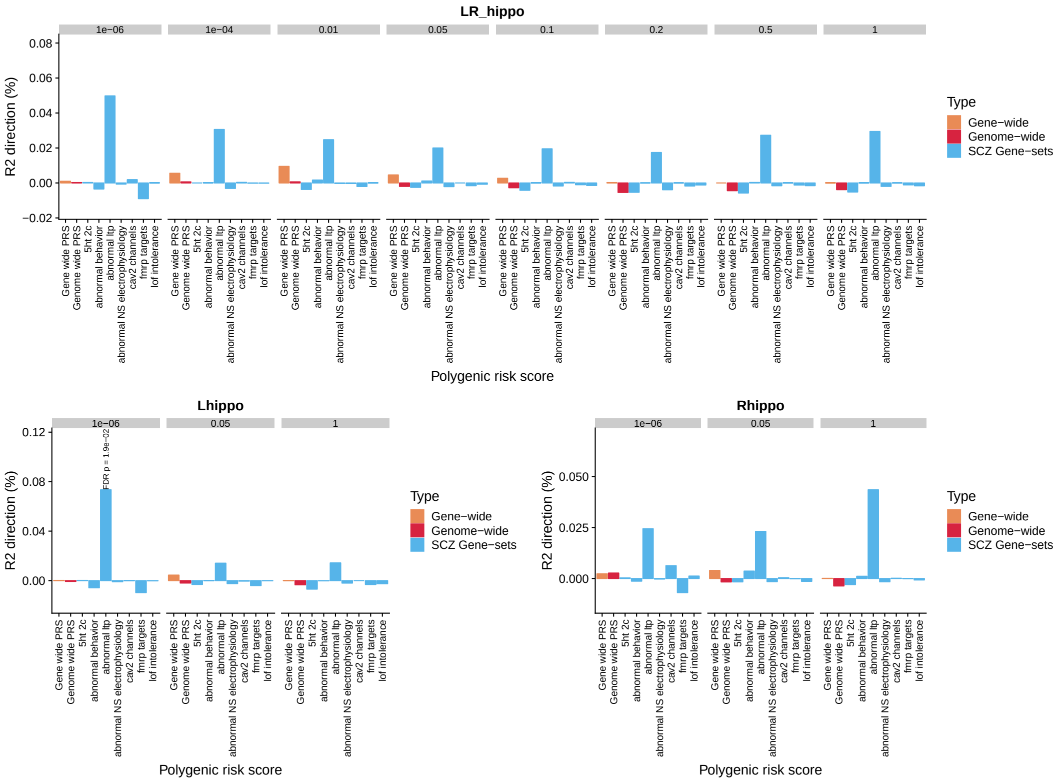


FIGURE 4.3: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on hippocampal brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the hippocampus. The top of each plot displays the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction for the association with the hippocampus subcortical brain region specified in the title of the plot.

Within the caudate nucleus, apart from 'abnormal nervous system electrophysiology' at a Pt of 1e-06, no gene-set PRS captured more information over the genic-wide or the genome-wide PRS.

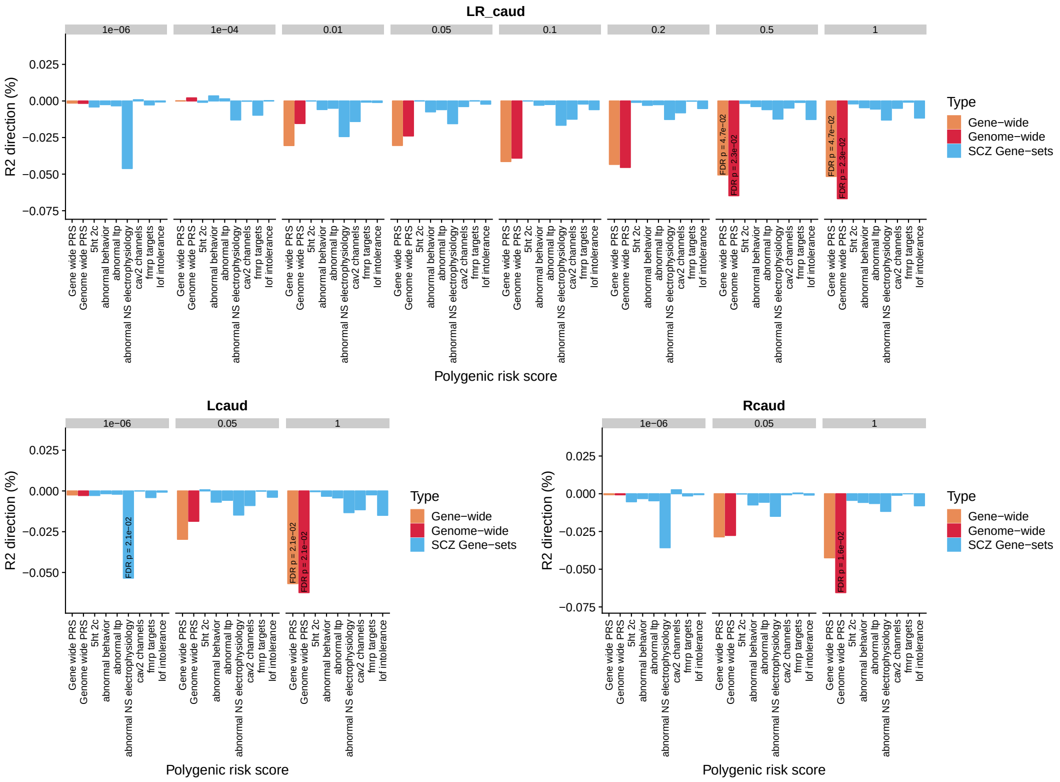


FIGURE 4.4: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on caudate nucleus brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the caudate nucleus. The top of each plot displays the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction for the association with the caudate nucleus subcortical brain region specified in the title of the plot.

within the pallidum, 'Abnormal nervous system electrophysiology' captured more variation over the genome-wide and gene-centric PRS at a Pt of 1e-06. Depending on the gene-set and on which Pt the PRS was derived from, the standardised coefficient of the PRS was found to have instances of both positive and negative directions of effect on the size of the pallidum. The signal for the genome-wide PRS appears at a Pt of 0.01 and is maintained for the remaining Pt. The signal for 'Abnormal nervous system electrophysiology' only appears at a Pt of 1e-06.

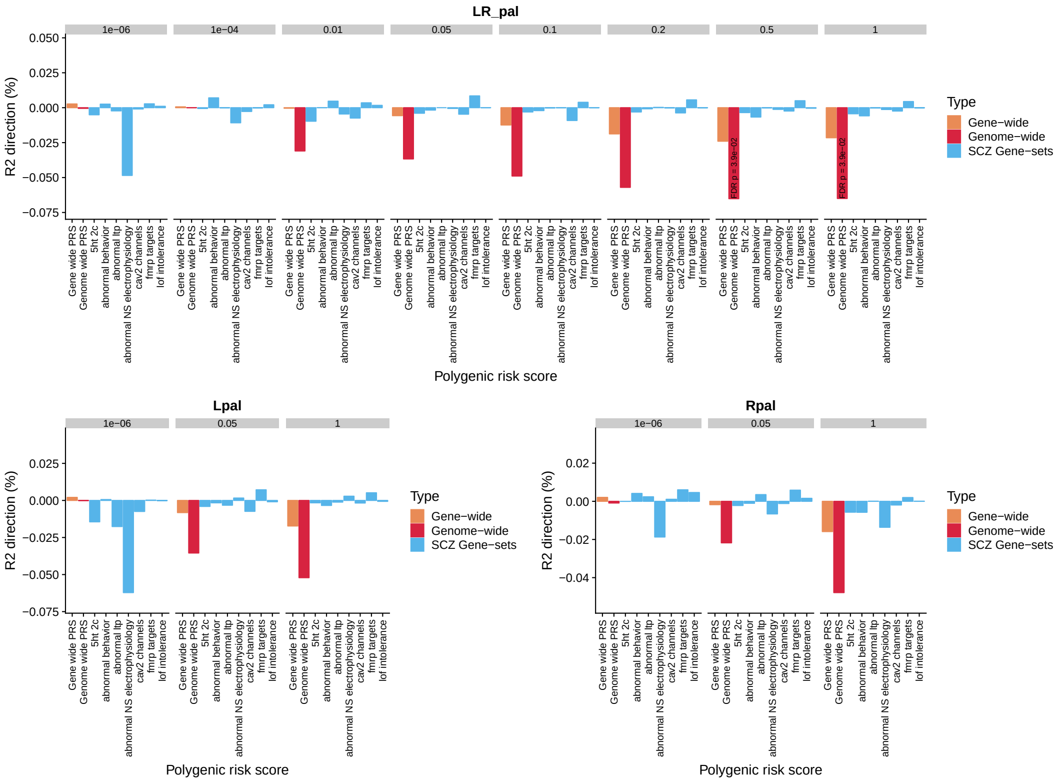


FIGURE 4.5: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on pallidum brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the pallidum. The top of each plot displays the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction for the association with the palidum subcortical brain region specified in the title of the plot.

The amygdala brain region contained the largest dichotomy between the effect sizes of the PRS on the size of the amygdala (Figure 4.6). As the Pt increased, the standardised coefficient of the genome-wide PRS on the size of the amygdala changed direction. Additionally, unlike the majority of the subcortical brain regions, the largest amount of variation explained by the gene-wide PRS on the full amgdala size was seen at a Pt of 0.01. Additionally, the gene-wide PRS captured more variation over all gene-set PRS at a Pt of 1e-04 within the full amygdala brain region and a Pt of 1e-06 within the left hemisphere of the amygdala. No PRS displays a consistent signal across all Pt.

All other plots of each brain region can be found within Appendix B.

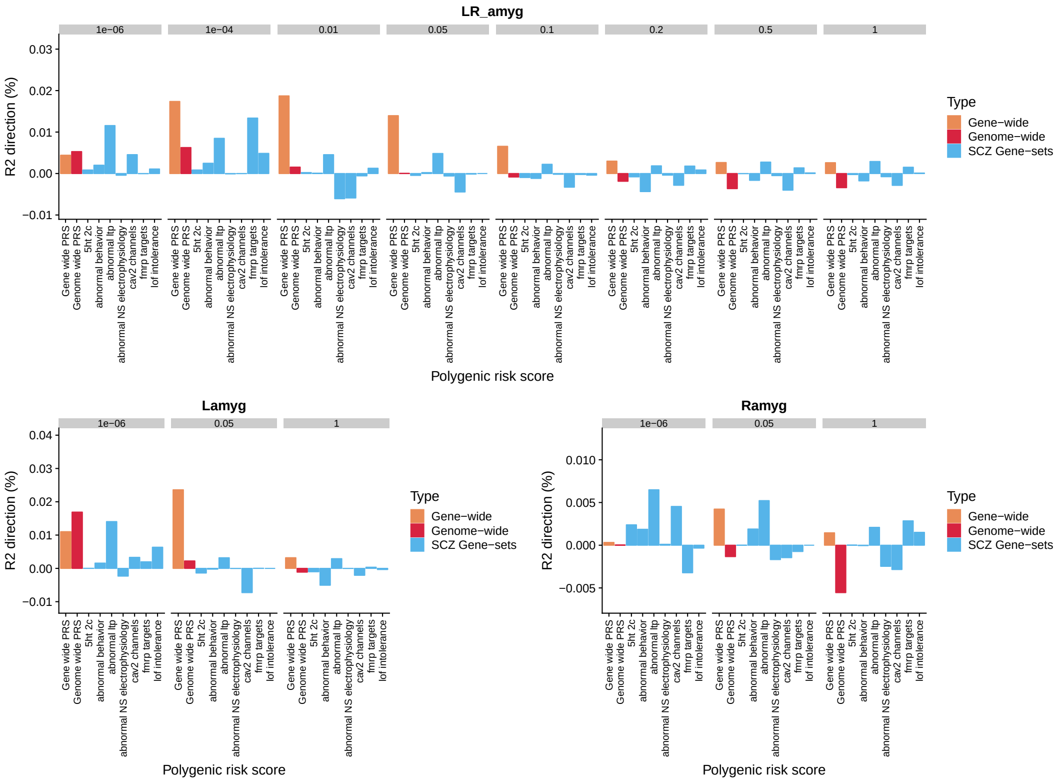


FIGURE 4.6: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on amygdala brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the amygdala. The top of each plot displays the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction for the association with the amygdala subcortical brain region specified in the title of the plot.

4.2.3 Discussion

There was no increase in the association of the schizophrenia PRS to any subcortical brain region volume as the power of the schizophrenia PRS increased. This finding is in-line with previous high-powered studies (similar sample size to schizophrenia GWA studies which have found common genetic variants associated with schizophrenia (Pardiñas et al., 2018) which have used schizophrenia PRSs to test their association to subcortical brain volumes (Merwe et al., 2019). The absence of association as the quality of the schizophrenia information increases on the UK Biobank cohort of around 18,000 individuals suggests that the approach of increasing the power of the PRS will not provide a robust association of a schizophrenia PRS with a subcortical brain volume.

The lack of association found within this study does not however, exclude the possibility that there is an association between the common genetic risk of schizophrenia and subcortical brain region sizes. As brain volumes can be heterogeneous even within that of schizophrenia patients, it is logical to hypothesise that different biological mechanisms of schizophrenia may affect some brain regions more so than others (Brugger and Howes, 2017). The analysis using the schizophrenia gene-set PRS and SurPRSe made it possible to investigate this further.

Overall, the results of the gene-set PRS analysis were inconclusive. In the hippocampus, the 'abnormal ltp' gene-set captured more variation than any other genome-wide, gene-centric or gene-set PRS. However, the only significant PRS was found at a low Pt, where only a fraction of the SNPs located within the gene-set would have been used for the association analysis. Other brain regions including the caudate nucleus and the pallidum did not provide enough evidence to conclude that the genome-wide PRS captured more variation over the gene-set PRS.

It is becoming apparent that a lot of work will be required to disentangle the contribution of common genetic risk to brain anatomy. It may be prudent to continue the investigation into whether there is a common genetic link between schizophrenia and brain anatomy using more focused techniques and experiments. For example, Stauffer et. al. (Stauffer et al., 2021) found an association of a genome-wide PRS to grey and white microstructure (aka at the cellular level) instead of macrostructure which was measured in this study.

It may also be beneficial to investigate the contribution of environmental factors with respect to subcortical brain region sizes in individuals with schizophrenia. For example, it has previously been shown that individiduals who were considered to have a high risk of developing schizophrenia (based on positive schizophrenia cases in familial history) were found to differ significantly from a control population for environmental risk factors including family conflict and stressful life events (Walder et al., 2014). In combination with the observation that exposure to stress may contribute to reductions in hippocampal volume via hypercortisolemia (Lawrie et al., 2008), it is viable that a causal link between stress in schizophrenia patients and reduced hippocampal brain volume in schizophrenia patients may exist.

The results displayed here do not disprove the idea that partitioning the genetic risk schizophrenia will allow for a better understanding of the effect (if any) of the genetics of schizophrenia on brain volume. Gene-set analysis is a complex field in of itself and the inconclusive results here may simply be a reflection that the input gene-sets may need refining. However, it can be argued that by using UK Biobank, a resource of a healthy population, you are removing environmental factors associated with schizophrenia including anti-psychotic treatment and increased cannabis use as factors within this analysis.

SurPRSe appeared to produce gene-set PRSs which were accurate. As observed in Figures 4.3, 4.4, and 4.5 when an FDR p-value is observed for a PRS in one Pt, the direction of effect is the same across all other Pt for that same PRS. In addition, the direction of effect for most gene-set PRS across all other Pt broadly was the same direction of effect for the PRS with a p-value that passed FDR correction. One difference between the gene-set PRS and the genome-wide PRS was the apparent increase in signal at the lower Pt for some gene-sets. There are two occurrences which could explain these results. The first is that the gene-set is particularly large and a subset of SNPs within the gene-set is driving the signal. For future work, it would be useful to identify which SNPs are within the gene-set at the lower thresholds and perform a stringent gene-set enrichment test on these SNPs to examine whether a known biological pathway exists that matches the signal observed. In the second instance, it might be due to the properties of a small gene-set PRS. In these cases, the number of SNPs contributing to the PRS might be extremely small at the lower thresholds. The signal observed might be an anomaly of having insufficient data to describe the PRS, but further work using for example, simulations would be required to see if this was the case.

In conclusion, it appears that in most cases a gene-set PRS does not provide information above and beyond what a genome-wide or a gene-centric PRS would provide. However, with a larger sample size and further investigation, a gene-set PRS may provide insight into the effect of schizophrenia genetics on the hippocampal brain volume.

Furthermore, A summary on the current state of brain imaging and schizophrenia can be found based on the reaction to the paper by Alnaes *et al.* (2019). Essentially, the authors attempted to show whether brain structure variability between individuals with schizophrenia and a healthy cohort (UK Biobank) was caused by the schizophrenia PRS (Alnæs et al., 2019). They concluded that the PRS was not capturing the genetic or environmental factors responsible for this difference. In response, De Peri *et al.* (De Peri and Vita, 2019) stated that the study did not incorporate the reverse causation hypothesis (e.g. inclusion of anti-psychotic medication as a covariate) into their analysis. As a reply, Alnaes and Westlye (2019) stated that there is a lack of harmonised protocols for clinical phenotypes across the samples available to them. They further argued that many environmental factors and "indeed most constituents of life itself" affect the structure of the brain, and so in order to understand the genetic impact of neuropsychiatric traits on brain anatomy, a strategy to catalogue these environmental variables must be created (Alnæs and Westlye, 2019).

NOTE OF WORK

Steluta Grama produced the PRS using SurPRSe. All other work including but not limited to association analysis, plotting, interpretation and presentation was performed by myself.

5 Using polygenic risk score approaches to investigate the common-variant genetic architecture of cognition in schizophrenia

5.1 Introduction

Cognitive impairment is common in schizophrenia, and is a predictor of poor functional outcome (Green, 2006). The cause of cognitive impairment in schizophrenia has been hypothesised to have a genetic component due to observations that cognitive performance is impaired in the relatives of patients with schizophrenia (Seidman et al., 2015). Investigating the genetic component of cognitive impairment in schizophrenia patients has been challenging, as secondary factors including illness-related behaviours (e.g. poor nutrition and substance abuse) and consequences of treatment (e.g. anti-psychotics) have also been shown to affect cognition (Green, Llerena, and Kern, 2015; Keefe et al., 2007a; Keefe et al., 2007b).

Multiple Measurements of cognitive performance (Deary, Johnson, and Houlihan, 2009; Kremen et al., 2013; Davis, Haworth, and Plomin, 2009; Polderman et al., 2015b) and educational attainment (a proxy measurement for cognitive performance) (Krapohl et al., 2014) are highly heritable. Additionally, almost all of these cognitive measurements correlate substantially and positively (Plomin and Deary, 2015). General Intelligence (an index of the co-variance between multiple cognitive tests) was one of the first traits to show that the genetic influence contributed to substantial individual differences of the trait (Deary, Johnson, and Houlihan, 2009). Kremen et. al. (2013) also found through the use of twin studies, that a combined metric of measurements for cognitive impairment (including episodic memory, Executive function, and verbal/language) were found to have a heritability of approximately 40%

to 48%. A recent **G**enome **W**ide **A**ssociation (GWA) study has identified 206 genetic loci associated with general intelligence within the general population (**I**ntelligence **Q**uotient (IQ)) (Savage et al., 2018).

There is evidence that supports the hypothesis of genetic overlap between schizophrenia, IQ and cognitive performance in schizophrenia through twin studies (Lemvigh et al., 2020), GWA studies (Ohi et al., 2018), and **P**olygenic **R**isk **S**cores (PRSs) (Hubbard et al., 2016; Richards et al., 2019).

LD Score **R**egression (LDSC) has been used to measure the genetic variant correlation (r_g) between general cognitive function GWA studies and schizophrenia GWA studies (Ohi et al., 2018). Several studies have reported an r_g of approximately -0.2 between these two traits (Trampush et al., 2017; Sniekers et al., 2017; Lam et al., 2017; Davies et al., 2018; Savage et al., 2018).

PRS has shown an association of cognitive PRS with schizophrenia, and conversely, a schizophrenia PRS has shown an association with various measurements of IQ including IQ, attention, processing speed, working memory, problem solving and social cognition (Hubbard et al., 2016). So far, no study has examined the association of schiophrenia PRS or cognition PRS with a cognition within schizophrenia phenotype.

Lemvigh et al. (2020) performed a twin study that showed cognitive deficits within schizophrenia was heritable, and found that some components of cognitive functions associated with schizophrenia liability were independent of IQ. This suggests both a shared genetic etiology between schizophrenia and cognitive performance within schizophrenia patients. It also shows that although IQ correlates strongly to cognitive measurements(40%), there is a suggestion that some aspects of cognition may be an indication of specific risk factors for schizophrenia, separate from IQ measurements. Discerning where these risk variants are located could provide an indication of which cognitive and/or schizophrenia biological processes are involved, and potentially, the direction of causality (i.e. does the risk for cognitive deficits contribute towards schizophrenia liability or does schizophrenia liability contribute to cognitive deficits).

The largest schizophrenia GWA study to date at the time of writing, the CLOZUK meta-analysis (Pardiñas et al., 2018; 40,675 cases, 64,643 controls; see chapter 2) identified six gene-sets significantly associated to schizophrenia (Targets of FMRP, Abnormal behavior (MP:0004924), $5 - HT_2C$ receptor complex, Abnormal nervous system electrophysiology (MP:0002272), Voltage-gated calcium channel complexes, Abnormal long-term potentiation (MP:0002207) and suggests that the risk for the disorder converges onto physiological, molecular and behavioural pathways (Pardiñas et al., 2018). These gene sets collectively captured 30% of the total **S**ingle

Nucleotide **P**olymorphism (SNP) heritability of schizophrenia; a disproportionately high amount for the proportion of all SNPs included in these gene-sets (Pardiñas et al., 2018). Taking a similar approach, the latest IQ GWA study at the time of writing (IQ3; 269,876 individuals; see 2) identified six gene-sets which were significantly associated with general intelligence in the population(neurogenesis, neuron differentiation, central nervous system neuron differentiation, regulation of nervous system development, positive regulation of nervous system development, and regulation of synapse structure or activity). After conditional analyses on these gene-sets, three gene-sets were found to be independently associated to IQ(regulation of nervous system development, central nervous system neuron differentiation, and regulation of synapse structure or activity) and altogether accounted for the association found in the other three gene-sets (Savage et al., 2018).

A significant gene-set from a gene-set analysis in the context of (for example) schizophrenia, infers that SNPs located within a group of related genes have, on average, more significant association test statistics for schizophrenia than expected either by chance, or than all other SNPs within genes in the remainder of the genome. This is advantageous if applied to PRS analyses because it allows the PRS to convey more information about any one individual's genetic profile. Within the classical model of polygenic disease, the polygenic risk score for any individual conveys the total liability of a trait (for example schizophrenia) to a single value estimate that should lie on a spectrum from low to high genetic risk for schizophrenia. This reduction of information to a single numerical figure loses potentially important information about that individual's genetic profile for schizophrenia. If PRSs are defined across gene sets significantly associated to schizophrenia, it can describe how the risk for schizophrenia (and the subphenotypes of schizophrenia) varies across different biological processes and pathways (Choi et al., 2022).

For example, if one wanted to test the association between the genetic propensity for IQ and a cognitive phenotype within schizophrenia, an IQ GWA study would be used to inform the PRS, and the PRS would be tested for any association with a cognitive phenotype recorded for individuals with schizophrenia. However, within this model, every allele available would be included within the PRS, the majority of which (when defined within the polygenic model) would have varying contributions to the genetic risk of schizophrenia, of IQ and, of cognitive performance in schizophrenia patients. However, some of these alleles may have a larger contribution to biological processes of schizophrenia separate from the biological processes of cognitive performance within schizophrenia and of IQ. These alleles are still included within the analysis and therefore any signal for the potential association of IQ to cognitive performance within schizophrenia patients may be diluted. The interpretation of how schizophrenia influences cognition is limited by the reduction of the genetic signal to a single estimate per individual.

Instead, it may be beneficial to limit the alleles within the IQ PRS to a set of alleles that have been shown to be associated with schizophrenia. The analysis is testing whether alleles that contribute to schizophrenia may also contribute to the cognitive performance within schizophrenia. Additionally, it may be possible to discern which biological processes relevant to schizophrenia are also relevant for cognitive performance within schizophrenia, and which processes have no relevance for cognitive performance within schizophrenia.

5.1.1 Aims

Within this study, I aim to test whether gene-set PRS gives a better insight into the pathogenesis of schizophrenia, specifically, it's relationship to cognition. The direction of causation between cognition and schizophrenia is unknown at the time of writing.

I will use gene-sets significantly associated with schizophrenia, gene-sets significantly associated to IQ, a high-powered schizophrenia GWA study and a high-powered IQ GWA study to inform the PRS. By using the genetic liability to one trait, and gene-sets associated to the other, I aim to provide more insight about the relationship between IQ liability, schizophrenia liability and the liability of the cognitive phenotypes observed within schizophrenia patients. The null hypothesis is that both schizophrenia PRS and IQ PRS will account for the same amount of variability within cognition within schizophrenia patients, and show the same direction of effect.

5.2 Materials and Methods

5.2.1 Samples

CardiffCOGS

Exploration of the cognition within schizophrenia phenotype using PRS requires a target data set of patients diagnosed with schizophrenia where a robust metric of cognition within schizophrenia has been recorded.

CardiffCOGS is a sample of 1,024 UK-based participants diagnosed with schizophrenia, schizoaffective depressed, schizoaffective bipolar or other psychotic disorder

(Pardiñas et al., 2018; Lynham et al., 2018a). These participants were recruited via secondary care NHS mental health services in England and Wales. To determine the reliability of a schizophrenia phenotype, all individuals were interviewed using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al., 1990). Both this interview and available clinical records were reviewed by trained raters to give a consensus lifetime DSM-IV diagnosis (American Psychiatric Association, 2000).

Cognitive Assessment and Outcome

All individuals in CardiffCOGS had a cognitive assessment by trained psychologists using the MCCB Nuechterlein2008. A detailed explanation of the MATRICS **C**ognitive **C**onsensus **B**attery (MCCB) applied to CardiffCOGS is described in Lynham *et al.* (Lynham et al., 2018a). Briefly, seven domains of cognition are measured from ten tasks including Speed of processing, verbal/visual learning and reasoning and problem solving. Z scores for each task were derived using the mean and standard deviation of healthy controls matched for age and sex. Domain and composite scores were calculated following the procedures in the MCCB manual, whereby composite cognitive scores are derived from five or more domain cognitive measures if they are present in each individual.

For this study I selected the MCCB composite score, a measure of generalized cognitive functioning, as our primary outcome (Figure 5.1. This battery was specifically designed to be the accepted diagnostic tool for assessing cognitive change in individuals with schizophrenia (Nuechterlein et al., 2008b).

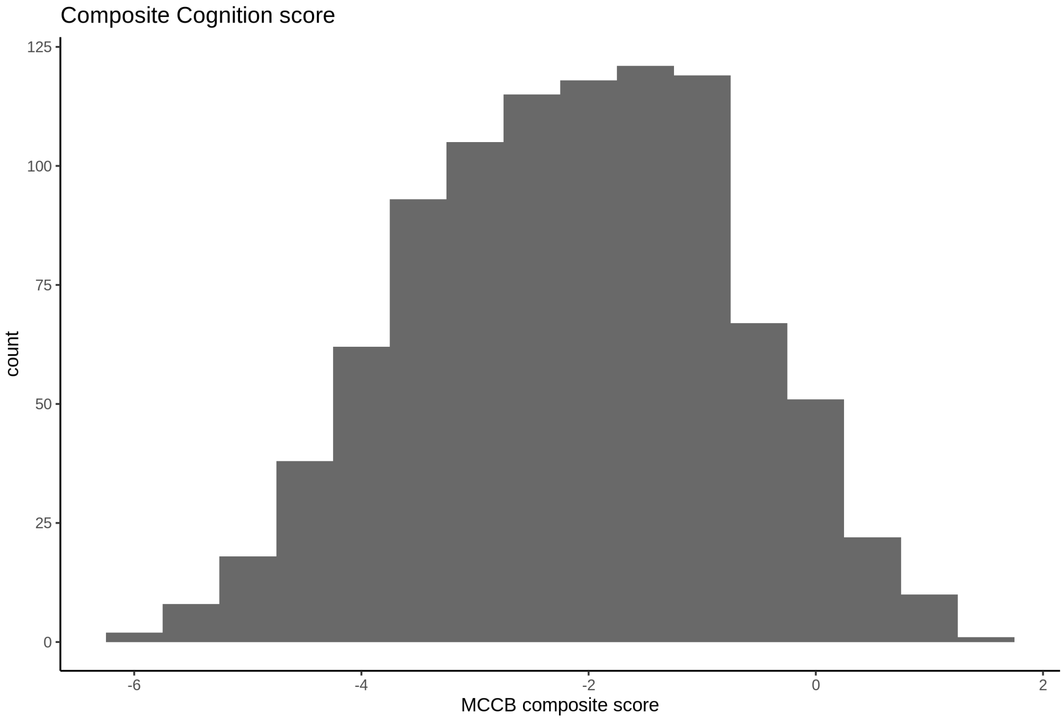


FIGURE 5.1: **Distribution of the MCCB composite score across all schizophrenia patients used within this study**

Further recruitment and genotype information for these individuals has been extensively described elsewhere (Pardiñas et al., 2018; Lynham et al., 2018a).

5.2.2 Schizophrenia Genome Wide Association Study

To investigate the genetic effects of schizophrenia on the phenotype of cognition within schizophrenia patients, a sufficiently powered training set (GWA study) is required.

CardiffCOGS samples were removed from the CLOZUK meta-analysis (40,675 cases and 64,643 controls; Pardinas2018) to create new GWA summary statistics referred to as SCZminusCOGS (39,950 cases and 64,643 controls; see Chapter 2). The fixed-effects procedure in METAL (Willer, Li, and Abecasis, 2010) was used to perform the meta-analysis, with a filter on the INFO score > 0.9.

The CLOZUK meta-analysis was previously shown to confer over 90% of polygenic contribution to the association signal observed within the GWA study (Using LD Score v1.0) (Pardiñas et al., 2018; Bulik-Sullivan et al., 2015). As CLOZUK itself conferred over 80% of the polygenic contribution, SCZminusCOGS should confer between 80-90% of the polygenic contribution to the association signal observed within the SCZminusCOGS.

5.2.3 IQ Genome Wide Association Study

I obtained IQ GWA study results from Savage *et al.* (2018) in the form of summary statistics from the Complex trait genetics lab (IQ3, n = 269,876; `http://ctg.cncr.nl/software/summary_statistics`; see Chapter 2).

5.2.4 Gene-set Analysis

To expand the information available for schizophrenia, IQ and cognition within schizophrenia from the PRS beyond a single estimate per individual, gene-sets significantly associated to schizophrenia and IQ need to be defined.

134 **C**entral **N**ervous **S**ystem (CNS) related gene-sets were taken directly from Pardiñas *et al*. These gene-sets were analysed with the gene-set analysis tool MAGMA (Leeuw et al., 2015) and the SNPs within these gene-sets were collectively found to capture a disproportionate amount of the SNP heritability as compared with all other annotated genes (30% of the total heritability, and 46% of the genic heritability). This colection of genes was found to be enriched for common variation in schizophrenia as compared with all other annotated genes (P=8.57×10–13) (Pardiñas et al., 2018).

A gene-set containing loss of function intolerant genes was included. This gene-set was defined using a gene-level constraint measure (**p**robability of being **L**oss of Function **Intolerant** (pLI) $\leq$ 0.9) provided by the **Ex**ome **A**ggregation **C**onsortium (ExAC) (Lek et al., 2016b). pLI was calculated by analysing the proportion of the observed number of SNPs that were to be rare (**M**inor **A**llele **F**requency (MAF) < 10%) in ExAC with the expected number of SNPs (the expected number of SNPs were quantified using a selection neutral, sequence-context based mutational model (Samocha et al., 2014)). Using MAGMA, loss of function intolerant genes were found to be enriched for common variation in schizophrenia as compared with all annotated genes (P=4.1×10–16) (Pardiñas et al., 2018).

Our testing sample (CardiffCOGS) contained samples from the GWA study used in Pardinas *et al.* (2018), so the gene-set analysis was performed de novo in SCZminusCOGS in order to ensure that the gene-sets which were enriched within the () GWA were also enriched and/or contained the lowest p-values (for association with the common variation in schizophrenia) compared to all available annotated gene-sets.

Enrichment of all 135 gene-sets in SCZminusCOGS was assessed using MAGMA v1.06 (Leeuw et al., 2015). After accounting for **L**inkage **D**isequilibrium (LD), p-values of all SNPs inside genes were combined to create gene-wide p-values. Genes were defined allowing for a window 10kb downstream and 35kb upstream of the gene to capture any signal found in regulatory regions (Pardiñas et al., 2018; Cox and Lee, 2008). Competitive gene-set p-values were then derived from the gene p-values after accounting for LD between genes, gene size and gene-set density. 'Competitive' gene-set analysis hypothesises that the gene-set being analysed is enriched to the same degree to the phenotype as a background gene-set (in this case, the combination of all available annotated gene-sets). Multiple correction testing was performed using the Westfall-Young **F**amily-**W**ise **E**rror **R**ate (FWER) procedure (100,000 re-samplings, alpha threshold = 0.05; Benjamini and Hochberg, 1995)

Six Gene-sets for IQ were taken directly from Savage *et al.* (2018). They were derived by using the gene-set association analysis tool MAGMA to test the association of pre-defined gene-sets to IQ. The source of the gene-sets came from three different groups. For the first group, 7,246 gene-sets representing biological and metabolic pathways were derived from nine separate resources defined within MsigDB (Liberzon et al., 2011). The second group were derived from gene expression values for 53 tissues from GTEx (Ardlie, Deluca, and Segre, 2015) and the third group was derived from cell-type specific gene expression within 24 types of brain cells (Skene et al., 2018).

5.2.5 Derivation of PRS

PRSs need to be created within the gene-sets defined previously. Background PRSs should be defined to observe whether the gene-set PRS confer polygenic signal above and beyond the polygenic signal they are expected to confer, and whether the gene-set PRS confer more information (e.g. direction of effect of the PRS on the phenotype of cognition within schizophrenia patients). As the pool of SNPs for the gene-set PRS can only be within genic regions, a whole genome PRS limited to this pool of SNPs within genic regions should be created as a fair comparison to gene-set PRS.

Another potential source of extra information which can be extracted from gene-set PRS beyond whole genome PRS is to combine all SNPs within all schizophrenia and IQ gene-sets together (for both traits individually and a collective schizophrenia and IQ set). If all these gene-sets are observed to be have the same direction of effect on the phenotype of cognition within schizophrenia patients, it may increase the power of the respective PRS.

All PRSs were performed using **S**upercomputing (with) **P**olygenic **R**isk **S**core evaluation (SurPRSe). The **Q**uality **C**ontrol (QC) stages used the CardiffCOGS genotypes as a reference data set. SNPs with ambiguous alleles, a low minor allele frequency (MAF < 0.1, low quality (**INFO**rmation score (INFO) < 0.9), or mapping to the extended region of linkage disequilibrium surrounding **M**ajor **H**istocompatibility **C**omplex (MHC) region of chromosome 6 (24MB - 35MB) were removed. The MHC region was removed because of the complex LD structure within this region. Significantly associated SNPs within the MHC have not been consistently replicated across different GWA studies, even within the same ethnic group (Mokhtari and Lachman, 2016). SNPs were pruned to account for LD, removing SNPs within 500kb and r2 > 0.1 of another associated SNP with a higher training set p-value. SNPs were also removed if they had a Hardy-Weinberg equilibrium mid-p test p-value below 1e-06. PRS were assigned to CardiffCOGS individuals using the PLINK 1.9 –score command within **Su**pe**r**computing (with) **P**olygenic **R**isk **S**core **e**valuation (SurPRSe) and three **P** value **t**hreshold s (Pts) were used (Pt < 5e-08, 0.05, 1; (2015)). The Pt of 0.05 has been shown to capture the most amount of variation of case-control status for a schizophrenia PRS, while the Pt of 5e-08 and 1 are supplementary thresholds to ensure that the SurPRSe is working correctly, and examine whether the gene-set PRS have different statistical properties to the genome-wide PRS. PRSs were calculated by summing the number of associated alleles for each index SNP, weighted by their coefficient of effect size (beta).

PRSs in CardiffCOGS were corrected for the first five population principal components. Tucker et al (2014) have shown that for GWA studies, the first five PCs are sufficient to account for population stratification. Linear regressions were used to test association between PRS and cognition, indexed by the MCCB. For each regression, covariates for age at the SCAN interview and sex were included in the model; individuals were removed if a cognitive composite score was missing. PRS p-values were corrected for multiple comparisons (**F**alse **D**iscovery **R**ate (FDR) p) using the Benjamini-Hochberg FDR procedure (Benjamini and Hochberg, 1995).

Genome-wide and Gene-centric PRS

Genome-wide and gene-centric PRSs were tested in CardiffCOGS; for the gene-centric PRS, a window 35 **K**ilo-**B**ase (KB) upstream and 10 KB downstream of each gene was included, enabling some of the association signal from SNPs within regulatory regions adjacent to the gene to be captured (Banaschewski et al., 2015; Maston, Evans, and Green, 2006). These regions were defined and SNPs outside of the genic regions were removed before LD pruning procedures.

Gene-set polygenic risk scores

For each gene-set, SNPs were limited to the gene boundaries including the above flanking sequences. PRSs were calculated as above.

I also defined schizophrenia and intelligence collated sets by, respectively, combining all genes contained within the associated schizophrenia and the IQ gene-sets. These collated sets were also tested for association with the MCCB composite cognitive score.

5.2.6 Correlation between gene-set PRS

All the gene sets defined here are not mutually exclusive from each other. In order to gauge the size of these overlaps, the Pearson correlation coefficient was calculated between all gene-set PRS including the schizophrenia and IQ collated sets. This is not a direct test of whether each gene set PRS is mutually exclusive from one another, but it will give an indication if an overlap of genes is responsible for a similar R2 value observed between a schizophrenia gene set PRS and an IQ gene-set PRS.

5.2.7 Power of gene-set PRS

At the time of writing, there is no published evidence that gene-set PRS confer comparable power to that of the genome-wide and genic PRSs. The power of each gene set PRS was calculated.

A two sample z test was performed using the power.z.test() function found in R. This calculation requires an alpha, sigma, the N number and the effect size. Alpha was set to 0.05, N was set to 725 (The number of samples with a definitive schizophrenia diagnosis), sigma was set as the standard error of the linear regression between cognition with schizophrenia patients and the PRS, multiplied by the square root of N. The effect size was the Beta coefficient of the linear regression between cognition with schizophrenia patients and the PRS.

5.3 Results

5.3.1 Gene-set analysis

In the original schizophrenia GWA study, targets of the **F**ragile X **M**ental **R**etardation **P**rotein (FMRP), abnormal behavior, abnormal nervous system electrophysiology, abnormal long-term potentiation, voltage-gated calcium channel complexes, 5-HT2C receptor complex and loss of function intolerant genes were associated to schizophrenia. In SCZminusCOGS, three gene-sets were associated with schizophrenia after adjusting for multiple testing correction. These were loss of function intolerant genes, FMRP targets and abnormal behavior (see Table 5.1). These gene-sets were also the most significantly associated to a large schizophrenia GWA study (Pardiñas et al., 2018), which included the CardiffCOGS samples.

TABLE 5.1: Functional gene set analysis in SZ-COGS

Gene set	Number of Genes	P-value	MT corrected P-value
LoF intolerant genes	2903	1.5e-16	<1.0e-06
FMRP targets	794	1.1e-09	<1.0e-06
Abnormal behavior	1925	2.3e-04	2.9e-02

P-value = gene-set p-value as derived in MAGMA [26], MT corrected P-value = gene-set p-value corrected for multiple testing using Westfall-Young family wise error rate as defined in MAGMA (Leeuw et al., 2015).

5.3.2 Polygenic Risk Scoring

Association of schizophrenia polygenic risk scores with cognitive ability.

In total, thirteen PRSs were tested for association with cognition at three significance thresholds (See table 5.2 and Supplementary Table C.1). These were the three gene sets associated with schizophrenia, six gene sets associated with IQ, the schizophrenia collated gene set, the IQ collated gene set, the gene-centric gene set and the genome-wide gene set.

A genome-wide PRS conferring the risk of schizophrenia alleles was tested for association with the MCCB composite cognition score. Across all three Pt within SCZminusCOGS, a schizophrenia genome-wide PRS was not significantly associated with a lower cognitive measure (FDR < 0.05; Figure 5.2).

Three gene-set PRS containing gene-sets was associated to schizophrenia were also tested for association with cognitive ability within CardiffCOGS (Figure 5.2). No

TABLE 5.2: Number of SNPs within each PRS at different Pt

	Significance thresholds		
PRS	5e-08	0.05	1
CNS Neuron Differentiation	8	362	1184
Neurogenesis	39	2422	8269
Neuron differentiation	23	1639	5639
Positive regulation of NS development	13	892	2955
Regulation of NS development	26	1424	4792
Regulation of synapse structure or activity	7	566	1807
IQ collated set	41	2611	8912
Abnormal behavior	34	3167	10779
FMRP targets	36	1856	5896
LoF intolerant genes	67	4798	15637
SCZ collated set	80	6602	22298
Gene-centric	115	13206	47563
Genome-wide	153	21147	75696

The number under each significance threshold indicates the number of SNPs found at each threshold for each respective PRS. The number of SNPs at a Pt of 1e-06,1e-04,0.01,0.1,0.2 and 0.5 can be found in the Appendix at Table C.1

association was found in any gene-set PRS. Similarly, no association was found in the collated schizophrenia set.

In contrast to the generally negative findings with schizophrenia sets, schizophrenia risk alleles in five sets identified from the GWA study of IQ (neurogenesis, positive regulation of nervous system development, regulation of nervous system development, regulation of synapse structure or activity and the collated IQ set), were significant at Pt = 0.05. Moreover, these sets captured more variation in cognitive ability than either the genome-wide or the gene-centric PRS at Pt = 0.05 (Figure 5.2) and survived FDR multiple correction testing (all five FDR p = 0.043, see Supplementary table C.2). There are five PRS with the same FDR p-value because of method of FDR multiple testing correction. Under this procedure, the p-values are first ordered from the highest to the lowest p-value. The FDR value is calculated by following: $numberofsamples/positionofthep - value * p - value$. However, it only outputs the cumulative minimum of these values. For example, if $39/2 * p - value$ is less than $39/1 * p - value$ then the $39/2p - value$ is output instead)

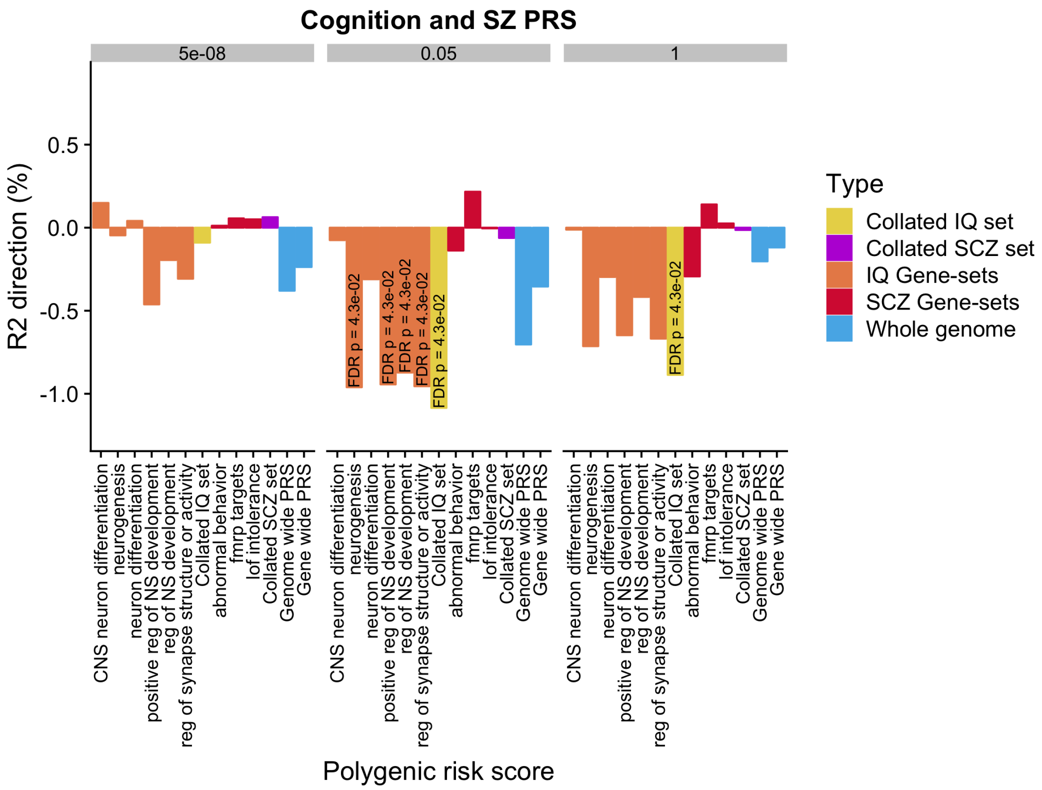


FIGURE 5.2: **Proportion of variance explained for schizophrenia PRS**. The subheadings at the top of each bar plot indicate the Pt of which the PRS was created. The R2 direction (%) axis indicates the amount of variation explained (in the form of a percentage) by the respective PRS with the MCCB cognitive score in the CardiffCOGS individuals with the direction of effect incorporated in. For example an R squared below zero reflect the direction of effect is consistent with a higher burden of schizophrenia risk alleles is associated with lower cognitive ability. Associations surviving FDR correction are reflected by notation of the FDR p-value. CNS = Central Nervous System. reg = regulation. NS = Nervous System.

Association of IQ polygenic risk scores with MCCB composite cognitive score

Twelve of the thirteen gene sets tested using the IQ PRS were associated with cognitive score in people with schizophrenia (Figure 5.3), the exception being central nervous system neuron differentiation. Association for the genome-wide PRS was seen across all three p-value thresholds (Pnom = 2.84e-12,FDR p = 2.32e-10, R2 = 5.00%, Pt = 0.05), but the genome wide PRS narrowly captured less variation in the MCCB composite cognitive score over the genic wide PRS (Pnom = 2.53e-12, FDR p = 9.88e-11, R2 = 5.64%, Pt = 0.05; Figure 5.3).

No gene-set PRS captured more variation in the MCCB composite cognitive score than the genome-wide PRS or the gene-centric PRS. All results for figure 5.3 can be found in the Appendix at table C.3.

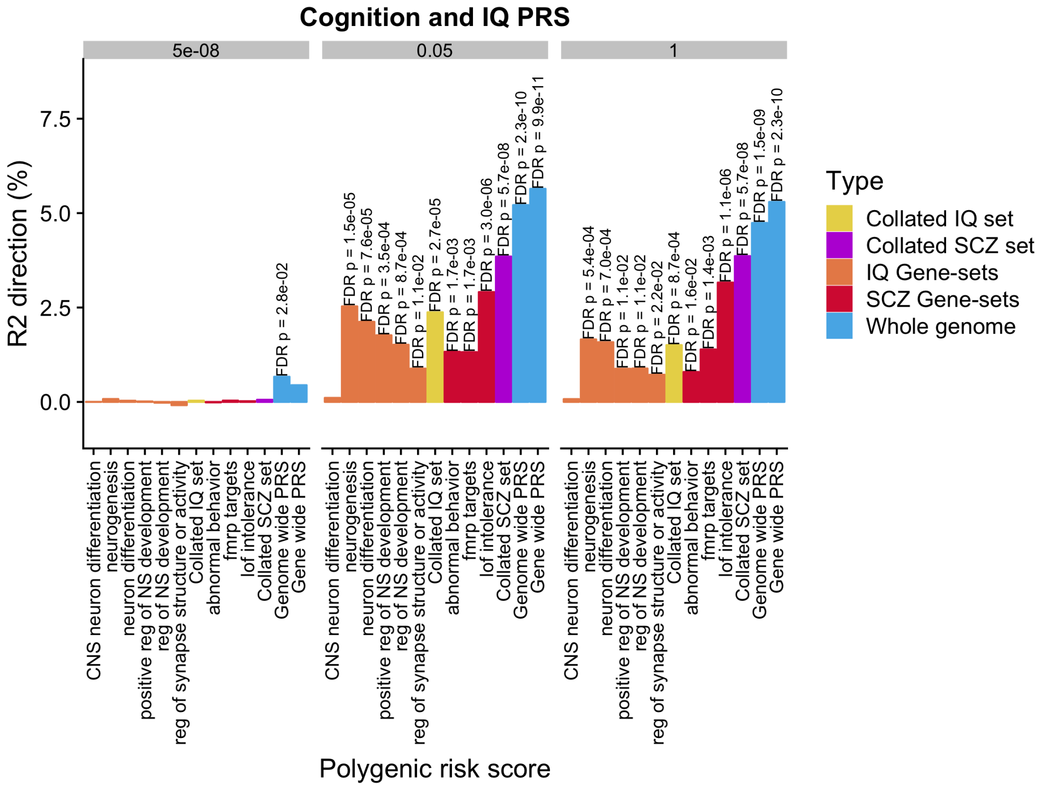
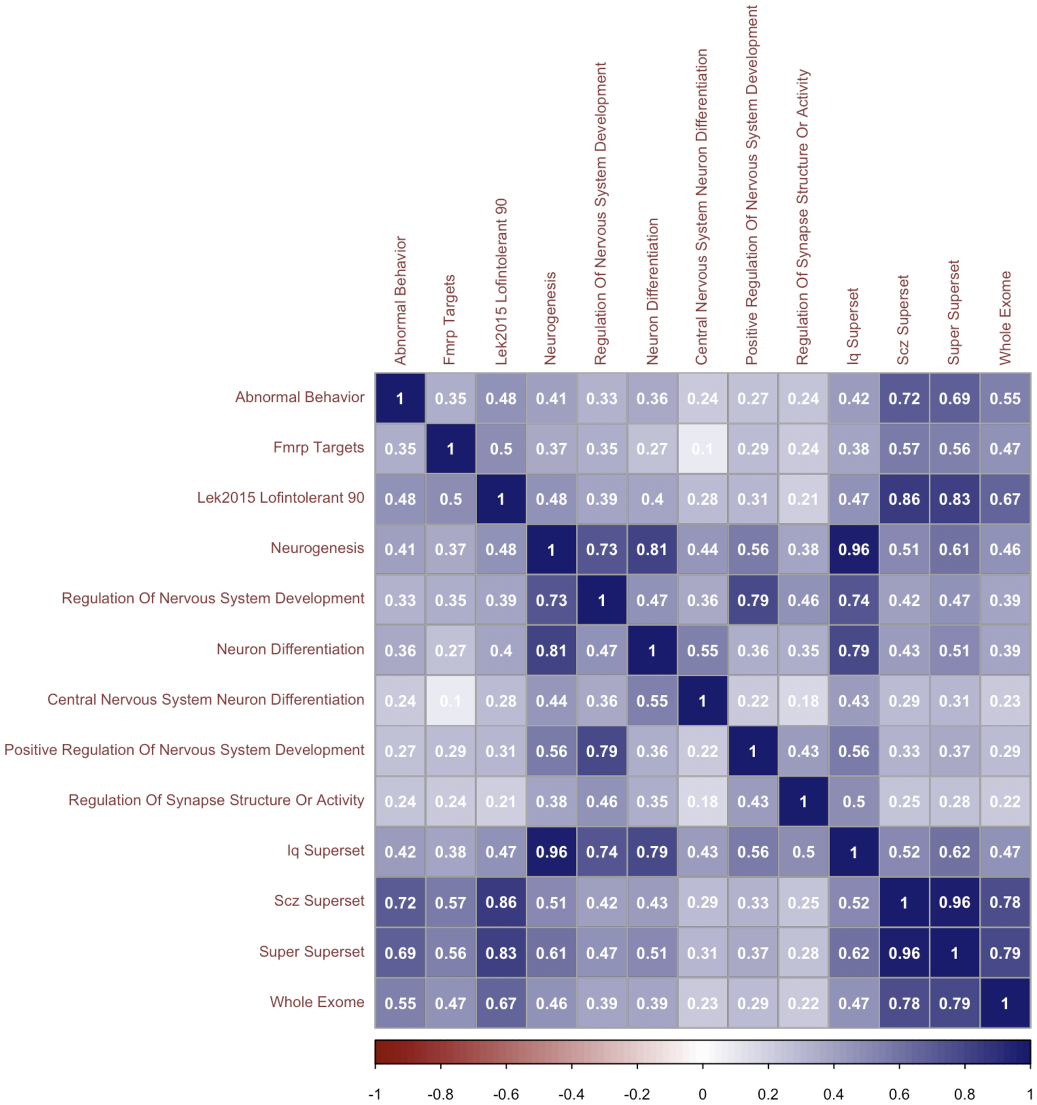


FIGURE 5.3: **Proportion of variance in cognition in people with schizophrenia explained by IQ PRS.** P-value threshold for selecting risk alleles from the training GWA studies are designated as Pt. R2 values below zero reflect the direction of effect is consistent with a higher propensity of IQ alleles is associated with lower cognitive ability. Associations surviving FDR correction are reflected by notation of the FDR p-value.

Correlation of PRS

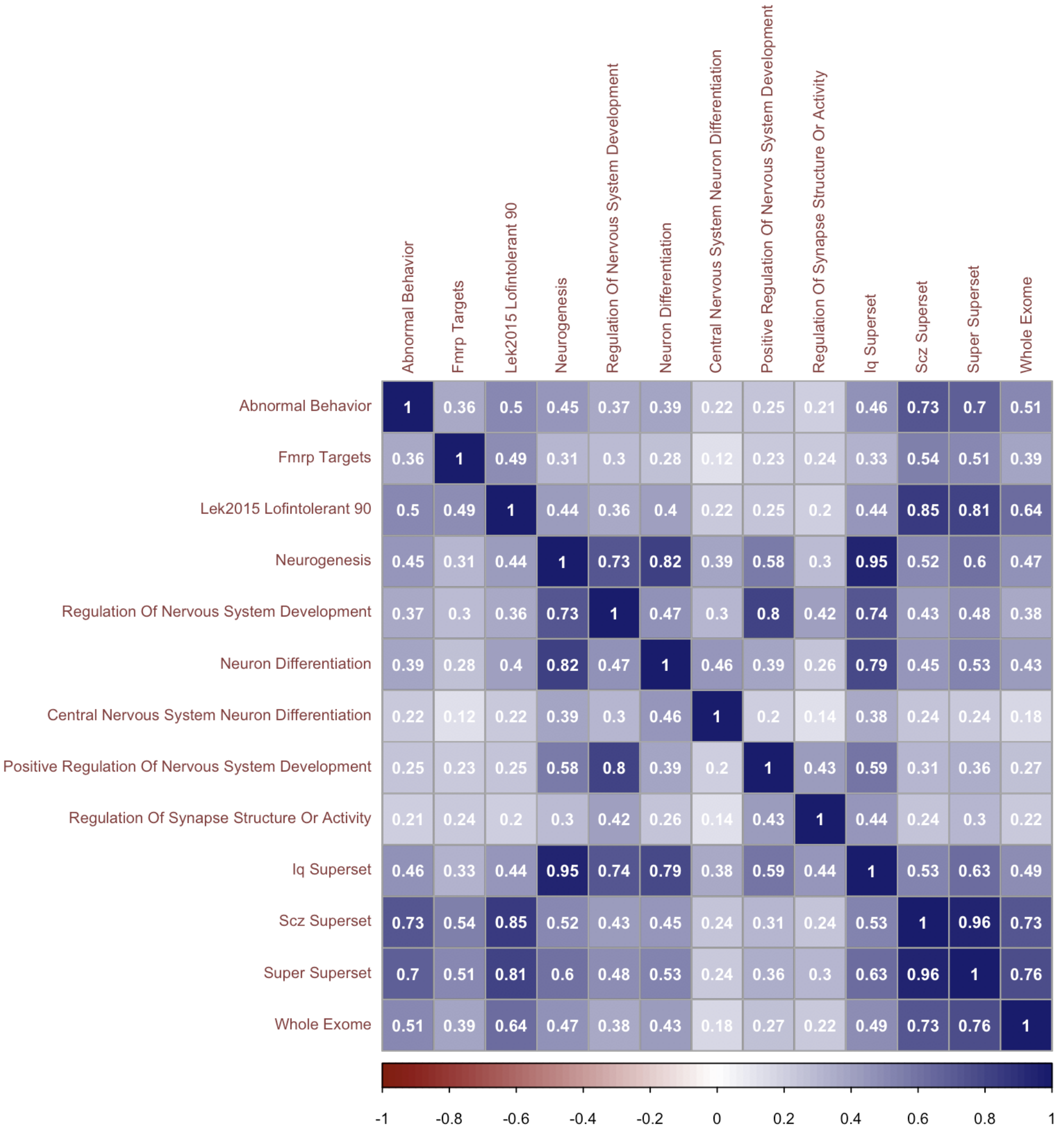
When IQ alleles were used to inform the PRSs, the Collated IQ PRS contained a correlation coefficient of 0.52 with the Collated schizophrenia PRS when IQ alleles were used to inform the PRS (Figure 5.4). The collated schizophrenia set was generally more correlated with the schizophrenia gene-set PRSs, than the IQ gene-set PRSs.

When schizophrenia alleles were used to inform the PRSs, a similar pattern of correlation coefficients as described above was observed (Figure 5.5).



	Abnormal Behavior	Fmrp Targets	Lek2015 Lofintolerant 90	Neurogenesis	Regulation Of Nervous System Development	Neuron Differentiation	Central Nervous System Neuron Differentiation	Positive Regulation Of Nervous System Development	Regulation Of Synapse Structure Or Activity	Iq Superset	Scz Superset	Super Superset	Whole Exome
Abnormal Behavior	1	0.35	0.48	0.41	0.33	0.36	0.24	0.27	0.24	0.42	0.72	0.69	0.55
Fmrp Targets	0.35	1	0.5	0.37	0.35	0.27	0.1	0.29	0.24	0.38	0.57	0.56	0.47
Lek2015 Lofintolerant 90	0.48	0.5	1	0.48	0.39	0.4	0.28	0.31	0.21	0.47	0.86	0.83	0.67
Neurogenesis	0.41	0.37	0.48	1	0.73	0.81	0.44	0.56	0.38	0.96	0.51	0.61	0.46
Regulation Of Nervous System Development	0.33	0.35	0.39	0.73	1	0.47	0.36	0.79	0.46	0.74	0.42	0.47	0.39
Neuron Differentiation	0.36	0.27	0.4	0.81	0.47	1	0.55	0.36	0.35	0.79	0.43	0.51	0.39
Central Nervous System Neuron Differentiation	0.24	0.1	0.28	0.44	0.36	0.55	1	0.22	0.18	0.43	0.29	0.31	0.23
Positive Regulation Of Nervous System Development	0.27	0.29	0.31	0.56	0.79	0.36	0.22	1	0.43	0.56	0.33	0.37	0.29
Regulation Of Synapse Structure Or Activity	0.24	0.24	0.21	0.38	0.46	0.35	0.18	0.43	1	0.5	0.25	0.28	0.22
Iq Superset	0.42	0.38	0.47	0.96	0.74	0.79	0.43	0.56	0.5	1	0.52	0.62	0.47
Scz Superset	0.72	0.57	0.86	0.51	0.42	0.43	0.29	0.33	0.25	0.52	1	0.96	0.78
Super Superset	0.69	0.56	0.83	0.61	0.47	0.51	0.31	0.37	0.28	0.62	0.96	1	0.79
Whole Exome	0.55	0.47	0.67	0.46	0.39	0.39	0.23	0.29	0.22	0.47	0.78	0.79	1

FIGURE 5.4: **Correlation of gene-set PRS using IQ alleles**. The x-axis displays the heatmap scale of the Pearson's correlation coefficient for each comparison. Each numerical Pearson's correlation coefficient is also displayed for each comparison. IQ3 was used as the training set for each PRS. In the diagram: Iq Superset = Collated IQ PRS, Scz Superset = Collated schizophrenia PRS, Super Superset = all alleles contained with the IQ and SCZ collated sets combined into one PRS. Whole exome = Gene-centric PRS.



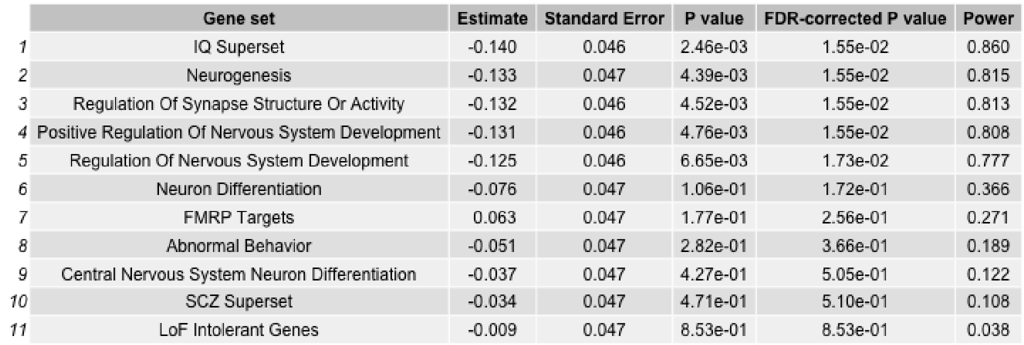
	Abnormal Behavior	Fmrp Targets	Lek2015 Lofintolerant 90	Neurogenesis	Regulation Of Nervous System Development	Neuron Differentiation	Central Nervous System Neuron Differentiation	Positive Regulation Of Nervous System Development	Regulation Of Synapse Structure Or Activity	Iq Superset	Scz Superset	Super Superset	Whole Exome
Abnormal Behavior	1	0.36	0.5	0.45	0.37	0.39	0.22	0.25	0.21	0.46	0.73	0.7	0.51
Fmrp Targets	0.36	1	0.49	0.31	0.3	0.28	0.12	0.23	0.24	0.33	0.54	0.51	0.39
Lek2015 Lofintolerant 90	0.5	0.49	1	0.44	0.36	0.4	0.22	0.25	0.2	0.44	0.85	0.81	0.64
Neurogenesis	0.45	0.31	0.44	1	0.73	0.82	0.39	0.58	0.3	0.95	0.52	0.6	0.47
Regulation Of Nervous System Development	0.37	0.3	0.36	0.73	1	0.47	0.3	0.8	0.42	0.74	0.43	0.48	0.38
Neuron Differentiation	0.39	0.28	0.4	0.82	0.47	1	0.46	0.39	0.26	0.79	0.45	0.53	0.43
Central Nervous System Neuron Differentiation	0.22	0.12	0.22	0.39	0.3	0.46	1	0.2	0.14	0.38	0.24	0.24	0.18
Positive Regulation Of Nervous System Development	0.25	0.23	0.25	0.58	0.8	0.39	0.2	1	0.43	0.59	0.31	0.36	0.27
Regulation Of Synapse Structure Or Activity	0.21	0.24	0.2	0.3	0.42	0.26	0.14	0.43	1	0.44	0.24	0.3	0.22
Iq Superset	0.46	0.33	0.44	0.95	0.74	0.79	0.38	0.59	0.44	1	0.53	0.63	0.49
Scz Superset	0.73	0.54	0.85	0.52	0.43	0.45	0.24	0.31	0.24	0.53	1	0.96	0.73
Super Superset	0.7	0.51	0.81	0.6	0.48	0.53	0.24	0.36	0.3	0.63	0.96	1	0.76
Whole Exome	0.51	0.39	0.64	0.47	0.38	0.43	0.18	0.27	0.22	0.49	0.73	0.76	1

FIGURE 5.5: **Correlation of gene-set PRS using SCZ alleles**. The x-axis displays the heatmap scale of the Pearson's correlation coefficient for each comparison. Each numerical Pearson's correlation coefficient is also displayed for each comparison. SCZminusCOGS was used as the training set for each PRS. In the diagram: Iq Superset = the collated IQ PRS, Scz Superset = the collated schizophrenia PRS, Super Superset = all alleles contained with the IQ and SCZ collated sets combined into one PRS. Whole exome = Gene-centric PRS.

Power of gene-set PRS

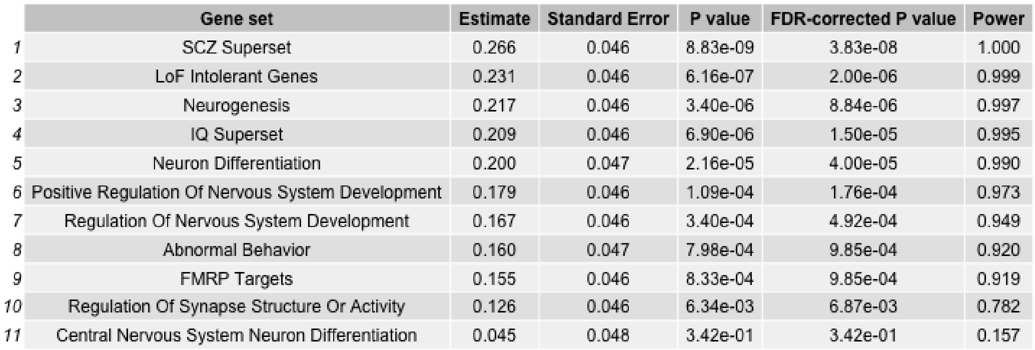
All gene-set PRS that produced an FDR corrected P-value below 0.05, produced a Power statistic above 0.77. The results observed was consistent when the training data set was schizophrenia (Table 5.3) or IQ3 (Table 5.4).

The Power statistic in this context is the probability of the regression between cognition and the gene-set PRS to be able to detect a significant effect when one exists. The Power of any hypothesis test (including the above regression) lies between 0 and 1. The closer the Power statistic is to 1, the more likely the regression will be able to detect an existing significant effect. In general, the threshold where the power for the hypothesis test is considered 'good' is above 0.8. Two gene-set PRS with a FDR corrected P-value below 0.05 (Regulation of Synapse Structure or Activity (IQ3) Regulation of Nervous System Development (SCZ)) failed to meet this threshold.



	Gene set	Estimate	Standard Error	P value	FDR-corrected P value	Power
1	IQ Superset	-0.140	0.046	2.46e-03	1.55e-02	0.860
2	Neurogenesis	-0.133	0.047	4.39e-03	1.55e-02	0.815
3	Regulation Of Synapse Structure Or Activity	-0.132	0.046	4.52e-03	1.55e-02	0.813
4	Positive Regulation Of Nervous System Development	-0.131	0.046	4.76e-03	1.55e-02	0.808
5	Regulation Of Nervous System Development	-0.125	0.046	6.65e-03	1.73e-02	0.777
6	Neuron Differentiation	-0.076	0.047	1.06e-01	1.72e-01	0.366
7	FMRP Targets	0.063	0.047	1.77e-01	2.56e-01	0.271
8	Abnormal Behavior	-0.051	0.047	2.82e-01	3.66e-01	0.189
9	Central Nervous System Neuron Differentiation	-0.037	0.047	4.27e-01	5.05e-01	0.122
10	SCZ Superset	-0.034	0.047	4.71e-01	5.10e-01	0.108
11	LoF Intolerant Genes	-0.009	0.047	8.53e-01	8.53e-01	0.038

TABLE 5.3: **Power of Gene-set PRS when SCZ was the training set**. Estimate, standard error, P-value and FDR-corrected P value are the values output from the linear regression between the cognition phenotype and the respective Gene set PRS. Power indicates the two sample z test as performed using the power.z.test() function found in R. This calculation requires an alpha, sigma, the N number and the effect size. Alpha was set to 0.05, N was set to 725 (The number of samples with a definitive schizophrenia diagnosis), sigma was set as the standard error multiplied by the square root of N. The effect size was the Estimate.



	Gene set	Estimate	Standard Error	P value	FDR-corrected P value	Power
1	SCZ Superset	0.266	0.046	8.83e-09	3.83e-08	1.000
2	LoF Intolerant Genes	0.231	0.046	6.16e-07	2.00e-06	0.999
3	Neurogenesis	0.217	0.046	3.40e-06	8.84e-06	0.997
4	IQ Superset	0.209	0.046	6.90e-06	1.50e-05	0.995
5	Neuron Differentiation	0.200	0.047	2.16e-05	4.00e-05	0.990
6	Positive Regulation Of Nervous System Development	0.179	0.046	1.09e-04	1.76e-04	0.973
7	Regulation Of Nervous System Development	0.167	0.046	3.40e-04	4.92e-04	0.949
8	Abnormal Behavior	0.160	0.047	7.98e-04	9.85e-04	0.920
9	FMRP Targets	0.155	0.046	8.33e-04	9.85e-04	0.919
10	Regulation Of Synapse Structure Or Activity	0.126	0.046	6.34e-03	6.87e-03	0.782
11	Central Nervous System Neuron Differentiation	0.045	0.048	3.42e-01	3.42e-01	0.157

TABLE 5.4: **Power of Gene-set PRS when IQ3 was the training set**. Estimate, standard error, P-value and FDR-corrected P value are the values output from the linear regression between the cognition phenotype and the respective Gene set PRS. Power indicates the two sample z test as performed using the power.z.test() function found in R. This calculation requires an alpha, sigma, the N number and the effect size. Alpha was set to 0.05, N was set to 725 (The number of samples with a definitive schizophrenia diagnosis), sigma was set as the standard error multiplied by the square root of N. The effect size was the Estimate.

5.4 Discussion

I have explored the relationship between the genetic risk of schizophrenia, the genetic predisposition towards general IQ, and the variation in cognitive performance in schizophrenia patients using PRS approaches. I started by discovering gene-sets which were significantly associated to the novel GWA study, SCZminusCOGS (described in Chapter 2). I then found that IQ gene-set PRS captured more variation for the cognition within schizophrenia phenotype than the schizophrenia gene-set PRS and the genome-wide PRS when schizophrenia alleles were used. Furthermore, when IQ alleles were used, most gene-set PRS and all genome-wide PRS were positively associated with the cognition within schizophrenia phenotype.

In a recent study by Richards *et al.* (2019), an IQ PRS was found to be significantly positively associated with a derivation of the 'general intelligence factor', *g* in individuals with schizophrenia. However, no association was found with the same phenotype when the risk of schizophrenia informed the PRS. The absence of any association of the schizophrenia PRS is a result I replicate and expand upon with the gene-set and the genome-wide PRSs here (with the addendum that all individuals within the testing data-set used here was a subset of the samples in Richards *et al.* (2019)). If cognitive impairment is a functional outcome of schizophrenia and there is evidence to show that the cause of this was genetic, it would seem counter-intuitive that the genetic risk of schizophrenia has no part to play in the biological processes that cause cognitive impairment in individuals with schizophrenia. This is counter-intuitive because if the genetic risk of schizophrenia does not cause a sub-phenotype of schizophrenia, from where does the genetic risk originate and how is it specific to cognitive deficits observed within schizophrenia patients?

Previous studies have observed negative correlations and associations between schizophrenia PRS and cognition or **E**ducational **A**ttainment (EA) (Hubbard et al., 2016; Hagenaars et al., 2016b; Hill et al., 2016; Toulopoulou et al., 2010; Fowler et al., 2012). The same observation is observed here, but I also observe that schizophrenia alleles appear to predict cognition within schizophrenia sets in IQ defined sets, but not schizophrenia defined sets. In addition, IQ alleles predict cognition within schizophrenia patients regardless of the gene-set they are located within. I propose a theory that these observations may be explained via pleiotropy.

Within the schizophrenia defined gene sets, schizophrenia alleles are depleted for pleiotropic effects on cognition (at the very least, within individuals with schizophrenia) compared with the schizophrenia alleles located within the IQ defined sets. Thus, I begin to see evidence for specific phenotypic effects of subsets of schizophrenia risk alleles defined by biological pathways.

The extensive evidence for genetic correlations across cognitive and psychiatric phenotypes implies that widespread pleiotropy exists across these traits (Lee et al., 2013). Additionally, a recent Phe-WAS on a schizophrenia PRS displayed pleiotropy across several mental health disorders (Zheutlin et al., 2019). This does not imply that any given individual allele has pleiotropic effects. The observation that schizophrenia alleles in the biological pathways most highly enriched for schizophrenia liability do not influence cognitive ability, whereas IQ associated alleles within the same gene-set do, implies that the specific mechanisms underpinning these two phenotypes are non-overlapping, and that for these most strongly enriched pathways, schizophrenia liability is not broadly mediated by effects on cognition.

The observation that schizophrenia risk alleles in gene sets associated with IQ are also associated with cognition in schizophrenia suggests that schizophrenia risk alleles in IQ associated gene sets are likely to be pleiotropic.

This in turn suggests that the mechanisms underpinning schizophrenia and cognition in processes most robustly implicated in cognition may be substantially overlapping, and that for this set of schizophrenia risk alleles, their effects are mediated by effects on cognition.

The above observations contain the assumption that each gene set PRS is independent from one another. Evidence of independent associations of the schizophrenia gene-sets to schizophrenia and the IQ gene sets to IQ was shown within the MAGMA gene-set conditional analysis performed in the original publications (Pardiñas et al., 2018; Savage et al., 2018). However, there is no previous evidence showing that the schizophrenia gene-sets are independent from the IQ gene-sets. No high correlations were found between the collated schizophrenia gene-set PRSs and the collated IQ gene-set PRSs, supporting the assumption of independence.

Further analysis is required to support the observations in the PRS analysis here. More direct statistical tests exist for pleiotropy observed within a set of SNPs including, for example **G**eneralised **S**ummary-data-based **M**endelian **R**andomisation (GSMR) (Zhu et al., 2018) and Phe-WAS (Zheutlin et al., 2019). The investigation into pleiotropy between schizophrenia and cognitive phenotypes has already provided evidence to further our understanding of these two traits. Lam et al. (2019b) examined the correlations between groups of SNPs to identify which subsets of SNPs contained concordant effect sizes for educational attainment, cognition and schizophrenia liability. MAGMA pathway analysis was then performed to highlight early neurodevelopmental pathways that characterize concordant allelic variation, and adulthood synaptic pruning pathways, which may be contribute to the positive genetic association observed between educational attainment and schizophrenia (Lam et al., 2019b).

Legge et al. (Legge et al., 2021) used factor analysis to examine the relationship between the phenotypic dimensions of schizophrenia (eg negative and positive symptoms) and the PRS for schizophrenia, intelligence and other neuropsychiatric disorders. They found that while both schizophrenia and intelligence PRS was associated to current cognitive ability, only the intelligence PRS was associated with premorbid IQ. In addition, when adding premorbid IQ as a covariate with current cognitive ability, the schizophrenia PRS remained significant, but the intelligence PRS did not. This suggests that current cognitive ability in schizophrenia is partly a function of premorbid IQ, influenced by both schizophrenia and intelligence SNPs. It may be interesting to test the association of both intelligence and schizophrenia gene-set PRS created here, on the phenotypes of current cognitive ability and premorbid IQ to further dissect the relationship between cognition and intelligence and schiophrneia variants, as well as to test whether similar results found here are replicated with a different phenotype describing cognition within schizophrenia.

In addition, it would be beneficial to replicate the results observed here within another cohort (for example UK Biobank) to ensure that the associations observed here are not simply statistical noise within the CardiffCOGS cohort.

The size of CardiffCOGS is also a limitation for this study. While the cognition phenotype within schizophrenia patients is well defined, only approximately one thousand individuals had their cognitive phenotype recorded and less actually had a **D**iagnostic (and) **S**tatistical **M**anual (of Mental Disorders) (DSM)-IV defined diagnosis of schizophrenia. However, I did perform power calculations and observed that enough individuals were included within each linear regression to detect authentic signal when significant associations were found.

In summary, gene-set PRSs appears to be a useful tool to investigate the relationship between schizophrenia and one of it's endophenotypes: cognition. I have provided further evidence that the variance observed in IQ within the general population has similar genetic causes to the variance observed within cognition in schizophrenia patients. I theorise that the variation in schizophrenia risk may play a role in the variation observed within cognition of schizophrenia patients through pleiotropic alleles, but further work is required to support this.

6 Discussion and Future work

6.1 Conclusions

The primary aim of this thesis was to investigate whether a gene-set PRS method could help understand the role that the common genetic risk of schizophrenia plays in the pathogenesis of schizophrenia and its associated phenotypes. This started with an investigation into how to produce a gene-set PRS given that there is no standardised method for producing a gene-set PRS. A gene-set PRS bioinformatics workflow named SurPRSe was created and its ability to produce gene-set PRS was tested against existing software. SurPRSe was then applied to two schizophrenia related research questions: 1. Is the lack of association observed between the common genetic risk of schizophrenia and subcortical brain region volumes due to too much noise within the genome-wide PRS? 2. Does the common genetic risk for schizophrenia have any role to play with cognitive impairment observed in schizophrenia patients?

With regards to the first question, the results were inconclusive. While it appeared apparent that even a maximally powered genome-wide PRS would not be able to capture the schizophrenia common genetic component affecting subcortical brain region volumes (if indeed one exists), the ability of a gene set PRS to capture this variation will require further replication and investigation due to the complexities of gene-set analysis and the statistical rigour required to produce confident results.

With regards to the second question, the outlook for the use of gene-set PRS to disentangle the cross-disorder common genetics between neuropsychiatric traits was promising. The gene-set PRS were able to provide a hypothesis on how common schizophrenia risk affects cognitive impairment in individuals with schizophrenia, conclusions must be caveated by the knowledge that this is a novel technique and that this technique reduces the amount of information available compared to a genome-wide PRS. The the effect sizes for each SNP within the PRS are consistently very small, so the power of each gene-set PRS should be investigated further than displayed here.

In the next section I will provide an overview on the reasons why the creation of SurPRSe was required and the limitations involved in its use. I will then expand on the explanations regarding the two research questions that SurPRSe was applied to. Finally, I will ponder the future of using gene-set PRS in schizophrenia research.

6.2 Discussion

6.2.1 SurPRSe

Benchmarking bioinformatics tools is becoming ever more important in the era of considerably larger and more complex genetic data sets (Aniba, Poch, and Thompson, 2010). The main issue within the system of the creation of bioinformatics workflows and tools can be summarised by the term 'workflow' or 'tool'. Outside of bioinformatics and biology, the term 'workflow' is not used to describe a set of coded software instructions to perform a specific task. The term used is 'program' in which there is an expectation that the program is built robustly with a large amount of safeguards to ensure the program does not fail when it is run by multiple users. Within the current academic system, a bioinformatics tool or workflow only requires the tool to pass the peer-review system without any analysis of the code-base itself. In addition, within bionformatics, very little resources are allocated to ensure the workflow is running as expected. In most cases the workflow is produced by one or two authors and contains limited to no safeguards. In addition, in the field of neuropsychiatric genetics, many bioinformatics tools are based on data that is either open access, owned by the group which created the bioinformatics tools itself or on simulated data.

For example, in the case of the software PRSice-2 (Choi and O'Reilly, 2019), the only data used to test the efficacy for general use was UK Biobank as the genotyped data (Sudlow et al., 2015), a GWA study by the GIANT consortium (Locke et al., 2015) and simulated training sets with a heritability of 0.2 and 0.6 (Choi and O'Reilly, 2019). The simulated data compared PRSice-2 to other PRS software in lassosum (Mak et al., 2017) and LDpred (Vilhjálmsson et al., 2015), but only displayed information on the ability of PRSice-2 to computationally scale-up much better than existing alternatives, a pattern we observed here when PRSet was compared to SurPRSe, and the predictive power of each PRS. The tool is advertised as being able to perform "automation of PRS analyses applied to large-scale genotype-phenotype data", but has only been systematically tested on three data sets at the time of writing, and no mention has been made of its gene-set PRS tool PRSet (Choi and O'Reilly, 2019). The difference between designing a tool that works for one polygenic risk score analysis compared to a tool that is versatile is vast, and can contain many computational bugs which may not be observed without sufficient testing of the tool against a baseline which incorporates all known instances of "large-scale genotype-phenotype data" (Choi and O'Reilly, 2019).

SurPRSe was built for two reasons. 1) The insufficient testing and constant development of existing gene-set PRS tools was inappropriate for use within a three year project and 2) the majority of the computational work involved the QC and standardisation of the input data sets into the PRS, a process which has limitations in existing software. The computational work to produce the scoring within a PRS is simple enough to be performed outside of existing software if required, but has been well documented and used in the bioinformatics tool PLINK 1.9 (Chang et al., 2015). However, as described in chapter 2, fourteen data sets were used to produce the multiple PRSs used within this thesis and performing individual QC steps across each data set to set it up for a simple scoring algorithm would be an inefficient use of time. In addition, there is no system in place to remove computational artifacts from explaining the results, all the QC steps would be performed using individual scripts suited towards each data set. These artifacts can have profound effects on the interpretation of the PRSs as evidenced in section (include unit testing example from appendix) due to the low effect sizes and r-squared values generally seen when analysing post-GWAS data on complex neuropsychiatric traits including schizophrenia and cognition. SurPRSe does not guarantee the removal of computational artifacts from the results, but does provide an automated, standardised procedure that is suited towards the data stored within the databank at Cardiff university, and was applied to all fourteen data sets within this thesis.

There are however, some limitations to using SurPRSe across the thesis as opposed to alternative software or individualised PRS analyses. The most influential is the loss of power that is associated with automating any procedure. For example, it is indeed possible to included insertions and deletions within the PRS as described in Chapter 3. However, in most analyses, these SNPs have low MAFs and/or INFO scores and are removed as a part of the standardised QC procedures. These SNPs can also be a part of the exclusion criteria in the creation of a PRS (Ripke et al., 2014; Hess et al., 2019). There was no cost effective benefit to include the option to keep these SNPs within SurPRSe, but it may have had a minor effect on the interpretation of genome-wide PRS and especially gene-set PRS when the power of the data is reduced significantly. In addition, as with other bioinformatics workflows, the resources and time to turn SurPRSe into a robust program was not available, and so must come under the nomenclature of a 'workflow' instead.

6.2.2 Gene-set PRS

The creation of gene-set PRS may have wide-ranging implications in the understanding of neuropsychiatric traits. The aggregation of variants into biologically meaningful pathways on an individual level basis could not only dissect schizophrenia from its subphenotypes, but also stratify individuals based on their predisposition towards schizophrenia. This could provide many clinical benefits including, for example, the efficacy of a particular treatment or antipsychotic on a patient with schizophrenia. While there has been some success in defining gene-sets associated with schizophrenia from a population (Pardiñas et al., 2018), the next hurdle for clinical application is the interpretation of the biological significance of these gene-sets on an individual level basis.

In the context of this project, the main limitation of gene-set analysis that had to be overcome was the absence of gold-standard standardisation across different gene-set database resources, as this creates problems when defining hypotheses that require comparisons across multiple traits. If you were to compare a genome-wide PRS between schizophrenia and cognition for example, the analysis could be defined within one to two regression or genetic correlation analyses. As long as the population and array effects are accounted for within both the schizophrenia and cognition data sets, the analysis structure is robustly defined as the total risk of schizophrenia compared to the total predisposition with cognition. However, as soon as you partition the genetic components of each trait, the hypotheses may become less interpretative. For example even in the direct comparison of a loss of function intolerance gene-set, the analysis structure of the above becomes the genetic disposition towards schizophrenia within the loss of function intolerant genes compared to the genetic disposition towards cognition within the loss of function intolerant genes. The robustness of this comparison now relies upon the definition of the loss of function intolerant genes gene-set and the association of the loss of function intolerant genes with schizophrenia and cognition. The latter is usually solved by performing a gene-set association analysis on the trait of interest to define gene-sets which are associated with either schizophrenia or cognition. In the case of the robustness of the definition of the loss of function intolerant genes gene-set, there is a wider methodological problem which is detailed below.

In the case of breast cancer, it was found that SNPs near the FGFR2 gene demonstrate associations at $p < 10\text{-}300$ (Michailidou et al., 2017). It was also found that 86 of the top 100 pathways associated to breast cancer all contain the FGFR2 gene, and so gene-sets that are un-associated with breast cancer may be artificially driven to the top of the results if they contain the FGFR2 gene within the gene-set (Sun et al., 2019; Michailidou et al., 2017). In the context of schizophrenia and cognition, the loss of function intolerance gene-set is very large, and so may also may show association to cognition simply due to the appearance of cognitive associated genes within the loss of function intolerance genes gene-set. Because gene-sets are un-standardised across the bioinformatics field, there is no systematic method to prove that this is not true, and so the association of each gene-set to each trait is solely reliable on the robustness of each individual gene-set analysis that was performed for each trait. If certain genes or gene-sets are not included, this may have an effect on the interpretation of the results.

A PRS is limited in its ability to confer causal genetic effects by design. It includes variants that may have no true association with, for example, schizophrenia, in order to increase the power of the common polygenic component observed with the schizophrenia trait. The issue is that a PRS is, at the time of writing, the only method which can reliably capture the common genetic component of schizophrenia at an individual level and so it is currently one of the best methods available to discover causal associations of schizophrenia SNPs with other traits. Briefly, a **M**endelian **R**andomisation (MR) experiment requires genetic instruments that are definitively causal to the outcome trait in order to separate out confounding effects from the genetic effects (Davies, Holmes, and Davey Smith, 2018). However, as the effect sizes of even the most associated schizophrenia SNP are small, very few validated risk alleles are available so a MR study would be unsuitable for defining the causal effects of schizophrenia, especially when compared to another complex trait like cognition. There has been some evidence using MR to link the genetic components of other traits towards schizophrenia. A recent MR study showed that the genetic liability towards cannabis use was causal to the genetic risk of schizophrenia, but only ten cannabis-associated SNPs were used as genetic instruments and these SNPs did not pass genome-wide significance (Vaucher et al., 2018). There is also evidence that interleukin-6 effects and low C-reactive protein may increase the risk of schizophrenia within a two-sample MR based experiment (Hartwig et al., 2017).

What is clear is that it is of vital importance to set out a clear hypothesis before the use of any gene-set PRS when applied to a neuropsychiatric trait. While the interpretation of gene-set PRS may be confused due to annotation problems, I have shown here that gene-set PRSs are quite versatile in terms of the research questions that they can be applied to. In the first instance whereby the common genetic risk of schizophrenia was tested for association with subcortical brain volume size, the hypothesis was that gene-set PRSs would explain more variation in subcortical brain region sizes over that of a genome-wide PRS. The scientific basis behind this hypothesis was that there may be heterogeneity within the schizophrenia common alleles affecting different brain structures. As a gene-set PRS incorporates a biological meaning within its risk score, it is the ideal analysis method to test this hypothesis.

Gene-set PRSs can also be used to investigate the genetic architecture between three or more traits. The aim within the second study of my thesis was to examine whether only a part of the genetic predisposition towards schizophrenia and cognition influences the observed phenotype of cognitive impairment within schizophrenia patients. As the gene-set PRS is able to dissect the genetic risk into meaningful biological pathways, it is ideal for this type of cross trait analysis. The added benefit is that you are able to isolate alleles that confer the genetic risk of one trait but are found to be associated to another in order to investigate the combined genetic architecture of both traits. A gene-set PRS may also provide insights into the pleiotropic effects of any selected trait, but the method would need to be refined, which goes beyond the remit of this thesis.

6.2.3 Future work

SurPRSe and PRSAVE

SurPRSe was built to suit the data sets used in this thesis (See Chapter 2) on the computational architecture suited towards Cardiff university. It was however, also built in a way so that other individuals would be able to use it. There is therefore scope to make SurPRSe into software given the right resources and time. SurPRSe is however, quite complex in terms of its computational architecture (aka the codebase), and considering that there are other software options already out there in PRSice (Choi and O'Reilly, 2019) and LDpred (Vilhjálmsson et al., 2015), what may be more beneficial is to develop PRSAVE, the shiny app that visualises the results of PRS association analyses (See chapter 3). As the main issue with using gene-set PRS is the interpretation of the results, creating apps which allow an interactive visualisation of the results can provide simultaneous analysis of the initial hypothesis, while providing information on whether the analysis has worked correctly. When the results are being discussed in a collaborative environment, more focus can be directed on the biological implications of the results, especially if everyone within the room is able to use and understand the app.

Gene-set PRS applied to subcortical brain volumes

The absence of any association of the common genetic risk towards schizophrenia with subcortical brain volumes is a tricky obstacle to overcome in psychiatric genetics. There is an indication of gene-set PRS being a suitable analysis method to investiate the relationship between these two traits but a better refinement of the PRS and a replication is required before committing to this analysis structure. What can be concluded is that genome-wide PRS should no longer be used when examining the relationship between the common genetic risk of schizophrenia an subcortical brain region volumes.

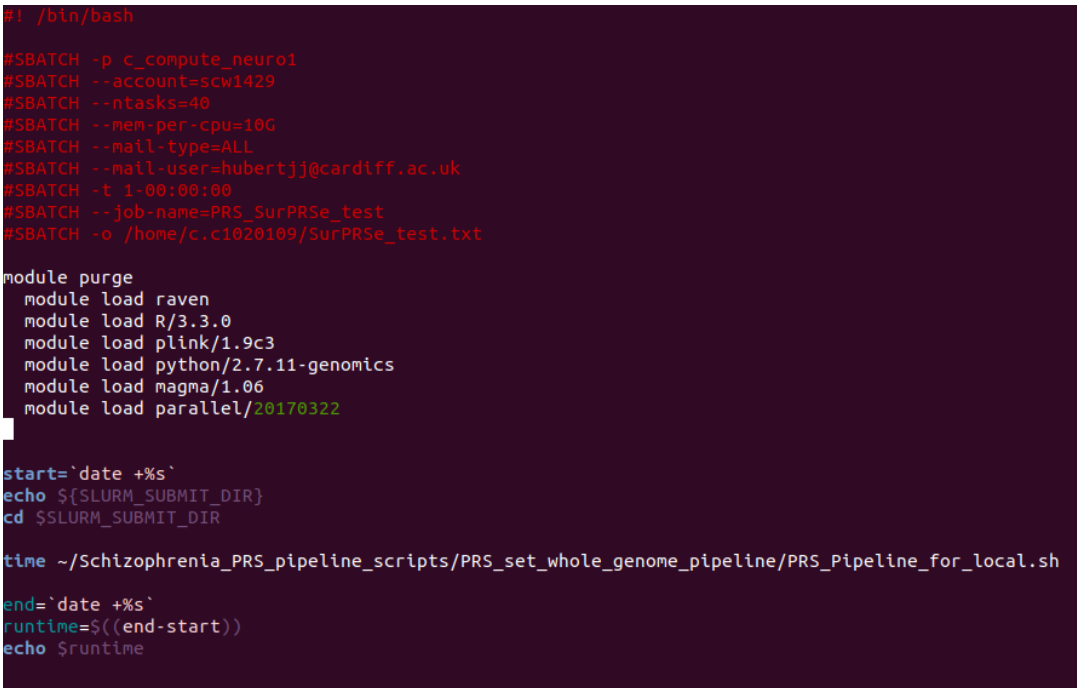
Gene-set PRS applied to cognition within schizophrenia patients

Replication will be required in order to support the findings from Chapter 5. One potential route for replication is through the use of UK Biobank. It may be useful to investigate the genetic architecture of schizophrenia with other traits given the right samples are available, but even so, it may also been useful to compare the gene-set PRS with a trait unrelated to either schizophrenia and cognition, to get a better indication of the quality of the signal that was observed here.

A Appendix for Chapter 3

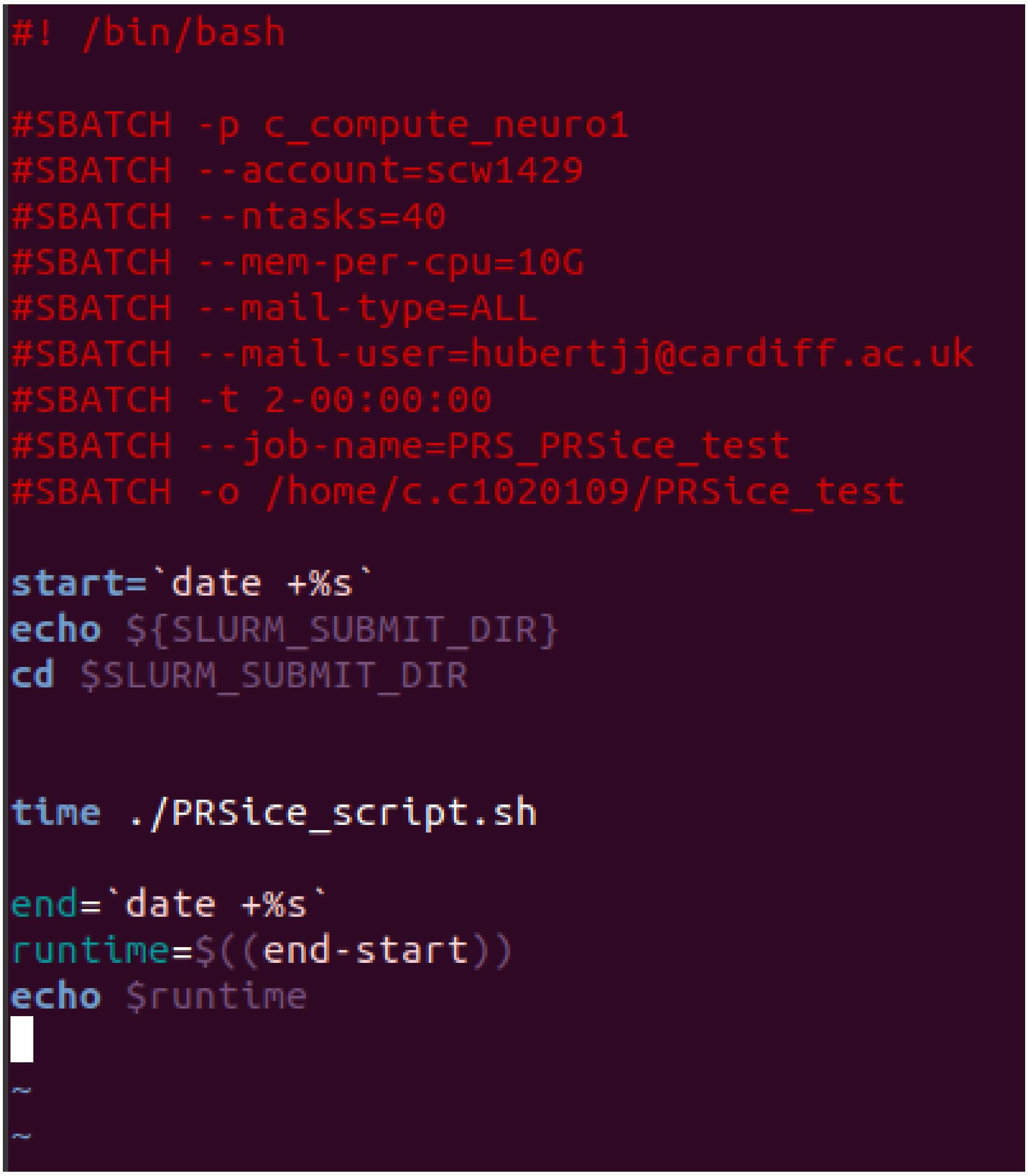
A.1 PRSetvsSurPRSe

A.1.1 Components of SurPRSe



```
#! /bin/bash

#SBATCH -p c_compute_neuro1
#SBATCH --account=scw1429
#SBATCH --ntasks=40
#SBATCH --mem-per-cpu=10G
#SBATCH --mail-type=ALL
#SBATCH --mail-user=hubertjj@cardiff.ac.uk
#SBATCH -t 1-00:00:00
#SBATCH --job-name=PRS_SurPRSe_test
#SBATCH -o /home/c.c1020109/SurPRSe_test.txt

module purge
  module load raven
  module load R/3.3.0
  module load plink/1.9c3
  module load python/2.7.11-genomics
  module load magma/1.06
  module load parallel/20170322


start=`date +%s`
echo ${SLURM_SUBMIT_DIR}
cd $SLURM_SUBMIT_DIR

time ~/Schizophrenia_PRS_pipeline_scripts/PRS_set_whole_genome_pipeline/PRS_Pipeline_for_local.sh

end=`date +%s`
runtime=$((end-start))
echo $runtime
```

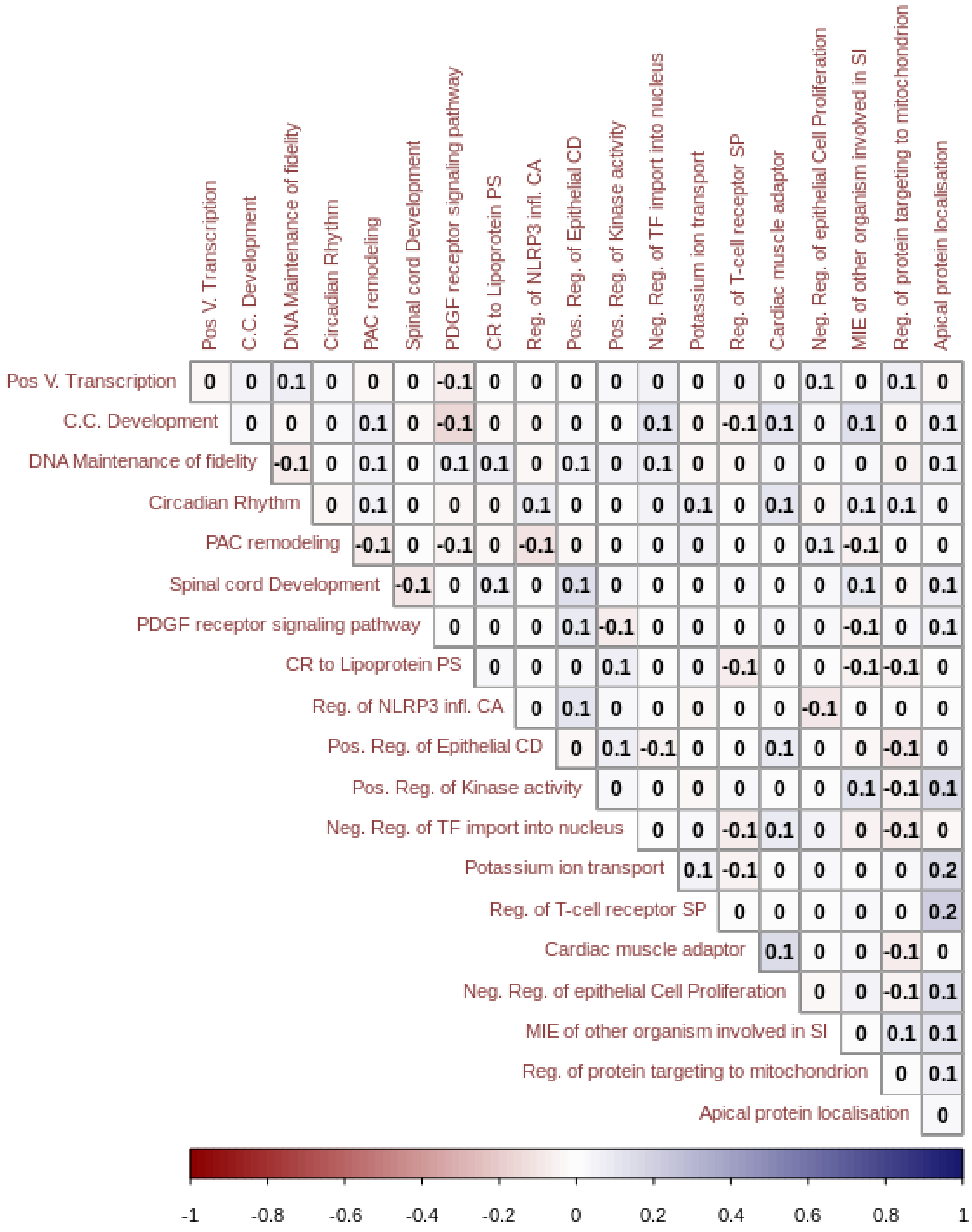
FIGURE A.1: Submission script to Hawk for the runthrough testing of SurPRSe. SurPRSe is run directly after the "time" command.



```
#! /bin/bash

#SBATCH -p c_compute_neuro1
#SBATCH --account=scw1429
#SBATCH --ntasks=40
#SBATCH --mem-per-cpu=10G
#SBATCH --mail-type=ALL
#SBATCH --mail-user=hubertjj@cardiff.ac.uk
#SBATCH -t 2-00:00:00
#SBATCH --job-name=PRS_PRSice_test
#SBATCH -o /home/c.c1020109/PRSice_test

start=`date +%s`
echo ${SLURM_SUBMIT_DIR}
cd $SLURM_SUBMIT_DIR


time ./PRSice_script.sh

end=`date +%s`
runtime=$((end-start))
echo $runtime

~
~
```

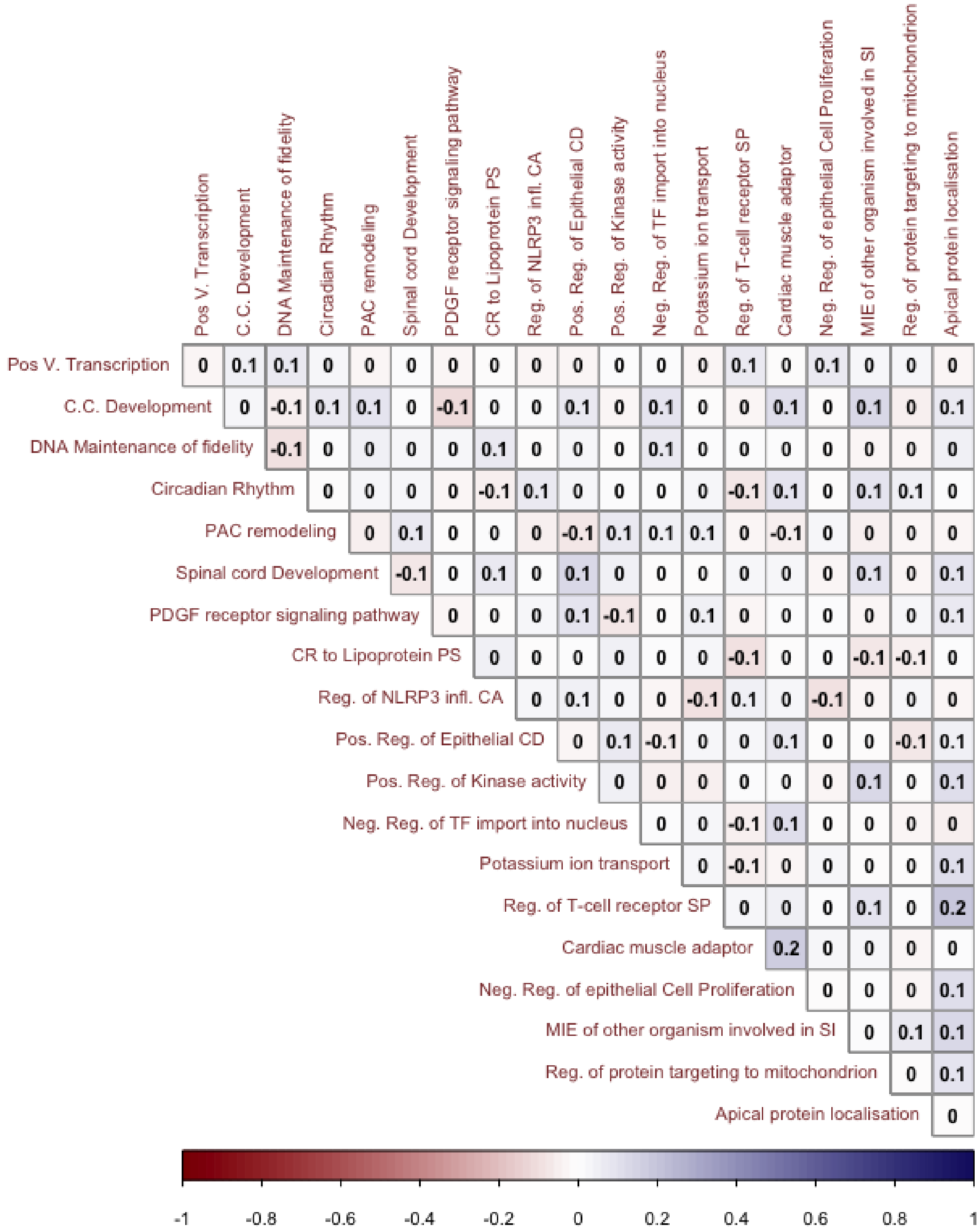
FIGURE A.2: Submission script to Hawk for the runthrough testing of PRSet. PRSet is run directly after the "time" command

A.1.2 Comparison of gene-set PRS



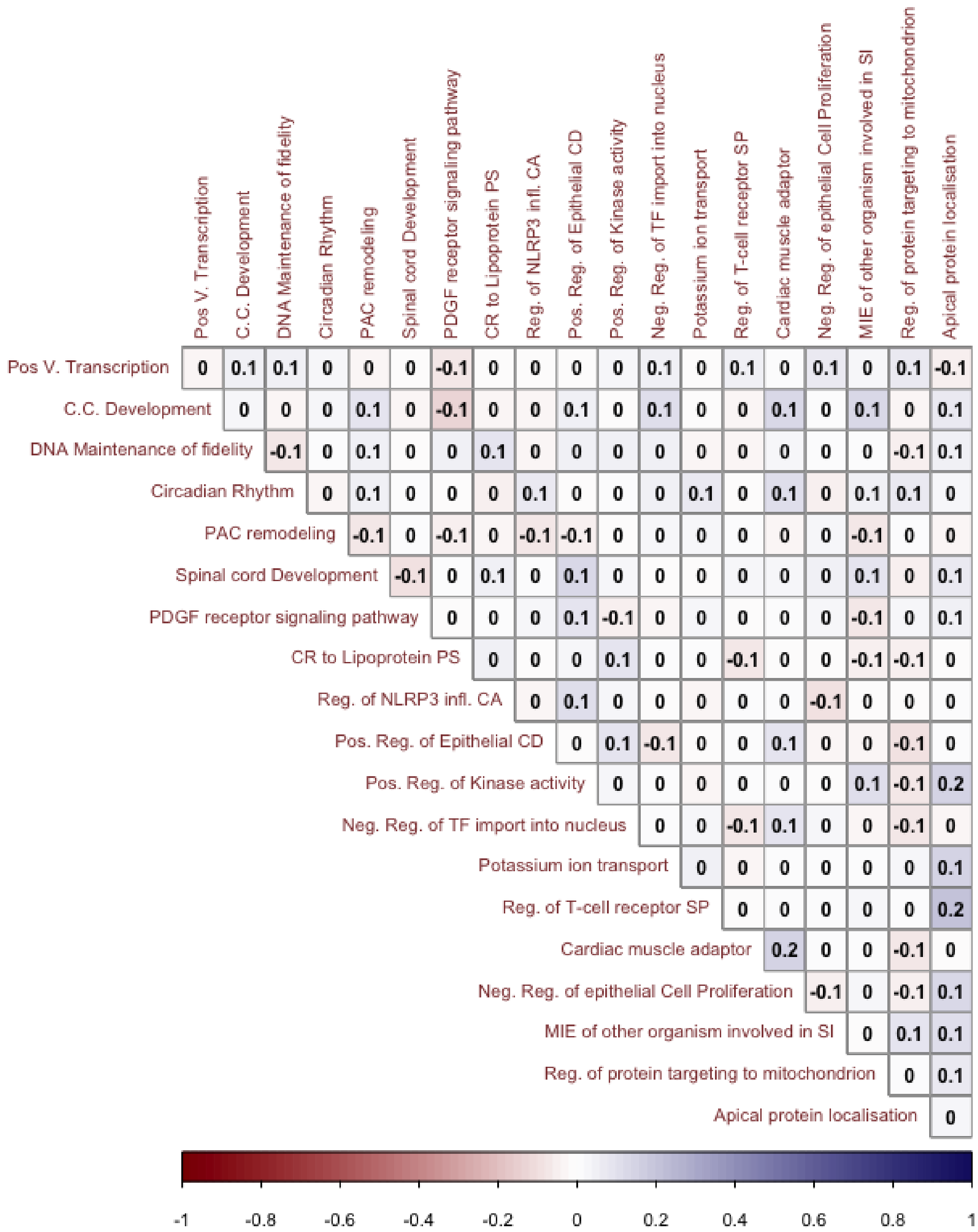
	Pos V. Transcription	C.C. Development	DNA Maintenance of fidelity	Circadian Rhythm	PAC remodeling	Spinal cord Development	PDGF receptor signaling pathway	CR to Lipoprotein PS	Reg. of NLRP3 infl. CA	Pos. Reg. of Epithelial CD	Pos. Reg. of Kinase activity	Neg. Reg. of TF import into nucleus	Potassium ion transport	Reg. of T-cell receptor SP	Cardiac muscle adaptor	Neg. Reg. of epithelial Cell Proliferation	MIE of other organism involved in SI	Reg. of protein targeting to mitochondrion	Apical protein localisation
Pos V. Transcription	0	0	0.1	0	0	0	-0.1	0	0	0	0	0	0	0	0	0.1	0	0.1	0
C.C. Development		0	0	0	0.1	0	-0.1	0	0	0	0	0.1	0	-0.1	0.1	0	0.1	0	0.1
DNA Maintenance of fidelity			-0.1	0	0.1	0	0.1	0.1	0	0.1	0	0.1	0	0	0	0	0	0	0.1
Circadian Rhythm				0	0.1	0	0	0	0.1	0	0	0	0.1	0	0.1	0	0.1	0.1	0
PAC remodeling					-0.1	0	-0.1	0	-0.1	0	0	0	0	0	0	0.1	-0.1	0	0
Spinal cord Development						-0.1	0	0.1	0	0.1	0	0	0	0	0	0	0.1	0	0.1
PDGF receptor signaling pathway							0	0	0	0.1	-0.1	0	0	0	0	0	-0.1	0	0.1
CR to Lipoprotein PS								0	0	0	0.1	0	0	-0.1	0	0	-0.1	-0.1	0
Reg. of NLRP3 infl. CA									0	0.1	0	0	0	0	0	-0.1	0	0	0
Pos. Reg. of Epithelial CD										0	0.1	-0.1	0	0	0.1	0	0	-0.1	0
Pos. Reg. of Kinase activity											0	0	0	0	0	0	0.1	-0.1	0.1
Neg. Reg. of TF import into nucleus												0	0	-0.1	0.1	0	0	-0.1	0
Potassium ion transport													0.1	-0.1	0	0	0	0	0.2
Reg. of T-cell receptor SP														0	0	0	0	0	0.2
Cardiac muscle adaptor															0.1	0	0	-0.1	0
Neg. Reg. of epithelial Cell Proliferation																0	0	-0.1	0.1
MIE of other organism involved in SI																	0	0.1	0.1
Reg. of protein targeting to mitochondrion																		0	0.1
Apical protein localisation																			0

FIGURE A.3: **Gene set wide comparison of PRS between SurPRSe and PRSet Pt = 1.** Each box signifies the correlation coefficient between the gene set PRS produced by SurPRSe and PRSet at a Pt of 1. The y-axis labels signify the identifiers of the gene set PRS from PRSet and the x-axis signifies PRS from SurPRSe. The diagonal describes direct comparisons of gene-set PRS. The legend and colour of the boxes indicates the correlation coefficient where red is negative and blue is positive. The numerical correlation coefficient is also displayed within each box.



	Pos V. Transcription	C.C. Development	DNA Maintenance of fidelity	Circadian Rhythm	PAC remodeling	Spinal cord Development	PDGF receptor signaling pathway	CR to Lipoprotein PS	Reg. of NLRP3 infl. CA	Pos. Reg. of Epithelial CD	Pos. Reg. of Kinase activity	Neg. Reg. of TF import into nucleus	Potassium ion transport	Reg. of T-cell receptor SP	Cardiac muscle adaptor	Neg. Reg. of epithelial Cell Proliferation	MIE of other organism involved in SI	Reg. of protein targeting to mitochondrion	Apical protein localisation
Pos V. Transcription	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0	0	0
C.C. Development		0	-0.1	0.1	0.1	0	-0.1	0	0	0.1	0	0.1	0	0	0.1	0	0.1	0	0.1
DNA Maintenance of fidelity			-0.1	0	0	0	0	0.1	0	0	0	0.1	0	0	0	0	0	0	0
Circadian Rhythm				0	0	0	0	-0.1	0.1	0	0	0	0	-0.1	0.1	0	0.1	0.1	0
PAC remodeling					0	0.1	0	0	0	-0.1	0.1	0.1	0.1	0	-0.1	0	0	0	0
Spinal cord Development						-0.1	0	0.1	0	0.1	0	0	0	0	0	0	0.1	0	0.1
PDGF receptor signaling pathway							0	0	0	0.1	-0.1	0	0.1	0	0	0	0	0	0.1
CR to Lipoprotein PS								0	0	0	0	0	0	-0.1	0	0	-0.1	-0.1	0
Reg. of NLRP3 infl. CA									0	0.1	0	0	-0.1	0.1	0	-0.1	0	0	0
Pos. Reg. of Epithelial CD										0	0.1	-0.1	0	0	0.1	0	0	-0.1	0.1
Pos. Reg. of Kinase activity											0	0	0	0	0	0	0.1	0	0.1
Neg. Reg. of TF import into nucleus												0	0	-0.1	0.1	0	0	0	0
Potassium ion transport													0	-0.1	0	0	0	0	0.1
Reg. of T-cell receptor SP														0	0	0	0.1	0	0.2
Cardiac muscle adaptor															0.2	0	0	0	0
Neg. Reg. of epithelial Cell Proliferation																0	0	0	0.1
MIE of other organism involved in SI																	0	0.1	0.1
Reg. of protein targeting to mitochondrion																		0	0.1
Apical protein localisation																			0

FIGURE A.4: **Gene set wide comparison of PRS between SurPRSe and PRSet at Pt = 0.2.** Each box signifies the correlation coefficient between the gene set PRS produced by SurPRSe and PRSet at a Pt of 0.2. The y-axis labels signify the identifiers of the gene set PRS from PRSet and the x-axis signifies PRS from SurPRSe. The diagonal describes direct comparisons of gene-set PRS. The legend and colour of the boxes indicates the correlation coefficient where red is negative and blue is positive. The numerical correlation coefficient is also displayed within each box.



	Pos V. Transcription	C.C. Development	DNA Maintenance of fidelity	Circadian Rhythm	PAC remodeling	Spinal cord Development	PDGF receptor signaling pathway	CR to Lipoprotein PS	Reg. of NLRP3 infl. CA	Pos. Reg. of Epithelial CD	Pos. Reg. of Kinase activity	Neg. Reg. of TF import into nucleus	Potassium ion transport	Reg. of T-cell receptor SP	Cardiac muscle adaptor	Neg. Reg. of epithelial Cell Proliferation	MIE of other organism involved in SI	Reg. of protein targeting to mitochondrion	Apical protein localisation
Pos V. Transcription	0	0.1	0.1	0	0	0	-0.1	0	0	0	0	0.1	0	0.1	0	0.1	0	0.1	-0.1
C.C. Development		0	0	0	0.1	0	-0.1	0	0	0.1	0	0.1	0	0	0.1	0	0.1	0	0.1
DNA Maintenance of fidelity			-0.1	0	0.1	0	0	0.1	0	0	0	0	0	0	0	0	0	-0.1	0.1
Circadian Rhythm				0	0.1	0	0	0	0.1	0	0	0	0.1	0	0.1	0	0.1	0.1	0
PAC remodeling					-0.1	0	-0.1	0	-0.1	-0.1	0	0	0	0	0	0	-0.1	0	0
Spinal cord Development						-0.1	0	0.1	0	0.1	0	0	0	0	0	0	0.1	0	0.1
PDGF receptor signaling pathway							0	0	0	0.1	-0.1	0	0	0	0	0	-0.1	0	0.1
CR to Lipoprotein PS								0	0	0	0.1	0	0	-0.1	0	0	-0.1	-0.1	0
Reg. of NLRP3 infl. CA									0	0.1	0	0	0	0	0	-0.1	0	0	0
Pos. Reg. of Epithelial CD										0	0.1	-0.1	0	0	0.1	0	0	-0.1	0
Pos. Reg. of Kinase activity											0	0	0	0	0	0	0.1	-0.1	0.2
Neg. Reg. of TF import into nucleus												0	0	-0.1	0.1	0	0	-0.1	0
Potassium ion transport													0	0	0	0	0	0	0.1
Reg. of T-cell receptor SP														0	0	0	0	0	0.2
Cardiac muscle adaptor															0.2	0	0	-0.1	0
Neg. Reg. of epithelial Cell Proliferation																-0.1	0	-0.1	0.1
MIE of other organism involved in SI																	0	0.1	0.1
Reg. of protein targeting to mitochondrion																		0	0.1
Apical protein localisation																			0

FIGURE A.5: **Gene set wide comparison of PRS between SurPRSe and PRSet at Pt = 0.5.** Each box signifies the correlation coefficient between the gene set PRS produced by SurPRSe and PRSet at a Pt of 0.5. The y-axis labels signify the identifiers of the gene set PRS from PRSet and the x-axis signifies PRS from SurPRSe. The diagonal describes direct comparisons of gene-set PRS. The legend and colour of the boxes indicates the correlation coefficient where red is negative and blue is positive. The numerical correlation coefficient is also displayed within each box.

A.1.3 Direct comparison of SNPs within Gene-set PRS for SurPRSe vs PRSet

Cardiac development gene-set PRS

TABLE A.1: SNPs not included in SurPRSe but incorrectly included in PRSet with an INFO score <= 0.9 within the cardiac development gene-set

SNP	CHR	BP	INFO
rs1008673	22	25994013	0.737
rs10230746	7	31112592	0.780
rs1027604	10	1335849	0.836
rs10647195	21	46543189	0.837
rs10857892	1	112063002	0.864
rs111327786	12	52357808	0.887
rs112281202	10	1522853	0.751
rs11250503	10	1460041	0.750
rs11324447	17	73967822	0.721
rs113614592	15	89384191	0.807
rs11394383	9	19125920	0.757
rs114324639	10	1695763	0.834
rs11445913	17	31923135	0.806
rs11648854	16	50287079	0.739
rs11657396	17	31816537	0.762
rs11741244	5	7405952	0.873
rs11781115	8	26681450	0.730
rs11816776	10	1622660	0.896
rs12247154	10	1335557	0.749
rs12416363	10	1659473	0.711
rs12690837	7	45717943	0.846
rs12774968	10	1404073	0.900
rs12809597	12	52356323	0.759
rs1370155	15	35082141	0.803
rs1415796	1	112049586	0.871
rs143614067	20	10649319	0.803
rs147375828	11	935794	0.836
rs149648382	17	32479946	0.760
rs1553041	17	31609251	0.888
rs17027794	1	112026449	0.748

Continued on next page

Table A.1 – continued from previous page

SNP	CHR	BP	INFO
rs17192422	17	32019411	0.754
rs17293755	10	1353714	0.787
rs17783478	17	32160611	0.729
rs181140053	22	41905819	0.774
rs187540407	17	31477741	0.870
rs1889912	17	32141378	0.850
rs191604810	10	1582959	0.795
rs199808408	17	31950259	0.880
rs200159741	10	1413979	0.771
rs201163009	2	54380401	0.813
rs2026756	8	132032301	0.889
rs2173523	10	1735559	0.720
rs2228948	2	158593905	0.729
rs2241757	2	25064079	0.894
rs2267740	7	31137003	0.757
rs2283675	22	51179000	0.720
rs2461130	7	45695331	0.799
rs2645983	17	31674764	0.882
rs26738	5	7465056	0.835
rs2779205	17	15868203	0.800
rs28479305	22	26015973	0.773
rs317325	17	32162702	0.812
rs34205959	17	32156598	0.857
rs34271210	13	25079141	0.816
rs34476044	5	148236793	0.859
rs34787014	17	40077384	0.900
rs35038759	16	4130719	0.868
rs35460688	15	59010671	0.827
rs35863579	17	73956991	0.753
rs369844976	16	71770454	0.898
rs371927	20	43263036	0.734
rs3788157	21	46510708	0.855
rs3935891	17	31819214	0.717
rs4098461	1	236920370	0.850
rs4268746	16	50281671	0.855
rs4732853	8	26610651	0.890
rs4838712	10	135077626	0.734
rs4880914	10	1746219	0.733

Continued on next page

Table A.1 – continued from previous page

SNP	CHR	BP	INFO
rs56018738	5	7447646	0.716
rs56237504	9	32399970	0.827
rs56735090	14	69407394	0.828
rs5782582	10	1575483	0.842
rs5844848	22	29764745	0.710
rs5852699	3	132039496	0.741
rs59402890	16	50289434	0.824
rs60068179	1	236920415	0.764
rs60937573	19	39147786	0.801
rs61834370	10	1502242	0.710
rs61852569	10	90736014	0.823
rs62149818	2	70958693	0.835
rs62377693	5	159366329	0.713
rs6560762	10	1745254	0.803
rs67876598	17	31834588	0.738
rs7084465	10	1524224	0.705
rs71362891	17	31855186	0.850
rs7209082	17	32345908	0.771
rs7216322	17	31745046	0.842
rs72762997	10	1493513	0.808
rs72850167	11	970847	0.869
rs7333856	13	25021211	0.775
rs7483870	11	976019	0.874
rs74887187	17	32297751	0.721
rs75490400	17	33925919	0.723
rs76534753	7	31093003	0.745
rs77050686	16	71777203	0.891
rs7726571	5	7813630	0.893
rs7816340	8	26665587	0.815
rs78372398	2	25099494	0.899
rs8064532	17	79479469	0.820
rs8078040	17	31490595	0.825
rs9511299	13	25048780	0.844
rs9616824	22	51180934	0.805
rs9910792	17	79478916	0.868

TABLE A.2: SNPs not included in SurPRSe but included in PRSet with multiple reference or alternative alleles within the cardiac development gene-set

SNP	CHR	BP	INFO	A1	A2
rs10652281	1	244577487	0.986	T	TAATA
rs10714753	5	7829665	0.963	GA	G
rs11087625	20	4222606	0.965	T	TG
rs11317030	10	76426588	0.920	CT	C
rs113310373	16	4139119	0.979	G	GT
rs11352304	16	50340229	0.931	C	CA
rs113762370	1	112079429	0.922	CCCTT	C
rs11391543	3	132046637	0.822	CAA	C
rs113983256	10	1472934	0.960	C	CA
rs11408138	14	69411880	0.925	C	CA
rs139062111	17	31570057	0.930	CAT	C
rs139115987	3	123108341	0.978	C	CAA
rs139151090	17	32240110	0.907	TC	T
rs139412870	5	148249891	0.994	C	CA
rs142469537	16	4053517	0.928	GA	G
rs145164979	10	1541198	0.986	TACTC	T
rs145164979	10	1541198	0.986	TATTC	T
rs148961416	15	89353510	0.913	T	TA
rs199822191	17	31780302	0.973	GAA	G
rs200177703	5	7551123	0.952	T	TTCCC
rs200448617	5	7699183	0.958	C	CAACAGTAA
rs201684609	3	179297239	0.919	CAT	C
rs34354539	11	10328747	0.973	TC	T
rs34400308	8	26624863	0.969	A	AT
rs34415845	9	19127037	0.955	C	CA
rs34529627	1	244609971	1.000	CTAAT	C
rs34687232	10	1427654	0.997	C	CA
rs35124782	4	100259601	0.928	T	TAC
rs35939984	10	1414519	0.973	A	AG
rs36033430	2	70908243	0.974	AC	A
rs36034937	11	125543369	0.999	T	TAC
rs373296689	5	7538978	0.912	C	CT
rs3831638	20	43279078	0.936	CT	C
rs3834026	1	203114193	0.985	CCA	C
rs55941346	15	89396984	0.960	GATAC	G

Continued on next page

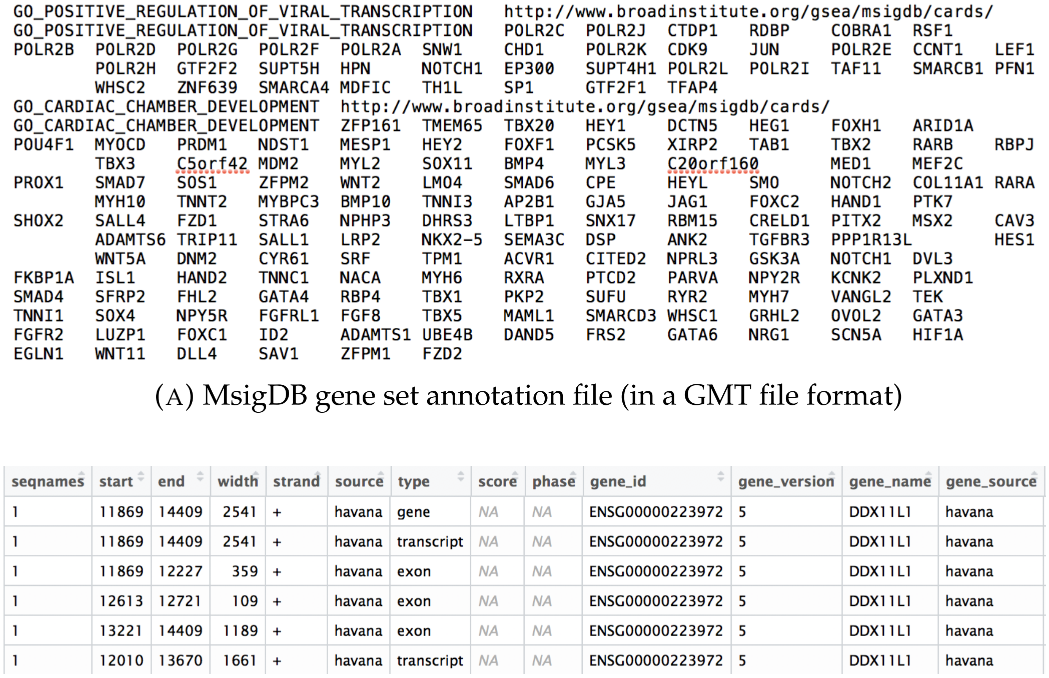
Table A.2 – continued from previous page

SNP	CHR	BP	INFO	A1	A2
rs55975325	10	1470487	0.937	A	AT
rs56278937	16	4108158	0.959	T	TA
rs57924167	10	1503867	0.950	TC	T
rs5845503	22	41891630	0.913	C	CT
rs59228224	1	229567683	0.987	TC	T
rs60661777	1	112054626	0.994	C	CA
rs70955994	4	100209493	0.953	GGT	G
rs72403882	20	4203821	0.904	T	TTTTCA
rs76397255	2	54482703	0.981	GGGGCCC	G

SNP’s that were not included due to the differing clumping procedures are as follows: rs56261301, rs2676782, rs2676723, rs11599764, rs4880906, rs10762577, rs11202921, rs7079111, rs11220182, rs79762772, rs1045476, rs2601777, rs2532019, rs2530897, rs7222081, rs8065313, rs319780, rs1497360, rs62067935, rs7218455, rs34831989, rs35246024, rs1434588, rs71379403, rs72811176, rs4795838, rs1414845, rs2214449, rs381390, rs71324410, rs2838808, rs738140, rs12997, rs11691159, rs2357954, rs72800779, rs10182300, rs2024452, rs4147544, rs7669660, rs1036174, rs1474190, rs4444938, rs1035798, rs3131300, rs12668955, rs1521470, rs1041321, rs4442231, rs72850137, rs10902256, rs61869001, rs1542874, rs884949, rs58007136, rs11250472, rs10751798, rs17293817, rs7923036, rs11250533, rs10903462, rs2813407, rs62650669, rs11250569, rs17221736, rs72764953, rs545174378, rs2813398, rs4497332, rs34547816, rs4345891, rs7907759, rs7097804, rs12149798, rs2238443, rs60392977, rs10852639, rs12444920, rs2108987, rs409963, rs11076807, rs709024, rs57361540, rs6052456, rs1809715, rs6867567, rs6555474, rs554557988, rs326174, rs2779208, rs7867814, rs3802335, rs750769, rs6753096, rs718163, rs11779546, rs741051, rs2302475, rs67679919, rs11652197, rs1579185, rs9674863, rs319783, rs319761, rs73982485, rs9890913, rs1497363, rs72821105, rs4795782, rs10853156, rs12451584, rs4795796, rs11870839, rs8075499, rs9890512, rs4795800, rs56308205, rs12453488, rs8071379, rs72827212, rs11650553, rs8069370, rs1490921, rs73986743, rs72818954, rs7212577, rs11869615, rs56177227, rs9892726, rs111813043, rs7847742, rs2026740, rs10813814, rs72828073, rs7166484, rs973009, rs893009, rs2008065, rs2076199, rs2007720, rs2177013, rs391815, rs11701974, rs1084051, rs728962, rs4785211, rs72782139, rs11675345, rs10186140, rs6545389, rs13410397, rs843645, rs17189820, rs761900, rs8003964, rs2074814, rs3821257, rs12929547, rs7359455, rs7219716, rs12467259, rs175510, rs7090732, rs16931332, rs11001115, rs7503278, rs4932426, rs12902384, rs12904298, rs61852568, rs4934433, rs1926197, rs13832, rs10030920, rs72679847, rs1693457, rs13118443, rs12035791, rs11102296, rs111453287, rs2800895, rs13443, rs66850653, rs479195, rs10097729, rs309997, rs1042713, rs3857420, rs13189358, rs12516689, rs962242, rs13001245, rs1372115, rs11741191, rs6790272, rs6686206, rs11803533, rs2282366, rs3127454, and rs3124056

A.2 PRSet

PRSet is a tool within the software PRSice (Euesden, Lewis, and O'Reilly, 2015) that aims to calculate, apply, evaluate and plot the results of gene set PRS analyses. As with SurPRSe, PRSet takes in a training data set and a testing data set as an input for the software. To obtain the definitions of the gene sets for PRS analyses, PRSet uses either an MSigDB file and a `.gtf` file (Figure A.6) or a `.bed` file (Figure A.7).



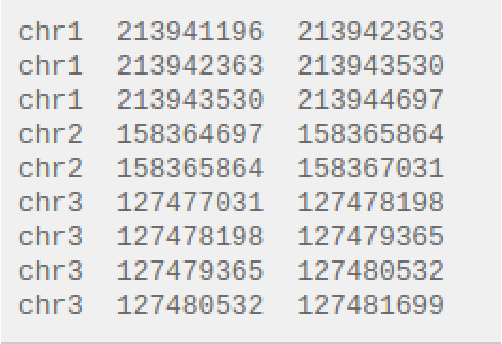
```
GO_POSITIVE_REGULATION_OF_VIRAL_TRANSCRIPTION   http://www.broadinstitute.org/gsea/msigdb/cards/
GO_POSITIVE_REGULATION_OF_VIRAL_TRANSCRIPTION   POLR2C  POLR2J  CTDP1   RDBP    COBRA1  RSF1
POLR2B  POLR2D  POLR2G  POLR2F  POLR2A  SNW1    CHD1    POLR2K  CDK9    JUN     POLR2E  CCNT1   LEF1
        POLR2H  GTF2F2  SUPT5H  HPN     NOTCH1  EP300   SUPT4H1 POLR2L  POLR2I  TAF11   SMARCB1 PFN1
        WHSC2   ZNF639  SMARCA4 MDFIC   TH1L    SP1     GTF2F1  TFAP4
GO_CARDIAC_CHAMBER_DEVELOPMENT  http://www.broadinstitute.org/gsea/msigdb/cards/
GO_CARDIAC_CHAMBER_DEVELOPMENT  ZFP161  TMEM65  TBX20   HEY1    DCTN5   HEG1    FOXH1   ARID1A
POU4F1  MYOCD   PRDM1   NDST1   MESP1   HEY2    FOXF1   PCSK5   XIRP2   TAB1    TBX2    RARB    RBPJ
        TBX3    C5orf42 MDM2    MYL2    SOX11   BMP4    MYL3    C20orf160       MED1    MEF2C
PROX1   SMAD7   SOS1    ZFPM2   WNT2    LMO4    SMAD6   CPE     HEYL    SMO     NOTCH2  COL11A1 RARA
        MYH10   TNNT2   MYBPC3  BMP10   TNNI3   AP2B1   GJA5    JAG1    FOXC2   HAND1   PTK7
SHOX2   SALL4   FZD1    STRA6   NPHP3   DHRS3   LTBP1   SNX17   RBM15   CRELD1  PITX2   MSX2    CAV3
        ADAMTS6 TRIP11  SALL1   LRP2    NKX2-5  SEMA3C  DSP     ANK2    TGFBR3  PPP1R13L        HES1
        WNT5A   DNM2    CYR61   SRF     TPM1    ACVR1   CITED2  NPRL3   GSK3A   NOTCH1  DVL3
FKBP1A  ISL1    HAND2   TNNC1   NACA    MYH6    RXRA    PTCD2   PARVA   NPY2R   KCNK2   PLXND1
SMAD4   SFRP2   FHL2    GATA4   RBP4    TBX1    PKP2    SUFU    RYR2    MYH7    VANGL2  TEK
TNNI1   SOX4    NPY5R   FGFRL1  FGF8    TBX5    MAML1   SMARCD3 WHSC1   GRHL2   OVOL2   GATA3
FGFR2   LUZP1   FOXC1   ID2     ADAMTS1 UBE4B   DAND5   FRS2    GATA6   NRG1    SCN5A   HIF1A
EGLN1   WNT11   DLL4    SAV1    ZFPM1   FZD2
```

(A) MsigDB gene set annotation file (in a GMT file format)

seqnames	start	end	width	strand	source	type	score	phase	gene_id	gene_version	gene_name	gene_source
1	11869	14409	2541	+	havana	gene	NA	NA	ENSG00000223972	5	DDX11L1	havana
1	11869	14409	2541	+	havana	transcript	NA	NA	ENSG00000223972	5	DDX11L1	havana
1	11869	12227	359	+	havana	exon	NA	NA	ENSG00000223972	5	DDX11L1	havana
1	12613	12721	109	+	havana	exon	NA	NA	ENSG00000223972	5	DDX11L1	havana
1	13221	14409	1189	+	havana	exon	NA	NA	ENSG00000223972	5	DDX11L1	havana
1	12010	13670	1661	+	havana	transcript	NA	NA	ENSG00000223972	5	DDX11L1	havana

(B) `.gtf` file format, include supplementary link to script and explanation that the raw format is tricky to read into R.

FIGURE A.6: **PRSet gene set input option one.**



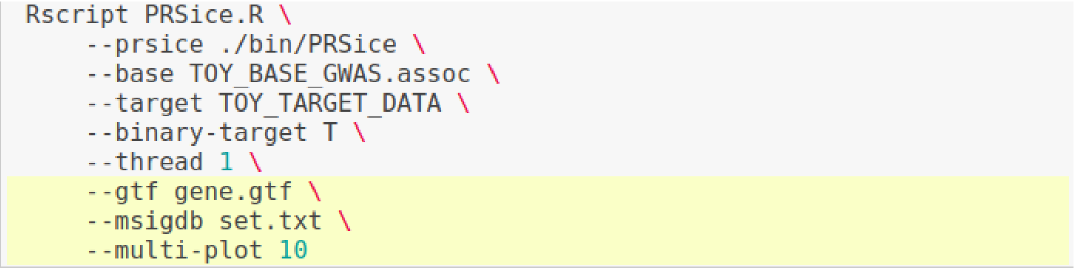
```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
chr2  158365864  158367031
chr3  127477031  127478198
chr3  127478198  127479365
chr3  127479365  127480532
chr3  127480532  127481699
```

FIGURE A.7: **PRSice2 gene-set input option two.** The input must be defined as the `.bed` file format described according to the Ensembl genome browser. The first column describes the chromosome or scaffold identifier. The second column describes the first feature in standard chromosomal co-ordinates. The third column describes the second feature in standard chromosomal co-ordinates. In respect to PRSice2, the name assigned to the `.bed` file is interpreted as the gene-set for the analysis. Further information on `.bed` files can be found at `https://www.ensembl.org/info/website/upload/bed.html`

PRSet is a complex and comprehensive command line tool, designed to encompass all the necessary tools required to perform a PRS analysis from the QC of the input data sets to the regression analysis of the PRS with the trait of interest. Separate from SurPRSe, it is able to obtain the optimal Pt for each PRS and adjust the method for which the PRS is calculated.

PRSet is written in the R statistical software and C++ programming languages. The R script controls the higher level functions of PRSet while the C++ code controls the lower level functions including memory allocation and computational processing to create very efficient production of PRS. PRSet is designed to be run as an executable on a personal desktop, but can easily be adapted to run in a supercomputer environment.

As stated in Figure A.8, PRSet is similar to SurPRSe in that a header script (this time an R script) is used to run the program, while instead of a configuration script, the control of the analysis is listed as arguments to the header R script.



```
Rscript PRSice.R \
    --prsice ./bin/PRSice \
    --base TOY_BASE_GWAS.assoc \
    --target TOY_TARGET_DATA \
    --binary-target T \
    --thread 1 \
    --gtf gene.gtf \
    --msigdb set.txt \
    --multi-plot 10
```

FIGURE A.8: **PRSet example.** Taken from `https://choishingwan.github.io/PRSice/quick_start_prset/`

If the regression analyses is performed, PRSet is able to visualise the results using the −−multi−plot <N> argument. Further details can be found in Section 3.4.4.

B Appendix for Chapter 4

B.1 Genome-wide PRS and subcortical brain region sizes

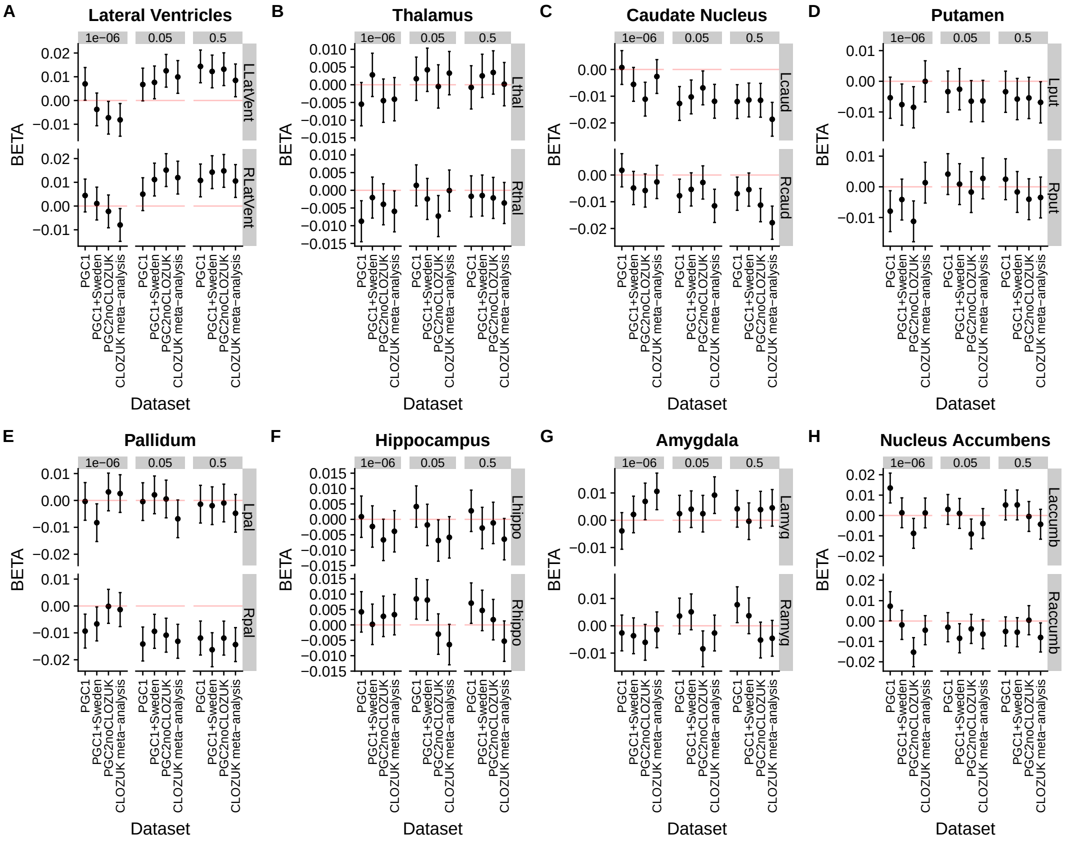


FIGURE B.1: **Comparisons of associations of polygenic risk scores to subcortical brain volumes in UK Biobank samples**. **A-H** Titles of each plot indicate the subcortical brain region. The three subtitles of each plot indicate the Pt. The top section of each plot is the left hemisphere of the specified brain region (indicated by the prefix 'L') and the bottom section is the right hemisphere of the specified brain region (indicated by the prefix 'R'). The x-axis indicates the training data-set used for the PRS. At each Pt the GWA studies increase in power from left to right. The BETA title on the x-axis indicated the standardised coefficient of the linear regression. The red line indicates a null value across each plot. Error bars displayed here are standard error bars. If the error bar does not cross the null, it does **NOT** indicate that the association within the linear regression model was significant.

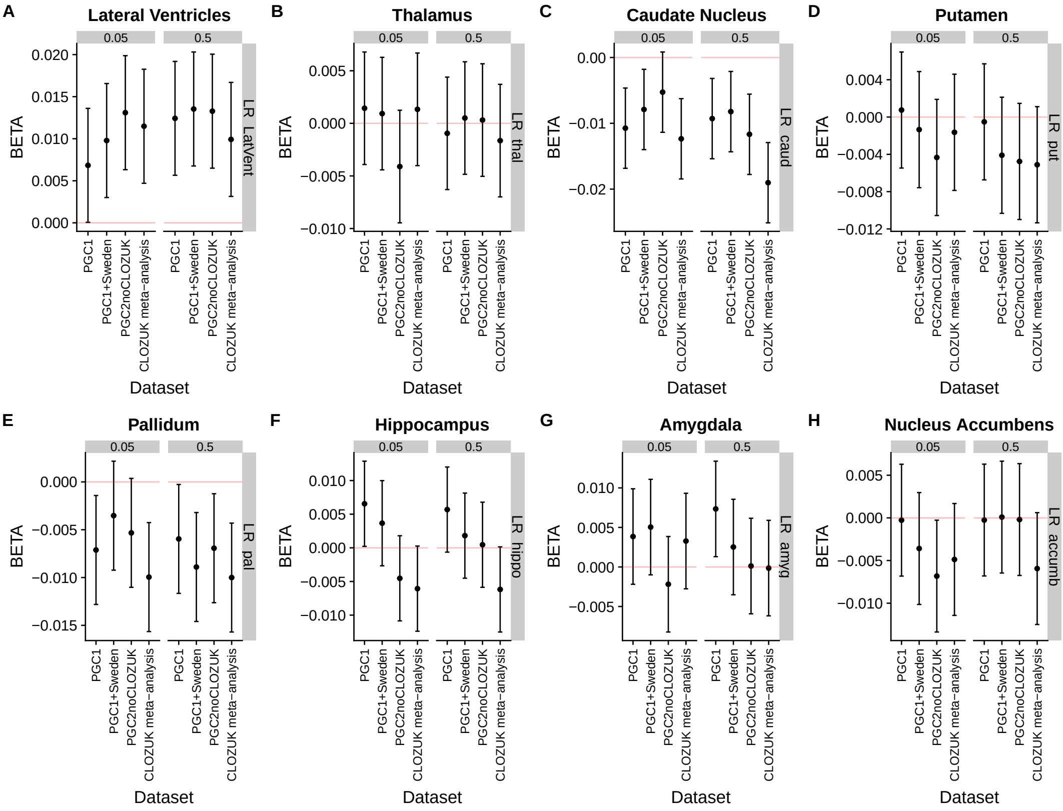


FIGURE B.2: **Comparisons of associations of polygenic risk scores to subcortical brain volumes in UK Biobank samples**. **A-H** Titles of each plot indicate the subcortical brain region. The three subtitles of each plot indicate the Pt. The 'LR' notation on the right side of the y-axis indicates that the brain region has been averaged between the left and right hemisphere. The x-axis indicates the training data-set used for the PRS. At each Pt the GWA studies increase in power from left to right. The BETA title on the x-axis indicated the standardised coefficient of the linear regression. The red line indicates a Null value across each plot. Error bars displayed here are standard error bars. If the error bar does not cross the null, it does **NOT** indicate that the association within the linear regression model was significant.

B.2 Gene-set PRS and subcortical brain region sizes

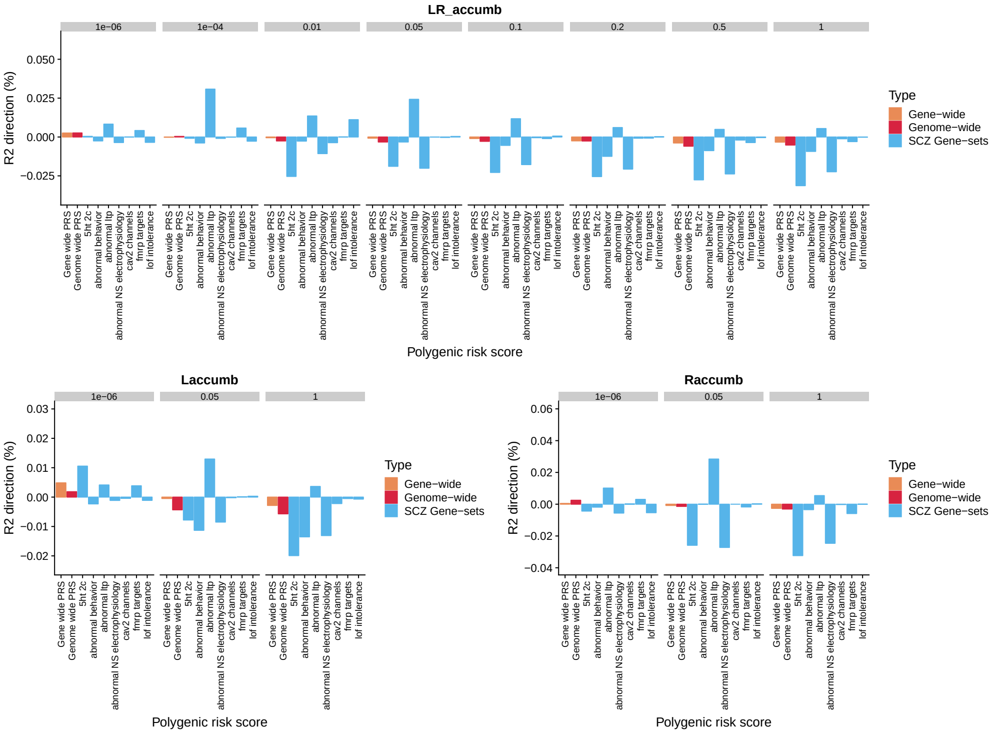


FIGURE B.3: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on nucleus accumbens brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the nucleus accumbens. The subheadings to each plot are the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction.

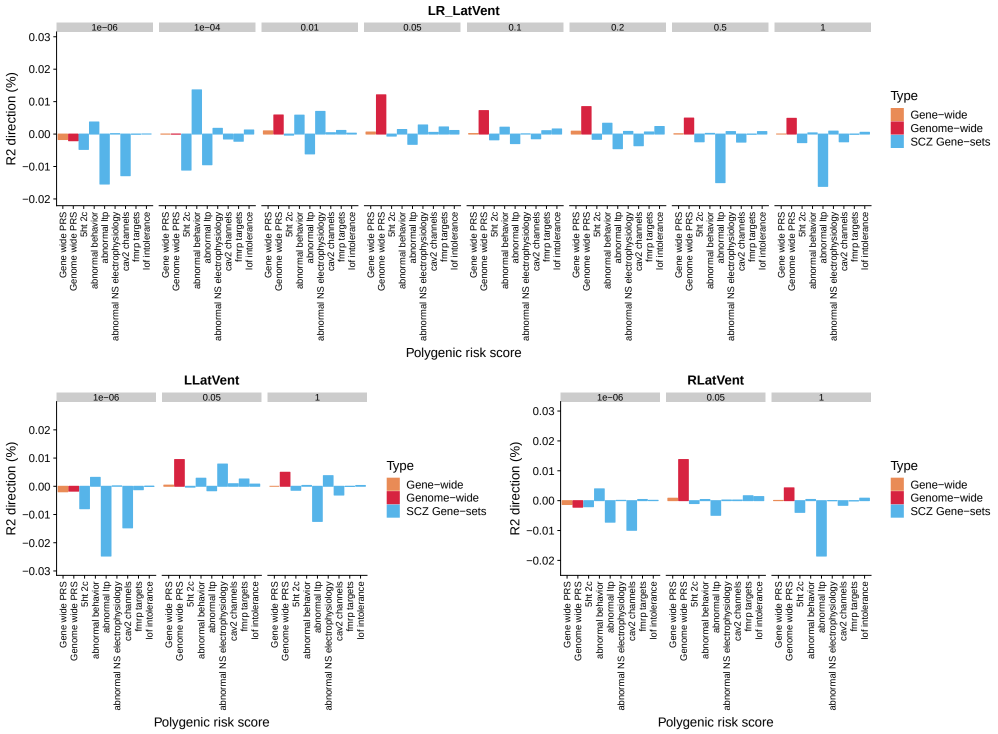


FIGURE B.4: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on lateral ventricles brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the lateral ventricles. The subheadings to each plot are the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction.

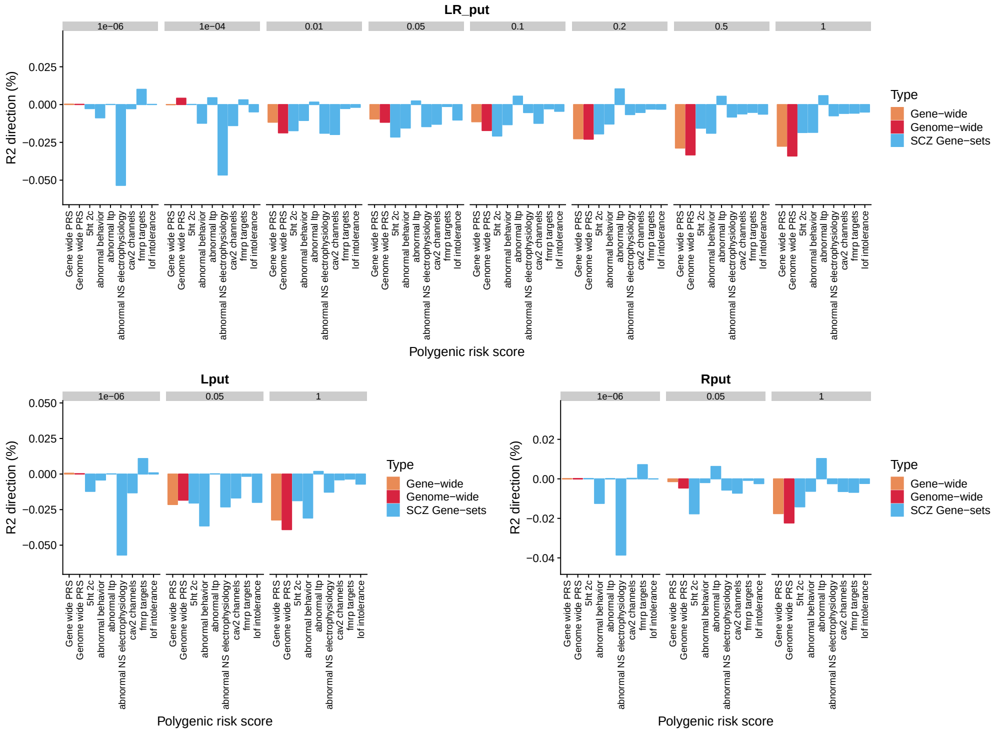


FIGURE B.5: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on putamen brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the putamen. The subheadings to each plot are the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction.

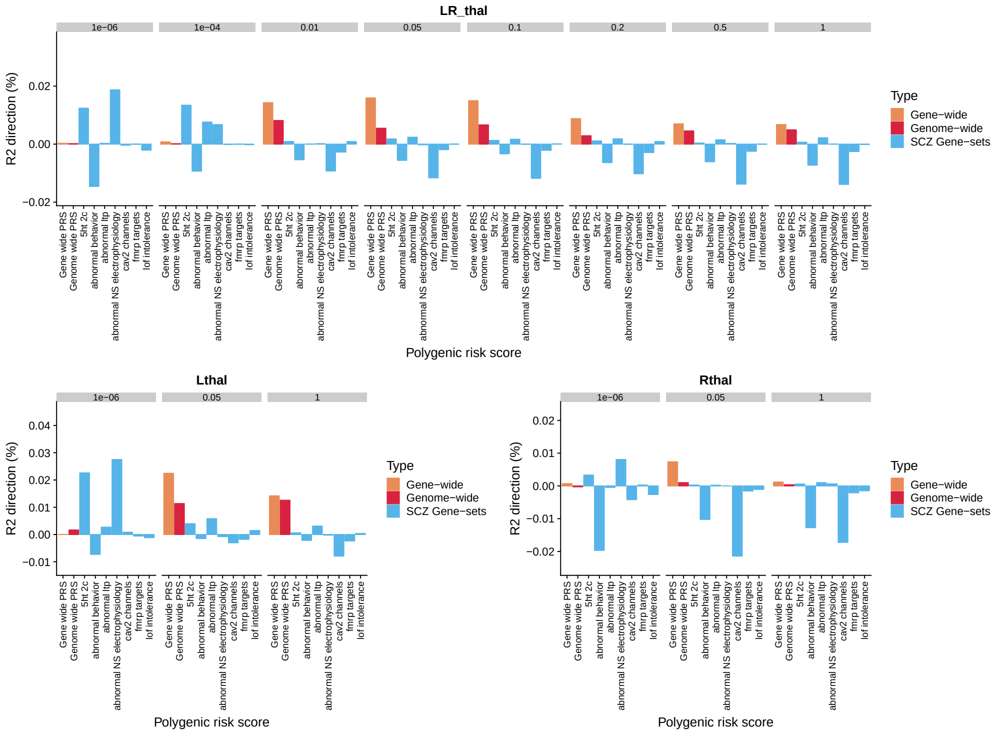


FIGURE B.6: **Comparisons of the variation explained by the genome-wide, gene-centric and gene-set PRS on Thalamus brain volume.** The left hemisphere is the bottom left plot and the right hemisphere is the bottom right plot. The top plot is the averaged left and right hemisphere of the thalamus. The subheadings to each plot are the Pt of the PRS. The y-axis signifies the value of the r-squared in the form of a percentage with the direction of the standardised regression coefficient incorporated in. All gene-set PRS appear in blue, the genome-wide PRS is red and gene-wide PRS is orange. If the FDR p-value is noted down, that gene-set PRS passed multiple testing correction.

C Appendix for Chapter 5

C.1 Gene-set analysis

TABLE C.1: Number of genes per gene-set

	Significance thresholds								
PRS	5e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1
Abnormal behavior	34	67	282	1523	3167	4341	5938	8760	10779
FMRP targets	36	65	193	948	1856	2544	3399	4876	5896
CNS Neuron Differentiation	8	13	49	190	362	498	679	1000	1184
Neurogenesis	39	61	245	1216	2422	3343	4643	6803	8269
Neuron differentiation	23	35	154	802	1639	2276	3164	4655	5639
Positive regulation of NS development	13	22	99	438	892	1218	1692	2411	2955
Regulation of NS development	26	40	153	724	1424	1932	2693	3896	4792
Regulation of synapse structure or activity	7	14	72	302	566	746	1035	1486	1807
IQ collated set	41	67	267	1304	2611	3605	4992	7316	8912
LoF intolerant genes	67	125	472	2375	4798	6513	8822	12851	15637
SCZ collated set	80	157	599	3195	6602	9075	12426	18257	22298
Gene-centric	115	238	957	6059	13206	18694	26158	38849	47563
Genome-wide	153	325	1347	9525	21147	29887	42025	62228	75696

C.2 Polygenic Risk Scores

C.2.1 Schizophrenia

TABLE C.2: Regression results of IQ PRS with congnition in schizophrenia

PRS	BETA	SE	tvalue	P value	R squared	lower CI	upper CI	Significance_thresholds	Type	FDR p
extended_geneset_SCORE_abnormal_behavior_5e-08	0.01	0.05	0.31	0.75	0.00	-0.08	0.11	5e-08	SCZ Gene-sets	0.77
extended_geneset_SCORE_abnormal_behavior_0.05	-0.05	0.05	-1.08	0.28	0.00	-0.14	0.04	0.05	SCZ Gene-sets	0.44
extended_geneset_SCORE_abnormal_behavior_1	-0.07	0.05	-1.57	0.12	0.00	-0.17	0.02	1	SCZ Gene-sets	0.25
extended_geneset_SCORE_GO_NEUROGENESIS_5e-08	-0.03	0.05	-0.62	0.53	0.00	-0.12	0.06	5e-08	IQ Gene-sets	0.63
extended_geneset_SCORE_GO_NEUROGENESIS_0.05	-0.13	0.05	-2.86	0.00	0.01	-0.22	-0.04	0.05	IQ Gene-sets	0.04
extended_geneset_SCORE_GO_NEUROGENESIS_1	-0.12	0.05	-2.46	0.01	0.01	-0.21	-0.02	1	IQ Gene-sets	0.07
extended_geneset_SCORE_FMRP_targets_5e-08	0.03	0.05	0.69	0.49	0.00	-0.06	0.13	5e-08	SCZ Gene-sets	0.62
extended_geneset_SCORE_FMRP_targets_0.05	0.06	0.05	1.35	0.18	0.00	-0.03	0.16	0.05	SCZ Gene-sets	0.35
extended_geneset_SCORE_FMRP_targets_1	0.05	0.05	1.09	0.28	0.00	-0.04	0.14	1	SCZ Gene-sets	0.44
extended_geneset_SCORE_GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_5e-08	-0.06	0.05	-1.28	0.20	0.00	-0.15	0.03	5e-08	IQ Gene-sets	0.36
extended_geneset_SCORE_GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_0.05	-0.13	0.05	-2.72	0.01	0.01	-0.22	-0.03	0.05	IQ Gene-sets	0.04
extended_geneset_SCORE_GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_1	-0.09	0.05	-1.88	0.06	0.00	-0.18	0.00	1	IQ Gene-sets	0.20
extended_geneset_SCORE_GO_NEURON_DIFFERENTIATION_5e-08	0.03	0.05	0.59	0.56	0.00	-0.07	0.12	5e-08	IQ Gene-sets	0.64
extended_geneset_SCORE_GO_NEURON_DIFFERENTIATION_0.05	-0.08	0.05	-1.62	0.11	0.00	-0.17	0.02	0.05	IQ Gene-sets	0.25
extended_geneset_SCORE_GO_NEURON_DIFFERENTIATION_1	-0.07	0.05	-1.58	0.11	0.00	-0.17	0.02	1	IQ Gene-sets	0.25
extended_geneset_SCORE_GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION_5e-08	0.05	0.05	1.12	0.26	0.00	-0.04	0.14	5e-08	IQ Gene-sets	0.44
extended_geneset_SCORE_GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION_0.05	-0.04	0.05	-0.79	0.43	0.00	-0.13	0.05	0.05	IQ Gene-sets	0.59
extended_geneset_SCORE_GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION_1	-0.02	0.05	-0.32	0.75	0.00	-0.11	0.08	1	IQ Gene-sets	0.77
extended_geneset_SCORE_Lek2015_LoFintolerant_90_5e-08	0.03	0.05	0.65	0.52	0.00	-0.06	0.12	5e-08	SCZ Gene-sets	0.63
extended_geneset_SCORE_Lek2015_LoFintolerant_90_0.05	-0.01	0.05	-0.19	0.85	0.00	-0.10	0.08	0.05	SCZ Gene-sets	0.85
extended_geneset_SCORE_Lek2015_LoFintolerant_90_1	0.02	0.05	0.46	0.64	0.00	-0.07	0.11	1	SCZ Gene-sets	0.72
extended_geneset_SCORE_GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_5e-08	-0.09	0.05	-1.98	0.05	0.00	-0.19	-0.00	5e-08	IQ Gene-sets	0.17
extended_geneset_SCORE_GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_0.05	-0.13	0.05	-2.83	0.00	0.01	-0.22	-0.04	0.05	IQ Gene-sets	0.04
extended_geneset_SCORE_GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_1	-0.11	0.05	-2.34	0.02	0.01	-0.20	-0.02	1	IQ Gene-sets	0.08
extended_geneset_SCORE_GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY_5e-08	-0.08	0.05	-1.61	0.11	0.00	-0.17	0.02	5e-08	IQ Gene-sets	0.25
extended_geneset_SCORE_GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY_0.05	-0.13	0.05	-2.85	0.00	0.01	-0.22	-0.04	0.05	IQ Gene-sets	0.04
extended_geneset_SCORE_GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY_1	-0.11	0.05	-2.38	0.02	0.01	-0.20	-0.02	1	IQ Gene-sets	0.08
extended_geneset_SCORE_IQ_Superset_5e-08	-0.04	0.05	-0.87	0.38	0.00	-0.13	0.05	5e-08	Collated IQ set	0.55
extended_geneset_SCORE_IQ_Superset_0.05	-0.14	0.05	-3.04	0.00	0.01	-0.23	-0.05	0.05	Collated IQ set	0.04
extended_geneset_SCORE_SCZ_Superset_5e-08	0.03	0.05	0.73	0.47	0.00	-0.06	0.13	5e-08	Collated SCZ set	0.61
extended_geneset_SCORE_IQ_Superset_1	-0.13	0.05	-2.74	0.01	0.01	-0.22	-0.04	1	Collated IQ set	0.04
extended_geneset_SCORE_SCZ_Superset_0.05	-0.03	0.05	-0.72	0.47	0.00	-0.13	0.06	0.05	Collated SCZ set	0.61
extended_geneset_SCORE_SCZ_Superset_1	-0.02	0.05	-0.34	0.73	0.00	-0.11	0.08	1	Collated SCZ set	0.77
extended.genic.genome_SCORE_whole_genome_5e-08	-0.07	0.05	-1.41	0.16	0.00	-0.16	0.03	5e-08	Whole genome	0.32
extended.genic.genome_SCORE_whole_genome_0.05	-0.08	0.05	-1.73	0.08	0.00	-0.17	0.01	0.05	Whole genome	0.23
extended.genic.genome_SCORE_whole_genome_1	-0.05	0.05	-1.00	0.32	0.00	-0.14	0.04	1	Whole genome	0.48
All.genome_SCORE_whole_genome_5e-08	-0.08	0.05	-1.79	0.07	0.00	-0.17	0.01	5e-08	Whole genome	0.22
All.genome_SCORE_whole_genome_0.05	-0.11	0.05	-2.44	0.01	0.01	-0.21	-0.02	0.05	Whole genome	0.07
All.genome_SCORE_whole_genome_1	-0.06	0.05	-1.31	0.19	0.00	-0.15	0.03	1	Whole genome	0.36

C.2.2 IQ

TABLE C.3: Regression results of IQ PRS with congnition in schizophrenia.

PRS	BETA	SE	tvalue	P value	R squared	lower CI	upper CI	Significance_thresholds	Type	FDR p
extended_geneset_SCORE_abnormal_behavior_5e-08	-0.02	0.05	-0.4	0.66	2e-04	-0.11	0.07	5e-08	SCZ Gene-sets	0.74
extended_geneset_SCORE_abnormal_behavior_0.05	0.17	0.05	3.85	1e-04	0.02	0.09	0.26	0.05	SCZ Gene-sets	1.72e-03
extended_geneset_SCORE_abnormal_behavior_1	0.14	0.05	3.01	2e-03	1e-04	0.05	0.2	1	SCZ Gene-sets	1.69e-02
extended_geneset_SCORE_GO_NEUROGENESIS_5e-08	0.04	0.05	0.96	0.33	1e-03	0.05	0.13	5e-08	IQ Gene-sets	0.48
extended_geneset_SCORE_GO_NEUROGENESIS_0.05	0.22	0.04	5.11	3.96e-07	0.03	0.14	0.32	0.05	IQ Gene-sets	1.51e-05
extended_geneset_SCORE_GO_NEUROGENESIS_1	0.17	0.04	4.01	6.64e-06	0.02	0.09	0.27	1	IQ Gene-sets	5.4e-04
extended_geneset_SCORE_FMRP_targets_5e-08	0.04	0.04	0.89	0.37	8.80e-04	-0.05	0.13	5e-08	SCZ Gene-sets	0.49
extended_geneset_SCORE_FMRP_targets_0.05	0.13	0.05	2.86	4.33e-04	9.03e-03	0.04	0.21	0.05	SCZ Gene-sets	1.77e-03
extended_geneset_SCORE_FMRP_targets_1	0.13	0.04	3.08	2.09e-03	0.01	0.05	0.22	1	SCZ Gene-sets	1.40e-03
extended_geneset_SCORE_GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_5e-08	-0.03	0.05	-0.61	0.54	4.19e-04	-0.12	0.06	5e-08	IQ Gene-sets	0.64
extended_geneset_SCORE_GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_0.05	0.16	0.04	3.55	4.12e-04	0.01	0.07	0.24	0.05	IQ Gene-sets	8.73e-03
extended_geneset_SCORE_GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_1	0.13	0.04	2.98	2.96e-03	9.79e-03	0.05	0.22	1	IQ Gene-sets	1.17e-02
extended_geneset_SCORE_GO_NEURON_DIFFERENTIATION_5e-08	0.02	0.05	0.52	0.60	3.25e-04	-0.07	0.12	5e-08	IQ Gene-sets	0.68
extended_geneset_SCORE_GO_NEURON_DIFFERENTIATION_0.05	0.20	0.05	4.28	2.16e-05	0.02	0.11	0.29	0.05	IQ Gene-sets	7.60-05
extended_geneset_SCORE_GO_NEURON_DIFFERENTIATION_1	0.17	0.05	3.68	2.49e-04	0.02	0.08	0.27	1	IQ Gene-sets	7.01e-04
extended_geneset_SCORE_GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION_5e-08	0.01	0.05	-0.25	0.80	7.67e-05	-0.11	0.08	5e-08	IQ Gene-sets	8.21e-01
extended_geneset_SCORE_GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION_0.05	0.03	0.05	0.72	0.47	6.01e-04	-0.06	0.13	0.05	IQ Gene-sets	5.72e-01
extended_geneset_SCORE_GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION_1	0.04	0.05	0.88	0.38	9.20e-04	-0.05	0.14	1	IQ Gene-sets	0.49
extended_geneset_SCORE_Lek2015_LoFintolerant_90_5e-08	0.02	0.05	0.40	0.69	1.92e-03	-0.07	0.11	5e-08	SCZ Gene-sets	0.74
extended_geneset_SCORE_Lek2015_LoFintolerant_90_0.05	0.20	0.05	4.38	1.36e-05	0.02	0.11	0.08	0.29	SCZ Gene-sets	3.02e-06
extended_geneset_SCORE_Lek2015_LoFintolerant_90_1	0.21	0.05	4.66	3.72e-06	0.03	0.12	0.30	1	SCZ Gene-sets	1.18e-06
extended_geneset_SCORE_GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_5e-08	0.07	0.05	1.35	0.18	2.7e-03	-0.03	0.17	5e-08	IQ Gene-sets	0.27
extended_geneset_SCORE_GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_0.05	0.21	0.05	4.12	4.27e-05	0.02	0.11	0.30	0.05	IQ Gene-sets	3.59e-04
extended_geneset_SCORE_GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT_1	0.16	0.05	3.11	1.98e-03	0.01	0.06	0.26	1	IQ Gene-sets	1.12e-02
extended_geneset_SCORE_GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY_5e-08	-0.02	0.05	-0.39	0.70	1.70e-04	-0.11	0.07	5e-08	IQ Gene-sets	0.74
extended_geneset_SCORE_GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY_0.05	0.13	0.04	2.93	3.52e-03	0.019.44e-03	0.04	0.22	0.05	IQ Gene-sets	1.13e-02
extended_geneset_SCORE_GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY_1	-0.11	0.04	2.54	0.01	7.11e-03	0.026	0.20	1	IQ Gene-sets	2.25e-02
extended_geneset_SCORE_IQ_Superset_5e-08	0.04	0.05	0.83	0.41	7.70e-04	-0.05	1.30	5e-08	Collated IQ set	0.52
extended_geneset_SCORE_IQ_Superset_0.05	0.22	0.04	4.95	9.01e-07	0.03	0.13	0.31	0.05	Collated IQ set	2.79e-05
extended_geneset_SCORE_IQ_Superset_1	0.17	0.04	3.85	1.29e-04	0.02	0.08	0.26	1	Collated IQ set	8.79e-04
extended_geneset_SCORE_SCZ_Superset_5e-08	0.04	0.05	0.89	0.37	8.81e-04	-0.05	0.13	5e-08	Collated SCZ set	0.49
extended_geneset_SCORE_SCZ_Superset_0.05	0.26	0.04	5.86	6.82e-09	0.04	0.17	0.35	0.05	Collated SCZ set	5.78e-08
extended_geneset_SCORE_SCZ_Superset_1	0.26	0.04	-5.84	7.81e-09	0.04	0.17	0.35	1	Collated SCZ set	5.78e-08
extended.genic.genome_SCORE_whole_genome_5e-08	0.10	0.05	2.24	0.03	5.55e-03	0.01	0.19	5e-08	Whole genome	5.68e-02
extended.genic.genome_SCORE_whole_genome_0.05	0.31	0.04	7.03	4.54e-12	0.05	0.22	0.40	0.05	Whole genome	9.85e-11
extended.genic.genome_SCORE_whole_genome_1	0.30	0.04	6.70	3.93e-11	0.05	0.21	0.38	1	Whole genome	2.31e-10
All.genome_SCORE_whole_genome_5e-08	0.11	0.05	2.47	0.01	6.74e-03	0.02	0.20	5e-08	Whole genome	2.82e-02
All.genome_SCORE_whole_genome_0.05	0.32	0.05	7.10	2.84e-12	0.05	0.23	0.41	0.05	Whole genome	2.31e-10
All.genome_SCORE_whole_genome_1	0.29	0.04	6.54	1.13e-10	0.05	0.21	0.38	1	Whole genome	1.5e-09

Bibliography

Ahmed, Zia, Waleed Arshad, and Waqas Mahmood (2018). "Preference in using Agile Development with Larger Team Size". In: *International Journal of Advanced Computer Science and Applications*.

Allardyce, Judith et al. (2017). "Psychosis and the level of mood incongruence in Bipolar Disorder are related to genetic liability for Schizophrenia". In: *doi.org*, p. 160119. DOI: 10.1101/160119. URL: http://www.biorxiv.org/content/early/2017/09/19/160119?{\%}3Fcollection=.

Allardyce, Judith et al. (2018). "Association Between Schizophrenia-Related Polygenic Liability and the Occurrence and Level of Mood-Incongruent Psychotic Symptoms in Bipolar Disorder". In: *JAMA Psychiatry* 75.1, p. 28. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2017.3485. URL: http://www.ncbi.nlm.nih.gov/pubmed/29167880http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5833541http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/jamapsychiatry.2017.3485.

Alnæs, Dag and Lars T. Westlye (2019). "Factors Associated With Brain Heterogeneity in Schizophrenia—Reply". In: *JAMA Psychiatry*. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2019.1855. URL: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2738764.

Alnæs, Dag et al. (2019). "Brain Heterogeneity in Schizophrenia and Its Association With Polygenic Risk". In: *JAMA Psychiatry* 76.7, p. 739. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2019.0257. URL: http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/jamapsychiatry.2019.0257.

Altschul, Stephen F et al. (1990). "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3, pp. 403–410. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602.

American Psychiatric Association (2000). "Diagnostic and statistical manual of mental disorders". In: *American Psychiatric Association* 4th ed., text rev.

Aniba, Mohamed Radhouene, Olivier Poch, and Julie D Thompson (2010). "Issues in bioinformatics benchmarking: the case study of multiple sequence alignment". In: *Nucleic Acids Research* 38.21, pp. 7353–7363. ISSN: 0305-1048. DOI: 10.1093/nar/gkq625. URL: https://doi.org/10.1093/nar/gkq625.

Ardlie, Kristin G, David S Deluca, and Ayellet V Segre (2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." eng. In: *Science (New York, N.Y.)* 348.6235, pp. 648–660. ISSN: 1095-9203 (Electronic). DOI: 10.1126/science.1262110.

Ashburner, Michael et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: http://www.ncbi.nlm.nih.gov/pubmed/10802651http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3037419http://www.nature.com/doifinder/10.1038/75556.

Assmann, Gerd, Paul Cullen, and Helmut Schulte (2002). "Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study." eng. In: *Circulation* 105.3, pp. 310–315. ISSN: 1524-4539 (Electronic). DOI: 10.1161/hc0302.102575.

Auton, Adam and Gonçalo R. Abecasis (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393. URL: http://www.nature.com/articles/nature15393.

Auton, Adam et al. (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393. URL: http://www.nature.com/doifinder/10.1038/nature15393.

Avramopoulos, D (2018). "Recent Advances in the Genetics of Schizophrenia". In: *Molecular Neuropsychiatry* 4.1, pp. 35–51. ISSN: 2296-9209. DOI: 10.1159/000488679. URL: https://www.karger.com/DOI/10.1159/000488679.

Baggen, Robert et al. (2011). "Standardized code quality benchmarking for improving software maintainability". In: *Software Quality Journal* 20, pp. 1–21. DOI: 10.1007/s11219-011-9144-9.

Baker, Emily et al. (2018). "POLARIS: Polygenic LD-adjusted risk score approach for set-based analysis of GWAS data." In: *Genetic epidemiology* 42.4, pp. 366–377. ISSN: 1098-2272. DOI: 10.1002/gepi.22117. URL: http://www.ncbi.nlm.nih.gov/pubmed/29532500http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6001515.

Banaschewski, T et al. (2015). "Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways". In: *Nature Neuroscience* 18.2, pp. 199–209. DOI: 10.1038/nn.3922. URL: http://www.nature.com/reprints/index.html..

Bécamel, Carine et al. (2002). "Synaptic multiprotein complexes associated with 5-HT(2C) receptors: a proteomic approach". eng. In: *The EMBO journal* 21.10, pp. 2332–2342. ISSN: 0261-4189. DOI: `10.1093/emboj/21.10.2332`. URL: `https://pubmed.ncbi.nlm.nih.gov/12006486https://www.ncbi.nlm.nih.gov/pmc/articles/PMC126011/`.

Benjamini, Yoav and Yosef Hochberg (1995). *Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing*. Vol. 57, pp. 289–300. DOI: `10.2307/2346101`.

Bergen, Sarah E et al. (2019). "Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia." eng. In: *The American journal of psychiatry* 176.1, pp. 29–35. ISSN: 1535-7228 (Electronic). DOI: `10.1176/appi.ajp.2018.17040467`.

Blake, Judith A et al. (2003). "MGD: the Mouse Genome Database." In: *Nucleic acids research* 31.1, pp. 193–5. ISSN: 1362-4962. URL: `http://www.ncbi.nlm.nih.gov/pubmed/12519980http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC165494`.

Bleuler, Eugen (1911). "Dementia Praecox oder Gruppe der Schizophrenien". In: *Deuticke*.

Bleuler, Eugine and Carl Jung (1908). "Komplexe und Krankheitsursachen bei Dementia Praecox." In: *Zentralblatt für Nervenheilkunde und Psychiatrie* 31, pp. 220–227.

Bohlken, Marc M et al. (2016). "Genetic Variation in Schizophrenia Liability is Shared With Intellectual Ability and Brain Structure." eng. In: *Schizophrenia bulletin* 42.5, pp. 1167–1175. ISSN: 1745-1701 (Electronic). DOI: `10.1093/schbul/sbw034`.

Braber, Anouk den et al. (2013). "Heritability of subcortical brain measures: A perspective for future genome-wide association studies". In: *NeuroImage* 83, pp. 98–102. ISSN: 1053-8119. DOI: `10.1016/J.NEUROIMAGE.2013.06.027`. URL: `http://www.sciencedirect.com/science/article/pii/S1053811913006642?via{\%}3Dihub`.

Bracher-Smith, Matthew et al. (2022). "Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank." eng. In: *Schizophrenia research* 246, pp. 156–164. ISSN: 1573-2509 (Electronic). DOI: `10.1016/j.schres.2022.06.006`.

Braff, D L (1993). "Information processing and attention dysfunctions in schizophrenia." eng. In: *Schizophrenia bulletin* 19.2, pp. 233–259. ISSN: 0586-7614 (Print). DOI: `10.1093/schbul/19.2.233`.

Brugger, Stefan P. and Oliver D. Howes (2017). "Heterogeneity and Homogeneity of Regional Brain Structure in Schizophrenia". In: *JAMA Psychiatry* 74.11, p. 1104. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2017.2663. URL: http://www.ncbi.nlm.nih.gov/pubmed/28973084http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5669456http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/jamapsychiatry.2017.2663.

Bulik-Sullivan, Brendan K et al. (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3, pp. 291–295. ISSN: 1061-4036. DOI: 10.1038/ng.3211. arXiv: 15334406. URL: http://www.nature.com/doifinder/10.1038/ng.3211.

Calandreau, L et al. (2010). "Differential impact of polysialyltransferase ST8SiaII and ST8SiaIV knockout on social interaction and aggression." eng. In: *Genes, brain, and behavior* 9.8, pp. 958–967. ISSN: 1601-183X (Electronic). DOI: 10.1111/j.1601-183X.2010.00635.x.

Camacho, Christiam et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1, p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421. URL: http://www.biomedcentral.com/1471-2105/10/421.

Cao, Hengyi, Hang Zhou, and Tyrone D Cannon (2021). "Functional connectome-wide associations of schizophrenia polygenic risk." eng. In: *Molecular psychiatry* 26.6, pp. 2553–2561. ISSN: 1476-5578 (Electronic). DOI: 10.1038/s41380-020-0699-3.

Caseras, X et al. (2015). "Association between genetic risk scoring for schizophrenia and bipolar disorder with regional subcortical volumes." In: *Translational psychiatry* 5.12, e692. ISSN: 2158-3188. DOI: 10.1038/tp.2015.195. URL: http://www.ncbi.nlm.nih.gov/pubmed/26645627http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5068590.

Chang, C C et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1, p. 7. ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8. URL: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0047-8.

Chen, Guo-Bo et al. (2017). "Across-cohort QC analyses of GWAS summary statistics from complex traits". In: *European Journal of Human Genetics* 25.1, pp. 137–146. ISSN: 1476-5438. DOI: 10.1038/ejhg.2016.106. URL: https://doi.org/10.1038/ejhg.2016.106.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F O'Reilly (2020). "Tutorial: a guide to performing polygenic risk score analyses". In: *Nature Protocols* 15.9, pp. 2759–2772. ISSN: 1750-2799. DOI: 10.1038/s41596-020-0353-1. URL: https://doi.org/10.1038/s41596-020-0353-1.

Choi, Shing Wan and Paul F O'Reilly (2019). "PRSice-2: Polygenic Risk Score software for biobank-scale data". eng. In: *GigaScience* 8.7, giz082. ISSN: 2047-217X. DOI: 10.1093/gigascience/giz082. URL: https://www.ncbi.nlm.nih.gov/pubmed/31307061https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6629542/.

Choi, Shing Wan et al. (2022). "The power of pathway-based polygenic risk scores". In: *Research Square*. ISSN: 2693-5015. DOI: 10.21203/rs.3.rs-643696/v1. URL: https://doi.org/10.21203/rs.3.rs-643696/v1.

Chow, E W, A S Bassett, and R Weksberg (1994). "Velo-cardio-facial syndrome and psychotic disorders: implications for psychiatric genetics". eng. In: *American journal of medical genetics* 54.2, pp. 107–112. ISSN: 0148-7299. DOI: 10.1002/ajmg.1320540205. URL: https://www.ncbi.nlm.nih.gov/pubmed/8074160https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3142271/.

Clifton, N E et al. (2017). "Schizophrenia copy number variants and associative learning". In: *Molecular Psychiatry* 22.2, pp. 178–182. ISSN: 1476-5578. DOI: 10.1038/mp.2016.227. URL: https://doi.org/10.1038/mp.2016.227.

Collins, Rory (2012). "What makes UK Biobank special?" In: *The Lancet* 379.31, pp. 1173–1174. DOI: 10.1016/S0140. URL: http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(12)60404-8.pdf.

Cosgrove, Donna et al. (2018). "Effects of MiR-137 genetic risk score on brain volume and cortical measures in patients with schizophrenia and controls." eng. In: *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 177.3, pp. 369–376. ISSN: 1552-485X (Electronic). DOI: 10.1002/ajmg.b.32620.

Cox, D. D. and J. S. Lee (2008). "Pointwise testing with functional data using the Westfall-Young randomization method". In: *Biometrika* 95.3, pp. 621–634. ISSN: 0006-3444. DOI: 10.1093/biomet/asn021. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asn021.

Dahoun, T et al. (2017). "The impact of Disrupted-in-Schizophrenia 1 (DISC1) on the dopaminergic system: a systematic review". In: *Translational Psychiatry* 7.1, e1015–e1015. ISSN: 2158-3188. DOI: 10.1038/tp.2016.282. URL: https://doi.org/10.1038/tp.2016.282.

Darnell, Jennifer C. et al. (2011). "FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism". In: *Cell* 146.2, pp. 247–

261. ISSN: 00928674. DOI: 10.1016/j.cell.2011.06.013. URL: http://www.ncbi.nlm.nih.gov/pubmed/21784246http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3232425http://linkinghub.elsevier.com/retrieve/pii/S0092867411006556.

Davies, Gail et al. (2018). "Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function". In: *Nature Communications* 9.1, p. 2098. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04362-x. URL: https://doi.org/10.1038/s41467-018-04362-x.

Davies, Neil M, Michael V Holmes, and George Davey Smith (2018). "Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians". In: *BMJ* 362. ISSN: 0959-8138. DOI: 10.1136/bmj.k601. URL: https://www.bmj.com/content/362/bmj.k601.

Davis, Oliver S P, Claire M A Haworth, and Robert Plomin (2009). "Dramatic increase in heritability of cognitive development from early to middle childhood: an 8-year longitudinal study of 8,700 pairs of twins". eng. In: *Psychological science* 20.10, pp. 1301–1308. ISSN: 1467-9280. DOI: 10.1111/j.1467-9280.2009.02433.x. URL: https://pubmed.ncbi.nlm.nih.gov/19732386https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4040420/.

De Peri, Luca and Antonio Vita (2019). "Factors Associated With Brain Heterogeneity in Schizophrenia". In: *JAMA Psychiatry*. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2019.1852. URL: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2738763.

Deary, Ian J, W Johnson, and L M Houlihan (2009). "Genetic foundations of human intelligence." eng. In: *Human genetics* 126.1, pp. 215–232. ISSN: 1432-1203 (Electronic). DOI: 10.1007/s00439-009-0655-4.

Derks, Eske M et al. (2012). "Investigation of the Genetic Association between Quantitative Measures of Psychosis and Schizophrenia: A Polygenic Risk Score Analysis". In: *PLOS ONE* 7.6, e37852. URL: https://doi.org/10.1371/journal.pone.0037852.

Desikan, Rahul S et al. (2017). "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score". In: *PLOS Medicine* 14.3, e1002258. URL: https://doi.org/10.1371/journal.pmed.1002258.

Dickinson, Dwight et al. (2020). "Distinct Polygenic Score Profiles in Schizophrenia Subgroups With Different Trajectories of Cognitive Development." eng. In: *The American journal of psychiatry* 177.4, pp. 298–307. ISSN: 1535-7228 (Electronic). DOI: 10.1176/appi.ajp.2019.19050527.

Dudbridge, Frank (2013). "Power and Predictive Accuracy of Polygenic Risk Scores". In: *PLoS Genetics* 9.3. ISSN: 15537390. DOI: 10.1371/journal.pgen.1003348.

Erp, T G M van et al. (2016). "Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium". In: *Molecular Psychiatry* 21.4, pp. 547–553. ISSN: 1359-4184. DOI: 10.1038/mp.2015.63. URL: http://www.ncbi.nlm.nih.gov/pubmed/26033243http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4668237http://www.nature.com/doifinder/10.1038/mp.2015.63.

Erp, Theo G.M. van et al. (2018). "Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium". In: *Biological Psychiatry* 84.9, pp. 644–654. ISSN: 0006-3223. DOI: 10.1016/J.BIOPSYCH.2018.04.023. URL: https://www.sciencedirect.com/science/article/pii/S0006322318315178.

Escott-Price, V et al. (2019). "Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles." eng. In: *The journal of prevention of Alzheimer's disease* 6.1, pp. 16–19. ISSN: 2426-0266 (Electronic). DOI: 10.14283/jpad.2018.46.

Escott-Price, Valentina et al. (2015). "Common polygenic variation enhances risk prediction for Alzheimer's disease". In: *Brain* 138.12, pp. 3673–3684. ISSN: 0006-8950. DOI: 10.1093/brain/awv268. URL: http://www.ncbi.nlm.nih.gov/pubmed/26490334http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5006219https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awv268.

Euesden, Jack, Cathryn M. Lewis, and Paul F. O'Reilly (2015). "PRSice: Polygenic Risk Score software". In: *Bioinformatics* 31.9, pp. 1466–1468. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu848.

Fischer, M (1973). "Genetic and environmental factors in schizophrenia. A study of schizophrenic twins and their families." eng. In: *Acta psychiatrica Scandinavica. Supplementum* 238, pp. 9–142. ISSN: 0065-1591 (Print).

Foley, Sonya F. et al. (2017). "Multimodal Brain Imaging Reveals Structural Differences in Alzheimer's Disease Polygenic Risk Carriers: A Study in Healthy Young Adults". In: *Biological Psychiatry* 81.2, pp. 154–161. ISSN: 00063223. DOI: 10.1016/j.biopsych.2016.02.033. URL: http://www.ncbi.nlm.nih.gov/pubmed/27157680http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5177726https://linkinghub.elsevier.com/retrieve/pii/S0006322316311143.

Fowler, Tom et al. (2012). "A Population-Based Study of Shared Genetic Variation Between Premorbid IQ and Psychosis Among Male Twin Pairs and Sibling Pairs From Sweden". In: *Archives of General Psychiatry* 69.5, p. 460. ISSN: 0003-990X. DOI: `10.1001/archgenpsychiatry.2011.1370`. URL: `http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archgenpsychiatry.2011.1370`.

Frank, J et al. (2015). *Identification of increased genetic risk scores for schizophrenia in treatment-resistant patients.* eng. DOI: `10.1038/mp.2014.56`.

Franke, Barbara et al. (2016). "Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept". In: *Nature Neuroscience* 19.3, pp. 420–431. ISSN: 1097-6256. DOI: `10.1038/nn.4228`. URL: `http://www.nature.com/doifinder/10.1038/nn.4228`.

Frazer, Kelly A et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164, pp. 851–861. ISSN: 1476-4687. DOI: `10.1038/nature06258`. URL: `https://doi.org/10.1038/nature06258`.

Friston, K J and C D Frith (1995). "Schizophrenia: a disconnection syndrome?" eng. In: *Clinical neuroscience (New York, N.Y.)* 3.2, pp. 89–97. ISSN: 1065-6766 (Print).

Fromer, Menachem et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks". In: *Nature* 506.7487, pp. 179–184. ISSN: 0028-0836. DOI: `10.1038/nature12929`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/24463507http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4237002http://www.nature.com/doifinder/10.1038/nature12929`.

Fujito, Naoko T et al. (2018). "Positive selection on schizophrenia-associated ST8SIA2 gene in post-glacial Asia". In: *PLOS ONE* 13.7, e0200278. URL: `https://doi.org/10.1371/journal.pone.0200278`.

Fusar-Poli, Laura et al. (2022). "Polygenic risk scores for predicting outcomes and treatment response in psychiatry: hope or hype?" In: *International Review of Psychiatry*, pp. 1–13. ISSN: 0954-0261. DOI: `10.1080/09540261.2022.2101352`. URL: `https://doi.org/10.1080/09540261.2022.2101352`.

Ge, Tian et al. (2019). "Polygenic prediction via Bayesian regression and continuous shrinkage priors". In: *Nature Communications* 10.1, p. 1776. ISSN: 2041-1723. DOI: `10.1038/s41467-019-09718-5`. URL: `https://doi.org/10.1038/s41467-019-09718-5`.

Gejman, Pablo V, Alan R Sanders, and Kenneth S Kendler (2011). "Genetics of schizophrenia: new findings and challenges." eng. In: *Annual review of* *genomics and human genetics* 12, pp. 121–144. ISSN: 1545-293X (Electronic). DOI: 10.1146/annurev-genom-082410-101459.

Glasheen, Cristie (Substance Abuse et al. (2016). *Impact of the DSM-IV to DSM-5 Changes on the National Survey on Drug Use and Health*. Substance Abuse and Mental Health Services Administration (US). URL: http://www.ncbi.nlm.nih.gov/pubmed/30199183.

Glessner, Joseph T et al. (2010). "Strong synaptic transmission impact by copy number variations in schizophrenia". In: *Proceedings of the National Academy of Sciences* 107.23, pp. 10584–10589. DOI: 10.1073/pnas.1000274107. URL: https://doi.org/10.1073/pnas.1000274107.

Gottesman, I I and J Shields (1967). "A polygenic theory of schizophrenia." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 58.1, pp. 199–205. ISSN: 0027-8424 (Print). DOI: 10.1073/pnas.58.1.199.

Grama, Steluta et al. (2020). "Polygenic risk for schizophrenia and subcortical brain anatomy in the UK Biobank cohort." eng. In: *Translational psychiatry* 10.1, p. 309. ISSN: 2158-3188 (Electronic). DOI: 10.1038/s41398-020-00940-0.

Green, Michael F (2006). "Cognitive impairment and functional outcome in schizophrenia and bipolar disorder". In: *J Clin Psychiatry* 6767.9, pp. 3–8. URL: http://www.psychiatrist.com/JCP/article/{_}layouts/ppp.psych.controls/BinaryViewer.ashx?Article=/jcp/article/Pages/2006/v67s09/v67s0901.aspx{\&}Type=Article.

Green, Michael F and Philip D Harvey (2014). "Cognition in schizophrenia: Past, present, and future." In: *Schizophrenia research. Cognition* 1.1, e1–e9. ISSN: 2215-0013. DOI: 10.1016/j.scog.2014.02.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/25254156http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4171037.

Green, Michael F, Katiah Llerena, and Robert S Kern (2015). "The "Right Stuff" Revisited: What Have We Learned About the Determinants of Daily Functioning in Schizophrenia?" In: *Schizophrenia bulletin* 41.4, pp. 781–5. DOI: 10.1093/schbul/sbv018. URL: http://www.ncbi.nlm.nih.gov/pubmed/25750248.

Grove, Jakob et al. (2019). "Identification of common genetic risk variants for autism spectrum disorder." In: *Nature genetics* 51.3, pp. 431–444. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0344-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/30804558http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6454898.

Hagenaars, S P et al. (2016a). "Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112151) and 24 GWAS consortia". In: *Molecular Psychiatry* 21, p. 1624. URL: http://dx.doi.org/10.1038/mp.2015.225http://10.0.4.14/mp.2015.225https://www.nature.com/articles/mp2015225{\#}supplementary-information.

Hagenaars, S P et al. (2016b). "Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112151) and 24 GWAS consortia". In: *Molecular Psychiatry* 21.11, pp. 1624–1632. ISSN: 1359-4184. DOI: 10.1038/mp.2015.225. URL: http://www.nature.com/articles/mp2015225.

Haijma, Sander V. et al. (2013). "Brain Volumes in Schizophrenia: A Meta-Analysis in Over 18 000 Subjects". In: *Schizophrenia Bulletin* 39.5, pp. 1129–1138. ISSN: 1745-1701. DOI: 10.1093/schbul/sbs118. URL: http://www.ncbi.nlm.nih.gov/pubmed/23042112http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3756785https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/sbs118.

Hamshere, M L et al. (2011). "Polygenic dissection of the bipolar phenotype." eng. In: *The British journal of psychiatry : the journal of mental science* 198.4, pp. 284–288. ISSN: 1472-1465 (Electronic). DOI: 10.1192/bjp.bp.110.087866.

Han, Buhm et al. (2016). "A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping". eng. In: *Human molecular genetics* 25.9, pp. 1857–1866. ISSN: 1460-2083. DOI: 10.1093/hmg/ddw049. URL: https://pubmed.ncbi.nlm.nih.gov/26908615https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4986332/.

Hartwig, Fernando Pires et al. (2017). "Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study". eng. In: *JAMA psychiatry* 74.12, pp. 1226–1233. ISSN: 2168-6238. DOI: 10.1001/jamapsychiatry.2017.3191. URL: https://www.ncbi.nlm.nih.gov/pubmed/29094161https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6583386/.

Henriksen, Mads G, Julie Nordgaard, and Lennart B Jansson (2017). "Genetics of Schizophrenia: Overview of Methods, Findings and Limitations." In: *Frontiers in human neuroscience* 11, p. 322. ISSN: 1662-5161. DOI: 10.3389/fnhum.2017.00322. URL: http://www.ncbi.nlm.nih.gov/pubmed/28690503http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5480258.

Hess, Jonathan L et al. (2019). "A polygenic resilience score moderates the genetic risk for schizophrenia". In: *Molecular Psychiatry*. ISSN: 1476-5578.

DOI: 10.1038/s41380-019-0463-8. URL: https://doi.org/10.1038/s41380-019-0463-8.

Hibar, Derrek P et al. (2015). "Common genetic variants influence human subcortical brain structures". In: *Nature* 520.7546, pp. 224–229. ISSN: 1476-4687. DOI: 10.1038/nature14101. URL: https://doi.org/10.1038/nature14101.

Hill, Matthew J et al. (2014). "Transcriptional consequences of schizophrenia candidate miR-137 manipulation in human neural progenitor cells." eng. In: *Schizophrenia research* 153.1-3, pp. 225–230. ISSN: 1573-2509 (Electronic). DOI: 10.1016/j.schres.2014.01.034.

Hill, W David et al. (2016). "Age-Dependent Pleiotropy Between General Cognitive Function and Major Psychiatric Disorders." In: *Biological psychiatry* 80.4, pp. 266–273. ISSN: 1873-2402. DOI: 10.1016/j.biopsych.2015.08.033. URL: http://www.ncbi.nlm.nih.gov/pubmed/26476593http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4974237.

Holzman, Philip S. (1989). "The use of eye movement dysfunctions in exploring the genetic transmission of schizophrenia". In: *European Archives of Psychiatry and Neurological Sciences* 239.1, pp. 43–48. ISSN: 0175758X. DOI: 10.1007/BF01739743.

Howie, Bryan N, Peter Donnelly, and Jonathan Marchini (2009). "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies". In: *PLOS Genetics* 5.6, e1000529. URL: https://doi.org/10.1371/journal.pgen.1000529.

Hubbard, Leon et al. (2016). "Evidence of Common Genetic Overlap Between Schizophrenia and Cognition." In: *Schizophrenia bulletin* 42.3, pp. 832–42. ISSN: 1745-1701. DOI: 10.1093/schbul/sbv168. URL: http://www.ncbi.nlm.nih.gov/pubmed/26678674http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4838093.

Ingvar, D H and G Franzén (1974). "Abnormalities of cerebral blood flow distribution in patients with chronic schizophrenia." eng. In: *Acta psychiatrica Scandinavica* 50.4, pp. 425–462. ISSN: 0001-690X (Print). DOI: 10.1111/j.1600-0447.1974.tb09707.x.

Jablensky, Assen (2010). "The diagnostic concept of schizophrenia: its history, evolution, and future prospects." eng. In: *Dialogues in clinical neuroscience* 12.3, pp. 271–287. ISSN: 1294-8322 (Print). DOI: 10.31887/DCNS.2010.12.3/ajablensky.

Jonas, Katherine G et al. (2019). "Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders". In: *Translational Psychiatry* 9.1,

p. 300. ISSN: 2158-3188. DOI: 10.1038/s41398-019-0612-5. URL: https://doi.org/10.1038/s41398-019-0612-5.

Jones, Hannah J et al. (2016). "Phenotypic Manifestation of Genetic Risk for Schizophrenia During Adolescence in the General Population". In: *JAMA Psychiatry* 73.3, pp. 221–228. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2015.3058. URL: https://doi.org/10.1001/jamapsychiatry.2015.3058.

Kanehisa, M and S Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." eng. In: *Nucleic acids research* 28.1, pp. 27–30. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/28.1.27.

Keefe, Richard S. E. et al. (2007a). "Neurocognitive Effects of Antipsychotic Medications in Patients With Chronic Schizophrenia in the CATIE Trial". In: *Archives of General Psychiatry* 64.6, p. 633. ISSN: 0003-990X. DOI: 10.1001/archpsyc.64.6.633. URL: http://www.ncbi.nlm.nih.gov/pubmed/17548746 http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archpsyc.64.6.633.

Keefe, Richard S.E. et al. (2007b). "Effects of Olanzapine, Quetiapine, and Risperidone on Neurocognitive Function in Early Psychosis: A Randomized, Double-Blind 52-Week Comparison". In: *American Journal of Psychiatry* 164.7, pp. 1061–1071. ISSN: 0002-953X. DOI: 10.1176/ajp.2007.164.7.1061. URL: http://www.ncbi.nlm.nih.gov/pubmed/17606658http://psychiatryonline.org/doi/abs/10.1176/ajp.2007.164.7.1061.

Kelly, S et al. (2018). "Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group." eng. In: *Molecular psychiatry* 23.5, pp. 1261–1269. ISSN: 1476-5578 (Electronic). DOI: 10.1038/mp.2017.170.

Keshavan, Matcheri S et al. (2020). "Neuroimaging in Schizophrenia." eng. In: *Neuroimaging clinics of North America* 30.1, pp. 73–83. ISSN: 1557-9867 (Electronic). DOI: 10.1016/j.nic.2019.09.007.

Khera, Amit V et al. (2018). "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". In: *Nature Genetics* 50.9, pp. 1219–1224. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0183-z. URL: https://doi.org/10.1038/s41588-018-0183-z.

Kirov, G et al. (2012). "De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia". In: *Molecular Psychiatry* 17.2, pp. 142–153. ISSN: 1476-5578. DOI: 10.1038/mp.2011.154. URL: https://doi.org/10.1038/mp.2011.154.

Kirov, George et al. (2009). "Support for the involvement of large copy number variants in the pathogenesis of schizophrenia". In: *Human Molecular Genetics* 18.8, pp. 1497–1503. ISSN: 1460-2083. DOI: 10.1093/hmg/ddp043. URL: http://www.ncbi.nlm.nih.gov/pubmed/19181681http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2664144https://academic.oup.com/hmg/article/18/8/1497/583108.

Kochunov, Peter et al. (2018). "Integration of routine QA data into mega-analysis may improve quality and sensitivity of multisite diffusion tensor imaging studies." eng. In: *Human brain mapping* 39.2, pp. 1015–1023. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.23900.

Kowalec, Kaarina et al. (2021). "Increased schizophrenia family history burden and reduced premorbid IQ in treatment-resistant schizophrenia: a Swedish National Register and Genomic Study." eng. In: *Molecular psychiatry* 26.8, pp. 4487–4495. ISSN: 1476-5578 (Electronic). DOI: 10.1038/s41380-019-0575-1.

Kraguljac, Nina V et al. (2021). "Neuroimaging Biomarkers in Schizophrenia." eng. In: *The American journal of psychiatry* 178.6, pp. 509–521. ISSN: 1535-7228 (Electronic). DOI: 10.1176/appi.ajp.2020.20030340.

Krapohl, Eva et al. (2014). "The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence". In: *Proceedings of the National Academy of Sciences* 111.42, 15273 LP –15278. DOI: 10.1073/pnas.1408777111. URL: http://www.pnas.org/content/111/42/15273.abstract.

Kremen, William et al. (2013). "Early identification and heritability of mild cognitive impairment". In: *International journal of epidemiology* 43. DOI: 10.1093/ije/dyt242.

Kröcher, Tim et al. (2015). "Schizophrenia-like phenotype of polysialyltransferase ST8SIA2-deficient mice." eng. In: *Brain structure & function* 220.1, pp. 71–83. ISSN: 1863-2661 (Electronic). DOI: 10.1007/s00429-013-0638-z.

Lam, Max et al. (2017). "Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets." eng. In: *Cell reports* 21.9, pp. 2597–2613. ISSN: 2211-1247 (Electronic). DOI: 10.1016/j.celrep.2017.11.028.

Lam, Max et al. (2019a). "Comparative genetic architectures of schizophrenia in East Asian and European populations". In: *Nature Genetics* 51.12, pp. 1670–1678. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0512-x. URL: https://doi.org/10.1038/s41588-019-0512-x.

Lam, Max et al. (2019b). “Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways”. eng. In: *American journal of human genetics* 105.2, pp. 334–350. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2019.06.012. URL: https://pubmed.ncbi.nlm.nih.gov/31374203https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6699140/.

Lander, Eric S. et al. (2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822, pp. 860–921. ISSN: 00280836. DOI: 10.1038/35057062. URL: https://doi.org/10.1038/35057062.

Lawrie, Stephen M et al. (2008). “Brain structure and function changes during the development of schizophrenia: the evidence from studies of subjects at increased genetic risk.” eng. In: *Schizophrenia bulletin* 34.2, pp. 330–340. ISSN: 0586-7614 (Print). DOI: 10.1093/schbul/sbm158.

Le Tourneau, Christophe et al. (2015). “Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial.” eng. In: *The Lancet. Oncology* 16.13, pp. 1324–1334. ISSN: 1474-5488 (Electronic). DOI: 10.1016/S1470-2045(15)00188-6.

LeBlanc, Marissa et al. (2018). “A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework”. In: *BMC Genomics* 19.1, p. 494. ISSN: 1471-2164. DOI: 10.1186/s12864-018-4859-7. URL: http://www.ncbi.nlm.nih.gov/pubmed/29940862http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6019513https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4859-7.

Lee, P H et al. (2016). “Partitioning heritability analysis reveals a shared genetic basis of brain anatomy and schizophrenia.” eng. In: *Molecular psychiatry* 21.12, pp. 1680–1689. ISSN: 1476-5578 (Electronic). DOI: 10.1038/mp.2016.164.

Lee, S H et al. (2013). “Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs”. In: *Nature Genetics* 45.9, pp. 984–994. ISSN: 1061-4036. DOI: 10.1038/ng.2711. URL: http://www.ncbi.nlm.nih.gov/pubmed/23933821http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3800159http://www.nature.com/doifinder/10.1038/ng.2711.

Lee, Sang Hong et al. (2011). “Estimating missing heritability for disease from genome-wide association studies”. eng. In: *American journal of human genetics* 88.3, pp. 294–305. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2011.02.002. URL:

https://www.ncbi.nlm.nih.gov/pubmed/21376301https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059431/.

Lee, Seunggeun et al. (2012). "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies". In: *American Journal of Human Genetics* 91.2, pp. 224–237. ISSN: 00029297. DOI: 10.1016/j.ajhg.2012.06.007.

Leeuw, Christiaan A. de et al. (2015). "MAGMA: Generalized Gene-Set Analysis of GWAS Data". In: *PLoS Computational Biology* 11.4, pp. 1–19. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004219.

Legge, Sophie E et al. (2020). "Clinical indicators of treatment-resistant psychosis." eng. In: *The British journal of psychiatry : the journal of mental science* 216.5, pp. 259–266. ISSN: 1472-1465 (Electronic). DOI: 10.1192/bjp.2019.120.

Legge, Sophie E et al. (2021). "Associations Between Schizophrenia Polygenic Liability, Symptom Dimensions, and Cognitive Ability in Schizophrenia." eng. In: *JAMA psychiatry* 78.10, pp. 1143–1151. ISSN: 2168-6238 (Electronic). DOI: 10.1001/jamapsychiatry.2021.1961.

Leipzig, Jeremy (2016). "A review of bioinformatic pipeline frameworks". In: *Briefings in Bioinformatics* 18.3, bbw020. ISSN: 1467-5463. DOI: 10.1093/bib/bbw020. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw020.

Lek, Monkol et al. (2016a). "Analysis of protein-coding genetic variation in 60,706 humans". In: *Nature* 536.7616, pp. 285–291. ISSN: 1476-4687. DOI: 10.1038/nature19057. URL: https://doi.org/10.1038/nature19057.

Lek, Monkol et al. (2016b). "Analysis of protein-coding genetic variation in 60,706 humans". In: *Nature* 536.7616, pp. 285–291. ISSN: 0028-0836. DOI: 10.1038/nature19057. URL: http://www.nature.com/articles/nature19057.

Lemvigh, Cecilie K et al. (2020). "Heritability of specific cognitive functions and associations with schizophrenia spectrum disorders using CANTAB: a nation-wide twin study". In: *Psychological Medicine*, pp. 1–14. ISSN: 0033-2917. DOI: DOI:10.1017/S0033291720002858. URL: https://www.cambridge.org/core/article/heritability-of-specific-cognitive-functions-and-associations-with-schizophrenia-spectrum-disorders-using-cantab-a-nationwide-twin-study/346FD540108463CC3F1D6B9BCEF6D5EC.

Lencz, T et al. (2007). "Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia." eng. In: *Molecular psychiatry* 12.6, pp. 572–580. ISSN: 1359-4184 (Print). DOI: 10.1038/sj.mp.4001983.

Lencz, T et al. (2014). “Molecular genetic evidence for overlap between general cognitive ability and risk for schizophrenia: a report from the Cognitive Genomics consorTium (COGENT).” In: *Molecular psychiatry* 19.2, pp. 168–74. ISSN: 1476-5578. DOI: 10.1038/mp.2013.166. URL: http://www.ncbi.nlm.nih.gov/pubmed/24342994http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3968799.

Lewis, Cathryn M and Evangelos Vassos (2020). “Polygenic risk scores: from research tools to clinical instruments”. In: *Genome Medicine* 12.1, p. 44. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00742-5. URL: https://doi.org/10.1186/s13073-020-00742-5.

Li, Yun et al. (2009). “Genotype imputation”. eng. In: *Annual review of genomics and human genetics* 10, pp. 387–406. ISSN: 1545-293X. DOI: 10.1146/annurev.genom.9.081307.164242. URL: https://pubmed.ncbi.nlm.nih.gov/19715440https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925172/.

Li, Zhixiu et al. (2021). “Polygenic Risk Scores have high diagnostic capacity in ankylosing spondylitis”. In: *Annals of the Rheumatic Diseases*, annrheumdis–2020–219446. DOI: 10.1136/annrheumdis-2020-219446. URL: http://ard.bmj.com/content/early/2021/07/16/annrheumdis-2020-219446.abstract.

Liberzon, A. et al. (2011). “Molecular signatures database (MSigDB) 3.0”. In: *Bioinformatics* 27.12, pp. 1739–1740. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr260. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr260.

Lin, Dan-Yu and Patrick F Sullivan (2009). “Meta-analysis of genome-wide association studies with overlapping subjects.” In: *American journal of human genetics* 85.6, pp. 862–72. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2009.11.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/20004761http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2790578.

Lippert, Christoph et al. (2017). “Identification of individuals by trait prediction using whole-genome sequencing data”. In: *Proceedings of the National Academy of Sciences* 114.38, pp. 10166–10171. DOI: 10.1073/pnas.1711125114. URL: https://doi.org/10.1073/pnas.1711125114.

Liu, Shu et al. (2020). “MIR137 polygenic risk is associated with schizophrenia and affects functional connectivity of the dorsolateral prefrontal cortex.” eng. In: *Psychological medicine* 50.9, pp. 1510–1518. ISSN: 1469-8978 (Electronic). DOI: 10.1017/S0033291719001442.

Liu, Wei et al. (2021). *An Improved Genome-Wide Polygenic Score Model for Predicting the Risk of Type 2 Diabetes*. URL: `https://www.frontiersin.org/articles/10.3389/fgene.2021.632385`.

Lloyd-Jones, Luke R et al. (2019). "Improved polygenic prediction by Bayesian multiple regression on summary statistics". In: *Nature Communications* 10.1, p. 5086. ISSN: 2041-1723. DOI: `10.1038/s41467-019-12653-0`. URL: `https://doi.org/10.1038/s41467-019-12653-0`.

Lobo, Ingrid (2008). "Basic Local Alignment Search Tool (BLAST)". In: *Nature Education* 1.1, p. 215.

Locke, Adam E et al. (2015). "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.7538, pp. 197–206. ISSN: 1476-4687. DOI: `10.1038/nature14177`. URL: `https://doi.org/10.1038/nature14177`.

Lubman, D I et al. (2002). "Incidental radiological findings on brain magnetic resonance imaging in first-episode psychosis and chronic schizophrenia." eng. In: *Acta psychiatrica Scandinavica* 106.5, pp. 331–336. ISSN: 0001-690X (Print). DOI: `10.1034/j.1600-0447.2002.02217.x`.

Lynham, Amy J et al. (2018a). "Examining cognition across the bipolar / schizophrenia diagnostic spectrum". In: *Journal of Psychiatry and Neuroscience*. URL: `http://orca.cf.ac.uk/105165/http://orca.cf.ac.uk/policies.html`.

– (2018b). "Examining cognition across the bipolar/schizophrenia diagnostic spectrum". In: *Journal of Psychiatry & Neuroscience* 43.4, pp. 245–253. ISSN: 11804882. DOI: `10.1503/jpn.170076`. URL: `http://jpn.ca/vol43-issue4/43-4-245/`.

Lynham, Amy J et al. (2018c). "Examining cognition across the bipolar/schizophrenia diagnostic spectrum." In: *Journal of psychiatry & neuroscience : JPN* 43.4, pp. 245–253. ISSN: 1488-2434. DOI: `10.1503/JPN.170076`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/29947606http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6019354`.

Mak, Timothy Shin Heng et al. (2017). "Polygenic scores via penalized regression on summary statistics". In: *Genetic Epidemiology* 41.6, pp. 469–480. ISSN: 0741-0395. DOI: `10.1002/gepi.22050`. URL: `https://doi.org/10.1002/gepi.22050`.

Márquez-Luna, Carla et al. (2017). "Multiethnic polygenic risk scores improve risk prediction in diverse populations." In: *Genetic epidemiology* 41.8, pp. 811–823. ISSN: 1098-2272. DOI: `10.1002/gepi.22083`. URL: `http://www.ncbi.nl` m.nih.gov/pubmed/29110330http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5726434.

Marshall, Christian R et al. (2017). "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects." eng. In: *Nature genetics* 49.1, pp. 27–35. ISSN: 1546-1718 (Electronic). DOI: 10.1038/ng.3725.

Martin, Alicia R et al. (2019). "Clinical use of current polygenic risk scores may exacerbate health disparities". eng. In: *Nature genetics* 51.4, pp. 584–591. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0379-x. URL: https://pubmed.ncbi.nlm.nih.gov/30926966https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6563838/.

Maston, Glenn A., Sara K. Evans, and Michael R. Green (2006). "Transcriptional Regulatory Elements in the Human Genome". In: *Annual Review of Genomics and Human Genetics* 7.1, pp. 29–59. ISSN: 1527-8204. DOI: 10.1146/annurev.genom.7.080505.115623. URL: http://www.ncbi.nlm.nih.gov/pubmed/16719718http://www.annualreviews.org/doi/10.1146/annurev.genom.7.080505.115623.

McEntyre, Jo and Jim Ostell (2002). *The NCBI Handbook*. Ed. by Jo McEntyre and Jim Ostell. Bethesda (MD): National Center for Biotechnology Information (US). URL: https://www.ncbi.nlm.nih.gov/books/NBK21088/.

McGrath, John J et al. (2014). "A Comprehensive Assessment of Parental Age and Psychiatric Disorders". In: *JAMA Psychiatry* 71.3, pp. 301–309. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2013.4081. URL: https://doi.org/10.1001/jamapsychiatry.2013.4081.

McInnes, Leland, John Healy, and James Melville (2020). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv*. arXiv: 1802.03426.

Merwe, C. van der et al. (2019). "Polygenic risk for schizophrenia and associated brain structural changes: A systematic review". In: *Comprehensive Psychiatry* 88, pp. 77–82. ISSN: 0010-440X. DOI: 10.1016/J.COMPPSYCH.2018.11.014. URL: https://www.sciencedirect.com/science/article/pii/S0010440X18301998.

Michailidou, Kyriaki et al. (2017). "Association analysis identifies 65 new breast cancer risk loci". eng. In: *Nature* 551.7678, pp. 92–94. ISSN: 1476-4687. DOI: 10.1038/nature24284. URL: https://www.ncbi.nlm.nih.gov/pubmed/29059683https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5798588/.

Milaneschi, Y et al. (2016). "Polygenic dissection of major depression clinical heterogeneity." eng. In: *Molecular psychiatry* 21.4, pp. 516–522. ISSN: 1476-5578 (Electronic). DOI: 10.1038/mp.2015.86.

Mistry, Sumit et al. (2018). "The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review." eng. In: *Schizophrenia research* 197, pp. 2–8. ISSN: 1573-2509 (Electronic). DOI: 10.1016/j.schres.2017.10.037.

Mokhtari, Ryan and Herbert M Lachman (2016). "The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review". eng. In: *Journal of clinical & cellular immunology* 7.6, p. 479. ISSN: 2155-9899. DOI: 10.4172/2155-9899.1000479. URL: https://www.ncbi.nlm.nih.gov/pubmed/28180029https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5293234/.

Morgulis, Aleksandr et al. (2008). "Database indexing for production MegaBLAST searches". In: *Bioinformatics* 24.16, pp. 1757–1764. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btn322. URL: http://www.ncbi.nlm.nih.gov/pubmed/18567917http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2696921https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn322.

Mullins, Niamh et al. (2021). "Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology". In: *Nature Genetics* 53.6, pp. 817–829. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00857-4. URL: https://doi.org/10.1038/s41588-021-00857-4.

Nelson, Sarah C et al. (2012). "Is 'forward' the same as 'plus'?...and other adventures in SNP allele nomenclature". eng. In: *Trends in genetics : TIG* 28.8, pp. 361–363. ISSN: 0168-9525. DOI: 10.1016/j.tig.2012.05.002. URL: https://pubmed.ncbi.nlm.nih.gov/22658725https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6099125/.

Newcombe, Paul J et al. (2019). "A flexible and parallelizable approach to genome-wide polygenic risk scores". In: *Genetic Epidemiology* 43.7, pp. 730–741. ISSN: 0741-0395. DOI: https://doi.org/10.1002/gepi.22245. URL: https://doi.org/10.1002/gepi.22245.

Ng, M Y M et al. (2009). "Meta-analysis of 32 genome-wide linkage studies of schizophrenia." eng. In: *Molecular psychiatry* 14.8, pp. 774–785. ISSN: 1476-5578 (Electronic). DOI: 10.1038/mp.2008.135.

Nikolaus, Susanne, Hans-Wilhelm Müller, and Hubertus Hautzel (2016). "Different patterns of 5-HT receptor and transporter dysfunction in neuropsychiatric disorders–a comparative analysis of in vivo imaging findings." eng. In: *Reviews in the neurosciences* 27.1, pp. 27–59. ISSN: 2191-0200 (Electronic). DOI: 10.1515/revneuro-2015-0014.

Noordzij, M et al. (2010). "Measures of Disease Frequency: Prevalence and Incidence". In: *Nephron Clinical Practice* 115.1, pp. c17–c20. DOI: 10.1159/000286345. URL: https://www.karger.com/DOI/10.1159/000286345.

Nuechterlein, Keith H. et al. (2008a). "The MATRICS Consensus Cognitive Battery, Part 1: Test Selection, Reliability, and Validity". In: *American Journal of Psychiatry* 165.2, pp. 203–213. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.2007.07010042. URL: http://www.ncbi.nlm.nih.gov/pubmed/18172019http://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2007.07010042.

– (2008b). "The MATRICS Consensus Cognitive Battery, Part 1: Test Selection, Reliability, and Validity". In: *American Journal of Psychiatry* 165.2, pp. 203–213. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.2007.07010042. URL: http://www.ncbi.nlm.nih.gov/pubmed/18172019http://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2007.07010042.

O'Brien, Heath E et al. (2018). "Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders". In: *Genome Biology* 19.1, p. 194. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1567-1. URL: https://doi.org/10.1186/s13059-018-1567-1.

O'Connell, R A et al. (1989). "Single photon emission computed tomography (SPECT) with [123I]IMP in the differential diagnosis of psychiatric disorders." eng. In: *The Journal of neuropsychiatry and clinical neurosciences* 1.2, pp. 145–153. ISSN: 0895-0172 (Print). DOI: 10.1176/jnp.1.2.145.

O'Donovan, Michael C et al. (2008). "Identification of loci associated with schizophrenia by genome-wide association and follow-up". In: *Nature Genetics* 40.9, pp. 1053–1055. ISSN: 1546-1718. DOI: 10.1038/ng.201. URL: https://doi.org/10.1038/ng.201.

Ohi, Kazutaka et al. (2018). "Genetic Overlap between General Cognitive Function and Schizophrenia: A Review of Cognitive GWASs". eng. In: *International journal of molecular sciences* 19.12, p. 3822. ISSN: 1422-0067. DOI: 10.3390/ijms19123822. URL: https://pubmed.ncbi.nlm.nih.gov/30513630https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6320986/.

Okada, N et al. (2016). "Abnormal asymmetries in subcortical brain volume in schizophrenia." In: *Molecular psychiatry* 21.10, pp. 1460–6. ISSN: 1476-5578. DOI: 10.1038/mp.2015.209. URL: http://www.ncbi.nlm.nih.gov/pubmed/26782053http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5030462.

Pardiñas, Antonio F. et al. (2018). "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection". In: *Nature Genetics* 50, pp. 381–389. ISSN: 1061-4036. DOI: 10.1038/s4

1588-018-0059-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/29483656 http://www.nature.com/articles/s41588-018-0059-2.

Plomin, R and I J Deary (2015). "Genetics and intelligence differences: five special findings". eng. In: *Molecular psychiatry* 20.1, pp. 98–108. ISSN: 1476-5578. DOI: 10.1038/mp.2014.105. URL: https://pubmed.ncbi.nlm.nih.gov/25224258https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270739/.

Pocklington, Andrew J. et al. (2015). "Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia". In: *Neuron* 86.5, pp. 1203–1214. ISSN: 10974199. DOI: 10.1016/j.neuron.2015.04.022. URL: http://dx.doi.org/10.1016/j.neuron.2015.04.022.

Polderman, Tinca J C et al. (2015a). "Meta-analysis of the heritability of human traits based on fifty years of twin studies". In: *Nature Genetics* 47.7, pp. 702–709. ISSN: 1546-1718. DOI: 10.1038/ng.3285. URL: https://doi.org/10.1038/ng.3285.

– (2015b). "Meta-analysis of the heritability of human traits based on fifty years of twin studies". In: *Nature Genetics* 47.7, pp. 702–709. ISSN: 1061-4036. DOI: 10.1038/ng.3285. URL: http://www.nature.com/articles/ng.3285.

Potkin, Steven G et al. (2020). "The neurobiology of treatment-resistant schizophrenia: paths to antipsychotic resistance and a roadmap for future research". In: *npj Schizophrenia* 6.1, p. 1. ISSN: 2334-265X. DOI: 10.1038/s41537-019-0090-z. URL: https://doi.org/10.1038/s41537-019-0090-z.

Power, Robert A et al. (2013). "Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings." eng. In: *JAMA psychiatry* 70.1, pp. 22–30. ISSN: 2168-6238 (Electronic). DOI: 10.1001/jamapsychiatry.2013.268.

Privé, Florian et al. (2019). "Making the Most of Clumping and Thresholding for Polygenic Scores." eng. In: *American journal of human genetics* 105.6, pp. 1213–1221. ISSN: 1537-6605 (Electronic). DOI: 10.1016/j.ajhg.2019.11.001.

Purcell, S M et al. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." In: *Nature* 460.7256, pp. 748–52. ISSN: 1476-4687. DOI: 10.1038/nature08185. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/19571811{\%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3912837.

Purcell, Shaun M et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". In: *Nature* 506.7487, pp. 185–190. ISSN: 1476-4687. DOI: 10.1038/nature12975. URL: https://doi.org/10.1038/nature12975.

Quick, Corbin et al. (2020). "Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations". In: *Genetic Epidemiology* 44.6, pp. 537–549. ISSN: 0741-0395. DOI: `https://doi.org/10.1002/gepi.22326`. URL: `https://doi.org/10.1002/gepi.22326`.

Radulescu, Eugenia et al. (2020). "Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain". In: *Molecular Psychiatry* 25.4, pp. 791–804. ISSN: 1476-5578. DOI: `10.1038/s41380-018-0304-1`. URL: `https://doi.org/10.1038/s41380-018-0304-1`.

Rammos, Alexandros et al. (2019). "The role of polygenic risk score gene-set analysis in the context of the omnigenic model of schizophrenia". In: *Neuropsychopharmacology* 44.9, pp. 1562–1569. ISSN: 1740-634X. DOI: `10.1038/s41386-019-0410-z`. URL: `https://doi.org/10.1038/s41386-019-0410-z`.

Ranlund, Siri et al. (2018). "A polygenic risk score analysis of psychosis endophenotypes across brain functional, structural, and cognitive domains." eng. In: *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 177.1, pp. 21–34. ISSN: 1552-485X (Electronic). DOI: `10.1002/ajmg.b.32581`.

Rees, E. et al. (2014a). "Analysis of copy number variations at 15 schizophrenia-associated loci". In: *The British Journal of Psychiatry* 204.2, pp. 108–114. ISSN: 0007-1250. DOI: `10.1192/bjp.bp.113.131052`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/24311552http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3909838http://bjp.rcpsych.org/cgi/doi/10.1192/bjp.bp.113.131052`.

Rees, E et al. (2014b). "Evidence that duplications of 22q11.2 protect against schizophrenia". eng. In: *Molecular psychiatry* 19.1, pp. 37–40. ISSN: 1476-5578. DOI: `10.1038/mp.2013.156`. URL: `https://www.ncbi.nlm.nih.gov/pubmed/24217254https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3873028/`.

Rees, Elliott, Michael C O'Donovan, and Michael J Owen (2015). "Genetics of schizophrenia". In: *Current Opinion in Behavioral Sciences* 2, pp. 8–14. ISSN: 2352-1546. DOI: `10.1016/J.COBEHA.2014.07.001`. URL: `https://www.sciencedirect.com/science/article/pii/S2352154614000035`.

Rees, Elliott et al. (2020). "De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia." eng. In: *Nature neuroscience* 23.2, pp. 179–184. ISSN: 1546-1726 (Electronic). DOI: `10.1038/s41593-019-0565-2`.

Rehman, Faiz (2011). "Schedules for clinical assessment in neuropsychiatry". In: *BMJ* 342, p. c7160. ISSN: 0959-8138. DOI: 10.1136/bmj.c7160. URL: http://www.bmj.com/lookup/doi/10.1136/bmj.c7160.

Reus, L M et al. (2017). "Association of polygenic risk for major psychiatric illness with subcortical volumes and white matter integrity in UK Biobank". In: *Scientific reports* 7, p. 42140. DOI: 10.1038/srep42140. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5301496/pdf/srep42140.pdf.

Richards, Alexander L et al. (2019). "The relationship between polygenic risk scores and cognition in schizophrenia". In: *Schizophrenia Bulletin*.

Richards, Alexander L et al. (2020). "The Relationship Between Polygenic Risk Scores and Cognition in Schizophrenia". In: *Schizophrenia Bulletin* 46.2, pp. 336–344. ISSN: 0586-7614. DOI: 10.1093/schbul/sbz061. URL: https://doi.org/10.1093/schbul/sbz061.

Ripke, Stephan et al. (2011). "Genome-wide association study identifies five new schizophrenia loci". In: *Nature Genetics* 43.10, pp. 969–978. ISSN: 10614036. DOI: 10.1038/ng.940. URL: http://www.ncbi.nlm.nih.gov/pubmed/21926974http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3303194.

Ripke, Stephan et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". In: *Nature genetics* 45.10, pp. 1–26. ISSN: 1546-1718. DOI: 10.1038/ng.2742.Genome-wide. URL: http://www.nature.com/ng/journal/v45/n10/abs/ng.2742.html.

Ripke, Stephan et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511.7510, pp. 421–427. ISSN: 14764687. DOI: 10.1038/nature13595. arXiv: NIHMS150003. URL: http://www.nature.com/doifinder/10.1038/nature13595.

Risch, N and K Merikangas (1996). "The future of genetic studies of complex human diseases." eng. In: *Science (New York, N.Y.)* 273.5281, pp. 1516–1517. ISSN: 0036-8075 (Print). DOI: 10.1126/science.273.5281.1516.

Sakaue, Saori et al. (2020). "Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction". In: *Nature Communications* 11.1, p. 1569. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15194-z. URL: https://doi.org/10.1038/s41467-020-15194-z.

Samocha, Kaitlin E et al. (2014). "A framework for the interpretation of de novo mutation in human disease". In: *Nature Genetics* 46.9, pp. 944–950. ISSN: 1546-1718. DOI: 10.1038/ng.3050. URL: https://doi.org/10.1038/ng.3050.

Sand, Philipp G (2007). "A lesson not learned: allele misassignment." In: *Behavioral and brain functions : BBF* 3, p. 65. ISSN: 1744-9081. DOI: 10.1186/1744-9081-3-65. URL: http://www.ncbi.nlm.nih.gov/pubmed/18154681http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2231368.

Sanders, Alan R et al. (2008). "No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics." eng. In: *The American journal of psychiatry* 165.4, pp. 497–506. ISSN: 0002-953X (Print). DOI: 10.1176/appi.ajp.2007.07101573.

Satizabal, Claudia L. et al. (2017). "Genetic Architecture of Subcortical Brain Structures in Over 40,000 Individuals Worldwide". In: *bioRxiv*. DOI: https://doi.org/10.1101/173831. arXiv: 173831. URL: http://www.biorxiv.org/content/early/2017/08/28/173831?{\%}3Fcollection=.

Sato, Daiki X and Masakado Kawata (2018). "Positive and balancing selection on SLC18A1 gene associated with psychiatric disorders and human-unique personality traits". In: *Evolution Letters* 2.5, pp. 499–510. ISSN: 2056-3744. DOI: https://doi.org/10.1002/evl3.81. URL: https://doi.org/10.1002/evl3.81.

Savage, Jeanne E. et al. (2018). "Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence". In: *Nature Genetics* 50.7, pp. 912–919. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0152-6. URL: http://www.nature.com/articles/s41588-018-0152-6.

Schijven, Dick et al. (2018). "Comprehensive pathway analyses of schizophrenia risk loci point to dysfunctional postsynaptic signaling". In: *Schizophrenia Research* 199, pp. 195–202. ISSN: 0920-9964. DOI: https://doi.org/10.1016/j.schres.2018.03.032. URL: https://www.sciencedirect.com/science/article/pii/S092099641830183X.

Schür, Remmelt R et al. (2016). "Brain GABA levels across psychiatric disorders: A systematic literature review and meta-analysis of (1) H-MRS studies." eng. In: *Human brain mapping* 37.9, pp. 3337–3352. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.23244.

Seeman, Mary V (2016). "Schizophrenogenic Mother". In: *Encyclopedia of Couple and Family Therapy*. Ed. by Jay Lebow, Anthony Chambers, and Douglas C Breunlin. Cham: Springer International Publishing, pp. 1–2. ISBN: 978-3-319-15877-8. DOI: 10.1007/978-3-319-15877-8_482-1. URL: https://doi.org/10.1007/978-3-319-15877-8{_}482-1.

Segrè, Ayellet V et al. (2010). "Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits." eng. In: *PLoS genetics* 6.8. ISSN: 1553-7404 (Electronic). DOI: 10.1371/journal.pgen.1001058.

Seibert, Tyler M et al. (2018). "Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts". In: *BMJ* 360, j5757. DOI: 10.1136/bmj.j5757. URL: http://www.bmj.com/content/360/bmj.j5757.abstract.

Seidman, Larry J. et al. (2015). "Factor structure and heritability of endophenotypes in schizophrenia: Findings from the Consortium on the Genetics of Schizophrenia (COGS-1)". In: *Schizophrenia Research* 163.1-3, pp. 73–79. ISSN: 09209964. DOI: 10.1016/j.schres.2015.01.027. URL: http://www.ncbi.nlm.nih.gov/pubmed/25682549http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5944296https://linkinghub.elsevier.com/retrieve/pii/S0920996415000572.

Sherry, S T et al. (2001). "dbSNP: the NCBI database of genetic variation." In: *Nucleic acids research* 29.1, pp. 308–11. ISSN: 1362-4962. URL: http://www.ncbi.nlm.nih.gov/pubmed/11125122http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC29783.

Shihabuddin, L et al. (1996). "Ventricular enlargement associated with linkage marker for schizophrenia-related disorders in one pedigree." eng. In: *Molecular psychiatry* 1.3, pp. 215–222. ISSN: 1359-4184 (Print).

Shulze, Ralf (2004). *Meta-analysis: A Comparison of Approaches*. Hogrefe & Huber.

Singh, Tarjinder et al. (2022). "Rare coding variants in ten genes confer substantial risk for schizophrenia". In: *Nature* 604.7906, pp. 509–516. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04556-w. URL: https://doi.org/10.1038/s41586-022-04556-w.

Skene, Nathan G et al. (2018). "Genetic identification of brain cell types underlying schizophrenia". In: *Nature Genetics* 50.6, pp. 825–833. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0129-5. URL: https://doi.org/10.1038/s41588-018-0129-5.

Smeland, Olav B. and Ole A. Andreassen (2018). "How can genetics help understand the relationship between cognitive dysfunction and schizophrenia?" In: *Scandinavian Journal of Psychology* 59.1, pp. 26–31. ISSN: 00365564. DOI: 10.1111/sjop.12407. URL: http://doi.wiley.com/10.1111/sjop.12407.

Smith, D J et al. (2016). "Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci". In: *Molecular Psychiatry* 21.6, pp. 749–757. ISSN: 1359-4184. DOI: `10.1038/mp.2016.49`. arXiv: `/dx.doi.org/10.1101/032417 [bioRxiv doi: http:]`. URL: `http://www.nature.com/doifinder/10.1038/mp.2016.49`.

Sniekers, S et al. (2017). "Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence". In: *Nature Genetics* 49, pp. 1107–1112. ISSN: 1061-4036. DOI: `10.1038/ng.3869`. URL: `http://www.nature.com/doifinder/10.1038/ng.3869`.

SNP FAQ Archive [Internet] (2005). Bethesda (MD): National Center for Biotechnology Information (US). URL: `https://www.ncbi.nlm.nih.gov/books/NBK44455/`.

Stauffer, Eva-Maria et al. (2021). "Grey and white matter microstructure is associated with polygenic risk for schizophrenia". In: *Molecular Psychiatry* 26.12, pp. 7709–7718. ISSN: 1476-5578. DOI: `10.1038/s41380-021-01260-5`. URL: `https://doi.org/10.1038/s41380-021-01260-5`.

Stein, Jason L et al. (2012). "Identification of common variants associated with human hippocampal and intracranial volumes." eng. In: *Nature genetics* 44.5, pp. 552–561. ISSN: 1546-1718 (Electronic). DOI: `10.1038/ng.2250`.

Stone, James M et al. (2009). "Cortical dopamine D2/D3 receptors are a common site of action for antipsychotic drugs–an original patient data meta-analysis of the SPECT and PET in vivo receptor imaging literature." eng. In: *Schizophrenia bulletin* 35.4, pp. 789–797. ISSN: 1745-1701 (Electronic). DOI: `10.1093/schbul/sbn009`.

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–50. ISSN: 0027-8424. DOI: `10.1073/pnas.0506580102`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/16199517http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1239896`.

Sudlow, Cathie et al. (2015). "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLOS Medicine* 12.3, e1001779. ISSN: 1549-1676. DOI: `10.1371/journal.pmed.1001779`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/25826379http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4380465http://dx.plos.org/10.1371/journal.pmed.1001779`.

Sugrue, Leo P and Rahul S Desikan (2019). "What Are Polygenic Scores and Why Are They Important?" In: *JAMA* 321.18, pp. 1820–1821. ISSN: 0098-7484.

DOI: 10.1001/jama.2019.3893. URL: https://doi.org/10.1001/jama.2019.3893.

Sul, Jae Hoon, Lana S. Martin, and Eleazar Eskin (2018). "Population structure in genetic studies: Confounding factors and mixed models". In: *PLOS Genetics* 14.12. Ed. by Gregory S. Barsh, e1007309. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007309. URL: http://dx.plos.org/10.1371/journal.pgen.1007309.

Sullivan, Patrick F, Kenneth S Kendler, and Michael C Neale (2003). "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies." eng. In: *Archives of general psychiatry* 60.12, pp. 1187–1192. ISSN: 0003-990X (Print). DOI: 10.1001/archpsyc.60.12.1187.

Sun, Ryan et al. (2019). "Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic". In: *PLOS Genetics* 15.3, e1007530. URL: https://doi.org/10.1371/journal.pgen.1007530.

Swantje, Müller Catrin et al. (2010). "Quantitative proteomics of the Cav2 channel nano-environments in the mammalian brain". In: *Proceedings of the National Academy of Sciences* 107.34, pp. 14950–14957. DOI: 10.1073/pnas.1005940107. URL: https://doi.org/10.1073/pnas.1005940107.

Tandon, Rajiv et al. (2013). "Definition and description of schizophrenia in the DSM-5". In: *Schizophrenia Research* 150.1, pp. 3–10. ISSN: 09209964. DOI: 10.1016/j.schres.2013.05.028. URL: http://www.ncbi.nlm.nih.gov/pubmed/23800613https://linkinghub.elsevier.com/retrieve/pii/S0920996413002831.

Tansey, K E et al. (2016). "Common alleles contribute to schizophrenia in CNV carriers". In: *Molecular Psychiatry* 21.8, pp. 1085–1089. ISSN: 1359-4184. DOI: 10.1038/mp.2015.143. URL: http://www.nature.com/doifinder/10.1038/mp.2015.143.

Tesli, M et al. (2014). "Polygenic risk score and the psychosis continuum model." eng. In: *Acta psychiatrica Scandinavica* 130.4, pp. 311–317. ISSN: 1600-0447 (Electronic). DOI: 10.1111/acps.12307.

Thompson, Paul M et al. (2020). "ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries". In: *Translational Psychiatry* 10.1, p. 100. ISSN: 2158-3188. DOI: 10.1038/s41398-020-0705-1. URL: https://doi.org/10.1038/s41398-020-0705-1.

Tian, Dongmei et al. (2020). "GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals". In: *Nucleic Acids Research*

48.D1, pp. D927–D932. ISSN: 0305-1048. DOI: 10.1093/nar/gkz828. URL: https://doi.org/10.1093/nar/gkz828.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

Toulopoulou, Timothea et al. (2010). "Impaired Intellect and Memory". In: *Archives of General Psychiatry* 67.9, p. 905. ISSN: 0003-990X. DOI: 10.1001/archgenpsychiatry.2010.99. URL: http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archgenpsychiatry.2010.99.

Trampush, J W et al. (2017). "GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium". In: *Molecular Psychiatry* 22.3, pp. 336–345. ISSN: 1476-5578. DOI: 10.1038/mp.2016.244. URL: https://doi.org/10.1038/mp.2016.244.

Tucker, George, Alkes L Price, and Bonnie Berger (2014). "Improving the power of GWAS and avoiding confounding from population stratification with PC-Select". eng. In: *Genetics* 197.3, pp. 1045–1049. ISSN: 1943-2631. DOI: 10.1534/genetics.114.164285. URL: https://pubmed.ncbi.nlm.nih.gov/24788602https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4096359/.

Valk, Ralf J P van der et al. (2015). "A novel common variant in DCST2 is associated with length in early life and height in adulthood." eng. In: *Human molecular genetics* 24.4, pp. 1155–1168. ISSN: 1460-2083 (Electronic). DOI: 10.1093/hmg/ddu510.

Vassos, Evangelos et al. (2017). "An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis". In: *Biological Psychiatry* 81.6, pp. 470–477. ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2016.06.028. URL: https://doi.org/10.1016/j.biopsych.2016.06.028.

Vaucher, J et al. (2018). "Cannabis use and risk of schizophrenia: a Mendelian randomization study". eng. In: *Molecular psychiatry* 23.5, pp. 1287–1292. ISSN: 1476-5578. DOI: 10.1038/mp.2016.252. URL: https://www.ncbi.nlm.nih.gov/pubmed/28115737https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5984096/.

Vilhjálmsson, Bjarni J. et al. (2015). "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores". In: *American Journal of Human Genetics* 97.4, pp. 576–592. ISSN: 15376605. DOI: 10.1016/j.ajhg.2015.09.001.

Visscher, Peter M et al. (2012). "Five years of GWAS discovery." eng. In: *American journal of human genetics* 90.1, pp. 7–24. ISSN: 1537-6605 (Electronic). DOI: 10.1016/j.ajhg.2011.11.029.

Visscher, Peter M. et al. (2017). "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1, pp. 5–22. ISSN: 0002-9297. DOI: 10.1016/J.AJHG.2017.06.005. URL: https://www.sciencedirect.com/science/article/pii/S0002929717302409.

Walder, Deborah J et al. (2014). "Genetic liability, prenatal health, stress and family environment: Risk factors in the Harvard Adolescent Family High Risk for Schizophrenia Study". In: *Schizophrenia Research* 157.1, pp. 142–148. ISSN: 0920-9964. DOI: https://doi.org/10.1016/j.schres.2014.04.015. URL: https://www.sciencedirect.com/science/article/pii/S092099641400187X.

Walsh, Tom et al. (2008). "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia." eng. In: *Science (New York, N.Y.)* 320.5875, pp. 539–543. ISSN: 1095-9203 (Electronic). DOI: 10.1126/science.1155174.

Ward, Joey et al. (2017). "Genome-wide analysis in UK Biobank identifies four loci associated with mood instability and genetic correlation with major depressive disorder, anxiety disorder and schizophrenia". eng. In: *Translational psychiatry* 7.11, p. 1264. ISSN: 2158-3188. DOI: 10.1038/s41398-017-0012-7. URL: https://www.ncbi.nlm.nih.gov/pubmed/29187730https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5802589/.

Warland, Anthony et al. (2019). "Schizophrenia-associated genomic copy number variants and subcortical brain volumes in the UK Biobank". In: *Molecular Psychiatry*, pp. 1–9. ISSN: 1359-4184. DOI: 10.1038/s41380-019-0355-y. URL: http://www.nature.com/articles/s41380-019-0355-y.

Weinberger, D R (1987). "Implications of normal brain development for the pathogenesis of schizophrenia." eng. In: *Archives of general psychiatry* 44.7, pp. 660–669. ISSN: 0003-990X (Print). DOI: 10.1001/archpsyc.1987.01800190080012.

Weiner, Daniel J et al. (2017). "Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders". In: *Nature Genetics* 49.7. ISSN: 1061-4036. DOI: 10.1038/ng.3863.

Wheeler, Anne L and Aristotle N Voineskos (2014). "A review of structural neuroimaging in schizophrenia: from connectivity to connectomics." eng. In: *Frontiers in human neuroscience* 8, p. 653. ISSN: 1662-5161 (Print). DOI: 10.3389/fnhum.2014.00653.

Willer, C. J., Y. Li, and G. R. Abecasis (2010). "METAL: fast and efficient meta-analysis of genomewide association scans". In: *Bioinformatics* 26.17, pp. 2190–2191. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq340. URL: http://www.ncbi.nlm.nih.gov/pubmed/20616382http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2922887https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq340.

Wing, J K et al. (1990). "SCAN. Schedules for Clinical Assessment in Neuropsychiatry." In: *Archives of general psychiatry* 47.6, pp. 589–93. ISSN: 0003-990X. URL: http://www.ncbi.nlm.nih.gov/pubmed/2190539.

Wockner, L F et al. (2015). "Brain-specific epigenetic markers of schizophrenia". eng. In: *Translational psychiatry* 5.11, e680–e680. ISSN: 2158-3188. DOI: 10.1038/tp.2015.177. URL: https://www.ncbi.nlm.nih.gov/pubmed/26575221 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5068768/.

Wray, Naomi R et al. (2013). "Pitfalls of predicting complex traits from SNPs". In: *Nature Reviews Genetics* 14.7, pp. 507–515. ISSN: 1471-0064. DOI: 10.1038/nrg3457. URL: https://doi.org/10.1038/nrg3457.

Wray, Naomi R. et al. (2018). "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model". In: *Cell* 173.7, pp. 1573–1580. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2018.05.051. URL: https://www.sciencedirect.com/science/article/pii/S0092867418307141?via{\%}3Dihub.

You, Chong et al. (2021). "Polygenic Scores and Parental Predictors: An Adult Height Study Based on the United Kingdom Biobank and the Framingham Heart Study." eng. In: *Frontiers in genetics* 12, p. 669441. ISSN: 1664-8021 (Print). DOI: 10.3389/fgene.2021.669441.

Zhang, Jian-Ping et al. (2019). "Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis." eng. In: *The American journal of psychiatry* 176.1, pp. 21–28. ISSN: 1535-7228 (Electronic). DOI: 10.1176/appi.ajp.2018.17121363.

Zheutlin, Amanda B. et al. (2019). "Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems". In: *American Journal of Psychiatry* 176.10, pp. 846–855. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.2019.18091085. URL: http://ajp.psychiatryonline.org/doi/10.1176/appi.ajp.2019.18091085.

Zhu, Zhihong et al. (2016). "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets". In: *Nature Genetics* 48.5, pp. 481–487. ISSN: 1061-4036. DOI: 10.1038/ng.3538. URL: http://www.nature.com/articles/ng.3538.

Zhu, Zhihong et al. (2018). "Causal associations between risk factors and common diseases inferred from GWAS summary data". In: *Nature Communications* 9.1, p. 224. ISSN: 2041-1723. DOI: `10.1038/s41467-017-02317-2`. URL: `http://www.nature.com/articles/s41467-017-02317-2`.