Interpretation and improvement of the current genetic epidemiology methodology for schizophrenia

Alexandros Rammos

BSc in Neuroscience

MSc in Neuroscience

A thesis submitted for the degree of Doctor of Philosophy (PhD) to the School of Genetics and Microbiology, Department of Genetics, Trinity College Dublin

December 2017

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Summary

Recent technological advances have allowed for the development and implementation of new methodologies in the investigation of the genetic epidemiology of schizophrenia. The study of the genetic aetiology underpinning schizophrenia has proven a challenging task for scientists for the last quarter century, with new methods developed in the last decade focusing on common variance that could help explain the genetic background of the disorder. Two of the most prominent methods in that field are the Polygenic Risk Scores (PRS) and the Genetic Restricted Maximum Likelihood (GREML) approach, applied through the GCTA (Genome-wide complex trait analysis) software. Three separate studies were carried out in the context of this thesis to investigate the means through which these methods operate and devise ways to optimise their function, as well as, improve their interpretability, in the context of schizophrenia research.

The first study focused on extending a gene-set-based approach in PRS analysis, using experimentally derived gene-sets as a basis for polygenic scores, in an effort to improve the interpretability of PRS analysis results and translate results from this type of analysis into a tool of better comprehending the genetic architecture of the disorder. Results from that analysis indicated that this was indeed a viable way of analysis and managed to identify specific gene-sets that contributed in excess in the genetic background of the disorder.

The second study investigated PRS application in the field, by utilising three of the most prominent PRS methodologies and comparing them in extended simulation scenarios with a range of different parameters. These simulations showed that no

method was able to capture all the variability in the simulated scenarios, as noise seemed to significantly impair the polygenic signal. Further simulations also showed that increasing the sample twentyfold in the simulations did not improve the estimates.

In the third study, GREML was applied in a population cohort to calculate heritability estimates for two polygenic characteristics. Subsequently, the GREML sensitivity to cryptic population substructure was investigated. Finally, a comparison between GREML and GREML-IBD, a recent extension of the original method, which takes into account rare variants to calculate heritability estimates, is made.

This thesis highlighted potential methodological limitations of two of the most commonly used approaches in schizophrenia research and through their implementation on both population and clinical-based samples proposed novel means of improving them. As the field of psychiatric genetic enhances the current knowledge on the genetic architecture of schizophrenia, research focusing on the understanding of the strengths and caveats of its methodology is necessary towards the advancement of the field.

Acknowledgements

First and foremost my sincere thanks go to Kristin K Nicodemus for constantly and reliably offering for her guidance and expertise throughout this PhD.

To Kevin Mitchell for his invaluable support, encouragement and guidance throughout this project.

To the Psychiatric Genomics Consortium and the Generation Scotland Cohort for granting permission of access to data for this project.

To the staff of SURFsara (LISA) and ECDF (Eddie) for providing computational resources necessary for the completion of this thesis.

To all the staff at the Smurfit Institute of Genetics who has helped me along the way.

I would like to thank my family and especially my parents and sister, for their encouragement and confidence in my ability.

And last but not least, a special thanks to Foteini for her unwavering support, patience and constant supply of coffee.

Publications

Rammos A, Mitchell KJ, Nicodemus KK. Comparing the effectiveness of current methods of polygenic score measurement. Genetic Epidemiology. 2017 Aug 21. doi: 10.1002/gepi.22062

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism-Spectrum Disorders
BD	Bipolar Disorder
BDNF	Brain Derived Neurotrophic Factor
CHD8	Chromodomain Helicase DNA Binding Protein 8
DISC1	Disrupted-in-Schizophrenia 1
DRD2/3/4	Dopamine Receptor D2/3/4
DSM	Diagnostic and Statistical Manual of Mental Disorders
FMR1	Fragile X Metal Retardation 1
FMRP	Fragile X Mental Retardation Protein
GCTA	Genome-wide Complex Trait Analysis
GS	Generation Scotland
GREML	Genetic- relatedness-matrix Restricted Maximum Likelihood
GRM	Genetic Relatedness Matrix
GWAS	Genome-wide Association Studies
HLA	Human Leukocyte Antigen
IBD	Identity by Descend
ICD	International Classification of Disorders
LD	Linkage Disequilibrium
LOD	Logarithm of Odds
PCA	Principle Component Analysis
PGC	Psychiatric Genomics Consortium
PRS	Polygenic Risk Score
SNP	Single Nucleotide Polymorphisms
TCF4	Transcription Factor 4
TDR	True Discovery Rate
TNF	Tumor Necrosis Factor

List of tables

Chapter 1

Table Number	Title	Page Number
1.1	Criteria for Schizophrenia in DSM V (2013), adapted from DSM V and Tandon et al (2013)	4
1.2	Schizophrenia Genome-wide Association Studies published up to 2017	30
1.3	Schizophrenia-Related Genome-wide Association Studies published up to 2016	32
1.4	Major Polygenic Risk Score Studies since 2009	42

Chapter 2

Table Number	Title	Page Number
2.1	Nested R^2 results for all individual studies for each gene-set	66

Chapter 3

Table Number	Title	Page Number
3.1	Median and Mean of PRS Nested R^2 by number of causative SNPs	91

Chapter 4

Table Number	Title	Page Number
4.1	Initial GRM REML Analysis	112
4.2	GRM REML “Raw” Analysis	113
4.3	GRM REML DBSCAN Analysis	114
4.4	GRM-IBD Initial Analysis (1 cM)	115
4.5	GRM-IBD “Raw” Analysis	116
4.6	GRM-IBD DBSCAN Analysis	117

List of figures

Chapter 1

Figure Number	Title	Page Number
1.1	Lifetime risk for schizophrenia in relation to degree of genetic relatedness. The risk increases as the degree of relatedness increases. (Adapted from Gottesman, 1991)	10

Chapter 2

Figure Number	Title	Page Number
2.1	Leave-One Out Cross-Validation Process	59
2.2	Flowchart for Polygenic Score Generation in Each Leave-One-Out Iteration	61
2.3	Overlap of gene-sets	64
2.4	R^2 and *p*-values from meta-analysis of all gene-sets	67
2.5	Q-Q Plot of $-\log_{10}$ P-value in the PGC2 Sample of 39 studies	69

Chapter 3

Figure Number	Title	Page Number
3.1	Example of LD pruning and clumping	80
3.2	Process of Sample selection and simulation of phenotypes	84
3.3	Violin Plot of Median PRS Nested R^2 by chromosome	89
3.4	Violin Plot of Median PRS Nested R^2 by sample size	90

Chapter 3 (continued)

Figure Number	Title	Page Number
3.5	Median PRS nested R^2 across P-value thresholds at the scenario with the least predictive power	92
3.6	Median PRS nested R^2 across P-value thresholds at the scenario with the most predictive power	93
3.7	Violin Plot of Median PRS nested R^2 by method	94
3.8	Median PRS nested R^2 in the extended sample model (N=40,000) by method including the true value of 0.50	96

Chapter 4

Figure Number	Title	Page Number
4.1	Flowchart of analysis for GCTA-GREML sensitivity	108

Contents

Chapter 1: Schizophrenia: Epidemiology and Genetic Epidemiology……..1

1. Introduction……..2

1.1. Overview and Clinical Features……2

1.1.1 Epidemiology of Schizophrenia……..5

1.1.2 Environmental Risk Factors of Schizophrenia……6

1.2 Genetic Epidemiology of Schizophrenia……9

1.2.1 Family Studies……...10

1.2.1.1 Paternal Age Effects/Inbreeding Effects……11

1.2.1.2 Overlapping Aetiology with other Psychiatric Disorders……...13

1.2.2 Twin Studies……..14

1.2.3 Adoption Studies……15

1.2.4 Linkage Studies……..16

1.2.4.1 Chromosome 5q22-q31……..17

1.2.4.2 Chromosome 6q21-q22……..18

1.2.4.3 Chromosome 6p24-p22……..18

1.2.4.4 Chromosome 8p22-p21……..19

1.2.4.5 Chromosome 13q14-q32……19

1.2.4.6 Chromosome 22q……20

1.2.4.7 The Present and Future of Linkage Studies……20

1.2.5 Association Analysis: The Candidate Gene Approach……..21

1.2.5.1 The HLA Region……22

1.2.5.2 *AKT1*……..23

1.2.5.3 *BDNF*…….23

1.2.5.4 *DISC1*…….24

1.2.5.5 Dopamine Hypothesis *(DRD2, DRD3 & SLC6A3*)……...25

1.2.5.6 *DRD4*……..26

1.2.5.7 *HTR2A*……27

1.2.5.8 *KCNN3*……...28

1.2.5.9 *TNF*…….28

1.2.6 Association Analysis: Genome-Wide Association Studies……...29

1.2.7 Epistatic Effects…….36

1.2.8 Copy number variants (CNVs)……..38

1.3 The polygenic Risk Score (PRS)……..39

1.3.1 Initial Concept and implementation……..39

1.3.2 PRS Applications in Schizophrenia, Psychiatric Genetics and other Complex Traits……..41

1.3.3 Challenges, criticism and Alternative Approaches……...43

1.4 GREML applied through GCTA……..46

1.4.1 Overview……...46

1.4.2 GREML applied through GCTA applications……..47

1.4.3 Challenges and Criticism……..49

1.5 Thesis Rationale……...50

1.5.1 Thesis Outline……...50

1.5.2 Aims of the Thesis……51

Chapter 2: The Role of Polygenic Molecular Gene-Sets in the Schizophrenia Working Group Genome-Wide Study of the Psychiatric Genomics Consortium (PGC2)……...52

2.1 Introduction……...53

2.1.1 Neuronal Gene-sets……....55

2.1.2 Non-Neuronal Gene-sets……57

2.1.3 Aims……...57

2.2 Methods……....58

2.2.1 The Schizophrenia Working Group of the Psychiatric Genomics Consortium 2 Case-Control GWAS……...58

2.2.2 Leave-One-Out (LOO) Polygenic Risk Score Analysis……60

2.2.3 Statistical Analysis……62

2.2.4 Meta-Analysis……...62

2.2.5 Simulation and Validation Studies……....63

2.3 Results ……64

2.3.1 Gene-Set Characteristics……...65

2.3.2 Polygenic Risk Score Analysis……66

2.3.3 Simulation and Validation Studies……70

2.4 Discussion……71

2.4.1 Gene-Set Analysis……71

2.4.2 Floor Effect……...73

2.4.3 Limitations……75

2.5 Conclusions……76

Chapter 3: Comparison of Current Methods of Polygenic Score Generation…....77

3.1 Introduction……..78

3.1.1 Background……...78

3.1.2 Aims ……81

3.2 Methods……82

3.2.1 Initial Sample……82

3.2.2 Simulation of Phenotypes……83

3.2.3 Polygenic Score Creation……..85

3.2.4 Linkage Disequilibrium Pruning and Thresholding……..85

3.2.5 Linkage Disequilibrium Clumping……85

3.2.6 True Discovery Rate Weights……86

3.2.7 Extended Simulation Application Sample……86

3.3 Results……...88

3.3.1 Effect of different LD structure in PRS estimates……88

3.3.2 Effect of Increasing Sample Size……..90

3.3.3 Effect of Different Number of Causal SNPs……91

3.3.4 Effect of Differing P-Value Threshold Levels……..91

3.3.5 Comparison of methods……93

3.3.6 Extended simulation application……95

3.4 Discussion……96

3.4.1 Limitations……99

3.5 Conclusions…….100

Chapter 4: Can possible cryptic population stratification affect GCTA GREML estimates? Examination of two traits in the Generation Scotland cohort……102

4.1 Introduction…….103

4.1.1 Background……..103

4.1.2 Aims……....105

4.2 Methods……..106

4.2.1 Generation Scotland……106

4.2.2 Outcome Variables……..106

4.2.3 Analysis Plan……...107

4.2.3.1 Software Used……..109

4.2.3.2 Basic Analysis……..109

4.2.3.3 Clustering Analyses……..110

4.3 Results……112

4.3.1 GRM REML……....112

4.3.1.1 Raw Approach……..113

4.3.1.2 DBSCAN……..114

4.3.2 GRM-IBD REML……....114

4.3.2.1 Raw Approach……..115

4.3.2.2 DBSCAN……..116

4.4 Discussion……...118

4.4.1 Limitations……...118

4.5 Conclusions……121

Chapter 5: Conclusion……122

5.1 Rationale and Aims of the Thesis……...123

5.2 Synopsis……..124

5.3 Sources of Data……...126

5.3.1 PGC2-schizophrenia...……126

5.3.2 Generation Scotland….……128

5.3.3 Potential Future Sources of Data…….130

5.3.3.1 Population Biobank: UK Biobank……... 130

5.3.3.2 Clinical Biobank: The String of Pearls Initiative…….131

5.4 Methodological Implications and Future Directions……..131

5.4.1 Polygenic Risk Scores…….132

5.4.2 GREML……...133

References……135

Appendices……...165

Appendix 2.1 Individual PGC Study Details……166

Appendix 2.2 Graphic representation of the LOO permutation process……..167

Appendix 2.3 Gene ontology Enrichment……168

Appendix 3.1 Script for weight generation taken from Mak et al……...170

Appendix 3.2A Table of all median nested R^2 of simulations for all conditions on the N= 500 samples……172

Appendix 3.2B Table of all median nested R^2 of simulations for all conditions on the N= 1000 samples……..174

Appendix 3.2C Table of all median nested R^2 of simulations for all conditions on the

N= 2500 samples……..176

Appendix 3.3 Table of median nested R^2 of simulations in the 50,000 samples…….178

Appendix 4.1A K-nearest neighbour distance plots for DBSCan ε neighbourhood detection. SNP GRMs K-nearest neighbour distance plots…….179

Appendix 4.1B K-nearest neighbour distance plots for DBSCan ε neighbourhood detection. IBD GRMs K-nearest neighbour distance plots……180

Appendix 4.2 Script to convert GRMs from GCTA into a meaningful input for DBSCAN……..181

CHAPTER 1:

Schizophrenia: Epidemiology and Genetic Epidemiology

1. Introduction

Severe mental illness, including schizophrenia, is a major public health issue due to its economic, psychological and social impact on the population. Despite the research interest in the study of the mechanisms behind schizophrenia genetics, a number of methodological issues in accounting for the genetic component of schizophrenia highlight the limitations of the current understanding of genetic architecture of schizophrenia.

The main aim of this chapter is to explore the variations in measurement of schizophrenia symptomatology, epidemiology and genetics using a range of sources and approaches, through the investigation of the strengths and limitations of previously implemented procedures. The outline of this thesis will also be elaborated and the overall aims will be presented.

1.1. Overview and Clinical Features

Schizophrenia, or dementia precox, is a serious and extremely debilitating mental health disorder, characterised by persistent abnormal beliefs, hallucinations, disorganised thought and speech, as well as avolition and distorted emotional response (Plomin , 2008). It is the one of the most severe forms of psychopathology and is considered to be the most debilitating among mental health disorders (Ustün 1999; Ustün et al, 1999).

Defining schizophrenia and its diagnostic criteria has been an ongoing process with diagnostic manuals such as the International Classification of Disorders (ICD, 1992) and the Diagnostic and Statistical Manual of Mental Disorders (DSM) being constantly

updated to match current scientific knowledge. The main reason for this has been the heterogeneity of this disorder. Recent definitions of schizophrenia in the DSM (DSM IV-TR) have been shown to be reliable with a significant degree of validity (Tandon et al, 2009; Nasrallah et al, 2009). This instrument has been shown to be highly stable, with 80-90 percent of individuals receiving a diagnosis based on these criteria retaining it for 1-10 years (Bromet et al, 2011). The success of that instrument was made evident by the fact that the new edition of the Manual (DSM V, 2013) retained essentially the same criteria for schizophrenia (Tandon et al, 2013). In detail, the diagnostic criteria that are currently required for a diagnosis of schizophrenia by the DSM-V manual are presented in Table 1.1 below.

Table 1.1 Criteria for Schizophrenia in DSM V (2013), adapted from DSM V and Tandon et al (2013).

<table>
<tr><td colspan="2">Criteria for Schizophrenia in DSM V (2013)</td></tr>
<tr><td><u>Criterion A:</u>
Two (or more) of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated)
At least one of these should include 1 to 3:
1. Delusions
2. Hallucinations
3. Disorganized speech
4. Grossly disorganised or catatonic behaviour
5. Negative symptoms, (i.e. diminished emotional expression or avolition)</td><td><u>Criterion B:</u>
For a significant portion of the time since the onset of the disturbance, one or more major areas of functioning, such as work, interpersonal relations, or self-care, are markedly below the level achieved prior to the onset (or when the onset is in childhood or adolescence, failure to achieve expected level of interpersonal, academic, or occupational achievement).</td></tr>
<tr><td><u>Criterion C:</u>
Continuous signs of the disturbance persist for at least 6 months. This 6-month period must include at least 1 month of symptoms (or less if successfully treated) that meet Criterion A (i.e., active-phase symptoms) and may include periods of prodromal or residual symptoms.
During these prodromal or residual periods, the signs of the disturbance may be manifested by only negative symptoms or by two or more symptoms listed in Criterion A.</td><td><u>Criterion D:</u>
Schizoaffective disorder and depressive or bipolar disorder with psychotic features have been ruled out because either 1. no major depressive or manic episodes have occurred concurrently with the active phase symptoms; or 2. if mood episodes have occurred during active-phase symptoms, their total duration has been brief relative to the duration of the active and residual periods.</td></tr>
<tr><td><u>Criterion E:</u>
Substance/general medical condition exclusion: The disturbance is not attributed to the direct physiological effects of a substance (e.g., a drug of abuse, medication) or another medical condition.</td><td><u>Criterion F:</u>
If there is a history of autism spectrum disorder or other communication disorder of childhood onset, the additional diagnosis of schizophrenia is made only if prominent delusions or hallucinations are also present for at least
1 month (or less if successfully treated).</td></tr>
</table>

1.1.1 Epidemiology of Schizophrenia

Investigation on the epidemiological factors defining schizophrenia and psychosis-spectrum disorders has been an ongoing process over the last century. Prevalence of schizophrenia seems to differ across different countries and, although migration seems to play a definitive role into it (Saha et al, 2005), countries from higher latitude tended to have significantly higher prevalence of schizophrenia, compared with countries of lower latitude. Lifetime risk of schizophrenia was estimated to have a median value of 4 per 1000 persons (McGrath et al, 2008).

A major meta-analysis (McGrath et al, 2004), including all studies reporting the incidence of schizophrenia from 1965 to 2001 was performed; urbanicity or gender did not seem to be statistically significant factors in terms of the incidence of the disorder. However, migration status appeared as an important factor, in terms of developing a mental disorder. A separate meta-analysis conducted by Cantor-Graae et al (2005), yielded a relative risk of 2.9 (95% CI=2.5-3.4) to develop schizophrenia for migrants, irrespective of whether they were first or second generation.

There also seem to be distinct differences on the basis of gender in psychosis-spectrum disorders and the ways they manifest themselves. Age of onset of the disorder is the most distinct among these characteristics, with males typically developing symptoms in average 3-4 years earlier than females (Hafner et al, 1989; Hafner et al, 1992; Hafner et al, 1994), a finding independent of cultural background or definition of schizophrenia (Leung et al, 2000). In terms of symptomatology, female patients tend to be more susceptible to affective psychosis-spectrum disorders (Amminger et al, 2000), while males tend to usually have a poorer prognosis (Goldstein, 1988; Tseliou et al, 2015) and

greater deficits in social ability both before first admission (Dworkin, 1990) and later on during the life course (Thorup et al, 2007; Hui et al, 2014).

1.1.2 Environmental Risk Factors of Schizophrenia

There has been a substantial body of research aiming to investigate which environmental insults might be detrimental to the development of schizophrenia and how these might have an influence on the disorder itself and subsequent symptomatology. These studies are usually of retrospective nature and utilise record linkage and hospital records as a measure of the association between psychiatric outcomes and early life events. One of the most famous studies conducted in this manner was about the Dutch Famine that occurred during the Nazi occupation of Netherlands during World War II (Stein et al, 1975; Susser et al, 1996). The Dutch Famine had a particularly devastating effect on children who were affected by it during their mother's pregnancy. Compared to a control matched cohort, those in the second trimester of pregnancy at the time of exposure to the famine had a twofold risk of developing schizophrenia (Susser et al, 1996; Hoek et al, 1998). Additionally, there was a similar effect size in the increase of non-clinical schizoid personality traits in the rest of the cohort born during that period (Hoek et al, 1996). These findings along with neurological defects in the sample, well above the population norm (Stein et al, 1975), provide a link between negative long-term outcomes and malnutrition in the prenatal period (Brown et al, 2008).

Risk factors that have consistently been shown to be linked with the development of a psychosis-spectrum disorder in later life include: 1)season of birth, with a relative risk increase of 10 percent for children born in winter (Davies et al, 2003), 2) birth

complications including complications of the pregnancy itself, abnormal foetal development, or delivery complications (Geddes et al, 1995), 3) paternal age (Zammit et al, 2003), with increased paternal age reported of raising the relative risk as much as 3 times in comparison with younger parents (Brown et al, 2002), and 4) infection of the mother during the second trimester of the pregnancy (Yolken et al, 1995).

There have also been a multitude of studies that have aimed to investigate how childhood conditions may play a role in later life mental health outcomes. Such studies have looked at how environmental factors, both at household and at school level might influence mental health in later life; such factors have included minority status in the community (Bourque et al, 2010) personal or family history of migration (Cantor-Graae & Selten, 2005), parental communication deviance (de Sousa et al, 2014), exposure to childhood adversity and trauma mediated by childhood abuse or neglect (Bendall, et al, 2007; Morgan & Fisher, 2006). A recent meta-analysis has indicated that there is no specific type of childhood adversity that is stronger than the others, but it mostly the age of exposure and multi-victimization that seem to be more strongly related to psychosis risk (Varese et al, 2012).

Finally, research on environmental factors of schizophrenia has branched at looking at proximal factors of schizophrenia; that is environmental insults occurring during or just before the onset of the disorder. These are insults that are not causal to the development of the disorder per se; rather they act as catalyst events that contribute to the onset of the disorder and are only important in the event of underlying susceptibility. These types of events have been shown to potentially be able to increase the risk almost three-fold; however research on adult life events has been generally limited with very few

studies assessing their role in psychosis (Beards et al, 2013).

Based on these findings and the extrapolated effect that the environment can play in the development of a mental disorder, there have been two key theories that have tried to explain this phenomenon; the Diathesis-Stress model, first developed in the 19th century and later considered and applied in mental health disorders (Monroe et al, 1991) and the Differential Susceptibility model first proposed in 1997 (Belsky, 1997). The Diathesis-Stress model (Monroe et al, 1991; Zuckerman, 1999) proposes that some individuals, who have some predisposing factor, either behavioural, genetic or endophenotypic, are more vulnerable to environmental stressors which might subsequently culminate in the development of mental illness. Differential Susceptibility, which is an alternative hypothesis, builds on the Diathesis-Stress model (Belsky & Pluess, 2009), expanding on it on the basis of positive psychology (Diener & Biswas-Diener, 2008). It proposes an evolutionary approach according to which "vulnerable individuals" identified by the Diathesis-Stress model are simply individuals with greater developmental plasticity and therefore more likely to be influenced by their environmental influences (Ellis, 2011). This is consistent with the evolutionary model of development, which proposes that the environment regulates the individual's survival strategies in stressful conditions, although these strategies may prove counterproductive for the individual or the society in the long run (Hinde et al, 1990; Main, 2000). Such modes of plasticity have been observed in many studies including Gluckman et al (2008), which showed that foetal malnutrition resulted in insulin resistance, protective in an environment with scarce resources, while having the potential to increase the risk of cardiovascular attacks in a rich one.

It is important to consider both the Stress-Diathesis and Differential Susceptibility hypotheses not only from a purely psychological/behavioural framework but also on the basis of the underlying biology they are hinting at, given that both theories are essentially, at least partially, alluding to Gene-Environment interactions, which will be discussed later in this chapter and have been shown to be relevant on the manifestations of, at least some mental health disorders (Caspi et al, 2002).

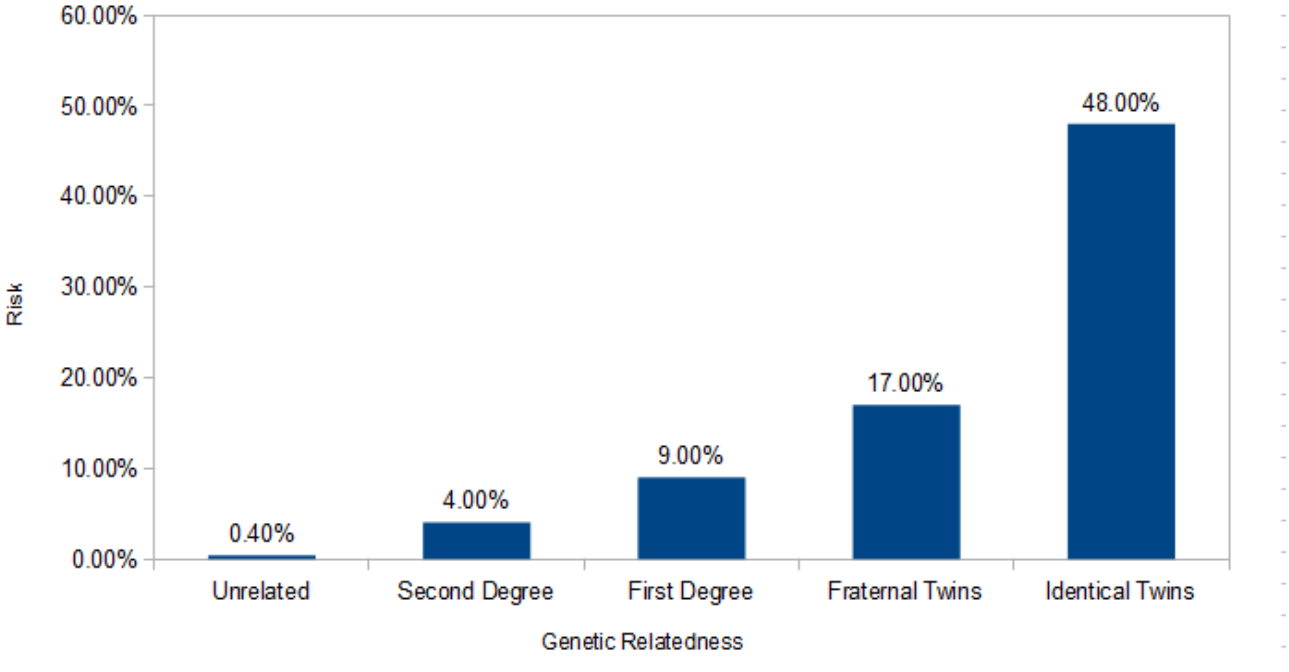
1.2 Genetic Epidemiology of Schizophrenia.

While there has been on-going research investigating the underlying environmental factors that underpin psychosis-spectrum disorders, an equally large body of work has been trying to investigate the biological and genetic aetiology of the disorder. This section will review the different methods that have been used to understand the complex genetic architecture of the disorder. This will be discussed in detail and will be presented in chronological order beginning with family studies, factors including paternal age and inbreeding effects, as well as co-morbidity with other mental health disorders. Afterwards, twin and adoption studies will be elaborated upon. Subsequently, more contemporary genetic methods will be discussed: candidate gene studies, linkage studies as well genome-wide association studies. Finally, this part will conclude with description of epistatic effects (gene (x) gene and gene (x) environment), followed by a look at next generation sequencing. Polygenic Risk Scores as an approach to understand the underlying architecture of schizophrenia will not be included here as the second part of the introduction will focus solely on them.

1.2.1 Family Studies

Family studies are often used as a means to identify the level of risk of relatives developing mental disorders that members of their family are suffering from. The most frequent type of experimental design in family studies are case-control family studies. These, are studies that measure the relative risk of a family member of a person with a mental illness to develop that disorder as opposed to a member of a family with no incidences of that mental disorder. Family studies have also, historically, been used to determine the risk of a disorder being inherited to the offspring. One of the hallmark characteristics of schizophrenia has been its complex inheritance pattern, first observed through this type of studies. Gottesman (1991) presents a comprehensive picture of familial inheritance in schizophrenia presented also here, adapted in Figure 1.1 below.

Figure 1.1 Lifetime Risk for schizophrenia in relation to degree of genetic relatedness. The risk increases as the degree of relatedness increases. (Adapted from Gottesman, 1991).



In comparison to the risk of less than 1 percent in the general population, schizophrenia risk increases as the degree of familial relatedness increases: 4 percent for second degree relatives and 9 percent for first degree relatives. There seems to be some interesting differences between first degree relatives, with parents having a median risk of 6 percent, siblings 9 percent and offspring 13 percent (Plomin et al, 2008). The low percentage of parents could be attributed to the fact that individuals with a schizophrenia diagnosis are less likely to marry. In contrast, the high percentage of offspring showcases the fact that when patients with schizophrenia do marry their children are also highly likely to be suffering from the disorder.

Additionally, family designs have allowed for retrospective studies of children whose mothers were suffering from the disorder and had been labelled as a "high risk" group for developing the disorder. Such a study design by Parnas and colleagues (1993) followed up 200 children of such origin for more than 30 years and showed a risk of 16 percent in comparison to a control group, as well as a higher likelihood of a perinatal event (Cannon et al, 1993) and attention deficit in childhood (Hollister et al, 1994).

1.2.1.1 Paternal Age Effects/Inbreeding Effects

Two additional factors closely linked with familial structure that have been well established as role-players in the development of schizophrenia are paternal age and inbreeding.

There have been two major retrospective studies in the last decade that have thoroughly investigated the effects of paternal age. The first, a record linkage (Zammit et al, 2003)

of 50,000 adolescents for schizophrenia admissions between 1970 and 1996 highlighted that there was an increased risk for schizophrenia linked with advanced paternal age in a dose-response relationship. More specifically, with every 10-year increment of paternal age, there was also an increase in the odds for developing the disorder by 1.3 times. Another similar retrospective cohort (Malaspina et al, 2001) was also able to identify a similar pattern of risk increase in 88,000 Israeli-born individuals (658 cases), with risk increasing proportionally to the age of the father. This elevated risk on the basis of the age of the father has been attributed to de novo genetic mutations arising in the paternal germ cells with increasing frequency, as paternal age increases. Additionally, it is worth remarking here that such de novo mutations have been identified and linked to a host of other more rare neurological defects (Malaspina, 2001).

Inbreeding as a practice is regarded as a taboo and is generally socially frowned upon in Western societies. Thus, most data for inbreeding effects in the development of schizophrenia are derived from non-Western societies where this practice is more commonly observed and more socially accepted. Recently, two studies from Arabic nations were published on that matter; the first (Bener et al, 2012), was conducted as a case control recall design in Qatar and showed a considerably higher inbreeding coefficient among psychiatric patients in comparison to a matched control sample (41% consanguinity vs 29% in the control sample). The second study (Mansour et al, 2010) conducted in Egypt, showed similar estimates for consanguinity in that sample, adding self-reported measures of inbreeding in a case-control design. This is consistent with biological decline due to inbreeding and concentration of deleterious alleles in inbred populations.

1.2.1.2 Overlapping Aetiology with other Psychiatric Disorders

There has been an extensive discussion on the overlap of schizophrenia with other psychiatric disorders and most prominently Bipolar Disorder (BPD) and Autism-Spectrum Disorders (ASD), with a number of family studies investigating whether schizophrenia aggregates with those disorders in families and up to what extend there is a shared heritability element between those three broad categories of mental health disorders. There has been no evidence of overlap in family studies between ASD and schizophrenia; despite that, there seem to be common molecular and genetic pathways between the two disorders (Carroll et al, 2009), which warrant further investigation as to the links between the two disorders. BPD and schizophrenia, on the other hand, not only share common clinical features (Hulshoff Pol et al, 2012) and a degree of genetic overlap (Carroll et al, 2009; Lencz et al, 2013) but also appear to share some familial features. Schizophrenia proband family members are at an increased risk of schizophrenia, schizoaffective disorder (an intermediate phenotype disorder with both psychotic and affective elements) and recurrent unipolar depressive episodes (Gershon et al, 1988; Maier et al, 1993). Bipolar proband family members are also at increased risk of the disorder itself but also of schizoaffective disorder and recurrent unipolar episodes (Gershon et al, 1982; Winokur et al. 1995). Thus, there seems to be a significant indirect overlap of schizophrenia and BPD in their propensity to increase the risk for schizoaffective disorder and recurrent unipolar episodes in proband families.

1.2.2 Twin Studies

The next major study design that has been used to determine genetic influences and separate them from environmental influences is the twin study design. Identical (monozygotic) twins share 100 percent of their genetic make-up. Fraternal (dizygotic) twins, on the other hand share, on average 50 percent of their genes, similar to first degree relatives. Twin studies utilise the equal environments assumption, which assumes that both types of twins are roughly affected on the same level by the environment, when reared in the same family (Derks et al, 2006). The first study to utilise the twin design, comparing monozygotic and dizygotic twins was conducted in 1924 by Merriman and aimed to estimate the genetic influence in intelligence quotient (Merriman, 1924). Although the twin study design has been extensively implemented in many fields, psychiatric genetics being one of them, there have been some criticisms pertaining to the equal environments assumption, regarding two main issues that arise from it. First, twins tend to be born about a month prematurely in comparison to non-twin births (Phillips, 1993) and secondly twins tend to have a 30 percent lower birth weight than non-twin babies (MacGillivray et al, 1988).

Schizophrenia studies on the basis of the twin design have been critical in establishing the level of heritability for the disorder. The concept of heritability was generated from the question of how much genetic influences contribute to a trait. Heritability is the statistic estimate of that genetic effect size and can be calculated as the proportion of phenotypic variance accounted by genetic differences. Sullivan and colleagues (2003) carried out a meta-analysis of 14 twin studies and through the use of a liability threshold model were able to establish a heritability of liability at 80 percent. In essence, this indicates that roughly 80 percent of variance can be solely attributed to

genetic factors. The liability threshold model is a classical model implemented in many disorders that posits that the genetic risk has a normal distribution, but manifests as a disorder only when a critical threshold of risk factors is accumulated (Smith et al, 1974).

Nevertheless, it is worth noting that schizophrenia has only ~50 percent concordance in monozygotic twins, providing a clear context where environmental effects also factor heavily in schizophrenia. Based on the twin design, there have also been additional research attempts to understand the factors behind the complementary non-genetic liability, that may lead to only one twin developing the disorder. Two studies focusing on pairs of discordant twins have demonstrated that differences in manifesting the disorder seemed to be linked to birth complications and subtle brain structure differences among them (Torrey et al, 1994; Mosher et al, 1971).

1.2.3 Adoption Studies

Adoption studies constitute another major type of study design that has been used extensively in psychiatric genetics. These studies were developed on the basis of improving understanding of whether characteristics and behavioural traits are inherited or acquired through shared environmental influences within the familial context. Adoption creates a two-fold opportunity for the researcher: Initially, this allows to examine individuals who, although genetically related, do not share a common environment and thus, to calculate an accurate estimate of genetic influence irrespective of environment. Additionally, it establishes family members that do share family environment, but are not genetically related. Similarity between those individuals and their adopted family constitutes a purely environmental influence. The first adoption study was conducted for schizophrenia (Heston, 1966), in an attempt to highlight the

importance of genetics in the aetiology of the disorder. In this study 47 adopted-away children of women that were hospitalised for schizophrenia were interviewed. Of those 47, five had developed schizophrenia (10.6 %), an incidence markedly higher than the population average and the matched control group of that study (0 %). These results have been elaborated further in two additional extensive studies. The first by Tienari and colleagues (2004), showed similar results to the initial study and followed a similar study design, demonstrating a 10 percent incidence of schizophrenia for adopted-away children of probands and about 1 percent incidence for control adoptees. The second study by Kety (1987; Kety et al, 1994), implemented a reverse design focusing instead on adoptees that had been diagnosed with schizophrenia (N=47) and an equally sized non-psychiatric control group. The parents of the proband group had a 5 percent rate of schizophrenia (14/279 biological first degree relatives), while there were almost no first degree biological relatives with schizophrenia for the control group (1/234). For both groups there were no people with schizophrenia in their respective adopted families, highlighting the overwhelming importance of genetic influences over shared environment in the development of the disorder.

1.2.4 Linkage Studies

With the advent of new technological advances in the eighties, genetic linkage analysis was one of the initial tools used to localise genes linked to disorders in chromosomes. This approach is based on the initial observation that genes that reside physically close on a chromosome remain linked during meiosis. By following the segregation of alleles from affected individuals to offspring, loci relevant to a disorder could be identified (Riley, 2004). There were two types of analysis that tried to identify linkage, parametric and non-parametric analyses, on the basis of logarithm of odds (LOD) scores. These

strategies proved very successful in identifying Mendelian inheritance disorder loci, but not in schizophrenia which, as is evident from figure 1.1, does not follow typical Mendelian patterns. Despite that, there were a number of regions where linkage was established in a subgroup of studies. However, as even the largest linkage studies (Badner & Gershon, 2002; Lewis et al, 2003; Segurado et al, 2003; Ng et al, 2009), failed to replicate significant findings and pinpoint specific loci implicated in the disorder, linkage analysis fell into disuse, although there has been a resurgence of the technique recently, in the wake of Whole Genome Sequencing technological advances (Ott et al, 2015). Some of the most prominent regions of linkage that were proposed as loci of interest for schizophrenia during that era are discussed below.

1.2.4.1 Chromosome 5q22-q31

The first study that was successful in identifying a positive locus that had a link with schizophrenia was in 1988 by Sherrington and colleagues. Through the use of genetic linkage, they were able to identify a locus in the long arm of chromosome 5 that was linked to schizophrenia in seven Icelandic and British families that had multiple incidences of schizophrenia. However, the linkage was not replicated either by other groups or themselves (Kennedy et al, 1988; St Clair et al, 1989; Detera-Wadleigh et al, 1989; Macciardi et al, 1992). In contrast to that, there have been two subsequent studies that have found some additional suggestive evidence of a potential linkage between that region and schizophrenia. Straub et al (1997) were able to identify high linkage with that region in an Irish sample of narrow defined cases. Nevertheless this result was not as significant when a broader psychosis diagnosis was considered. Schwab et al (1997) also found evidence of linkage from markers in the same region in 14 families from Germany, with an observed LOD score of 1.8 which was, however, decreased to 1.27 in

an independent sample of 40 families.

1.2.4.2 Chromosome 6q21-q22

Evidence for potential susceptibility in locus 6q21-q22 came from an initial study examining 53 US families (Cao et al, 1997). It is noteworthy that this result was replicated internally in an independent sample of 63 families from the National Institute of Mental Health (NIMH) Schizophrenia Genetics initiative. A follow-up study by the same group (Martinez et al, 1999) was also able to replicate the finding, albeit at a marginal significance level using 43 new pedigrees. However, later meta-analyses of genome scans of linkage found no evidence implicating that region with the disorder (Badner and Gershon, 2002; Lewis et al, 2003; Segurado et al, 2003; Ng et al, 2009).

1.2.4.3 Chromosome 6p24-p22

Evidence for linkage in area 6p24-p22 first came up from Straub and colleagues (1995) that used 265 Irish pedigrees. The evidence for that linkage was stronger when "broad" psychosis-spectrum diagnoses were included and even stronger when non-psychotic psychiatric diagnoses were considered in the analysis. Multiple reports implicating that region were published over the next years, after the initial linkage was established (Moises et al, 1995; Levinson et al, 1996; Maziade et al, 1997; Lindholm et al, 1999), with varying levels of success in replicating the initial finding; in most cases, the finding was replicated in parts of the sample but not in all pedigrees. From subsequent meta-analyses of genome scans of linkage, only one (Lewis et al, 2003) found some weak links to that region.

1.2.4.4 Chromosome 8p22-p21

Region 8p22-p21 was first implicated to be in linkage with schizophrenia in the Maryland family sample (Pulver et al, 1995). In this study, they examined 520 loci for potential susceptibility loci for schizophrenia. The finding of this region as a potential locus connected to schizophrenia was further explored in a multi-site replication collaborative study (Levinson et al, 1996). However, despite positive replication results and later confirmation from a number of big meta-analyses (Badner & Gershon, 2002; Lewis et al, 2003), the evidence for linkage in the region was not as strong, in terms of LOD size and not the same locus was linked in all studies that found evidence in the general region.

1.2.4.5 Chromosome 13q14-q32

The initial investigation where this chromosomal region was first pinpointed as a potential candidate for linkage was a mixed sample of British and Japanese pedigrees (Lin et al, 1995). This linkage became of particular interest to researchers, as within it, there was the gene coding for a serotonin receptor *(HTR2A*). The follow-up study by the same group (Lin et al, 1997) was able to only partially replicate the initial finding, as only a European subset of the sample supported the linkage. Furthermore, the markers that were linked to the finding were too far apart in that sub-sample, with a null region between them. Further studies on that region by other research groups also found mixed results (Kalsi et al, 1996; Shaw et al, 1998; Brzustowicz et al, 1999), with no consensus regions and linkage reports positive results spreading over a wide region.

1.2.4.6 Chromosome 22q

Initial discovery of potential linkage for chromosome 22q also came from the Maryland Family Sample (Pulver et al, 1994). This finding yielded two positive replicates on the same year (Coon et al, 1994; Polymeropoulos et al, 1994) from Utah and England/Wales respectively. This linkage finding was of particular interest at the time, as that region has been notably linked with Velo-cardio-facial Syndrome (VCFS). VCFS, also known as DiGeorge syndrome is a syndrome caused by microdeletion of part of the long arm of chromosome 22, with genes being affected by that deletion ranging between 30 and 50 (Maynard et al, 2008). Prominent features of the disorder include cardiac defects, abnormal facial expression, aplasia of the thymus, palate abnormality and hypoparathyroidism. Additionally, 90 percent of those affected have learning difficulties (Lindsay, 2001) and a high percentage of people born with VCFS are also schizophrenia probands (Murphy et al, 1999). Nevertheless, subsequent linkage studies in the region have failed to uncover any further strong evidence supporting linkage with that region (Jorgensen et al, 2002; Mowry et al, 2004). A meta-analysis conducted for genome scans also failed to consistently implicate that region in any tangible manner (Lewis et al, 2003; Ng et al, 2009).

1.2.4.7 The Present and Future of Linkage Studies

With newer, faster and cheaper techniques for genome-wide association studies developed in the first decade of the 20th century, linkage studies took a back seat in the developments of psychiatric genetics. A review of the meta-analyses of linkage genome scans for schizophrenia (Crow, 2007) demonstrated that even large meta-analyses failed to identify consensus sites of linkage. This period was a turning point for schizophrenia

research which shifted gears and instead of trying to identify rare variants with a large effect, became more interested in the identification of multiple common loci with small effects. Recently, there has been some resurgence in the interest for linkage analysis for the investigation of rare variants associated with complex traits that have a high degree of penetrance. This has been made possible with the combination of linkage with whole genome sequencing and has led to the identification of susceptibility genes for familial hypertension (Louis-Dit-Picard et al, 2012) and hearing impairment (Santos Cortez et al, 2013), among others. It remains to be seen whether this new resurgence of linkage will be extended into psychiatric genetics once more, and how this will further enable investigation of rare familial mutations that might be linked to psychosis-spectrum disorders.

1.2.5 Association Analysis: The Candidate Gene Approach

Back in the late 20th century, as newer types of technology started to become available, there was an upsurge of research that aimed to investigate the genetic aetiology of schizophrenia and identify genes that could be implicated in the disorder, beyond linkage analysis. One of the initial approaches was the candidate gene approach. This approach, tested for heterogeneity in pre-determined small localised areas, usually in case-control designs on the basis of either a pre-existing theoretical framework or an initial linkage study, that implicated a region showing some promise. These were implemented in the general population and did not use family pedigrees, which was the hallmark of linkage studies. They were able to use a larger number of markers compared to linkage studies but were limited to small areas of the genome due to the technological and financial limits of the era before genome-wide association scans. Below some of the more prominent candidate genes from that era are presented along

with some additional later evidence regarding possible connections with schizophrenia.

1.2.5.1 The HLA Region

One of the first genomic areas that were investigated by the use of the candidate gene rationale was the Human Leukocyte Antigen (HLA) region of the human genome that had been implicated as having a critical role in the development of multiple disorders (McGuffin & Stuart, 1986). Initial investigation of the region was able to uncover associations of the region with Narcolepsy (Honda, 1988), Multiple Sclerosis (Hillert & Olerup, 1993) and other neurological or neuropsychiatric phenotypes. Pertaining to the psychiatric genetics of schizophrenia, in the last two decades of the last century more than 60 association studies between schizophrenia and that region were conducted. The association with the area had, on average a modest effect on risk (RR of 1.7-2.2) and despite multiple attempts no specific reasons for the implication of that region were identified. This region remains an important focal point for psychosis studies to this day, as recent Genome-Wide Association Studies (GWAS) (Purcell et al, 2009; Ripke et al, 2011; Ripke et al, 2014) also did report Single Nucleotide Polymorphisms (SNPs) in that region in their findings. Explanation as to the reasons implicating this region with psychosis-spectrum disorders have focused on the commonality of pathways between inflammation and schizophrenia. There has been some evidence suggestive of a measure of neuro-inflammation in schizophrenia (Bechter et al, 2010, Muller and Schwarz, 2010; Meyer et al, 2011) and elevated level of cytokines in the blood of probands (Potvin et al, 2008) compared to controls. Additional evidence of the importance of the region was presented by Sekar and colleagues (2016) which showed that the association of the disorder with the region is, in part driven from the gene *C4* (component 4) found within the region, which plays a critical role in human immune

response. These findings point towards a potential commonly shared pathway between inflammation and schizophrenia that even after almost four decades of looking at the HLA region bears further investigation.

1.2.5.2 *AKT1*

Active interest in the *AKT/GSK3β* signalling cascade for schizophrenia research did not materialise until 2004. Previously, it was well established that this signalling cascade was a molecular target of lithium and thus an important target in the research for mood disorders. Initial interest was generated from the findings of Emamian et al (2004), who reported evidence for reduced *AKT1* levels and *GSK3β* phosphorylation in peripheral blood cells of individuals with schizophrenia. This was expanded in the same report with an association analysis on the locus of the gene where results of moderate significance were detected. This finding was further elaborated into a number of other association studies, and a total of 13 association studies with that gene were reported (Farrell et al, 2015). Despite not showing up in the recent PGC mega-GWAS (Ripke et al, 2014), *AKT1* signalling has remained of interest to psychiatric genetics, supported by a number of findings in addition to the original reports and the association analyses that followed. Tan et al (2008) linked the *AKT1* variant from Emamian et al (2004) with cognitive abilities (IQ, executive functioning) in healthy controls, while Thiselton et al (2008) also found evidence of reduced *AKT1* levels in the prefrontal cortex, hippocampus and peripheral blood of patients suffering from schizophrenia.

1.2.5.3 *BDNF*

Brain Derived Neurotrophic Factor (*BDNF*) was first implicated with schizophrenia on the basis of its critical role in neuronal development and in neuroplasticity (Green et al,

2011). It was initially implicated with schizophrenia in a study conducted by Sasaki and colleagues (1997) where they examined an association of a SNP site located in proximity to *BDNF* between 60 probands and an equal number of controls, with no significant evidence of association. Research on *BDNF* and possible connections to mental health continued on, mainly due to its central role in neural development, rather than some hallmark finding implicating it with psychosis. Eventually, such a finding was uncovered as Rosa et al (2006) implicated the Val66Met substitution polymorphism with schizophrenia risk in a family study. This polymorphism, subsequently was very important to *BDNF* studies in psychosis and resulted in many studies investigating cognitive deficits in psychosis, (Egan et al, 2003; Ho et al, 2006, Baig et al, 2010; Zhang et al, 2012). In a recent meta-analysis of all studies examining the association of the polymorphism with aspects of cognition (Ahmed et al, 2015), there was no significant difference between carriers of the Met allele and Val homozygotes.

1.2.5.4 *DISC1*

Disrupted-in-Schizophrenia 1 (*DISC1*) was initially discovered in a Scottish family that was reported to present with a surprisingly increased incidence of schizophrenia and was followed up for 3 decades (Blackwood et al, 2001). The gene was first isolated and identified in 2000 (Millar et al, 2000) and a translocation at that locus was shown to be a major contributor to the development of schizophrenia or other mental disorders within that family (Porteous et al, 2011). However, this discovery was shown to be important only in rare familial variants of schizophrenia and not in population sporadic cases, as *DISC1* was not prominent in either of the big GWAS conducted recently (Purcell et al, 2009; Ripke et al, 2014) and there have been questions over its overall

usefulness (Sullivan, 2013). Nevertheless, as its catalytic role in these rare sporadic cases of schizophrenia is indisputable, *DISC1* has been used as a knockout target in mouse modelling of schizophrenia (Tomoda et al, 2016). Despite the obvious caveats of using a rare mutant as a truly representative model of schizophrenia (Wong et al, 2016), it still remains the most popular way to investigate schizophrenia in animal models (Tomoda et al, 2016). Further research into the biological functions of *DISC1* have revealed a very broad spectrum of mechanisms that the *DISC1* gene may be a part of, including neural development and neuronal signalling (Bradshaw & Porteous, 2012).

1.2.5.5 Dopamine Hypothesis (*DRD2*, *DRD3 & SLC6A3*)

The concept of a dopaminergic network that was a major contributor in schizophrenia started developing in the 1960s, when a study (Carlsson et al, 1963) uncovered that one of the actions of first generation antipsychotic drugs was the blocking of dopaminergic receptors. This generated the first hypothesis on the involvement of dopamine in psychosis (van Rossum, 1966). Essentially, it proposed that actions of dopaminergic receptors were central to the development of schizophrenia and that they could very well be one of the underlying causes for its manifestation. Since this initial report, there was a massive amount of research literature surrounding this hypothesis for the next 30 years. In the era of association analysis, it was natural that dopamine related genes would be among the first to be investigated regarding their link with schizophrenia, in order to investigate whether the Dopamine Hypothesis could be validated, in terms of genetic association.

DRD2 (Dopamine Receptor D2) was among the first to be investigated, as a number of effective antipsychotic medications are antagonists of this receptor (Moriguchi et al,

2013). In terms of association studies, 3 SNPs were of particular interest to researchers, as they seemed to be linked with functional domains of the receptor, rs1799732, rs1801028 and rs1800497. Meta-analyses performed regarding the 3 SNPs (Glatt et al, 2004; Glatt et al, 2003; Yao et al, 2014; respectively) did not show any association between these loci and schizophrenia, indicating that there was no common variation link between *DRD2* and schizophrenia.

DRD3 (Dopamine Receptor D3), was also investigated in connection to possible association with schizophrenia, with association being originally reported in 1992 by Crocq and colleagues. However since, there have been mixed results in a number of studies, with no significant evidence in larger samples that have subsequently been examined (for example, Nunokawa et al, 2010) with regards to that gene.

Finally, *SLC6A3* (Dopamine Transporter Gene) was also investigated, as it would be a good candidate to be affected in schizophrenia, on the basis of the dopamine hypothesis. However, even early reports regarding its possible involvement (Li et al, 1994), failed to find any association. In the recent analysis of candidate genes, through the use of the PGC (Psychiatric Genomics Consortium) dataset (Farrell et al, 2015), consisting of 34,241 cases and 45,604 controls, none of the 3 aforementioned genes that were investigated due to their ties to the dopamine hypothesis showed any level of significance.

1.2.5.6 *DRD4*

Dopamine Receptor D4 (*DRD4*), despite being also a dopaminergic receptor was not implicated in the dopamine hypothesis. The initial association report on 115

schizophrenia cases and an equal number of matched controls did not show statistically significant results (Sommer et al, 1993). In spite the fact that the gene has not been as influential in schizophrenia genetics research as others described here, there have also been some positive results regarding its possible role in regulating the response to antipsychotic medication (Hwu et al, 1998). Additionally, polymorphisms in *DRD4* have been shown to be related to other psychiatric conditions, including Attention Deficit Hyperactivity Disorder (ADHD), Substance Abuse and Stress (Ptacek et al, 2011).

1.2.5.7 *HTR2A*

Serotonin Receptor 2A (*HTR2A*) was initially proposed as a candidate gene for schizophrenia on the basis of pharmacokinetics of a number of antipsychotics that were shown to be able to block *HTR2A* (Leysen et al, 1992). An initial report from Inayama and colleagues (1996) seemed to have potential, as a polymorphism was identified, that was significantly positively associated with the development of schizophrenia in an initial cohort of 62 schizophrenia cases and 92 controls. A number of later reports looking at the *HTR2A*, localised a possible variant at the *HTR2A*–1354C/T SNP, as a potential confounder of schizophrenia and, additionally as a potential link between schizophrenia and bipolar disorder. A recent meta-analysis by Gu et al (2013) pooled together all studies that looked at the association of the polymorphism in schizophrenia and bipolar disorder and demonstrated that in the pooled sample, as well as most individual samples there was no significant association of the SNP with either disorder.

1.2.5.8 *KCNN3*

KCNN3 (Small Conductance Calcium-Activated Potassium Channel 3) was initially considered as a candidate gene due to an initial report by Chandy et al (1998), that identified a number of CAG repeats in the *KCNN3* gene. As such repeats had been demonstrated at the time to be causatively linked to other hereditary neuronal disorders, such as Huntington's disease (MacDonald et al, 1996) and some types of ataxia (Giunti et al, 1994), it was thought that a similar genetic construct might explain the underlying heritability of schizophrenia. This was further corroborated by some earlier evidence indicating the presence of potential excess CAG repeats in patients with schizophrenia (Morris et al, 1995; O' Donovan et al, 1996). This initial discovery was followed by a number of association (Bonnet-Brilhault et al, 1999; Joober et al, 1999; Wittekindt et al, 1999; Laurent et al, 2003) and linkage studies (McInnis et al, 1999; Stober et al, 2000), but in both types of study, no evidence was found suggestive of a potential link between the disorder and the CAG repeats within *KCNN3*, despite the original hypothesis.

1.2.5.9 *TNF*

Initial interest in *TNF* (Tumor Necrosis Factor) was generated by the fact that its protein product is a pro-inflammatory cytokine, with neurotrophic or neurotoxic qualities that can potentially induce an inflammatory response in the nervous system (Loddick et al, 1999). This, coupled with the fact that there was evidence of inflammatory dysregulation in schizophrenia (Altamura et al, 1999), led to the preliminary investigations of *TNF* as a candidate gene. Indeed, in the first association study by Boin and colleagues (2001) it was reported that there were significant differences between cases and controls at the -G308A SNP, located within the *TNF*

gene, in their sample. Subsequent investigations of this polymorphism yielded mixed results. A meta-analysis was carried out by the group (Sacchetti et al, 2007) that published the initial report and included all studies that had investigated the polymorphism until then. The result of the meta-analysis was a weak association (p= 0.05) between the variant and schizophrenia. In a more recent report of Farrell et al (2015), *TNF* was one of the few candidate genes that did show genome-wide significance. This does not necessarily validate *TNF*, however, as the gene is located within the MHC complex region which is significantly associated with schizophrenia but also has a very LD (Linkage Disequilibrium) - rich structure. Thus, although elements of the region might be genome-wide associated with schizophrenia, TNF does not necessarily have to be one of them.

1.2.6 Association Analysis: Genome-Wide Association Studies

The main problem of association studies up to before 2007, was the fact that they had to be limited to a region or a subset of regions due to technological and financial limitations, thus using the candidate gene approach described in the previous section. This changed in 2007, when the first GWAS for schizophrenia was published (Lencz et al, 2007), and such a type of analysis was made possible and widely available, owing to advances in technology. In this section, a review of all previously conducted GWAS relevant to schizophrenia from 2007 and beyond will be presented both independently as well as on the basis of their contribution to the progress of the field. For reference, Table 1.2, below, presents all GWAS studies that have been conducted by 2016 in schizophrenia and Table 1.3, other GWAS that have implicated psychosis-spectrum disorders in some manner.

Table 1.2 Schizophrenia Genome-wide Association Studies published up to 2017

First Author	Year	Phenotype	Sample	Replication	Max Significance	Gene
			Schizophrenia Study			
Lencz, T	2007	Schizophrenia	178 Cases/ 144 Controls	(N)	$4x10^{-7}$	CSF2RA
Shifman, S	2008	Schizophrenia (sex – specific)	660 Cases/ 2271 Controls	(Y)	$3x10^{-5}$	REELN
Kirov, G	2008	Schizophrenia	574 Trios/ 605 Controls	(N)	10^{-6}	CCDC60
Sullivan, PF	2008	Schizophrenia	738 Cases/ 733 Controls	(N)	$1.7x10^{-6}$	-
O' Donovan, MC	2008	Schizophrenia	479 Cases/ 2937 Controls	(Y)	$1.8x10^{-6}$	ZNF804A
Need, AC	2009	Schizophrenia	871 Cases/ 863 Controls	(Y)	$1.3x10^{-6}$	ADAMTSL3
Purcell, SM	2009	Schizophrenia	3332 Cases/3587 Controls	(Y)	$4.7x10^{-8}$	NOTCH4 (MHC)
Shi, J	2009	Schizophrenia	2681 Cases/ 2663 Controls	(Y)	$4.6x10^{-7}$	CENTG2
Stefansson, H	2009	Schizophrenia	2663 Cases/ 13498 Controls	(Y)	10^{-12}	PRSS16 (MHC)
Athanasiu, L	2010	Schizophrenia	201 Cases/ 305 Controls	(Y)	10^{-5}	PCLO
Ott, J	2010	Schizophrenia	14 Cases/ 23 Controls	(N)	NA	NA
Ikeda, M	2010	Schizophrenia	575 Cases/ 564 Controls	(Y)	$6x10^{-6}$	SULT6B1
Yamada, K	2011	Schizophrenia	120 Trios	(Y)	$8x10^{-4}$	ELAVL2
Alkelai A	2011	Schizophrenia	155 Cases/ 176 Controls	(Y)	10^{-7}	DOCK4
Rietschel, M	2011	Schizophrenia	1169 Cases/ 3714 Controls	(Y)	$4.5x10^{-7}$	ARGHAP18(MHC)
Chen , J	2011	Schizophrenia	1658 Cases/ 1655 Controls	(Y)	10^{-3}	PTPN21
Alkelai A (2^{nd})	2011	Schizophrenia	189 Cases in 57 Families	(Y)	10^{-11}	LRRFIP1
Ripke, S	2011	Schizophrenia	9394 Cases/ 12462 Controls	(Y)	10^{-11}	MIR137
Yue, WH	2011	Schizophrenia	764 Cases/ 1599 Controls	(Y)	$4x10^{-6}$	ZKSCAN4 (MHC)
Shi, Y	2011	Schizophrenia	3570 Cases/ 6468 Controls	(Y)	$3x10^{-6}$	-
Liou, YJ	2012	Schizophrenia (treatment resistant)	522 Cases/ 806 Controls	(Y)	$2x10^{-7}$	SLAMF1
ISGC & WTCCC2	2012	Schizophrenia	1606 Cases/ 1794 Controls	(Y)	10^{-9}	MHC
Levinson, DF	2012	Schizophrenia (family based)	1218 Cases/ 990 Controls	(N)	NA	-
Betcheva, ET	2012	Schizophrenia	188 Cases/ 376 Controls	(Y)	10^{-7}	HHAT

Aberg, KA	2013	Schizophrenia (family based)	11185 Cases/ 10768 Controls	(Y)	10^{-7}	BRD1
Ripke, S	2013	Schizophrenia	8832 cases/12067 controls	(Y)	$2x10^{10}$	22 loci
Wong, EH	2013	Schizophrenia	481 Cases/ 2025 Controls	(Y)	$4x10^{-8}$	RENBP
Ripke, S	2014	Schizophrenia	36989 Cases/ 113075 Controls	(Y)	10^{-30}	108 loci
Li, J	2014	Schizophrenia (treatment resistant)	79 Cases (TR)/ 95 Cases	(Y)	$5x10^{-6}$	DDC
Goes, FS	2015	Schizophrenia	592 Cases/505 Controls	(Y)	$5x10^{-6}$	TBX1, GLN1, COMT
Kim LH	2016	Schizophrenia	350 Cases/ 700 Controls	(N)	$6x10^{-8}$	MECR
Yu, H	2016	Schizophrenia	4384 Cases/ 5770 Controls	(Y)	10^{-9}	6 Loci
Li Z	2017	Schizophrenia	7,699 cases/ 18,327 controls	(Y)	10^{-30}	109 Loci

Notes: [1] *First Author is the first author of the initial publication made using the sample described;* [2] *Only the initial sample of the GWAS is listed and not the replication samples (if any);*[3]*Includes the* p-v*alue of the most significant SNP in the study sample GWAS.* [4] *This column reports best linked gene on the basis of the results; for the largest GWAS sample, the number of genome-wide associated SNPs are reported.*

Table 1.3 Schizophrenia-Related Genome-wide Association Studies published up to 2016 (non-exhaustive)

First Author	Year	Phenotype
Volpi S	2008	Antipsychotics effect on QT after 14 days
Potkin SG	2008	Mean oxygen level of DPFC (Recall)
Alkelai A	2009	Response to Antipsychotic Treatment
Aberg K	2009	Extrapyramidal Side Effects (Antipsychotics)
Kim, S	2010	Brain Architecture of Psychiatric Disorders
Huang, J	2010	Cross-Disorder GWAS of Psychatric Disorders
Wang, KS	2010	Schizophrenia – Bipolar Disorder Comparison
Aberg K	2010	Antipsychotics effect on QT
Greenbaum, L	2010	Tardive Dyskinesia
Curtis, D	2010	Schizophrenia – Bipolar Disorder Comparison
McClay, JL	2010	Neurocognition and Treatment Response
Kendler, KS	2011	Alcohol Dependence and Schizophrenia
Ma, X	2011	Fluid Intelligence in Schizophrenia
Wang, KS	2011	Age of Onset in Schizophrenia
Bakken, TE	2011	Cortical Thickness in Schizophrenia
Le Blanc, M	2011	Neurocognition in Schizophrenia
Athanasiu, L	2012	BMI and Antipsychotic Treatment
Wang, KS	2012	Thought Disorder in Schizophrenia
Bergen, SE	2012	Schizophrenia – Bipolar Disorder Comparison
Fanous, AH	2012	Symptom Dimensions in Schizophrenia
Clark, SL	2012	Symptom Severity in Antipsychotic Treatment
Borgium, AD	2013	Interaction with maternal Cytomegalovirus
Xu, C	2013	Negative Symptoms in Schizophrenia
Smoller, JW	2013	Cross-Disorder GWAS of Psychatric Disorders
Ikeda, M	2013	Schizophrenia and methamphetamine psychosis
Hashimoto, R	2013	Cognitive Decline in Schizophrenia
Hass, J	2013	Hippocampal Volume in Schizophrenia
McGrath, LM	2013	Cross-Disorder GWAS of Psychatric Disorders
Wang, Q	2013	Grey Matter Volume in First Episode Psychosis
Sleiman, P	2013	Cross-Disorder Meta-analysis of Psychatric Disorders
Lencz, T	2013	Cross-Disorder GWAS of Psychatric Disorders
Ruderfer, DM	2013	Cross-Disorder GWAS of diagnosis and symptoms
Hashimoto, R	2014	Brain Volume in Schizophrenia
Avramopoulos, D	2015	Infection in Schizophenia and Bipolar Disorder
Hatzimanolis. A	2015	Neurocognition of Healthy Individuals
Chavarria-Siles, I	2016	White Matter Intergrity in Schizophrenia

QT: QT interval; DLPFC: Dorsolateral Prefrontal Cortex; BMI: Body Mass Index.

Notes: [(1)] *First Author is the first author of the study described.*

One of the big hurdles that GWAS faced was finding effects with power significant enough to be detected in smaller samples; this issue was amplified by the fact that because GWAS performed associations for hundreds of thousands or even millions of SNPs, their results needed to be corrected for multiple testing. This led to the significance level for genome-wide studies to be set at 10^{-8}, to account for testing for all SNPs that were available in each study. The first published GWAS for schizophrenia by Lencz and colleagues (2007) included a relatively small sample (300 people in total) and failed to detect genome-wide significant results. However, it was very important for the field for showcasing the feasibility of a schizophrenia GWAS and at the same time highlighting a problem that would become one of the hallmark characteristics of future GWAS: the search for increasingly bigger sample sizes, in order to detect the small effects of individual SNPs that needed to be exceedingly significant to be considered valid. The next schizophrenia GWAS by Shifman and colleagues (2008) also failed to find genome-wide significant results. However, it was able to detect an interesting interaction of a SNP in the *REELN* gene with gender that showed some promise. In the same year (2008), three more GWAS for schizophrenia were published. The first, by Kirov and colleagues, utilised a family-based design (trios design), where cases' family members are used as the control sample, in essence, transforming the initial family design into a case-control study. This study failed to find any SNPs with genome-wide significance, reaching a maximum level of significance of 10^{-6}. The second, by Sullivan and colleagues, also failed to find any genome-wide significant SNPs and reached a maximum level of significance similar to those in the Kirov study. The final GWAS published in 2008 was another important milestone for genome-wide association studies in schizophrenia. This study, by O' Donovan and colleagues, did not reach

genome-wide significance, but was the first to use an independent replication sample, a standard practice in schizophrenia GWAS nowadays. It was able to replicate its most significant findings which included a hit in *ZNF804A*. This gene has since been strongly implicated as a schizophrenia candidate gene with strong ties to cognitive constructs (Lencz et al, 2010; Nicodemus et al, 2014), and has also been suggested as a candidate gene in a large consortium study (Ripke et al, 2013). In 2009, another 4 schizophrenia GWAS were published, all of which had additionally some form of replication of their findings in other datasets. Two of these studies (Purcell et al, 2009; Stefansson et al, 2009) implicated a SNP found in, or near, genes within the MHC region of the genome, a region with a lot of genomic information and very complex architecture. Out of the four studies, none found a SNP with genome-wide significance and only after meta-analysing their original study with the replication sample, were the p-values significant enough to exceed the threshold of 10^{-8}. In 2010, there were another 3 GWAS studies (Athanasiu et al, 2010; Ott et al, 2010; Ikeda et al, 2010), that performed a GWAS in three distinct and very different populations; Athanasiu and colleagues studied a sample from Sweden and used a larger cohort as his "test" set; Ott and colleagues reported on an initial pilot cohort of a Sardenian isolated population, while Ikeda et al was the first publication in a Japanese population. Later on in 2011, 8 GWAS of schizophrenia were published, including the first GWAS of schizophrenia in Chinese populations (Yue et al, 2011; Shi et al, 2011), as well as two studies in Arab-Israeli populations (Alkelai et al, 2011a; Alkelai et al, 2011b). In 2012, 4 GWAS were published, with a genome-wide hit in the MHC region from the Irish Schizophrenia Genomics Consortium (ISGC & WTCCC2, 2012). This year, a GWAS for treatment-resistant schizophrenia was also published (Liou et al, 2012), reporting possible genetic differences among cases with a treatment-resistant phenotype, further investigated in Li

et al (2015). On the same year, an additional GWAS on Ashkenazi Jews was also published (Goes et al, 2015). Following these, two further studies were published on Chinese populations (Yu et al, 2016; Li et al, 2017) as well as a smaller GWAS on a Korean population that yielded surprisingly positive results (Kim et al, 2016).

During these years of GWAS implementation in the field of schizophrenia research, it became increasingly apparent that in order to detect significant genome-wide effects, larger sample sizes were needed. As this was also true for the investigation of other psychiatric disorders as well, the PGC (Psychiatric Genomics Consortium) was formed to enable researchers to pool their resources together and come up with larger sample sizes than any individual research group would be able to. This led to the inflation of populations in studies progressively, as the years went on. The first publication by the Consortium (Ripke et al, 2011), though having a relatively large study population, failed to implicate any gene with genome-wide significance. The same was true for the interim study of the Consortium (Ripke et al, 2013), that while reaffirming the *ZNF804A* hit from O'Donovan et al (2008), again failed to produce genome-wide significant results. Finally, a mega analysis by the PGC in 2014, was successful at pinpointing 108 susceptibility loci (Ripke et al, 2014), in the largest study of schizophrenia genetics to date with more than 35,000 cases and 100,000 controls.

At the same time as all the studies described above, there were additional investigations of schizophrenia genetics, performed not by GWAS of the disorder itself, but rather, in studies that either directly or indirectly implicated schizophrenia (Table 1.3). These investigations included studies that focused on whether different response to medication is genetically driven (Volpi et al, 2008; Aberg et al, 2010; Greenbaum et al,

2010; Athanasiu et al 2012), genetic differences of cognitive (McClay et al, 2010; Ma et al, 2011; Hashimoto et al. 2013), clinical (Wang et al, 2011; Fanous et al, 2012; Clarck et al, 2012; Xu et al, 2013) or brain (Potkin et al, 2008; Kim et al, 2010; Bakken et al, 2011; Hass et al, 2013, Wang et al, 2013; Hashimoto et al, 2014; Chavarna-Siles et al, 2016) endophenotypes and, finally, cross-disorder studies (Huang et al, 2010; Curtis et al, 2010; Bergen et al, 2012; Sleiman et al; 2013; McGrath et al, 2013; Lencz et al, 2013; Ruderfer et al, 2013). All these studies helped in developing an appreciation of the genetic background of schizophrenia and led increasingly to the realisation that the genetics of schizophrenia were not driven by variants with large effects; rather it is theorised that it is driven by the accumulation of small genetic effects that contribute toward the disorder in a liability-threshold model along with additional burden by possible epistatic effects, either through gene-gene or gene-environment interactions.

1.2.7 Epistatic Effects

In this section, current evidence and hypotheses for the possibility of epistatic interactions in schizophrenia will be briefly described; the polygenic risk score (PRS), which is an important aspect of the current understanding of schizophrenia genetics, will be described in more detail in the second part of the introduction.

Since the advent of modern age genomics, there have been several studies aiming to look at gene-gene interactions and how these might factor in psychosis. The initial approach to these interactions has been to examine epistasis on the basis of previously known gene interactions and thus aim only at a small subset of SNP-SNP terms. Within the boundaries of that initial pre-requisite, there has been a significant body of research that has found interactions that might influence schizophrenia or cognitive sub-domains

of the disorder such as *the MAPK – CNR1* interaction on brain volume abnormalities on marijuana dependent patients with schizophrenia (Onwuameze et al, 2013), the *BDNF - NTRK2* interaction in the heritability of schizophrenia with paranoid elements in a Chinese population (Lin et al, 2013), and three-way interactions between *NRG1, AKT1 and ERBB4* in schizophrenia heritability (Nicodemus et al, 2010a), among others. Additionally, there have been attempts to expand this line of research by trying to include multiple candidate genes within interactions; these include a study by Nicodemus et al (2010b) that investigated for epistasis among *DISC1* and 5 of the genes whose products interact directly with the *DISC1* protein, and another by Andreasen et al (2012) looking for epistasis among 14 previously implicated candidate genes.

One of the biggest caveats of searching for epistasis among genes without prior knowledge of a specific interaction is the dimensionality of the problem. 2-SNP interactions in a dataset of all SNPs are vast, even if there is a very moderate number of SNPs, making it simply not possible to reach statistical significance when testing for every possible 2-SNP combination. Machine Learning approaches such as the Random Forest algorithm, allow for an alternative and not as computationally or statistically inefficient method. An alternative model of searching for gene – gene interaction was put forward in Nicodemus et al (2014), where, after detecting a strong additive heritability element within a pathway of genes regulated by *ZNF804A*, all possible 2-SNP iterations were investigated in an initial "training" dataset. Then the top interaction terms were tested for significance in an additional cohort.

Finally, there has been a separate effort in investigating the interaction of genetic

factors with environmental influences in psychosis-spectrum disorders (van Os et al, 2008; van Os et al, 2009; EU-GEI, 2014). However, there has not been any concrete evidence about gene-environment interactions to date in that regard, although there is an increasing talk on how current methodologies might shape up new ways of hunting for these interactions (Geoffroy et al, 2013; EU-GEI, 2014; Vinkhuyzen & Wray, 2015).

1.2.8 Copy Number Variants (CNV)

Copy number variants can be described as large deletions and duplications of the genome with a varying number of copies among individuals in the populations These variations can be observed in the general population and are considered to be a source of individual genetic variation (Redon et al, 2009). In 2008, a number of studies examined the hypothesis that CNVs can contribute to the burden of disease of schizophrenia. ISC(2008) found that cases had an increase of 1.15 times to have a CNV, a result reaffirmed by two further studies (Stefansson et al, 2008, Walsh et al, 2008) that showed a significant increase of microdeletions and microduplications in individuals with schizophrenia. These results indicated that rare variants also contribute to the heritability of schizophrenia, even if only in a small number of cases (Sebat et al, 2009). These results were further corroborated (Kirov et al, 2009, 2011), but CNV studies' sample size remained small compared to larger GWAS. This was addressed when the largest CNV analysis to date (Marshal et al, 2017) was recently published (21,094 cases and 20,227 controls), with the results mirroring the initial investigations (CNV odds ratio for cases = 1.11) , while reaffirming the role of CNVs in the aetiology of schizophrenia.

1.3 The Polygenic Risk Score (PRS)

1.3.1 Initial Concept and implementation

The concept proposing that the genetic liability of schizophrenia lies in a polygenic complex inheritance pattern is not new; in fact, the foundations for a polygenic model of schizophrenia were first laid by a 1967 report by Gottesman and Shields. In that report, they proposed that schizophrenia should be biologically viewed as a liability-threshold characteristic, in which, a sufficient accumulation of genetic factors that individually were not necessarily deleterious, could lead to the manifestation of the disorder. In the late 00s, as GWAS attempted and failed to identify single common variants that were Genome-wide significant, while the effect size of these that were detected was quite low, a new construct was introduced by Purcell and colleagues in 2009; a polygenic score construct that would try to integrate all common SNP inheritance in a single metric. In that initial report, they proposed a polygenic risk score (PRS) comprised of all (almost) independent SNPs (74,062 SNPs) in their discovery sample, which was their male subpopulation (2176 cases and 1642 controls) that was subsequently applied in a target "test" sample of the female subpopulation (1146 cases and 1945 controls) of the sample. PRS was positively correlated with schizophrenia, with a R^2 value of 3%. The same construct with corrections for population stratification, was later applied in independent samples from other studies, serving as test sets, while the discovery sample was redefined as the entirety of the initial cohort. Again, the PRS from the discovery sample of the group was significantly associated with schizophrenia in these independent cohorts, with R^2 values increasing in parallel with the inclusion p-value threshold of the polygenic score. In addition, when the construct was applied in a bipolar disorder case-control sample, it was associated with

bipolar disorder caseness, albeit with a lower R^2 value, but not with caseness of non-psychiatric disorders, such as heart disease, diabetes and hypertension. This application demonstrated not only the relevance of the polygenic score as a construct to capture all the common SNP additive heritability from a given sample, but also its potential usefulness in examining genetic commonality between multiple disorders. This initial successful application of PRS in schizophrenia and the relative ease of use in applying it through the PLINK software (Purcell et al, 2007), gave subsequently rise to an increasing number of scientific reports that included PRS, both in the field of psychiatric genetics of schizophrenia and elsewhere.

A polygenic score is usually constructed from the following formula:

$$\text{PRS} = \sum_i (w_i * targetAlleles)$$

, where PRS is the polygenic risk score, *wi* are the weights assigned from the discovery set and are usually either regression betas in continuous traits or the natural logarithm of the odds ratio in binary traits. Afterwards, two regression models are calculated:

$$M1: Outcome = PRS + Cov$$

$$M2: Outcome = Cov$$

, where outcome signifies the disorder of interest and Cov other covariates that are included in the model beyond the polygenic score itself (for example the count of missing SNPs for each individual). Finally, the amount of phenotypic variance explainable by common variation is explained by subtracting the R^2 of M2 (reduced model) from that of M1 (full model).

Next, a number of studies using PRS will be presented, both in the context of schizophrenia and other psychiatric disorders.

1.3.2 PRS Applications in Schizophrenia, Psychiatric Genetics and other Complex Traits

As described above, PRS were initially applied in the schizophrenia sample described in the 2009 Purcel study. The method was immediately influential on reporting and investigating GWAS results in schizophrenia and other complex genetic disorders or traits. Below in Table 1.4 is a brief overview of selected publications since the first implementation of PRS that have used the method on a number of different traits with varying degrees of success. This is not meant to be an exhaustive list of publications that have applied a PRS as part of their methodology; rather it is meant as a demonstration of the diversity of research in which PRS have been applied.

Table 1.4 Major Polygenic Risk Score Studies since 2009

First Author	Year	Phenotype
		PRS Studies
Painter, JN	2011	Endometriosis
Davies, G	2011	Crystalised and Fluid Cognition (Intelligence)
Peterson, RE	2011	Body Mass Index
Ramdas, WD	2011	Open Angle Glaucoma Determinants
Hamshere, ML	2011	SCZ PRS differentiation between Type I and Type II BPD
Demirkan, A	2011	Circulating Lipid Levels
Demirkan, A	2011	Depression and Anxiety
Machiela, MJ	2011	Breast and Prostate Cancer Prediction
Middeldorp, CM	2011	Personality PRS to predict BPD and MDD
Simonson, MA	2011	Cardiovascular Disease Risk
Jung, JY	2011	Cross-disorder PRS of ASD and Autoimmune Disorders
Frank, J	2012	Alcohol Dependence
Taal, HR	2012	Blood Pressure
Varghese, JS	2012	Mammographic Breast Density PRS and Breast Cancer
Derks, EM	2012	Symptom Dimensions of SCZ
Pierce, BL	2012	Pancreatic Cancer
Sabuncu, MR	2012	Alzheimer PRS and Cortical Thickness in healthy controls
Otowa, T	2012	Panic Disorder
Fanous, AH	2012	PRS of SCZ disorganised symptoms
Ripke, S	2013	SCZ
Uher, R	2013	Antidepressant Efficacy in MDD
Van Scheltinga, AFT	2013	SCZ PRS prediction of Total Brain Volume
Smoller JW	2013	Cross-Disorder PRS among psychiatric disorders
McIntosh, AM	2013	SCZ PRS prediction of cognitive aging in healthy controls
Belksy, DW	2013	Asthma
Hamshere, ML	2013	ADHD PRS prediction of comorbid aggression
Whalley, HC	2013	MDD PRS association with White Matter Integrity
Heilmann, SS	2013	Androgenetic Alopecia
Vink, JM	2014	Overlap of smoking with drinking behaviour
Kirkpatrick, RM	2014	General Cognitive Ability
Ripke, S	2014	SCZ
Nicodemus, KK	2014	SCZ PRS prediction of working memory
Cui, J	2014	Antibody levels in Rheumatoid Arthritis
Chang, SC	2014	Depression long term phenotype
Solovieff, N	2014	PTSD
Hargreaves, A	2014	SCZ PRS prediction of Memory and Attention Dimensions
Mullins, N	2014	MDD PRS prediction of Suicidal Ideation
Ruderfer, DM	2014	BPD PRS prediction of manic symptoms in SCZ
Groen-Blokhuis MM	2014	ADHD PRS prediction of general population attention problems
Walton, E	2014	SCZ PRS association with prefrontal deficiency
Stringer, S	2014	SCZ PRS prediction of Immune Disorder Disease Status
Byrne, EM	2014	MDD PRS prediction of Post-Partum Depression
Fransen, E	2015	Age-Related Hearing Impairment
Yu, D	2015	Obsessive-Compulsive Disorder PRS
Martin, J	2015	ADHD PRS relationship with ADHD CNVs
Escott-Price, V	2015	Parkinson Disease PRS and Age of Onset
Aminoff, SR	2015	BPD PRS in BPD subgroups
Szulkin, R	2015	Prediction of Risk to Prostate Cancer
Ahn, K	2016	Childhood Onset SCZ
Hettige, NC	2016	PRS prediction of Antipsychotic Dosage in SCZ
Coleman, C	2016	Coeliac Disease
Jones, HJ	2016	SCZ PRS prediction of Anxiety in Adolescence
Lupton, MK	2016	Alzheimer PRS and Hippocampal Volume

PRS: Polygenic Risk Score; SCZ: Schizophrenia; BPD: Bipolar Disorder; MDD: Major Depressive Disorder; ASD: Autism-Spectrum Disorders; ADHD: Attention Deficit Hyperactivity Disorder; PTSD: Post-Traumatic Stress Disorder.

What is evident from Table 1.4, is that PRS have been extremely useful as tools in multiple fields. Their main uses have been to either investigate additive SNP heritability or as predictive constructs in possible genetic cross-disorder implications. A very good example of the latter use is in the investigation published by the Cross-Disorder Group of the Psychiatric Genomics consortium (2013) whereby the authors created polygenic scores for 5 different psychiatric conditions to investigate possible cross-diagnostic genetic factors.

1.3.3 Challenges, Criticism and Alternative Approaches

As this method has been integral to several high profile publications, it has also become the subject of scrutiny, especially in recent years (Dudridge, 2013). There have been, mainly, three major points of criticism as to the application of PRS.

The first point of criticism has highlighted the fact that as a method it undermines the sample size by forcing a split into discovery and test cohorts. Ripke et al (2014) proposed an alternative approach to that, which would add power to the score. The sample at their disposal was comprised of 49 studies of European Ancestry. In this sample, they performed a leave-one study out polygenic score analysis where the score coefficients were calculated from 48 of these and then applied to the odd study out. This process was repeated 49 times, essentially running 49 different PRS analyses and then meta-analysed to produce the final PRS estimates. By going through this process, they strengthened the pooled coefficient of significance and were able to observe heritability estimates independently on each of the samples.

The second point, of contention stems from the fact that as a method, polygenic scores require from the scientist applying it multiple parameters that may have an impact to the end result of the analysis. These include SNP selection on the basis of Linkage Disequilibrium and p-value thresholds of inclusion.

For the first, a few recent studies have been published, aiming to either perform pruning in a different way or to eschew the process all-together, in favour of applying a form of statistical weight to the polygenic score. On the subject of LD pruning, LD-based clumping has been proposed as an alternative process, which is also able to conciliate between datasets that are genotyped on slightly different platforms by combining SNP results that are in perfect LD (Shi et al, 2011). Additionally, there has recently been a novel method, LDpred, which has been proposed by Vilhjalmsson et al (2015) whereby all SNPs are used with weights that are calculated on the basis of LD information from exterior information. This method showed a moderate improvement in prediction in larger datasets, but was not able to outperform LD pruning under specific conditions. An additional methodological approach has been put forward by Mak et al (2016) objecting to the inferences made by LDpred. Instead, they are proposing a mode of weighing SNPs on the basis of either local true discovery rate based on Kernell density estimation. Again this method seems to be working on par with LD pruning in most cases, without either having necessarily an advantage over the other. The second parameter which has been a point of debate is the p-value threshold of inclusion of SNPs as there is the danger of adding too many SNPs in the polygenic scores, resulting in excess noise in the model. A recent development in that regard is PRSice, a tool developed by Euesden J et al (2015), which can calculate the optimal p-value threshold

for a polygenic score and generate graphics representing polygenic scores at various different thresholds. This tool could solve the matter of PRS p-value thresholds, however, as the PRSice is able to identify the most significant model among an arbitrarily large number of potential model, there exists the very real danger of over-fitting.

A third point of contention concerning polygenic scores is based on the fact that there are additional methods to calculate additive heritability beyond the PRS construct. A recent publication by Pan et al (2015) proposes such a method by introducing the aSPU (adaptive Sum of Power Score), an alternative construct that uses exponents of weights to the polygenic score and subsequently selects the most appropriate one.

An additional alternative approach has been proposed over the last few years over polygenic scores and has been demonstrated to be a reliable alternative way to account for additive heritability: LD Score regression (Bulik Sullivan et al, 2015) and, more specifically, Stratified LD Score Regression for Partitioned Heritability (Finucane et al, 2015). LD Score regression (Bulik Sullivan et al, 2015) was modelled on the premise that variants with elevated linkage disequilibrium with their neighbouring SNPs, the more likely it is to be tagging a causal variant and by regressing the χ^2 of each snp against their LD Score, the authors were able to account for bias due to inflation in each SNP. Stratified LD Score Regression for Partitioned Heritability (Finucane et al, 2015) built on that and demonstrated a way to apply LD Score regression to look for specific enrichments of polygenic contribution in functional and cell-specific elements using only summary GWAS statistics and LD Scores.

The upsurge of different methodologies, trying to investigate a common theme (the

correct application of PRS) from different angles, has caused some confusion in the field. Due to the absence of a singular consensus as to whether there is an optimal method to create a polygenic score, more scientific teams either default to the original method (Purcel et al, 2009) or implement the method that fits their data best, producing potentially spurious results.

1.4 GREML applied through GCTA

1.4.1 Overview

The origins of the GREML (Genetic-relatedness-matrix Restricted Maximum Likelihood) method that the GCTA (Genome-wide Complex Trait Analysis) tool uses can be traced back to the original research by Patterson and Thompson (1971), that introduced a method for estimating inter-block weights by maximizing the likelihood of a subgroup of points. GCTA (Yang et al, 2011) utilises a refined version of GREML (Gilmour et al, 1995) taking into account average information from a matrix, in conjunction with a Best Linear Unbiased Prediction (BLUP) (Henderson, 1975; Meuwissen et al, 2001) to fit a linear model either on the trait of interest or on the basis of a liability-threshold model, for a binary polygenic trait, as is the proposed case with schizophrenia (Gottesman, 1967). Through this process, the authors posit that all genetic contribution to phenotypic variance by common SNPs is quantified. The original group that developed the initial method, made a novel application of it the following year (Benjamin et al, 2011) by examining how the method might be applied on continuous social behaviour phenotypes and further expanded on its use (Lee et al, 2012a), using it as a tool of estimation of genetic correlation between disorders. Visscher et al (2014) also published a theoretical background for the method developed,

providing insight on means for optimally designing experimental protocols that would facilitate the use of this method. Finally in 2012, the research group collaborated with the Psychiatric Genomics Consortium and the International Schizophrenia Consortium to demonstrate how this method might be applicable in the large consortium schizophrenia databases to accurately predict schizophrenia common SNP heritability estimates. Since then, there has been an upsurge of published research reports that have used the GCTA tool in order to estimate the proportion of phenotypic variance captured by all SNPs. Recently, an extension of GREML, GREML-IBD(Identity-by-Descent) was released, that tries to reconcile common and rare variant research through the use of similarity matrices for IBD segments using whole-genome sequencing data and can potentially be of use in detecting rare previously unreported variants (Evans et al, 2017).

1.4.2 GREML applied through GCTA applications

As the application of the methodology is fairly straightforward and the software is openly available to everyone, a number of studies in every aspect of psychiatric genetics have attempted to incorporate this method in their research. In the field of schizophrenia genetics, it was initially applied in Lee et al (2012b), as it was already described above. Subsequently, it was incorporated as part of the main investigation in all subsequent GWAS reports by the Psychiatric Genomics Consortium as a measure of the level of variability explained by all the SNPs incorporated in the sample (Ripke et al, 2013; Ripke et al, 2014). Beyond psychosis, it has also been used in research relevant to psychiatric genetics, with the aim to quantify missing common SNP heritability in Parkinson's disease (Keller et al, 2012) as well as the proportion of differences in antidepressant response (Tansey et al, 2013). Davis et al (2013) also put

the method to use as a means of calculating the common heritability in Tourette Syndrome and Obsessive-Compulsive Disorder, and additionally found a degree of correlation between the two disorders. Trzaskowski et al (2013) by applying GREML through GCTA calculated the amount of variability conferred by common SNPs in anxiety traits manifested in childhood. In the same year, a report that used it to calculate common SNP heritability was published, thus quantifying the amount of variability in five addiction-related behaviours (McGue et al, 2013). Plomin et al (2013) also demonstrated that a huge amount of variability on cognitive abilities was attributable to common SNP variation. During that year, two more relevant reports were published calculating common SNP heritability for childhood callous-unemotional behaviour (Viding et al, 2013) and the likelihood of reporting life-events (Power et al, 2013). 2014 was another year where the popularity of the method continued to grow and was featured in studies including Borderline Personality traits (Lubke et al, 2014) and Social Communication traits in Autism-Spectrum Disorders (Pourcain et al, 2014). In 2015, reports that incorporated GREML through GCTA were also numerous including ADHD in a Norwegian population (Zayats et al, 2015), Alcohol Dependence (Mbarek et al, 2015) and Cannabis use age of onset (Genome of the Netherlands Consortium, 2015). Additionally, a study by Palmer and colleagues (2015) utilised GREML to examine co-heritability of common SNPs among multiple common addiction disorders. Finally, Davies and colleagues (2016) further examined possible co-heritability between cognitive ability and educational attainment.

In addition to the its numerous applications in the field of psychiatric genetics, GCTA has been linked in a significant amount of scientific output in other areas, aiming to estimate the common SNP heritability in paediatric obesity (Llewellyn et al, 2013),

self-reported subjective well-being (Rietvield et al, 2013), the cross-section family socio-economic status and intelligence (Trzaskowski et al, 2014; Marioni et al, 2014), the cross section of education level and health behaviours (Boardman et al, 2015), age-related macular degeneration (Hall et al, 2015), drug response in arthritis (Umicevic-Mirkov et al, 2015), epigenetic age acceleration (Levine et al, 2015), Multiple System Atrophy (Federoff et al, 2016) and a host of social-demographic outcomes (Domingue et al, 2016).

1.4.3 Challenges and Criticism

As it can be surmised from the above narrative, GREML applied through GCTA has become a mainstay in many fields and considered in many cases a de facto way of accounting for common SNP heritability. However, the method has recently come under scrutiny and a recent report by Kumar et al (2015) heavily criticised the method as a given for any problem of missing SNP heritability. Additionally, they pointed out that GCTA is very sensitive to even small changes in the initial matrix and highly susceptibility to population stratification, as well as highly sensitive to sample and SNP selection and measurement errors in the phenotype. The group that is responsible for the development of GCTA and the application of GREML, responded to that initial report, dismissing most of the critique that was expressed in the Kumar report, but stressing that GREML estimates of binary disease traits, and not quantitative traits, must be used with caution (Yang et al, 2011).

1.5 Thesis Rationale

1.5.1 Thesis Outline

In this thesis, three different projects were carried out in an effort to investigate the current methods used in the genetic epidemiology of schizophrenia and determine ways they could be improved. The first project focused on the use of pathway analysis of schizophrenia through polygenic scores, carrying out analyses similar to what has been previously described in Nicodemus et al (2015). The main difference in this study is that this method is applied on a binary outcome rather being extended to a larger database (PGC2); that has been previously used in the Nicodemus et al (2015) publication. In addition, an attempt was made to use multiple pathways, compare them and evaluate their biological validity in the context of schizophrenia. The second project focused on different methodologies that have been previously used to construct a polygenic score aiming to compare them, first on a simulated dataset derived from GS and subsequently in a more extensive simulation environment to examine under what conditions each method operates optimally. Finally, in the third study, the main aim was to investigate the GREML methodology as applied through GCTA in the context of binary characteristics. In the conclusion, the findings from these studies are evaluated in the context of recent developments in the field and steps forward to further current understanding of the genetic architecture of schizophrenia are proposed.

1.5.2 Aims of the Thesis

The overarching aim of this PhD Thesis was to investigate current methodologies that are being applied currently in schizophrenia genetics, and more specifically, polygenic scores and GREML. The study-specific aims were as follows:

I) To discern whether specific molecular pathways are major contributors to the polygenic score of schizophrenia.

II) To determine whether there is an optimally powerful method of constructing a polygenic score or if various different methods are more advantageous on the basis of the dataset/ the characteristics of the disorder.

III) To determine whether GREML applied through the GCTA software is a viable way to measure heritability in psychiatric disorders and the effect of population stratification in its function.

CHAPTER 2:

The Role of Polygenic Molecular Gene-Sets in the Schizophrenia Working Group Genome-Wide Study of the Psychiatric Genomics Consortium (PGC2)

2.1 Introduction

Genome-wide association studies (GWAS) have been used to investigate the genetic underpinnings of schizophrenia, highlighting putative biological pathways at play. These studies have identified multiple individual genes and can be used to locate classes of variants that show an excess in schizophrenia cases. A recent GWAS study (Ripke et al, 2014) has identified over a hundred unique common single nucleotide polymorphisms (SNPs) that occur at higher frequency in cases versus controls with consistent genome-wide significance levels. However, the increased disease risk associated with any one of these SNPs is very small.

Despite multiple loci being individually weakly associated with schizophrenia, the underlying genomic architecture, as defined by additive or interaction effects between variants in any individual, remains unclear (Mitchell, 2015). A number of methods have been applied in an attempt to capture cumulative common variation that might confer vulnerability to the disorder. Among the most prominent of these methods, polygenic risk scores (PRS; Purcell et al, 2009) have been shown to be capable of measuring most common additive variation and have also been able to pinpoint specific gene-sets that might be underlying cognitive traits (Nicodemus et al, 2014).

An additional measure of gene-set involvement in polygenic variance that has been consistently implemented is gene-set enrichment analysis (Subramanian et al, 2005). However, despite the prominence of the application of enrichment analysis in GWAS (O'Dushlaine et al, 2015; Pouget et al, 2016) and exome-sequencing studies (Curtis, 2016), these analyses in general lack the means of estimating the level of contribution

of these gene-sets to the amount of variance explained; instead their main function is to provide an indication of whether the gene-set of interest is more enriched, in terms of GWAS p-values than it would be expected by chance alone. This allows for a general estimation of a gene-set involvement, but is neither an able nor a sufficient measure to quantify such an involvement.

This study aimed to explore how the application of a gene-set focused polygenic score analysis might uncover gene-sets functionally driving the polygenic score and quantify their contribution. For this purpose, the focus was on eight gene-sets, six of which are centred on neuronal genes previously implicated in schizophrenia genetics or biology. It was hypothesised that at least some of these sets would be associated with schizophrenia case-control status. The remaining two were non-neuronal gene-sets associated with cancer and cardiac disease, were hypothesised as not being associated with schizophrenia case-control status. The rationale behind the choice of each included gene-set is presented below. For comparison, the behaviour of the polygenic risk score under H_0 was also examined.

2.1.1 Neuronal Gene-sets

The criteria for the choice of the six neuronal gene-sets included being experimentally derived, for example genes previously implicated in schizophrenia and being sufficiently large in size. In detail, gene-sets were selected from recent studies (Forrest et al, 2013; Steinberg et al, 2013; Hill et al, 2014; Sugathan et al 2014) that are based on a single gene with strong evidence for association with schizophrenia. SNPs in the gene *TCF4* (Transcription Factor 4) have been shown to be genome-wide significantly associated with risk for schizophrenia (Ripke et al, 2011; Ripke et al, 2014), and haploinsufficiency of this gene causes Pitt-Hopkins syndrome, with associated severe cognitive deficits (Amiel et al, 2007; Sweatt et al, 2013) as well as risk of psychosis (Stefansson et al, 2009). The *TCF4* gene-set was created on the basis of the differential expression of genes in neuroblastoma cells after the knockdown of *TCF4* from Forrest and colleagues (2013). A total of 1052 autosomal genes (5652 SNPs) demonstrating differential expression were included in the gene-set.

FMR1 (Fragile X Metal Retardation 1) is a gene coding for FMRP (Fragile X Mental Retardation Protein), whose loss of function results in the Fragile X syndrome (Verheij et al, 1993), a very serious developmental disorder, often co-morbid with autism spectrum disorders (ASD). Additionally, *FMR1* mutations have been shown to be linked with cognitive impairment and earlier age of onset in schizophrenia (Kovacs et al, 2013). The FMRP gene-set was created on the basis of functional gene-sets based on developmental expression of genes contingent on FMRP expression (Steinberg et al, 2013). For the purpose of this analysis, all four gene-subsets described were combined into one all-encompassing gene-set which was used for this analysis, containing 680 autosomal genes (5833 SNPs).

miR-137 is a microRNA with high levels of expression in the brain and in neural stem cells (Guella et al, 2014). Transcriptional targets of *miR-137*, such as *ZNF804A* and *CACNA1C*, as well as the gene itself, have been implicated with schizophrenia, thus increasing the appeal of studying regulation gene-sets stemming from changes in expression of *miR-137* (Kim et al, 2012; Kwon et al, 2011, Ripke et al, 2011). The third and fourth molecular gene-sets were chosen on the basis of the findings from Hill et al (2014), where two gene-sets were generated on the basis of upregulated (817 genes, 7796 SNPs) and down-regulated (761 genes, 8533 SNPs) genes after overexpression of *miR-137* in neural progenitor cells in vitro.

CHD8 (Chromodomain Helicase DNA Binding Protein 8) codes for a DNA helicase that suppresses gene expression by affecting chromatin restructure. It has been found to be a significant contributor in autism susceptibility (Wilkinson et al, 2015) and a key component of the CHARGE syndrome (a congenital deaf-blindness syndrome), through its interaction with *CHD7* (Batsukh et al, 2010). It has also recently been shown that through a rare variant it may be contributing to schizophrenia risk (Kimura et al, 2016).

The two final neural gene-sets were generated from the findings of Sugathan et al (2014), where *CHD8* reduction in neural progenitor cells led to the creation of two gene-sets, one of upregulated (1140 genes and 8807 SNPs) and one of down-regulated (616 genes, 4986 SNPs) genes. For the latter two gene-sets the decision to split them into down-regulated and upregulated gene-sets was undertaken on the basis of the findings from their respective experimental reports (Sugathan et al, 2014; Hill et al, 2014) that described a more pronounced response under one of the conditions.

2.1.2 Non-Neuronal Gene-sets

As null "control" gene-sets, two gene-sets that were related to coronary artery disease and cancer were selected. The list of genes for these was curated in the coronary artery disease database (http://www.bioguo.org/CADgene/) and the Atlas of Genetics and Cytogenetics in Oncology and Haematology (atlasgeneticsoncology.org). Those gene-sets had in them 534 and 459 genes, respectively (with 8078 and 7316 SNPs). The rationale for using these non-neuronal gene-sets was mainly to serve as null "control" gene-sets of roughly equal size to their neuronal counterparts.

2.1.3 Aims

The main aim of the study was to identify and test polygenic scores based on the biologically-validated neuronal gene-sets that were expected to show genetic links with schizophrenia; an additional aim was to compare them with polygenic scores using non-neuronal gene-sets that were hypothesised as not showing significant association with schizophrenia. A major question was whether polygenic scores based on these biologically-based gene sets would be able to account for more variation explained than randomly-selected SNPs. Statistical and bio-informatic approaches were also used to examine the behaviour of these gene-sets in the PGC data.

2.2 Methods

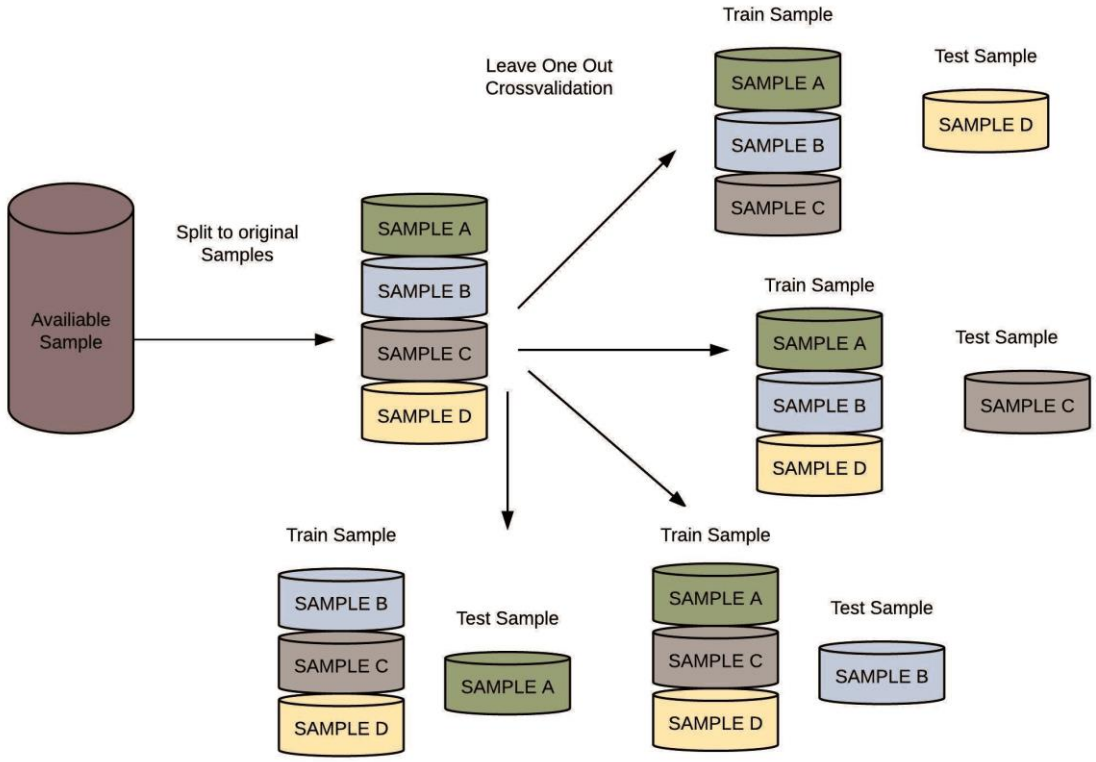
2.2.1 The Schizophrenia Working Group of the Psychiatric Genomics Consortium 2 Case-Control GWAS

Sample composition and selection is described in detail in Ripke et al (2014). In brief, cases were mainly selected based on a diagnosis of either schizophrenia or schizoaffective disorder, as the two disorders tend to aggregate together in family studies (Kendler et al, 1993) and there is a low inter-rater reliability across the two groups on the basis of their initial diagnosis (Faraone et al, 1996). The quality of diagnosis for cases was assessed through a questionnaire that examined quality control and structured diagnosis procedures for each study (Ripke et al, 2014). Studies with different case ascertainment procedures were also considered and included in the final sample (Hamshere et al, 2012). For 2 of the studies that were included in the sample, cases were included on the basis of clozapine uptake and a prior diagnosis of treatment-resistant schizophrenia (Ripke et al, 2013).

In total, 39 different studies were included in the final sample, which constitutes the largest sub-sample of the PGC2 that was available for secondary data analysis. The sample, in which this research was conducted, was composed of 29,125 cases and 34,836 controls of European ancestry. In that sample, there were 36,318 males, 22,061 females and 5,582 participants with no sex information. In the interest of having the largest possible sample possible, individuals with missing sex information were kept in the analysis. Details of subject composition for each individual study and how these were collected can be found in Ripke et al. (2014). Details of the individual studies can

be found in the appendix (Appendix 2.1). Genotypes were imputed using the 1000 Genomes project dataset (August 2012, 30,069,288 variants, release "v3.macGT1") as a reference for the imputation process, through the use of the IMPUTE2/SHAPEIT software (Howie et al, 2011). In terms of quality control, the following were considered as essential: SNP missingness < 0.05 (before sample removal); subject missingness < 0.02; autosomal heterozygosity deviation (|Fhet| < 0.2); SNP missingness < 0.02 (after sample removal); difference in SNP missingness between cases and controls < 0.02; and SNP Hardy-Weinberg equilibrium (p-value > 10^{-6} in controls or p-value > 10^{-10} in cases). The quality control was performed before the data were handed to the researcher.

Figure 2.1 Leave-One Out Cross-Validation Process

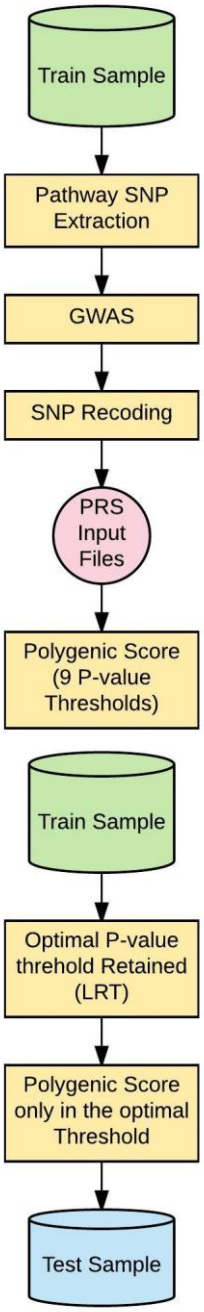


Example of leave one out cross-validation process for a sample containing 4 datasets. The same process was followed with the 39 PGC datasets.

2.2.2 Leave-One-Out (LOO) Polygenic Risk Score Analysis

For the implementation of the main analysis, two datasets were created for each of the 39 studies; one with every dataset but the one held out, serving as the training set and one with only the participants from the study of interest which would be used as the independent testing set, similar to the study design for polygenic risk scores in the original PGC2 manuscript (Ripke et al, 2014). Figures 2.1 and 2.2 illustrate (a) the process of leave-one out cross-validation and (b) the flowchart that was followed in each iteration of the cross-validation process. For each study, first a GWAS was performed in the training set to calculate the p-value and *natural logarithm* (odds ratio) of each individual SNP relative to case-control status in all datasets except the one that was left out as the test data set for the polygenic score. Subsequently, in order to make sure that SNPs in the training set were coding the same reference allele as the risk allele in the test study; all SNPs were coded as risk by selecting in every instance the allele which had an odds ratio larger than 1. Afterwards, polygenic scores were created for nine different p-value cut-off thresholds (0.0001, 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50). These were generated for the training set in order to reduce the need of correction for multiple testing on the held-out test set. A logistic regression model was fitted for each of these 9 polygenic scores in each of the 39 training sets that included 38 studies, including covariates (count of missing genotypes, principal components and study indicators, as in the original PGC2 study (Ripke et al, 2014)). In each study, the largest test statistic from the nine polygenic scores in the training set was used to select the single polygenic score to be tested on each of the 39 held-out test sets. Subsequently, a single polygenic score was created on each of the held-out test sets, on the same p-value cut-off that was able to produce the most significant results in the respective training set. (See also appendix 2.2)

Figure 2.2 Flowchart for Polygenic Score Generation in Each Leave-One-Out Iteration



Flowchart of the process followed in each iteration of the leave-one-out cross-validation. PRS input files are the polygenic scoring file and the individual SNP p-v*alue file. PRS: Polygenic Risk Score; LRT: Likelihood Ratio Test*

2.2.3 Statistical Analysis

All analyses were performed in PLINK 1.90 (Chang et al, 2015) for polygenic score generation and genetic data manipulation and in R 3.2.4 (R Core Team, 2016) for the generation of regression models and additional analysis. Additionally, the R package fmsb (Nakazawa, 2015) was used to calculate Nagelkerke's R^2. MetaP (*Dongliang* G, Duke Institute For Genome Sciences & Policy, NC, USA) was used to calculate Stouffer's z p-value meta-analysis estimates (Stouffer et al, 1949). For the gene-sets described above, first, all available SNPs within 20Kb of genes within the gene-sets were identified and extracted from the overall dataset. Afterwards, the SNP set was linkage disequilibrium (LD) pruned in PLINK. LD pruning uses a sliding window process, where LD between SNPs is examined and, for every pair of SNPs that are in LD above some user-defined threshold within that window, one is removed. A sliding window of 50 SNPs was used, a sliding step of 5 SNPs and an r^2 threshold of inclusion at 0.25. For the regression analysis, the same principal components as the ones originally used in Ripke et al (2014) were utilised, to control for population stratification, also adding the study indicators as covariates. Finally, likelihood ratio tests between nested regression models in R were used and calculated the Nagelkerke R^2 as well as the p-value for the polygenic score in each of the 39 held-out test datasets.

2.2.4 Meta-Analysis

To estimate the significance of the results in the overall sample, a meta-analysis of the 39 results from the test sets only was performed, collected from all the studies through the use of Stouffer's z p-value in metaP, also accounting for directionality of effect and sample size. Because each training set would have different *natural logarithm*(odds ratios) and p-values, each polygenic score based on the training sets were different; for

this reason, a standard meta-analysis is not strictly appropriate, thus p-values were combined across held-out test sets. For the R^2 values the median, interquartile range and range were chosen from the held-out test sets as the metrics that would be able to accurately depict the performance of the gene-sets in the overall dataset.

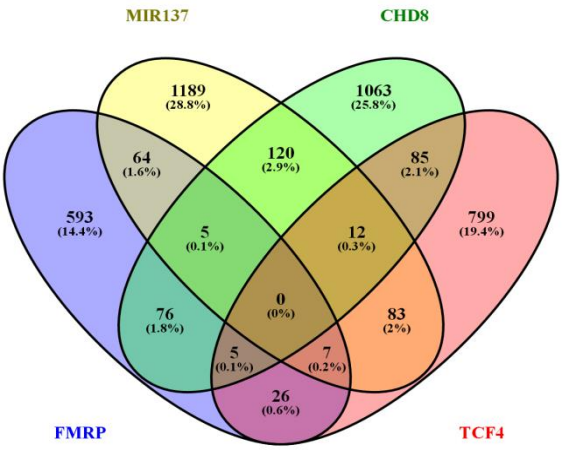
2.2.5 Simulation and Validation Studies

Two additional studies were performed to examine the performance of the methodology used and the influence of genic versus non-genic SNPs. The use of genic SNPs has been demonstrated before to produce slightly inflated results (Schork, 2013). The first analysis was a standard experiment-wise permutation test that was conducted on the *TCF4* gene-set, consisting of permuting the phenotype 100 times and rerunning the entire experimental pipeline, leaving a single study out at a time, on these randomly generated phenotypes. If the pipeline is robust to type I error, only 5 percent of these permuted experiment-wise results should show a significant result with the gene-set at $\alpha = 0.05$, as would be expected by chance. For the second analysis, 10 random subsets of genic SNPs were selected, defined as SNPs found either a) within genes, b) 5 Kb upstream of genes or c) 1 Kb downstream of genes, of a mean size of 5, 000 SNPs and an equal number of non-genic SNP subsets, defined as SNPs not included in the genic subset. The pipeline was run with the main outcome variable, using all methods as previously outlined. The purpose of performing this analysis was to establish if there is a systematic enrichment of genic SNPs sets showing significant results, regardless of the gene or gene-set in which they were embedded.

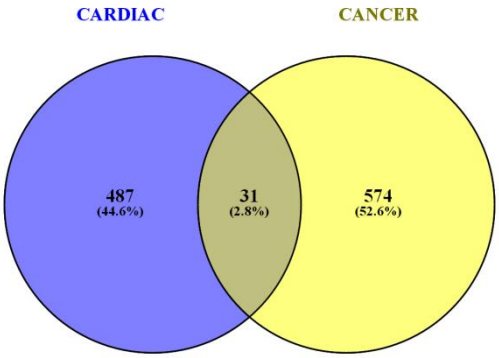
2.3 Results

Figure 2.3 Overlap of gene-sets.

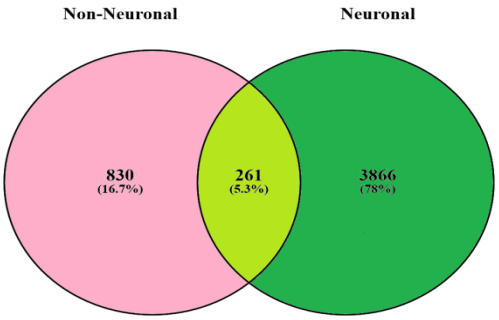
A



B



C



A: neuronal gene-sets; B: non-neuronal gene-sets; C: Combination of the two. Percentages in the graph indicate the percentage of the total genes found in each overlapping segment. miR-137 and CHD8 indicate all the genes for both the down- and the up-regulated gene-sets as there was no overlap between the two.

2.3.1 Gene-Set Characteristics

Initially, the analysis investigated if there was any significant overlap among the neuronal and non-neuronal gene-sets. Little overlap was found among the gene-sets, indicating that potentially significant results would not be driven by similar sets of SNPs and would therefore be independent of each other (Figure 2.3). In the neuronal gene-sets, there was an overlap of 15 percent between FMRP and the three other gene-sets combined, an overlap of 19.7 percent between miR137 and the other gene-sets, an overlap of 22.1 percent between *CHD8* and the other gene-sets and an overlap of 20.8 percent between *TCF4* and the other gene-sets. Non-neuronal gene-sets showed minimal overlap between them, with only 31 genes being shared among all of them (less than 10%). Finally neuronal and non-neuronal sets had an overlap of a total of 261 genes (5.3% of genes used in the study). In addition, it was investigated whether any of the gene-sets that were selected were enriched for any specific Gene Ontology Term (Gene-Ontology Consortium, 2015) (Appendix 2.3), which would indicate a possible implication of specificity for a cellular process. From the six neuronal gene-sets that were investigated only the FMRP gene-set showed an enrichment for neuronal processes and more prominently for nervous system development enrichment (p-value = 10^{-60}) and generation of neurons (p-value = 10^{-43}). For the remaining five gene-sets that were used, the broad terms “biological process” and “cellular process” were those that came as the top GO terms for them, indicating that these gene-sets were implicated in multiple biological processes.

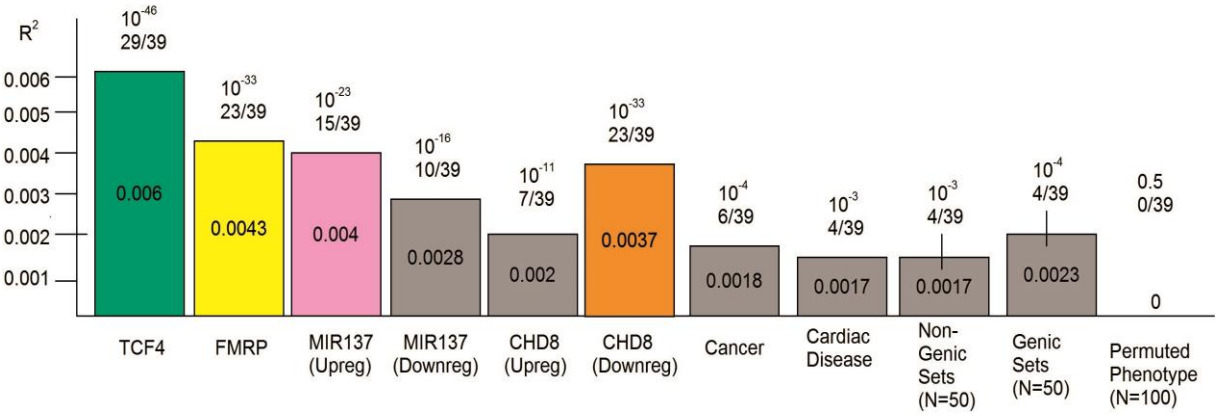
2.3.2 Polygenic Risk Score Analysis

Table 2.1 Nested R^2 results for all individual studies for each gene-set.

Study details			Gene-sets							
Study ID	Case (N)	Control (N)	***TCF4***	**FMRP**	***MIR137* (up)**	***MIR137* (down)**	***CHD8* (up)**	***CHD8* (down)**	**Cancer**	**Heart disease**
clm2	3426	4085	**0.00400**	**0.00132**	**0.00266**	0.00034	0.00046	0.00000	**0.00470**	0.00130
mgs2	2638	2482	**0.00840**	**0.00741**	**0.00534**	**0.00596**	**0.00469**	**0.00415**	**0.00401**	**0.00515**
clo3	2105	1975	**0.01693**	**0.01226**	0.01090	0.01455	0.01039	**0.04791**	0.01627	0.00143
s234	1980	2274	**0.00584**	**0.00232**	**0.00207**	0.00084	**0.00258**	**0.00113**	0.00053	**0.00179**
swe5	1764	2581	**0.00636**	**0.00761**	**0.00618**	**0.00397**	**0.00498**	**0.00379**	**0.00333**	**0.00325**
irwt	1291	1006	**0.00980**	**0.01104**	**0.00764**	0.00005	0.00312	**0.01148**	**0.01386**	**0.00609**
gras	1067	1169	**0.00820**	**0.00433**	**0.00765**	**0.00230**	0.00407	**0.00965**	0.00044	0.00246
swe6	975	1145	**0.00970**	**0.00366**	**0.02518**	0.00414	0.00202	**0.00305**	0.00175	0.00221
ajsz	894	1594	**0.00434**	**0.00592**	0.00172	**0.00850**	0.00249	**0.00154**	0.00227	0.00001
aber	719	697	**0.00623**	**0.00832**	0.00314	0.00093	0.00171	**0.01189**	0.00026	0.00070
ucla	700	607	**0.00835**	**0.00573**	0.00251	0.00266	0.00005	0.00174	0.00007	0.00327
uktr	649	649	0.00911	0.00047	0.03779	0.03675	0.00208	**0.00358**	0.00690	0.00031
pewb	574	1812	**0.00603**	**0.00211**	**0.00317**	0.00158	0.00100	**0.00565**	**0.00686**	0.00108
cou3	530	678	**0.00725**	0.00817	**0.01952**	0.00010	0.00551	**0.00599**	0.00131	0.00455
lemu	516	516	0.00011	0.00023	0.00017	0.00935	0.00013	0.00001	0.00182	0.00425
uclo	509	485	**0.00528**	**0.01287**	0.00618	**0.01092**	0.00459	0.00105	0.00000	0.00158
lie5	497	389	**0.00912**	0.00144	0.00028	0.00018	0.00125	**0.00495**	0.00101	0.00638
denm	471	456	0.00068	**0.01327**	0.00062	0.00025	0.00001	0.00010	0.00220	0.00312
asrb	456	287	**0.00402**	0.00187	0.00745	0.00503	0.00061	0.00010	0.00183	0.00249
munc	421	312	**0.00558**	**0.01080**	0.00006	**0.00737**	0.00234	0.00158	0.00397	0.00134
cati	397	203	**0.01392**	**0.01451**	**0.02506**	0.00855	0.00108	**0.02120**	0.00005	0.00137
caws	396	284	0.00268	**0.00539**	0.00722	0.00231	0.00011	**0.00994**	0.00153	0.00613
top8	377	403	**0.00772**	**0.01392**	**0.00601**	0.00016	0.00024	**0.00394**	0.00003	0.00363
edin	367	284	**0.00528**	**0.04107**	0.00202	**0.01422**	0.00085	**0.01435**	0.00378	0.00280
port	346	215	0.00016	0.00135	0.00185	0.00487	0.00000	**0.00586**	0.00038	0.00021
umeb	341	577	**0.00684**	**0.00864**	0.00409	**0.01769**	**0.00659**	**0.00752**	0.00195	0.00122
msaf	325	139	0.00026	0.00064	0.00169	0.00080	0.00009	**0.00335**	0.00158	0.00009
ersw	265	319	**0.00635**	**0.00846**	0.00327	0.00351	0.00085	0.00003	0.00015	0.00675
dubl	264	839	**0.00921**	0.00025	**0.00434**	0.00224	0.00141	**0.01166**	0.02123	**0.00845**
egcu	234	1152	**0.00291**	0.00213	0.00041	0.00047	**0.00449**	0.00037	**0.00520**	0.00027
swe1	215	210	0.00133	0.00027	0.00237	0.00326	0.00221	**0.03613**	0.00243	0.01048
buls	195	608	**0.00579**	**0.00950**	**0.01389**	0.00028	**0.00979**	0.00280	0.00821	0.00119
umes	193	704	**0.01759**	0.00096	0.00041	0.00078	0.00084	0.00155	0.00026	0.00156
zhh1	190	190	0.00023	0.00111	**0.01705**	0.00005	0.00063	0.00041	0.00031	0.00195
lacw	157	245	**0.02095**	**0.02732**	0.00884	**0.01177**	**0.02024**	**0.01505**	0.00536	0.00354
pews	150	236	0.00076	0.00004	0.00038	**0.01161**	0.00008	0.00192	0.00126	0.00095
lie2	133	269	**0.00948**	0.00167	**0.01504**	0.00286	0.01548	0.00407	0.00096	0.00015
butr	70	70	**0.00577**	0.00397	0.00210	0.00075	0.00495	0.00218	0.00251	0.00005
cims	67	65	0.00005	0.00006	0.00602	0.00830	0.02008	0.00455	0.00226	0.00164

Table of results in each individual study; the first column indicates the PGC2 label used for each study. The table is sorted by the number of cases. Highlighted boxes had a level of significance $p<0.05$. Details for each study and their respective size can also be found in Appendix 2.1 and the original PGC2 (Ripke et al, 2014) study.

Figure 2.4 R^2 and p-values from meta-analysis of all gene-sets.



Numbers on top of the bars denote the meta-analysed Stouffer's z p-value *for the gene-set and the number of polygenic scores that were significant in independent, held-out test studies. For the genic and non-genic sets, the statistics represent the median of 50 sets; the line above the box represents the range of these sets for the 50 iterations of each. The final box is the median results for 100 permuted phenotype iterations*

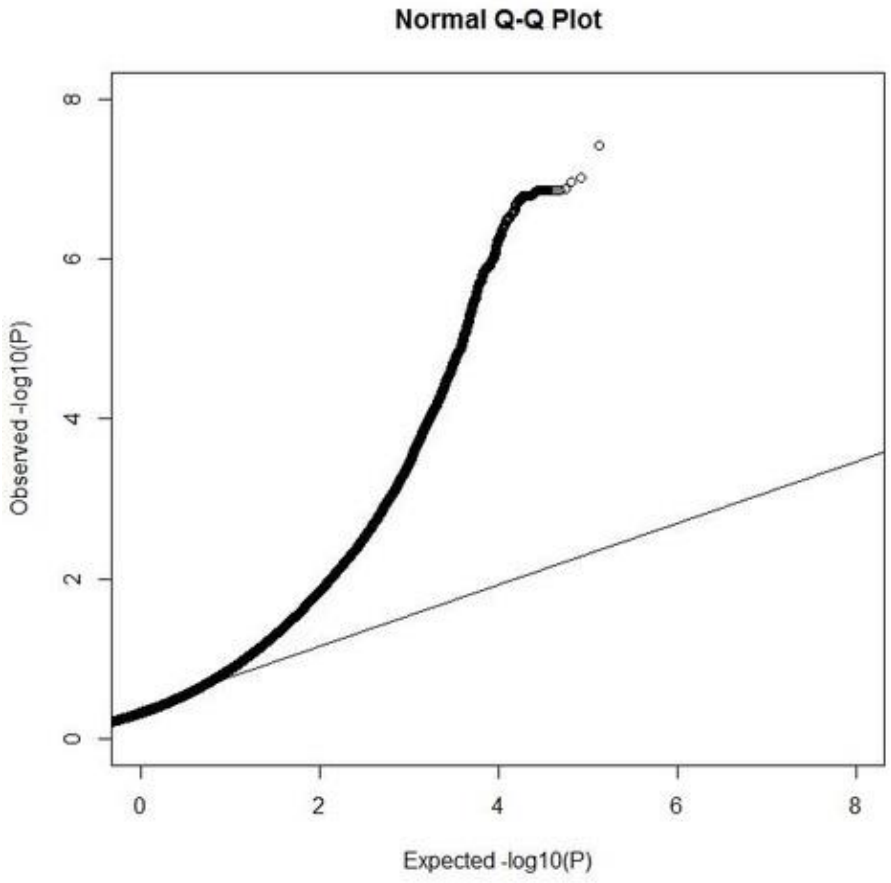
Results of the polygenic score analyses are presented in Table 2.1 and Figure 2.4. *TCF4* gene-set-weighted polygenic scores reached the highest level of significance in the meta-analysis of studies (Stouffer's z p-value = 10^{-46}). This particular gene-set was also the one where most of the individual studies, as independent test sets in the LOO, were significant (29/39), showing evidence for association at p-value < 0.05 (uncorrected, as only one polygenic score was tested in each of the held-out test sets; Table 2.1). This gene-set was able to explain the highest percentage of variability among the studies described (Nagelkerke R^2 = 0.6%; Figure 2.4). As in the original PGC2 study (Ripke et al, 2014), *TCF4* itself was found to be genome-wide-significantly associated with schizophrenia and thus may be driving the results. To test this, 12 SNPs within *TCF4* were removed and the analysis was repeated, with results in the same level of magnitude in terms of significance (Stouffer's z p-value = 10^{-40}) and effect size

(Nagelkerke R^2 = 0.6%). FMRP gene-set-weighted polygenic scores were also highly significant (Stouffer's z p-value = 10^{-33}) with 23/39 individual independent test set results showing evidence for association; it explained 0.43 percent of the variability of schizophrenia case-control status. For the two *miR-137*-regulated gene-sets, there were also consistent levels of significance, albeit lower than for either *TCF4* or FMRP, with Stouffer's z p-value = 10^{-23} for the gene-set upregulated in the overexpression of *miR-137* and Stouffer's z p-value = 10^{-11} for the gene-set down-regulated in the overexpression of *miR-137*. These gene-sets explained 0.4 percent and 0.28 percent of the variability of schizophrenia case status. For the *CHD8* gene-set polygenic scores, the gene-set of down-regulated genes in the absence of *CHD8* was highly significant (Stouffer's z p-value = 10^{-33}) and explained 0.37 percent of the variability. The polygenic scores created from genes that were upregulated in the absence of *CHD8* were also significant (Stouffer's z p-value = 10^{-11}), but only a small number of individual held-out test sets were individually significant (7/39) and the overall effect explained 0.2 percent of the variability.

Interestingly, gene-sets that were created from the non-neuronal sources were weakly, but still statistically significantly, associated with the outcome (Stouffer's z p-value = $10^{-4,}$ and 10^{-3}, respectively). Six and four out of the 39 polygenic scores were significant at p-value < 0.05, uncorrected, in those analyses, respectively. To address this issue and to examine the distribution of p-values across all of the gene-sets investigated, p-value bins corresponding to deciles under the null hypothesis were created, where p-values are distributed ~$U(0,1)$. For all the gene-sets investigated, there seemed to be a similar distribution of SNPs among individual SNP p-value bins. Additionally, there seemed to be an increased proportion of SNPs in the top ten percent bin for all of the gene-sets,

consistent with the quantile-quantile (Q-Q) plot from the PGC2 mega-analysis (Ripke et al, 2014) and the Q-Q plot demonstrating a strong deviation from the expected (Figure 2.5).

Figure 2.5 Q-Q Plot of -$\log_{10}$ P-value in the PGC2 Sample of 39 studies



Expected versus Observed P-values in the selected sample which included 39 studies from the PGC2 original sample.

2.3.3 Simulation and Validation Studies

In the simulation study, for the 100 runs with permuted phenotypes, the type I error rate at $\alpha = 0.05$ was as expected under the null hypothesis of no association, with 4 of 100 of them having a Stouffer's *z* p-value value of less than 0.05 (i.e., type I error of 4%). To further investigate these results, it was examined whether the small effect that was detected in the non-neuronal gene-sets might be related to either some small number of genes in those gene-sets that were linked to schizophrenia, or if the effect was due to the inclusion of genic SNPs versus non-genic SNPs. To that end, 50 random subsets of 5000 SNPs from genic and non-genic SNPs were generated, respectively, and implemented the same experimental protocol of 39 leave-one-out analyses and combining p-values as with the gene-set analysis described before. On average, all sets of SNPs that were tested had a level of significance ranging from 10^{-2} to 10^{-7} with no individual set exceeding significance of that observed among neuronal sets in this study. Genic sets were consistently but only slightly more significant than non-genic sets (median Stouffer's *z* p-value = 10^{-4} versus 10^{-3}). The R^2 values were also higher in the genic set with a median value of 0.0021 against a 0.0016 value for the non-genic set. Finally, a full polygenic score of the sample was run to have a measure of comparison of the full gene-set versus the gene-specific gene-sets that were used. A total of 3,848,785 SNPs were used to generate the polygenic score which yielded a median nested R^2 value of 0.24.

2.4 Discussion

In this study, polygenic risk scores were used to investigate whether a number of neuronal gene-sets of interest play a significant role in the common genetic architecture of schizophrenia. There was significant heterogeneity among the gene-sets that were used, with a number of them, including the *TCF4* gene-set, the FMRP gene-set, the gene-set upregulated in the presence of excess *MIR-137* and the gene-set down-regulated in the absence of *CHD8* shown to be associated with schizophrenia. In contrast, the apparently significant effects that were observed in the non-neuronal gene-sets (cancer and coronary artery disease), as well as the gene-set down-regulated in the presence of excess *miR-137* and the gene-set upregulated in the absence of *CHD8*, were not actually higher than a floor effect observed with random sets of genic SNPs.

2.4.1 Gene-Set Analysis

The *TCF4* gene-set was the most significant among those investigated, with a Stouffer's *z* p-value of 10^{-46}. The nested R^2 effect observed was three times that of any set of random SNPs of the same size. The result retained its significance and magnitude of effect size even after removing SNPs within the *TCF4* gene itself, indicating that the observed relationship exists between genes of the gene-set and the phenotype above and beyond the effect that *TCF4* by itself might also exert. There is consistent evidence for the role of *TCF4* itself in the common polygenic background of schizophrenia (Ripke et al, 2011; Ripke et al, 2014) and due to the nature of SNPs implicated (non-coding genetic elements) to the pathway of genes influenced by *TCF4* expression (Harrison, 2015).

The FMRP gene-set was also very strongly associated with schizophrenia. This is an intriguing finding, as FMRP has primarily been implicated in autism spectrum disorders. There are commonalities among both the clinical features and genomics of major psychiatric disorders and a recent cross-disorder Mega-GWAS (Smoller et al, 2013) that indicated that common variation pre-disposing to mental illnesses might be shared to some degree among major psychiatric disorders. Additional evidence of the involvement of FMRP targets to schizophrenia can be observed from rare variant studies that have consistently implicated FMRP pathways with schizophrenia (Purcell et al, 2014; Fromer et al, 2014; Richards et al, 2016).

In the two *miR-137* gene-sets that were investigated, there was a strong and positive effect only on the gene-set that was up-regulated after *miR-137* over-expression. The down-regulated gene-set, although reaching levels of nominal significance, did not show an effect stronger than what would be expected by chance on a similar set of genic SNPs. This result is consistent with findings of other studies of miR137 expression, that seem to indicate that up-regulation of the gene seems to be linked with pathways suspected of being implicated with psychosis (such as the Major Histocompatibility Complex) (Collins et al, 2014).

Finally, from the *CHD8* gene-sets created on the basis of knockdown, the down-regulated gene-set was the one that showed evidence of significance. *CHD8* has not previously been centrally implicated in psychosis as it is associated with a congenital disorder (CHARGE syndrome) and linked to autism (Wilkinson et al, 2015). However, there is a reasonable argument to be made on the basis of common susceptibility to mental health disorders that genes that are central to other major mental health

disorders might also affect, on a lesser scale, schizophrenia. There has been recent evidence on rare variants on the gene itself (Kenny et al, 2014; Kimura et al, 2016) being implicated in psychosis, which adds to the pre-existing notion of the cross-disorder nature of *CHD8* and pathways associated with it. The down-regulated gene-set that was strongly associated with schizophrenia in the present study was also the one that Sugathan and colleagues (2014), from where the initial genetic gene-set was taken, reported to be significantly enriched in autism-related genes.

2.4.2 Floor Effect

In addition to the investigation of specific gene-sets with schizophrenia, a systematic floor effect in polygenic scores was observed. This observation is consistent with predictions that would be made based on the recently proposed omnigenic model of complex traits (Boyle et al, 2017). This model states that most genes expressed in cells that are relevant to the biology of an illness contribute to heritability and PRS because of the likely interaction of multiple signaling pathways within cells that support their biological functions. In light of this hypothesis, implicating a greater number of SNPs than the ordinary polygenic model would suggest, the results support the hypothesis by demonstrating a weak polygenic effect extant in every random subset of genes. This omnigenic effect is also supported by Figure 2.5, which demonstrates a marked overall increase in SNP test statistics versus expected values, as well as the Q-Q plot in the original PGC2 report(Ripke et al, 2014) that also showed a very similar effect across an increased number of observations. Finally, the enrichment analysis conducted for the gene-sets investigated indicated an enrichment for broadly expressed genes, which also corroborates the principle finding of the omnigenic model(Boyle et al, 2017) for schizophrenia.

Genic SNP sets seemed to explain slightly more variation than their non-genic counterparts. This indicates that studies implementing a pathway stratagem should be mindful of both these effects when trying to assess whether their gene-set explains more variation than a random subset of genic SNPs, with significant differences being observed on the basis of SNP localization (Schork et al, 2013).

This study showed that several of the target putative core gene-sets investigated were highly significantly associated with schizophrenia, with the strongest effect being observed for the *TCF4* core gene-set. Even though most of the genes in these sets are not associated with risk in current GWAS datasets, they may be pheripheral genes, part of the network suggested by the omnigenic model. These findings strongly indicate that, despite a very widespread, possibly even omnigenic contribution to risk, it is possible to identify subsets of genes making relatively larger contributions - putative core genes - which may implicate specific biochemical pathways or molecular processes with selectively greater roles in pathogenesis. Our analyses were based on a somewhat arbitrary selection of target gene-sets, in that they relied on prior discoveries and appropriate experimental datasets. The findings of this study do not exhaustively reveal the underlying molecular architecture of schizophrenia risk. More generally, though, the method developed here allows a quantification of the contribution of specified core gene-sets as well as potentially identifying pheripheral genes, and should be practically applicable in the selection of sets of SNPs that yield the greatest signal to noise in the construction of a PRS.

2.4.3 Limitations

There are some aspects of the study presented in this chapter which limits its applicability. First of all, as has been indicated in the introductory chapter as well as Chapter 3, there are a number of different ways to approach the generation of a polygenic score (for example Shi et al, 2011 or Mak et al, 2016). As the focus of this study was to apply specific gene-set generated polygenic risk scores in the PGC Schizophrenia data-set and the amount of analysis required for each was quite resource intensive, I did not include further methods of polygenic score generation in the analysis, beyond the original methodology (Purcell, 2009). Despite that, the results presented, should be quite precise with regards to the pathways examined, given the high level of confidence that the Stouffer's z p-value indicated. However the exact level of variance explained by those pathways may fluctuate from method to method used, so it should be thought of more as a clear indication of the pathways working above and beyond the omnigenic effect.

An additional limitation to this study is that although we did examine a number of different biologically validated pathways, there are still other similar pathways that also merit investigation in a similar manner but due to time and computational power limitations. These could include pathways from prominent candidate genes described in the introductory chapter such as *DISC1* (Blackwood et al, 2001). or *BDNF* (Green et al, 2011) or they could also include gene-sets derived from genes implicated in other psychiatric disorders, in order to investigate their cross-disorder properties.

2.5 Conclusions

The main aim of this study was to create polygenic scores from a number of different gene-sets that have been previously implicated experimentally in schizophrenia and subsequently test them in the context of the PGC schizophrenia data-set. Subsequently those gene sets were tested against each other, against random subsets of genic and non-genic SNPs and against two sets taken from unrelated disorders (cardiac disease and cancer). The results from this study indicate that:

1) A number of the gene-sets investigated were significantly associated with schizophrenia. The strongest effect was linked to the *TCF4* gene-set, while significant effects were observed for the FMRP, MIR137 upregulated and *CHD8* downregulated gene-sets.
2) A floor effect of R^2 values was discovered in the PGC2 cohort as any set of random SNPs would give a low but significant estimate of variance explained.
3) A difference between genic and non-genic SNPs, in terms of both p-value and R^2 was observed, with genic SNPs explaining more variance in any given subset. These latter two points should be taken into account when investigating cohorts with Q-Q plots that deviate from expected.

CHAPTER 3:

Comparison of Current Methods of Polygenic Score Generation

3.1 Introduction

3.1.1 Background

Genome-Wide Association studies, despite identifying a number of common variants that contribute genome-wide to the increase of risk, have thus far not been able to account for the majority of common additive variance. One of the reasons for that is that many common SNPs are very weakly associated with the phenotype and thus unable to reach a nominal genome-wide significance that would single them out. Indeed, expanded GWAS studies conducted by consortia that increased sample sizes have been able to expand upon initial findings and detect multiple significant loci, where almost none were present before (Ripke et al, 2014). However, even these genome-wide studies have a large proportion of common variation missing.

Polygenic risk scores (PRS) were introduced as constructs that would be able to account for variability explainable by additive common variation that, on its own, would be of too small of an effect to sufficiently detect with current sample sizes and statistical methodologies. The score is expressed itself as linear additive sum for each individual with:

$$\text{PRS} = \sum_i (w_i * targetAlleles),$$

where w_i is a weight calculated on the basis of the natural logarithm of the odds ratio or the regression coefficient of each SNP in an independent discovery sample (Purcell et al, 2009).

Since then, the polygenic score construct has been effectively implemented in a range of different traits and conditions ranging from asthma (Belsky et al, 2013) to body mass index (Peterson et al, 2011), and from cognition (Kirkpatrick et al, 2014) to psychiatric conditions such as bipolar disorder (Aminoff et al, 2015) and schizophrenia (Ripke et al, 2014). As the use of polygenic scores has increased over the last decade, there have been several proposed improvements suggested on how to optimally build a polygenic score and to best account for the maximal amount of variance explainable by additive common SNP variation.

The first suggested improvement on the basis that the initial process used to account for SNP linkage disequilibrium, Linkage Disequilibrium (LD) pruning (Purcell et al, 2009), did not take an informed approach to how SNP pruning was performed. This method, using a sliding window technique would look at the first SNP within its window, then, find which SNPs were in LD with it and discard them, before moving to the next SNP window. This process did not take into account the functionality of the SNPs and could potentially lead to loss of useful genetic information. Thus, an alternative process of LD clumping on the basis of the most prominent SNPs within a region was proposed (Shi et al, 2011). This process clumped together SNPs on the basis of a previous genome-wide association study (GWAS), taking into account the p-v*alues* of all SNPs within the clump and selecting the strongest signals, in an effort to rationalise selection. This process of clumping has also been demonstrated to be very useful when combining studies on different genotyping platforms on a meta-analysis, as to combine the statistical power of all SNPs found within a clump (Shi et al, 2011). Figure 3.1 further demonstrates how SNP selection would occur for each of the two methods in a sample chromosomal window.

Figure 3.1 Example of LD pruning and Clumping

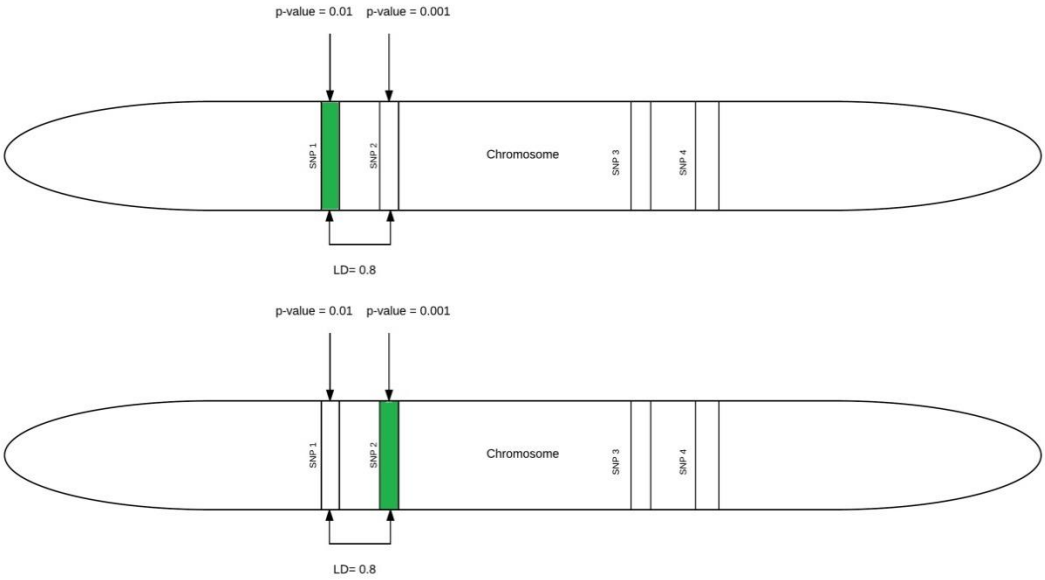


Figure 3.1: SNP selection on the basis of a) LD pruning or b) LD clumping.When pruning is used, the first SNP in LD with others for each window is used (here SNP 1) , whereas in clumping is SNP with the higher p-v*alue among those in LD are selected (here SNP 2).*

Finally, Mak et al (2016) proposed a different method of SNP weighing that would retain more information by including all of the SNPs and weighting them. The basis of these weights was the true discovery rates which were obtained by either (a) maximum likelihood of the z-values distribution or (b) kernel density estimation of the z-values distribution. These methods have an advantage over typical weighting methods such as those put forward by Vilhamson et al (2015) of being non-parametric and therefore independent from the nature of the data. Detailed scripts of how these weights were calculated can be found in the appendix of the original report and in Appendix 3.1.

These methods, including different parameter settings of the original LD pruning methodology, have never been tested against each other on a number of different simulated sample scenarios before. This comparison between these methodologies will provide researchers with a guideline as to under what conditions polygenic risk scores might optimally work, how effective the measurement would be compared to the real underlying variation and what method to use under each condition.

3.1.2 Aims

The aim of this study was to examine how different polygenic score methods would behave under a number of different conditions in a simulation study using real genotyping data. More specifically, a comparison of methods was implemented in terms of (i) sample size, (ii) LD structure and (iii) underlying polygenic architecture. This methodological approach was undertaken aiming to observe how each method would operate on a given pre-decided set of conditions as well as observe how different thresholds of pruning and thresholding would affect those results. A second aim was to investigate how these methods would operate in a larger sample that is similar to those from current biobanking efforts.

3.2 Methods

3.2.1 Initial Sample

The initial sample comprised of 7,372 unrelated individuals from the Generation Scotland (GS) cohort (Smith et al, 2013). Generation Scotland is a family based study that collected sociodemographic and genetic data from about 24,000 volunteers across Scotland between 2006 and 2011. DNA from 20,000 of those individuals has been analysed by high density genome-wide genotyping (Illumina OmniExpress SNP GWAS (700k)). Quality Control (QC) analyses have been performed. To select only unrelated individuals, GCTA (Yang et al, 2011) was used to remove individuals with a similarity of more than 0.025. The quality control was performed before the data were handed to the researcher.

In order to study the effect of differing population size on the polygenic score, three random subsamples of varying size were selected (500, 1000 and 2500 individuals). These sizes were selected as to represent a small, a medium and a large individual GWA study. Furthermore, as it has been previously reported that different LD structures may affect GWAS estimates and consequently PRS results (Laurie et al, 2010), three chromosomes with different density in genes and therefore LD (Smith et al, 2005) were selected; chromosome 13, which has been characterised as gene-poor (327 protein-coding genes in 114 Mb), chromosome 19, which has been previously described as gene-rich (1472 protein-coding genes in 58 Mb), and chromosome 15 which would serve as middle point between the two (613 genes in 102 Mb) (Farrel et al, 2014).

3.2.2 Simulation of Phenotypes

For each of the nine sample/chromosome combinations described, three distinct phenotypes were generated. To generate the phenotypes, independent SNPs for each chromosome were randomly selected in each iteration of the simulations. Initially the three chromosomes were LD pruned at a pair-wise LD value of 0.01 (1 percent) leaving only SNPs independent of each other in each chromosome. For each iteration, 3 subsets of 20, 100 and 200 SNPs were randomly selected. As these would be the source of variation between simulations, there was no thresholding of the SNPs on the basis of their Minor Allele Frequency (MAF). Subsequently, phenotypes were simulated using the LDAK software (Speed and Balding, 2014). LDAK simulates polygenic phenotypes by computing genetic contributions and subsequently drawing effect sizes from a Gaussian distribution. Afterwards, LDAK adds noise to produce phenotypes with the predefined amount of variance explained by the sum of the additive variance. In order for the results between the three phenotypes to be comparable, the total variation explained by the sum of SNPs in each phenotype was set to be 0.5. Thus, SNPs individually would be explaining 0.025, 0.005, and 0.0025 of the total variance. Figure 3.2 demonstrates the process of selecting the datasets and creating the simulated phenotypes for each dataset. The process was repeated 500 times with random SNPs selected each time to ensure variation between iterations.

Figure 3.2 Process of Sample selection and simulation of phenotypes.

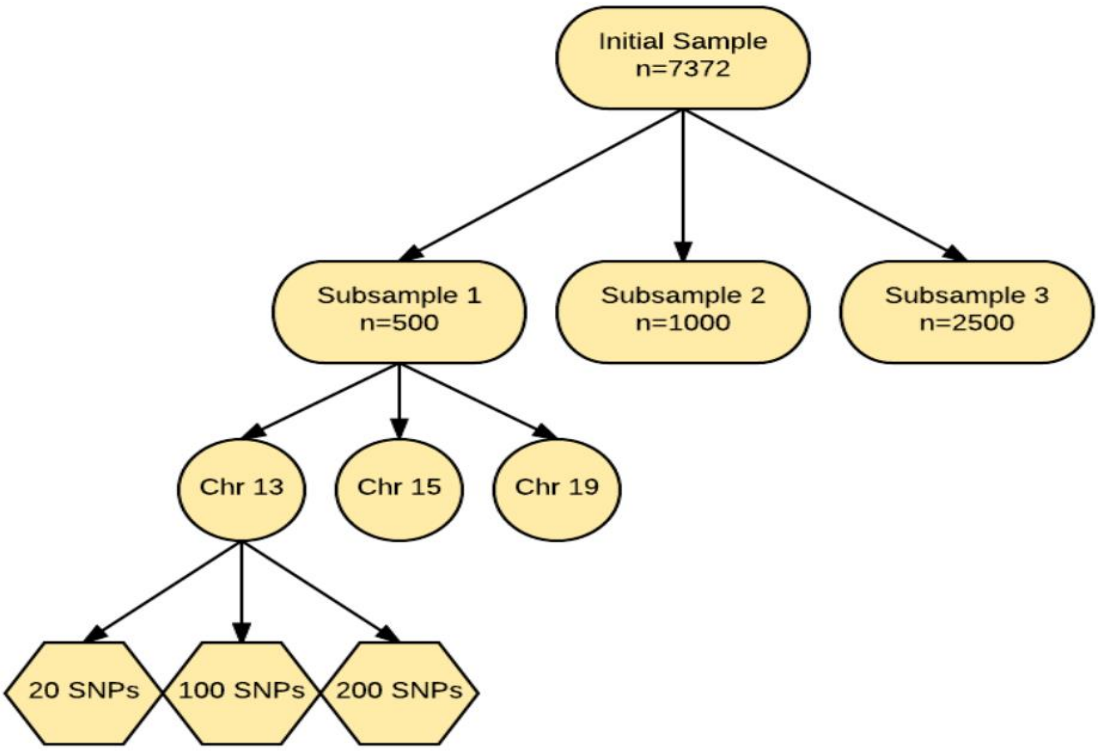


Figure 3.2: Tree diagram of the process followed; the sample was subsampled for 500, 1000 and 2500 individuals. For each subsample, 3 chromosomes were selected as separate datasets: chromosome 13, 15 and 19. Finally, in each of these chromosomes 3 phenotypes were created, each one containing 20, 100 and 200 SNPs. This yielded a total of 27 phenotypes for each iteration of the process which were tested on the three methods. The process was repeated 500 times.

3.2.3 Polygenic Score Creation

Polygenic scores for all methods described below were generated through the use of Plink 1.9 (Chang et al, 2014). For the purposes of calculating the R^2 values, regression models were fitted with the simulated phenotype as an outcome and the polygenic score and the count of missing phenotypes as covariates. To calculate nested R^2 estimates, the R^2 of a reduced model that did not include the polygenic score was subtracted from the R^2 of the full model such as:

$$\Delta R^2 = {R^2}_{Full} - {R^2}_{Reduced}$$

3.2.4 Linkage Disequilibrium Pruning and Thresholding

The first method that was applied on the simulated datasets was linkage disequilibrium (LD) pruning, implemented in Plink 1.9 (Chang et al, 2014). To do this, equal sized discovery and target datasets were created. The discovery samples were LD pruned at 3 different LD r^2 thresholds (0.1, 0.25 and 0.5). The sliding window size and step were kept constant at 50 SNPs and 5 SNPs, respectively. Afterwards, polygenic scores were generated on the basis of 11 different p-value thresholds (0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1) on the target sets.

3.2.5 Linkage Disequilibrium Clumping

The second method applied on the simulated datasets was Linkage Disequilibrium (LD) clumping that operates by creating clumps of SNPs and then selecting the ones with the higher degree of significance in each clump on the basis of prior GWA statistic.

Clumps were created in three r^2 thresholds in the discovery set (0.1, 0.25 and 0.5) to be consistent with the LD pruning thresholds. Subsequent polygenic scores generated were applied to the target samples.

3.2.6 True Discovery Rate Weights

The third method that was applied on the simulated datasets was local True Discovery Rate (TDR) weights, as originally proposed by Mak et al (2016), where their application was able to approximate the results of the best possible p-value threshold, which would be very advantageous in studies when this would not be easily estimated. Both methods that have been put forward by Mak et al (2016) were applied to the datasets: weighing by the maximum likelihood of z-values distribution and weighing by kernel density estimation of the z-value distribution. The R scripts originally provided by the authors to generate these weights were used for this analysis and can be found in Appendix 3.1.

3.2.7 Extended Simulation Application Sample

To create an extended sample, the most informative chromosome on the basis of the results (which was chromosome 19) and a number of SNPs approximating those observed in a real polygenic situation (N=200) were selected. Using the shapeit (Delaneau et al, 2012) software, phased haplotypes were created from the available data of 7372 individuals who were then randomly combined to generate a cohort of 100,000 individuals. From this new sample, a random subsample of 40,000 individuals was selected as the discovery GWAS and a separate subsample of an equal number was selected as the target polygenic score application sample. The phenotype was once again created using LDAK (Speed and Balding, 2014). Causal SNPs were selected on

the basis of their Minor Allele Frequency (MAF) ranging between 0.4 and 0.5 and were kept constant between simulations, while phenotypic variance between simulations was ensured by adding a different distribution of the noise vector to each iteration of the simulations. In total, 500 such iterations were created. The reasoning behind creating this larger sample was to create a situation similar to recent bio-banking studies with sample sizes exceeding 20,000 individuals and to investigate whether application in such a sample would influence the results.

3.3 Results

First, to check that this phenotype simulation was working correctly, the polygenic scores for all 27 different scenarios that were created were calculated, using only the SNPs assigned as causative in the model each time. The results were within the range of 0.48 to 0.51, indicating a high validity of the simulated models, given the true value of 0.50.

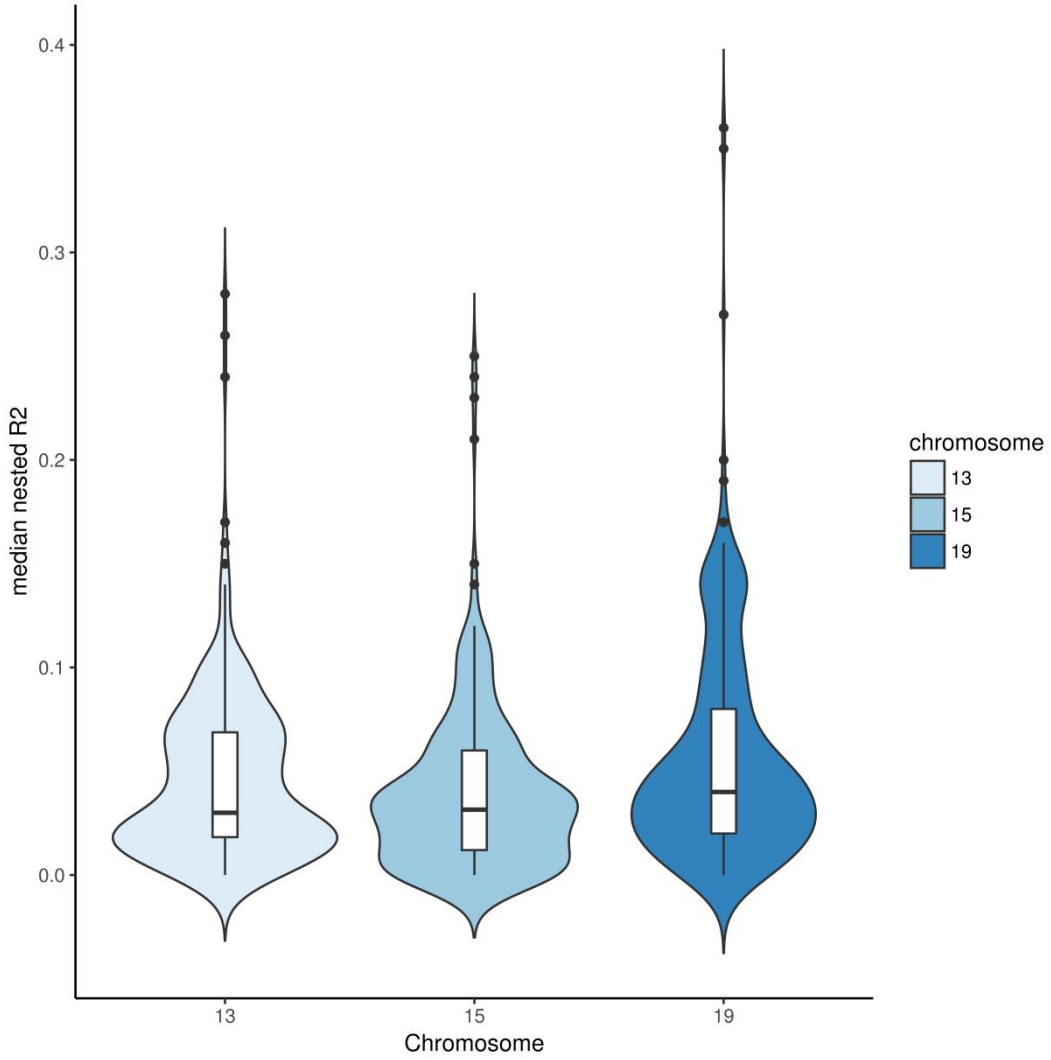
Initially, the results on the basis of the 3 conditions that were set out are presented, irrespective of method. Thus, the results on the basis of (1) LD structure /chromosome selection, then (2) on the basis of sample sizes and finally (3) on the basis of phenotype are presented. Afterwards the least predictive and most predictive scenarios are presented for discussion. In the next section, a comparison across the methods that were tested is presented, before finally the methods are compared again in the larger sample. Detailed tables of the median of all simulated conditions are available in Appendices 3.2 and 3.3.

3.3.1 Effect of different LD structure in PRS estimates

Figure 3.3 demonstrates how different LD structure, determined by the three chromosomes with different genic content that were used (Chromosome 13, Chromosome 15, Chromosome 19) affected Polygenic Risk Score estimates. Greater genic content and, as a result increasing Linkage Disequilibrium, resulted in an increase in R^2. When all methods were taken in consideration, the median R^2 value for chromosome 13 was 0.033, 0.035 for chromosome 15 and 0.0415 for chromosome 19. These differences indicate that a more conserved structure with higher amounts of LD

may lead to better detection of true effects, regardless of method, as when SNP dimensionality is reduced, a portion of the effect will still remain due to linkage disequilibrium effects.

Figure 3.3 Violin Plot of Median PRS Nested R^2 by chromosome.

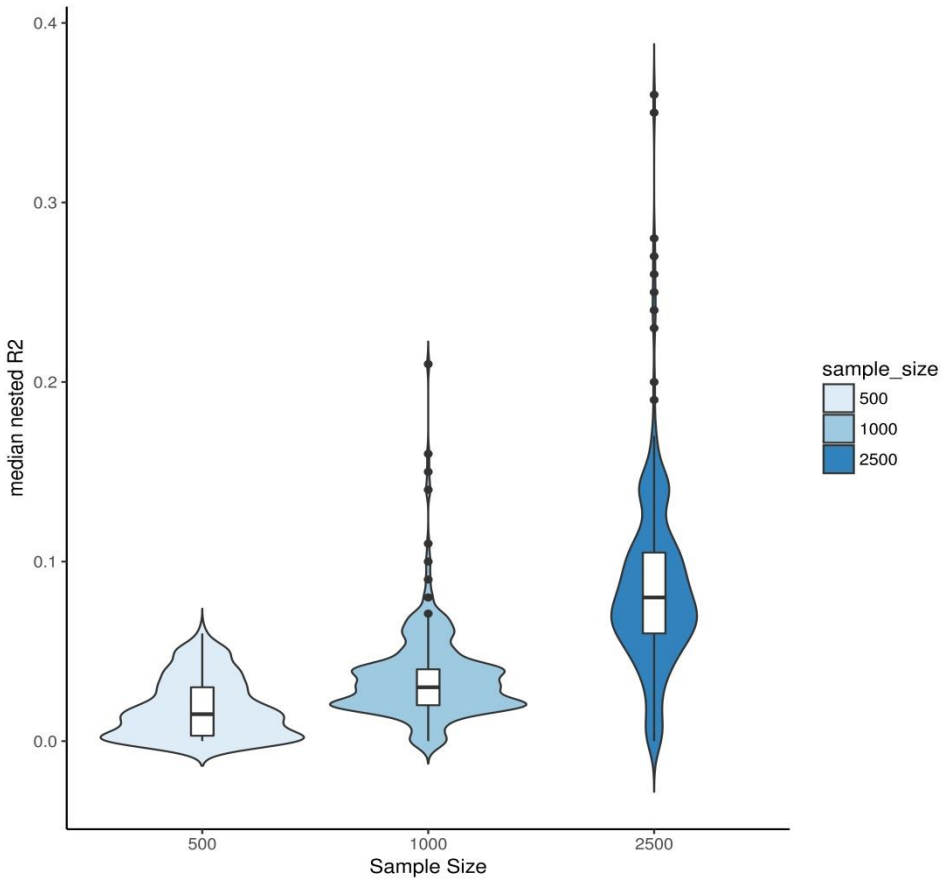


Comparison of median R^2 by chromosome using all methods and irrespective of sample size and target phenotype.

3.3.2 Effect of Increasing Sample Size

There was a very sharp increase of PRS estimates on the basis of increasing sample size. Figure 3.4 demonstrates how the median variation explained by the polygenic effect was almost tripled (from 1.3 to 3 percent) when the sample size was increased from 500 individuals to 1000. It was doubled again to 7 percent when the sample was further increased to 2500 individuals, showing a steady linear increase of variation explained which parallels the sample size increase. When examining individual methods, the same trend is true univocally for all methods that were investigated, demonstrating an effect of sample size to estimates that is independent of the method used.

Figure 3.4 Violin plot of Median PRS Nested R^2 by sample size.



Comparison of median R^2 by sample size using all methods and irrespective of chromosome and target phenotype.

3.3.3 Effect of Different Number of Causal SNPs

There was no overarching effect due to SNP size affecting R^2 in all instances. Table 3.1 below shows the mean and median values of R^2 on the basis of the number of causative SNPs. Both mean and median tendencies were investigated for a pattern but none was evident. To better understand how the three different phenotypes operated, p-value thresholding was investigated using all three different phenotypes.

Table 3.1 Median and Mean of PRS Nested R^2 by number of causative SNPs.

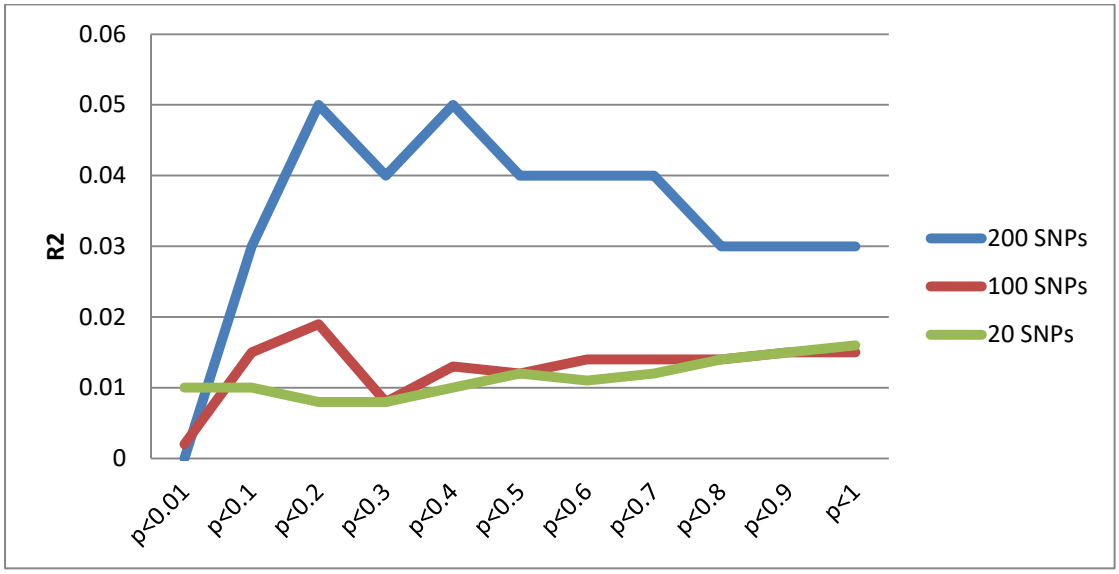
Causative SNPs	Median Nested R^2	Mean Nested R^2
200 SNPs	0.05	0.05
100 SNPs	0.03	0.05
20 SNPs	0.04	0.05

3.3.4 Effect of Differing P-Value Threshold Levels

Overall, 10 different threshold levels were applied to the polygenic score at 3 different LD pruning levels, resulting in 30 different pruning/thresholding results per scenario. In Figures 3.5 and 3.6, the effects of thresholding are presented in the datasets were the lowest and the highest level of R^2 were recorded. Figure 3.5 demonstrates the results for the sample that that had the worst predictive power of the true variation. Consistent with the results described above, that was the sample consisting of 500 individuals at chromosome 13. In this threshold, the scenario with 200 SNPs performed consistently

better than 20 or 100 SNPs throughout the process with a maximum nested R^2 of 0.043, while both 100 and 20 SNPs reached a maximum at a little over 0.02. Regarding thresholding in this scenario, the sample with 200 causative SNPs tended to perform better between 0.1<p-value<0.6 while the other two methods performed better at the highest thresholds (0.8<p-value<1).

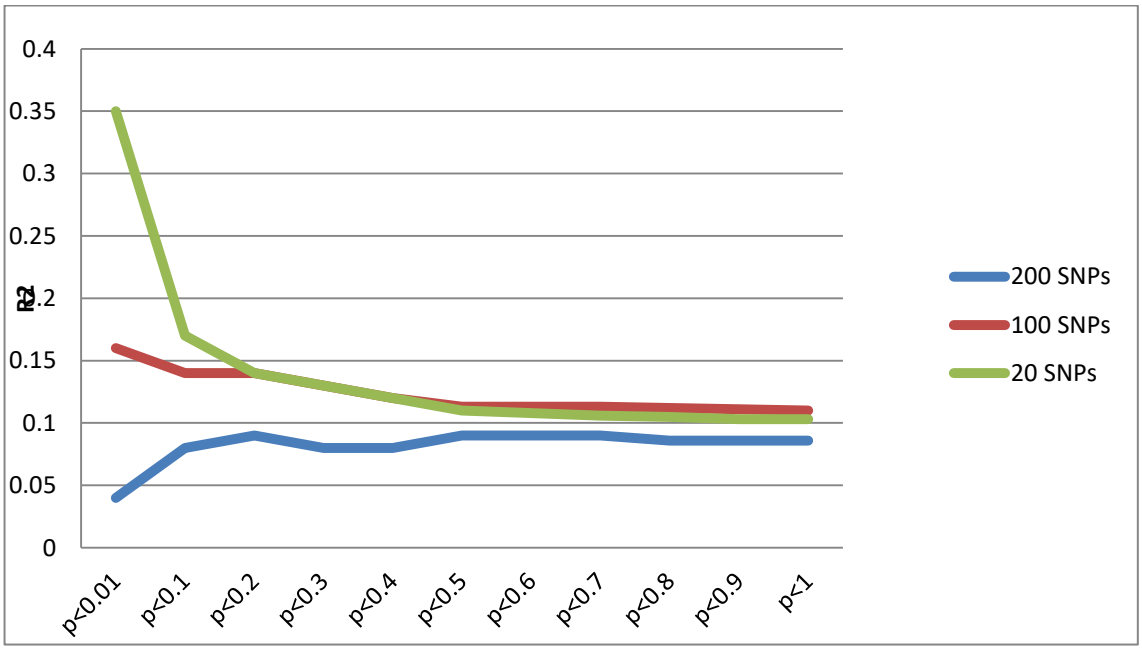
Figure 3.5 Median PRS nested R^2 across P-value thresholds at the scenario with the least predictive power



The results in that threshold were in stark contrast with those in the "best" scenario, that is, where the best results, in terms of outcome were produced (Figure 3.6). Here, the sample where 20 SNPs were used outperformed the other two, with R^2 reaching 0.35, while the other two reached a maximum nested R^2 of 0.15 (100 SNPs) and 0.1 (200 SNPs) respectively. Regarding thresholding, the lowest threshold used (p-value<0.01) was where the maximum level of R^2 was recorded for both the 20 and 100 SNP phenotypes, while the 200 SNP phenotype maximised its nested R^2 at p-value<0.2. It is

of note that although in this scenario the 20 SNP phenotypes worked better in the lowest thresholds, it goes down and follows the results of the other 2 phenotypes in higher inclusion p-value thresholds.

Figure 3.6 Median PRS nested R ^{2} across P-value thresholds at the scenario with the most predictive power.

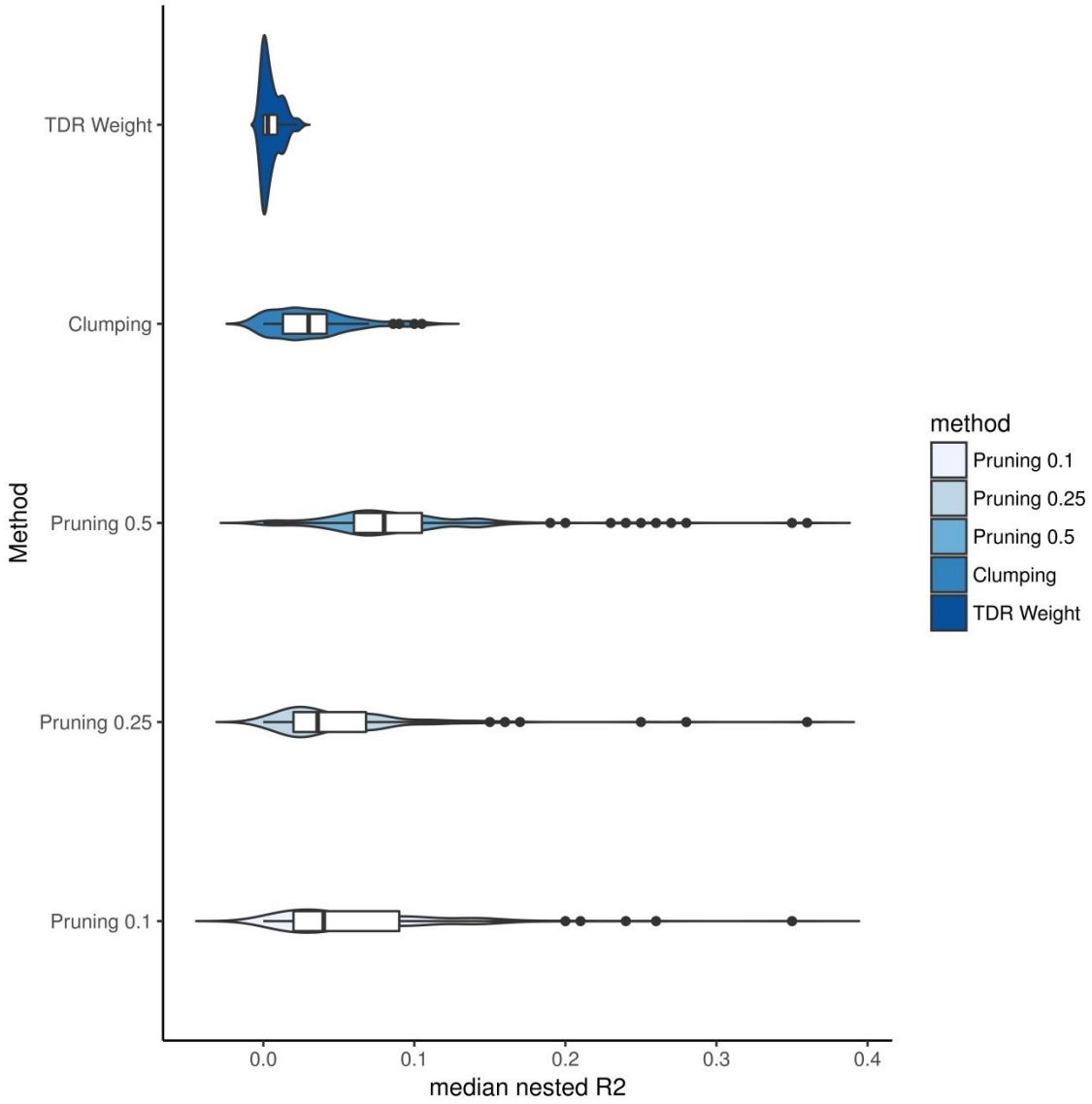


3.3.5 Comparison of methods

Figure 3.7 displays the results for the methods that were used in these simulations ranked from least to most successful in detecting the true outcome across all sample combinations. Weighted PRS had a median nested R^2 of 0.0025, LD clumping had a median nested R^2 of 0.028 (irrespective of clumping threshold), while LD pruning had a median nested R^2 between 0.028 and 0.04 (depending on the pruning threshold). Comparison of the three LD pruning thresholds showed that more aggressive pruning enabled better predictability for the model with pruning = 0.1 outperforming the other two at almost all of the scenario.

None of the methods were able to approximate the true R^2 value of 0.5 with a substantial difference between the true value and the results that on average any single method was able to produce.

Figure 3.7 Violin Plot of Median PRS nested R^2 by method

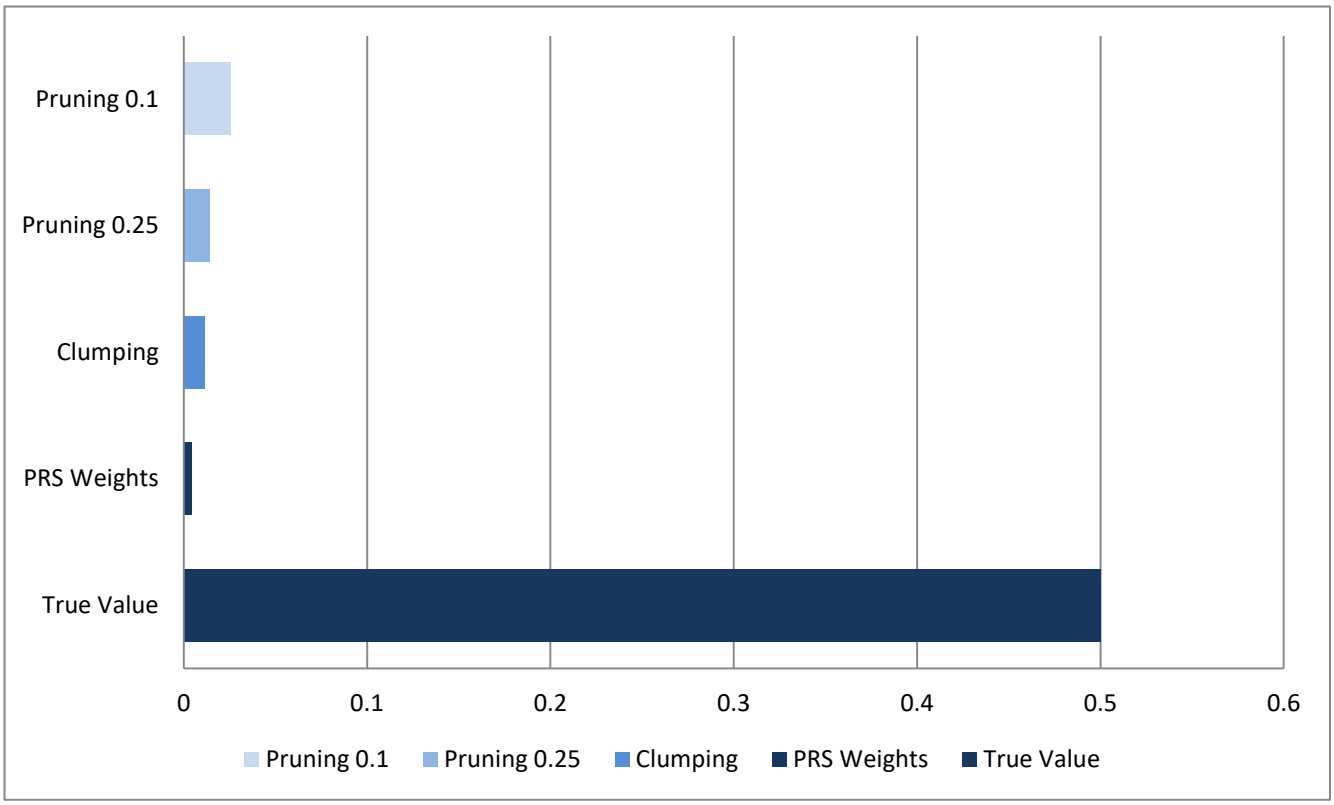


Comparison of median R^2 by method irrespective of chromosome, phenotype and sample size

3.3.6 Extended simulation application

For the extended simulation application, as the target was to approximate conditions of a real-life GWAS, a population of 100,000 was created through random pairs of haplotype combination and a sample of 80,000 was drawn from it. Half of those individuals were used as a discovery sample and half as a test sample. A total of 500 iterations of simulations were run, with a random distribution from a noise vector applied in each. Results of these simulations are presented in Figure 3.8a, compared to the real outcome. No method achieved a median nested R^2 greater than 0.1 and thus, despite reaching high levels of statistical significance (p-value$<10^{-10}$ on average) failed to capture the true amount of variation. In the comparison among the methods, maximum dimensionality reduction in the simulations seemed to yield the best results with the maximum pruning and thresholding approach (pruning all SNPs in LD of more than 0.1 and from those remaining taking only the SNPs with a p-value of less than 0.01 in the initial GWAS) finding a median R^2 result of 0.052 (or 5.2%), which was the maximum amount of variation that any of the methods was able to explain in the extended simulations.

Figure 3.8 Plot of Median PRS nested R^2 in the extended sample model (N=40,000) by method including the true value of 0.50



Comparison of median R^2 by method in the extended sample size simulation in chromosome 19.

3.4 Discussion

The aim of these simulations was to compare between methods currently being employed in polygenic score analysis across a number of different conditions. Additionally, those methods were implemented in an extended sample with a size equivalent to a modern day GWAS under optimal conditions. None of the methods were able to give optimal results in the conditions that were initial set. In the initial set of simulations, only the model where 20 SNPs were causative to the polygenic characteristic in the most stringent conditions of LD pruning (0.1) and thresholding ($p<0.01$) was able to yield results approximating the true value (0.35 versus 0.5). This

was achieved in the scenario where the population size (N=2500) was maximised in the most LD-rich of the three chromosomes that were selected (Chromosome 19). However, in the extended simulation where the population size was increased to 100,000 individuals and created a cohort of 40,000 as the discovery sample and an equal-sized cohort as the target sample, there was no improvement in the nested R^2 in any of investigated methods. Moreover, the result was swamped by the noise that the increased sample size seemed to amplify at the expense of the true signal.

These results are consistent with the models presented in Purcell (2009), which suggested that variants get swamped by noise as more SNPs are included in the model for most of the performed simulations. Indeed, these simulations are advantaged by the fact that they were simulated on samples that have larger number of people and markers than those examined in the initial PRS simulations by Purcell et al (2009). Additionally, a number of markers were used on a single chromosome in each simulation, as opposed to the whole genome. In the analysis, all the SNPs were taken into consideration; whereas models presented in the original report were only based on SNPs in linkage equilibrium (however the authors report a similar result pattern when all SNPs that were available to them were included). Finally, in the original report, the point was not to see whether risk scores would be able to detect results in a given sample; rather it was to select from a range of models (which included a range of options such as exponential models in it) and select the one that would best discern the underlying polygenic construct. This report does not contradict that a model along the lines of the original report is a viable solution to account for additivity; if only causative SNPs are included in the model, the correct nested R^2 value can be very reliably found. The problem comes from very high amounts of noise in data and the fact that current

methods fail to amplify the noise to signal ratio in such a way as to permit to draw meaningful conclusions from PRS results. This is amplified by the fact that, as surmised by Dudbridge et al (2013): "*polygenic risk score analyses are performed opportunistically*". This leads to results that are inconsistent between studies and offer additive variance estimates that ultimately can be contradicted even within the same datasets with tweaks in PRS parameters.

As polygenic scores have become integral to today's genetic research, there is a question as to how to maximise their efficacy. The set of simulations presented here indicates that increased sample sizes that will help with the identification of genome-wide significant hits in GWAS will not help improve the overall polygenic score modelling if all the information is included in it. This conclusion may seem counterintuitive given the fact that gradual increase of the PGC-schizophrenia sample has led to better results (Ripke et al, 2011; Ripke et al, 2013; Ripke et al, 2014). However, this might not be due to the sample size increase in the PRS application set, but, rather, due to bigger and better discovery sets which have led to more accurate GWAS estimates, These, in turn, enable the creation of better informed PRS, that although capturing a larger percentage of the true effect, still cannot come close to the true effect estimates. On the basis of the results presented here, different approaches are warranted if polygenic scores are to remain relevant in future genetic studies. One such approach would be to implement stricter cut-off criteria and only select SNPs that are genome-wide significant or at least in close proximity to that, which would substantially reduce noise in model. However, this approach would possibly endanger the loss of potentially useful information in SNPs with lower MAFs which would not be within the top ranked ones in a given GWAS. Another alternative approach would

be that suggested in Nicodemus et al (2014), proposing the inclusion of only specific molecular pathways in the score that are thought to be implicated in the disorder, thus keeping only information that is already known to be biologically meaningful. In this manner more precise conclusions could be drawn from polygenic score analysis and thus help in applying the results of modern genetic epidemiology in a translational manner.

3.4.1 Limitations

Despite the wide range of simulations undertaken in this chapter, there is a number of issues that need to be taken into consideration when considering these results.

First of all, as reported in the introduction there are other methods that could have been also been used in comparison to the ones proposed here, allowing for an even broader overview. These include but are not limited to methods such as LDPRED (Vilhjalmsson et al, 2014), LD Score Regression (Bullik-Sullivan et al, 2015) and PRSice (Eusden et al, 2015).

Next, the simulations themselves could also be designed to include the whole genome instead of (or in addition to) focusing on a single chromosome at a time. The reason this was not selected was twofold; first, there were concerns about the computational time and power that simulations including all chromosomes would take. Additionally, as LD is confined within the same chromosome and it was not within the scope of this project to take into account long range linkage disequilibria.

Finally, regarding the simulations themselves, although the simulated models strive to be as close to real-life conditions, there are two closely linked parameters that are

inflated in the simulations presented. The first one has to do with the amount of R^2 that can be explained by each SNP individually, as due to having a low number of SNPs in only one chromosome at a time, the individual R^2 values were well above those expected to be found in a realistic polygenic trait. Additionally, the maximum nimber of SNPs included in the models (200 SNPs) may be overly conservative given that the omnigenic model of schizophrenia (Boyle et al, 2017) proposes a very expansive polygenic background.

3.5 Conclusions

The main aim of the research presented here was to provide a comparison between methods that are currently being used for generating polygenic risk scores and determine which one would yield better results under a number of different conditions. These included different chromosomal linkage equilibrium and genic structure, different sample size and different underlying genotype. Finally, the methods under investigation were re-examined in a larger sample which would resemble recent large scale GWAS samples. All simulation conditions were repeated 500 times with variation stemming (a) from SNP selection in the first set of simulations or (b) from a random noise vector stemming in the secondary large simulation.

The main findings of the simulations that were conducted under the above conditions suggest that:

1. Sample size increase in the initial simulations increased the estimated median nested R^2 in a linear fashion

2. More complex LD patterns also resulted in an increase of the estimated median nested R^2.

3. Underlying genotypic variance did not yield consistent results, with R^2 not following a specific pattern under all simulated conditions. The underlying genotype with 200 SNPs was the one that yielded the best results under the majority of simulated models.

4. Direct comparison of methods showed that weighted and clumped scores underperformed when compared with scores pruned under the most stringent thresholds. However, even that method was unable to detect that the polygenic score was able to explain more than 10 percent of variance, on average, across simulations. Additionally, when looking at the scenario with the most predictive power, LD pruning at the lowest threshold was able to explain between 10 and 30 percent of the variance, depending from the underlying genotype of the trait.

6. In the secondary large simulation, all methods underperformed well below the linear improvement that could be predicted on the basis of the initial simulations on the basis of the sample size.

CHAPTER 4:

Can possible cryptic population stratification affect GCTA GREML estimates?

Examination of two traits in the Generation Scotland cohort.

4.1 Introduction

4.1.1 Background

The proportion of phenotypic variance explainable by common additive genetic variation has been the focus of multiple studies in the field of genetics (Sullivan et al, 2008; O'Donovan et al, 2008). Through constructs such as the polygenic risk scores (Purcell, 2009), there have been attempts to quantify this narrow sense heritability but a large amount predicted by earlier twin studies was still missing. The GREML (Genetic-relatedness-matrix Restricted-Maximum-Likelihood) application by Yang et al (2011) managed to reconcile the genetic data with previous studies, producing estimates that appear to be closer to the true heritability estimates of traits, and thus explaining away some of the missing heritability previously reported (Yang et al, 2011). Through the use of GWAS-sized samples consisting of thousands of individuals, GCTA attempts to extract an extremely small signal of genetic similarity among the noise from hundreds of thousands of SNPs. GCTA: a) generates a genetic relatedness matrix (GRM) from GWAS case and control genome-wide SNP data, b) prunes all pairwise comparisons for relatedness greater than approximately third cousins, c) checks how many more SNPs in common the ostensibly 'unrelated' cases have relative to controls and subsequently d) attributes any observed increased average genetic similarity of the case cohort relative to the control cohort to the underlying characteristic or disorder. Through a mixed-linear-model and restricted-maximum-likelihood approach, GREML extrapolates from the observed difference in average similarity between cases and controls an estimate for the total contribution of common variants to heritability of the disorder.

GCTA estimates have been used as evidence of the need for GWAS to expand in sample size and quality, in order to reveal the identities of a large number of common causal variants that may be hidden below genome-wide significance thresholds. This issue has prompted a closer inspection of GCTA to determine whether the extrapolation of GCTA results to such strong conclusions is justified. It has also brought some level of criticism to GCTA with some questioning the stability and reliability of these estimates (Krishna Kumar et al, 2015) and others proposing alternatives such as a Principal Component Analysis (PCA) to overcome shortcomings of the initial method (Dadousis et al, 2014).

Although GCTA allows for the implementation of a variety of methods to control for potentially confounding issues, such as the removal of related individuals through the exclusion of one member of every pair with relatedness greater than 0.025 (thought to be approximately equivalent to third cousins) and the inclusion of twenty principal components to control for within-sample population stratification, there are still potential confounding issues; as the GCTA signal could in fact be driven by an increased number of clusters of distantly related individuals among cases. The presence of cryptic population substructure differences between cases and controls would not be controlled by the PCA which could only attempt to reconciliate differences within populations but not across.

Recently, Evans et al (2017) also proposed a novel approach for GCTA, the GREML-IBD (Identity By Descent) which attempts to reconcile the previously established

GREML methodology with rare variant research through the use of similarity matrices for IBD segments across whole-genome sequencing data, in order to detect previously unreported rare variants in the population. However, they also reported that their inclusion thresholds for relatives could have been insufficient and that inclusion criteria merit further investigation.

4.1.2 Aims

The aim of this project was to investigate whether there is potential cryptic clustering within case and control populations after removal of relatives from GCTA and whether this could affect GREML estimates of GCTA. To achieve that, a number of clustering approaches was applied to both a GRM and a GRM-IBD and clustered individuals were removed to examine whether these might be driving the overall signal. The characteristics selected for this study were height and g (general intelligence coefficient). These were chosen due to the previously established polygenic nature of both traits (Lango Allen et al, 2010 – height; Davies et al, 2011 – intelligence, respectively). Additionally, g is of particular interest to the psychiatric genetics field as it has been demonstrated that schizophrenia polygenic scores are associated with g and that, in turn, g polygenic scores are associated with schizophrenia (Hubbard et al, 2016). Finally, the GRM IBD was included in this analysis to investigate its usefulness and its potential to explain the heritability in the same manner as the SNP-wise analysis.

4.2 Methods

4.2.1 Generation Scotland

The subsample from the Generation Scotland (GS) cohort used for this analysis included 7,372 unrelated individuals (Smith et al, 2013). Socio-demographic and genetic data were collected from about 24,000 volunteers across Scotland since 2006 with the data collection phase ending in 2011. DNA samples from 20,000 of those individuals was analysed through the use of high density genome-wide genotyping (Illumina OmniExpress SNP GWAS (700k)) and subsequently went through Quality Control (QC) analyses. The quality control was performed before the data were handed to the researcher.

4.2.2 Outcome Variables

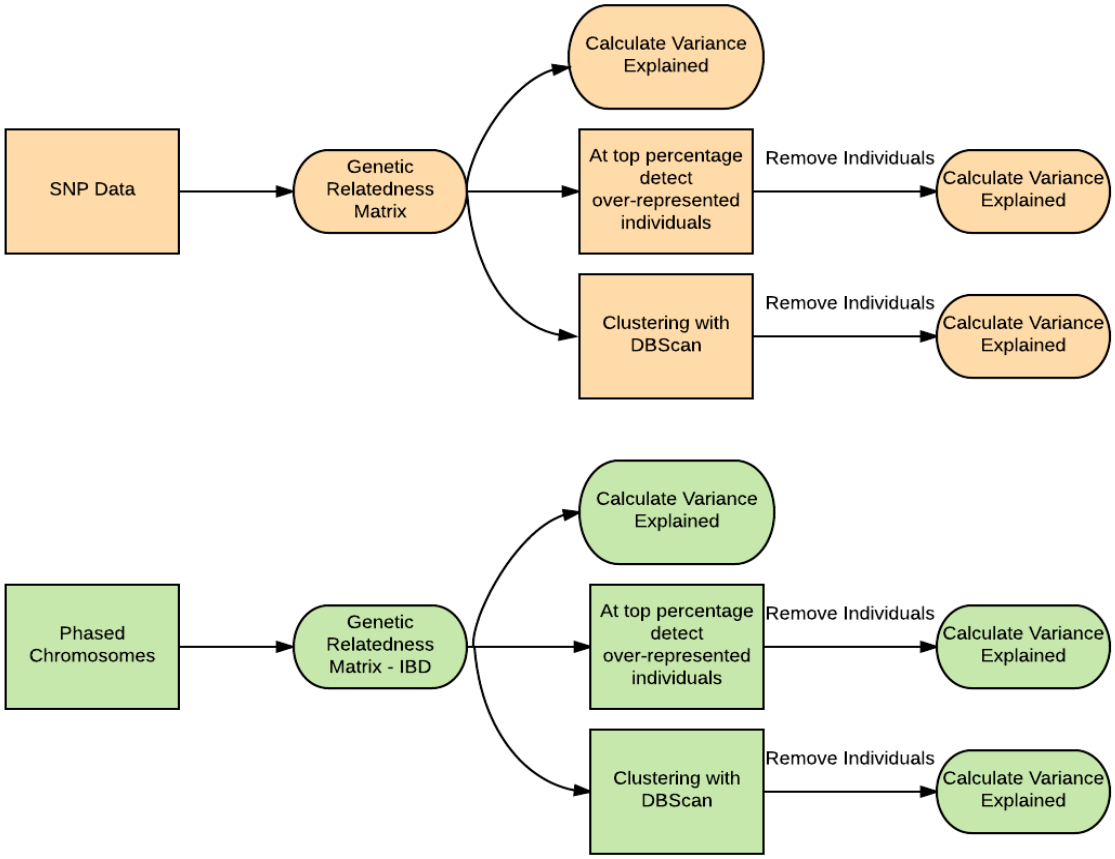
Height and g were selected as outcome variables of choice due to their established polygenic nature (Allen et al, 2010; Davies et al, 2011, respectively). Cognitive abilities for the generation of the g variable were assessed using four tests. Verbal ability was assessed using the Mill Hill Vocabulary Scale (Raven 1965). Immediate and delayed scores from the recall section of the Wechsler Logical Memory test were summed to provide a measure of verbal declarative memory (Wechsler 1997). The Wechsler Digit Symbol Coding test was used to measure processing speed (Wechsler 1997). Finally, executive function was measured using the letter-based phonemic verbal fluency test (Lezak 1995). The intelligence coefficient was extrapolated by performing a PCA on these four domains and taking the first unrotated principal component as an indicator of general intelligence (Smith et al, 2013). Height measurements were available for 6,390 individuals from unrelated individuals with genetic data. Cognitive measurements were

available for 6,413 unrelated individuals with genetic data. For the purposes of this analysis, both variables needed to be binarised. Towards that end, the variables were split as binary variables in the top and bottom third of their distribution to create a binary variable which would act as a proxy of the underlying continuous trait. Individuals with height above 172 cm were selected to be the in the "high" height group and individuals under 163 cm to be in the "low" height group. The splits for the g variable high and low groups were above 0.46 and below -0.35 accordingly.

4.2.3 Analysis Plan

Upon the definition of the two characteristics, two different approaches were chosen to investigate potential cryptic population stratification: a frequency-based "raw" approach and an agnostic clustering algorithm that could potentially identify subtle clustering differences between cases and controls. The process of analyses implemented is demonstrated in Figure 4.1 and described in detail below.

Figure 4.1 Flowchart of analysis for GCTA-GREML sensitivity



4.2.3.1 Software Used

R 3.3.1(R Core Team, 2016), plink 1.9 (Chang et al, 2015) and GCTA (Yang et al, 2011) were used for the main analysis, data preparation and manipulation.

As the genetic data were initially in genotype form, they were converted to phased haplotypes in SHAPEIT V2 (Delaneau et al, 2011) and segments were calculated using FISHR2 (Bjelland et al, 2017). The R statistical package was used for the implementation of the raw approach. GCTA-IBD (Evans et al, 2017) was used for the analysis and discovery of clusters in the haplotype data. Finally, R and the package dbscan (Hahsler & Piekenbrock, 2017) were used for the DBSCAN analysis in both the chromosomal block and the genotype SNP analysis.

4.2.3.2 Basic Analysis

One basic analysis and two types of clustering were implemented for each data type. After each clustering implementation, individuals in distinct clusters were removed and the GREML analysis was rerun without them. Before the implementation of the analyses, GCTA was used to remove from the sample individuals with a degree of relatedness above 0.025. Although a more relaxed parameter has been suggested (0.05), recent evidence (Evans et al, 2017) suggests that more stringent criteria may be necessary.

For the basic analysis, the SNP data from the cohort were used to create a genetic relatedness matrix, as described by GCTA (Yang et al, 2011). Subsequently, from this matrix the amount of variance explained by additive common variation was calculated. This was used as a base measure of variance explained by the outcome measures and a means of comparison as to whether subsequent removal of population substructure affected GCTA estimates. In the same manner, a GREML-IBD was created as described by Evans et al (2017) and the amount of variance explained by IBD segments was calculated as a baseline. In detail, to calculate phased data SHAPIT2 (Delaneau et al, 2013) was used. Subsequently, FISHR2 (Bjelland et al, 2017) was used to identify shared segments across all pairs of individuals. The parameters used were the same that Evans and colleagues (2017) had used (err_hom 4 --err_het 1 --min_snp 128 --min_cm_initial 1 --min_cm_final 1 --window 50 --gap 100 --h_extend --w_extend –homoz --emp--ma--threshold 0.06 --emp--pie--threshold 0.015 --count.gap.errors TRUE), as they reported that at that length (1 centiMorgan) and above the false positive rate of segments was low (<0.05). Finally, the length of all segments shared by each pair was summed in Mb (Megabases) and divided by the total length of the genome. This created a GRM whose elements would represent the amount of IBD variance shared between individuals.

4.2.3.3 Clustering Analyses

In the first clustering analysis, the GRM was split into a “case” and a “control” GRM. On the basis of those new matrices, the top 5 percent of frequencies was isolated. Within that percentage the instances that each individual ID appears were calculated. This provided a measure of representation of each individual, in terms of the top 5 percent of similarity. From that measure of representation at the top percentage,

individuals that were distributed beyond the 99th percentile were selected and subsequently removed. The process was repeated for both case and control matrices. Subsequently REML was rerun. This approach was used for both SNP and IBD matrices.

In the second clustering analysis, the DBSCAN algorithm was used. The DBSCAN family of clustering algorithms (Simoudis et al, 1997) have been used for k-nearest neighbour search through the use of a k-dimensional tree. The DBSCAN clustering process begins from an arbitrary starting point, examining the epsilon neighbourhood of that point (points within a set "ε" area). If the number of points found within the epsilon neighbourhood is greater to the value initially set as "minPts", a cluster is created. Once the creation of the cluster is established, DBSCAN expands the cluster to include all the points in the epsilon neighbourhoods of each other cluster member as well. The process stops once there are no more points within the ε vicinity of any of the cluster points. This process continues for the whole sample until all points have been either assigned to a cluster or classified as noise (i.e. no minimum points found within their epsilon neighbourhood). In this analysis, the DBSCAN algorithm was implemented in the case and the control GRMs separately. The parameters used in this analysis were a minPts value of 3 (as the algorithm suggests to empirically use a minPts value equal to the number of dimensions of the data plus one) and a epsilon neighbourhood of 0.118-0.120 (SNP GRMs) and 0.41-0.42 (IBD GRMs), on the basis of the knee of the k-nearest neighbour distance plot (Appendix 4.1). Members of clusters identified were removed before the REML analysis was re-run. This approach was also used for both SNP and IBD matrices. As DBSCAN needs half matrices to be implemented, a

transformation process was implemented in the GRMs to convert them in the ideal format for the DBSCAN analysis (Appendix 4.2).

4.3 Results

4.3.1 GRM REML

Initially REML analysis was performed in the GRM of the sample for both the binarised new traits as well as the initial continuous traits to ensure that there was a similar (or increased) amount of variance explained in both instances.

Table 4.1 Initial GRM REML Analysis

Characteristic	Sample Size	Variance explained	Standard Error	P-Value
Height	6390	0.375	0.055	1.15×10^{-13}
Height Binarised	4131	0.492	0.084	6.1×10^{-10}
G	6413	0.404	0.054	3.3×10^{-16}
g Binarised	4337	0.494	0.079	1.2×10^{-11}

As it can be observed in Table 4.1, the differences between cases and controls are even more pronounced in the binarised sample; this is due to the fact that the binary characteristic has a clearer separation (as all middle values were discarded from the sample), therefore making the differences among the top and the bottom of the sample more pronounced. The binarised results are less significant than the continuous traits,

but that is expected due to the fact that the samples were reduced by one third to create the binary traits.

4.3.1.1 Raw Approach

For the raw approach, the top 5% of the GRM for case-case and control-control comparisons were isolated. A frequency chart from those was generated and the 1% top hits were kept. These individuals had an above-average amount of similarity with a high percentage of their peers within their binarised allocated group and were removed from the cohort before rerunning the REML analysis, to investigate how that would affect the amount of variance explained. As the analysis was performed equally in the balanced high and the low matrices, the same number of individuals was removed from each (15 "cases" and 15 "controls" for g and 13 "cases and 14 "controls" for height).

Table 4.2 GRM REML "Raw" Analysis

Characteristic	Sample Size	Variance explained	Standard Error	P-Value
Height Binarised	4131	0.492	0.084	6.1×10^{-10}
Height Binarised (Top Individuals Removed)	4104	0.491	0.085	8.8×10^{-10}
g Binarised	4337	0.494	0.079	1.2×10^{-11}
g Binarised (Top Individuals Removed)	4307	0.498	0.080	1.2×10^{-11}

There was no shift in either variance explained or significance of results from the removal of those individuals in either of the two characteristics examined.

4.3.1.2 DBSCAN

Using the pre-specified epsilon neighbourhood and minimum points, DBSCAN was run in both case-case and control-control similarity sub matrices. For height, no clustering was detected at either the high or low groups with all points within defined as noise points from the algorithm. For g, 1 cluster of 6 individuals was identified in the low g group while no clustering was observed in the high g group.

Table 4.3 GRM REML DBSCAN Analysis

Characteristic	Sample Size	Variance explained	Standard Error	P-Value
g Binarised	4337	0.4939	0.079	1.2×10^{-11}
g Binarised (Cluster Removed)	4331	0.4943	0.079	1.2×10^{-11}

Again, as with the raw approach above no specific change was observed for g when the cluster of individuals from the low group was removed, indicating a robustness of GRM-REML to the types of clusters detected by DBScan.

4.3.2 GRM-IBD REML

From the IBD similarity lists, GRM-IBD matrices were generated. As these were similar to GRMs for regular REML analysis, REML was implemented using GCTA. For this estimation, only the binary variables were used. Below is a table of these results (Table 4.4).

Table 4.4 GRM-IBD Initial Analysis (1 cM)

Characteristic	Sample Size	Variance explained (IBD)	Standard Error	P-Value
Height Binarised	4131	0.524	0.144	4.2×10^{-5}
g Binarised	4337	0.515	0.221	1.05×10^{-6}

These results indicate that IBD variation within segments accounts for 52% of the variance explained in height and 51% of the variance explained in g. There is a significant decrease in p-value and an increase of the standard error of those calculations, indicating a lower amount of certainty in these. Indeed, as Evans et al (2017) indicated, these varied on the basis of the cM length of chunk definition with increase of minimum chunk length resulting in more erratic prediction results with lower degrees of confidence.

4.3.2.1 Raw Approach

The raw analysis followed a similar process to the one implemented for the GRM matrix, with the top 5% of the GRM-IBD for case-case and control-control comparisons isolated. Subsequently, a frequency chart from those was generated and the 1% top hits were kept. High percentage similar individuals were removed from the sample and REML was rerun in the reduced matrices.

Table 4.5 GRM-IBD "Raw" Analysis

Characteristic	Sample Size	Variance explained (IBD)	Standard Error	P-Value
Height Binarised	4131	0.524	0.144	4.2×10^{-5}
Height Binarised (Top Individuals Removed)	4097	0.502	0.165	4.17×10^{-5}
g Binarised	4337	0.515	0.221	1.05×10^{-6}
g Binarised (Top Individuals Removed)	4298	0.486	0.231	1.05×10^{-6}

The removal of a small proportion of individuals in this analysis had a bigger effect than in the SNP GRM. However, overall estimates remained high while the standard errors remained consistently high.

4.3.2.2 DBSCAN

Finally, using the pre-specified epsilon neighbourhood and minimum points, the DBSCAN algorithm was applied in both case-case and control-control IBD similarity sub-matrices**.** In the height data, no clusters were observed in the low height matrix of individuals. In the "high" height matrix, 2 and 3 clusters with a total of 15 and 24 individuals were observed. These individuals were removed and REML was run two times, for 2 and 3 clusters removed, accordingly. In the g analysis, one cluster was detected at all times for both the low and high g matrices (a total of 16 individuals). These clusters were removed and REML was run for the reduced sample.

Table 4.6 GRM-IBD DBSCAN Analysis

Characteristic	Sample Size	Variance explained (IBD)	Standard Error	P-Value
Height Binarised	4131	0.524	0.144	4.2×10^{-5}
Height Binarised (2 Clusters removed)	4116	0.519	0.152	4.2×10^{-5}
Height Binarised (3 Clusters removed)	4107	0.510	0.149	4.2×10^{-5}
g Binarised	4337	0.515	0.221	1.05×10^{-6}
g Binarised (1+1 clusters removed)	4321	0.513	0.229	1.05×10^{-6}

Removal of the clusters did not alter the estimate significantly with estimates of the new REML-IBD analyses being very close to the ones from the initial analysis.

4.4 Discussion

Two polygenic characteristics within the Generation Scotland cohort were investigated using REML and REML-IBD, and the proportion of variance explained by common and rare variants within those two characteristics was determined. Subsequently, two different clustering approaches were implemented on the common and the rare variant matrices to determine if population patterns detected from these methods could have an effect on the initial estimates. No method used in either REML or REML-IBD analysis significantly altered the results. However, IBD GRMs were more unstable and more likely to be influenced by such subtle structural changes than SNP GRMs. This is the first analysis to date attempting to compare and test those two types of matrices under different conditions and the first to test IBD-GRM for binary characteristics. Finally, the algorithm of DBSCAN is presented here as a useful tool for genetic analysis.

4.4.1 Limitations

There are several methodological aspects that may limit the applicability of these results. First of all, although the objective of this study was to observe the behaviour of GCTA-GREML in a binary characteristic with a clear polygenic background and an underlying liability threshold model, no such characteristic was found in the data available, with a sufficient sample size to produce reliable REML estimates. Thus the characteristics chosen, although polygenic and heritable, do not present the underlying structure and possibly population segregation characteristics that a trait such as schizophrenia might demonstrate.

This project demonstrated the application of the DBSCAN algorithm as a tool for detecting clusters within genetic relationship matrices. The parameters used were chosen on the basis of the dbscan R package manual. However, the algorithm is very sensitive to changes in its parameters and when run on different epsilon neighbourhood parameters, the number of clusters in its output can significantly vary.

There are other approaches to clustering that could have been used and have shown potential in detecting subtle patterns in dense matrices such as GRMs or IBD GRMs. MCLTribe (van Dongen, 2000; Enright et al, 2002) is a fast scalable clustering algorithm that identifies cluster structure in graphs through a random walk process, followed by matrix squaring, inflation and scaling until equilibrium is reached, at which point the graph is separated in clusters. In the field of genetics it has been previously successfully used in the detection of families of genes and proteins (Dolan et al, 2007; Bibollet-Bahena et al, 2017) and its random walk pattern-finding approach could be a useful way of navigating through a GRM. Additionally, Random Forest (Breiman, 2001) approaches to clustering individuals through the use of SNP data (but not similarity matrices) could provide an alternate way to find stable clusters of individuals with specific shared genetic background. Random Forest is a known machine-learning tool for genomic research (Chen and Ishwaran, 2012), so applying it in this type of analysis should be relatively straightforward.

Furthermore, it is worth noting that although GRM-IBD was one of the types of dataset that was used here, there are several other ways to build GRMs that attempt to integrate common SNP information with rare variation. In the original GRM-IBD report, the

authors propose LD-stratified GRMs (LDMS), as well as combinations of GRM-SNP and GRM-LDMS with GRM-IBD as ways to evaluate common and rare variation. The combined construct comprised of GRM-LDMS and GRM-IBD seemed to be the most accurate in predicting common and rare variation partitions of the total variance (Evans et al, 2017). Given its success in that report it would be useful to also include it in future comparisons of common and rare variation.

Finally, the GRM-IBD method was extremely prone to overestimate results and increase in variance when bigger chunks (>1cM) were used as a cut-off for IBD segments. This was also referenced in the original report with authors discussing the possibility of stricter initial cut-off points for family members (instead of 0.05 which was the value used in that study) as a way to reduce false positives which are bound to increase in IBD studies (Evans et al, 2017).

4.5 Conclusions

The main aim of the research presented in this Chapter was to examine two polygenic characteristics from the GS cohort using GREML and the novel GREML-IBD methods and investigate whether predictions of the two methods would remain stable when two different approaches to detect subtle population substructure were implemented.

The results of this study indicate that:

1) Both characteristics were highly polygenic in nature with common SNPs explaining 37 percent of height variability and 40 percent of g variability.

2) The two approaches that were implemented to detect population substructure did not alter the results significantly of either GREML or GREML-IBD.

3) When compared with its common variance counterpart, GREML-IBD produced more unstable estimates and was likely to change on the basis of original chunk length results.

CHAPTER 5:
Conclusion

5.1 Rationale and Aims of the Thesis

The investigation of the genetic architecture of schizophrenia has been at the forefront of psychiatric research for the last quarter of the century. Remarkable advancements have been made, however the particulars of its genetic structure remain unknown. Two of the most prominent methods currently employed in the quest to uncover how common variants might contribute to it, are the polygenic score (Purcell et al, 2009) and GREML applied by GCTA (Yang et al, 2011). Both these methods rely on a number of different parameters which may affect their results profoundly. It is therefore critical to examine their function and to illuminate the strengths and caveats of each method, thus enhancing the current understanding of schizophrenia genetics.

This PhD thesis makes an original contribution to the field in terms of investigating and testing primarily polygenic risk scores (Purcell et al, 2009) and secondarily GREML as applied through GCTA (Yang et al, 2011) in different scenarios and employing different parameters. The overall aim of the thesis was to optimise prediction of polygenic scores and GREML and define parameters under which they might optimally operate, in order to inform on current methodological applications on the field. The specific study aims were to:

- Apply these methods in real and simulated datasets as to ascertain their use under different conditions and
- Investigate the use of variation in the methodologies in improving the prediction and accuracy of the methods themselves.

5.2 Synopsis

In this thesis three separate and distinct research projects were undertaken with the aim to investigate the genetic architecture of schizophrenia as well as the current methods that are employed in common variance analyses. Two of these research projects investigated polygenic risk scores and one investigated GREML applied through GCTA.

For the first research project, polygenic risk scores (PRS) were investigated in the context of specific molecular pathways that could inform as to how these might be implicated in the architecture of schizophrenia. A number of different gene-sets were used to create polygenic risk scores on that basis, including genes regulated from FMRP (Steinberg et al, 2013), miR137 (Hill et al, 2013), *TCF4* (Forrest et al, 2013) and *CHD8* (Sugathan et al, 2014). The data were analysed within the PGC2-schizophrenia dataset for secondary analysis, which included 29,125 cases and 34,836 controls from 39 different centres (Ripke et al, 2014 describes the initial cohort). A leave-one-out approach was employed as a means of maximising the usefulness of the data available. Results indicated that a number of these pathways were implicated in the disorder, more so than the floor effect that was detected to implicate any random subset of SNPs of roughly equal size, as well as, the inflation of the floor effect that seemed to be present in random subsets of genic SNPs. The existence of this effect further indicates a possible omnigenic genetic background for schizophrenia (Boyle et al, 2017).

The second investigation was implemented with the aim to carefully examine polygenic

risk scores and investigate current approaches and parameters being employed at constructing them. The methods that were tested were Linkage Disequilibrium pruning and p-value thresholding (Purcell et al, 2009), LD clumping (Shi et al, 2011) and the weighing scheme proposed by Mak et al (2016). The Generation Scotland cohort (Smith et al, 2013) was used as the basis for constructing simulated phenotypes under different conditions to examine the behaviour of these approaches. Results of this investigation indicated that while all methods were able to detect that the effect was present, none of them could capture all the effect and most, grossly underestimated the contribution of common variance in the R^2 of the simulated traits. This was further demonstrated in a larger simulation that was conducted in a sample size similar to current biobanking efforts, emphasising the need for more precise and noise-resistant tools of polygenic score measurement.

In the third and final investigation of this thesis, an investigation of the other method currently employed to account for the sum of additive common variance effects, GREML, as applied through the GCTA software (Yang et al, 2011) was conducted. More specifically, the effect of population substructure in its estimates was explored through the application of two different approaches, a "raw" frequency-based approach and a clustering algorithm approach (DBSCAN in Simoudis et al, 1997). Furthermore, these clustering approaches were investigated in the context of the novel GREML-IBD approach that incorporates elements of the GREML approach to detect effects of rare variation. Again, the Generation Scotland cohort (Smith et al, 2013) was the basis for this investigation, with the characteristics of height and g (general intelligence), utilised as outcomes due to their inherent polygenic nature. Results of this investigation suggested that the GREML and GREML-IBD estimates remained relatively stable

when these sub-structures were removed.

5.3 Sources of Data

For the three analyses that were performed in the present thesis, two main sources of data were employed: The PGC2-schizophrenia cohort for secondary analysis (Ripke et al, 2014) and the Generation Scotland cohort (Smith et al, 2013). Both cohorts had strengths and limitations which are presented below. Furthermore, alterative contemporary data sources and their usefulness in the context of schizophrenia research are discussed.

5.3.1 PGC2-schizophrenia

The first dataset that was used was the PGC2-schizophrenia cohort. The PGC (Psychiatric Genomics Consortium) was created in 2007 with the aim to gather genetic data for psychiatric disorders for the purpose of conducting mega-analyses of genome-wide association studies for psychiatric disorders. The consortium has contributed significant results not only in the field of schizophrenia (Ripke et al, 2014) but also in the fields of a range of psychiatric disorders, recently including diagnoses as diverse as post-traumatic stress disorder (Logue et al, 2015) and anxiety (McGrath et al, 2013). The cohort that the aforementioned research project was implemented in constitutes a subsample of the initial cohort that was presented in Ripke and colleagues (2014); made available to researchers for secondary data analysis. Thirty-nine different studies were combined to make up the final cohort. Case ascertainment was implemented differently in many of these studies and the process of recruitment was assessed by as subgroup of investigators from the PGC Schizophrenia Working Group. Cases included had either schizophrenia or schizoaffective disorder, due to the fact that inter-rater reliability for

the two disorders is often low between groups (Faraone et al, 1996). Controls were collected from the same countries as the cases, with a large number of controls not screened for schizophrenia due to the fact that its prevalence in population samples is low.

The usefulness of this dataset for the current research project is self-evident, given that it would allow for a direct investigation of the genetic architecture of schizophrenia in the largest schizophrenia sample reported to date. To better utilise the sample structure and maximise the fact that it comprised of 39 different samples, a leave-one-out replication process was employed. This allowed for a more effective analysis than a simple singular split of the sample in discovery/target partitions. Additionally, the use of Meta-P (Ge, 2012) which takes into account sample size and effect directionality, allowed for a meta-analysed significance value that showed an effect of the gene-sets investigated that spread across the included populations.

However, there were a number of caveats that occur during the utilisation of a consortium sample from multiple study centres. Initially, there was the issue of the resulting mega-sample having an amplified population stratification effect, due to differing population with varied population size. For the purposes of this thesis, this was addressed through the use of a Principal Component Analysis (PCA) and the application of these Principal Components in both the polygenic score and GWAS processes that were implemented in the samples. Additionally, controls were selected from the same population but were not matched to cases on the basis of demographic characteristics. This may lead to other factors beyond the disorder itself to factor in the genetic differences between cases and controls. Again this was addressed, in part,

through the PC analysis, but as a number of those covariates were not known, it could not be further explored. Additionally, case ascertainment was not the same for all studies, with different criteria (DSM-IV, ICD-10) being used, while in other studies, specific endophenotypes were exclusively included. One such example is the CLOZUK sample, whose cases derived from a Clozapine trial where they were recruited on the basis of a diagnosis of treatment-resistant schizophrenia (Hamshere et al, 2013).

5.3.2 Generation Scotland

The Generation Scotland (GS) cohort, which was first conceptualised back in 2003, is a population cohort implemented in Scotland. Participants were identified and invited to participate from lists made available through GP practices. The initial wave of the cohort included individuals aged 35-65 from the Glasgow and Tayside areas, while the 2010 follow-up wave also included individuals from the Ayrshire, Arran and Northeast Scotland aged 18-65. Each participant was also required to invite a first degree relative aged 18-65 in order to participate. A total of 23,960 individuals were recruited in the study, but only a small subsample of those was unrelated. Beyond the genetic data, clinical measurements were taken from the participants; these included standardised physical and cognitive measurements. Finally, the data of the participants were linked to their medical records, giving researchers access to previously reported health or mental health issues.

For the purpose of this thesis, the GS cohort was used in the implementation of the second and third core analyses. In the polygenic score comparisons analyses (Chapter 3), random subsamples from the unrelated individuals were selected and subsequently simulated phenotypes were fit on the basis of the genotypes of these individuals. In the

second part of that investigation, the haplotypes from the GS cohort were randomly combined to create a larger independent cohort, to allow further investigation to the claims that PRS estimates would improve as sample sizes increased. In the third analysis (Chapter 4), the GS cohort of unrelated individuals was used to investigate the effect of population substructure in the estimates of GREML and IBD-GREML as those are applied by the GCTA software (Yang et al, 2011). The two variables that were selected for this analysis (height and g) were measured as part of the physical and cognitive assessment of the GS cohort and were chosen on the basis of their polygenic nature (Allen et al, 2010; Davies et al, 2011).

Unfortunately, there can be caveats in the use of a population cohort towards the investigation of a rare polygenic disorder. Schizophrenia, which was the main characteristic of interest in this thesis could not be directly investigated in the context of this cohort, as the prevalence of schizophrenia in the general population is quite low (McGrath et al, 2008) and therefore not present in sufficient numbers in the sample. Furthermore, the questionnaire of schizotypy that was completed by the cohort members, as a means to assess schizotypal traits in the general population (Raine, 1995), has since been revised (Fonseca-Pedrero et al, 2017) and shown to yield inconsistent results on the basis of cultural constructs (Liu et al, 2017). Finally, regarding the representativeness of the sample, despite the fact that the cohort was designed to capture the population of Scotland in terms of key demographic characteristics, there was considerable variation between the cohort respondents and the general population of Scotland in terms of age, gender, employment status and self-reported depression (Smith et al, 2013), as it true in most voluntary cohort studies.

5.3.3 Potential Future Sources of Data

Beyond the data sources that were described above, a range of other data sources could be potentially useful for similar investigations. There are two types of Data Biobanks that are currently being organised worldwide: a) Population Biobanks, which are large scale prospective studies, with multiple phenotypes and measurement points, aiming to investigate not only disorders but also determinants of those disorders over time and on the basis of environmental insults, and b) Clinical Biobanks, which focus on specific disorders and try to establish a consistent sampling process across disorders, as well as, create matched control cohorts for them. Below one example from each category will be briefly described in the context of its potential in terms of schizophrenia research.

5.3.3.1 Population Biobank: UK Biobank

The UK Biobank sample was recruited between 2006 and 2010 and includes 500,000 individuals aged between 40 and 69 (Allen et al, 2012). The aim of UK Biobank is to create a resource that will be available for all researchers and allow for the investigation of multiple phenotypes within its cohort, as well as, allow for the longitudinal investigation of outcomes, using detailed follow-ups on health and mental health outcomes of the participants. In terms of research relevant to this thesis, despite not being able to conduct schizophrenia-focused research in the same manner as in the Consortium sample, due to only having 1078 individuals with the disorder included, there are a number of other analyses that can only be performed in UK Biobank, purely due to the large number of phenotypes available in it. A new study published in 2017 investigated the relationship between a diagnosis of schizophrenia and lifestyle characteristics in UK Biobank (Firth et al, 2017). Another study used polygenic scores for schizophrenia derived from the PGC-2 (Ripke et al, 2014) to investigate how risk

for schizophrenia may be linked with other UK Biobank cohort characteristics (Smeland et al, 2017).

5.3.3.2 Clinical Biobank: The String of Pearls Initiative

The Parelsnoer Institute (PSI; http://www.parelsnoer.org) is a collaboration of the eight Dutch University Medical Centers (UMCs). It aims to create an infrastructure for collection of clinical and biological data from patients with chronic diseases. Each one of the diseases that are investigated constitutes one of the "Pearls" of the initiative. Within each pearl multiple phenotypes are quantified, including imaging and biometric results, and all resources are made available for the institutes participating in the initiative. The PSI does not have a dedicated branch for mental health yet; however, a similar initiative for mental health disorders could offer researchers a plethora of phenotypes and facilitate research into psychiatric cognitive and clinical endophenotypes.

5.4 Methodological Implications and Future Directions

Two methods were investigated in the context of this thesis; the polygenic risk score (Purcell et al, 2009) and GREML applied through GCTA (Yang et al, 2011). Below the implications deriving from the findings of this thesis for the two methodologies are elaborated. Furthermore, current methodological developments in the field are discussed in the context of how they may be incorporated in future research.

5.4.1 Polygenic Risk Scores

In Chapter 2, the potential of Gene-set specific polygenic risk scores was demonstrated. This thesis reaffirmed the role of a number of gene-sets, by using new experimentally evaluated sets of genes regulated by FMRP, *TCF4*, miR137 and *CHD8* and demonstrated the effect of these gene-sets in predicting schizophrenia, above and beyond similarly-sized sets of genes related to heart disease and cancer. Future work regarding the incorporation of biological information in polygenic scores could lead to a weighing scheme on the basis of biological information derived from experimental conditions to bolster the polygenic signal and decrease noise.

Moreover, a floor effect for schizophrenia risk scores was demonstrated through the selection of subsets with random SNPs. This ties conceptually with the newly proposed omnigenic model for schizophrenia and its theoretical framework, proposing that thousands of genes contribute to the common polygenic background of the disorder (Boyle et al, 2017). An alternative explanation of the floor effect could indicate some sort of artificial inflation of the sample and would stress the need for rigorous re-evaluation of the samples on the basis of unexplored systematic bias in the existing samples. The floor effect observed, is reaffirmed by the Q-Q plot presented in the study (Figure 2.5), as well as the Q-Q plot of the original PGC2 study (Ripke et al, 2014) which indicate an inflation of p-values at all thresholds above 0.01.

Regarding the comparison of methods conducted in the third Chapter, none of the methods examined were able to optimally detect the true polygenic signal. Future work in the field of refining polygenic risk score methodology would also test some of the additional methods currently in use for the generation of polygenic risk scores; PRSice

(Eusden et al, 2015) and LDPRED (Vilhjalmsson et al, 2014). Both methods are described in Chapter one and would be valuable additions in future comparative simulations. Finally, the models that were created operated under a conservative polygenic model. Given current considerations (Boyle et al, 2017) and the results from the second Chapter, which indicate a broader polygenic background, an additional model with 1000-5000 causative SNPs of very low effect sizes per chromosome could be proposed as an additional simulated condition.

5.4.2 GREML

In Chapter 4, the idea behind population cryptic relatedeness was explored in the GS cohort through the use of a "raw" frequency-based approach and DBSCAN for the purpose of detecting clustering. There are a number of future directions for this study going forward. First of all, despite the fact that the analysis did not demonstrate significant differences, it was performed on a truly continuous characteristic and not on a binary construct such as schizophrenia. Therefore, the analysis conducted here should be replicated in a case-control psychiatric cohort to test the hypothesis of this analysis on a differentially structured sample. The idea behind the analysis would be better suited for such a cohort, as recruiting would be conducted in a different manner in psychiatric cases and controls and thus account for further genetic differences between the two. Furthermore, there are other clustering algorithms that could be investigated within the context of this analysis, previously summarised in Chapter 4. Finally, it is of note that the new GREML-IBD report by Evans and colleagues (2017) presents a number of different approaches in the construction of a GRM which merit further investigation, with regards to the optimal way a similarity matrix could be created in order to maximise use of the available genetic information.

In conclusion, the potential of current genetic analysis methods is demonstrated throughout this thesis. The results presented here highlight the fact that, if understood and applied cautiously, these methods can be valuable research tools, contributing towards an in depth understanding of the genetic architecture of schizophrenia.

References

Aberg, K. A., Liu, Y., Bukszár, J., McClay, J. L., Khachane, A. N., Andreassen, O. A., … van den Oord, E. J. (2013). A Comprehensive Family-Based Replication Study of Schizophrenia Genes. *JAMA Psychiatry*, *70*(6), 573. https://doi.org/10.1001/jamapsychiatry.2013.288

Åberg, K., Adkins, D. E., Bukszár, J., Webb, B. T., Caroff, S. N., Miller, D. D., … van den Oord, E. J. C. G. (2010). Genomewide Association Study of Movement-Related Adverse Antipsychotic Effects. *Biological Psychiatry*, *67*(3), 279–282. https://doi.org/10.1016/j.biopsych.2009.08.036

Åberg, K., Adkins, D. E., Liu, Y., McClay, J. L., Bukszár, J., Jia, P., … van den Oord, E. J. C. G. (2012). Genome-wide association study of antipsychotic-induced QTc interval prolongation. *The Pharmacogenomics Journal*, *12*(2), 165–172. https://doi.org/10.1038/tpj.2010.76

Ahmed, A. O., Mantini, A. M., Fridberg, D. J., & Buckley, P. F. (2015). Brain-derived neurotrophic factor (BDNF) and neurocognitive deficits in people with schizophrenia: A meta-analysis. *Psychiatry Research*, *226*(1), 1–13. https://doi.org/10.1016/j.psychres.2014.12.069

Alkelai, A., Greenbaum, L., Rigbi, A., Kanyas, K., & Lerer, B. (2009). Genome-wide association study of antipsychotic-induced parkinsonism severity among schizophrenia patients. *Psychopharmacology*, *206*(3), 491–499. https://doi.org/10.1007/s00213-009-1627-z

Alkelai, A., Lupoli, S., Greenbaum, L., Giegling, I., Kohn, Y., Sarner-Kanyas, K., … Lerer, B. (2011). Identification of new schizophrenia susceptibility loci in an ethnically homogeneous, family-based, Arab-Israeli sample. *The FASEB Journal*, *25*(11), 4011–4023. https://doi.org/10.1096/fj.11-184937

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., … Collins, R. (2012). UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, *1*(3), 123–126. https://doi.org/10.1016/j.hlpt.2012.07.003

Altamura, A. C., Boin, F., & Maes, M. (1999). HPA axis and cytokines dysregulation in schizophrenia: potential implications for the antipsychotic treatment. *European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology*, *10*(1), 1–4.

American Psychiatric Association, & American Psychiatric Association (Eds.). (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR* (4th ed., text revision). Washington, DC: American Psychiatric Association.

American Psychiatric Association, & American Psychiatric Association (Eds.). (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed). Washington, D.C: American Psychiatric Association.

Amiel, J., Rio, M., Pontual, L. de, Redon, R., Malan, V., Boddaert, N., … Colleaux, L. (2007). Mutations in *TCF4*, Encoding a Class I Basic Helix-Loop-Helix Transcription Factor, Are Responsible for Pitt-Hopkins Syndrome, a Severe Epileptic Encephalopathy

Associated with Autonomic Dysfunction. *The American Journal of Human Genetics*, *80*(5), 988–993. https://doi.org/10.1086/515582

Aminoff, S. R., Tesli, M., Bettella, F., Aas, M., Lagerberg, T. V., Djurovic, S., … Melle, I. (2015). Polygenic risk scores in bipolar disorder subgroups. *Journal of Affective Disorders*, *183*, 310–314. https://doi.org/10.1016/j.jad.2015.05.021

Amminger, G. P., Pape, S., Rock, D., Roberts, S. A., Squires-Wheeler, E., Kestenbaum, C., & Erlenmeyer-Kimling, L. (2000). The New York High-Risk Project: comorbidity for axis I disorders is preceded by childhood behavioral disturbance. *The Journal of Nervous and Mental Disease*, *188*(11), 751–756.

Andreasen, N. C., Wilcox, M. A., Ho, B.-C., Epping, E., Ziebell, S., Zeien, E., … Wassink, T. (2012). Statistical epistasis and progressive brain change in schizophrenia: an approach for examining the relationships between multiple genes. *Molecular Psychiatry*, *17*(11), 1093–1102. https://doi.org/10.1038/mp.2011.108

Athanasiu, L., Brown, A. A., Birkenaes, A. B., Mattingsdal, M., Agartz, I., Melle, I., … Djurovic, S. (2012). Genome-wide association study identifies genetic loci associated with body mass index and high density lipoprotein-cholesterol levels during psychopharmacological treatment — a cross-sectional naturalistic study. *Psychiatry Research*, *197*(3), 327–336. https://doi.org/10.1016/j.psychres.2011.12.036

Athanasiu, L., Mattingsdal, M., Kähler, A. K., Brown, A., Gustafsson, O., Agartz, I., … Andreassen, O. A. (2010). Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. *Journal of Psychiatric Research*, *44*(12), 748–753. https://doi.org/10.1016/j.jpsychires.2010.02.002

Avramopoulos, D., Pearce, B. D., McGrath, J., Wolyniec, P., Wang, R., Eckart, N., … Pulver, A. E. (2015). Infection and Inflammation in Schizophrenia and Bipolar Disorder: A Genome Wide Study for Interactions with Genetic Variation. *PLOS ONE*, *10*(3), e0116696. https://doi.org/10.1371/journal.pone.0116696

Badner, J. A., & Gershon, E. S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular Psychiatry*, *7*(4), 405–411. https://doi.org/10.1038/sj.mp.4001012

Baig, B. J., Whalley, H. C., Hall, J., McIntosh, A. M., Job, D. E., Cunningham-Owens, D. G., … Lawrie, S. M. (2010). Functional magnetic resonance imaging of BDNF val66met polymorphism in unmedicated subjects at high genetic risk of schizophrenia performing a verbal memory task. *Psychiatry Research: Neuroimaging*, *183*(3), 195–201. https://doi.org/10.1016/j.pscychresns.2010.06.009

Bakken, T. E. (2011). Association of Genetic Variants on 15q12 With Cortical Thickness and Cognition in Schizophrenia. *Archives of General Psychiatry*, *68*(8), 781. https://doi.org/10.1001/archgenpsychiatry.2011.81

Batsukh, T., Pieper, L., Koszucka, A. M., von Velsen, N., Hoyer-Fender, S., Elbracht, M., … Pauli, S. (2010). *CHD8* interacts with CHD7, a protein which is mutated in CHARGE syndrome. *Human Molecular Genetics*, *19*(14), 2858–2866. https://doi.org/10.1093/hmg/ddq189

Beards, S., Gayer-Anderson, C., Borges, S., Dewey, M. E., Fisher, H. L., & Morgan, C. (2013). Life Events and Psychosis: A Review and Meta-analysis. *Schizophrenia Bulletin*, *39*(4), 740–747. https://doi.org/10.1093/schbul/sbt065

Bechter, K., Reiber, H., Herzog, S., Fuchs, D., Tumani, H., & Maxeiner, H. G. (2010). Cerebrospinal fluid analysis in affective and schizophrenic spectrum disorders: Identification of subgroups with immune responses and blood–CSF barrier dysfunction. *Journal of Psychiatric Research*, *44*(5), 321–330. https://doi.org/10.1016/j.jpsychires.2009.08.008

Belsky, D. W., Sears, M. R., Hancox, R. J., Harrington, H., Houts, R., Moffitt, T. E., … Caspi, A. (2013). Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. *The Lancet Respiratory Medicine*, *1*(6), 453–461. https://doi.org/10.1016/S2213-2600(13)70101-2

Belsky, J. (1997). Theory testing, effect-size evaluation, and differential susceptibility to rearing influence: the case of mothering and attachment. *Child Development*, *68*(4), 598–600.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. *Psychological Bulletin*, *135*(6), 885–908. https://doi.org/10.1037/a0017376

Bendall, S., Jackson, H. J., Hulbert, C. A., & McGorry, P. D. (2007). Childhood Trauma and Psychotic Disorders: a Systematic, Critical Review of the Evidence. *Schizophrenia Bulletin*, *34*(3), 568–579. https://doi.org/10.1093/schbul/sbm121

Bener, A., Dafeeah, E. E., & Samson, N. (2012). Does consanguinity increase the risk of schizophrenia? Study based on primary health care centre visits. *Mental Health in Family Medicine*, *9*(4), 241–248.

Benjamin, D. J., Cesarini, D., van der Loos, M. J. H. M., Dawes, C. T., Koellinger, P. D., Magnusson, P. K. E., … Visscher, P. M. (2012). The genetic architecture of economic and political preferences. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(21), 8026–8031. https://doi.org/10.1073/pnas.1120666109

Bentall, R. P., de Sousa, P., Varese, F., Wickham, S., Sitko, K., Haarmans, M., & Read, J. (2014). From adversity to psychosis: pathways and mechanisms from specific adversities to specific symptoms. *Social Psychiatry and Psychiatric Epidemiology*, *49*(7), 1011–1022. https://doi.org/10.1007/s00127-014-0914-0

Bergen, S. E., O'Dushlaine, C. T., Ripke, S., Lee, P. H., Ruderfer, D. M., Akterin, S., … Sullivan, P. F. (2012). Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Molecular Psychiatry*, *17*(9), 880–886. https://doi.org/10.1038/mp.2012.73

Betcheva, E. T., Yosifova, A. G., Mushiroda, T., Kubo, M., Takahashi, A., Karachanak, S. K., … Nakamura, Y. (2013). Whole-genome-wide association study in the Bulgarian population reveals HHAT as schizophrenia susceptibility gene: *Psychiatric Genetics*, *23*(1), 11–19. https://doi.org/10.1097/YPG.0b013e3283586343

Bibollet-Bahena, O., Okafuji, T., Hokamp, K., Tear, G., & Mitchell, K. J. (2017). A dual-strategy expression screen for candidate connectivity labels in the developing thalamus. *PLOS ONE*, *12*(5), e0177977. https://doi.org/10.1371/journal.pone.0177977

Bjelland, D. W., Lingala, U., Patel, P. S., Jones, M., & Keller, M. C. (2017). A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data.

European Journal of Human Genetics, *25*(5), 617–624. https://doi.org/10.1038/ejhg.2017.6

Blackwood, D. H., Fordyce, A., Walker, M. T., St Clair, D. M., Porteous, D. J., & Muir, W. J. (2001). Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *American Journal of Human Genetics*, *69*(2), 428–433.

Boardman, J. D., Domingue, B. W., & Daw, J. (2015). What can genes tell us about the relationship between education and health? *Social Science & Medicine*, *127*, 171–180. https://doi.org/10.1016/j.socscimed.2014.08.001

Boin, F., Zanardini, R., Pioli, R., Altamura, C. A., Maes, M., & Gennarelli, M. (2001). Association between −G308A tumor necrosis factor alpha gene polymorphism and schizophrenia. *Molecular Psychiatry*, *6*(1), 79–82. https://doi.org/10.1038/sj.mp.4000815

Bonnet-Brilhault, F., Laurent, C., Campion, D., Thibaut, F., Lafargue, C., Charbonnier, F., … Mallet, J. (1999). No evidence for involvement of KCNN3 (hSKCa3) potassium channel gene in familial and isolated cases of schizophrenia. *European Journal of Human Genetics*, *7*(2), 247–250. https://doi.org/10.1038/sj.ejhg.5200278

Børglum, A. D., Demontis, D., Grove, J., Pallesen, J., Hollegaard, M. V., Pedersen, C. B., … Mors, O. (2014). Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Molecular Psychiatry*, *19*(3), 325–333. https://doi.org/10.1038/mp.2013.2

Bose, K. S., & Sarma, R. H. (1975). Delineation of the intimate details of the backbone conformation of pyridine nucleotide coenzymes in aqueous solution. *Biochemical and Biophysical Research Communications*, *66*(4), 1173–1179.

Bourque, F., van der Ven, E., & Malla, A. (2011). A meta-analysis of the risk for psychotic disorders among first- and second-generation immigrants. *Psychological Medicine*, *41*(05), 897–910. https://doi.org/10.1017/S0033291710001406

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Bradshaw, N. J., & Porteous, D. J. (2012). DISC1-binding proteins in neural development, signalling and schizophrenia. *Neuropharmacology*, *62*(3), 1230–1241. https://doi.org/10.1016/j.neuropharm.2010.12.027

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bromet, E. J., Kotov, R., Fochtmann, L. J., Carlson, G. A., Tanenberg-Karant, M., Ruggero, C., & Chang, S. (2011). Diagnostic Shifts During the Decade Following First Admission for Psychosis. *American Journal of Psychiatry*, *168*(11), 1186–1194. https://doi.org/10.1176/appi.ajp.2011.11010048

Brown, A. S., Schaefer, C. A., Wyatt, R. J., Begg, M. D., Goetz, R., Bresnahan, M. A., … Susser, E. S. (2002). Paternal Age and Risk of Schizophrenia in Adult Offspring. *American Journal of Psychiatry*, *159*(9), 1528–1533. https://doi.org/10.1176/appi.ajp.159.9.1528

Brown, A. S., & Susser, E. S. (2008). Prenatal Nutritional Deficiency and Risk of Adult

Schizophrenia. *Schizophrenia Bulletin*, *34*(6), 1054–1063. https://doi.org/10.1093/schbul/sbn096

Brzustowicz, L. M., Honer, W. G., Chow, E. W. C., Little, D., Hogan, J., Hodgkinson, K., & Bassett, A. S. (1999). Linkage of Familial Schizophrenia to Chromosome 13q32. *The American Journal of Human Genetics*, *65*(4), 1096–1103. https://doi.org/10.1086/302579

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, … Neale, B. M. (2015). LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nature Genetics*, *47*(3), 291–295. http://doi.org/10.1038/ng.3211

Cannon, T. D., & Mednick, S. A. (1993). The schizophrenia high-risk project in Copenhagen: three decades of progress. *Acta Psychiatrica Scandinavica. Supplementum*, *370*, 33–47.

Cantor-Graae, E., & Selten, J.-P. (2005). Schizophrenia and Migration: A Meta-Analysis and Review. *American Journal of Psychiatry*, *162*(1), 12–24. https://doi.org/10.1176/appi.ajp.162.1.12

Cao, Q., Martinez, M., Zhang, J., Sanders, A. R., Badner, J. A., Cravchik, A., … Gejman, P. V. (1997). Suggestive Evidence for a Schizophrenia Susceptibility Locus on Chromosome 6q and a Confirmation in an Independent Series of Pedigrees. *Genomics*, *43*(1), 1–8. https://doi.org/10.1006/geno.1997.4815

Carlsson, A., & Lindqvist, M. (1963). Effect Of Chlorpromazine Or Haloperidol On Formation Of 3methoxytyramine And Normetanephrine In Mouse Brain. *Acta Pharmacologica Et Toxicologica*, *20*, 140–144.

Carroll, L. S., & Owen, M. J. (2009). Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Medicine*, *1*(10), 102. https://doi.org/10.1186/gm102

Caspi, A. (2002). Role of Genotype in the Cycle of Violence in Maltreated Children. *Science*, *297*(5582), 851–854. https://doi.org/10.1126/science.1072290

Chandy, K. G., Fantino, E., Wittekindt, O., Kalman, K., Tong, L. L., Ho, T. H., … Gargus, J. J. (1998). Isolation of a novel potassium channel gene hSKCa3 containing a polymorphic CAG repeat: a candidate for schizophrenia and bipolar disorder? *Molecular Psychiatry*, *3*(1), 32–37.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0047-8

Chavarria-Siles, I., White, T., de Leeuw, C., Goudriaan, A., Lips, E., Ehrlich, S., … Posthuma, D. (2016). Myelination-related genes are associated with decreased white matter integrity in schizophrenia. *European Journal of Human Genetics*, *24*(3), 381–386. https://doi.org/10.1038/ejhg.2015.120

Chen, J., Lee, G., Fanous, A. H., Zhao, Z., Jia, P., O'Neill, A., … International Schizophrenia Consortium. (2011). Two non-synonymous markers in PTPN21, identified by genome-wide association study data-mining and replication, are associated with schizophrenia. *Schizophrenia Research*, *131*(1–3), 43–51. https://doi.org/10.1016/j.schres.2011.06.023

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*,

99(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

Clark, S. L., Souza, R. P., Adkins, D. E., Åberg, K., Bukszár, J., McClay, J. L., … van den Oord, E. J. C. G. (2013). Genome-wide association study of patient-rated and clinician-rated global impression of severity during antipsychotic treatment: *Pharmacogenetics and Genomics*, *23*(2), 69–77. https://doi.org/10.1097/FPC.0b013e32835ca260

Collins, A. L., Kim, Y., Bloom, R. J., Kelada, S. N., Sethupathy, P., & Sullivan, P. F. (2014). Transcriptional targets of the schizophrenia risk gene MIR137. *Translational Psychiatry*, *4*(7), e404. https://doi.org/10.1038/tp.2014.42

Coon, H., Holik, J., Hoff, M., Reimherr, F., Wender, P., Myles-Worsley, M., … Byerley, W. (1994). Analysis of chromosome 22 markers in nine schizophrenia pedigrees. *American Journal of Medical Genetics*, *54*(1), 72–79. https://doi.org/10.1002/ajmg.1320540112

Crocq, M. A., Mant, R., Asherson, P., Williams, J., Hode, Y., Mayerova, A., … Schwartz, J. C. (1992). Association between schizophrenia and homozygosity at the dopamine D3 receptor gene. *Journal of Medical Genetics*, *29*(12), 858–860.

Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)*, *381*(9875), 1371–1379. https://doi.org/10.1016/S0140-6736(12)62129-1

Crow, T. J. (2007). How and Why Genetic Linkage Has Not Solved the Problem of Psychosis: Review and Hypothesis. *American Journal of Psychiatry*, *164*(1), 13–21. https://doi.org/10.1176/ajp.2007.164.1.13

Curtis, D. (2016). Pathway analysis of whole exome sequence data provides further support for the involvement of histone modification in the aetiology of schizophrenia: *Psychiatric Genetics*, *26*(5), 223–227. https://doi.org/10.1097/YPG.0000000000000132

Curtis, D., Vine, A. E., McQuillin, A., Bass, N. J., Pereira, A., Kandaswamy, R., … Gurling, H. M. D. (2011). Case–case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes: *Psychiatric Genetics*, *21*(1), 1–4. https://doi.org/10.1097/YPG.0b013e3283413382

Dadousis, C., Veerkamp, R. F., Heringstad, B., Pszczola, M., & Calus, M. P. (2014). A comparison of principal component regression and genomic REML for genomic prediction across populations. *Genetics Selection Evolution*, *46*(1). https://doi.org/10.1186/s12711-014-0060-x

Davies, G., Marioni, R. E., Liewald, D. C., Hill, W. D., Hagenaars, S. P., Harris, S. E., … Deary, I. J. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Molecular Psychiatry*, *21*(6), 758–767. https://doi.org/10.1038/mp.2016.45

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., … Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*, *16*(10), 996–1005. https://doi.org/10.1038/mp.2011.85

Davies, G., Welham, J., Chant, D., Torrey, E. F., & McGrath, J. (2003). A systematic review and meta-analysis of Northern Hemisphere season of birth studies in schizophrenia. *Schizophrenia Bulletin*, *29*(3), 587–593.

Davis, L. K., Yu, D., Keenan, C. L., Gamazon, E. R., Konkashbaev, A. I., Derks, E. M., … Scharf, J. M. (2013). Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. *PLoS Genetics*, *9*(10), e1003864. https://doi.org/10.1371/journal.pgen.1003864

de Sousa, P., Varese, F., Sellwood, W., & Bentall, R. P. (2014). Parental Communication and Psychosis: A Meta-analysis. *Schizophrenia Bulletin*, *40*(4), 756–768. https://doi.org/10.1093/schbul/sbt088

Delaneau, O., Marchini, J., & Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*(2), 179–181. https://doi.org/10.1038/nmeth.1785

Derks, E. M., Dolan, C. V., & Boomsma, D. I. (2006). A Test of the Equal Environment Assumption (EEA) in Multivariate Twin Studies. *Twin Research and Human Genetics*, *9*(3), 403–411. https://doi.org/10.1375/183242706777591290

Detera-Wadleigh, S. D., Goldin, L. R., Sherrington, R., Encio, I., Miguel, C. de, Berrettini, W., … Gershon, E. S. (1989). Exclusion of linkage to 5qll–13 in families with schizophrenia and other psychiatric disorders. *Nature*, *340*(6232), 391–393. https://doi.org/10.1038/340391a0

Diener, E., & Biswas-Diener, R. (n.d.). *The science of optimal happiness.* Boston, MA: Blackwell.

Dolan, J., Walshe, K., Alsbury, S., Hokamp, K., O'Keeffe, S., Okafuji, T., … Mitchell, K. J. (2007). The extracellular Leucine-Rich Repeat superfamily; a comparative survey and analysis of evolutionary relationships and expression patterns. *BMC Genomics*, *8*(1), 320. https://doi.org/10.1186/1471-2164-8-320

Domingue, B. W., Wedow, R., Conley, D., McQueen, M., Hoffmann, T. J., & Boardman, J. D. (2016). Genome-Wide Estimates of Heritability for Social Demographic Outcomes. *Biodemography and Social Biology*, *62*(1), 1–18. https://doi.org/10.1080/19485565.2015.1068106

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348

Dworkin, R. H. (1990). Patterns of sex differences in negative symptoms and social functioning consistent with separate dimensions of schizophrenic psychopathology. *The American Journal of Psychiatry*, *147*(3), 347–349. https://doi.org/10.1176/ajp.147.3.347

Egan, M. F., Kojima, M., Callicott, J. H., Goldberg, T. E., Kolachana, B. S., Bertolino, A., … Weinberger, D. R. (2003). The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell*, *112*(2), 257–269.

Ellis, B. J., Boyce, W. T., Belsky, J., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2011). Differential susceptibility to the environment: An evolutionary–neurodevelopmental theory. *Development and Psychopathology*, *23*(01), 7–28. https://doi.org/10.1017/S0954579410000611

Emamian, E. S., Hall, D., Birnbaum, M. J., Karayiorgou, M., & Gogos, J. A. (2004). Convergent evidence for impaired AKT1-GSK3β signaling in schizophrenia. *Nature Genetics*, *36*(2), 131–137. https://doi.org/10.1038/ng1296

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584.

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*, *31*(9), 1466–1468. https://doi.org/10.1093/bioinformatics/btu848

European Network of National Networks studying Gene-Environment Interactions in Schizophrenia (EU-GEI). (2014). Identifying Gene-Environment Interactions in Schizophrenia: Contemporary Challenges for Integrated, Large-scale Investigations. *Schizophrenia Bulletin*, *40*(4), 729–736. https://doi.org/10.1093/schbul/sbu069

Evans, L., Tahmasbi, R., Vrieze, S., Abecasis, G., Das, S., Bjelland, D., … Keller, M. (2017). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. https://doi.org/10.1101/115527

Fanous, A. H., Zhou, B., Aggen, S. H., Bergen, S. E., Amdur, R. L., Duan, J., … Levinson, D. F. (2012). Genome-Wide Association Study of Clinical Dimensions of Schizophrenia: Polygenic Effect on Disorganized Symptoms. *American Journal of Psychiatry*, *169*(12), 1309–1317. https://doi.org/10.1176/appi.ajp.2012.12020218

Faraone, S. V., Blehar, M., Pepple, J., Moldin, S. O., Norton, J., Nurnberger, J. I., … Tsuang, M. T. (1996). Diagnostic accuracy and confusability analyses: an application to the Diagnostic Interview for Genetic Studies. *Psychological Medicine*, *26*(2), 401–410.

Farrell, C. M., O'Leary, N. A., Harte, R. A., Loveland, J. E., Wilming, L. G., Wallin, C., … Pruitt, K. D. (2014). Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Research*, *42*(D1), D865–D872. https://doi.org/10.1093/nar/gkt1059

Farrell, M. S., Werge, T., Sklar, P., Owen, M. J., Ophoff, R. A., O'Donovan, M. C., … Sullivan, P. F. (2015). Evaluating historical candidate genes for schizophrenia. *Molecular Psychiatry*, *20*(5), 555–562. https://doi.org/10.1038/mp.2015.16

Federoff, M., Price, T. R., Sailer, A., Scholz, S., Hernandez, D., Nicolas, A., … Houlden, H. (2016). Genome-wide estimate of the heritability of Multiple System Atrophy. *Parkinsonism & Related Disorders*, *22*, 35–41. https://doi.org/10.1016/j.parkreldis.2015.11.005

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., … Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, *47*(11), 1228–1235. http://doi.org/10.1038/ng.3404

Firth, J., Stubbs, B., Vancampfort, D., Schuch, F. B., Rosenbaum, S., Ward, P. B., … Yung, A. R. (2017). The Validity and Value of Self-reported Physical Activity and Accelerometry in People With Schizophrenia: A Population-Scale Study of the UK Biobank. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sbx149

Fonseca-Pedrero, E., Cohen, A., Ortuño-Sierra, J., de Álbeniz, A. P., & Muñiz, J. (2017). Dimensional Structure and Measurement Invariance of the Schizotypal Personality Questionnaire – Brief Revised (SPQ-BR) Scores Across American and Spanish Samples. *Journal of Personality Disorders*, *31*(4), 522–541. https://doi.org/10.1521/pedi_2016_30_266

Forrest, M. P., Waite, A. J., Martin-Rendon, E., & Blake, D. J. (2013). Knockdown of

Human *TCF4* Affects Multiple Signaling Pathways Involved in Cell Survival, Epithelial to Mesenchymal Transition and Neuronal Differentiation. *PLoS ONE*, *8*(8), e73169. https://doi.org/10.1371/journal.pone.0073169

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., … O'Donovan, M. C. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature*, *506*(7487), 179–184. https://doi.org/10.1038/nature12929

Ge, D. (2012). *MetaP*. Retrieved from http://people.genome.duke.edu/~dg48/metap.php

Geddes, J. R., & Lawrie, S. M. (1995). Obstetric complications and schizophrenia: a meta-analysis. *The British Journal of Psychiatry: The Journal of Mental Science*, *167*(6), 786–793.

Geoffroy, P. A., Etain, B., & Houenou, J. (2013). Gene X Environment Interactions in Schizophrenia and Bipolar Disorder: Evidence from Neuroimaging. *Frontiers in Psychiatry*, *4*. https://doi.org/10.3389/fpsyt.2013.00136

Gershon, E. S., DeLisi, L. E., Hamovit, J., Nurnberger, J. I., Maxwell, M. E., Schreiber, J., … Guroff, J. J. (1988). A controlled family study of chronic psychoses. Schizophrenia and schizoaffective disorder. *Archives of General Psychiatry*, *45*(4), 328–336.

Gershon, E. S., Hamovit, J., Guroff, J. J., Dibble, E., Leckman, J. F., Sceery, W., … Bunney, W. E. (1982). A family study of schizoaffective, bipolar I, bipolar II, unipolar, and normal control probands. *Archives of General Psychiatry*, *39*(10), 1157–1167.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, *51*(4), 1440. https://doi.org/10.2307/2533274

Giunti, P., Sweeney, M. G., Spadaro, M., Jodice, C., Novelletto, A., Malaspina, P., … Harding, A. E. (1994). The trinucleotide repeat expansion on chromosome 6p (SCA1) in autosomal dominant cerebellar ataxias. *Brain: A Journal of Neurology*, *117 (Pt 4)*, 645–649.

Glatt, S. J., Faraone, S. V., & Tsuang, M. T. (2003). Schizophrenia is not associated with DRD4 48-base-pair-repeat length or individual alleles: results of a meta-analysis. *Biological Psychiatry*, *54*(6), 629–635.

Glatt, S. J., Faraone, S. V., & Tsuang, M. T. (2004). DRD2 ?141C insertion/deletion polymorphism is not associated with schizophrenia: Results of a meta-analysis. *American Journal of Medical Genetics*, *128B*(1), 21–23. https://doi.org/10.1002/ajmg.b.30007

Gluckman, P. D., Hanson, M. A., Beedle, A. S., & Raubenheimer, D. (2008). Fetal and neonatal pathways to obesity. *Frontiers of Hormone Research*, *36*, 61–72. https://doi.org/10.1159/000115337

Goes, F.S., McGrath, J., Avramopoulos D., Wolyniec, P., Pirooznia, M., Ruczinski, I., Nestadt G,… Pulver, A.E. (2015). Genome-Wide Association Study of Schizophrenia in Ashkenazi Jews. *Am J Med Genet Part B*, *168B*, 649–659.

Goldstein, J. M. (1988). Gender differences in the course of schizophrenia. *The American Journal of Psychiatry*, *145*(6), 684–689. https://doi.org/10.1176/ajp.145.6.684

Gottesman, I. I. (1991). *Schizophrenia Genesis: The Origins of Madness.* W.H. Freeman and Co.

Gottesman, I. I., & Shields, J. (1967). A polygenic theory of schizophrenia. *Proceedings of*

the National Academy of Sciences, *58*(1), 199–205. https://doi.org/10.1073/pnas.58.1.199

Green, M. J., Matheson, S. L., Shepherd, A., Weickert, C. S., & Carr, V. J. (2011). Brain-derived neurotrophic factor levels in schizophrenia: a systematic review with meta-analysis. *Molecular Psychiatry*, *16*(9), 960–972. https://doi.org/10.1038/mp.2010.88

Greenbaum, L., Alkelai, A., Rigbi, A., Kohn, Y., & Lerer, B. (2010). Evidence for association of the GLI2 gene with tardive dyskinesia in patients with chronic schizophrenia. *Movement Disorders*, *25*(16), 2809–2817. https://doi.org/10.1002/mds.23377

Gu, L., Long, J., Yan, Y., Chen, Q., Pan, R., Xie, X., … Su, L. (2013). *HTR2A* −1438A/G polymorphism influences the risk of schizophrenia but not bipolar disorder or major depressive disorder: A meta-analysis. *Journal of Neuroscience Research*, *91*(5), 623–633. https://doi.org/10.1002/jnr.23180

Guella, I., Sequeira, A., Rollins, B., Morgan, L., Torri, F., van Erp, T. G. M., … Vawter, M. P. (2013). Analysis of *miR-137* expression and rs1625579 in dorsolateral prefrontal cortex. *Journal of Psychiatric Research*, *47*(9), 1215–1221. https://doi.org/10.1016/j.jpsychires.2013.05.021

Häfner, H., Maurer, K., Löffler, W., Fätkenheuer, B., an der Heiden, W., Riecher-Rössler, A., … Gattaz, W. F. (1994). The epidemiology of early schizophrenia. Influence of age and gender on onset and early course. *The British Journal of Psychiatry. Supplement*, (23), 29–38.

Häfner, H., Riecher, A., Maurer, K., Löffler, W., Munk-Jørgensen, P., & Strömgren, E. (1989). How does gender influence age at first hospitalization for schizophrenia? A transnational case register study. *Psychological Medicine*, *19*(4), 903–918.

Häfner, H., Riecher-Rössler, A., Maurer, K., Fätkenheuer, B., & Löffler, W. (1992). First onset and early symptomatology of schizophrenia. A chapter of epidemiological and neurobiological research into age and sex differences. *European Archives of Psychiatry and Clinical Neuroscience*, *242*(2–3), 109–118.

Hahsler M, P. M. (2017). *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R package version 1.1-1*. Retrieved from https://CRAN.R-project.org/package=dbscan

Hall, J. B., Cooke Bailey, J. N., Hoffman, J. D., Pericak-Vance, M. A., Scott, W. K., Kovach, J. L., … Bush, W. S. (2015). Estimating cumulative pathway effects on risk for age-related macular degeneration using mixed linear models. *BMC Bioinformatics*, *16*(1). https://doi.org/10.1186/s12859-015-0760-4

Hamshere, M. L., Walters, J. T. R., Smith, R., Richards, A. L., Green, E., Grozeva, D., … O'Donovan, M. C. (2013). Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Molecular Psychiatry*, *18*(6), 708–712. https://doi.org/10.1038/mp.2012.67

Harrison, P. J. (2015). Recent genetic findings in schizophrenia and their therapeutic relevance. *Journal of Psychopharmacology*, *29*(2), 85–96. https://doi.org/10.1177/0269881114553647

Hashimoto, R., Ikeda, M., Ohi, K., Yasuda, Y., Yamamori, H., Fukumoto, M., … Takeda, M.

(2013). Genome-Wide Association Study of Cognitive Decline in Schizophrenia. *American Journal of Psychiatry*, *170*(6), 683–684. https://doi.org/10.1176/appi.ajp.2013.12091228

Hashimoto, R., Ikeda, M., Yamashita, F., Ohi, K., Yamamori, H., Yasuda, Y., … Ozaki, N. (2014). Common variants at 1p36 are associated with superior frontal gyrus volume. *Translational Psychiatry*, *4*(10), e472. https://doi.org/10.1038/tp.2014.110

Hashimoto, R., Ohi, K., Yasuda, Y., Fukumoto, M., Yamamori, H., Kamino, K., … Takeda, M. (2013). The *KCNH2* gene is associated with neurocognition and the risk of schizophrenia. *The World Journal of Biological Psychiatry*, *14*(2), 114–120. https://doi.org/10.3109/15622975.2011.604350

Hass, J., Walton, E., Kirsten, H., Liu, J., Priebe, L., Wolf, C., … Ehrlich, S. (2013). A Genome-Wide Association Study Suggests Novel Loci Associated with a Schizophrenia-Related Brain-Based Phenotype. *PLoS ONE*, *8*(6), e64872. https://doi.org/10.1371/journal.pone.0064872

Hatzimanolis, A., Bhatnagar, P., Moes, A., Wang, R., Roussos, P., Bitsios, P., … Avramopoulos, D. (2015). Common genetic variation and schizophrenia polygenic risk influence neurocognitive performance in young adulthood. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *168*(5), 392–401. https://doi.org/10.1002/ajmg.b.32323

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2), 423–447.

Heston, L. L. (1966). Psychiatric Disorders in Foster Home Reared Children of Schizophrenic Mothers. *The British Journal of Psychiatry*, *112*(489), 819–825. https://doi.org/10.1192/bjp.112.489.819

Hill, M. J., Donocik, J. G., Nuamah, R. A., Mein, C. A., Sainz-Fuertes, R., & Bray, N. J. (2014). Transcriptional consequences of schizophrenia candidate *miR-137* manipulation in human neural progenitor cells. *Schizophrenia Research*, *153*(1–3), 225–230. https://doi.org/10.1016/j.schres.2014.01.034

Hillert, J., & Olerup, O. (1993). Multiple sclerosis is associated with genes within or close to the HLA-DR-DQ subregion on a normal DR15,DQ6,Dw2 haplotype. *Neurology*, *43*(1), 163–168.

Hinde, R. A., & Stevenson-Hinde, J. (1990). Attachment: Biological, Cultural and Individual Desiderata. *Human Development*, *33*(1), 62–72. https://doi.org/10.1159/000276503

Ho, B.-C., Milev, P., O'Leary, D. S., Librant, A., Andreasen, N. C., & Wassink, T. H. (2006). Cognitive and Magnetic Resonance Imaging Brain Morphometric Correlates of Brain-Derived Neurotrophic Factor Val66Met Gene Polymorphism in Patients With Schizophrenia and Healthy Volunteers. *Archives of General Psychiatry*, *63*(7), 731. https://doi.org/10.1001/archpsyc.63.7.731

Hoek, H. W., Brown, A. S., & Susser, E. (1998). The Dutch famine and schizophrenia spectrum disorders. *Social Psychiatry and Psychiatric Epidemiology*, *33*(8), 373–379.

Hoek, H. W., Susser, E., Buck, K. A., Lumey, L. H., Lin, S. P., & Gorman, J. M. (1996). Schizoid personality disorder after prenatal exposure to famine. *The American Journal of Psychiatry*, *153*(12), 1637–1639. https://doi.org/10.1176/ajp.153.12.1637

Hollister, J. M., Mednick, S. A., Brennan, P., & Cannon, T. D. (1994). Impaired autonomic

nervous system-habituation in those at genetic risk for schizophrenia. *Archives of General Psychiatry*, *51*(7), 552–558.

Honda, K., Hirayama, K., Kikuchi, I., Nagato, H., Tamai, H., & Sasazuki, T. (1988). HLA and Silicosis in Japan. *New England Journal of Medicine*, *319*(24), 1610–1610. https://doi.org/10.1056/NEJM198812153192418

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3: Genes|Genomes|Genetics*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Huang, J., Perlis, R. H., Lee, P. H., Rush, A. J., Fava, M., Sachs, G. S., … Smoller, J. W. (2010). Cross-Disorder Genomewide Analysis of Schizophrenia, Bipolar Disorder, and Depression. *American Journal of Psychiatry*, *167*(10), 1254–1263. https://doi.org/10.1176/appi.ajp.2010.09091335

Hubbard, L., Tansey, K. E., Rai, D., Jones, P., Ripke, S., Chambert, K. D., … Zammit, S. (2016). Evidence of Common Genetic Overlap Between Schizophrenia and Cognition. *Schizophrenia Bulletin*, *42*(3), 832–842. https://doi.org/10.1093/schbul/sbv168

Hui, C. L.-M., Li, A. W.-Y., Leung, C.-M., Chang, W.-C., Chan, S. K.-W., Lee, E. H.-M., & Chen, E. Y.-H. (2014). Comparing illness presentation, treatment and functioning between patients with adolescent- and adult-onset psychosis. *Psychiatry Research*, *220*(3), 797–802. https://doi.org/10.1016/j.psychres.2014.08.046

Hulshoff Pol, H. E., van Baal, G. C. M., Schnack, H. G., Brans, R. G. H., van der Schot, A. C., Brouwer, R. M., … Kahn, R. S. (2012). Overlapping and segregating structural brain abnormalities in twins with schizophrenia or bipolar disorder. *Archives of General Psychiatry*, *69*(4), 349–359. https://doi.org/10.1001/archgenpsychiatry.2011.1615

Hwu, H. G., Hong, C. J., Lee, Y. L., Lee, P. C., & Lee, S. F. (1998). Dopamine D4 receptor gene polymorphisms and neuroleptic response in schizophrenia. *Biological Psychiatry*, *44*(6), 483–487.

Ikeda, M., Aleksic, B., Kinoshita, Y., Okochi, T., Kawashima, K., Kushima, I., … Iwata, N. (2011). Genome-Wide Association Study of Schizophrenia in a Japanese Population. *Biological Psychiatry*, *69*(5), 472–478. https://doi.org/10.1016/j.biopsych.2010.07.010

Ikeda, M., Okahisa, Y., Aleksic, B., Won, M., Kondo, N., Naruse, N., … Iwata, N. (2013). Evidence for Shared Genetic Risk Between Methamphetamine-Induced Psychosis and Schizophrenia. *Neuropsychopharmacology*, *38*(10), 1864–1870. https://doi.org/10.1038/npp.2013.94

Inayama, Y., Yoneda, H., Sakai, T., Ishida, T., Nonomura, Y., Kono, Y., … Asaba, H. (1996). Positive association between a DNA sequence variant in the serotonin 2A receptor gene and schizophrenia. *American Journal of Medical Genetics*, *67*(1), 103–105. https://doi.org/10.1002/(SICI)1096-8628(19960216)67:1<103::AID-AJMG18>3.0.CO;2-S

Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2. (2012). Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia. *Biological Psychiatry*, *72*(8), 620–628. https://doi.org/10.1016/j.biopsych.2012.05.035

Joober, R., Benkelfat, C., Brisebois, K., Toulouse, A., Lafrenière, R. G., Turecki, G., … Rouleau, G. A. (1999). Lack of association between the hSKCa3 channel gene CAG

polymorphism and schizophrenia. *American Journal of Medical Genetics*, *88*(2), 154–157.

Jorgensen, T. H., Degn, B., Wang, A. G., Vang, M., Gurling, H., Kalsi, G., … Ewald, H. (2002). Linkage disequilibrium and demographic history of the isolated population of the Faroe Islands. *European Journal of Human Genetics*, *10*(6), 381–387. https://doi.org/10.1038/sj.ejhg.5200816

Kalsi, G., Sherrington, R., Mankoo, B. S., Brynjolfsson, J., Sigmundsson, T., Curtis, D., … Gurling, H. M. (1996). Linkage study of the D5 dopamine receptor gene (DRD5) in multiplex Icelandic and English schizophrenia pedigrees. *The American Journal of Psychiatry*, *153*(1), 107–109. https://doi.org/10.1176/ajp.153.1.107

Keller, M. F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simon-Sanchez, J., … for the International Parkinson's Disease Genomics Consortium (IPDGC) and The Wellcome Trust Case Control Consortium 2 (WTCCC2). (2012). Using genome-wide complex trait analysis to quantify "missing heritability" in Parkinson's disease. *Human Molecular Genetics*, *21*(22), 4996–5009. https://doi.org/10.1093/hmg/dds335

Kendler, K. S., & Diehl, S. R. (1993). The genetics of schizophrenia: a current, genetic-epidemiologic perspective. *Schizophrenia Bulletin*, *19*(2), 261–285.

Kendler, K. S., Kalsi, G., Holmans, P. A., Sanders, A. R., Aggen, S. H., Dick, D. M., … Gejman, P. V. (2011). Genomewide Association Analysis of Symptoms of Alcohol Dependence in the Molecular Genetics of Schizophrenia (MGS2) Control Sample: SYMPTOMS OF AD IN THE MGS2 CONTROL SAMPLE. *Alcoholism: Clinical and Experimental Research*, *35*(5), 963–975. https://doi.org/10.1111/j.1530-0277.2010.01427.x

Kennedy, J. L., Giuffra, L. A., Moises, H. W., Cavalli-Sforza, L. L., Pakstis, A. J., Kidd, J. R., … Kidd, K. K. (1988). Evidence against linkage of schizophrenia to markers on chromosome 5 in a northern Swedish pedigree. *Nature*, *336*(6195), 167–170. https://doi.org/10.1038/336167a0

Kenny, E. M., Cormican, P., Furlong, S., Heron, E., Kenny, G., Fahey, C., … Morris, D. W. (2014). Excess of rare novel loss-of-function variants in synaptic genes in schizophrenia and autism spectrum disorders. *Molecular Psychiatry*, *19*(8), 872–879. https://doi.org/10.1038/mp.2013.127

Kety, S. S. (1987). The significance of genetic factors in the etiology of schizophrenia: results from the national study of adoptees in Denmark. *Journal of Psychiatric Research*, *21*(4), 423–429.

Kety, S. S., Wender, P. H., Jacobsen, B., Ingraham, L. J., Jansson, L., Faber, B., & Kinney, D. K. (1994). Mental illness in the biological and adoptive relatives of schizophrenic adoptees. Replication of the Copenhagen Study in the rest of Denmark. *Archives of General Psychiatry*, *51*(6), 442–455.

Kim, A. H., Parker, E. K., Williamson, V., McMichael, G. O., Fanous, A. H., & Vladimirov, V. I. (2012). Experimental validation of candidate schizophrenia gene ZNF804A as target for hsa-*miR-137*. *Schizophrenia Research*, *141*(1), 60–64. https://doi.org/10.1016/j.schres.2012.06.038

Kim, L.H., Park, B.L., Cheong, H.S., Namgoong, S., Kim, J.O., … Woo, S.-I.. (2015). Genome-Wide Association Study With the Risk of Schizophrenia in a

Korean Population. *Am J Med Genet Part B ,171,* 257–265.

Kim, S., & Webster, M. J. (2010). Correlation analysis between genome-wide expression profiles and cytoarchitectural abnormalities in the prefrontal cortex of psychiatric disorders. *Molecular Psychiatry*, *15*(3), 326–336. https://doi.org/10.1038/mp.2008.99

Kimura, H., Wang, C., Ishizuka, K., Xing, J., Takasaki, Y., Kushima, I., … Ozaki, N. (2016). Identification of a rare variant in *CHD8* that contributes to schizophrenia and autism spectrum disorder susceptibility. *Schizophrenia Research*, *178*(1–3), 104–106. https://doi.org/10.1016/j.schres.2016.08.023

Kirkpatrick, R. M., McGue, M., Iacono, W. G., Miller, M. B., & Basu, S. (2014). Results of a "GWAS Plus:" General Cognitive Ability Is Substantially Heritable and Massively Polygenic. *PLoS ONE*, *9*(11), e112390. https://doi.org/10.1371/journal.pone.0112390

Kirov, G., Zaharieva, I., Georgieva, L., Moskvina, V., Nikolov, I., Cichon, S., … O'Donovan, M. C. (2009). A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Molecular Psychiatry*, *14*(8), 796–803. https://doi.org/10.1038/mp.2008.33

Kirov, G., Grozeva, D., Norton, N., Ivanov, D., Mantripragada, K. K., Holmans, P., … O'Donovan. (2009). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Human Molecular Genetics*, *18*(8), 1497–1503. http://doi.org/10.1093/hmg/ddp043

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., … Owen, M. J. (2012). *De novo* CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry*, *17*(2), 142–153. http://doi.org/10.1038/mp.2011.154

Kovács, T., Kelemen, O., & Kéri, S. (2013). Decreased fragile X mental retardation protein (FMRP) is associated with lower IQ and earlier illness onset in patients with schizophrenia. *Psychiatry Research*, *210*(3), 690–693. https://doi.org/10.1016/j.psychres.2012.12.022

Krishna Kumar, S., Feldman, M. W., Rehkopf, D. H., & Tuljapurkar, S. (2016). Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences*, *113*(1), E61–E70. https://doi.org/10.1073/pnas.1520109113

Kwon, E., Wang, W., & Tsai, L.-H. (2013). Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and *TCF4* as *miR-137* targets. *Molecular Psychiatry*, *18*(1), 11–12. https://doi.org/10.1038/mp.2011.170

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838. https://doi.org/10.1038/nature09410

Laurent, C., Niehaus, D., Bauché, S., Levinson, D. F., Soubigou, S., Pimstone, S., … Mallet, J. (2003). CAG repeat polymorphisms in KCNN3 (HSKCa3) and PPP2R2B show no association or linkage to schizophrenia: KCNN3 and PPP2R2B CAG Repeats and Schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *116B*(1), 45–50. https://doi.org/10.1002/ajmg.b.10797

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., … for the GENEVA Investigators. (2010). Quality control and quality assurance in genotypic data

for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602. https://doi.org/10.1002/gepi.20516

LeBlanc, M., Kulle, B., Sundet, K., Agartz, I., Melle, I., Djurovic, S., … Andreassen, O. A. (2012). Genome-wide study identifies PTPRO and WDR72 and FOXQ1-SUMO1P1 interaction associated with neurocognitive function. *Journal of Psychiatric Research*, *46*(2), 271–278. https://doi.org/10.1016/j.jpsychires.2011.11.001

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., … Wray, N. R. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, *45*(9), 984–994. https://doi.org/10.1038/ng.2711

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics (Oxford, England)*, *28*(19), 2540–2542. https://doi.org/10.1093/bioinformatics/bts474

Lencz, T., Guha, S., Liu, C., Rosenfeld, J., Mukherjee, S., DeRosse, P., … Darvasi, A. (2013). Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms3739

Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M., … Malhotra, A. K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences*, *104*(50), 19942–19947. https://doi.org/10.1073/pnas.0710021104

Leung, A., & Chue, P. (2000). Sex differences in schizophrenia, a review of the literature. *Acta Psychiatrica Scandinavica. Supplementum*, *401*, 3–38.

Levine, M. E., Lu, A. T., Bennett, D. A., & Horvath, S. (2015). Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging*, *7*(12), 1198–1211. https://doi.org/10.18632/aging.100864

Levinson, D. F., Mowry, B. J., Sharpe, L., & Endicott, J. (1996). Penetrance of schizophrenia-related disorders in multiplex families after correction for ascertainment. *Genetic Epidemiology*, *13*(1), 11–21. https://doi.org/10.1002/(SICI)1098-2272(1996)13:1<11::AID-GEPI2>3.0.CO;2-9

Levinson, D. F., Shi, J., Wang, K., Oh, S., Riley, B., Pulver, A. E., … Holmans, P. A. (2012). Genome-Wide Association Study of Multiplex Schizophrenia Pedigrees. *American Journal of Psychiatry*, *169*(9), 963–973. https://doi.org/10.1176/appi.ajp.2012.11091423

Lewis, C. M., Levinson, D. F., Wise, L. H., DeLisi, L. E., Straub, R. E., Hovatta, I., … Helgason, T. (2003). Genome Scan Meta-Analysis of Schizophrenia and Bipolar Disorder, Part II: Schizophrenia. *The American Journal of Human Genetics*, *73*(1), 34–48. https://doi.org/10.1086/376549

Leysen, J. E., Janssen, P. M., Gommeren, W., Wynants, J., Pauwels, P. J., & Janssen, P. A. (1992). In vitro and in vivo receptor binding and effects on monoamine turnover in rat brain regions of the novel antipsychotics risperidone and ocaperidone. *Molecular Pharmacology*, *41*(3), 494–508.

Lezak MD. (1995). *Neuropsychological Assessment* (Third Edition). Oxford University Press.

Li, J., & Meltzer, H. Y. (2014). A genetic locus in 7p12.2 associated with treatment resistant schizophrenia. *Schizophrenia Research*, *159*(2–3), 333–339. https://doi.org/10.1016/j.schres.2014.08.018

Li, M., Luo, X.-J., Xiao, X., Shi, L., Liu, X.-Y., Yin, L.-D., … Su, B. (2013). Analysis of common genetic variants identifies *RELN* as a risk gene for schizophrenia in Chinese population. *The World Journal of Biological Psychiatry*, *14*(2), 91–99. https://doi.org/10.3109/15622975.2011.587891

Li, T., Yang, L., Wiese, C., Xu, C. T., Zeng, Z., Giros, B., … Liu, X. (1994). No association between alleles or genotypes at the dopamine transporter gene and schizophrenia. *Psychiatry Research*, *52*(1), 17–23.

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., … Shi, Y. (2017) Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat Genet. 49(11)*, 1576-1583. doi: 10.1038/ng.3973

Lin, M. W., Curtis, D., Williams, N., Arranz, M., Nanko, S., Collier, D., … Gill, M. (1995). Suggestive evidence for linkage of schizophrenia to markers on chromosome 13q14.1-q32. *Psychiatric Genetics*, *5*(3), 117–126.

Lin, M. W., Sham, P., Hwu, H. G., Collier, D., Murray, R., & Powell, J. F. (1997). Suggestive evidence for linkage of schizophrenia to markers on chromosome 13 in Caucasian but not Oriental populations. *Human Genetics*, *99*(3), 417–420.

Lin, Z., Su, Y., Zhang, C., Xing, M., Ding, W., Liao, L., … Cui, D. (2013). The Interaction of BDNF and NTRK2 Gene Increases the Susceptibility of Paranoid Schizophrenia. *PLoS ONE*, *8*(9), e74264. https://doi.org/10.1371/journal.pone.0074264

Lindholm, E., Ekholm, B., Balciuniene, J., Johansson, G., Castensson, A., Koisti, M., … Jazin, E. (1999). Linkage analysis of a large Swedish kindred provides further support for a susceptibility locus for schizophrenia on chromosome 6p23. *American Journal of Medical Genetics*, *88*(4), 369–377.

Lindsay, R. S., Kobes, S., Knowler, W. C., Bennett, P. H., & Hanson, R. L. (2001). Genome-wide linkage analysis assessing parent-of-origin effects in the inheritance of type 2 diabetes and BMI in Pima Indians. *Diabetes*, *50*(12), 2850–2857.

Liou, Y.-J., Wang, H.-H., Lee, M.-T. M., Wang, S.-C., Chiang, H.-L., Chen, C.-C., … Wu, J.-Y. (2012). Genome-Wide Association Study of Treatment Refractory Schizophrenia in Han Chinese. *PLoS ONE*, *7*(3), e33598. https://doi.org/10.1371/journal.pone.0033598

Liu, S., Mellor, D., Ling, M., Saiz, J. L., Vinet, E. V., Xu, X., … Byrne, L. K. (2017). The Schizotypal Personality Questionnaire-Brief lacks measurement invariance across three countries. *Psychiatry Research*, *258*, 544–550. https://doi.org/10.1016/j.psychres.2017.08.088

Llewellyn, C. H., Trzaskowski, M., Plomin, R., & Wardle, J. (2013). Finding the missing heritability in pediatric obesity: the contribution of genome-wide complex trait analysis. *International Journal of Obesity*, *37*(11), 1506–1509. https://doi.org/10.1038/ijo.2013.30

Loddick, S. A., & Rothwell, N. J. (1999). Mechanisms of tumor necrosis factor alpha action on neurodegeneration: interaction with insulin-like growth factor-1. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(17), 9449–9451.

Logue, M. W., Amstadter, A. B., Baker, D. G., Duncan, L., Koenen, K. C., Liberzon, I., …

Uddin, M. (2015). The Psychiatric Genomics Consortium Posttraumatic Stress Disorder Workgroup: Posttraumatic Stress Disorder Enters the Age of Large-Scale Genomic Collaboration. *Neuropsychopharmacology*, *40*(10), 2287–2297. https://doi.org/10.1038/npp.2015.118

Louis-Dit-Picard, H., Barc, J., Trujillano, D., Miserey-Lenkei, S., Bouatia-Naji, N., Pylypenko, O., … Jeunemaitre, X. (2012). KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nature Genetics*, *44*(4), 456–460. https://doi.org/10.1038/ng.2218

Lubke, G. H., Laurin, C., Amin, N., Hottenga, J. J., Willemsen, G., van Grootheest, G., … Boomsma, D. I. (2014). Genome-wide analyses of borderline personality features. *Molecular Psychiatry*, *19*(8), 923–929. https://doi.org/10.1038/mp.2013.109

Ma, X., Deng, W., Liu, X., Li, M., Chen, Z., He, Z., … Li, T. (2011). A genome-wide association study for quantitative traits in schizophrenia in China. *Genes, Brain and Behavior*, *10*(7), 734–739. https://doi.org/10.1111/j.1601-183X.2011.00712.x

Macciardi, F., Kennedy, J. L., Ruocco, L., Giuffra, L., Carrera, P., Marino, C., … Ferrari, M. (1992). A genetic linkage study of schizophrenia to chromosome 5 markers in a northern Italian population. *Biological Psychiatry*, *31*(7), 720–728.

MacDonald, M. E., & Gusella, J. F. (1996). Huntington's disease: translating a CAG repeat into a pathogenic mechanism. *Current Opinion in Neurobiology*, *6*(5), 638–643.

MacGillivray, I., Thompson, B., & Campbell, D. M. (Eds.). (1988). *Twinning and twins*. Chichester ; New York: Wiley.

Maier, W., Lichtermann, D., Minges, J., Hallmayer, J., Heun, R., Benkert, O., & Levinson, D. F. (1993). Continuity and discontinuity of affective disorders and schizophrenia. Results of a controlled family study. *Archives of General Psychiatry*, *50*(11), 871–883.

Main, M. (2000). The organized categories of infant, child, and adult attachment: flexible vs. inflexible attention under attachment-related stress. *Journal of the American Psychoanalytic Association*, *48*(4), 1055-1096; discussion 1175-1187. https://doi.org/10.1177/00030651000480041801

Mak, T. S. H., Kwan, J. S. H., Campbell, D. D., & Sham, P. C. (2016). Local True Discovery Rate Weighted Polygenic Scores Using GWAS Summary Data. *Behavior Genetics*, *46*(4), 573–582. https://doi.org/10.1007/s10519-015-9770-2

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., & Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, *41*(6), 469–480. https://doi.org/10.1002/gepi.22050

Makar, A. B., McMartin, K. E., Palese, M., & Tephly, T. R. (1975). Formate assay in body fluids: application in methanol poisoning. *Biochemical Medicine*, *13*(2), 117–126.

Malaspina, D. (2001). Paternal factors and schizophrenia risk: de novo mutations and imprinting. *Schizophrenia Bulletin*, *27*(3), 379–393.

Malaspina, D., Harlap, S., Fennig, S., Heiman, D., Nahon, D., Feldman, D., & Susser, E. S. (2001). Advancing paternal age and the risk of schizophrenia. *Archives of General Psychiatry*, *58*(4), 361–367.

Mansour, H., Fathi, W., Klei, L., Wood, J., Chowdari, K., Watson, A., … Nimgaonkar, V. L. (2010). Consanguinity and increased risk for schizophrenia in Egypt. *Schizophrenia Research*, *120*(1–3), 108–112. https://doi.org/10.1016/j.schres.2010.03.026

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., … the CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, *49*(1), 27–35. http://doi.org/10.1038/ng.3725

Marioni, R. E., Davies, G., Hayward, C., Liewald, D., Kerr, S. M., Campbell, A., … Deary, I. J. (2014). Molecular genetic contributions to socioeconomic status and intelligence. *Intelligence*, *44*, 26–32. https://doi.org/10.1016/j.intell.2014.02.006

Martinez, M., Goldin, L. R., Cao, Q., Zhang, J., Sanders, A. R., Nancarrow, D. J., … Gejman, P. V. (1999). Follow-up study on a susceptibility locus for schizophrenia on chromosome 6q. *American Journal of Medical Genetics*, *88*(4), 337–343.

Maynard, T. M., Meechan, D. W., Dudevoir, M. L., Gopalakrishna, D., Peters, A. Z., Heindel, C. C., … LaMantia, A.-S. (2008). Mitochondrial localization and function of a subset of 22q11 deletion syndrome candidate genes. *Molecular and Cellular Neuroscience*, *39*(3), 439–451. https://doi.org/10.1016/j.mcn.2008.07.027

Maziade, M., Bissonnette, L., Rouillard, E., Martinez, M., Turgeon, M., Charron, L., … Mérette, C. (1997). 6p24-22 region and major psychoses in the Eastern Quebec population. Le Groupe IREP. *American Journal of Medical Genetics*, *74*(3), 311–318.

Mbarek, H., Milaneschi, Y., Fedko, I. O., Hottenga, J.-J., de Moor, M. H. M., Jansen, R., … Vink, J. M. (2015). The genetics of alcohol dependence: Twin and SNP-based heritability, and genome-wide association study based on AUDIT scores. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *168*(8), 739–748. https://doi.org/10.1002/ajmg.b.32379

McClay, J. L., Adkins, D. E., Åberg, K., Bukszár, J., Khachane, A. N., Keefe, R. S. E., … van den Oord, E. J. C. G. (2011). Genome-Wide Pharmacogenomic Study of Neurocognition As an Indicator of Antipsychotic Treatment Response in Schizophrenia. *Neuropsychopharmacology*, *36*(3), 616–626. https://doi.org/10.1038/npp.2010.193

McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. *Epidemiologic Reviews*, *30*(1), 67–76. https://doi.org/10.1093/epirev/mxn001

McGrath, J., Saha, S., Welham, J., El Saadi, O., MacCauley, C., & Chant, D. (2004). A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Medicine*, *2*(1). https://doi.org/10.1186/1741-7015-2-13

McGrath, L. M., Cornelis, M. C., Lee, P. H., Robinson, E. B., Duncan, L. E., Barnett, J. H., … Smoller, J. W. (2013). Genetic predictors of risk and resilience in psychiatric disorders: A cross-disorder genome-wide association study of functional impairment in major depressive disorder, bipolar disorder, and schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *162*(8), 779–788. https://doi.org/10.1002/ajmg.b.32190

McGrath, L. M., Weill, S., Robinson, E. B., Macrae, R., & Smoller, J. W. (2012). Bringing a developmental perspective to anxiety genetics. *Development and Psychopathology*, *24*(04), 1179–1193. https://doi.org/10.1017/S0954579412000636

McGue, M., Zhang, Y., Miller, M. B., Basu, S., Vrieze, S., Hicks, B., … Iacono, W. G.

(2013). A Genome-Wide Association Study of Behavioral Disinhibition. *Behavior Genetics*, *43*(5), 363–373. https://doi.org/10.1007/s10519-013-9606-x

McGuffin, P., & Sturt, E. (1986). Genetic markers in schizophrenia. *Human Heredity*, *36*(2), 65–88.

McInnis, M. G., Breschel, T. S., Margolis, R. L., Chellis, J., MacKinnon, D. F., McMahon, F. J., … DePaulo, J. R. (1999). Family-based association analysis of the hSKCa3 potassium channel gene in bipolar disorder. *Molecular Psychiatry*, *4*(3), 217–219.

Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs*, *33*(5), i-57. https://doi.org/10.1037/h0093212

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Meyer, U., Schwarz, M. J., & Müller, N. (2011). Inflammatory processes in schizophrenia: A promising neuroimmunological target for the treatment of negative/cognitive symptoms and beyond. *Pharmacology & Therapeutics*, *132*(1), 96–110. https://doi.org/10.1016/j.pharmthera.2011.06.003

Millar, J. K., Wilson-Annan, J. C., Anderson, S., Christie, S., Taylor, M. S., Semple, C. A., … Porteous, D. J. (2000). Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Human Molecular Genetics*, *9*(9), 1415–1423.

Mitchell, K. J. (Ed.). (2015). *The genetics of neurodevelopmental disorders*. Hoboken, New Jersey: Wiley-Blackwell.

Moises, H. W., Yang, L., Kristbjarnarson, H., Wiese, C., Byerley, W., Macciardi, F., … Helgason, T. (1995). An international two–stage genome–wide search for schizophrenia susceptibility genes. *Nature Genetics*, *11*(3), 321–324. https://doi.org/10.1038/ng1195-321

Monroe, S. M., & Simons, A. D. (1991). Diathesis-stress theories in the context of life stress research: implications for the depressive disorders. *Psychological Bulletin*, *110*(3), 406–425.

Morgan, C., & Fisher, H. (2006). Environment and Schizophrenia: Environmental Factors in Schizophrenia: Childhood Trauma--A Critical Review. *Schizophrenia Bulletin*, *33*(1), 3–10. https://doi.org/10.1093/schbul/sbl053

Moriguchi, S., Bies, R. R., Remington, G., Suzuki, T., Mamo, D. C., Watanabe, K., … Uchida, H. (2013). Estimated Dopamine D2 Receptor Occupancy and Remission in Schizophrenia: Analysis of the CATIE Data. *Journal of Clinical Psychopharmacology*, *33*(5), 682–685. https://doi.org/10.1097/JCP.0b013e3182979a0a

Morris, A. G., Gaitonde, E., McKenna, P. J., Mollon, J. D., & Hunt, D. M. (1995). CAG repeat expansions and schizophrenia: association with disease in females and with early age-at-onset. *Human Molecular Genetics*, *4*(10), 1957–1961.

Mosher, L. R., Pollin, W., & Stabenau, J. R. (1971). Identical twins discordant for schizophrenia. Neurologic findings. *Archives of General Psychiatry*, *24*(5), 422–430.

Mowry, B. J., Holmans, P. A., Pulver, A. E., Gejman, P. V., Riley, B., Williams, N. M., … Levinson, D. F. (2004). Multicenter linkage study of schizophrenia loci on chromosome 22q. *Molecular Psychiatry*, *9*(8), 784–795. https://doi.org/10.1038/sj.mp.4001481

Müller, N., & Schwarz, M. J. (2010). Immune System and Schizophrenia., *6*(3), 213–220.

Murphy, K. C., Jones, L. A., & Owen, M. J. (1999). High rates of schizophrenia in adults

with velo-cardio-facial syndrome. *Archives of General Psychiatry*, *56*(10), 940–945.

Nakazawa, M. (2015). *Fmsb: Functions for Medical Statistics Book with some Demographic Data*. Retrieved from https://CRAN.R-project.org/package=fmsb

Need, A. C., Ge, D., Weale, M. E., Maia, J., Feng, S., Heinzen, E. L., … Goldstein, D. B. (2009). A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia. *PLoS Genetics*, *5*(2), e1000373. https://doi.org/10.1371/journal.pgen.1000373

Ng, M., Levinson, D., Faraone, S., Suarez, B., DeLisi, L., Arinami, T., … Lewis, C. (2009). Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Molecular Psychiatry*, *14*(8), 774–785. http://doi.org/10.1038/mp.2008.135

Nicodemus, K. K., Callicott, J. H., Higier, R. G., Luna, A., Nixon, D. C., Lipska, B. K., … Weinberger, D. R. (2010). Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Human Genetics*, *127*(4), 441–452. https://doi.org/10.1007/s00439-009-0782-y

Nicodemus, K. K., Hargreaves, A., Morris, D., Anney, R., Gill, M., Corvin, A., & Donohoe, G. (2014). Variability in Working Memory Performance Explained by Epistasis vs Polygenic Scores in the *ZNF804A* Pathway. *JAMA Psychiatry*, *71*(7), 778. https://doi.org/10.1001/jamapsychiatry.2014.528

Nicodemus, K. K., Law, A. J., Radulescu, E., Luna, A., Kolachana, B., Vakkalanka, R., … Weinberger, D. R. (2010). Biological Validation of Increased Schizophrenia Risk With NRG1, ERBB4, and AKT1 Epistasis via Functional Neuroimaging in Healthy Controls. *Archives of General Psychiatry*, *67*(10), 991. https://doi.org/10.1001/archgenpsychiatry.2010.117

Nunokawa, A., Watanabe, Y., Kaneko, N., Sugai, T., Yazaki, S., Arinami, T., … Someya, T. (2010). The dopamine D3 receptor (DRD3) gene and risk of schizophrenia: Case–control studies and an updated meta-analysis. *Schizophrenia Research*, *116*(1), 61–67. https://doi.org/10.1016/j.schres.2009.10.016

O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., … Owen, M. J. (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics*, *40*(9), 1053–1055. https://doi.org/10.1038/ng.201

O'Donovan, M. C., Guy, C., Craddock, N., Bowen, T., McKeon, P., Macedo, A., … Owen, M. J. (1996). Confirmation of association between expanded CAG/CTG repeats and both schizophrenia and bipolar disorder. *Psychological Medicine*, *26*(06), 1145. https://doi.org/10.1017/S0033291700035868

O'Dushlaine, C., Rossin, L., Lee, P. H., Duncan, L., Parikshak, N. N., Newhouse, S., … Breen, G. (2015). Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, *18*(2), 199–209. https://doi.org/10.1038/nn.3922

Onwuameze, O. E., Nam, K. W., Epping, E. A., Wassink, T. H., Ziebell, S., Andreasen, N. C., & Ho, B.-C. (2013). MAPK14 and CNR1 gene variant interactions: effects on brain volume deficits in schizophrenia patients with marijuana misuse. *Psychological Medicine*, *43*(03), 619–631. https://doi.org/10.1017/S0033291712001559

Ott, J., Macciardi, F., Shen, Y., Carta, M. G., Murru, A., Triunfo, R., … Siniscalco, M.

(2010). Pilot Study on Schizophrenia in Sardinia. *Human Heredity*, *70*(2), 92–96. https://doi.org/10.1159/000313844

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, *16*(5), 275–284. https://doi.org/10.1038/nrg3908

Palmer, R. H. C., Brick, L., Nugent, N. R., Bidwell, L. C., McGeary, J. E., Knopik, V. S., & Keller, M. C. (2015). Examining the role of common genetic variants on alcohol, tobacco, cannabis and illicit drug dependence: genetics of vulnerability to drug dependence: Genetics of vulnerability to drug dependence. *Addiction*, *110*(3), 530–537. https://doi.org/10.1111/add.12815

Pan, W., Kwak, I.-Y., & Wei, P. (2015). A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. *The American Journal of Human Genetics*, *97*(1), 86–98. https://doi.org/10.1016/j.ajhg.2015.05.018

Parnas, J., Cannon, T. D., Jacobsen, B., Schulsinger, H., Schulsinger, F., & Mednick, S. A. (1993). Lifetime DSM-III-R diagnostic outcomes in the offspring of schizophrenic mothers. Results from the Copenhagen High-Risk Study. *Archives of General Psychiatry*, *50*(9), 707–714.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Peterson, R. E., Maes, H. H., Holmans, P., Sanders, A. R., Levinson, D. F., Shi, J., … Webb, B. T. (2011). Genetic risk sum score comprised of common polygenic variation is associated with body mass index. *Human Genetics*, *129*(2), 221–230. https://doi.org/10.1007/s00439-010-0917-1

Phillips, D. I. (1993). Twin studies in medical research: can they tell us whether diseases are genetically determined? *Lancet (London, England)*, *341*(8851), 1008–1009.

Plomin, R. (2008). *Behavioral genetics* (5th ed). New York: Worth Publishers.

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral genetics* (Sixth edition). New York: Worth Publishers.

Plomin, R., Haworth, C. M. A., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2, & Davis, O. S. P. (2013). Common DNA Markers Can Account for More Than Half of the Genetic Influence on Cognitive Abilities. *Psychological Science*, *24*(4), 562–568. https://doi.org/10.1177/0956797612457952

Polymeropoulos, M. H., Coon, H., Byerley, W., Gershon, E. S., Goldin, L., Crow, T. J., … Delisi, L. E. (1994). Search for a schizophrenia susceptibility locus on human chromosome 22. *American Journal of Medical Genetics*, *54*(2), 93–99. https://doi.org/10.1002/ajmg.1320540203

Porteous, D. J., Millar, J. K., Brandon, N. J., & Sawa, A. (2011). DISC1 at 10: connecting psychiatric genetics and neuroscience. *Trends in Molecular Medicine*, *17*(12), 699–706. https://doi.org/10.1016/j.molmed.2011.09.002

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Fallon, J. H., Nguyen, D. D., … Macciardi, F. (2009). A Genome-Wide Association Study of Schizophrenia Using Brain Activation as a Quantitative Phenotype. *Schizophrenia Bulletin*, *35*(1), 96–108. https://doi.org/10.1093/schbul/sbn155

Potvin, S., Stip, E., Sepehry, A. A., Gendron, A., Bah, R., & Kouassi, E. (2008). Inflammatory Cytokine Alterations in Schizophrenia: A Systematic Quantitative

Review. *Biological Psychiatry*, *63*(8), 801–808. https://doi.org/10.1016/j.biopsych.2007.09.024

Pouget, J. G., Gonçalves, V. F., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Spain, S. L., Finucane, H. K., Raychaudhuri, S., … Knight, J. (2016). Genome-Wide Association Studies Suggest Limited Immune Gene Enrichment in Schizophrenia Compared to 5 Autoimmune Diseases. *Schizophrenia Bulletin*, *42*(5), 1176–1184. https://doi.org/10.1093/schbul/sbw059

Power, R. A., Wingenbach, T., Cohen-Woods, S., Uher, R., Ng, M. Y., Butler, A. W., … McGuffin, P. (2013). Estimating the heritability of reporting stressful life events captured by common genetic variants. *Psychological Medicine*, *43*(09), 1965–1971. https://doi.org/10.1017/S0033291712002589

Ptácek, R., Kuzelová, H., & Stefano, G. B. (2011). Dopamine D4 receptor gene DRD4 and its association with psychiatric disorders. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, *17*(9), RA215-220.

Pulver, A. E., Karayiorgou, M., Lasseter, V. K., Wolyniec, P., Kasch, L., Antonarakis, S., … Mallet, J. (1994). Follow-up of a report of a potential linkage for schizophrenia on chromosome 22q12-q13.1: Part 2. *American Journal of Medical Genetics*, *54*(1), 44–50. https://doi.org/10.1002/ajmg.1320540109

Pulver, A. E., Lasseter, V. K., Kasch, L., Wolyniec, P., Nestadt, G., Blouin, J.-L., … Kazazian, H. H. (1995). Schizophrenia: A genome scan targets chromosomes 3p and 8p as potential sites of susceptibility genes. *American Journal of Medical Genetics*, *60*(3), 252–260. https://doi.org/10.1002/ajmg.1320600316

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., … Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, *506*(7487), 185–190. https://doi.org/10.1038/nature12975

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. https://doi.org/10.1038/nature08185

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rainsford, K. D. (1975). The biochemical pathology of aspirin-induced gastric damage. *Agents and Actions*, *5*(4), 326–344.

Raven JC. (1965). *Guide to using the Mill Hill Vocabulary Scale with the Progressive Matrices Scales.*

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. http://doi.org/10.1038/nature05329

Richards, A. L., Leonenko, G., Walters, J. T., Kavanagh, D. H., Rees, E. G., Evans, A., … O'Donovan, M. C. (2016). Exome arrays capture polygenic rare variant contributions to schizophrenia. *Human Molecular Genetics*, *25*(5), 1001–1007. https://doi.org/10.1093/hmg/ddv620

Rietschel, M., Mattheisen, M., Degenhardt, F., Kahn, R. S., Linszen, D. H., Os, J. van, …

Cichon, S. (2012). Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe. *Molecular Psychiatry*, *17*(9), 906–917. https://doi.org/10.1038/mp.2011.80

Rietveld, C. A., Cesarini, D., Benjamin, D. J., Koellinger, P. D., De Neve, J.-E., Tiemeier, H., … Bartels, M. (2013). Molecular genetics and subjective well-being. *Proceedings of the National Academy of Sciences*, *110*(24), 9692–9697. https://doi.org/10.1073/pnas.1222171110

Riley, B. (2004). Linkage studies of schizophrenia. *Neurotoxicity Research*, *6*(1), 17–34.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. A., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. https://doi.org/10.1038/nature13595

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–1159. https://doi.org/10.1038/ng.2742

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., … Gejman, P. V. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, *43*(10), 969–976. doi:10.1038/ng.940

Rosa, A., Cuesta, M. J., Fatjó-Vilas, M., Peralta, V., Zarzuela, A., & Fañanás, L. (2006). The Val66Met polymorphism of the brain-derived neurotrophic factor gene is associated with risk for psychosis: Evidence from a family-based association study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *141B*(2), 135–138. https://doi.org/10.1002/ajmg.b.30266

Ruderfer, D. M., Fanous, A. H., Ripke, S., McQuillin, A., Amdur, R. L., Gejman, P. V., … Kendler, K. S. (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry*, *19*(9), 1017–1024. https://doi.org/10.1038/mp.2013.138

Sacchetti, E., Bocchio-Chiavetto, L., Valsecchi, P., Scassellati, C., Pasqualetti, P., Bonvicini, C., … Gennarelli, M. (2007). -G308A tumor necrosis factor alpha functional polymorphism and schizophrenia risk: Meta-analysis plus association study. *Brain, Behavior, and Immunity*, *21*(4), 450–457. https://doi.org/10.1016/j.bbi.2006.11.009

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A Systematic Review of the Prevalence of Schizophrenia. *PLoS Medicine*, *2*(5), e141. https://doi.org/10.1371/journal.pmed.0020141

Santos-Cortez, R. L. P., Lee, K., Azeem, Z., Antonellis, P. J., Pollock, L. M., Khan, S., … Leal. (2013). Mutations in KARS, Encoding Lysyl-tRNA Synthetase, Cause Autosomal-Recessive Nonsyndromic Hearing Impairment DFNB89. *The American Journal of Human Genetics*, *93*(1), 132–140. https://doi.org/10.1016/j.ajhg.2013.05.018

Sasaki, T., Dai, X. Y., Kuwata, S., Fukuda, R., Kunugi, H., Hattori, M., & Nanko, S. (1997). Brain-derived neurotrophic factor gene and schizophrenia in Japanese subjects. *American Journal of Medical Genetics*, *74*(4), 443–444.

Schmoldt, A., Benthe, H. F., & Haberland, G. (1975). Digitoxin metabolism by rat liver microsomes. *Biochemical Pharmacology*, *24*(17), 1639–1641.

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., … Dale, A. M. (2013). All SNPs Are Not Created Equal: Genome-Wide Association

Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLoS Genetics*, *9*(4), e1003449. https://doi.org/10.1371/journal.pgen.1003449

Schwab, S. G., Eckstein, G. N., Hallmayer, J., Lerer, B., Albus, M., Borrmann, M., … Wildenauer, D. B. (1997). Evidence suggestive of a locus on chromosome 5q31 contributing to susceptibility for schizophrenia in German and Israeli families by multipoint affected sib-pair linkage analysis. *Molecular Psychiatry*, *2*(2), 156–160.

Sebat, J., Levy, D. L., & McCarthy, S. E. (2009). Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends in Genetics*, *25*(12), 528–535. http://doi.org/10.1016/j.tig.2009.10.004

Segurado, R., Detera-Wadleigh, S. D., Levinson, D. F., Lewis, C. M., Gill, M., Nurnberger, J. I., … Akarsu, N. (2003). Genome Scan Meta-Analysis of Schizophrenia and Bipolar Disorder, Part III: Bipolar Disorder. *The American Journal of Human Genetics*, *73*(1), 49–62. https://doi.org/10.1086/376547

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., … McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183. http://doi.org/10.1038/nature16549

Shaw, S. H., Kelly, M., Smith, A. B., Shields, G., Hopkins, P. J., Loftus, J., … DeLisi, L. E. (1998). A genome-wide search for schizophrenia susceptibility genes. *American Journal of Medical Genetics*, *81*(5), 364–376.

Sherrington, R., Brynjolfsson, J., Petursson, H., Potter, M., Dudleston, K., Barraclough, B., … Gurling, H. (1988). Localization of a susceptibility locus for schizophrenia on chromosome 5. *Nature*, *336*(6195), 164–167. https://doi.org/10.1038/336164a0

Shi, H., Medway, C., Brown, K., Kalsheker, N., & Morgan, K. (2011). Using Fisher's method with PLINK "LD clumped" output to compare SNP effects across Genome-wide Association Study (GWAS) datasets. *International Journal of Molecular Epidemiology and Genetics*, *2*(1), 30–35.

Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., … Gejman, P. V. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*. https://doi.org/10.1038/nature08192

Shi, Y., Li, Z., Xu, Q., Wang, T., Li, T., Shen, J., … He, L. (2011). Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nature Genetics*, *43*(12), 1224–1227. https://doi.org/10.1038/ng.980

Shifman, S., Johannesson, M., Bronstein, M., Chen, S. X., Collier, D. A., Craddock, N. J., … Darvasi, A. (2008). Genome-Wide Association Identifies a Common Variant in the Reelin Gene That Increases the Risk of Schizophrenia Only in Women. *PLoS Genetics*, *4*(2), e28. https://doi.org/10.1371/journal.pgen.0040028

Simoudis, E., Han, J., & Fayyad, U. M. (Eds.). (1996). *KDD-96: proceedings*. Menlo Park, Calif: AAAI Press.

Sleiman, P., Wang, D., Glessner, J., Hadley, D., Gur, R. E., Cohen, N., … Janssen-CHOP Neuropsychiatric Genomics Working Group. (2013). GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia / Bipolar susceptibility locus. *Scientific Reports*, *3*, 3075. https://doi.org/10.1038/srep03075

Smeland, O. B., Frei, O., Kauppi, K., Hill, W. D., Li, W., Wang, Y., … for the NeuroCHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology)

Cognitive Working Group. (2017). Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function. *JAMA Psychiatry*, *74*(10), 1065. https://doi.org/10.1001/jamapsychiatry.2017.1986

Smith, B. H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S. M., … Morris, A. D. (2013). Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology*, *42*(3), 689–700.

Smith, A. V., Thomas, D. J., Munro, H. M., & Abecasis, G. R. (2005). Sequence features in regions of weak and strong linkage disequilibrium. *Genome Research*, *15*(11), 1519–1534. http://doi.org/10.1101/gr.4421405

Smith, C. (1974). Concordance in twins: methods and interpretation. *American Journal of Human Genetics*, *26*(4), 454–466.

Smith, R. J., & Bryant, R. G. (1975). Metal substitutions incarbonic anhydrase: a halide ion probe study. *Biochemical and Biophysical Research Communications*, *66*(4), 1281–1286.

Sommer, S. S., Lind, T. J., Heston, L. L., & Sobell, J. L. (1993). Dopamine D4 receptor variants in unrelated schizophrenic cases and controls. *American Journal of Medical Genetics*, *48*(2), 90–93. https://doi.org/10.1002/ajmg.1320480207

Speed, D., & Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research*, *24*(9), 1550–1557. https://doi.org/10.1101/gr.169375.113

St Clair, D., Blackwood, D., Muir, W., Baillie, D., Hubbard, A., Wright, A., & Evans, H. J. (1989). No linkage of chromosome 5q11-q13 markers to schizophrenia in Scottish families. *Nature*, *339*(6222), 305–309. https://doi.org/10.1038/339305a0

St Pourcain, B., Skuse, D. H., Mandy, W. P., Wang, K., Hakonarson, H., Timpson, N. J., … Smith, G. D. (2014). Variability in the common genetic architecture of social-communication spectrum phenotypes during childhood and adolescence. *Molecular Autism*, *5*(1), 18. https://doi.org/10.1186/2040-2392-5-18

Stamm, O., Latscha, U., Janecek, P., & Campana, A. (1976). Development of a special electrode for continuous subcutaneous pH measurement in the infant scalp. *American Journal of Obstetrics and Gynecology*, *124*(2), 193–195.

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., … Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*(7210), 232–236. http://doi.org/10.1038/nature07229

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., … Myin-Germeys, I. (2009). Common variants conferring risk of schizophrenia. *Nature*. https://doi.org/10.1038/nature08186

Stein, Z., & Susser, M. (1975). Fertility, fecundity, famine: food rations in the dutch famine 1944/5 have a causal relation to fertility, and probably to fecundity. *Human Biology*, *47*(1), 131–154.

Steinberg, J., & Webber, C. (2013). The Roles of FMRP-Regulated Genes in Autism Spectrum Disorder: Single- and Multiple-Hit Genetic Etiologies. *The American Journal of Human Genetics*, *93*(5), 825–839. https://doi.org/10.1016/j.ajhg.2013.09.013

Stöber, G., Meyer, J., Nanda, I., Wienker, T. F., Saar, K., Jatzke, S., … Beckmann, H. (2000). hKCNN3 which maps to chromosome 1q21 is not the causative gene in periodic catatonia, a familial subtype of schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience*, *250*(4), 163–168. https://doi.org/10.1007/s004060070020

Stouffer, S. A., Suchman, E. A., DeVinney, L., Star, S. A., & Williams, Jr., R. M. (1949). *The American Soldier: Adjustment During Army Life*. United States: Military Affairs/Aerospace Historian.

Straub, R. E., MacLean, C. J., O'Neill, F. A., Burke, J., Murphy, B., Duke, F., … Kendler, K. S. (1995). A potential vulnerability locus for schizophrenia on chromosome 6p24–22: evidence for genetic heterogeneity. *Nature Genetics*, *11*(3), 287–293. https://doi.org/10.1038/ng1195-287

Straub, R. E., MacLean, C. J., O'Neill, F. A., Walsh, D., & Kendler, K. S. (1997). Support for a possible schizophrenia vulnerability locus in region 5q22-31 in Irish families. *Molecular Psychiatry*, *2*(2), 148–155.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., … Talkowski, M. E. (2014). *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proceedings of the National Academy of Sciences*, *111*(42), E4468–E4477. https://doi.org/10.1073/pnas.1405266111

Sullivan, P. F. (2013). Questions about *DISC1* as a Genetic Risk Factor for Schizophrenia. *Molecular Psychiatry*, *18*(10), 1050–1052.

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a Complex Trait: Evidence From a Meta-analysis of Twin Studies. *Archives of General Psychiatry*, *60*(12), 1187. https://doi.org/10.1001/archpsyc.60.12.1187

Sullivan, P. F., Lin, D., Tzeng, J.-Y., van den Oord, E., Perkins, D., Stroup, T. S., … Close, S. L. (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Molecular Psychiatry*, *13*(6), 570–584. https://doi.org/10.1038/mp.2008.25

Susser, E., Neugebauer, R., Hoek, H. W., Brown, A. S., Lin, S., Labovitz, D., & Gorman, J. M. (1996). Schizophrenia after prenatal famine. Further evidence. *Archives of General Psychiatry*, *53*(1), 25–31.

Sweatt, J. D. (2013). Pitt–Hopkins Syndrome: intellectual disability due to loss of *TCF4*-regulated gene transcription. *Experimental & Molecular Medicine*, *45*(5), e21. https://doi.org/10.1038/emm.2013.32

Tan, H.-Y., Nicodemus, K. K., Chen, Q., Li, Z., Brooke, J. K., Honea, R., … Weinberger, D. R. (2008). Genetic variation in AKT1 is linked to dopamine-associated prefrontal cortical structure and function in humans. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI34725

Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., … Carpenter, W. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research*, *150*(1), 3–10. https://doi.org/10.1016/j.schres.2013.05.028

Tandon, R., Nasrallah, H. A., & Keshavan, M. S. (2009). Schizophrenia, "just the facts" 4.

Clinical features and conceptualization. *Schizophrenia Research, 110*(1–3), 1–23. https://doi.org/10.1016/j.schres.2009.03.005

Tansey, K. E., Guipponi, M., Hu, X., Domenici, E., Lewis, G., Malafosse, A., … Uher, R. (2013). Contribution of Common Genetic Variants to Antidepressant Response. *Biological Psychiatry*, *73*(7), 679–682. https://doi.org/10.1016/j.biopsych.2012.10.030

Tarentino, A. L., & Maley, F. (1975). A comparison of the substrate specificities of endo-beta-N-acetylglucosaminidases from Streptomyces griseus and Diplococcus Pneumoniae. *Biochemical and Biophysical Research Communications*, *67*(1), 455–462.

The Gene Ontology Consortium. (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Research*, *43*(D1), D1049–D1056. https://doi.org/10.1093/nar/gku1179

The Genome of the Netherlands Consortium, Minică, C. C., Dolan, C. V., Hottenga, J.-J., Pool, R., Fedko, I. O., … Vink, J. M. (2015). Heritability, SNP- and Gene-Based Analyses of Cannabis Use Initiation and Age at Onset. *Behavior Genetics*, *45*(5), 503–513. https://doi.org/10.1007/s10519-015-9723-9

The International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, *455*(7210), 237–241. http://doi.org/10.1038/nature07239

The R Core Team. (2016). R: The R project for statistical computing. (Version 2015, March 9). Retrieved from https://www.r-project.org/

Thiselton, D. L., Vladimirov, V. I., Kuo, P.-H., McClay, J., Wormley, B., Fanous, A., … Riley, B. P. (2008). AKT1 Is Associated with Schizophrenia Across Multiple Symptom Dimensions in the Irish Study of High Density Schizophrenia Families. *Biological Psychiatry*, *63*(5), 449–457. https://doi.org/10.1016/j.biopsych.2007.06.005

Thorup, A., Petersen, L., Jeppesen, P., Ohlenschlaeger, J., Christensen, T., Krarup, G., … Nordentoft, M. (2007). Gender differences in young adults with first-episode schizophrenia spectrum disorders at baseline in the Danish OPUS study. *The Journal of Nervous and Mental Disease*, *195*(5), 396–405. https://doi.org/10.1097/01.nmd.0000253784.59708.dd

Tienari, P., Wynne, L. C., Sorri, A., Lahti, I., Läksy, K., Moring, J., … Wahlberg, K.-E. (2004). Genotype-environment interaction in schizophrenia-spectrum disorder. Long-term follow-up study of Finnish adoptees. *The British Journal of Psychiatry: The Journal of Mental Science*, *184*, 216–222.

Tomoda, T., Sumitomo, A., Jaaro-Peled, H., & Sawa, A. (2016). Utility and validity of DISC1 mouse models in biological psychiatry. *Neuroscience*, *321*, 99–107. https://doi.org/10.1016/j.neuroscience.2015.12.061

Torrey, E. F., Taylor, E. H., Bracha, H. S., Bowler, A. E., McNeil, T. F., Rawlings, R. R., … Sjostrom, K. (1994). Prenatal origin of schizophrenia in a subgroup of discordant monozygotic twins. *Schizophrenia Bulletin*, *20*(3), 423–432.

Trzaskowski, M., Eley, T. C., Davis, O. S. P., Doherty, S. J., Hanscombe, K. B., Meaburn, E. L., … Plomin, R. (2013). First Genome-Wide Association Study on Anxiety-Related Behaviours in Childhood. *PLoS ONE*, *8*(4), e58676. https://doi.org/10.1371/journal.pone.0058676

Trzaskowski, M., Harlaar, N., Arden, R., Krapohl, E., Rimfeld, K., McMillan, A., … Plomin, R. (2014). Genetic influence on family socioeconomic status and children's

intelligence. *Intelligence*, *42*, 83–88. https://doi.org/10.1016/j.intell.2013.11.002

Tseliou, F., Johnson, S., Major, B., Rahaman, N., Joyce, J., Lawrence, J., … MiData Consortium. (2017). Gender differences in one-year outcomes of first-presentation psychosis patients in inner-city UK Early Intervention Services: 1-year outcomes of EI services by gender. *Early Intervention in Psychiatry*, *11*(3), 215–223. https://doi.org/10.1111/eip.12235

Umićević Mirkov, M., Janss, L., Vermeulen, S. H., van de Laar, M. A. F. J., van Riel, P. L. C. M., Guchelaar, H.-J., … Coenen, M. J. H. (2015). Estimation of heritability of different outcomes for genetic studies of TNFi response in patients with rheumatoid arthritis. *Annals of the Rheumatic Diseases*, *74*(12), 2183–2187. https://doi.org/10.1136/annrheumdis-2014-205541

Ustün, T. B. (1999). The global burden of mental disorders. *American Journal of Public Health*, *89*(9), 1315–1318. https://doi.org/10.2105/AJPH.89.9.1315

Ustün, T. B., Rehm, J., Chatterji, S., Saxena, S., Trotter, R., Room, R., & Bickenbach, J. (1999). Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet (London, England)*, *354*(9173), 111–115.

Van Dongen S. (2000). *A Cluster Algorithm for Graphs*. Amsterdam: National Research Institute for Mathematics and Computer Science.

van Os, J., Kenis, G., & Rutten, B. P. F. (2010). The environment and schizophrenia. *Nature*, *468*(7321), 203–212. https://doi.org/10.1038/nature09563

van Os, J., Rutten, B. P., & Poulton, R. (2008). Gene-Environment Interactions in Schizophrenia: Review of Epidemiological Findings and Future Directions. *Schizophrenia Bulletin*, *34*(6), 1066–1082. https://doi.org/10.1093/schbul/sbn117

van Rossum, J. M. (1966). The significance of dopamine-receptor blockade for the mechanism of action of neuroleptic drugs. *Archives Internationales De Pharmacodynamie Et De Therapie*, *160*(2), 492–494.

Varese, F., Smeets, F., Drukker, M., Lieverse, R., Lataster, T., Viechtbauer, W., … Bentall, R. P. (2012). Childhood Adversities Increase the Risk of Psychosis: A Meta-analysis of Patient-Control, Prospective- and Cross-sectional Cohort Studies. *Schizophrenia Bulletin*, *38*(4), 661–671. https://doi.org/10.1093/schbul/sbs050

Verheij, C., Bakker, C. E., de Graaff, E., Keulemans, J., Willemsen, R., Verkerk, A. J. M. H., … Oostra, B. A. (1993). Characterization and localization of the FMR-1 gene product associated with fragile X syndrome. *Nature*, *363*(6431), 722–724. https://doi.org/10.1038/363722a0

Viding, E., Price, T. S., Jaffee, S. R., Trzaskowski, M., Davis, O. S. P., Meaburn, E. L., … Plomin, R. (2013). Genetics of Callous-Unemotional Behavior in Children. *PLoS ONE*, *8*(7), e65789. https://doi.org/10.1371/journal.pone.0065789

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., … Zheng, W. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Vinkhuyzen, A. A. E., & Wray, N. R. (2015). Novel directions for G × E analysis in psychiatry. *Epidemiology and Psychiatric Sciences*, *24*(01), 12–19.

https://doi.org/10.1017/S2045796014000584

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G.-B., Lee, S. H., Wray, N. R., … Yang, J. (2014). Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genetics*, *10*(4), e1004269. https://doi.org/10.1371/journal.pgen.1004269

Volpi, S., Heaton, C., Mack, K., Hamilton, J. B., Lannan, R., Wolfgang, C. D., … Lavedan, C. (2009). Whole genome association study identifies polymorphisms associated with QT prolongation during iloperidone treatment of schizophrenia. *Molecular Psychiatry*, *14*(11), 1024–1031. https://doi.org/10.1038/mp.2008.52

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M. … Sebat, J.(2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science, 320(*5875*)*, 539-43. doi: 10.1126/science.1155174.

Wang, K.-S., Liu, X., Zhang, Q., Aragam, N., & Pan, Y. (2011). Genome-wide association analysis of age at onset in schizophrenia in a European-American sample. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *156*(6), 671–680. https://doi.org/10.1002/ajmg.b.31209

Wang, K.-S., Liu, X.-F., & Aragam, N. (2010). A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophrenia Research*, *124*(1–3), 192–199. https://doi.org/10.1016/j.schres.2010.09.002

Wang, K.-S., Zhang, Q., Liu, X., Wu, L., & Zeng, M. (2012). PKNOX2 is Associated with Formal Thought Disorder in Schizophrenia: a Meta-Analysis of Two Genome-wide Association Studies. *Journal of Molecular Neuroscience*, *48*(1), 265–272. https://doi.org/10.1007/s12031-012-9787-4

Wang, Q., Xiang, B., Deng, W., Wu, J., Li, M., Ma, X., … Li, T. (2013). Genome-Wide Association Analysis with Gray Matter Volume as a Quantitative Phenotype in First-Episode Treatment-Naïve Patients with Schizophrenia. *PLoS ONE*, *8*(9), e75083. https://doi.org/10.1371/journal.pone.0075083

Wechsler D. (1997). *Wechsler Memory Scale* (Third Edition).

Wiesmann, U. N., DiDonato, S., & Herschkowitz, N. N. (1975). Effect of chloroquine on cultured fibroblasts: release of lysosomal hydrolases and inhibition of their uptake. *Biochemical and Biophysical Research Communications*, *66*(4), 1338–1343.

Wilkinson, B., Grepo, N., Thompson, B. L., Kim, J., Wang, K., Evgrafov, O. V., … Campbell, D. B. (2015). The autism-associated gene chromodomain helicase DNA-binding protein 8 (*CHD8*) regulates noncoding RNAs and autism-related genes. *Translational Psychiatry*, *5*(5), e568. https://doi.org/10.1038/tp.2015.62

Winokur, G., Coryell, W., Keller, M., Endicott, J., & Leon, A. (1995). A family study of manic-depressive (bipolar I) disease. Is it a distinct illness separable from primary unipolar depression? *Archives of General Psychiatry*, *52*(5), 367–373.

Wittekindt, O., Schwab, S. G., Burgert, E., Knapp, M., Albus, M., Lerer, B., … Wildenauer, D. B. (1999). Association between hSKCa3 and schizophrenia not confirmed by transmission disequilibrium test in 193 offspring/parents trios. *Molecular Psychiatry*, *4*(3), 267–270.

Wong, A. H. C., & Josselyn, S. A. (2016). Caution When Diagnosing Your Mouse With Schizophrenia: The Use and Misuse of Model Animals for Understanding Psychiatric

Disorders. *Biological Psychiatry*, *79*(1), 32–38. https://doi.org/10.1016/j.biopsych.2015.04.023

Wong, E. H. M., So, H.-C., Li, M., Wang, Q., Butler, A. W., Paul, B., … Sham, P.-C. (2014). Common Variants on Xq28 Conferring Risk of Schizophrenia in Han Chinese. *Schizophrenia Bulletin*, *40*(4), 777–786. https://doi.org/10.1093/schbul/sbt104

World Health Organisation. (1992). *The ICD-10 classification of mental and behavioural disorders. Clinical descriptions and diagnostic guidelines.* World Health Organisation.

Xu, C., Aragam, N., Li, X., Villla, E. C., Wang, L., Briones, D., … Wang, K. (2013). BCL9 and C9orf5 Are Associated with Negative Symptoms in Schizophrenia: Meta-Analysis of Two Genome-Wide Association Studies. *PLoS ONE*, *8*(1), e51674. https://doi.org/10.1371/journal.pone.0051674

Yamada, K., Iwayama, Y., Hattori, E., Iwamoto, K., Toyota, T., Ohnishi, T., … Yoshikawa, T. (2011). Genome-Wide Association Study of Schizophrenia in Japanese Population. *PLoS ONE*, *6*(6), e20468. https://doi.org/10.1371/journal.pone.0020468

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yao, J., Pan, Y., Ding, M., Pang, H., & Wang, B. (2015). Association between *DRD2* (rs1799732 and rs1801028) and *ANKK1* (rs1800497) polymorphisms and schizophrenia: A meta-analysis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *168*(1), 1–13. https://doi.org/10.1002/ajmg.b.32281

Yolken, R. H., & Torrey, E. F. (1995). Viruses, schizophrenia, and bipolar disorder. *Clinical Microbiology Reviews*, *8*(1), 131–145.

Yoshimura, F., & Suzuki, T. (1975). Calcium-stimulated adenosine triphosphatase in the microsomal fraction of tooth germ from porcine fetus. *Biochimica Et Biophysica Acta*, *410*(1), 167–177.

Yu, H., Yan, H., Li, J., Li, Z., Zhang, X., Ma, Y., … Yue, W. (2017). Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population. *Mol Psychiatry, 22(7),* 954-960. doi: 10.1038/mp.2016.21

Yue, W.-H., Wang, H.-F., Sun, L.-D., Tang, F.-L., Liu, Z.-H., Zhang, H.-X., … Zhang, D. (2011). Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nature Genetics*, *43*(12), 1228–1231. https://doi.org/10.1038/ng.979

Zammit, S. (2003). Paternal age and risk for schizophrenia. *The British Journal of Psychiatry*, *183*(5), 405–408. https://doi.org/10.1192/bjp.183.5.405

Zayats, T., Athanasiu, L., Sonderby, I., Djurovic, S., Westlye, L. T., Tamnes, C. K., … Haavik, J. (2015). Genome-Wide Analysis of Attention Deficit Hyperactivity Disorder in Norway. *PLOS ONE*, *10*(4), e0122501. https://doi.org/10.1371/journal.pone.0122501

Zhang, X. Y., Chen, D. C., Xiu, M. H., Haile, C. N., Luo, X., Xu, K., … Kosten, T. R. (2012). Cognitive and serum BDNF correlates of BDNF Val66Met gene polymorphism in patients with schizophrenia and normal controls. *Human Genetics*, *131*(7), 1187–1195. https://doi.org/10.1007/s00439-012-1150-x

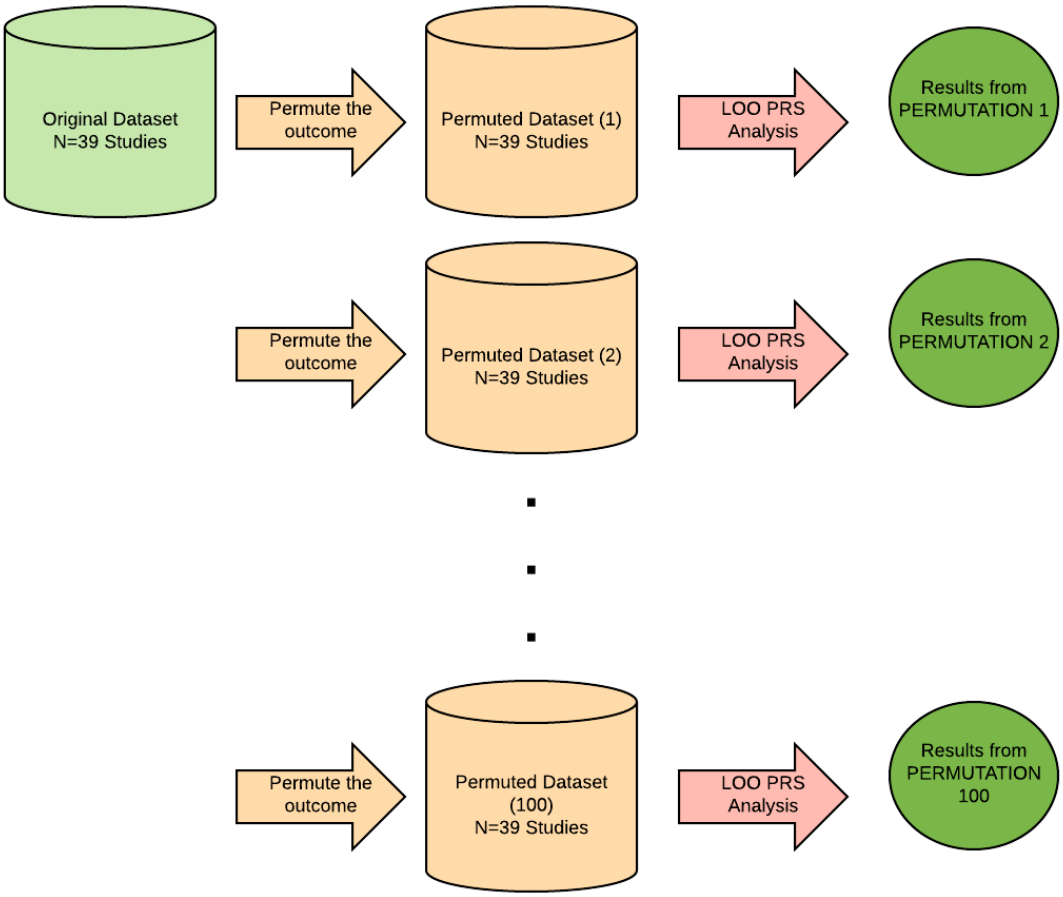
Zuckerman, M. (1999). *Vulnerability to psychopathology: a biosocial model* (1st ed). Washington, DC: American Psychological Association.

Appendices

Appendix 2.1 Individual PGC Study Details

Site	QC score	Array	Cases	Controls	Male
Umeå, Sweden	9	omni	341	577	0.503
Umeå, Sweden	9	omni	193	704	0.475
Norway (TOP)	9	A6.0	377	403	0.533
Edinburgh, UK	8	A6.0	367	284	0.633
Seven countries (PEIC, WTCCC2)	6	I1M	574	1812	0.557
Spain (PEIC, WTCCC2)	6	I1M	150	236	0.585
New York, US & Israel	7	A6.0	325	139	0.614
Ireland	9	A6.0	264	839	0.394
Ireland (WTCCC2)	9	A6.0	1291	1006	0.617
Germany (GRAS)	9	AXI	1067	1169	0.642
Estonia (EGCUT)	2	omni	234	1152	0.268
US, Australia (MGS)	9	A6.0	2638	2482	0.588
London, UK	8	A6.0	509	485	0.572
Sweden (Hubin)	3	omni	265	319	0.618
Bulgaria	8	A6.0	195	608	0.474
Israel	8	I1M	894	1594	0.701
Six countries, WTCCC controls	4	I550	157	245	0.918
New York, US	8	A500	190	190	0.577
Australia	9	I650	456	287	0.601
Cardiff, UK	9	A500	396	284	0.589
UK (CLOZUK)	0	I1M	3426	4085	0.88
UK (CLOZUK)	0	omni	2105	1975	0.629
Netherlands	7	I550	700	607	0.628
Portugal	9	A6.0	346	215	0.521
Boston, US (CIDAR)	9	omni	67	65	0.757
Munich, Germany	8	I317	421	312	0.569
Aberdeen, UK	9	A6.0	719	697	0.693
US (CATIE)	7	A500	397	203	0.767
Sweden	3	A5.0	215	210	0.527
Sweden	3	A6.0	1980	2274	
Sweden	3	omni	1764	2581	0.553
Sweden	3	omni	975	1145	0.543
Cardiff, UK (CogUK)	9	omni	530	678	0.554
NIMH CBDB	5	O25	133	269	0.547
NIMH CBDB	5	I550	497	389	0.627
Denmark	8	I650	471	456	0.583
Bulgaria (trios)	8	A6.0	649	649	0.502
Six countries (trios)	4	I650	516	516	0.556
Bulgaria (trios)	8	omni	70	70	0.595

Appendix 2.2 Graphic representation of the LOO permutation process



Appendix 2.3 Gene ontology Enrichment

TCF4	**P-value**
biological process	4.10E-16
cellular process	7.01E-15
single-organism process	6.32E-10
metabolic process	1.06E-09
cellular component organization or biogenesis	2.93E-09
single-organism cellular process	1.77E-08
organic substance metabolic process	5.30E-08
cellular metabolic process	1.36E-07
cellular component organization	2.72E-07
biological regulation	7.78E-07
FMRP	
nervous system development	3.36E-60
generation of neurons	6.97E-44
neurogenesis	2.79E-42
neuron projection development	2.02E-39
synaptic transmission	2.40E-36
trans-synaptic signalling	2.40E-36
synaptic signalling	2.40E-36
signalling	6.91E-36
cell communication	1.06E-35
single organism signalling	4.63E-35
MIR 137 (downregulated)	
cellular process	7.51E-14
biological process	9.37E-13
metabolic process	1.59E-10
cellular metabolic process	3.94E-10
organic substance metabolic process	7.84E-09
primary metabolic process	1.03E-08
cellular component organization or biogenesis	1.09E-07
cellular component organization	1.30E-07
single-organism process	2.37E-07
cellular protein metabolic process	1.04E-06
MIR 137 (upregulated)	
cellular process	2.23E-13
biological process	5.31E-10
cellular metabolic process	2.33E-08
metabolic process	2.57E-07
organic substance metabolic process	5.43E-06
primary metabolic process	2.22E-05
single-organism process	4.62E-05
cellular macromolecule metabolic process	8.22E-05
cellular component organization or biogenesis	1.55E-04
single-organism cellular process	2.03E-04
CHD8 **(downregulated)**	
biological process	1.16E-08
single-organism developmental process	9.85E-08
nervous system development	1.48E-07

anatomical structure development	1.73E-07
developmental process	2.09E-07
single-multicellular organism process	2.14E-07
multicellular organism development	2.73E-07
system development	1.29E-06
single-organism process	1.98E-06
multicellular organismal process	9.68E-05
***CHD8* (upregulated)**	
cellular process	5.23E-10
biological process	1.53E-09
cellular metabolic process	2.55E-09
cellular component organization or biogenesis	6.37E-07
metabolic process	1.82E-06
primary metabolic process	2.76E-06
cellular macromolecule metabolic process	6.37E-06
single-organism cellular process	9.59E-06
organic substance metabolic process	9.96E-06
sensory perception of chemical stimulus	1.83E-05
Cancer	
positive regulation of macromolecule metabolic process	1.65E-81
regulation of nucleobase-containing compound metabolic process	1.99E-81 4.46E-80
regulation of macromolecule metabolic process	4.57E-80
regulation of nitrogen compound metabolic process	4.64E-80
regulation of nucleic acid-templated transcription	1.33E-79
regulation of biosynthetic process	1.64E-79
regulation of RNA biosynthetic process	2.69E-79
regulation of cellular biosynthetic process	2.70E-79
positive regulation of metabolic process	1.56E-78
regulation of metabolic process	
Cardiac disease	
response to chemical	9.31E-130
response to stress	6.66E-120
response to organic substance	4.90E-117
response to stimulus	1.97E-115
regulation of biological quality	2.25E-111
response to oxygen-containing compound	1.20E-109
response to external stimulus	1.32E-107
regulation of multicellular organismal process	8.07E-105
single-multicellular organism process	7.11E-101
multicellular organismal process	4.23E-95

Appendix 3.1 Script for weight generation taken from Mak et al.

CODE FOR THE MAIN FUNCTIONS USED

```
library(boot)


mixture.halfnormal <- function (pvals , p, var , logp =T) {
z <- qnorm ( pvals /2, lower.tail =F)
sd <- sqrt ( var )
logf <- dnorm (z, sd=sd , log =T) - dnorm (z, log =T)
f <- exp( logf )
# print (f)
f2 <- p*f + (1-p)
logf2 <- log (f2)
if( logp ) return ( logf2 )
else return (exp( logf2 ))
}
ml.mixture.halfnormal <- function (pvals , init =c(0, 0) ,...) {
optimum <- optim ( par =init , fn= function (pars , pvals ) {
p <- inv.logit ( pars [1])
var <- 1 + exp( pars [2])
loglik <- sum( mixture.halfnormal (pvals , p=p, var = var ))
return (- loglik )
},
pvals =pvals ,...)
best.p <- inv.logit ( optimum$par [1])
best.var <- 1 + exp ( optimum$par [2])
return ( list (p= best.p, var= best.var , value =- optimum$value , optim = optimum ))
}
```

CALCULATING DATASETS
###fdr=fdr by kernel density estimation of z-values distribution
###fdr2=fdr by maximum likelihood of z-values distribution

```
for (i in 1:20){

a1=paste("pfile", i, sep="")
b1=paste("orfile",i, sep="")
a<-read.table(a1, head=T)
b<-read.table(b1, head=T)
b<-(b[,-3])
pvals<-a$V2
ml <- ml.mixture.halfnormal (pvals , init = rnorm (2, sd =10))
pi1 <- ml$p
mix.var <- ml$var
fp <- list ()
fp$pvals.order <- order (pvals)
fp$eval.points <- pvals [ fp$pvals.order ]
fp$fp <- mixture.halfnormal (fp$eval.points , p=pi1 , var = mix.var , log=F)
```

```
pi0 <- 1 - pi1
fdr2 <- pi0 / fp$fp


library ( stats )
library ("qvalue")
z <- qnorm ( pvals / 2)
z2 <- sort (c(z, -z))
pi0 <- pi0est (pvals , 0.5)
len <- length (z)
den.obj <- density (x=z2)
den.fun <- splinefun (den.obj$x , den.obj$y )
fp <- den.fun (z2)
zden <- dnorm (z2)
fdr <- zden * pi0$pi0 / fp
fdr <- fdr [1: len ]

tt1<-cbind(b,fdr)
tt2<-cbind(b,fdr2)
e1=paste("orfile1.",i, sep="")
e2=paste("orfile2.", i, sep="")
write.table(tt1,e1 ,quote=F, row.names=F, col.names=F)
write.table(tt2, e2 ,quote=F, row.names=F, col.names=F)
}
```

Appendix 3.2A Table of all median nested R^2 of simulations for all conditions on the N= 500 samples

Methods	n500.chr13.100	n500.chr13.200	n500.chr13.20	n500.chr15.100	n500.chr15.200	n500.chr15.20	n500.chr19.100	n500.chr19.200	n500.chr19.20
PRS Pruning 0.1 - Thresholding 0.01	0.002	0.03	0.05	0	0.004	0	0.037	0	0
PRS Pruning 0.1 - Thresholding 0.1	0.015	0.03	0.04	0	0.015	0.01	0.025	0.005	0.024
PRS Pruning 0.1 - Thresholding 0.2	0.025	0.05	0.02	0.023	0.04	0.007	0.032	0.014	0.01
PRS Pruning 0.1 - Thresholding 0.3	0.008	0.05	0.03	0.016	0.049	0.008	0.034	0.029	0.02
PRS Pruning 0.1 - Thresholding 0.4	0.013	0.05	0.03	0.014	0.047	0.011	0.029	0.018	0.018
PRS Pruning 0.1 - Thresholding 0.5	0.011	0.04	0.03	0.018	0.036	0.004	0.018	0.012	0.022
PRS Pruning 0.1 - Thresholding 0.6	0.011	0.04	0.03	0.016	0.038	0.004	0.016	0.005	0.023
PRS Pruning 0.1 - Thresholding 0.7	0.014	0.04	0.03	0.017	0.035	0.006	0.016	0.007	0.02
PRS Pruning 0.1 - Thresholding 0.8	0.014	0.03	0.03	0.016	0.035	0.005	0.015	0.008	0.02
PRS Pruning 0.1 - Thresholding 0.9	0.015	0.04	0.03	0.016	0.035	0.005	0.015	0.009	0.009
PRS Pruning 0.1 - Thresholding 1	0.015	0.04	0.03	0.015	0.036	0.005	0.015	0.009	0.009
PRS Pruning 0.25 - Thresholding 0.01	0.005	0	0.01	0	0	0	0.04	0	0.02
PRS Pruning 0.25 - Thresholding 0.1	0.015	0.019	0.008	0	0.02	0.026	0.034	0.003	0.042
PRS Pruning 0.25 - Thresholding 0.2	0.019	0.029	0.008	0.001	0.04	0.012	0.042	0.008	0.041
PRS Pruning 0.25 - Thresholding 0.3	0.012	0.028	0.008	0.001	0.038	0.015	0.047	0.018	0.046
PRS Pruning 0.25 - Thresholding 0.4	0.013	0.036	0.01	0.002	0.048	0.015	0.036	0.019	0.054
PRS Pruning 0.25 - Thresholding 0.5	0.013	0.026	0.012	0.004	0.046	0.021	0.032	0.013	0.048
PRS Pruning 0.25 - Thresholding 0.6	0.014	0.025	0.011	0.001	0.039	0.018	0.03	0.009	0.052
PRS Pruning 0.25 - Thresholding 0.7	0.015	0.025	0.012	0.001	0.034	0.016	0.028	0.01	0.054

PRS Pruning 0.25 - Thresholding 0.8	0.015	0.022	0.014	0	0.035	0.017	0.025	0.011	0.052
PRS Pruning 0.25 - Thresholding 0.9	0.015	0.024	0.015	0	0.033	0.018	0.025	0.013	0.052
PRS Pruning 0.25 - Thresholding 1	0.015	0.024	0.016	0	0.033	0.018	0.025	0.013	0.052
PRS Pruning 0.5 - Thresholding 0.01	0	0	0	0	0	0.01	0.02	0	0.02
PRS Pruning 0.5 - Thresholding 0.1	0.004	0.04	0.01	0	0.01	0	0.03	0	0.05
PRS Pruning 0.5 - Thresholding 0.2	0.009	0.05	0.004	0	0.02	0	0.03	0	0.05
PRS Pruning 0.5 - Thresholding 0.3	0.007	0.04	0.005	0	0.03	0	0.03	0	0.05
PRS Pruning 0.5 - Thresholding 0.4	0.014	0.05	0.003	0	0.04	0	0.02	0	0.06
PRS Pruning 0.5 - Thresholding 0.5	0.012	0.04	0.004	0	0.04	0	0.02	0	0.05
PRS Pruning 0.5 - Thresholding 0.6	0.014	0.04	0.002	0	0.04	0	0.02	0	0.05
PRS Pruning 0.5 - Thresholding 0.7	0.013	0.04	0.003	0	0.04	0	0.02	0	0.05
PRS Pruning 0.5 - Thresholding 0.8	0.013	0.03	0.004	0	0.04	0	0.02	0	0.05
PRS Pruning 0.5 - Thresholding 0.9	0.013	0.03	0.004	0	0.04	0	0.02	0	0.05
PRS Pruning 0.5 - Thresholding 1	0.013	0.03	0.004	0	0.04	0	0.02	0	0.05
PRS Clumping - 0.1 Clump	0	0.06	0.001	0.004	0	0.001	0.02	0.002	0.03
PRS Clumping - 0.25 Clump	0	0.05	0.002	0.004	0	0.001	0.02	0.002	0.03
PRS Clumping - 0.5 Clump	0	0.06	0.003	0.004	0	0.001	0.022	0.002	0.03
PRS weighted by kernel density estimation of z-values distribution	0	0	0	0	0	0.012	0	0.012	0.006
PRS weighted by maximum likelihood of z-values distribution	0	0	0	0	0	0.012	0	0.013	0.006

Appendix 3.2B Table of all median nested R^2 of simulations for all conditions on the N= 1000 samples

Methods	n1000.chr13.100	n1000.chr13.200	n1000.chr13.20	n1000.chr15.100	n1000.chr15.200	n1000.chr15.20	n1000.chr19.100	n1000.chr19.200	n1000.chr19.20
PRS Pruning 0.1 - Thresholding 0.01	0.02	0.01	0.14	0	0	0.21	0.05	0	0.14
PRS Pruning 0.1 - Thresholding 0.1	0.02	0.05	0.04	0.03	0.04	0.09	0.05	0.03	0.07
PRS Pruning 0.1 - Thresholding 0.2	0.02	0.07	0.03	0.04	0.06	0.06	0.04	0.03	0.04
PRS Pruning 0.1 - Thresholding 0.3	0.02	0.07	0.02	0.04	0.04	0.06	0.05	0.02	0.04
PRS Pruning 0.1 - Thresholding 0.4	0.02	0.08	0.02	0.04	0.04	0.05	0.05	0.03	0.04
PRS Pruning 0.1 - Thresholding 0.5	0.03	0.08	0.02	0.04	0.04	0.04	0.05	0.04	0.04
PRS Pruning 0.1 - Thresholding 0.6	0.02	0.08	0.02	0.04	0.04	0.04	0.05	0.04	0.04
PRS Pruning 0.1 - Thresholding 0.7	0.02	0.08	0.02	0.04	0.04	0.04	0.05	0.04	0.04
PRS Pruning 0.1 - Thresholding 0.8	0.02	0.07	0.02	0.04	0.04	0.04	0.05	0.05	0.04
PRS Pruning 0.1 - Thresholding 0.9	0.02	0.07	0.02	0.04	0.04	0.04	0.05	0.05	0.04
PRS Pruning 0.1 - Thresholding 1	0.02	0.07	0.02	0.04	0.04	0.04	0.05	0.05	0.04
PRS Pruning 0.25 - Thresholding 0.01	0.02	0.02	0.16	0	0	0.15	0.04	0.01	0.09
PRS Pruning 0.25 - Thresholding 0.1	0.013	0.054	0.059	0.02	0.03	0.043	0.026	0.04	0.06
PRS Pruning 0.25 - Thresholding 0.2	0.018	0.064	0.03	0.025	0.04	0.025	0.033	0.047	0.036
PRS Pruning 0.25 - Thresholding 0.3	0.021	0.065	0.033	0.021	0.03	0.022	0.035	0.042	0.043
PRS Pruning 0.25 - Thresholding 0.4	0.021	0.071	0.03	0.021	0.03	0.02	0.032	0.048	0.034
PRS Pruning 0.25 - Thresholding 0.5	0.022	0.07	0.028	0.022	0.035	0.02	0.03	0.052	0.033
PRS Pruning 0.25 - Thresholding 0.6	0.022	0.068	0.029	0.021	0.035	0.02	0.032	0.06	0.033
PRS Pruning 0.25 - Thresholding 0.7	0.022	0.067	0.029	0.019	0.033	0.02	0.03	0.06	0.034

PRS Pruning 0.25 - Thresholding 0.8	0.02	0.065	0.027	0.02	0.032	0.02	0.03	0.066	0.033
PRS Pruning 0.25 - Thresholding 0.9	0.02	0.066	0.027	0.02	0.031	0.024	0.032	0.067	0.032
PRS Pruning 0.25 - Thresholding 1	0.019	0.066	0.026	0.02	0.031	0.024	0.032	0.067	0.032
PRS Pruning 0.5 - Thresholding 0.01	0.02	0	0.1	0	0.02	0.11	0.03	0.03	0.1
PRS Pruning 0.5 - Thresholding 0.1	0.016	0.05	0.02	0.03	0.04	0.03	0.02	0.05	0.06
PRS Pruning 0.5 - Thresholding 0.2	0.014	0.06	0.02	0.03	0.04	0.02	0.02	0.05	0.04
PRS Pruning 0.5 - Thresholding 0.3	0.02	0.06	0.02	0.03	0.04	0.02	0.02	0.04	0.04
PRS Pruning 0.5 - Thresholding 0.4	0.02	0.07	0.02	0.03	0.04	0.02	0.02	0.05	0.03
PRS Pruning 0.5 - Thresholding 0.5	0.02	0.06	0.02	0.03	0.04	0.01	0.02	0.05	0.03
PRS Pruning 0.5 - Thresholding 0.6	0.02	0.06	0.02	0.03	0.04	0.01	0.02	0.06	0.03
PRS Pruning 0.5 - Thresholding 0.7	0.02	0.06	0.02	0.03	0.04	0.01	0.02	0.06	0.03
PRS Pruning 0.5 - Thresholding 0.8	0.02	0.06	0.02	0.03	0.04	0.01	0.02	0.06	0.03
PRS Pruning 0.5 - Thresholding 0.9	0.02	0.06	0.02	0.03	0.04	0.01	0.02	0.06	0.03
PRS Pruning 0.5 - Thresholding 1	0.02	0.06	0.02	0.03	0.04	0.01	0.02	0.07	0.03
PRS Clumping - 0.1 Clump	0.01	0.03	0.04	0.02	0.02	0.03	0.02	0.04	0.04
PRS Clumping - 0.25 Clump	0.013	0.03	0.04	0.02	0.02	0.03	0.02	0.04	0.04
PRS Clumping - 0.5 Clump	0.013	0.032	0.041	0.017	0.018	0.033	0.019	0.04	0.042
PRS weighted by kernel density estimation of z-values distribution	0.001	0.022	0.016	0	0.012	0	0.006	0.002	0
PRS weighted by maximum likelihood of z-values distribution	0	0.023	0.016	0	0.013	0	0.007	0.002	0

Appendix 3.2C Table of all median nested R^2 of simulations for all conditions on the N= 2500 samples

Methods	n2500.chr13.100	n2500.chr13.200	n2500.chr13.20	n2500.chr15.100	n2500.chr15.200	n2500.chr15.20	n2500.chr19.100	n2500.chr19.200	n2500.chr19.20
PRS Pruning 0.1 - Thresholding 0.01	0.17	0.05	0.26	0.12	0.06	0.24	0.19	0.06	0.35
PRS Pruning 0.1 - Thresholding 0.1	0.12	0.08	0.15	0.14	0.08	0.12	0.17	0.12	0.2
PRS Pruning 0.1 - Thresholding 0.2	0.1	0.08	0.13	0.11	0.09	0.1	0.16	0.14	0.16
PRS Pruning 0.1 - Thresholding 0.3	0.1	0.08	0.1	0.1	0.1	0.08	0.15	0.13	0.15
PRS Pruning 0.1 - Thresholding 0.4	0.1	0.08	0.1	0.09	0.11	0.08	0.14	0.14	0.15
PRS Pruning 0.1 - Thresholding 0.5	0.1	0.09	0.1	0.09	0.11	0.08	0.14	0.15	0.14
PRS Pruning 0.1 - Thresholding 0.6	0.1	0.09	0.1	0.09	0.11	0.07	0.14	0.15	0.14
PRS Pruning 0.1 - Thresholding 0.7	0.1	0.09	0.1	0.09	0.11	0.07	0.14	0.15	0.14
PRS Pruning 0.1 - Thresholding 0.8	0.1	0.09	0.1	0.09	0.11	0.07	0.14	0.14	0.14
PRS Pruning 0.1 - Thresholding 0.9	0.1	0.09	0.1	0.09	0.11	0.07	0.14	0.14	0.14
PRS Pruning 0.1 - Thresholding 1	0.1	0.09	0.09	0.09	0.11	0.07	0.14	0.14	0.14
PRS Pruning 0.25 - Thresholding 0.01	0.14	0.05	0.28	0.14	0.06	0.25	0.16	0.04	0.36
PRS Pruning 0.25 - Thresholding 0.1	0.084	0.064	0.14	0.06	0.064	0.11	0.14	0.08	0.17
PRS Pruning 0.25 - Thresholding 0.2	0.069	0.068	0.12	0.053	0.068	0.08	0.14	0.09	0.14
PRS Pruning 0.25 - Thresholding 0.3	0.067	0.075	0.1	0.047	0.075	0.07	0.13	0.08	0.13
PRS Pruning 0.25 - Thresholding 0.4	0.068	0.073	0.08	0.042	0.08	0.064	0.12	0.08	0.12
PRS Pruning 0.25 - Thresholding 0.5	0.068	0.074	0.084	0.038	0.077	0.063	0.113	0.09	0.11
PRS Pruning 0.25 - Thresholding 0.6	0.067	0.077	0.081	0.038	0.074	0.061	0.113	0.09	0.108
PRS Pruning 0.25 - Thresholding 0.7	0.068	0.078	0.078	0.039	0.07	0.061	0.113	0.09	0.106

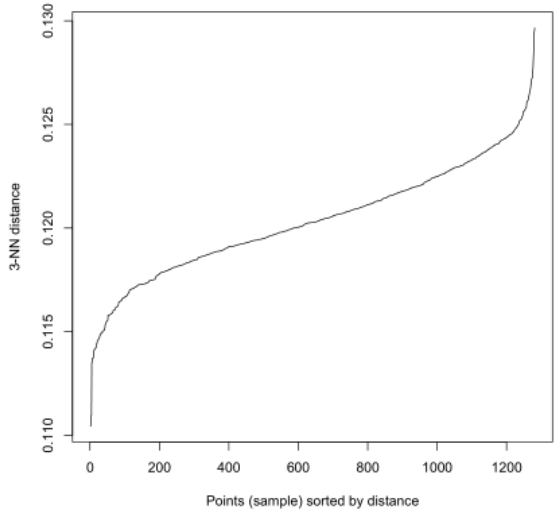
PRS Pruning 0.25 - Thresholding 0.8	0.067	0.078	0.076	0.039	0.07	0.06	0.112	0.086	0.105
PRS Pruning 0.25 - Thresholding 0.9	0.066	0.077	0.076	0.039	0.07	0.06	0.111	0.086	0.103
PRS Pruning 0.25 - Thresholding 1	0.066	0.077	0.076	0.038	0.07	0.06	0.11	0.086	0.103
PRS Pruning 0.5 - Thresholding 0.01	0.13	0.05	0.24	0.06	0.05	0.23	0.15	0.04	0.27
PRS Pruning 0.5 - Thresholding 0.1	0.08	0.05	0.11	0.04	0.06	0.12	0.11	0.07	0.12
PRS Pruning 0.5 - Thresholding 0.2	0.07	0.05	0.1	0.04	0.06	0.1	0.11	0.07	0.1
PRS Pruning 0.5 - Thresholding 0.3	0.06	0.05	0.08	0.03	0.06	0.08	0.1	0.06	0.09
PRS Pruning 0.5 - Thresholding 0.4	0.06	0.05	0.08	0.03	0.06	0.07	0.1	0.07	0.09
PRS Pruning 0.5 - Thresholding 0.5	0.05	0.05	0.07	0.03	0.06	0.07	0.09	0.07	0.08
PRS Pruning 0.5 - Thresholding 0.6	0.05	0.05	0.07	0.03	0.06	0.06	0.09	0.07	0.08
PRS Pruning 0.5 - Thresholding 0.7	0.05	0.05	0.07	0.03	0.06	0.06	0.09	0.07	0.08
PRS Pruning 0.5 - Thresholding 0.8	0.05	0.05	0.06	0.03	0.06	0.06	0.09	0.07	0.08
PRS Pruning 0.5 - Thresholding 0.9	0.05	0.05	0.06	0.03	0.06	0.06	0.09	0.07	0.07
PRS Pruning 0.5 - Thresholding 1	0.05	0.05	0.06	0.03	0.06	0.06	0.09	0.07	0.07
PRS Clumping - 0.1 Clump	0.07	0.04	0.09	0.02	0.05	0.1	0.05	0.02	0.06
PRS Clumping - 0.25 Clump	0.07	0.04	0.09	0.02	0.04	0.1	0.05	0.02	0.06
PRS Clumping - 0.5 Clump	0.07	0.04	0.086	0.015	0.047	0.105	0.05	0.022	0.06
PRS weighted by kernel density estimation of z-values distribution	0.006	0.004	0	0.003	0.003	0	0.015	0.01	0.006
PRS weighted by maximum likelihood of z-values distribution	0.006	0.004	0	0.003	0.003	0	0.015	0.01	0.006

Appendix 3.3 Table of median nested R^2 of simulations in the 50,000 samples

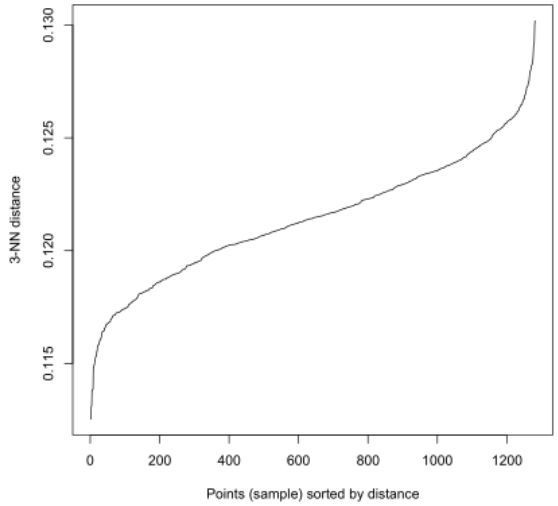
Methods	n50,000.chr19.200NPs
PRS Pruning 0.1 - Thresholding 0.01	0.052390011
PRS Pruning 0.1 - Thresholding 0.1	0.035909842
PRS Pruning 0.1 - Thresholding 0.2	0.031066906
PRS Pruning 0.1 - Thresholding 0.3	0.02855628
PRS Pruning 0.1 - Thresholding 0.4	0.028082561
PRS Pruning 0.1 - Thresholding 0.5	0.025842275
PRS Pruning 0.1 - Thresholding 0.6	0.024296134
PRS Pruning 0.1 - Thresholding 0.7	0.024296134
PRS Pruning 0.1 - Thresholding 0.8	0.023326912
PRS Pruning 0.1 - Thresholding 0.9	0.021465242
PRS Pruning 0.1 - Thresholding 1	0.019438226
PRS Pruning 0.25 - Thresholding 0.01	0.02517382
PRS Pruning 0.25 - Thresholding 0.1	0.01957765
PRS Pruning 0.25 - Thresholding 0.2	0.01752328
PRS Pruning 0.25 - Thresholding 0.3	0.01619387
PRS Pruning 0.25 - Thresholding 0.4	0.01519918
PRS Pruning 0.25 - Thresholding 0.5	0.01434611
PRS Pruning 0.25 - Thresholding 0.6	0.01361167
PRS Pruning 0.25 - Thresholding 0.7	0.01293208
PRS Pruning 0.25 - Thresholding 0.8	0.01232271
PRS Pruning 0.25 - Thresholding 0.9	0.01165571
PRS Pruning 0.25 - Thresholding 1	0.01078922
PRS Clumping - 0.1 Clump	0.0139
PRS Clumping - 0.25 Clump	0.0113
PRS Clumping - 0.5 Clump	0.0113
PRS weighted by kernel density estimation of z-values distribution	0.0043
PRS weighted by maximum likelihood of z-values distribution	0.0045

Appendix 4.1A K-nearest neighbour distance plots for DBSCan ε neighbourhood detection. SNP GRMs K-nearest neighbour distance plots

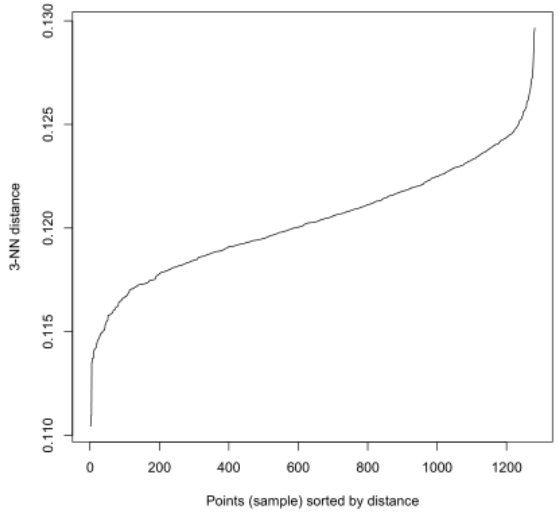
Height (group A > 1.72)



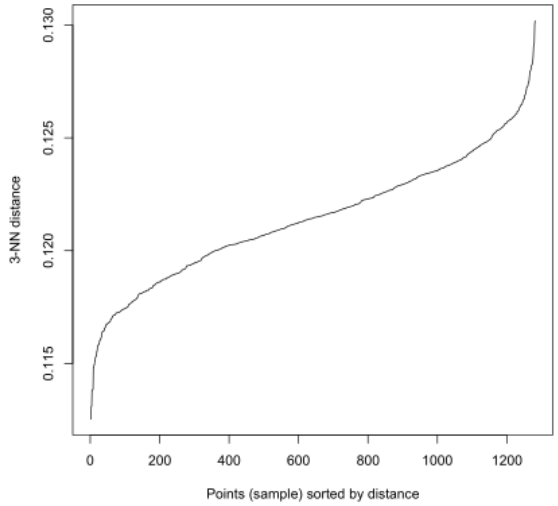
Height (group B < 1.63)



g (group A > 0.46)

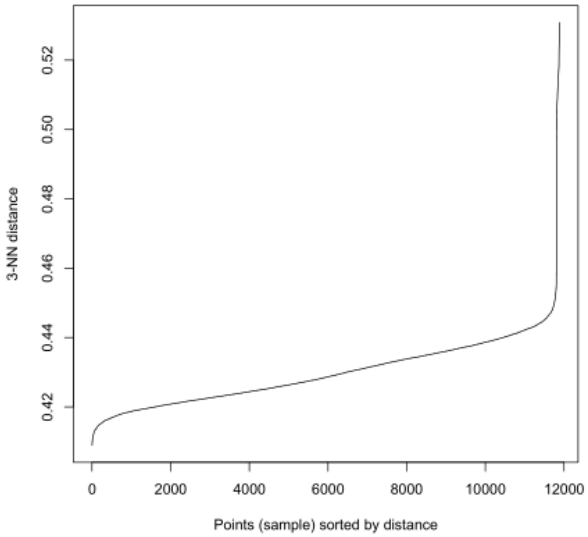


g (group B < -0.35)

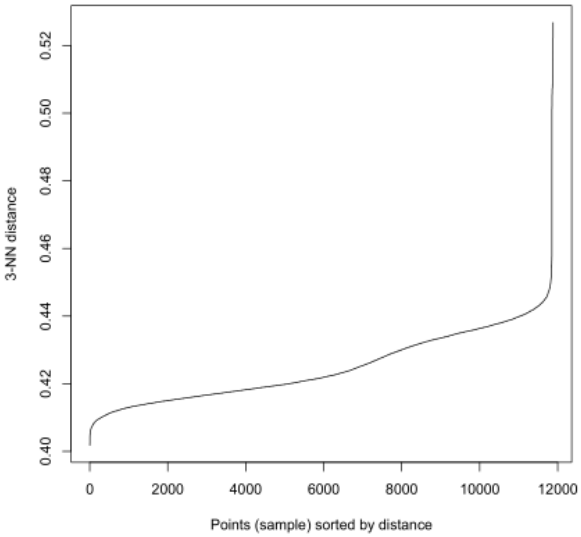


Appendix 4.1B K-nearest neighbour distance plots for DBSCan ε neighbourhood detection. IBD GRMs K-nearest neighbour distance plots

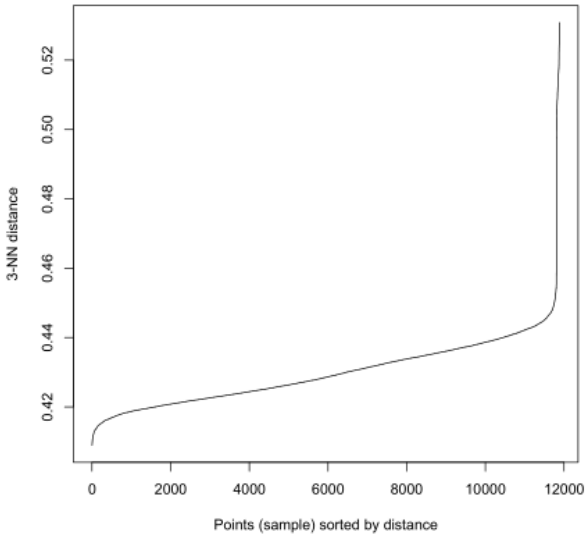
Height (group A > 1.72)



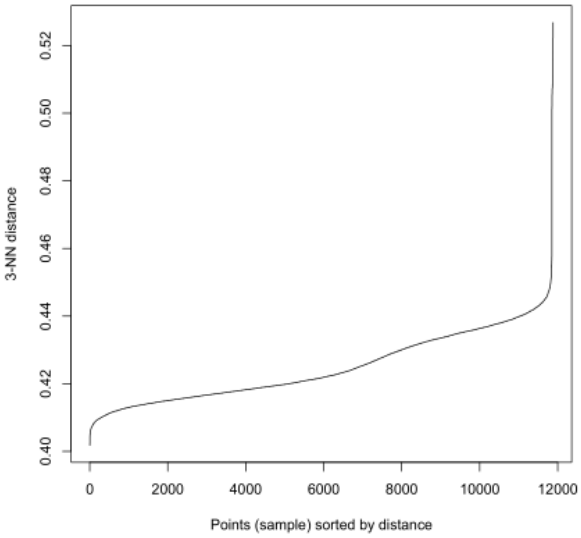
Height (group B < 1.63)



g (group A > 0.46)



g (group B < -0.35)



Appendix 4.2 Script to convert GRMs from GCTA into a meaningful input for DBSCAN.

```
#### Create function to read GCTA GRM in R ####
readGRM <- function(rootname)
{
bin.file.name <- paste(rootname, ".grm.bin", sep="")
n.file.name <- paste(rootname, ".grm.N.bin", sep="")
id.file.name <- paste(rootname, ".grm.id", sep="")
cat("Reading IDs\n")
id <- read.table(id.file.name, colClasses="character")
n <- dim(id)[1]
cat("Reading GRM\n")
bin.file <- file(bin.file.name, "rb")
grm <- readBin(bin.file, n=n*(n+1)/2, what=numeric(0), size=4)
close(bin.file)
cat("Reading N\n")
n.file <- file(n.file.name, "rb")
N <- readBin(n.file, n=n*(n+1)/2, what=numeric(0), size=4)
close(n.file)
cat("Creating data frame\n")
l <- list()
for(i in 1:n)
{
l[[i]] <- 1:i
}
col1 <- rep(1:n, 1:n)
col2 <- unlist(l)
grm <- data.frame(id1=col1, id2=col2, N=N, grm=grm)
ret <- list()
ret$grm <- grm
ret$id <- id
return(ret)
}
##### Merge GRM and phenotype file ####
a1<-readGRM("grm")
a2<-read.table("phenotypefile")
a44<-merge(a1$id, a2, by=c("V1","V2"))
a44$V2.y=NULL
a44$V2.x=NULL
id1<-rep(1:7873)
a5<-cbind(a44,id1)
names(a5) <- sub("^V3$", "pheno1", names(a5))
a6<-a5
names(a6) <- sub("^pheno1$", "pheno2", names(a6))
names(a6) <- sub("^id1$", "id2", names(a6))
a6$V1=NULL
a5$V1=NULL
b1<-merge(a1$grm, a5, by="id1")
b2<-merge(b1, a6, by="id2")
```

```
comparison<-b2$pheno1+b2$pheno2
comparison<-sub("2", "control/control", comparison)
comparison<-sub("4", "case/case", comparison)
c1<-cbind(b2,comparison)
c2<-subset(c1, subset=id1!=id2)
write.table(c2, "grm.txt", col.names=T, row.names=F, quote=F)
                    #### Extract case/case and control/control data ####
d1<-subset(c2, subset=c2$comparison=="case/case")
d2<-subset(c2, subset=c2$comparison=="control/control")
dim(d1)
dim(d2)
write.table(d1, "casescomparison.txt", quote=F, col.names=T, row.names=F)
write.table(d2, "controlscomparison.txt", quote=F, col.names=T, row.names=F)
d11<-cbind (d1$id1, d1$id2)
d111<-cbind (d11, d1$grm)
head(d111)
d21<-cbind (d2$id1, d2$id2)
d211<-cbind (d21, d2$grm)
head(d211)
write.table(d111, "cases", quote=F, col.names=F, row.names=F)
write.table(d211, "controls", quote=F, col.names=F, row.names=F)
q()

                    #### Generating half-matrix input for DBSCAN #####
x<-read.table('cases', head=F)
x.names <- sort(unique(c(x[[1]], x[[2]])))
x.dist <- matrix(0, length(x.names), length(x.names))
dimnames(x.dist) <- list(x.names, x.names)
x.ind <- rbind(cbind(match(x[[1]], x.names), match(x[[2]], x.names)),
cbind(match(x[[2]],
x.names), match(x[[1]], x.names)))
x.dist[x.ind] <- rep(x[[3]], 2)
```