Centre for Neuropsychiatric Genetics and Genomics
Canolfan Geneteg a Genomeg Niwroseiciatrig

Medical Research Council

The Genomic Stratification of Schizophrenia

Isabella Willcocks

Supervisors: Professor James Walters and Dr Antonio Pardiñas

A Thesis Presented for the Degree of Doctor of Philosophy

Centre for Neuropsychiatric Genetics and Genomics
Cardiff University
United Kingdom
December 2022

Summary

Common genetic variation is integral to the genetic architecture of schizophrenia, but until recently, the sample sizes and computational advances needed to detect them did not exist. Through the formation of international consortia such as the PGC, variants associated with schizophrenia can now be robustly detected, however it has come at the cost of phenotypic detail. As a result, genetics specific to more homogenous patient groups, for example individuals with treatment resistance, have remained a challenge to identify. In addition, the heterogenous nature of schizophrenia cohorts, the disorder itself, and the significant overlap with patients of other major psychiatric disorders, has made the identification of 'schizophrenia-unique' variation and neurobiology challenging. To achieve more personalised medicine approaches in schizophrenia, the stratification of individuals into more genetically and phenotypically homogenous groups will be vital.

I first examined the genetic differences between schizophrenia and bipolar disorder, using a recently published research method, the CC-GWAS. I identified 27 loci that were differentially associated with schizophrenia and bipolar disorder, with follow up interrogation of the summary statistics pointing to 26 of them being 'schizophrenia-unique' loci.

Following this I shifted the focus from cross-disorder to within, in an attempt to identify common genetic variation specific to treatment-resistant schizophrenia (TRS). I performed a direct case-case GWAS of ~40,000 individuals with TRS and non-TRS, identifying a genome-wide significant locus that was positively associated with TRS.

Finally, I investigated the relationship between clozapine and neutrophil counts, identifying a significant association between clozapine:norclozapine ratio and absolute neutrophil count, as well as associations with a small set of pharmacogenomic variants associated with the metabolism of clozapine.

This thesis therefore examined multiple ways in which patients with schizophrenia can be stratified into more homogenous subgroups, and the findings could have meaningful implications for research and, with time, patient care.

Personal Acknowledgements

This thesis would not have been possible without the ongoing support of a number of people, who I would like to take this opportunity to thank.

First off, I would like to thank my supervisors, Professor James Walters and Dr Antonio Pardiñas, for all of their patience and guidance throughout my studies. Their support and wealth of knowledge has not only allowed me to complete this thesis, but also to develop my skills as a researcher, something that I will be forever grateful to them both for as I begin my career in academic research.

Next, I would like to thank the wider psychosis department, for never passing up an opportunity to offer me help and guidance when I needed it, and for making me feel welcome in the group.

I would also like to thank my parents, Alex and Alison, and my partner, Luke, for all your ongoing love and support throughout this process and beyond, without which this thesis would not have been written. An extra thank you goes to Luke for all the cups of coffee, half drunk or otherwise, without which this thesis would undoubtedly not have been written either.

Finally, I would like to thank the Medical Research Council for funding my PhD studies, and all the research participants in the cohorts used in this thesis. Without your contributions of data and life experiences, this thesis would not have been possible.

Acknowledgement of Work by Others

Professors James Walters and Michael O'Donovan were key to the collection of phenotypic information used in research chapter 2, through their individual communication with the PIs of the schizophrenia working group of the PGC.

Dr Anders Kämpe and Professor Aarno Palotie generated and provided the FinnGen summary statistics used in research chapter 2.

Dr Kevin O'Connell and Professor Ole Andreassen provided the genotype and phenotype data for the CLOZNOR dataset for use in research chapter 2.

Dr Antonio Pardiñas provided the FINEMAP results for comparison in research chapter 1, and the ancestry inference R scripts that were used in research chapter 2. He also provided a list of TRS individuals for a subset of the datasets in research chapter 2 based on his work on a previous PGC secondary analysis.

Publications Resulting from This Work

Willcocks IR, Legge SE, Nalmpanti M, Mazzeo L, King A, Jansen J, Helthuis M, Owen MJ, O'Donovan MC, Walters JTR, Pardiñas AF (2021) Clozapine Metabolism is Associated With Absolute Neutrophil Count in Individuals With Treatment-Resistant Schizophrenia, *Frontiers in Pharmacology*, DOI:10.3389/fphar.2021.658734

Abbreviations

A list of the abbreviations used throughout this thesis are given below. Due to the sheer number, gene abbreviations will be defined within the body of text only.

Abbreviation	**Definition**
5-HT2A	5-hydroxy-tryptamine subtype 2A
AMD	Age-related macular degeneration
ANC	Absolute Neutrophil Count
AUROC	Area under the receiver operating characteristic
CADD	Combined Annotation Dependent Depletion
DNA	Deoxyribonucleic acid
DSM	Diagnostic and Statistical Manual of Mental Disorders
DZ	Dizygotic
ENIGMA	Enhancing NeuroImaging Genetics through Meta-Analysis
eQTL	Expression quantitative trait loci
FST	Fixation index
GWAS	Genome wide association study
H^2l	SNP-based heritability on the liability scale
HEIDI	Heterogeneity in dependent instruments
HIPO	Heritability informed power optimisation
HRC	Haplotype Reference Consortium
HWE	Hardy Weinberg Equilibrium
K	Lifetime disorder prevalence

Abbreviation	**Definition**
K	Lifetime population prevalence
LAVA	Local Analysis of [co]Variant Annotation
LD	Linkage disequilibrium
LOD	Logarithm of Odds
M	Estimated number of independent causal variants
MAC	Minor Allele Count
MAF	Minor Allele Frequency
MHC	Major histocompatibility complex
MOSTest	Multivariate omnibus statistical test
MRI	Magnetic resonance imaging
MTAG	Multi-trait analysis of GWAS
mtCOJO	Multi-trait conditional and joint analysis
MZ	Monozygotic
Neff	Effective sample size
NIMH	National Institute of Mental Health
NMDA	N-methyl-D-aspartate
Non-TRS	Non treatment resistant schizophrenia
OLS	Ordinary least squares
OR	Odds ratio
PCP	Phencyclidine
PGC	Psychiatric Genetics Consortium
PRS	Polygenic risk score
QC	Quality control

Abbreviation	**Definition**
Rg	Genetic correlation
SCHEMA	Schizophrenia Exome Sequencing Meta Analysis
SEM	Structural equation modelling
SMR	Summary-based mendelian randomisation
SNP	Single nucleotide polymorphism
TOPMED	Trans-Omics for Precision Medicine
TRS	Treatment-resistant schizophrenia
UTR	Untranslated region

Table of Contents

List of Figures

List of Tables

Introduction

The introduction of this thesis will provide an overview of research methods and techniques that have been instrumental in progressing our understanding of neuropsychiatric genetics, each of which will be followed by a summary of current findings in the schizophrenia genetics field. Whilst all the methods outlined in this section continue to be of utility today, it has been ordered chronologically to demonstrate the progression of the field as computational and technological advances have been made. Due to their relevance for the work completed in the thesis, the methods and findings discussed here will focus on common genetic variation, although it is well established that variants with a wide range of population allele frequencies, penetrance and effect sizes play a role in the overall genetic risk of schizophrenia.

Schizophrenia

In the mid-late 19th century, psychosis, manifesting in the form of hallucinations and delusions, was treated as a secondary disorder that could not exist in isolation, but instead had to be preceded by a period of primary psychiatric illness; most commonly 'melancholia' (most closely linked to modern day major depression) or 'mania' (most closely linked to modern day bipolar disorder). However, this view came to be challenged by western psychiatrists, who systematically documented the symptoms of thousands of their patients, ultimately concluding that whilst psychosis could indeed develop in the context of abnormal mood, for at least a subset of patients, this initial disruption was absent [1]. This culminated with the concept of *dementia praecox*, first put forward by Emil Kraepelin in the late 19th century [2], which posited for the first time in European literature that psychotic symptoms could develop in the absence of a disturbance of mood, and also identified the key role of cognitive deficits within the disorder. Kraepelin reviewed and amended this concept multiple times through a series of textbook revisions, and in the early 20th century, it became the foundation of the work of Eugen Bleuler, ultimately culminating in the concept of schizophrenia that we know today [3]. Coined from the combination of Greek words 'schizo' (split), and 'phrene' (mind), the term schizophrenia was meant to describe a state of fragmented and disorganised thinking. Unfortunately, popular culture has more commonly conflated the term with the idea of 'split personality', leading to a high degree of misunderstanding and stigma surrounding the condition, summarised effectively by Robert Kolker in his publication Hidden Valley Road [4]:

> "Bleuler chose this new word because its Latin root—schizo—implied a harsh, drastic splitting of mental functions. This turned out to be a tragically poor choice. Almost ever since, a vast swath of popular culture—from Psycho to Sybil to The Three Faces of Eve—has confused schizophrenia with the idea of split personality. That couldn't be further off the mark. Bleuler was trying to describe a split between a patient's exterior and interior lives—a divide between perception and reality. Schizophrenia is not about multiple personalities. It is about walling oneself off from consciousness, first slowly and then all at once, until you are no longer accessing anything that others accept as real."

Schizophrenia: Overview

Today, schizophrenia is classified as a severe neuropsychiatric disorder of (often) chronic course that causes substantial disruption to the lives of patients. Onset of symptoms typically occurs in early adulthood [5], with manifestation of the characteristic psychotic symptoms often being preceded by a period known as the prodrome [6]. In this stage of illness, a patient may exhibit a wide range of psychological or behavioural symptoms, such as social withdrawal, sleep disruption, memory issues and motivational deficits [7]. A formal diagnosis of schizophrenia is made based on the presence of symptoms that can be categorised into three broad groups, although there are a number of papers utilising factor analytic approaches that have suggested these symptom groups can be broken down into smaller, more specific symptom domains [8-10]. The first group of symptoms are so-called positive symptoms, which represent an addition of thoughts or behaviours not normally seen in healthy individuals. The most common are delusions, where the individual will hold a belief with absolute conviction despite a lack of evidence to support it, and hallucinations, where the patient will report smelling, hearing, seeing etc. sensations or things that are not there, the most common form in schizophrenia being auditory. In addition, the individual may exhibit disorganised thoughts and behaviours, for example positive thought disorders (pressure of speech, incoherence, derailment etc.) and inappropriate affect (the display of emotions that do not match the situation and context). The next group are negative symptoms, which in contrast represent a loss or detraction from behaviour observed in healthy individuals. Another factor analysis study in 2021 posited that these can be split further into two domains: negative symptoms of diminished expressivity, for example alogia (poverty of speech), and affective flattening (detached or reduced reactions to a situation that would normally elicit an emotional response), and negative symptoms of diminished motivation and pleasure, for example apathy and anhedonia (an inability to feel pleasure) [11]. The final group of symptoms, and commonly the most disruptive to normal functioning, are cognitive symptoms, which can include slow thinking, poor concentration and difficulty integrating thoughts, feelings, or behaviours. To make a diagnosis of schizophrenia based upon the guidelines set out by the DSM-V manual [12], a patient must display at least two of five key symptoms (delusions, hallucinations, disorganised speech, disorganised behaviour, and negative symptoms), with at least one of them being one of the first three symptoms listed. These symptoms need to have been present for a clinically significant amount of time, and organic causes of psychosis and psychosis related to substance abuse must be ruled out, for a formal diagnosis of schizophrenia to be made.

Schizophrenia has the potential to be a debilitating condition for patients and can have a substantial impact on the family members of those affected. A 2-year study of employment patterns in 1208 individuals with schizophrenia in the UK, France and Germany observed an employment rate of approximately 21.5%, with only 9.4% able to support themselves through their earnings alone without external help from either family or national benefit systems [13]. A more recent, registry-based study conducted in Norway of 8399 individuals with schizophrenia observed a combined full-time and part-time employment rate of 10.24% for males and 9.8% for females [14]. A systematic review of the international literature focusing primarily on more economically developed countries (MEDCs) reported a higher figure of 12-39%, although this is still significantly lower than the general population [15]. Individuals with schizophrenia have also been demonstrated to be at increased risk of making poor dietary choices, for example through the increased consumption of sugars and processed foods, and are at higher risk of obesity, nutritional deficiencies and metabolic disorders [16]. They are also at increased risk of exercising less, with a recent systematic review of 35 studies representing a total of 3,453 individuals observing that 56.6% of participants achieved the recommended 150 minutes of moderate physical activity per week, a figure that could well be inflated due to over half of the included studies relying on self-report [17]. The lower levels of exercise observed in schizophrenia patients versus controls is not fully understood, but factors could include the sedative effects of some antipsychotic medications, as well as certain key symptoms of schizophrenia itself such as lack of motivation and difficulties organising oneself. The increased prevalence of smoking in individuals with schizophrenia has also been well-documented: a review conducted by the National Institute of Mental Health (NIMH) of smoking prevalence and practices in individuals with psychiatric disorders reported a lifetime smoking prevalence of 69.4% in patients with psychotic disorders, versus 39.1% in individuals with no history of psychiatric disorders [18]. This increase in smoking has been attributed to attempts by patients to self-medicate and find relief for some of their symptoms and medication side effects, although this is by no means the only explanation for the link [19]. All of these factors, as well as others, play a significant role in a stark reduction in life expectancy, with a 2017 systematic review covering 247,603 patients from all continents with the exception of South America reporting a reduction in life expectancy of approximately 15% [20]. A substantial proportion of this excess mortality can be attributed to suicide, with an estimated suicide rate in patients with schizophrenia of 5% [21,22]. It has been reported that up to 40% of schizophrenia patients attempt suicide at least once within their lifetime [23]. A wide array of demographic and psychosocial risk factors have been identified, including longer duration of untreated psychosis, higher levels of premorbid functioning and major life events [24]. The economic cost of schizophrenia, both directly through hospital admissions, community care etc. and indirectly through reduced labour and unpaid care by friends and family etc. is vast, accounting for 3.5% of national healthcare expenditures [25] and with an overall total cost of

£11.8 billion a year in the UK alone [26]. This is despite a low estimated lifetime prevalence of 0.7% in the general population [27].

The pathophysiology of schizophrenia, even now, remains poorly elucidated, and cannot be attributed to major disruption or abnormalites in a single brain structure, as is the case with some neurological disorders such as Parkinson's disease, nor is it a mendelian disease attributable to a single or small number of genes. There are no diagnostic biomarkers for schizophrenia, and diagnosis is based solely on symptoms. Neuroimaging studies have identified a wide range of structural abnormalities in schizophrenia, both in the white matter tracts and grey matter nuclei of the brain, but questions still remain regarding when these abnormalities occur, how they progress or remain static throught the course of the disorder, and what micoranatomical disturbances are causing them [28]. In addition, major brain abnormalities are not common in people with schizophrenia; a study in 2013 of 1379 MRI scans found that only 11.1% of people with schizophrenia had clinically relevant pathology, not significantly different from the percentage in controls, 11.8% [29]. In addition, until recently, there has been significant variability across the results of neuroimaging studies. There are several factors that could explain this, including the limited sample sizes of many studies, the use of different imaging modalities and the amalgamation of samples who have been diagnosed with a range of psychotic / affective diagnoses, not just schizophrenia [30]. In addition, another review of the schizophrenia imaging literature has also pointed to antipsychotic medications, smoking and time in disease course being significant potential confounders [31]. These difficulties have been overcome somewhat by recent, large scale meta-analyses of MRI imaging by collaborations such as the ENIGMA (Enhancing NeuroImaging Genetics through Meta-Analysis) consortium, which has reported associations between schizophrenia and smaller hippocampus, amygdala, thalamas, nucleus accumbens and intracranical volumes, alongside increased pallidum and lateral ventricle volumes [32] Part of the work of this consortium has been to conduct research on individuals who have not received pharmaceutical treatment yet, as well as in individuals who have not yet been diagnosed.

In terms of neurobiological abnormalities in schizophrenia, again, there is limited current understanding, but there are a number of theories at least partially supported by current evidence. One of the first theories to be posited was the dopamine hypothesis. Dopamine is a neurotransmitter of the catecholamine family, alongside adrenaline and noradrenaline. There are four major dopaminergic pathways in the brain: the mesolimbic, the mesocortical, the nigrostriatal and the tuberoinfundibular. Collectively, these pathways have been implicated in a wide range of functions, including executive function, broader cognition, feelings of reward and pleasure, and voluntary motor movements. In schizophrenia, the pathways most implicated are the mesocortical and mesolibmic. Both of these pathways originate in the ventral tegmental area of the midbrain, just medial to the substantia nigra, before projecting to the prefrontal cortex and ventral striatum respectively. The classical dopamine hypothesis of schizophrenia stated that hyperactivity of dopaminergic neurons leads to the development of an array of schizophrenia symptoms , and was conceived based on the observation that the first typical antipsychotics, haloperidol and chlorpromazine, were antagonists of the D2 subtype of dopamine receptors [33]. It is also well-documented that antipsychotic DRD2 occupancy is significantly correlated with the dose required to exert therapeutic effects [34]. Over time, the hypothesis has been amended, with positive symptoms being attributed to hyperactivity of the mesolimbic pathway, and hypoactivity of

D1 receptors in the mesocortical pathway being associated with negative and cognitive symptoms [35]. Other theories have been developed over time in an attempt to better explain the diverse symptom profile of schizophrenia patients. One such theory was the glutamate hypothesis of schizophrenia, based partially on the observation that N-methyl-D-aspartate (NMDA) receptor antagonists, such a phenicyclidine (PCP) and its derivative ketamine, could induce behaviours similar to the positive symptoms of schizophrenia [36]. This led to the theory that NMDA receptor hypoactivity in the prefrontal cortex of the brain led to the manisfestation of psychotic symptoms. A recent review pointed to both of these hypotheses, alongside a third theory of hyperactivity of serotonergic 5-HT2A receptors in the cerebral cortex, playing an interconnected role in the development of positive symptoms in schizophrenia, as well as to a lesser extent negative and cognitive symptoms [37].

The role of genetic risk factors in schizophrenia is, at this point, undeniable. Schizophrenia has been shown to be a highly heritable disorder, with 60-80% of variation in the disorder as a phenotype being accounted for by inherited genetic factors [38-40]. It has also been shown to be highly polygenic, involving genome-wide genetic variation of a wide range of allele frequencies and effect sizes [41]. Evidence for this genetic component has come from an extensive range of literature, spanning from the early 20th century family and pedigree-based studies to the international, collaborative GWAS and whole genome sequencing studies of today. An overview of a subset of these methods, and how they have been applied to schizophrenia to better elucidate the genetics of the disorder, will be given below. Whilst outside the remit of this thesis, it is worth noting that a number of environmental risk factors have also been identified for schizophrenia, for example obstetric complications, urbanicity and patterns of migration [42].

Family, Twin and Adoption Studies

At the start of the 20th century, decades even before the elucidation of the double helix structure of DNA, the genetic component of psychiatric disorders was being demonstrated through a group of three related observational study designs: twin, family, and adoption. Twin studies, the advent of which is often credited to Francis Galton in 1875, assess the rate of concordance amongst monozygotic (MZ) twins, who are 100% genetically identical, and dizygotic (DZ) twins, who are 50% identical, in an attempt to quantify the genetic component of a disorder. Whilst a higher rate of concordance amongst monozygotic twins is suggestive of a genetic element, it is difficult to separate this from the effect of the greater degree of shared environment between MZ twins versus DZ twins. Shared environment is also a limitation present in family-based studies, where the relatives of an affected individual, at varying degrees of relatedness, are assessed for a clinical history of psychiatric disorders. The observation that psychiatric disorders would often be present in multiple family members could once again hint at, but not confirm, the role of genetics in disease risk. In order to pick apart the genetic and environmental influences, adoption studies assess concordance rates in both related individuals living apart in separate home environments, and non-related individuals living in the same environment. The heritability of psychiatric disorders was first demonstrated through these study types, long before the advent of molecular genetic techniques, and by exploiting international collaboration and health registries and electronic records, their utility has only grown with sample size.

Family, Twin and Adoption Studies: schizophrenia

The first study to demonstrate the heritability of schizophrenia was a sibling study conducted in 1916 by Ernst Rüdin, who undertook an investigation of 2,732 siblings of 755 probands diagnosed with schizophrenia [43]. Rüdin observed a familial relationship between schizophrenia and other psychoses, a substantially lower risk in the parents of probands than siblings, and a segregation pattern that did not conform to what would have been expected for a mendelian condition. This would prove to be consistent with other family studies conducted in the first half of the 20th century, such as the study conducted in 1938 by Francis Kallmann, who assessed 1087 patients with schizophrenia and close to 12,500 of their relatives, observing evidence of a family history of schizophrenia in 10% of cases [44]. This study represented a number of methodological improvements over the 1916 sibling study conducted by Rüdin, including the examination of second and third degree relatives instead of focusing only on siblings, as well as dividing the schizophrenia patients into four more homogenous subgroups [45].

The earliest twin studies of schizophrenia [46,47], have been criticised for a number of methodological reasons, for example the high prevalence of premature twins included in the studies [48]. Later twin studies [49,50] demonstrated more robustly the high concordance rates amongst both monozygotic and dizygotic twins and a meta-analysis of twin studies conducted between 1963 and 1987 reported concordance rates of 48% and 17% for MZ and DZ twins respectively [51]. These rates were very similar to another meta-analysis of twin studies conducted between 1992 and 1999, 41%–65% for MZ and 0%–28% for DZ twins [52].

The first adoption study conducted in schizophrenia was led by Leonard Heston in 1966, and examined the offspring of mothers with schizophrenia being raised in the foster home system [53]. This study of 58 children found an age-corrected rate of schizophrenia of just over 16% in the offspring of the mothers with schizophrenia, with no occurrences in the children acting as controls. The observation of higher rates of schizophrenia in the adopted away offspring of schizophrenia patients versus the adopted away offspring of healthy parents has been reproduced in several consequent studies [54-56] etc. Other studies, such as the one conducted by Wender and colleagues in 1976, examined the rate of schizophrenia in children from healthy parents adopted by parents who went on to develop schizophrenia, and found no evidence of increased risk [57].

Through the utilisation of registry data and other forms of information, it is now possible to conduct large-scale family, twin and adoption studies with sample sizes that would have been unattainable at the time the classical studies were undertaken. For example, in a Danish study of just over 31,500 twin pairs, concordance rates for schizophrenia were 33% and 7% in MZ and DZ twins respectively, and heritability was estimated at 79% [58]. However, as integral as these study designs were to the field of schizophrenia genetics throughout the 20th century, and continue to be of importance today, they cannot point to specific genes or genomic loci that are associated with schizophrenia.

Linkage Studies and LOD Scores

Genetic linkage studies again utilise family-based data collected from large, affected pedigrees, but this time to detect chromosomal segments that are transmitted together (i.e., co-segregate) with the disease phenotype of interest. Once a large pedigree has been established and both affected and unaffected individuals from within the family have been genotyped, they can be assessed to see if certain genetic markers occur at a higher rate within affected individuals than unaffected. If this is the case, the chromosomal region where that marker resides could contain a gene linked to the pathology of the disease. The method relies on the process of recombination, whereby portions of DNA are exchanged between chromatids during meiosis. If a genetic marker is in proximity of a disease-associated locus, the rate of recombination between the two will be low, and the likelihood of both being inherited together is increased. In the traditional, parametric approach, the probability that an etiologically important gene is linked to a genetic marker is assessed via a LOD (Logarithm of the Odds) score. The LOD score is a numeric comparison of the likelihood of obtaining the test data to the likelihood of observing the same data purely by chance. By convention, a LOD score of > 3 is considered suggestive evidence of linkage, as this score is analogous to a 1000 to 1 odd that the result did not occur by chance, or a p-value of 0.05 after accounting for the linkage structure of the human genome.

Linkage Studies and LOD Scores: Schizophrenia

With the benefit of hindsight, given its non-mendelian mode of transmission, the linkage study design was not particularly appropriate for use in the investigation of schizophrenia. However, a large number of linkage studies were conducted, and a small number of regions did gain support from multiple sources [59]. These included 22q12-q13 [60,61], 8p22-p21 [62,63] and 6p24-p22 [64,65], amongst others. However, there have also been a number of studies that have failed to replicate these regions, including in the case of the 22q12-q13 region, a study by the same authors who first reported it [66]. A meta-analysis approach was applied in 2009 to a collective sample of 3255 family pedigrees containing 7413 probands, which observed genome-wide evidence for loci on chromosomes 2q and 5q and suggestive evidence for chromosome 8p when restricting to a European only sample [67].

Candidate Gene Studies

Before it became possible through the development of array-based genotyping to examine the entire genome simultaneously, researchers would instead investigate variants within a single or small list of predefined genes for association with the disorder of interest. The selection of these candidate genes was based upon *a priori* knowledge or hypothesis of biological relevance to the disease phenotype, for example their role as therapeutic drug target. Unlike the family-based studies previously discussed, candidate gene studies relied on a case-control design, where an exposure, in this case genetic variants, is investigated in a group of people who have the outcome of interest, as well as a group who do not. Whilst case-control matching, for example in terms of sex, age, ethnicity or other relevant factors, will not necessarily be enough to prevent confounding [68], it is considered standard practise. This is because it increases the precision of estimated effect sizes in the fully controlled statistical model and allows for control of unmeasured confounders that are correlated (in the population) to the measured confounders. In this framework, if a genetic variant is present at higher frequencies in the case group than the control group, the genetic variant could be associated with the outcome under study.

Candidate Gene Studies: Schizophrenia

The candidate gene study was a driving force in the field of neuropsychiatric genetics during the latter half of the 20th century, and schizophrenia was no exception. Indeed, between 1965 and 2006, 1064 schizophrenia candidate gene studies were published [69], the genes with the highest number of studies being *DRD2, COMT, DRD3, HTR2A* and *BDNF* [70]. However, whilst the influence of the candidate gene literature in psychiatry should not be diminished, many of the findings of these early papers have proved unreproducible. A study in 2017 concluded that variants in the 25 most studied schizophrenia candidate genes, including the gene of the primary therapeutic drug target, *DRD2*, were no more associated with schizophrenia than variants in groups of noncandidate genes [71], reaching a similar consensus to an earlier invited review of the same 25 genes [72]. The work of Johnson and colleagues also involved the examination of an extended list of 86 candidate genes, all of which had been studied at least five times and had not originally been implicated by GWAS. There was some evidence to suggest these 86 genes were more associated with schizophrenia than other genes, and several of the most studied candidate genes (*NOTCH4, DRD2, KCNN3, GRM3* and *TNF*) were more strongly associated with schizophrenia that would be expected purely by chance when investigated in isolation. However, with the benefit of hindsight, it has become apparent that many candidate gene studies were simply too underpowered to detect associations of small effect size. Another fundamental issue for these studies was that, at a time when known biological relevance was needed for the selection of potential candidate genes, the biology of psychiatric disorders was largely unknown, an issue compounded by the particularly high degree of heterogeneity observed in schizophrenia. As a result, it was not often possible to select, with any great level of certainty, a functionally relevant gene. Positional candidate gene studies [73], in which the locus specific association testing was supplemented with evidence from linkage studies to better elucidate disorder related regions, did attenuate these limitations somewhat. For example, a series of studies conducted in 2002 identified four candidate schizophrenia risk genes with more compelling evidence: *PRODH* [74], *DTNBP1* [64], *NRG1* [75] and *G72* [76]. However, a convincing argument could still be made that hypothesis–free approaches, looking for disease associations within a greater number of genes, regardless of suspected biological relevance, would be more appropriate.

Genome Wide Association Studies

The potential utility of a study design that would allow testing for associations between disorders and genetic variants across the entire genome simultaneously, without the need to select target genes based on supposed biological relevance, was first discussed in 1996 [77]. At the time of that publication, five years before the first human genome was sequenced [78], it was a theoretical proposal, as the technology required to first identify hundreds of thousands of SNPs for study, and then test for their association with a disorder in thousands of people, had not at that stage been fully realized . However, with the development of SNP genotyping arrays, the first of which was prototyped by Affymetrix in 1998 allowing for the simultaneous genotyping of 500 SNPs [79], and the successful completion of the human genome draft in 2001, the foundations of genome wide association studies, commonly referred to simply as GWAS, were laid.

An important concept to be aware of when interpreting the results of a GWAS is that of linkage disequilibrium (LD). LD refers to the non-random association of alleles in a population due to the lack of chromosomal recombination between nearby haplotypes. It leads to variants in close proximity to each other being inherited together at higher rates than would be expected by chance. LD will therefore lead to many non-causal variants, with no effect on disorder aetiology, being significantly associated with a phenotype because of their elevated levels of correlation with the true causal variant. In this way, any SNP analysed in a GWAS is acting as an index for a region of the genome, and any inferences regarding causality should be made with regards to the region, or locus, and not the SNP specifically. Whilst LD is beneficial in the sense that it means it is not necessary to sequence the whole genome to identify genomic loci associated with the disorder under study, at least in the case of common genetic variation, it also has important implications when it comes to interpreting the results of a GWAS.

Fundamentally, GWAS are very similar to classic case-control epidemiological studies, where individuals with the disorder of interest (cases) are compared to a well-matched control group, to identify risk factors; in the case of GWAS, primarily common genetic variants. This is because, at low minor allele counts (MAC), the assumptions of logistic regression models are no longer met, and power and error rate are both impacted significantly [80]. Prior to association testing, it is vital to complete a rigorous quality control (QC) procedure. Errors in genotyping data can arise for a large number of reasons, including sample contamination or poor-quality DNA samples being collected, incomplete DNA hybridisation to the array, and issues with the DNA probes on the array itself. Any of these could cause spurious associations to arise within the data, and so it is paramount that QC protocols be stringent enough to account for these, whilst not being unnecessarily conservative and leading to genuine associations being excluded.

The largest commercially available arrays now contain over 4 million markers for study, but millions more can be assessed through the process of imputation. This is a process by which unobserved genotypes i.e. SNPs that were not present on the array, can be statistically inferred via comparison of the sample genotypes to a collection of known haplotypes, accessed via publicly available resources such as the International HapMap Project [81], 1000 Genomes [82] the Haplotype Reference Consortium [83], and TOPMED [84]. In brief, genotyped samples are compared to reference haplotypes at points where SNPs overlap in order to predict, taking advantage of linkage between markers, what SNP alleles would be present in regions not directly genotyped by the array. By doing this, genome-wide coverage can be massively increased, and disorder-associated SNPs not included on the array can still be investigated. At this point, the association testing can be conducted.

In the majority of cases, GWAS studies of binary phenotypes, such as disorder states, make use of a series of logistic regression models, where each SNP is assessed for statistically significant differences in allele frequencies between cases and controls. To ensure differences in allele frequencies due to genetic ancestry and other variables do not affect the results, a set of principal components will often be added as covariates to the regression model. Principal components are a set of newly generated variables that aim to reduce the dimensionality of data, or how many attributes or variables a dataset has, whilst still retaining as much information about the data as possible. In the case of GWAS, principal component analysis is performed on the genotypes of the study cohort and mostly reflect genetic variation due to ancestry. Individuals with similar values for the top principal components i.e. Those that explain the highest proportion of variation in the sample, will tend to have a similar genetic ancestry. Following completion of the regression analyses, each SNP will be assigned an odds ratio (OR) and a p-value. If the case cohort has been coded as 1 and the controls as 0, as is customarily done, an OR of greater than one is indicative of the effect allele being more common in the case cohort, and thus suggests that the SNP is tagging an area of the genome that is associated with the disorder phenotype.

Due to the millions of statistical tests that are conducted in a standard GWAS, a stringent p-value threshold is selected to control for false positive results; ordinarily, a threshold of $5x10^{-8}$ is used to denote 'genome-wide significance'. This threshold was first put forward by the authors of the 1996 paper outlining the theory of GWAS and is the p-value required to retain a < 5% false positive rate when completing 1,000,000 tests, a rough (but still considered valid) approximation for the number of uncorrelated SNPs across the genome after accounting for LD [77].

The GWAS method was first successfully implemented in a study of age-related macular degeneration (AMD) in 2005 [85]. That study, which investigated 116,204 SNPs from across the genome in a cohort of 96 cases and 50 controls, identified a significant association at one SNP. This was found to be an intronic polymorphism within the complement factor H (*CFH*) gene; to this day, this remains one of the most strongly associated AMD genetic variants identified. Whilst this success cannot be diminished, statistically significant associations were still relatively infrequent in the early years of GWAS. However, as genotyping costs reduced, and the formation of international consortia such as the Psychiatric Genomics Consortium allowed for the development of sample sizes that had never previously been attainable, associations of common risk alleles of small effect with disorders could now be robustly detected. As of the time of writing, over 5100 GWAS have been published, reporting more than 270,000 statistically significant associations [86]. A brief overview of GWAS in schizophrenia will be given below.

GWAS: schizophrenia

One of the first GWAS to be conducted in schizophrenia was completed in 2008, in a sample of 479 cases and 2937 controls [87]. The gene with the strongest evidence of association with schizophrenia in this study was *ZNF804A*, which encodes for a zinc finger binding protein expressed ubiquitously across the brain. The association was strengthened when a case cohort of bipolar disorder was added to boost the sample size, suggesting that it is a pleiotropic locus that is associated with psychotic disorders more broadly. Although the precise function of this protein is still not fully understood, a study conducted in 2016 suggested that the protein played a key role in neurite formation, the maintenance of dendritic spines and synaptic plasticity [88]. Dendrites are branch-like appendages that project from the cell body of neurons, receiving electrochemical signals from the axon terminals of surrounding neurons and conducting it to the soma, or cell body. In the year following this, three GWAS of schizophrenia were published [89-91], all of which implicated an area on chromosome 6 known as the major histocompatibility complex (MHC). This area, roughly 3.6MB in length, is a genetically complex region containing approximately 224 genes. The area is also known to have high levels of LD, making it difficult to identify specific causal genes when an association is detected in this region. However, the area is known to contain a large number of immunity related genes, pointing to a possible role of immunity-related processes in schizophrenia aetiology. Other notable loci included an intronic marker within *TCF4*, a transcription factor involved in foetal neurological development [92], and *NRGN*, which encodes for a protein expressed within dendritic spines that regulates the availability of calmodulin at the post-synaptic membrane. Calmodulin is an intermediate calcium-binding messenger protein that acts as a signal transducer for a number of proteins that cannot themselves directly bind to calcium. These GWAS represented previously unmatched sample sizes at the time of their publication but were still restricted in terms of statistical power. In 2011, the schizophrenia working group of the Psychiatric Genome-Wide Association Study (GWAS) Consortium (PGC) conducted its first GWAS on a total of 51,695 individuals [93]. This GWAS identified seven loci associated with schizophrenia, five of which were previously unreported. The most robust evidence was for an association within a microRNA, *MIR137*. This is a known regulator of neuronal development and adult neurogenesis, and four other schizophrenia loci achieving genome-wide significance in this study contained predicted targets of *MIR137*, lending evidence to the hypothesis that *MIR137* dysregulation could contribute to the schizophrenia phenotype. Other genome-wide significant results included loci within *CACNA1C*, encoding a subunit of L-type voltage dependent calcium channels, and *ANK3*, a gene encoding for an ankyrin protein found at the Nodes of Ranvier, the periodic gaps in the myelin sheath of some neurons that allows for rapid electrical transmission along axons. The combination of the dataset from this study and an additional Swedish cohort increased the sample size to 59,318 and led to the identification of 22 schizophrenia associated loci, over half of which were novel [94]. Genome-wide significant associations were found at a number of loci involved in calcium channel signalling, including the previously reported *CACNA1C* locus, as well as additional L-Type calcium channel units. Loci containing calcium homeostasis modulator genes, such as *CALHM1*, were also implicated. The MHC on chromosome 6 was again implicated, as was *MIR137*. In addition, 14 of the implicated genes from this study were targets of *MIR137*, including calcium channel genes (*CACNB2*), gene expression regulators (*VARS*) and immunity-related genes (*HLA-DQA1*). This study also estimated that a total of 8332 independent SNPs contribute to the genetic risk for schizophrenia, and account for about 50% of overall heritability.

These studies culminated in the second GWAS conducted by the schizophrenia working group of the PGC, a landmark study of 36,989 cases and 113,075 controls [95]. The aim of the working group was to combine all currently available genotyped schizophrenia samples into a single analysis, theorising that the major limiting factor in schizophrenia GWAS to this point was sample size. This proved to be correct, and using the newly combined PGC cohort, 108 genomic loci were identified, including for the first time the main therapeutic target of antipsychotics, *DRD2*. A number of associations were also observed within genes of glutamatergic transmission, calcium channel signalling, synaptic function and plasticity, and neurodevelopment. This GWAS has since been followed by a series of large-scale GWAS. The first, conducted in 2018 on a sample of 40,675 cases and 64,643 controls, identified 145 independent loci, 93 of which were significant in the 2014 study [96]. This was followed in 2019 by what was at the time the largest study of schizophrenia in East-Asians conducted, with 22,778 cases and 35,362 controls [97]. This identified 21 genome-wide-significant associations in 19 genetic loci, and a meta-analysis with European datasets from the PGC identified 208 associations in 176 loci, 53 of which had not been reported previously. The most recent GWAS conducted by the schizophrenia working group of the PGC built on these samples further, amalgamating data from 76,755 cases and 243,649 controls [98]. This identified 287 genomic loci associated with schizophrenia disease risk, including overlap with 107 of the 108 loci from the original PGC paper. This GWAS, released in conjunction with a rare variant analysis by the Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) Consortium [99], represents not just the largest schizophrenia GWAS to date, but also the most ancestrally diverse, comprising 74.3% European, 17.5% East Asian, 5.7% African American and 2.5% Latin American individuals. By combining the results of two gene prioritisation techniques, fine-mapping [100] and summary-based mendelian randomisation (SMR) [101], the authors published a list of 120 unique prioritised genes, 106 of which were protein coding. Some genes of particular interest were *GRIN2A*, which encodes for a subunit of glutamatergic NMDA receptor, and transcription factor *SP4*, which is both regulated by NMDA transmission and plays a role in the regulation of NMDA receptor abundance [102]. This GWAS, like many before it, identified concentrations of genes with suspected roles within the pre- and postsynaptic locations, with further genes being linked to synaptic organisation, differentiation, and transmission. It also found evidence for enrichment of schizophrenia risk genes in almost all tested brain regions, pointing to abnormal neuronal function throughout the whole brain versus a small number of specific brain structures. The GWAS listed here not only highlighted the importance of large-scale collaboration in attaining the sample sizes required to identify disorder associated loci, demonstrated neatly by the sheer number of genome wide significant loci identified in the most recent PGC GWAS versus the first (Figure 1), but also the complex genetic architecture of schizophrenia, and the sheer polygenicity of the disorder.

Post GWAS Analysis

GWAS have been instrumental in the advancement of our understanding of the genetics of schizophrenia, but they represent only the first step of an investigation. It is commonplace for researchers, following the completion of a GWAS, to perform a series of subsequent analyses to better elucidate the biological or functional meanings of the GWAS results. A subset of possible downstream analyses will be outlined below.

Gene Identification

The characteristic peaks of a Manhattan plot, a visualisation of GWAS summary statistics that allows for the assessment of each SNP’s association with the phenotype of interest simultaneously, are caused by clusters of adjacent variants all displaying statistically significant associations with the phenotype, as a result of LD. In this way, GWAS can point to genomic regions from which an association signal is coming but cannot identify either the true causal SNP or disorder associated genes without further interrogation of the results. Complicating the matter further is the fact, whilst the closest gene to a GWAS signal is indeed often the most credible gene candidate, this will not always be the case. A number of methods have been developed to identify disorder-related genes from the results of a GWAS.

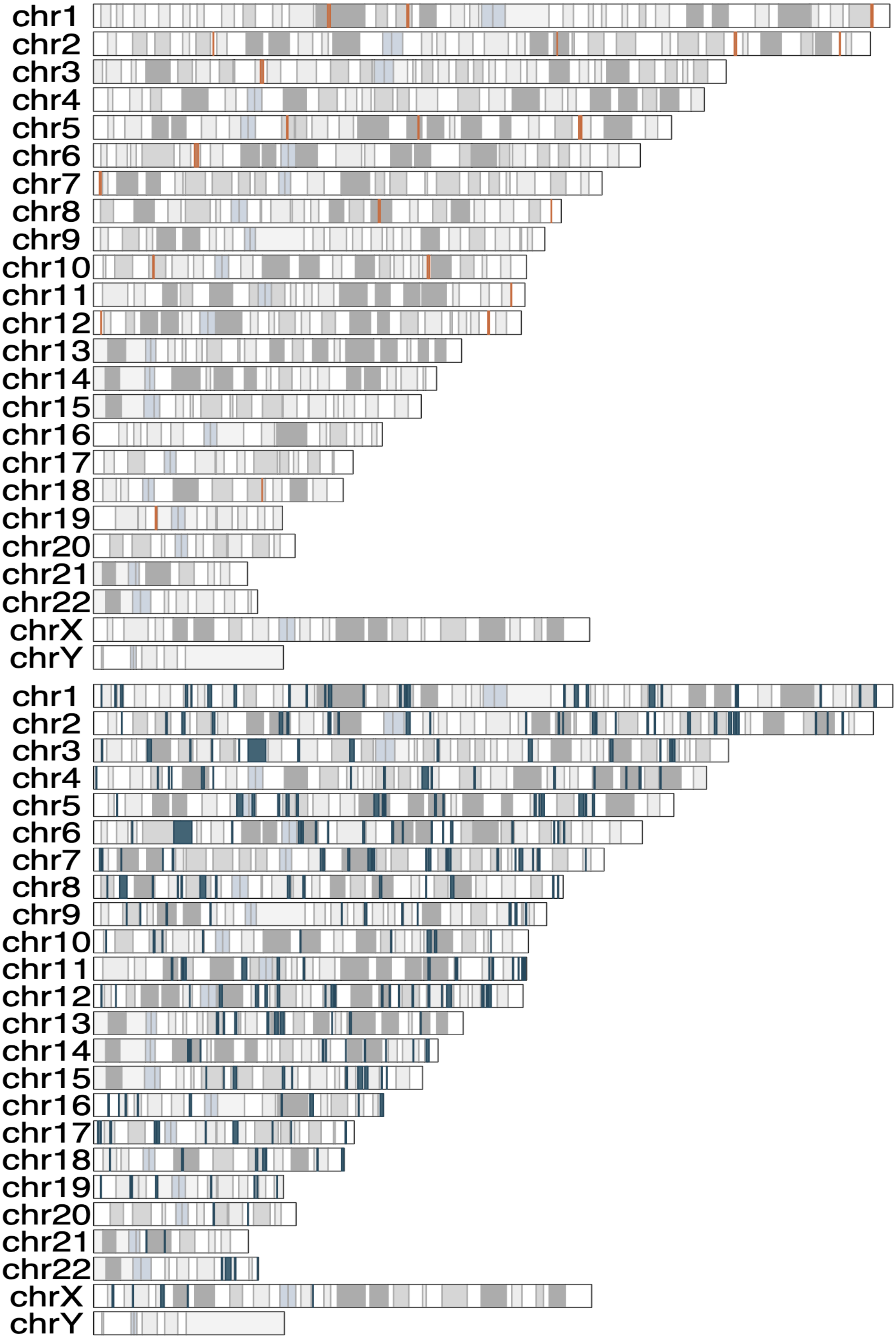

Figure 1: Karyotype plots, showing the number of genome wide significant loci identified in the original schizophrenia GWAS conducted by the PGC in 2011 (Top, highlighted in orange), and the newest GWAS conducted in 2022 (Bottom, highlighted in dark blue)

Gene Identification: Schizophrenia

The most recent schizophrenia GWAS from the PGC [98] utilised two primary methods to identify and prioritise genes: FINEMAP [100] and SMR [101]. FINEMAP is a Bayesian fine mapping method that involves the calculation of posterior probabilities for each genome-wide significant SNP, in an attempt to predict which is most likely to be the true causal variant. Using a shotgun stochastic search approach, iterations of the same locus will be repeated with a single candidate SNP added, removed, or exchanged, and each of the configurations are then compared to identify the most likely causal variant(s). The wider list of 628 genes (all of which contained at least one credible SNP) was refined following this by the selection of genes that contained at least one non-synonymous or untranslated region (UTR) variant that had an individual posterior probability of > 0.1. In 61 instances, the full 95% credible set of SNPs were restricted to within the boundaries of a single gene. SMR was used to assess whether the genome wide significant signals identified in the GWAS co-localised with expression quantitative trait loci (eQTLs) from both adult and foetal brain tissues as well as whole blood. The HEIDI (heterogeneity in dependent instruments) test [101] was used to reject co-localisations that occurred due to LD between eQTL variants and schizophrenia-associated variants. Following the implementation of further prioritisation techniques, for example chromatin conformation analysis, a further 55 genes were prioritised. By combining the findings of both FINEMAP and SMR, the authors of the 2022 GWAS developed a final list of 120 prioritised genes, 106 of which were protein coding [98].

Polygenic Risk Scores

A common follow-up analysis to GWAS is the generation of a polygenic risk score (PRS). This involves taking the summary statistics of a GWAS and using them to calculate the genetic liability to a trait in a separate, independent cohort. Each SNP is assigned a ‘weight’ based upon their estimated association with the phenotype of interest, and then the sum of these weights is calculated for each individual in the sample based on the number and combination of SNPs that they possess. It is common practice to then take these calculated scores and test for associations with disorder related traits, for example age of disease onset, measures of symptom severity, and outcomes such as educational attainment and employment status.

The application of polygenic risk scores in clinical settings can be thought of in terms of both their validity (how accurately does the PRS predict the phenotype it has been derived to predict) and their utility (what improvements in patient care / clinical decision making will the use of the PRS bring). The clinical validity of PRS can be assessed in a number of ways, for example the calculation of the median area under the receiver operating characteristic curve (AUROC), a model performance metric that assesses the predictive value of classification models. A review of the utility of PRS both personally and in a clinical setting identified 3 broad classes of PRS-informed interventions: therapeutic interventions, disease screening and life-planning [103].

Polygenic Risk Scores: schizophrenia

Due to schizophrenias extensive polygenicity and relatively low prevalence in the population, as well as the fact that diagnosis is based solely on symptoms, it is unlikely that PRS will ever be useful in the context of diagnosis. However, there are a range of other applications of PRS that have the potential to be of more significant use. For example, in research, it may be useful to have a quantitative approximation of disorder liability. In the most recent PGC schizophrenia GWAS [98], when comparing the lowest centile of the PRS to the highest, the individuals with the highest 1% of PRS had an odds ratio for schizophrenia of 39, and an odds ratio of 5.6 when compared to the remaining 99% pooled together. Another key potential application of PRS is in patient stratification, and the definition of more homogenous patient subgroups. For example, in a study comparing bipolar and schizophrenia cases, bipolar disorder PRS were significantly associated with increased prevalence of manic symptoms in schizophrenia cases [104]. In another study focusing on suicide attempt, a key factor in the reduced average life expectancy observed in schizophrenia, PRS for major depression were significantly positively associated with suicide attempt in cohorts of schizophrenia, bipolar disorder and major depression patients [105]. These findings suggest that common variants, to a certain extent, may be associated with specific symptom groups and outcomes, rather than disorders specifically. Lending further credence to this was a study that calculated schizophrenia PRS in 22q11.2 deletion carriers both with and without schizophrenia. The PRS was found to be significantly higher in carriers with psychotic symptoms that those who did not experience psychosis [106]. Another study in 2021 investigating whether genetic liability for schizophrenia is associated with specific symptom groups observed that schizophrenia PRS were associated with increased disorganised symptoms domain scores and decreased current cognitive ability, but not with the other symptom groups [11].

PRS has also been explored as a method of predicting treatment response. Schizophrenia PRS was found to be significantly associated with 12-week treatment outcomes in first-episode psychosis, with high PRS being associated with less improvement following antipsychotic treatment, and low PRS cases being almost twice as likely to be treatment responders [107]. In a study that examined disease progression over the course of 20 years, higher schizophrenia PRS was significantly associated with more severe negative symptoms, greater overall disease severity and higher levels of cognitive function [108]. These studies demonstrate how in the future it may become possible to better tailor treatment options to patients based on their PRS for schizophrenia. For example, if it is determined that a patient has a high likelihood of being treatment resistant, they can be prescribed clozapine in the first instance, rather than following two unsuccessful treatment trials with other anti-psychotic medications.

In addition, PRS has demonstrated the importance of trans-ancestry cohorts, with polygenic risk scores based on European samples being far less predictive of schizophrenia in Asian samples [97]. In this study, despite the genetic correlation between the Europeans only analysis and Asians only analysis being extremely high at 0.98, polygenic risk models based on the summary statistics of one genetic ancestry perform far worse when used in other genetic ancestries. When PRS were calculated for the East Asian cohort using the European-only summary statistics, despite the European effective sample size being close to triple that of the East Asian cohorts, the variance in schizophrenia risk explained by the PRS was reduced by a third. This was attributed to the fact that, whilst the majority of common variants were present in both ancestries, the allele frequency and LD structure could vary significantly, for example the complete lack of association found at the MHC region in the East Asian cohorts. This demonstrated the importance of building large cohorts of a wide range of genetic ancestries, as despite the high genetic correlation, there are still significant ancestry-related differences.

Finally, PRS have demonstrated very effectively that the genetic architecture of schizophrenia does not exist in isolation; large proportions of the common variants implicated in schizophrenia have also been linked to a wide range of other disorders. A study in 2014 demonstrated a substantial sharing of risk alleles between schizophrenia, bipolar disorder, autism spectrum disorder, major depression and obsessive compulsive disorder [109]. All of these disorders were demonstrated to be genetically correlated again several years later by the brainstorm consortium, whom compiled the summary statistics of 25 of the largest available GWAS to investigate the overlap of a wide range of psychiatric and neurodegenerative disorders [110]. Of these disorders, the most extensive overlap with schizophrenia was found with bipolar disorder, autism spectrum disorder and intellectual disability, an umbrella term used to describe limits in an individual's ability to learn and function at the expected level for their age.

GWAS Limitations

Whilst the impact that this method has had in schizophrenia research cannot be denied, it is important to be cognisant of several limitations when interpreting GWAS results. The first is that whilst GWAS results can be used in the identification of a locus of potential interest, it cannot formally identify the causal variant or gene without a formal fine-mapping effort. The use of generic commercial SNP arrays means that the variant(s) actually associated with the target phenotype may not even be genotyped, and the association signal may be wrongly attributed to another variant. Establishing causality can be even more challenging when the signal is coming from non-coding regions, where the functional importance of such variants can be difficult to discern [111].

Another important limitation to consider is the multiple testing burden that is necessitated by the study design. Whilst a genome-wide significance threshold of 5x10^{-8} is adopted to try and account for this, this is a Bonferroni corrected p-value that maintains a false positive rate of 5% based upon 1 million independent tests. It is not uncommon for modern GWAS to be conducted on ~8 million SNPs, and this number will likely only increase as GWAS moves from SNP arrays to whole genome sequencing (WGS) data, the current gold standard. The problem with multiple testing is therefore two-fold; whilst it could be argued that genuine signals are being disregarded because they do not reach the threshold, there is also an argument that as the size of GWAS increases, the current threshold is not stringent enough; indeed, some believe that as genotyping coverage improves, a higher threshold will become necessary [112]. Whilst this limitation can be overcome somewhat through increases in sample size boosting statistical power, there are some important caveats for this. For certain rare phenotypes, it will only ever be possible to gather limited samples for study, and it is also important that the drive for large sample sizes does not result in the inclusion of poorly characterised individuals who may end up obscuring the genuine causal signals. This is a concept that will be discussed extensively throughout research chapters one and two.

In addition to this, whilst it is a major focus of current research efforts, the majority of GWAS published to date in psychiatry have been primarily conducted in samples from European ancestries. Whilst restricting a sample to a single ancestry is beneficial in terms of controlling for population stratification, it immediately reduces the generalisability of the results found in these studies. For example, In the largest East Asian GWAS of schizophrenia to date, although the common variants were highly correlated in the Asian and European samples, there were some notable differences in terms of allele frequency; most significant of which was the association at the locus containing *CACNA2D2* (a calcium channel subunit), where the minor allele frequency (MAF) in the Asian cohort was 45% versus 0.7% in the European [97]. The association between schizophrenia and the MHC region of the genome was also not genome-wide significant in the analysis restricted to just East Asian participants, pointing to another ancestry specific difference. This is just one example of the importance of including individuals from a diverse range of genetic ancestries in GWAS samples, and accounting for population stratification in ancestrally diverse samples was a major point in the quality control procedures of research chapter two.

The final limitation of current GWAS work in psychiatry that will be discussed here is that, despite its unprecedented success in identifying common variants associated with psychiatric disorders, these variants still only account for a reasonably modest proportion of overall heritability. There are a few reasons that this may be happening. The one of most relevance to this thesis is that, in an effort to increase sample sizes to the level required to detect common variants of small effect, case cohorts may have become more heterogeneous, for example by employing a broader range of diagnostic methods to identify cases. Where historically recruitment to a case cohort would have most likely involved a standardised research diagnostic interview, it is now commonplace to also define cases based on clinical notes and diagnoses, as well as in some instances self-report. The potential effect of this was demonstrated in a study of major depressive disorder in the UK biobank [113]. The heritability of depression defined via 'minimal phenotyping', which relied on a series of self-report questions, was significantly lower than the heritability of a more strictly defined MDD phenotype, based on the Mental Health Questionnaire (MHQ). In the context of schizophrenia, self-reported diagnosis may prove to be more accurate due to the increased stigma that still surrounds the disorder versus depression, however sample heterogeneity has undoubtedly been introduced by the inclusion of individuals with treatment resistant schizophrenia in general schizophrenia samples. Although the literature is divided on whether TRS constitutes its own disorder subtype or is a matter of disease severity, recent large-scale studies have demonstrated that common genetic variation specific to TRS does exist [114]. On top of this, schizophrenia is a highly heterogeneous disorder in many other aspects, including symptom profiles, prognosis, response to individual antipsychotic medications and several measures of outcome such as cognition. As previously mentioned, it also displays significant genetic overlap with many other major psychiatric disorders. This has made the implementation of more personalised medicine approaches to schizophrenia treatment difficult.

Thesis Aims and Structure

As discussed, schizophrenia as a disorder is highly heterogeneous, a concept that Bleuler himself appeared to be cognisant of with his use of the term “Group of Schizophrenias” [3]. The amalgamation of extremely large samples through international collaborations such as the PGC has been fundamental to the success of identifying genetic variation associated with schizophrenia as a broad phenotype, but genetics specific to more homogenous patient groups, for example individuals with TRS, have remained difficult to elucidate. In addition, the heterogenous nature of schizophrenia samples, as well as their significant overlap with patients of genetically correlated disorders, has made the identification of ‘schizophrenia-unique’ loci, and as such schizophrenia specific neurobiology far from straightforward. Advances in personalised medicine approaches in schizophrenia will rely heavily on the identification of more genetically and phenotypically homogenous subsets of patients, as well as the stratification of individuals according to their response to antipsychotic treatment. It is gaps in this literature that this thesis will attempt to address.

This thesis has three research chapters. The first is an examination of the genetic differences between schizophrenia and bipolar disorder, through the use of a recently published research method, the CC-GWAS [115]. Following this, research chapter two focuses on the genetic differences between TRS and responsive schizophrenia (referred to throughout this thesis as non-TRS). Finally, the third research chapter will shift the focus to be solely TRS and will utilise pharmacokinetic and haematological data to better characterise the relationship between clozapine and neutrophils, in an attempt to improve understanding of the mechanisms underlying one of clozapine's most serious adverse side effects, agranulocytosis. A general discussion and conclusion chapter will follow.

Research Chapter One: Identification of Disorder-Specific Loci and Genes through Case-Case Genome-Wide Association Studies: An Application to schizophrenia and bipolar disorder

Chapter Summary

In psychiatry, schizophrenia and bipolar disorder are the most closely genetically correlated disorders, overlapping significantly in terms of symptoms, risk genes, outcome, and familial patterns of inheritance. Nevertheless, they remain diagnostically distinct disorders, with unique core symptoms and different effective treatment options. Previous studies of the genetic differences between schizophrenia and bipolar disorder have been hampered by the need to use individual level genetic data. Genotype data of this nature is not easily accessible for most, and requires very stringent sample matching in terms of, for example, genotyping array and cohort ancestry. However, more recently, an array of methods have been developed that allow for cross-disorder analyses using GWAS summary statistics, large-scale iterations of which are publicly available for all major psychiatric disorders as a result of international efforts to combine research cohorts such as the PGC. One such new method is the CC-GWAS [115], which was developed to identify common variants that are differentially associated between correlated pairs of disorders. The analysis outlined within this chapter identified 27 genome wide significant loci, 24 of which can be quite confidently considered loci unique to schizophrenia. Common genetic variants in these loci, on top of a pleiotropic background shared with bipolar disorder and psychiatric conditions more widely,

might contribute to biological processes leading to the preferential development of schizophrenia over bipolar or other related disorders.

Introduction

Bipolar disorder

Bipolar disorder is a major psychiatric disorder characterised by recurrent periods of mania and depression. There are two main subtypes of bipolar disorder; If the mania is severe, for example with psychotic features, this is characterized as Type 1, whilst less severe manic symptoms (termed hypomania) with concomitant periods of depression, distinguishes Type 2. Depressive episodes may also be present in patients with Type 1 bipolar disorder, but it is not a necessary feature for diagnosis. Bipolar disorder, much like schizophrenia, is a highly heritable disorder, with heritability estimates of 60-85% [116] and has a complex genetic architecture. In the most recent GWAS conducted by the bipolar disorder Working Group of the PGC, 64 genome-wide significant loci were identified, 33 of which were novel discoveries that had not been reported previously [117]. Much like the findings discussed in the introduction regarding schizophrenia, the bipolar disorder risk alleles discovered in this GWAS were enriched in gene sets involved in synaptic signalling, as well as other highly brain-expressed genes.

Pleiotropy / Phenotypic Overlap of Psychiatric Disorders

It has been well established that psychiatric disorders tend to be highly heritable, with very high proportions of variation in the disorders being attributed to genetic variation. For example, in a Swedish national study of eight psychiatric disorders in 4,408,646 full and half siblings, all of the disorders were found to be moderately to highly heritable, with schizophrenia, bipolar disorder, autism spectrum disorder and attention deficit hyperactivity disorder all displaying heritability of 51-80%, the highest being for ADHD [40]. Another study, which meta-analysed 2,748 publications of 17,804 phenotypes, collectively representing 14,558,903 twin pairs, demonstrated similarly high levels of heritability for a wide range of psychiatric phenotypes, with over half the examined studies focusing on psychiatric traits [39]. It has also been well established in past research that there is a high degree of genetic correlation between this set of disorders, the highest level existing between schizophrenia and bipolar disorder, which have a positive genetic correlation of 0.68 [118]. A study in 2019 of 232,964 cases of eight disorders identified a set of 109 loci that were associated with at least two psychiatric disorders [119]. 23 of these loci had pleiotropic effects on four or more of the eight disorders, including one that was associated with all 8 of the disorders; this was a locus on chromosome 18, indexed by SNP rs8084351 and attributed to the DCC gene. The protein coded for by this gene is a cell adhesion molecule that mediates axon guidance of neuronal growth cones and is vital for the proper developmental of neuronal projection fibres in the brain's white matter. Whilst this study and a large amount of other literature exists detailing the genetic similarities of psychiatric disorders, they remain diagnostically distinct disorders, with differences in representation i.e., Mania is not observed in archetypal schizophrenia, and treatment regimens. Whilst antipsychotic medications are used in the treatment of bipolar disorder in some cases, mood stabilisers such as lithium and valproate (also used as an anticonvulsant) are more commonplace. As such, genetic differences between disorders are likely to exist, and these have proven to be far more difficult to elucidate than genetic similarities. If specific genes that differentiate the psychiatric disorders could be identified, it could lead to the disorder-specific aetiology being better understood, ultimately resulting in the identification of novel drug targets in a field of medicine where treatment options have remained more or less static since the first psychiatric medications were developed.

Investigating Genetic Differences: Current Literature

The literature regarding specific genetic differences between schizophrenia and bipolar disorder is limited, but a small number of studies have been published to date. In 2011, a study of 506 bipolar cases, 523 schizophrenia cases and a shared cohort of 505 controls were used to assess 302,482 SNPs for differences in allele frequencies amongst cases [120]. One of the main aims of this work was to identify variants that, in the presence of a sufficient number of other pleiotropic risk variants, would cause one or the other disorder to develop. The results implicated voltage dependent calcium channel genes as being potentially interesting for follow-up analysis, but none of the variants surpassed genome-wide significance. In 2014, a cross-disorder analysis was conducted in 7129 schizophrenia cases versus 9252 bipolar disorder cases [104]. Whilst, again, no significant loci were identified with differential allele frequencies between cases, PRS analysis showed that bipolar disorder PRS were only associated with the manic factor in schizophrenia cases. Finally, in 2018, another case-case GWAS was conducted in a much larger cohort of 23,585 schizophrenia cases and 15,270 bipolar disorder cases [121]. With the boosted sample size, 2 genome-wide significant loci were identified as having divergent effects on schizophrenia and bipolar disorder (discussed further below). Both of these studies relied on access to individual-level genetic data, which is not readily available and restricts sample size due to the need for the cohorts to be strictly matched in terms of ancestry and genotyping array. Methods that rely on more readily available data formats, for example GWAS summary statistics, could therefore be highly advantageous in cross-disorder analyses.

A large number of methods have been developed recently to utilise GWAS summary statistics in the analysis of multiple disorders. For example, MTAG, or multi-trait analysis of GWAS, allows for the analysis of multiple related traits and can also account for overlapping samples [122]. HIPO, or Heritability Informed Power Optimisation, takes information from across phenotypes, but also across individual SNPs, to significantly increase the number of genome-wide significant associations that can be detected. [123]. The Multivariate Omnibus Statistical Test (MOSTest), developed for the multivariate analysis of regional brain morphology, identified 347 genomic loci associated with regional brain morphology in 26,502 participants of the UK Biobank, more than any previously reported study [124]. However, these methods are primarily focused on the improved detection of genetic similarities and pleiotropic loci between disorders by combining related traits together, not the genetic differences. Identifying loci with divergent effects on related traits is one listed use case for the Genomic SEM method [125], but again, it was primarily designed to analyse the joint genetic architecture of related complex traits. Another paper used a method called mtCOJO, or multi-trait conditional and joint analysis [126], to identify disorder specific SNP associations in psychiatric disorders by adjusting the summary statistics for each disorder for the effects of genetically correlated traits [127]. Using the largest schizophrenia GWAS available at the time [96], of the 130 genome-wide significant SNPs from this study, the significance of five was increased after adjusting for four other psychiatric disorders, and an additional eight showed an increase in effect size. In addition, 10 SNPs that were not significant in the original GWAS became significant in the conditional analysis. At the time of publication, this was the only example, to the best of knowledge, of a method that could utilise GWAS summary statistics to identify disorder specific loci. However, in 2021, a new method was published, that claimed to be able to effectively identify loci with divergent effects on disorder pairs through the utilisation of GWAS summary statistics [115]. This was called the CC-GWAS, and an overview of the method will be given below.

Case-Case GWAS (CC-GWAS) Method Overview

The CC-GWAS, or case-case genome wide association study, tests for differences in allele frequency between cases of two disorders using the summary statistics of two individual case-control GWAS [115]. For ease of explanation, a disorder pair will be referred to as A and B throughout this section. The method first converts the odds ratios for SNPs (i.e., from logistic regression in a case-control GWAS) to standardised observed scale betas based on a 50:50 case: control ascertainment (equation 5, [128]). Following this, it applies a set of weights to these betas, the calculation of which are based on one of three different methods. The first are termed 'OLS weights', and their proposed function is to minimise the expected squared difference between estimated and true A1B1 (A1 = cases of disorder A, B1 = cases of disorder B) effect sizes. These weights rely on a population-level quantity that the authors refer to as FST_{causal}, defined as the average normalised squared difference in allele frequencies of causal variants. This FST_{causal} is calculated based on a number of input parameters selected by the user of the method: The SNP-based heritability on the liability scale for both disorders (h^2l,A and h^2l,B), the lifetime population prevalence (K_A and K_B) of both disorders, the genetic correlation (R_g) and the estimated number of independent causal variants (m). An important assumption of the CC-GWAS method is that all 'm' SNPs impact both disorders with effect sizes that follow a bivariate normal distribution. The OLS weights are then calculated based on this FST_{causal} measure, with the addition of the sample sizes, the degree of sample overlap (overlapping controls increase the power of the method by allowing for a more direct comparison of the cases of both disorders) and the variance and covariance of the error terms of the calculated betas. These weights represent the primary CC-GWAS method; however, they were demonstrated through simulations by the authors to be susceptible to type I error in situations where the SNP has a non-zero effect size in the case-control GWAS, but an effect size of zero in the CC-GWAS. These are referred to by the method authors as 'stress test' SNPs, and in order to mitigate this, a second set of weights, known as exact weights, are also automatically calculated. These weights are sample-size independent and are based upon only the population prevalence ($1-K_A$ for disorder A, and -$1+K_B$ for disorder B). In addition to this, there is a third option, known as Delta weights. In the preprint version of this paper on BioRxiv [129] it was referred to as a 'naïve' method of particular interest to individuals who wanted to examine subtypes of the same disorder. In the finalised, published version of the paper, this designation is no longer present, and it is simply referred to as a 'simple method' that allocates a weight of +1 to disorder A and -1 for disorder B. The recommended p-value threshold for significance is different for each of the methods; 5E-08 for betas calculated using OLS weights, 1E-04 for the exact weights, and 1E-05 for delta weights. The summary statistics generated by the different methods are also recommended for different follow-up analyses; for clumping based on LD and polygenic risk score analysis for example, the OLS weighted results are recommended, and for genetic correlation analysis, the exact weighted results are preferred.

Following completion of the CC-GWAS analysis and the calculation of the betas and p values for each SNP, filtering is applied to any candidate CC-GWAS SNPs that surpass the 5E-08 threshold in the OLS weighted analysis. This filtering is designed to identify and discard false positive associations that can arise due to the differential tagging of a causal stress test SNP. There are three sets of criteria used in the filtering step, and the SNP is discarded when at least one of the three sets of criteria are met. The different sets of criteria are designed for varying sizes of input case-control GWAS.

For each candidate SNP, the 1Mb region around it is screened for the SNP with the largest product of case-control z-scores. This is selected as the potential stress test SNP (SNP max.zAzB). The first round of filtering is intended for GWAS of intermediate sample size, and consists of three steps:

1. The identified stress test SNP is likely to have the same population allele frequencies amongst both case cohorts, reflected by an exact-weighted p-value larger than 1E-04.
2. The stress test SNP is likely to be the causal SNP, reflected by absolute case-control z-scores almost as large as the largest absolute case-control z-scores in the region for both disorders.
3. The z-scores for the candidate CC-GWAS SNP in the case-control GWAS and the identified stress test SNP have a pattern consistent with differential tagging, defined as:

$$(zA_{\text{CCGWAS}}/zA_{max.zAzB}) - (zB_{\text{CCGWAS}}/zB_{max.zAzB})| < 1$$

Equation 1: Stress Test SNP Testing

All of these criteria must be met for the candidate SNP to be filtered out. If one or both of the GWAS are ‘underpowered’, with a Neff of less than 40,000 (a threshold set by the authors themselves), the stress test SNP is selected based on the largest maximum absolute case-control z-scores summed across the two disorders, and the candidate SNP is once again filtered out if the stress test SNP is likely to have the same population allele frequencies among cases of both disorders, reflected again by an exact weighted p-value larger than 1E-04.

If both input GWAS are well powered (Neff > 40,000), a different single criterion is applied, and is dependent on the power of the CC-GWAS compared to the power of the input GWAS. If it is significantly lower, the exact-weighted z-score of the candidate SNP will be much smaller than the z-scores of the input GWAS, and the candidate SNP will be filtered out accordingly. Once these steps have taken place, the analysis is complete, and the summary statistics are outputted.

The resulting summary statistics are supposedly analogous to performing an individual level case-case GWAS of a disorder pair. They do not represent SNPs associated with disorder A unique of disorder B (such as in genomic SEM [125]) but rather any SNP with a significant difference in allele frequency between the cases of the two disorders in either direction. No formal assessment of the nature of the association of each SNP is given by the method, for example if there is a zero effect in disorder A and a non-zero effect in disorder B, or effects in both disorders but in opposing directions. However, investigation of the odds ratios and p-values from the input case-control GWAS can often provide further elucidation, and the onus is on the user of the method to perform post-CC-GWAS analysis to investigate any significant loci further.

In the original study, schizophrenia [96] and bipolar disorder [130] were analysed, and 12 significant loci were identified. Of these, 5 were significantly associated with schizophrenia in the input GWAS, and the remaining 7 were deemed 'CC-GWAS specific'. This designation was assigned to a locus by the authors if none of the SNPs surpassing genome-wide significance had an $r^2 > 0.8$ with any of the genome-wide significant SNPs in the input case-control GWAS. They highlighted one of these loci as being of particular interest. The locus, defined by lead SNP rs1054972 on chromosome 19, is located within an exon of *KLF16*. The protein encoded for by this gene plays a role in transcription factor activity, and has been shown in *in vitro* study to act as a repressor of neurite outgrowth and axonal regeneration in central nervous system neurons [131]. The index SNP was non-significant in both of the input case-control GWAS, but based upon the p-values, it is suspected by the authors that this locus is associated with schizophrenia, rather than bipolar disorder, and could potentially be involved in the previously suggested mechanism of excessive synaptic pruning in schizophrenia [132]. In total, pairwise analyses of schizophrenia and seven other neuropsychiatric disorders were conducted, identifying 313 significant CC-GWAS loci summed across each of the pairs of disorders, resulting in 196 independent loci and 72 CC-GWAS specific loci. The CC-GWAS method does therefore seem to be a reliable, effective method for identifying loci that are differentially associated with pairs of disorders, even between disorder pairs that are moderately to highly heritable, such as psychiatric disorders.

Research Chapter Aims

Since the publication of the method, new GWAS of both schizophrenia and bipolar disorder have been conducted by the corresponding working groups of the PGC, representing significant increases in sample size over the GWAS used in the original paper. In addition, whilst the original paper examined the genetic differences across eight major psychiatric disorders, no follow-up analysis of the identified loci was conducted. In this chapter, the CC-GWAS method will be used in a focused investigation of the genetic differences between schizophrenia and bipolar disorder, to both identify disorder specific loci and genes that could offer novel biological insights into the aetiology of either or both disorders, and act as a validation of the method itself, investigating what analysis can be performed on the CC-GWAS summary statistics. In brief:

1. CC-GWAS analysis of schizophrenia and bipolar disorder will be performed, leveraging the increased power of the newer, much larger phase 3 PGC case-control GWAS, in order to maximise the power of the CC-GWAS method.
2. Following on from this, a series of follow up analysis will be performed on the resulting summary statistics. This included LD-based clumping of the results to identify GWS loci, polygenic risk score analysis, fine mapping, genetic correlation analysis and investigation of the results using the LAVA method.

The primary results discussed here are for the analysis of schizophrenia vs. bipolar disorder. However, the method was also used in an attempt to identify common variation specific to treatment-resistant schizophrenia (TRS) versus non-TRS. Although it was quickly realised that the two disorder subtypes were too highly genetically correlated for the CC-GWAS method, the analysis of TRS vs. non-TRS provided some key insights into the method itself, and so will be briefly discussed here. It also provides context for the next chapter of this thesis; when the disorders of interest are extremely highly correlated, a direct case-case GWAS using individual level data remains the gold standard.

Methods

All statistical analysis, data curation and data visualisations presented here were, unless otherwise specified, completed using the programming language R (v4.0.2) through the GUI RStudio (2021.09.0 Build 351). All the work that is about to be presented was completed by myself independently under the supervision of Dr. Pardiñas and Professor Walters, unless otherwise specifically indicated.

Case Control GWAS Datasets

Schizophrenia GWAS: The GWAS used here is the most recent study conducted by the schizophrenia working group of the PGC. At the time of the analysis, the article was in preprint [133], and has since been published [98]. In the time between preprint and publication, an additional set of 7,386 cases and 7,008 controls of African American and Latino ancestries were added. However, the summary statistics used at the time of this analysis did not yet contain these individuals, and so the results discussed here are based upon the European and East Asian PGC cohorts only (named the “core PGC dataset” in the published paper). The summary statistics used in this analysis also did not contain the 1,979 cases and

142,626 controls from deCODE genetics. This analysis was based on a set of 90 datasets that included 31,914 cases and 47,176 controls that had not been included in previous GWAS [96], for a total sample size of 67,390 cases, 94,015 controls and 7,585,076 SNPs. The precise makeup of the case cohorts varied, but in general, were a combination of patients with either a schizophrenia diagnosis, or schizoaffective disorder, which is a related disorder where psychotic symptoms and mood disturbances co-occur during the same episode. For the rest of the chapter, this GWAS will be referred to as PGC3 SCZ for ease.

Bipolar disorder GWAS: The most recent GWAS conducted by the bipolar disorder Working Group of the PGC (from here, referred to as PGC3 BD) was used in this analysis [117]. It contained a total of 57 bipolar disorder cohorts, primarily of European ancestry, collected in five waves by the bipolar disorder working group. The summary statistics were generated based on a total of 41,917 cases, a combination of bipolar disorder type 1 and type 2, 371,549 controls and 7,825,140 SNPs. The number of controls was boosted significantly by the inclusion of the deCODE genetics cohort, which added 192,602 controls to the sample alone.

A total of 664 cases and 45,497 controls overlapped between the two GWAS, as determined by Professor Stephan Ripke and colleagues following the deduplication of the schizophrenia and bipolar disorder GWAS; only the number of overlapping controls can be inputted as a parameter in the CC-GWAS methods, unless a third case-case GWAS is added, the so-called CC-GWAS+.

CC-GWAS Parameters

In order to conduct the method, in addition to the number of cases, controls and degree of control overlap, it is necessary to input the following parameters: the lifetime disorder prevalence (K), the SNP-based heritability on the liability scale (H^2l), the genetic correlation (Rg) and the estimated number of independently associated causal SNPs (m). An overview of the parameters used can be seen in Table 1 and were justified as follows:

PARAMETER	***SCHIZOPHRENIA***	***BIPOLAR DISORDER***
CASES	67390	41917
CONTROLS	94015	371549
K	1%	2%
H^2l	0.18	0.186
RG	0.68	
M	7350	
CONTROL OVERLAP	45497	

Table 1: An overview of the input parameters used for the CC-GWAS analysis of schizophrenia and bipolar disorder. 'K' = lifetime disorder prevalence, 'H2l' = the snp-based heritability on the liability scale, 'Rg' = the genetic correlation between the disorder pairs and 'M' = the estimated number of independently associated causal SNPs. Further justification of these parameters is provided in-text.

The lifetime disorder prevalence and SNP-based heritability on the liability scale were taken directly from the corresponding paper of each GWAS (K=1% and H^2l=0.18 for schizophrenia [133], and K=2% and H^2l=0.186 for bipolar disorder [117]). The genetic correlation (0.68) was again calculated via LD score regression [134], and the m number selected (7350) was calculated based upon the average of the estimates for schizophrenia and bipolar disorder put forward in the MiXeR paper [135] (8300 and 6400 for schizophrenia and bipolar disorder respectively). This was based on the suggestion of the CC-GWAS authors that if the m is estimated/known to be different between the disorders you are analysing, you should take an average of the two. The selection of the MiXeR paper as the reference of choice came from the analysis of TRS vs. non-TRS using this method, where several m's were selected in order to test the effects of changing the m number of the results, the selection of which can be considered quite arbitrary. Details of this will be discussed later in this chapter.

CC-GWAS Analysis

Schizophrenia was assigned as disorder A. No SNPs were deleted based upon missing values in the dataset, and no SNPs were deleted based on a MAF <= 0.01. No SNPs were deleted based on an odds ratio of > 2 or < 0.5, however 190,437 SNPs were deleted based on the Neff being < 2/3 of the maximum Neff. This resulted in a final number of 7,394,639 SNPs available for analysis in schizophrenia.

Bipolar disorder was assigned as disorder B. Again, no SNPs were deleted based on missingness, however 375 SNPs were deleted based on MAF <= 0.01. No SNPs had an odds ratio of > 2 or < 0.5, but 232,992 SNPs were deleted due to their Neff being < 2/3 of the maximum. Altogether this resulted in 7,590,773 SNPs available for analysis in bipolar disorder.

There was a total of 7,341,513 overlapping SNPs between the two datasets. No SNPs were deleted based on discordant chromosome or base pair positions between the datasets, and none were deleted based on differences in allele names. All references alleles were aligned correctly between both datasets and no reference alleles required changing. This is to be expected, as both GWAS were conducted using RICOPILI [136], the standardised QC, imputation and association testing pipeline utilised in the majority of PGC primary analyses.

Following the deletion of all strand ambiguous SNPs, of which there were 1,112,227, there was a final number of 6,229,286 SNPs available for the CC-GWAS analysis.

The OLS weights calculated for the analysis were 5.30e-01 for schizophrenia, and -5.17e-01 for bipolar disorder. The exact weights (sample size independent weights, calculated as 1-KA for disorder A and -1+KB for disorder B) were 9.9e-01 for schizophrenia and -9.8e-01 for bipolar disorder.

LD-based Clumping of Loci

The LD-based clumping of loci was conducted via PLINK V1.07 [137]. The LD reference used was the European ancestry-specific dataset available from phase 3 of 1000 genomes [82], with a gene locations list based on GRCh37 (accessed and downloaded via: https://www.cog-genomics.org/static/bin/plink/glist-hg19). The physical distance threshold for clumping was set to 3000kb, the significance threshold for index SNPs was set to p < 1e-04 and the LD threshold for clumping was r^2=0.1. A locus was considered to be significant if the p-value was less than 5e-08, in line with the p-value threshold used on the OLS-weighted CC-GWAS summary statistics. The less stringent threshold was selected primarily to be in line with the parameters used in previous PGC studies [96], but also to provide a list of nominally significant loci that may warrant further investigation in spite of not obtaining GWS.

SNP-based Heritability and Genetic Correlation Analysis

SNP-based heritability on the observed scale was calculated via LD score regression using the LDSC software V1.0.1 [134]. In all instances, the LD reference was the European ancestry-specific data from phase 3 of 1000 Genomes [82]. Prior to analysis, SNPs with an INFO score < 0.9 were excluded, and the datasets were trimmed to contain only those that are present in the third phase of the International HapMap Project [138]. This is a reference set of 1,440,616 SNPs genotyped in 1,184 individuals from 11 global populations.

For the genetic correlation analysis, the online resource LD-hub (http://ldsc.broadinstitute.org/, V1.9.3 was used [139]. In this project, the following trait groups were selected for analysis; smoking behaviour, psychiatric disorders, neurological diseases, education, personality traits, cognition and sleeping. The MHC is removed for all traits in LD-Hub, and all summary statistics included on the site are publicly available, non-sex stratified and predominantly based on cohorts of European genetic ancestry. As such, they are an appropriate match for each of the GWAS being used in this project. It is however worth noting that there is significant overlap between the cohorts of the GWAS used here and the GWAS that have been uploaded to LD-hub for schizophrenia [95], bipolar disorder [140] and the cross-disorder analysis conducted by the PGC in 2013 [118]

Locus-Specific SNP-based Heritability and Local Genetic Correlation Analysis (LAVA)

Local univariate SNP-based heritability's and bivariate local genetic correlations were conducted using LAVA (Local Analysis of [co]Variant Annotation) [141], a method and accompanying R package that first calculates the local SNP-based heritability of a locus for the phenotypes of interest, in order to confirm that there is sufficient genetic signal for bivariate analysis, before conducting local genetic correlation analysis. Known sample overlap is provided to the software so that it may be modelled as a residual covariance, removing potential upward bias caused by unaccounted for sample overlap. To reduce the overall number of tests conducted, but to avoid filtering out potentially interesting loci by being too stringent, a p-value threshold of 0.05 had to be surpassed by both disorders in the univariate analysis in order for the local genetic correlation to be calculated. By default, a minimum of two SNPs has to be present in the defined locus for LAVA to process it. Although primarily designed to detect loci that are shared between disorders, it also has potentially very interesting applications in the detection of differentially associated loci, in particular loci where the association is present in both disorders in opposing directions i.e., Associated with increased risk of one disorder whilst being 'protective' against another.

The loci were the same used in the original paper, which is a list of 2495 loci spanning the whole genome. They are each roughly 1Mb in length, generated using the European 1000 Genomes data [82]. The minimum number of SNPs per locus was set to 2500.

FINEMAP: schizophrenia

Fine-mapping of the PGC3 SCZ loci had already been completed by Dr Antonio Pardiñas as part of the publication, which allowed for a comparison of the 255 loci fine-mapped as part of the analysis (for simplicity, referred to as FINEMAP loci from this point forward). It was hypothesised that, given how schizophrenia specific the CC-GWAS results appeared to be, there would be a significant amount of overlap with the FINEMAP loci, both in terms of base pair boundary and genome-wide significant SNPs. The FINEMAP loci could also be used to further refine the list of genes identified by the CC-GWAS analysis.

FINEMAP loci were provided by Dr Pardiñas for this stage of the analysis. The MHC region was not fine-mapped in PGC3 SCZ, and so only 25 CC-GWAS loci could be compared. There were three broad research questions in this portion of the analysis:

1. How much of the total FINEMAP posterior probability of each locus is contained within the boundaries of each clumped CC-GWAS locus?
    a. Compare the BP boundaries of the CC-GWAS loci with the FINEMAP loci.
    b. Determine what proportion of the total posterior probability of a FINEMAP locus was accounted for by SNPs from just within the CC-GWAS BP boundaries.
2. How much of the total FINEMAP posterior probability of each locus is contained within CC-GWAS significant SNPs?
    a. Taking the full FINEMAP locus credible set, what proportion of the total posterior probability is accounted for by SNPs that were found to be GWS in the CC-GWAS analysis?
3. Do any of the CC-GWAS significant SNPs within each locus have:
    a. Have a listed impact of missense, reside within a UTR region, or have a CADD score of > 20, denoting particularly high levels of pathogenicity?
    b. Have a "posterior probability" or "posterior inclusion" values > 0.5?

FINEMAP: bipolar disorder

Fine mapping had not been completed as part of the PGC3 BD GWAS, and so was conducted for the purposes of this thesis using publicly available summary statistics. Locus information for the 64 significant loci from the GWAS was taken from the supplementary material, and all loci were fine mapped. The analysis was conducted using FINEMAP v1.4 [100], using the software's default parameters, for a single causal SNP at each locus, as the modelling of multiple causal signals per locus is error-prone without LD information computed using the GWAS sample itself. The LD reference was the publicly available European subset of the Haplotype Reference Consortium [83]. The loci were then compared against the CC-GWAS loci in the same way as the schizophrenia fine mapped loci.

Polygenic Risk Score Analysis: Risk Factors and Patient Outcomes in Cardiff COGS

Following on from the examination of the CC-GWAS loci, it was then decided that PRS should be generated using the CC-GWAS summary statistics as the base dataset. The primary aim of this was to determine whether the scores would offer any additional information above using the input case-control GWAS summary statistics. The hypothesis was that if the CC-GWAS results did indeed represent primarily a schizophrenia signal unique of its signal shared with bipolar disorder, they may display a stronger association with schizophrenia-associated variables than full schizophrenia GWAS PRS.

Cardiff COGS (Cardiff Cognition in Schizophrenia) is a primarily schizophrenia prevalence cohort made up of individuals with treatment resistant schizophrenia (TRS) and non-TRS, determined by the evidence of a lifetime prescription of clozapine (Total SCZ n = 817, TRS = 315, non-TRS = 502). As part of the ascertainment of this cohort, a detailed research diagnostic interview was conducted with each participant, and a wide range of risk factors and patient outcome variables are available for analysis, as described previously [142]. Due to its relatively large sample size and deep level of phenotypic information, it was considered a good target cohort in which to conduct PRS analysis.

Cardiff COGs makes up part of the PGC schizophrenia cohort, so firstly the CC-GWAS analysis had to be repeated using summary statistics generated with Cardiff COGs participants removed. The summary statistics were then trimmed to contain only SNPs with an INFO score of > 0.9 in Cardiff COGs, an MAF > 0.1, and non-ambiguous SNPs only. In addition, the MHC region was removed due to the high levels of LD present. PRS were then generated using PRSice-2 v2.3.3 [143] at three p-value thresholds: 1, 0.05 and 5E-08. Clump-kb was set to 250kb, clump-p was set to 1, and clump-r2 was set to 0.1. Of the 6,199,369 SNPs present in the CC-GWAS summary statistics, 3,880,365 SNPs were used to generate the PRS in Cardiff COGs.

To act as a comparator, schizophrenia and bipolar disorder PRS were generated using the same parameters described above. The base dataset for schizophrenia was PGC3 SCZ with Cardiff COGs removed, and for bipolar disorder the whole PGC3 BD could be used due to there being no overlap.

Following generation of the PRS, a series of regressions were conducted to assess the relationship of CC-GWAS PRS and a number of phenotypic variables. These included age at onset of psychosis, symptom domains, as outlined by Legge and colleagues [11], substance misuse, treatment resistance, course of illness, premorbid social adjustment, and educational attainment. As a base model, the dependent variable would be the phenotype, the independent the PRS, with the addition of seven covariates: Principal Component 1 – 5, generated in PLINK v1.07 [137], gender, and age at interview. In addition, where appropriate, additional covariates known to be associated with the phenotypic variable under study were added to the model; for the substance misuse and educational attainment variables, year of birth was also added, and for course of illness, duration of illness was added. Due to the nature of the phenotypic data, linear, logistic, and ordinal regression models were used as appropriate.

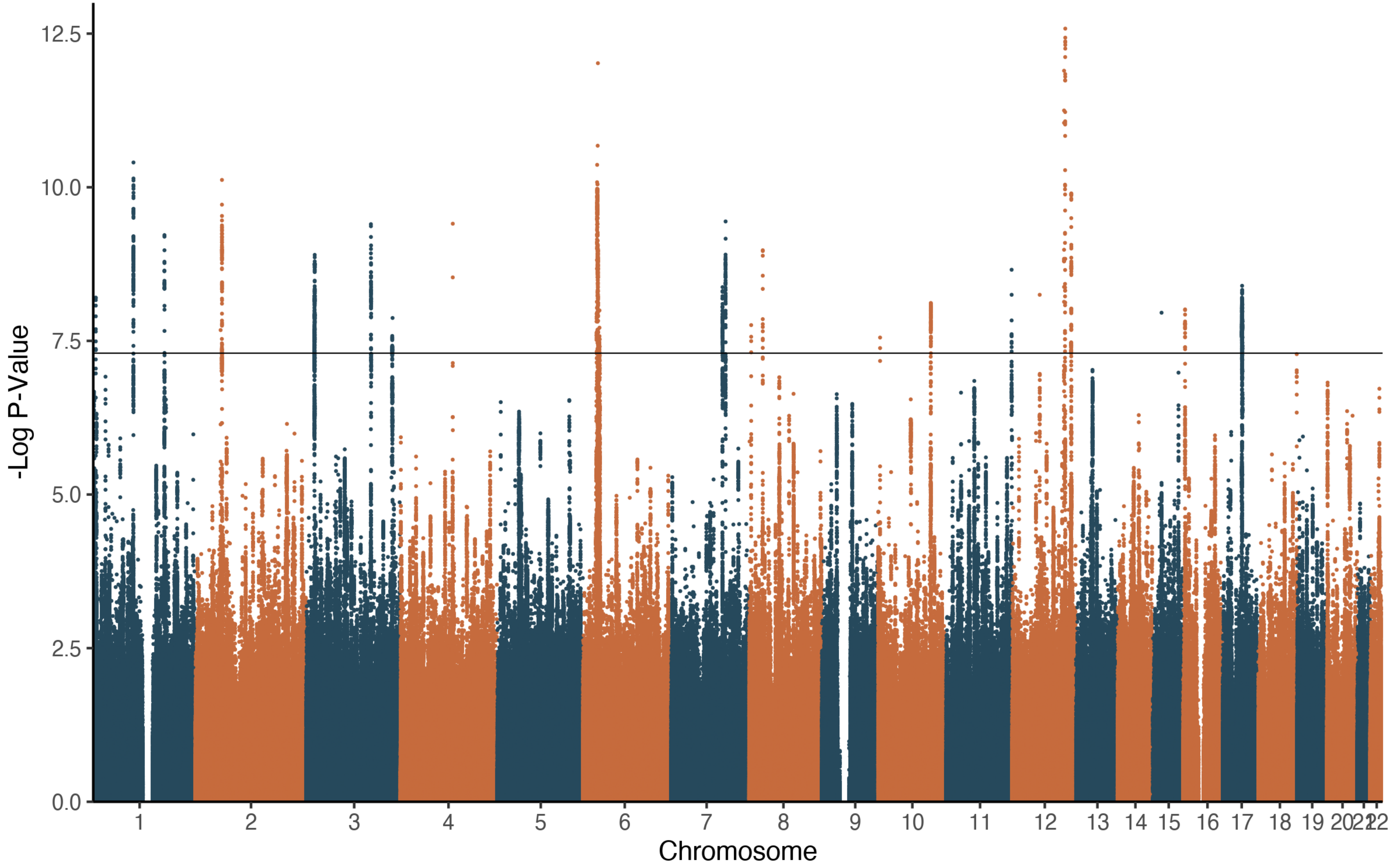

Figure 2: Manhattan Plot of the CC-GWAS results. The black line denotes a P-value of 5E-08. Generated using the 'ggman' R package in RStudio

Results

In total, 3099 SNPs were found to be associated with case-case status (schizophrenia cases versus bipolar cases; surpassing the author's recommended p-value threshold of < 5e-08 when utilising the OLS weights). Of these, 3 were filtered out based on potential differential tagging of a stress test SNP (details of this in the method overview section); all three were within 4Kb of each other on Chromosome 18. In total, there were therefore 3096 candidate CC-GWAS SNPs remaining at the end of the analysis. In Figure 1, a Manhattan Plot of the CC-GWAS summary statistics, calculated based upon the OLS weights, can be seen.

LD-based Clumping of Loci

In total, 27 loci surpassed genome-wide significance (P<5E-08), details of which can be seen in Table 2. The CC-GWAS results for these loci can be seen in Table 3

CHR	***SNP***	***P***	***START***	***STOP***	***GENES WITHIN LOCUS***
12	rs61942639	2.62E-13	110495648	111321512	VPS29, TCTN1, RAD9B, PPTC7, PPP1CC, IFT81, HVCN1, GPN3, FAM216A, CCDC63, C12orf76, ATP2A2, ARPC3, ANKRD13A, ANAPC7
6	rs9257566	9.58E-13	26175866	32107851	…
12	rs3764002	1.27E-12	108594044	108633649	WSCD2
1	rs4950119	3.95E-11	98186229	98604137	MIR2682, MIR137, MIR137HG, DPYD-AS2, DPYD
2	rs1518393	7.60E-11	57942987	58500141	VRK2, FANCL
12	rs4460848	1.26E-10	123424071	123909289	SETD8, SBNO1, RILPL2, PITPNM2, OGFOD2, MPHOSPH9, MIR8072, MIR4304, LOC100507091, CDK2AP1, C12orf65, ARL6IP4, ABCB9
7	rs37658	3.60E-10	110737149	111236477	LRRN3, IMMP2L
4	rs13107325	3.92E-10	102865304	103388441	SLC39A8, BANK1
3	rs1278493	3.99E-10	135669219	136508008	STAG1, PPP2R3A, PCCB, MSL2
1	rs9425755	6.01E-10	173389102	174979635	ZBTB37, SNORD77, SNORD81, SNORD80, SNORD78, SNORD79, SNORD75, SNORD76, SNORD74, SNORD47, SNORD44, SLC9C2, SERPINC1, RC3H1, RABGAP1L,

					PRDX6, MRPS14, LOC730159, LOC100506023, KLHL20, GPR52, GAS5, GAS5-AS1, DARS2, CENPL, CACYBP, ANKRD45
8	rs7838316	1.06E-09	27344719	27470597	MIR6843, EPHX2, CLU, CHRNA2
3	rs17273111	1.26E-09	17193348	17887988	TBC1D5
11	rs118031494	2.20E-09	133792743	133853008	IGSF9B
17	rs62062288	4.01E-09	43463493	44865603	WNT3, STH, SPPL2C, PLEKHM1, NSFP1, NSF, MIR4315-1, MIR4315-2, MGC57346, MAPT, MAPT-IT1,MAPT-AS1, LRRC37A4P, LRRC37A, LRRC37A2, LOC644172, KANSL1-AS1, KANSL1,CRHR1-IT1, CRHR1, ARL17A, ARL17B, ARHGAP27
7	rs2097942	4.22E-09	104476276	105040962	SRPK2, LINC01004, LHFPL3, LHFPL3-AS2, KMT2E-AS1, KMT2E
12	rs61937595	5.62E-09	57569478	57851182	STAC3, SHMT2, R3HDM2, NXPH4, NDUFA4L2, MIR1228, LRP1, INHBE, INHBC, GLI1, ARHGAP9
1	rs12562967	6.25E-09	6644723	7007010	ZBTB48, THAP3, TAS1R1, PHF13, LOC10050588, KLHL21, DNAJC11, CAMTA1
10	rs11191514	7.65E-09	104229588	104959852	WBP1L, TRIM8, TMEM180, SUFU, SFXN2, RPARP-AS1, NT5C2, CYP17A1, CNNM2, C10orf95, C10orf32-ASMT, C10orf32, AS3MT, ARL3, ACTR1A
16	rs62039173	9.73E-09	4413818	4596114	VASN, PAM16, NMRAL1, HMOX2, DNAJA3, CORO7, CORO7-PAM16, CDIP1, C16orf96
6	rs3130297	1.01E-08	29710380	32968136	…
15	rs7048	1.10E-08	43558004	44250313	ZSCAN29, WDR76, TUBGCP4, TP53BP1, TGM7, TGM5, STRC, SERINC4, SERF2, SERF2-C15ORF63, RNU6-28P, PPIP5K1, PIN4P1, PDIA3, MIR1282, MFAP1, MAP1A, LCMT2, HYPK, FRMD5, ELL3, CKMT1A, CKMT1B, CATSPER2, CATSPER2P1, ADAL
3	rs35746395	1.34E-08	180588841	181245594	SOX2-OT, LOC101928882, FXR1, DNAJC19
17	rs55938136	1.67E-08	43798360	43798360	CRHR1
8	rs10503253	1.75E-08	4177791	4220707	CSMD1
2	rs1822616	2.11E-08	55082530	55307811	RTN4, EML6
10	rs17731	2.79E-08	3795629	3828281	KLF6
10	rs79780963	3.54E-08	104952499	10495249	NT5C2

Table 2: The 27 regions that surpass p-value threshold of 5e-08. N refers to the number of SNPs that fall within the region, START refers to the bp position on the chromosome at which the region begins, STOP the ending base pair position, KB the total length of the region and GENES WITHIN LOCUS details the genes that fall within the region. For readability, the ranges column for the two loci on chromosome 6 have been left blank, due to them containing 167 and 246 genes respectively

SNP	***CHR***	***A1A0_B***	***A1A0_P***	***B1B0_B***	***B1B0_P***	***OLS_B***	***OLS_P***
rs61942639	12	0.025	1.64E-11	-0.00939	0.00279	0.0181	2.62E-13
rs9257566	6	0.0517	1.58E-36	0.0162	3.54E-07	0.019	9.58E-13
rs3764002	12	-0.0222	1.65E-09	0.0112	0.00033	-0.0176	1.27E-12
rs4950119	1	-0.0341	4.01E-20	-0.00332	0.291	-0.0164	3.95E-11
rs1518393	2	-0.0349	3.82E-21	-0.00446	0.16	-0.0162	7.60E-11
rs4460848	12	0.0365	3.20E-23	0.00664	0.0336	0.0159	1.26E-10
rs37658	7	0.0255	6.06E-12	-0.00446	0.171	0.0158	3.60E-10
rs13107325	4	-0.0391	2.90E-21	-0.00787	0.0121	-0.0166	3.92E-10
rs1278493	3	-0.0263	1.35E-12	0.00303	0.337	-0.0155	3.99E-10
rs9425755	1	0.0202	4.36E-08	-0.00892	0.0046	0.0153	6.01E-10
rs7838316	8	0.0285	3.72E-12	-0.00206	0.508	0.0162	1.06E-09
rs17273111	3	-0.0226	9.29E-10	0.00596	0.0585	-0.015	1.26E-09
rs118031494	11	0.0212	1.11E-08	-0.00699	0.0261	0.0148	2.20E-09
rs62062288	17	0.0234	7.53E-09	-0.00592	0.0597	0.0155	4.01E-09
rs2097942	7	-0.0259	2.69E-12	0.00158	0.612	-0.0146	4.22E-09
rs61937595	12	0.0292	1.32E-14	0.00151	0.632	0.0147	5.62E-09
rs12562967	1	-0.0139	0.000174	0.0136	1.64E-05	-0.0144	6.25E-09
rs11191514	10	0.0364	1.06E-22	0.00963	0.00215	0.0143	7.65E-09
rs62039173	16	0.0194	1.36E-07	-0.00761	0.0147	0.0142	9.73E-09
rs3130297	6	0.0442	8.28E-27	0.0157	8.69E-07	0.0153	1.01E-08
rs7048	15	0.0196	1.78E-06	-0.00912	0.00364	0.0151	1.10E-08
rs35746395	3	0.0264	1.02E-12	-0.000275	0.93	0.0141	1.34E-08
rs55938136	17	0.0235	1.23E-08	-0.00489	0.119	0.015	1.67E-08
rs10503253	8	-0.0245	4.37E-11	0.00193	0.535	-0.014	1.75E-08
rs1822616	2	-0.0156	2.40E-05	0.0109	0.000495	-0.0139	2.11E-08
rs17731	10	-0.0269	3.76E-13	-0.00097	0.757	-0.0138	2.79E-08
rs79780963	10	0.0359	2.53E-22	0.0104	0.000909	0.0137	3.54E-08

Table 3: CC-GWAS Results for the 27 loci identified by LD-based clumping of the CC-GWAS Results. A1A0 B/P refer to the summary statistics of the input schizophrenia case-control GWAS, B1B0 B/P refers to the summary statistics of the input bipolar disorder GWAS, and OLS B/P refers to the CC-GWAS results

As previously mentioned, interpretation of the results must be based upon the results from the input case-control GWAS to predict whether a locus is associated uniquely with schizophrenia or bipolar disorder. In this respect, regarding the genome wide significant loci:

- 23 / 27 index SNPs were significantly associated with schizophrenia in the input case-control GWAS, none of which displayed genome wide significance in the bipolar disorder GWAS, although a number were approaching this level of significance. As a result, they have been interpreted as ‘schizophrenia unique’ loci, at least with the current iteration of the input case-control GWAS.
- In PGC3 SCZ, 14 of these genome wide significant SNPs were positively associated with schizophrenia, and 9 negatively. Although the original input GWAS results were odds ratios, a positive beta is analogous to an odds ratio of greater than 1, and a negative beta is analogous to an odds ratio less than 1. If the exponent of the beta is calculated, giving the ratio OR(A): OR(B) with the CC-GWAS weightings applied, the results can be expressed in terms of a percentage change in effect size. For example, for rs61942639, the most significant result, the exponent of 0.0181 is 1.0183, indicating that there is a 1.82% increased effect size in schizophrenia relative to bipolar disorder. The results can therefore be interpreted as 9 loci that are uniquely associated with decreased risk of schizophrenia, and 14 that are uniquely associated with an increased risk of schizophrenia.
- The remaining 4 were not genome wide significant in either input GWAS, however the direction of the association can be inferred from the input case-control GWAS. For three of the SNPs (rs62039173, rs7048 and rs1822616), the SNP was more significant in PGC3 SCZ, sub genome wide significance, the first two listed positively associated with schizophrenia, the latter negatively. The final SNP (rs12562967) was more significantly associated with bipolar disorder and was negatively associated with schizophrenia and positively associated with bipolar disorder. This potentially

could represent a bipolar disorder specific effect, or a locus with divergent effects between the disorders, although the SNP was sub genome-wide significant in both input GWAS.

An example of how the results could be interpreted is as follows. The effect allele of the SNP rs7838316 on chromosome 8 was positively associated with schizophrenia in PGC3 SCZ (B = 0.0286, p=3.72e-12), and non-significantly associated with bipolar disorder (B=-0.002, p=0.508). This SNP was found to be positively associated in the CC-GWAS, and based on the input summary statistics, it can be inferred that the locus being tagged by this SNP is a schizophrenia unique locus associated with increased disorder risk. Taking the exponent of the CC-GWAS beta (1.0163), there is a 1.63% increase in effect size in schizophrenia. Based on the LD clumping procedure, this locus contains four potential gene candidates: *MIR6843, EPHX2, CLU, CHRNA2*. The SNP itself is an intronic variant of *CHRNA2*, which codes for a subunit of nicotinic acetylcholine receptors, and was found to be positive significantly associated in a GWAS of cannabis use disorder [144]

SNP-based Heritability and Genetic Correlation Analysis

In line with the recommendation of the authors, the summary statistics generated using the exact weights were used for this analysis. The SNP-based heritability on the observed scale for the CC-GWAS results of schizophrenia vs. bipolar disorder was 0.0345 (SE = 0.0014), with a genomic inflation factor (lambda) of 1.201. The lambda1000, calculated as (1 + (lambda - 1) * (1/case + 1/control) * 500) was 1.001. Of the 37 traits tested, 15 were found to be significantly genetically correlated with the CC-GWAS summary statistics. When applying a Bonferroni corrected p-value threshold of 0.00135 (0.05 / 37), 9 remain significant. Details of these traits can be seen in Table 4 and Figure 3. The results were significantly positively genetically correlated with schizophrenia (0.805), but not with bipolar disorder (0.0856). Whilst the CC-GWAS summary statistics should be not automatically interpreted as results specific to the disorder assigned as disorder A, given that 23 / 27 of the loci were associated with schizophrenia in the input GWAS, along with the high correlation with the schizophrenia GWAS on LD-Hub and lack of association with bipolar disorder, the results of these analyses do seem to represent mostly schizophrenia specific results. Of the 9 significantly correlated traits, 3 were negatively correlated with the CC-GWAS summary statistics: chronotype (-0.1248), intelligence (-0.1962) and subjective wellbeing (-0.2265). Based on our interpretation of the CC-GWAS statistics outlined above, this would mean that as genetic risk specific to schizophrenia increases, the listed phenotypes decrease (for the chronotype trait, this means that as risk specific for schizophrenia increases, the likelihood of being most active in the morning decreases)

TRAIT	RG	SE	P	H2
SCHIZOPHRENIA	0.8005	0.0153	0 *	0.4601
PGC CROSS-DISORDER ANALYSIS	0.5656	0.0371	2.24E-52 *	0.173
AMYOTROPHIC LATERAL SCLEROSIS	0.2383	0.0714	9.00E-04 *	0.0483
MAJOR DEPRESSIVE DISORDER	0.2063	0.0554	2.00E-04 *	0.1701
NEO-OPENNESS TO EXPERIENCE	0.1756	0.0805	0.0291	0.1063
SLEEP DURATION	0.1599	0.0369	1.50E-05 *	0.0555
AUTISM SPECTRUM DISORDER	0.138	0.0481	0.0041	0.4442
ALZHEIMER'S DISEASE	0.137	0.0685	0.0456	0.0452
NEUROTICISM	0.1263	0.0329	1.00E-04 *	0.0892
DEPRESSIVE SYMPTOMS	0.1178	0.0415	0.0045	0.0474
BIPOLAR DISORDER	0.0856	0.0407	0.0355	0.4381
ANOREXIA NERVOSA	0.0692	0.0284	0.0147	0.557
CHRONOTYPE	-0.1248	0.0321	1.00E-04 *	0.1008
INTELLIGENCE	-0.1962	0.0339	7.43E-09 *	0.1886
SUBJECTIVE WELL BEING	-0.2265	0.0348	7.72E-11 *	0.0251

Table 4: Genetic Correlation Analysis of schizophrenia vs. bipolar disorder CC-GWAS and 15 other phenotypes. A * denotes a p-value < 0.00135. The p-value for schizophrenia was so low that it was output as 0 by LD-hub. RG=Genetic Correlation, SE=Standard Error, H2= SNP-based heritability on the observed scale

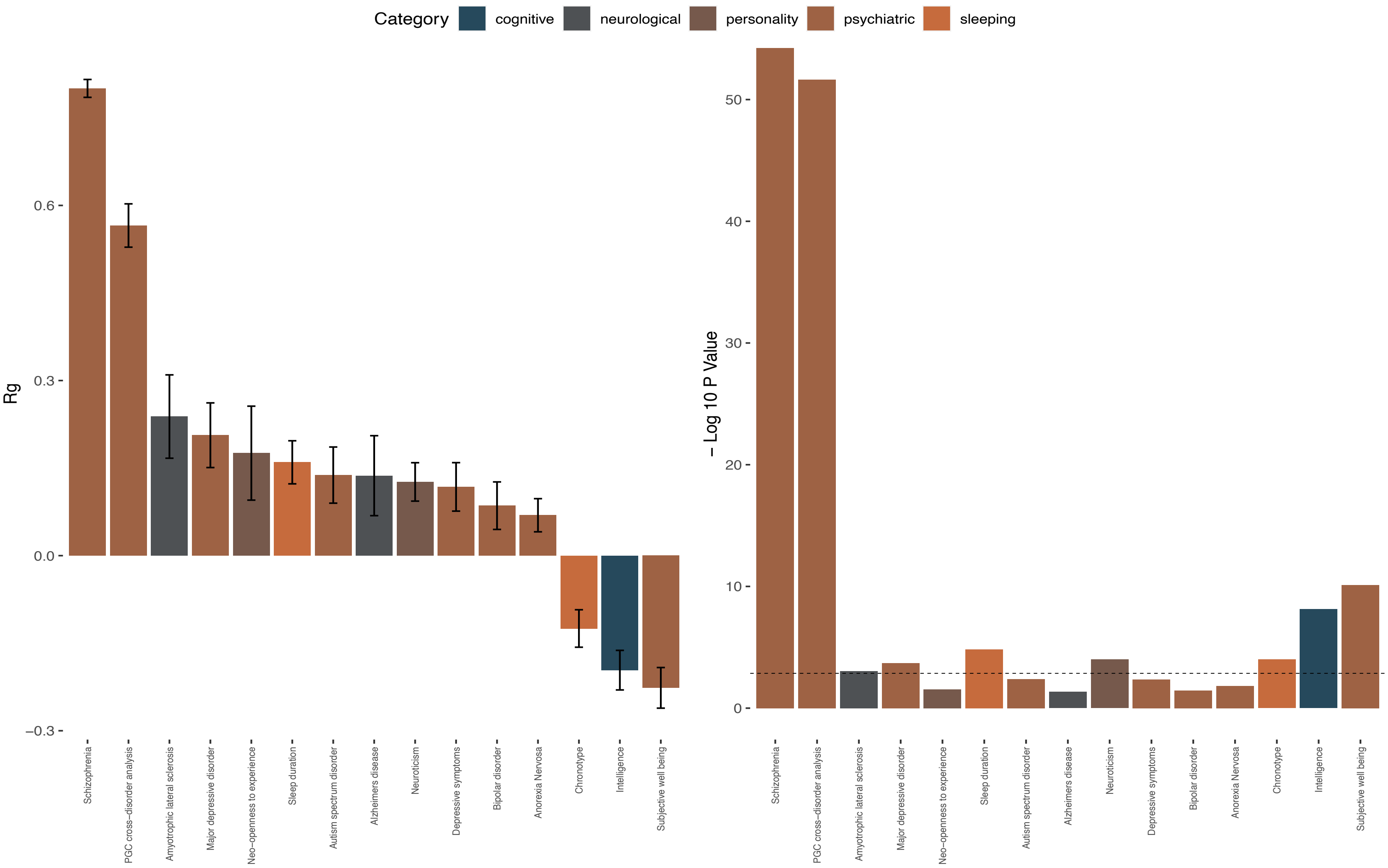

Figure 3: Results from the genetic correlation analysis of the CC-GWAS results with 15 other traits. The left-hand panel displays the genetic correlation (with SE bars), and the right-hand panel displays the p-values. The dashed line denotes the Bonferroni corrected p-value threshold for significance (0.05/37=0.00135). The colours denote the category of each trait on LD-Hub

FINEMAP: schizophrenia

Comparison of the CC-GWAS results with the FINEMAP results from PGC3 SCZ were conducted to try and identify the most likely causal gene for each CC-GWAS locus based on their overlap with a FINEMAP locus. Full results for this analysis can be seen in Table 5. The work outlined in this section sought to address the following questions:

1. How much of the total FINEMAP posterior probability of each locus is contained within the boundaries of each clumped CC-GWAS locus?

Two of the loci intersected with the MHC, and so could not be analysed further since this locus was not fine mapped in the PGC3 SZ study [98]. Of the remaining 25, 24 intersected / overlapped with the locus boundaries of a FINEMAP locus as defined in the PGC3 SZ study [98]. In 17 of the loci, the full posterior probability of the FINEMAP locus was accounted for by SNPs within the boundaries of a CC- GWAS locus. This is highly suggestive of the FINEMAP loci and the CC-GWAS loci capturing the same genetic signal at these overlapping loci.

2. How much of the total FINEMAP posterior probability of each locus is contained within CC-GWAS significant SNPs?

In three of the loci, the full posterior probability was contained within SNPs that had been significant in the CC-GWAS analysis. In a further 6, over half of the posterior probability was contained within SNPs that had been significant in the CC-GWAS.

CHR	START	STOP	FINEMAP LOCUS NO.	TOTAL NO. OF FINEMAP SNPS	TOTAL LOCUS PP	NO. OF FINEMAP SNPS WITHIN CCGWAS BOUNDARIES	PP CCGWAS BOUNDARIES ONLY	PROPORTION OF TOTAL PP	NO. OF SNPS OVERLAPPING WITH CCGWAS SIG. SNPS	PP CCGWAS SIG. SNPS ONLY	PROPORTION OF TOTAL PP
1	6644723	7007010	178	65	0.952	65	0.952	1	0	0	0
1	98186229	98604137	401	174	1.901	109	1.218	0.641	17	0.26	0.137
1	173389102	174979635	92	120	0.95	120	0.95	1	7	0.026	0.027
2	55082530	55307811	NO LOCUS OVERLAP								
2	57942987	58500141	204	93	1.907	93	1.907	1	42	0.83	0.435
3	17193348	17887988	82	144	0.95	144	0.95	1	119	0.825	0.868
3	135669219	136508008	14	51	0.954	51	0.954	1	25	0.729	0.764
3	180588841	181245594	207	120	1.902	120	1.902	1	0	0	0
4	102865304	103388441	5	1	0.992	1	0.992	1	1	0.992	1
6	26175866	32107851	MHC REGION. NOT ANALYSED FURTHER								
6	29710380	32968136	MHC REGION. NOT ANALYSED FURTHER								
7	104476276	105040962	18	23	0.95	23	0.95	1	1	0.001	0.001
7	110737149	111236477	23	32	0.951	32	0.951	1	3	0.378	0.398
8	4177791	4220707	43	9	0.956	9	0.956	1	4	0.66	0.691
8	27344719	27470597	29	257	1.908	112	0.987	0.517	11	0.962	0.504
10	3795629	3828281	28	3	1	3	1	1	2	0.86	0.86
10	104229588	104959852	201	355	2.853	349	2.745	0.962	73	0.826	0.29
10	104952499	104952499	SINGLE SNP WITHIN BOUNDS OF LOCUS 18								
11	133792743	133853008	216	88	3.84	40	1.03	0.268	6	0.709	0.185
12	57569478	57851182	19	1	0.994	1	0.994	1	1	0.994	1
12	108594044	108633649	108	15	1.983	3	0.984	0.496	3	0.984	0.496
12	110495648	111321512	208	1	0.994	1	0.994	1	1	0.994	1
12	123424071	123909289	405	5	5	2	2	0.4	0	0	0
15	43558004	44250313	39	46	0.951	44	0.817	0.858	0	0	0
16	4413818	4596114	128	81	0.951	81	0.951	1	19	0.16	0.168
17	43463493	44865603	68	1742	0.95	1742	0.95	1	1467	0.753	0.792
17	43798360	43798360	SINGLE SNP WITHIN BOUNDS OF LOCUS 26								

Table 5: The FINEMAP versus CC-GWAS results. ‘PP’= posterior probability. Columns 1-3 contain CC-GWAS locus information, columns 4-6 contain PGC3 SCZ FINEMAP locus information, columns 7-9 are in reference to the first question within the text and columns 10-12 are in reference to the second question within the text.

3. Do any of the CC-GWAS significant SNPs within each locus:
    a. Have a listed impact of missense (results in a different amino acid being encoded), occur within an untranslated region (UTR) (areas that occur on either side of a messenger RNA (mRNA) strand that play key roles in the post-transcriptional regulation of gene expression), or a CADD score of > 20, denoting high levels of potential deleteriousness based on conservation annotations?

In total, there were four CC-GWAS GWS SNPs that had a listed impact based on being located within an UTR or being a missense variant: rs13107325 (missense variant in SLC39A8), rs1043003 and rs17731 (UTR variants located in KLF6) and rs3764002 (missense variant in WSCD2). The two missense variants also had a CADD score of > 20 (23 and 27.6 respectively).

b. Have a "posterior probability" or "posterior inclusion" values > 0.5?

In total, 4 SNPs that were GWS in the CC-GWAS analysis had a posterior probability or posterior inclusion value of > 0. 5 signifying that these SNPs have over a 50% posterior probability of being the causal SNP within their respective locus. These included two of the SNPs described above, the missense variant in *SLC39A8* (rs13107325) and the UTR variant in *KLF6* (rs17731). The other two SNPs were both intronic variants within *R3HDM3* (rs61937595) and *ATP2A2* (rs4766428) respectively.

FINEMAP: bipolar disorder

Fine-mapping of the 64 GWS loci, excluding those in the MHC, from PGC3 BD was completed and compared to the CC-GWAS loci. However, the only overlap that occurred between the base pair boundaries of the PGC3 BD loci and the CC-GWAS loci were the loci within the MHC region, and so could not be compared further.

Locus-Specific SNP-based Heritability and Local Genetic Correlation Analysis

In an attempt to identify further loci that were differentially associated between schizophrenia and bipolar disorder, LAVA was first used to calculate the local observed heritability for both phenotypes at 2495 regions spanning the genome (details in the methods). This was done to determine which of the regions contain sufficient genetic signal in both disorders to calculate the local genetic correlation. Of 2495, there was determined to be significant (P<0.05) SNP-based heritability on the observed scale in both disorders in 1892, and so local genetic correlation analysis was then undertaken.

Applying a Bonferroni-corrected p-value threshold of 2.64e-5 (0.05 / 1892), 262 regions were significantly genetically correlated between schizophrenia and bipolar disorder (Figure 4). All these displayed positive genetic correlations, and so none of these would be considered as a differentially associated locus.

A potential application of the LAVA results would be for the further assessment of a CC-GWAS significant locus. A high positive genetic correlation at a CC-GWAS locus could potentially suggest that it is not actually differentially associated between the disorders, although there are some important caveats. Whilst the LAVA regions were all roughly 1Mb in length and contained 2500 SNPs at a minimum, no size constraints existed for the CC-GWAS loci. As such, some of them are very small, to the extent that two of the loci contain

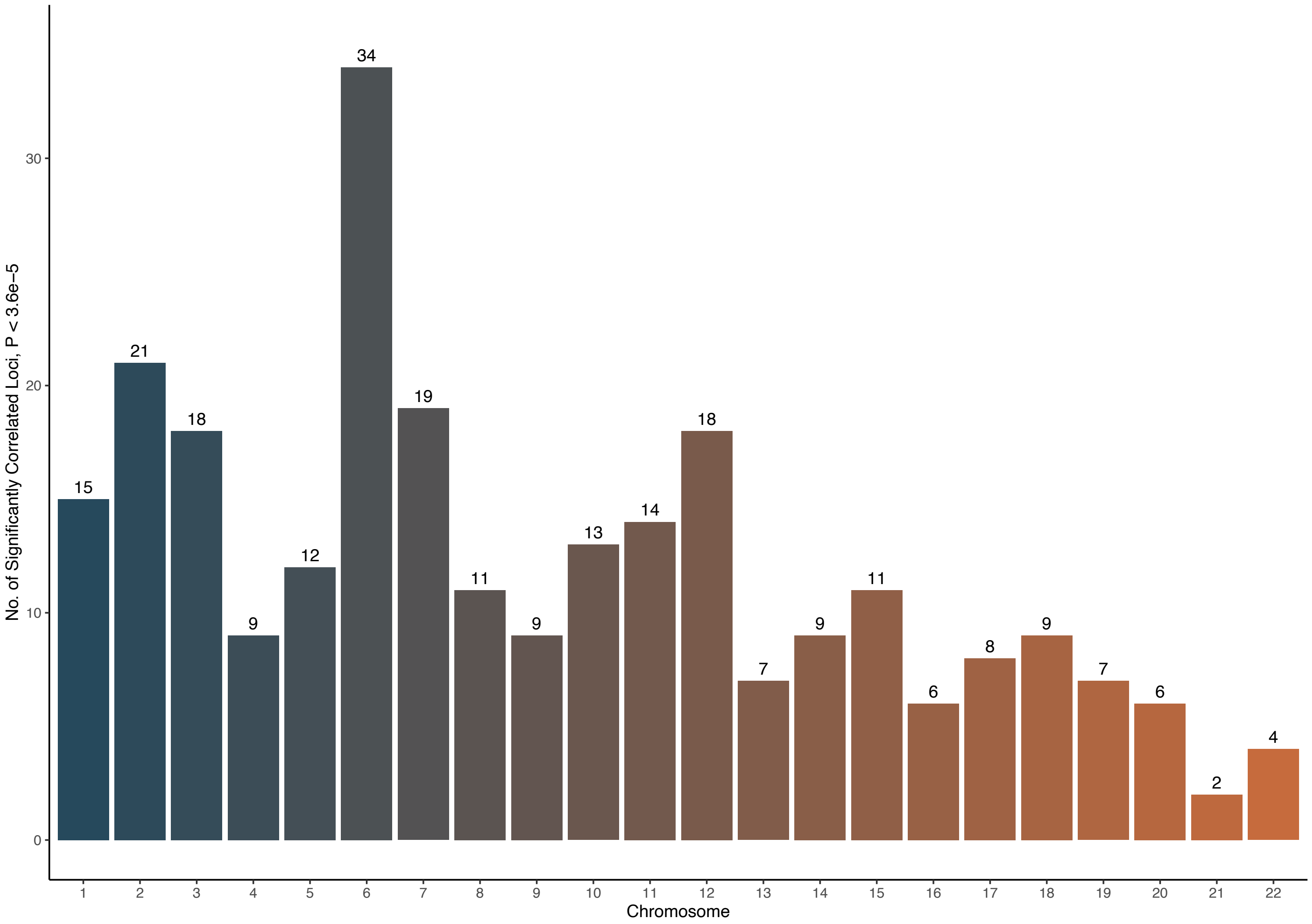

Figure 4: Graph showing the number of loci (as defined by the authors of the LAVA software package) by chromosome that were significantly genetically correlated between schizophrenia and bipolar disorder.

only a single SNP. It would therefore not be appropriate to say with confidence that the CC-GWAS locus and the overlapping LAVA region are capturing the same genetic signal. In addition, the p-value threshold for significance was different for the two sets of loci (5E-08 for the CC-GWAS analysis, and 3.6E-05 for the LAVA regions). Nevertheless, overlap of a CC-GWAS locus with a LAVA region could be indicative that it is not a truly divergent locus, and if the aim is to generate a list of genes that may potentially differentiate the two disorders, it may warrant their removal.

Of the 27 CC-GWAS loci, 10 overlapped with one or more of the 262 significant LAVA regions. Further analysis would be required to determine if this overlap merits their exclusion from future work.

Polygenic Risk Score Analyses: Risk Factors and Outcomes in Cardiff COGs

A summary of the results can be seen in table 6, with full regression results available in the appendix. As stated in the methods, 3 p-value thresholds were selected (1, 0.05 and 5E-08), and PRS were generated from schizophrenia, bipolar disorder and CC-GWAS summary statistics. In total, 9 regressions were generated per phenotype, for a total of 99 models. The results presented here are not corrected to account for this, as this section of the analysis was considered to be exploratory and hypothesis generating, and the findings will require future replication. The CC-GWAS PRS were associated with four phenotypes associated with schizophrenia: age at onset of psychosis (negative), negative symptoms of diminished expressivity (positive), disorganised symptoms (positive), and use of non-prescription drugs other than marijuana (negative).

PHENOTYPE	**PT_1**		**PT_0.05**		**PT_5E-08**	
	Estimate	**P**	**Estimate**	**P**	**Estimate**	**P**
AGE AT ONSET	-0.896	0.004	-0.829	0.009	0.002	0.996
NEGATIVE SYMPTOMS	0.09	0.016	0.116	0.002	0.028	0.456
DISORGANISED SYMPTOMS	0.132	0.0001	0.151	0.0001	0.041	0.276
NON-PRESCRIPTION DRUG USE	0.049	0.6	0.05	0.582	-0.18	0.047

Table 6 Summary table of significant results from the CC-GWAS PRS analysis in Cardiff COGs. 'Negative Symptoms' refers to negative symptoms of diminished expressivity, 'Disorganised Symptoms' refers to positive thought disorder and inappropriate affect and 'Non-prescription Drug Use' refers to regular (persistently for one month or repeatedly within one year) use of amphetamine, cocaine, heroin, LSD, solvents, benzodiazepine and ecstasy.

Discussion

In this chapter, the CC-GWAS method was used on the largest available case-control GWAS of schizophrenia and bipolar disorder conducted to date, for a total sample size of over 575,000 individuals, in an attempt to identify common variants with significantly different allele frequencies between the cases of both disorders. A total of 3096 SNPs, across 27 genomic loci, were identified, 24 of which can be inferred to be specific to schizophrenia. A comparison of the CC-GWAS loci with the FINEMAP loci of PGC3 SCZ allowed for the further prioritisation of genes contained within these loci, and also showed substantial locus overlap between the two methods. The CC-GWAS summary statistics were found to have an observable SNP-based heritability (0.0345), and displayed significant genetic correlations with 9 traits, which included schizophrenia, but not bipolar disorder. LAVA analysis was conducted in an attempt to identify further significant loci with divergent effects between the disorders, however all of the 262 significant loci displayed a positive correlation, conferring shared genetic signal. Finally, PRS analysis in a prevalence schizophrenia sample

(Cardiff COGs) identified associations with a range of phenotypes, including age of onset and disorganised symptoms, and were found to perform similarly to schizophrenia PRS.

Identified Genes

A subsection of the prioritised genes will be discussed further in the Appendix. For three of the CC-GWAS loci, the full posterior probability of the corresponding FINEMAP locus from PGC3 SCZ is accounted for by CC-GWAS significant SNPs and will be discussed first. For a further five, over 50% of the posterior probability is accounted for CC-GWAS significant SNPs, and the full credible SNP set of the corresponding FINEMAP locus falls within the base pair boundaries of a CC-GWAS locus and will also be discussed. With the exception of two (STAG1 and TBC1D5), each of these genes are in the final prioritised gene list generated as part of the work PGC3 SCZ. Finally, the locus for which there was no overlap with a PGC3 FINEMAP locus will also be discussed. A summary table of the CC-GWAS locus, with the gene determined to be most likely to be causal in each locus, can be seen in Table 7.

CHR	LEAD SNP	GENE
1	rs12562967	Intergenic
1	rs4950119	MIR137HG
1	rs9425755	DARS2
2	rs1518393	VRK2
2	rs1822616	RTN4
3	rs1278493	TBC1D5
3	rs17273111	STAG1
3	rs35746395	SOX2-OT
4	rs13107325	SLC39A8
6	rs3130297	MHC Region
6	rs9257566	MHC Region
7	rs2097942	KMT2E
7	rs37658	IMMP2L
8	rs10503253	CSMD1
8	rs7838316	GULOP
10	rs11191514	CNNM2
10	rs17731	KLF6

10	rs79780963	NT5C2
11	rs118031494	IGSF9B
12	rs3764002	WSCD2
12	rs4460848	MPHOSPH9
12	rs61937595	R3HDM2
12	rs61942639	ATP2A2
15	rs7048	LCMT2
16	rs62039173	CORO7
17	rs55938136	CRHR1
17	rs62062288	MAPT

Table 7 List of CC-GWAS loci with a single prioritised gene, based primarily on the results of PGC3 SCZ, as well as SNP position

Comparison of Results

Replication of Original CC-GWAS Results

In the paper outlining the CC-GWAS method [115], analysis of schizophrenia and bipolar disorder was conducted (based upon the largest publicly available case-control GWAS for each disorder at the time of publication [96,130]). Due to both of these GWAS being earlier studies conducted by the PGC, there is significant sample overlap between the GWAS used in the original paper and those used in this project. It would therefore be inappropriate to consider this section as an independent replication of the original study results. It is also worth noting that the input parameters for the calculation of FST_{causal} were different. For example, the m number (the expected number of independently associated causal SNPs) used in the original study was 10,000, whilst an m of 7350 was used here. This was due to the belief that the selection of 10,000 in the original paper was somewhat arbitrary, based only on the fact that schizophrenia and bipolar disorder are both suspected to be highly polygenic. The m number used in this analysis was the average of the estimates reported for schizophrenia and bipolar disorder by Frei and colleagues in the MiXeR paper [135]. MiXeR estimates the number of independent causal variants that account for 90% of the total heritability of a disorder based on causal mixture modelling of GWAS summary statistics incorporating additional information from LD structure, MAF and sample size. The population prevalence's, SNP-based heritability's and genetic correlation were also different; the justification of the selection of the input parameters in this project are given above. In the original paper, twelve loci surpassed genome wide significance (Table 8). Five of them were significant in the case-control GWAS of schizophrenia, whilst the remaining seven were deemed 'CC-GWAS specific' i.e. They had not reached significance in either of the input case-control GWAS. It Is worth noting however, that this does not mean the locus has not been significantly associated with the disorders in other GWAS, just not the two that the analysis is being conducted upon.

SNP	CHR	A1A0 BETA	A1A0 P	B1B0 BETA	B1B0 P	A1B1 BETA	A1B1 P	CC-GWAS SPECIFIC
RS2660304	1	0.0285	2.18E-18	0.00334	0.472	0.0141	2.23E-09	No
RS6701877	1	-0.0182	2.37E-08	0.0111	0.0173	-0.0146	5.81E-10	No
RS9866687	3	0.0124	0.000138	-0.0145	0.0017	0.013	4.05E-08	Yes
RS1278493	3	0.0198	1.21E-09	-0.00621	0.18	0.0135	1.24E-08	No
RS7790864	7	0.0146	7.18E-06	-0.0123	0.00793	0.0132	2.18E-08	Yes
RS11778040	8	-0.0196	1.66E-09	0.0051	0.273	-0.0129	4.85E-08	No
RS12554512	9	-0.00622	0.0554	0.0225	1.28E-06	-0.013	4.06E-08	Yes
RS3764002	12	0.0162	6.05E-07	-0.0154	0.000904	0.0155	6.33E-11	Yes
RS28637922	12	-0.0162	5.96E-07	0.0171	0.000228	-0.0162	8.14E-12	No
RS9319540	16	0.0122	0.000184	-0.0149	0.00126	0.013	3.67E-08	Yes
RS1054972	19	-0.0142	1.32E-05	0.0131	0.00474	-0.0133	1.75E-08	Yes
RS11696888	20	-0.0121	0.000194	0.018	0.000105	-0.0143	1.39E-09	Yes

Table 8:The genome-wide significant results from the original CC-GWAS paper. A1A0 BETA/P are in reference to the case-control schizophrenia GWAS used (Pardiñas et al, 2018), B1B0 BETA/P are in reference to the input bipolar disorder GWAS (Stahl et al, 2019). A1B1 BETA/P are in reference to the original CC-GWAS results (Peyrot and Price, 2021)

Table 9 displays the results for the 12 index SNPs from the original study from this current project. One of the loci could not be analysed as the index SNP, rs6701877, was missing from this project. However, the base pair position of rs6701877 does reside within the BP boundaries of one of the CC-GWAS loci outliner here, so it is possible the same association has been identified, just with a different index SNP. Of the other 11 loci, 5 remained significant in this current project, only one of which was a 'CC-GWAS Specific' loci. This was the locus located on chromosome 12, containing a single gene, WSCD2. Very little is known about the function of the protein product of this gene, making it difficult to predict the biological relevance of this protein to schizophrenia. However, there is reason to believe that this locus is not truly specific to the CC-GWAS analysis. The index SNP (rs3764002) is now significantly associated with schizophrenia in PGC3 SCZ [98] and was also significantly associated with schizophrenia in another GWAS of individuals with east Asian ancestry [97]. PGC3 SCZ contains ~ 20% East Asian cases, and there is considerable sample overlap between the study conducted by Lam et al. and PGC3 SCZ. As a result, it seems inappropriate to consider this locus as being specific to CC-GWAS, as there is reasonable evidence to support its association with schizophrenia. This is one example of why it is important to robustly investigate the results outputted by the CC-GWAS method.

SNP	CHR	A1A0 BETA	A1B0 P	B1B0 BETA	B1B0 P	A1B1 BETA	A1B1 P	REP
RS6701877	1	NA	NA	NA	NA	NA	NA	NA
RS3764002	12	-0.0222	1.65E-09	0.0112	0.00033	-0.0176	1.27E-12	Yes
RS2660304	1	-0.0339	5.14E-20	-0.00359	0.253	-0.0161	7.75E-11	Yes
RS28637922	12	0.0196	1.95E-07	-0.0106	0.000674	0.0159	2.39E-10	Yes
RS1278493	3	-0.0263	1.35E-12	0.00303	0.337	-0.0155	3.99E-10	Yes
RS11778040	8	0.0247	2.61E-11	-0.00166	0.596	0.0139	1.88E-08	Yes
RS11696888	20	0.0119	0.0013	-0.0107	0.000976	0.0118	2.65E-06	No
RS12554512	9	0.00891	0.0165	-0.0116	0.000222	0.0107	1.57E-05	No
RS9866687	3	-0.0101	0.00608	0.00973	0.0019	-0.0104	2.82E-05	No
RS7790864	7	-0.0114	0.00189	0.00598	0.0549	-0.00914	0.000227	No
RS1054972	19	0.0102	0.00585	-0.00373	0.251	0.00733	0.00359	No
RS9319540	16	-0.00833	0.0252	0.00399	0.202	-0.00648	0.00896	No

Table 9: Results for the 12 loci identified in the original CC-GWAS paper (Peyrot and Price, 2021) from the current analysis. A1A0 BETA/P is in reference to the input schizophrenia GWAS (Trubetskoy et al, 2022), B1B0 BETA/P is in reference to the input bipolar disorder GWAS (Mullins et al, 2021), A1B1 BETA/P is in reference to the current CC-GWAS analysis. REP= was the locus replicated from the original analysis

Comparison with Ruderfer et al. Results

In the direct case-case GWAS of 20,129 individuals with bipolar disorder and 33,426 individuals with schizophrenia conducted in 2018 by Ruderfer et al, two loci were identified with divergent effects on schizophrenia and bipolar disorder; a locus on chromosome 1, the lead SNP of which resides within the intron of *DARS2*, and a locus on chromosome 20 attributed to *ARFGEF2* [121]. In both cases, the MAF was found to be higher in bipolar disorder cases than schizophrenia cases. The locus on chromosome 20 was not replicated in this study, however *DARS2* was located within one of the ranges generated during the LD-based clumping procedure, with the index SNP rs9425755. The protein product of this gene is a mitochondrial enzyme that has been shown to be associated with leukoencephalopathy with brainstem and spinal cord involvement, an autosomal recessive disease associated with cerebellar ataxia, dorsal column dysfunction, and sometimes, cognitive deficits [145]. It has not to date been associated with neuropsychiatric disorders in GWAS prior to this. Three further loci were mentioned in the Ruderfer paper that did not surpass genome-wide significance, none of which were replicated in this analysis.

Comparison with Byrne et al. Results

In the analysis of Bryne and colleagues, using mtCOJO to generate schizophrenia summary statistics conditioned on the summary statistics of four other psychiatric disorders, 15 of the 130 schizophrenia loci identified in the input case control GWAS [96] were identified as being particularly specific to schizophrenia [127]. Of these 15, one locus displayed overlap with a CC-GWAS locus identified in this analysis and contains a single gene, *WSCD2*. This gene is explained in more detail in one of the previous sections regarding replication of the original CC-GWAS results. It is likely that had the study by Byrne and colleagues also made use of the PGC3 SCZ summary statistics, more overlap would have occurred between the two sets of results.

These results demonstrate the utility of the CC-GWAS method for identifying potential genetic differences between pairs of disorders. However, they also demonstrate the high level of care that must be taken when interpreting the results. A lack of association in the input GWAS for the CC-GWAS index SNP is not sufficient to designate that SNP as 'CC-GWAS specific' because, as demonstrated here, not only can the SNP be significantly associated in other GWAS of the disorders being analysed, but the genes attributed to the CC-GWAS SNPs can also have been identified in the input GWAS, just with a different index SNP. However, as long as these caveats are understood and the loci are suitably assessed through follow up analysis and review of the literature, the CC-GWAS can effectively identify loci with divergent effects in disorders, as well as loci that are specific to only one disorder. Local genetic correlation analysis, for example through LAVA, can be used in conjunction to provide additional information about any identified loci. It therefore presents the opportunity to generate sets of genes that could offer insights into the aetiology of specific disorders, and in turn future potential drug targets. This would allow for treatment options to become more specific for each disorder, which currently tend to be treated with the same set of medications. In addition to the identification of disorder-specific genes, the CC-GWAS summary statistics can be subjected to much the same post-hoc analysis as case-control GWAS summary statistics. Here, it was demonstrated that CC-GWAS results could be used for genetic correlation analysis, and the generation of PRS, but may other potential application exist.

Limitations: The CC-GWAS Method

There are a number of limitations of the CC-GWAS method that are outlined by the method authors themselves [115]. The main one of relevance here has been mentioned previously; the method does not provide any formal assessment of which of the disorders the SNP is associated with, and it is the responsibility of the user to deduce this. It is also very important not to generalise the results too much beyond the input case control GWAS, as you cannot guarantee that a SNP that displayed no association in the selected input GWAS did not display a significant association in another GWAS of the disorders. This can be mitigated somewhat by using the largest, most highly powered GWAS as the input, but again, the onus is on the user to investigate each observed association. Another limitation of relevance in this work is the effect of the power of the input GWAS on the results. In this case, the schizophrenia GWAS had significantly higher statistical power than then bipolar disorder GWAS, and so it is possibly not surprising that the loci that were identified here were schizophrenia specific. If a GWAS of similar power was available for bipolar disorder, it is possible that loci specific to this disorder could also have been identified, as it would be highly unreasonable to suggest that the genetics of bipolar disorder consist exclusively of loci shared with schizophrenia. The final limitation of relevance to the work of this thesis is the fact that the method can only be applied to disorders, or subtypes of disorders, with a genetic correlation of < 0.8. This negates its use, currently, in sex stratified analyses where the genetic correlation is normally very close to 1, but also of highly related disorder subtypes, for example treatment-resistant and treatment-responsive schizophrenia. However, a CC-GWAS of TRS vs. non-TRS was attempted as an exploratory analysis due to the ready availability of data to do so, as well as to attempt to test and validate the CC-GWAS method itself. No significant results were produced, but the insights were key, and so it will be discussed briefly here.

CC-GWAS: TRS vs. Non-TRS?

The TRS samples were sourced from the CLOZUK1 and CLOZUK2 cohorts (described in detail in previous studies [96,146]), which are made up of individuals with schizophrenia prescribed clozapine in the UK. The prescription of clozapine was preceded by two failed trials of alternative antipsychotics, in line with the guidelines set out by the National Institute for Health and Care Excellence (NICE)[147]. The controls are the same set used in Pardiñas et al. [96], all of which were sourced from publicly available datasets or via collaborations with UK-based sequencing projects. In total, this GWAS contained 10,501 TRS cases, 24,542 controls and 5,998,190 SNPs.

The individuals in the non-TRS GWAS were collected from a subset of the studies used by the schizophrenia Working Group of the PGC in their meta-analysis undertaken in 2014 [95] . Individuals who could be relatively confidently identified as treatment resistant based on clinical records were removed from 34 studies. The control individuals were a combination of public datasets and clinically ascertained individuals (described in detail previously [95]). In total, there was 20,325 cases, 30,122 controls and 10,435,339 SNPs.

These GWAS contained no overlapping individuals, either cases or controls, although it is possible that some additional TRS samples remain in the non-TRS GWAS. This is due to the fact that TRS continues to be an under-reported condition, and so there could well be individuals who would not be filtered out based on a review of clinical records. It is also possible it contains people who are in-fact treatment resistant, but have not been designated as such yet, due to them still being trialled on other antipsychotic medications. This is a major consideration of the work outlined in research chapter 2 and will be discussed in detail there.

The lifetime disorder prevalence for non-TRS (0.72%) was based on the estimate put forward by McGrath et al. in 2008 [27], and the prevalence for TRS (0.24%) was calculated to reflect the general consensus that TRS individuals make up around 30% of all schizophrenia cases. The SNP-based heritability estimates (0.21 and 0.22 for non-TRS and TRS respectively) and the genetic correlation (0.96) were calculated using LD score regression via the 'ldsc' software [134]. The initial m number of 10,000 was selected based upon the recommendation of the CC-GWAS authors to use this number when assessing disorders that are thought to be highly polygenic. In follow up analyses, two different m parameters were also tested; a 'high' estimate of 55,000, based upon the estimate put forward in the paper accompanying the SBayesS method [148], and a 'low' estimate of 8,300 based upon the estimate put forward in the paper accompanying the MiXeR method [135]. This was to test what manipulation of the m number would do to the output of the CC-GWAS method.

In total, 2 potential candidate CC-GWAS SNPs were identified: rs144433536 and rs1800628. Rs144433536 is located within *TNXB*, a gene that has been associated with Ehlers-Danlos

Syndrome, and rs1800628 is within a regulatory region on chromosome 6 that has been mapped to both *TNF* and *LTB*. However, both variants were filtered out as being a potentially false association due to differential tagging of a stress test SNP (explained in the method overview). Filtering is applied only to the candidate SNPs, not genome-wide, and the steps used for the filtering are dependent on the power and size of the case-control GWAS that are being used. In this case, these 2 SNPs were filtered out because the power of the CC-GWAS was significantly lower than that of the power of the input case-control GWAS, reflected by the z-scores of the CC-GWAS analysis and the corresponding case-control z-scores.

Therefore, no significant associations with case-case status were found between non-TRS and TRS in these analyses. However, it was decided that the analysis should be repeated again with different selections of the m number. The m number was selected for change because whilst the rest of the input parameters could be robustly justified, either due to them being calculated directly from the summary statistics of the input GWAS or backed up by a reference, the m number of 10,000 was somewhat arbitrary. Two further m numbers were selected for use: a ‘high’ m of 55,000 and a ‘low’ m of 8,300.

Sensitivity Analysis: m number

The overall results remained unchanged; the same two SNPs were initially labelled as candidate CC-GWAS SNPs before being filtered out due to the lack of power of the method in comparison with the input case-control GWAS. However, there were some differences caused by the change in the m number. Firstly, the m number significantly affected the values of the OLS weights that were applied to the betas of the input GWAS (Table 10). The second difference was the genomic control inflation factor (lambda) based upon the CC-GWAS summary statistics. For m = 10,000, the lambda was 1.216, for m=8300, it was 1.189, and for 55,000, it was 1.454. The genomic inflation factor is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median and thus reflects the extent of bulk inflation. A higher genomic inflation factor is therefore indicative of a higher false positive rate. In this case, as the m number increases, the statistics are inflated proportionally.

	OLS WEIGHTS	
M NUMBER	**SCZ**	**TRS**
10,000	2.69e-02	-1.06e-01
8,300	3.86e-02	-1.19e-01
55,000	-8.87e-03	-3.32e-02

Table 10: The OLS weights calculated for each iteration of the TRS versus non-TRS CC-GWAS analysis, by inputted M number

The same inflation is observed when you use the CC-GWAS OLS weighted summary statistics to calculate observed scale heritability using LDSC. When using the m=8300, h^2obs = 0.0908, however when you use m = 55,000, h^2obs greatly increases to 0.2803. When using the delta method (+1 for non-TRS and -1 for TRS, irrespective of the input parameters), the genomic inflation factor is 1.062 and the total observed scale heritability is 0.0135. This is significantly lower than the h^2obs calculated from the OLS weighted summary statistics but is identical to the findings of Pardiñas and colleagues using the test for interaction developed by Altman and Bland [114]. The examination of the m number conducted here demonstrates that care should be taken when selecting the m number to use to conduct the analysis, and caution should be exercised when interpreting follow up analyses such as heritability estimates /.

genetic correlation analysis based on the OLS weighted summary statistics. It was as a result of this work that the m number used in the schizophrenia vs. bipolar disorder analysis was selected.

It was concluded at the end of this analysis that due to the extremely high genetic correlation (0.96) of TRS and non-TRS, it is not currently feasible or appropriate to use the CC-GWAS method to identify genetic differences between them. This would likely remain the case if larger, better powered GWAS became available, although it is not impossible that improvements in phenotypic quality in TRS and non-TRS GWAS could lead to a decrease in the observed genetic correlation between the subtypes. Simulations in the original paper showed that as the genetic correlation moves closer to 1, the type I error rate sharply increases, accompanied by a significant decrease in power. This is particularly true for the delta method, which was suggested specifically for use when subtypes of the same are being examined. Therefore, for now, to detect genetic differences between TRS and non-TRS, a direct case-case GWAS, leveraging individual data, remains the most valid option.

Chapter Conclusion

To conclude, using the newly published CC-GWAS method, 27 loci that were differentially associated between schizophrenia and bipolar disorder were identified. Following a series of follow-up analyses, a list of genes has been developed that, potentially, represent the sites in which common variation can occur, in combination with pleiotropic genetic variation shared with other psychiatric disorders, in order to significantly increase an individual's schizophrenia disease risk. The method itself has been rigorously tested, and compared to other methods, and its utility in a series of post-GWAS analyses has been investigated. However, it is only appropriate for use when the disorder pair has a genetic correlation of <

0.8. For subtypes of the same disorder with much higher levels of genetic correlation, the direct case-case GWAS utilising individual level genotype data remains the gold standard. So as this thesis continues, and the focus shifts from differentiating schizophrenia from other psychiatric disorders to stratifying schizophrenia into treatment resistant and responsive subtypes, that will be the method of choice.

Research Chapter 2: An International GWAS Meta-Analysis of Treatment Resistant Schizophrenia

Chapter Summary

For those with treatment-resistant schizophrenia (TRS), who represent some of the most severely affected patients in psychiatry, treatment options remain limited. Clozapine remains the single evidence-based medication for the treatment of TRS, and up to 50% of individuals with TRS do not gain adequate therapeutic benefit from clozapine. Genetics have the potential to reveal new insights into the neurobiology underlying TRS, yet genomic differences that differentiate those with TRS from individuals who respond to treatment (referred to throughout this thesis as non-TRS) have not yet been identified. One barrier to such insights has been the lack of characterisation of TRS within schizophrenia genomic studies. In a recent PGC schizophrenia analysis, due to data limitations at the time, an indirect approach testing for differences between separate case-control GWAS of TRS and non-TRS did not reveal specific genetic variants associated with treatment resistance. However, a polygenic signal for TRS explaining 1-4% of the variance was identified, representing the first time TRS had been demonstrated to have a detectable heritability. This suggests that TRS-specific common variants may exist, and so further investigation of this phenotype in a genetic setting is warranted. Here, as part of a subsequent PGC secondary analysis, the aim was to build on previous work by conducting a direct case-case GWAS of TRS vs. Non-TRS, utilising the increased sample size of the most recent 3rd wave from the Schizophrenia Working Group of the PGC. Genotype data from the group was used, as well as available additional samples from other collaborators. Phenotype information was collected for each individual dataset to allow treatment resistance status to be determined. With the exception of one cohort in which treatment status had been clinician defined, this was based on evidence of a lifetime prescription of clozapine or OPCRIT ratings of ‘response to neuroleptic drugs’. All data underwent stringent QC using a combination of RICOPILI and DRAGON-Data pipelines, and cohorts were merged based on genotyping array and sample ancestry. Following imputation against the HRC reference panel, a series of TRS vs. non-TRS case-case GWAS were conducted, followed by meta-analysis with an additional set of summary statistics prepared and supplied by the analysts of FinnGen. It was possible to amalgamate a total of just under 19,000 TRS cases, and just over 22,500 non-TRS controls. A single genome wide significant locus on chromosome 1 was identified and is the first genomic region identified as being specifically associated with TRS at genome wide significance to date. Based on SNP position, the nearest gene was the pseudogene FMO7P, the wider protein family of which is involved in the metabolism of a range of antipsychotics. Further analyses could lead to better elucidation of the aetiology of treatment-resistant schizophrenia, ultimately potentially leading to improvements in patient care.

Introduction

In the previous chapter, a new method, the CC-GWAS, was used to examine the genetic differences between schizophrenia and bipolar disorder, identifying 27 loci that were differentially associated with the two correlated disorders. It was also employed in an attempt to identify common genetic variation differentially associated between two subtypes of schizophrenia, treatment-resistant schizophrenia (TRS) and non-treatment-resistant schizophrenia (referred to throughout this thesis as non-TRS). However, the CC-GWAS method can have inflated error rates in pairs of disorders/subtypes with a genetic correlation higher than 0.8. As such, it was not appropriate to implement the method when investigating TRS and non-TRS, the genetic correlation of which is likely to surpass this threshold, based on current evidence [114]. In situations of high genetic correlation, a direct case-case GWAS, requiring access to individual level genotype data, remains the gold standard, and that is the aim of the present chapter of this thesis.

Treatment-Resistant Schizophrenia

TRS is most widely defined as a failure of symptoms to respond to at least two antipsychotic medications, when prescribed at an adequate dose for sufficient duration for a therapeutic response [149]. Failure to respond is often defined as a lack of improvement in positive symptoms, although negative and cognitive symptoms can also be considered. It is estimated that 20-30% of patients with schizophrenia experience treatment resistance [150], with a recent review of first-episode psychosis cohorts totalling over 12,000 cases reporting a TRS prevalence of 24.8% [151]. Although the majority of treatment resistance seems to be present from disease onset, it is also possible for patients to develop TRS several years into antipsychotic treatment [152]. TRS has been found to be associated with a range of markers of poor outcome including higher levels of unemployment and more frequent hospitalisations, and are more likely to have a range of physical/psychiatric comorbidities [153]. In addition a review of 65 studies of clinical, social and economic associations with TRS identified high rates of smoking, alcohol and substance abuse, and suicidal ideation [154]. The same paper estimated that TRS is associated with 3-11 times greater annual costs than schizophrenia with symptomatic remission.

For individuals with TRS, there is a single licensed treatment option; the atypical antipsychotic clozapine. However, clozapine is associated with a wide range of adverse effects, and discontinuation rates are high. In a retrospective cohort study of 316 patients with TRS receiving their first course of clozapine, 45% of patients had discontinued clozapine within two years of initiation [155]. Another larger study utilising Finnish registry data found a similar rate, with 49.1% of 7037 patients receiving clozapine monotherapy discontinuing the regimen within a year [156]. Options for TRS that doesn't respond to clozapine (sometimes referred to as ultra-treatment-resistant schizophrenia), or individuals who have to discontinue because of issues of tolerability, are limited at this time. In fact, in cases of intolerability, it is often considered most effective to cautiously recommence clozapine therapy whilst monitoring its tolerability, except in the case of serious adverse effects such as agranulocytosis [156]. For cases of ultra TRS, or where clozapine cannot be reinitiated, little high quality evidence exists for any treatment regimen, although augmentation of antipsychotic treatment with electroconvulsive therapy (ECT) appears to be best supported by current evidence [157]. Complicating this matter further is the fact that the current understanding of the aetiology of TRS, and how it differs from schizophrenia that does respond to antipsychotic medication besides clozapine, is very limited. Indeed, there remains uncertainty as to whether TRS is a categorically distinct subtype of schizophrenia, or represents a more severe course of illness [158]. The identification of genetic variants or biomarkers specific to TRS has the potential to greatly improve our understanding of the condition and could also help guide consensus on whether TRS is its own distinct disorder subtype. Additionally, it could lead to earlier prediction and intervention of TRS and provide insights into possible new treatment options.

Neurobiology of Treatment Resistant Schizophrenia: Current Understanding

There is a limited literature around the basis and aetiology of TRS. The lack of consistent findings has likely arisen in part due to a historic lack of consensus regarding the definition of treatment resistance, as well as restricted sample sizes and heterogeneity of study designs [159]. One theory posits that TRS represents a subgroup of patients without the characteristic aberrations of the dopaminergic system classically associated with schizophrenia, and as such do not respond to antipsychotic treatment in the same way as individuals with non-TRS [160]. Another recent review of the literature concluded that TRS appears to be characterised by relatively normal dopaminergic transmission, aberrant glutamatergic signalling, and significant decreases in grey matter volume [158]. A small study of 71 Canadian patients concluded that first and second degree relatives of individuals with TRS had a significantly higher morbidity risk of schizophrenia spectrum disorders (MR=8.85) compared to relatives of patients with non-TRS (MR=2.45), and a significantly higher familial-loading score [161]. Candidate gene studies of TRS have provided no consistent replicable findings of genes differentially associated between patients with TRS and non-TRS, due primarily to small sample sizes, although genes such as *BDNF* [162], *5-HT2A* and *TPH1* [163], and DRD3 [164] have been implicated in small scale studies. Need and colleagues conducted an investigation of 2769 polymorphisms in 118 candidate genes utilising the data collected as part of the CATIE study [165], in which no association survived correction for multiple testing.

A small number of GWAS of TRS have been conducted to date, with limited success, again due to small sample sizes. A GWAS of 84 TRS patients reported no significant associations

[166], in both a GWAS and an interaction model with childhood trauma. A GWAS of 795 Han Chinese individuals with TRS versus 806 controls also reported no genome wide significant results but did observe nominal associations with variants located in the *RIPK4* and *NFKB1* genes [167]. A study of 79 TRS and 95 non-TRS schizophrenia patients reported a nominally significant association with a locus 70 kb upstream of L-dopa decarboxylase (*DDC*) [168]. Dopamine decarboxylase catalyses the decarboxylation of L-3,4-dihydroxyphenylalanine (DOPA) to dopamine, the primary neurotransmitter implicated in schizophrenia.

Finally, a study in 2022 with a total sample size of just under 85,500 people, conducted two case-control GWAS, one with a case cohort of TRS individuals, the other with non-TRS individuals, and then used a test for interaction [169] to quantify the differences in effect size of common genetic variants across the two GWAS [114]. Whilst this paper did not identify any specific genome wide significant results, it did demonstrate that treatment resistance in schizophrenia is a polygenic trait of detectable heritability and found that the summary statistics from the interaction analysis were significantly genetically correlated with numerous measures of intelligence, cognition and smoking behaviour. The treatment resistant phenotype was negatively associated with cognitive performance and educational attainment, and positively associated with measures of smoking behaviour. PRS calculated from the interaction results were also significantly positively associated with a history of taking clozapine in both prevalence and incidence schizophrenia cohorts. The conclusion of this paper was that TRS specific common genetic variants exist, but historically these associations have likely been concealed through the amalgamation of large GWAS samples of schizophrenia, in a drive for increasing sample sizes. A large-scale, meta-analytic study of TRS, leveraging the power of methods that rely on individual level data access, has to date not been conducted, and has the potential to identify genome-wide significant associations specific to TRS for the first time.

Chapter Aims and Hypotheses

The primary aim of this chapter is to take advantage of the large number of schizophrenia samples available through collaboration with the Schizophrenia Working Group of the PGC to conduct a direct, case-case GWAS of TRS versus non-TRS. Due to many cohorts not containing both TRS and non-TRS together in sufficient numbers to be analysed separately, a stringent, but not overly conservative, quality control procedure was implemented to allow for independent cohorts of schizophrenia cases to be combined, in line with work that has been completed in other disorders [170]. The combined cohorts were then be imputed, and treatment status defined for all participants with phenotypic information collected from the principal investigators of the schizophrenia working group of the PGC. Once treatment status had been defined, primarily based on evidence of clozapine prescription, association testing was completed on each of the assembled cohorts, followed by a meta-analysis.

Methods

All analyses outlined in this chapter were conducted on the LISA server (https://www.surf.nl/en/lisa-compute-cluster-extra-processing-power-for-research), the HPC system utilised by the PGC for all primary and secondary analyses. Visualisations were created using RV4.0.1. All work, besides the acknowledgements stated at the start of this thesis, was conducted by me under the supervision of Professor Walters and Dr Pardiñas, including all quality control, data merging, imputation, association testing, meta-analysis and post-GWAS interrogation.

Samples

In total, 41 datasets were collated for this analysis. This included 38 cohorts that had been collected as part of the collaborative efforts of the schizophrenia working group of the PGC, two new cohorts from Cardiff University that are independent of the PGC (CLOZUK3 and NCMH), and an additional cohort provided by colleagues from the University of Oslo (CLOZNOR). Summary statistics for another cohort, FinnGen, were also provided for use in this analysis, the GWAS having been conducted by Dr. Anders Kämpe. A full breakdown of the cohorts used in this analysis can be seen in table 11.

DATASET	*PGC DATASET CODE*	*ARRAY*	*TRS*	*NON.TRS*
ABERDEEN	scz_xaber_eur_sr-qc	A6.0	227	1544
ASRB	scz_xasrb_eur_sr-qc	I650	107	475
BERLIN	scz_bep1b	GSA	71	262
BOLOGNA	scz_serri_eur_sr-qc	PSYC	31	112
BOSTON, US (CIDAR)	scz_xcims_eur_sr-qc	OMEX	0	71
CLOZNOR	NA	GSA	143	875
CLOZUK	scz_xclm2_eur_sr-qc	I1M	3466	0
CLOZUK	scz_xclo3_eur_sr-qc	omni	2150	0
CLOZUK	scz_clz2a_eur_sr-qc	OMEX	5370	0
CLOZUK	NA	GSA	1438	0
COGS	scz_cgs1c_eur_sr-qc	OMEX	121	283
COGS	scz_xcou3_eur_sr-qc	omni	178	270
DENMARK	scz_xdenm_eur_sr-qc	I650	105	387
EDINBURGH	scz_xedin_eur_sr-qc	A6.0	0	368
FINNGEN	NA	-	2704	3811
F-SERIES AND SIB PAIRS	scz_xcaws_eur_sr-qc	A500	111	244
GERMANY	scz_xboco_eur_sr-qc	I550	289	1558
IRELAND	scz_xdubl_eur_sr-qc	A6.0	38	234
IRELAND	scz_xirwt_eur_sr-qc	A6.0	78	1222
ISRAEL	scz_xajsz_eur_sr-qc	omni	226	956
LUBECK	scz_geba1_eur_sr-qc	PSYC	372	0
MUNICH, GERMANY	scz_xmunc_eur_sr-qc	I317	166	271
NETHERLANDS	scz_xucla_eur_sr-qc	I550	0	705

NEW YORK, US, ISRAEL	scz_xmsaf_eur_sr-qc	A6.0	0	327
NEW YORK, US	scz_xzhh1_eur_sr-qc	A500	0	191
NIMH CBDB	scz_xlie2_eur_sr-qc	OMEX	0	137
NIMH CBDB	scz_xlie5_eur_sr-qc	I550	0	509
PEIC	scz_xpews_eur_sr-qc	A6.0	0	82
PEIC	scz_xpewb_eur_sr-qc	A6.0	0	597
PORTUGAL	scz_xport_eur_sr-qc	A6.0	22	328
SÃO PAULO	scz_sb2aa_eur_sr-qc	OMEX	112	142
SIX COUNTRIES	scz_xlacw_eur_sr-qc	I550	0	157
SWEDEN	scz_xersw_eur_sr-qc	omni	0	322
SWEDEN	scz_xswe1_eur_sr-qc	A5.0	60	161
SWEDEN	scz_xs234_eur_sr-qc	A6.0	402	1675
SWEDEN	scz_xswe5_eur_sr-qc	omni	433	1368
SWEDEN	scz_xswe6_eur_sr-qc	omni	228	865
TOP	scz_to10c_eur_sr-qc	OMEX	223	95
TOP	scz_xtop8_eur_sr-qc	A6.0	25	351
UCL	scz_xuclo_eur_sr-qc	A6.0	134	386
US (CATIE)	scz_xcati_eur_sr-qc	A500	0	409
			18,979	**22,523**

Table 11: Full list of all samples collected for this analysis, including information about their genotyping array, the number of TRS individuals, and the number of non-TRS individuals

Quality Control

All samples were subjected to the same quality control procedures, with the exception of FinnGen, which was supplied as summary statistics (an overview of the QC applied to FinnGen can be seen here: https://finngen.gitbook.io/documentation/methods/phewas). The QC was primarily performed using a combination of two preimputation QC pipelines, as well as some additional work undertaken using plink2 [171]. The two pipelines were RICOPILI (Rapid Imputation for COnsortias PIpeLIne), which is the analytic pipeline used by PGC working groups [136], and Dragon-Data [172], a pipeline developed at Cardiff University by colleagues within the department. Three rounds of QC were completed in total: on the individual batches of raw genotype data for each cohort, again following the combining of batches into their respective cohort, and again following the merging together of cohorts by genotype array into the final GWAS datasets. An outline of each step is given below.

Determination of Genotype Array

The array that each cohort had been genotyped on was determined using Dragon-Data, which compares the SNP coordinates to a reference panel of 391 genotyping platforms for the highest degree of overlap. The 391 arrays were taken from the Chipendium site (http://mccarthy.well.ox.ac.uk/chipendium/ui/, not currently available). RICOPILI also supplies a best guess genotyping platform, but the samples under study are only compared to a reference panel of 10 genotyping platforms, thus the Dragon-Data pipeline was used to allow for a better elucidation of the platforms.

Genotype Harmonisation

This step was conducted using the Dragon-Data pipeline, which utilises the 'Genotype Harmonizer' (GH) software [173] to resolve strand alignment, discordant alleles and coordinate mismatches for SNPs in common between the dataset and the imputation reference panel that will be used for imputation of the samples. In this case, that was the HRC1.1 reference panel [83]. This aids in the retention of the maximum possible number of SNPs, in order to boost the overall quality of imputation.

SNP / Sample QC

This was conducted via RICOPILI. Only samples and variants with a missing rate of < 0.02 were retained for analysis. By default, RICOPILI removes monomorphic SNPs, but this not completed for this analysis, in an attempt to maximise the number of overlapping SNPs between cohorts and so improve the quality of imputation. It has also been argued that the removal of monomorphic SNPs is not optimal in a meta-analytic setting, due to the fact that a SNP that is monomorphic in one study may not be so in others [174]. In addition, the default exclusion of any SNP with an MAF of < 0.01 was also removed, as it led to the excessive removal of SNPs from cohorts that were genotyped on arrays designed to capture rare variation. In addition, as a final QC step immediately prior to imputation, a minor allele count (MAC) threshold of 40 was applied, based on previous literature recommending MAC as a more appropriate parameter to threshold on versus MAF, as MAC tends to be more stable at lower sample sizes [80].

Heterozygosity

Plink2 was utilised for the calculation of methods-of-moments F coefficient estimates for each individual. The –het flag was used to compute the observed and expected homozygous/heterozygous genotype counts, and individuals whose F coefficient fell outside of the mean +/- 2.33 * the standard deviation of the cohort were excluded i.e., the most extreme 1% of values at either side of the distribution are omitted. Excessive levels of genome-wide heterozygosity or homozygosity can be indicative of poor-quality DNA or sample contamination, but homozygosity above expected can also be evidence of inbreeding within a population. This approach to heterozygosity was selected to avoid being overly conservative, in line with a recent paper that highlighted the platform and sample specific nature of the heterozygosity metric, recommending that thresholding based on the distribution of the measure within the cohort specifically was advantageous over utilising a single cut-off value [175].

Hardy Weinberg Equilibrium

Major deviations from HWE, particularly in cohorts of relatively homogenous genetic ancestry, can be indicative of errors in genotyping. However, genuine SNP-trait associations can also be expected to deviate from HWE to a certain extent, and so it is important to select a threshold that is not overly stringent. In this case, the selected threshold was 1x10E-6, and Plink2 was used to allow for the use of the '–midp' and '–keep-fewhet' options. The use of a mid p-value for HWE filtering has been demonstrated to have a lower type 1 error rate versus the use of standard two-sided p-values, as well as being better powered [176]. The '—keep-fewhet' modifier is used to preferentially retain variants that fail in the 'too-few-heterozygotes' direction, as this can be expected to occur in samples where population stratification is present. The modifier causes the threshold to only be applied in situations where the number of heterozygous genotypes for a variant is above the equilibrium value, and not below.

Sample Merging

Cohorts were merged using RICOPILI based on genotyping platform, to maximise the overlapping set of SNPs available for imputation. Due to the high number of cohorts that had been genotyped using Illumina platforms, these were separated into two groups. As a result of being genotyped on a range of different iterations of Omni Express array chips, combining them into a single group led to the overlap of a lower number of SNPs. If a cohort was the only one genotyped on a particular array, or there were not sufficient cohorts to generate a final dataset containing a reasonable number of individuals with TRS and non-TRS, they were merged with the batch that led to the lowest level of variant dropout, with the aim of retaining a minimum of 250,000 SNPs shared between cohorts for imputation.

One dataset, ‘xgras’, was the only dataset genotyped on an Axiom array, and led to an excessive level of SNP dropout (< 80,000 SNPs retained across all cohorts) irrespective of the cohorts it was merged with, and thus was excluded from further analysis.

Sample Ancestry, Kinship and PCA

Relatedness was calculated in plink2 using the KING algorithm [177], which allows for the accurate estimation of kinship coefficients between each pair in the sample even in the presence of population stratification. This was an improvement on existing methods of relatedness inference, for which the main assumption was homogenous population structure. Kinship coefficients calculated in KING are scaled as such that 0.5 would be indicative of a sample pair being monozygotic twins/ a duplicate, 0.25 first-degree relatives etc. For screening, it is then the recommendation to use the geometric mean between two thresholds as the cut-off value. A cut-off of 0.044, analogous to screening for fourth-degree relatives, was used in this case and one sample from each pair was then randomly selected for exclusion.

Linear Discriminant Analysis (LDA) modelling based on ancestry informative markers (AIMs) was used for the prediction of biogeographic ancestry in all of the samples with the exception of FinnGen. The genotype datasets were merged with a reference panel based on the Allen Ancient DNA Resource (AADR) (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data, version 50.0). This panel contains just over 1.2 million variants, and the AIMs panel was derived in plink by calculating an Fst statistic for each of these SNPs in all pairwise comparisons of ancestral populations and retaining those SNPs with the top 2.5% Fst in every comparison. The higher the Fst metric, the more differentiated the SNP is between population subgroups, and thus can be used to predict genetic ancestry. The overlapping SNPs between the genotype datasets and the top 2.5% SNPs from the AADR were then used as the AIMs for each of the eight datasets generated in the step outlined above.

Principal components were generated using the PC-AiR method [178] implemented in the 'GENESIS' R package, which utilises the kinship coefficients derived in the previous step to calculate principal components that are unaffected by family structure [179]. These principal components are calculated based on the AIMs derived above, and are then used in the LDA model to predict biogeographic ancestry, following methods previously outlined by our group[180]. The biogeographic groups are equivalent to those outlined by Huddart and colleagues for use in pharmacogenetics research [181]. There are nine in total: Americans, Central/South Asians, East Asians, Europeans, Near Easterns, Oceanians, and Sub-Saharan Africans, African Americans/Afro-Caribbeans and Latinos. The model is first trained on the AADR before being used on the target genotype datasets, and results in a set of probabilities for each individual of belonging to each of the nine biogeographic groups. If an individual had an LDA probability that surpassed 80%, they were assigned to that specific biogeographic group. If they did not surpass the 80% in any single group, they were categorised as 'admixed/unknown'.

Imputation

Imputation was conducted via RICOPILI, using the full HRC1.1 data as a reference panel [83]. Prephasing was conducted using Eagle v2.3.5 [182], and the imputation was conducted using Minimac3 [183]. The default post-imputation QC parameters provided by RICOPILI were utilised: MAF threshold 0.005, INFO score threshold 0.1 and genotype probability for making a best guess genotype call of 0.8. To boost the quality of imputation, the corresponding healthy controls for each cohort were retained at this stage, and removed once the dataset had been imputed, based on the findings of other large-scale consortia who have brought together and imputed previously genotyped data from a large number of research centres [184].

Phenotype Definition

To define treatment resistance within the samples, the PIs of the schizophrenia working group of the PGC were approached individually and asked to provide the following five variables:

1. Clozapine prescription information (Lifetime, Current etc.)
2. OPCRIT item 89 "Psychotic symptoms respond to neuroleptics".
    a. An alternative objective rating equivalent to OPCRIT item 89 (based on note review or clinician report) was also accepted.
3. Antipsychotic medication history – number, duration and types etc.
4. Age at onset and age at interview, in order to calculate duration of illness.
5. Diagnosis (ICD/DSM codes etc.)

Items 1 and 2 were collected to define TRS cases within each sample. Items 3 and 4 have not been utilised as of yet but were collected with the view of using it to define further TRS cases who may have not yet been prescribed clozapine. Many people are unable to tolerate clozapine or are not given the opportunity to take it by their clinician, and so medication/duration of illness information may allow for additional people with TRS to be identified and reclassified. Medication information was highly variable across cohorts, but included information about typical / atypical antipsychotics, depot injections, dosage information and treatment length information. Item 5 was collected to allow the cohorts to be restricted to just individuals with schizophrenia and schizoaffective disorder depressed subtype. A breakdown of what information was used to define treatment resistance within each sample can be seen in Table 12. In the four cohorts where both clozapine and OPCRIT89 were available, OPCRIT89 was used to identify additional individuals classified as TRS who had not been prescribed clozapine at the time of data collection. A small subset of datasets also contained information regarding response to clozapine, but this has also not been utilised at this time. TRS individuals were assigned a phenotype value of 2 (cases) and non-TRS individuals were assigned a 1 (controls).

DATASET	*PGC DATASET CODE*	*TRS DEFINED BY:*
ABERDEEN	scz_xaber_eur_sr-qc	CLOZAPINE
ASRB	scz_xasrb_eur_sr-qc	CLOZAPINE/OPCRIT89
BERLIN	scz_bep1b	CLOZAPINE
BOLOGNA	scz_serri_eur_sr-qc	CLOZAPINE/OPCRIT89
BOSTON, US (CIDAR)	scz_xcims_eur_sr-qc	NA
CLOZNOR	NA	CLOZAPINE
CLOZUK	scz_xclm2_eur_sr-qc	CLOZAPINE
CLOZUK	scz_xclo3_eur_sr-qc	CLOZAPINE
CLOZUK	scz_clz2a_eur_sr-qc	CLOZAPINE
CLOZUK	NA	CLOZAPINE
COGS	scz_cgs1c_eur_sr-qc	CLOZAPINE/OPCRIT89
COGS	scz_xcou3_eur_sr-qc	CLOZAPINE/OPCRIT89
DENMARK	scz_xdenm_eur_sr-qc	CLOZAPINE
EDINBURGH	scz_xedin_eur_sr-qc	NA
FINNGEN	NA	CLOZAPINE
F-SERIES AND SIB PAIRS	scz_xcaws_eur_sr-qc	OPCRIT89
GERMANY	scz_xboco_eur_sr-qc	CLOZAPINE

IRELAND	scz_xdubl_eur_sr-qc	CLOZAPINE
IRELAND	scz_xirwt_eur_sr-qc	CLOZAPINE
ISRAEL	scz_xajsz_eur_sr-qc	CLOZAPINE
LUBECK	scz_geba1_eur_sr-qc	CLINICIAN DEFINED
MUNICH, GERMANY	scz_xmunc_eur_sr-qc	CLOZAPINE
NETHERLANDS	scz_xucla_eur_sr-qc	NA
NEW YORK, US, ISRAEL	scz_xmsaf_eur_sr-qc	NA
NEW YORK, US	scz_xzhh1_eur_sr-qc	NA
NIMH CBDB	scz_xlie2_eur_sr-qc	NA
NIMH CBDB	scz_xlie5_eur_sr-qc	NA
PEIC	scz_xpews_eur_sr-qc	NA
PEIC	scz_xpewb_eur_sr-qc	NA
PORTUGAL	scz_xport_eur_sr-qc	OPCRIT89
SÃO PAULO	scz_sb2aa_eur_sr-qc	CLOZAPINE/IPAP
SIX COUNTRIES	scz_xlacw_eur_sr-qc	NA
SWEDEN	scz_xersw_eur_sr-qc	NA
SWEDEN	scz_xswe1_eur_sr-qc	CLOZAPINE
SWEDEN	scz_xs234_eur_sr-qc	CLOZAPINE
SWEDEN	scz_xswe5_eur_sr-qc	CLOZAPINE
SWEDEN	scz_xswe6_eur_sr-qc	CLOZAPINE
TOP	scz_to10c_eur_sr-qc	CLOZAPINE
TOP	scz_xtop8_eur_sr-qc	CLOZAPINE
UCL	scz_xuclo_eur_sr-qc	CLOZAPINE
US (CATIE)	scz_xcati_eur_sr-qc	NA

Table 12: A list of how treatment status was defined within each sample. 'NA' denotes that there is currently no available phenotype information for this sample

Association Testing and Meta Analysis

Association testing was completed in plink2. Eight separate TRS versus non-TRS GWAS, using an additive logistic model, were conducted, with principal components 1-15 included as covariates. In addition, the ancestry probabilities calculated as part of the ancestry inference section described above were included, in an attempt to better account for population stratification within the array batches. In line with the supplementary methods outlined by our group [185], the Europeans ancestry probability was omitted, in order to avoid collinearity with the regression intercept, which causes the association testing procedure to fail in plink2 due to regression models not being viable if the inputted predictor variables are colinear [186]. The meta-analysis of the eight GWAS and the summary statistics provided by FinnGen was conducted in plink2, using a standard error inverse-weight fixed effects model, in line with the latest GWAS from the schizophrenia working group of the PGC [98].

Further Refinement of the Samples

For 14 of the cohorts, totalling 4,611 samples, no phenotypic information was available, and in the group's previous research they had been utilised as non-TRS samples [114]. This was done to maximise the sample size, with the caveat that a significant proportion of those individuals will have been misclassified. The effect of their inclusion had been modelled to improve overall power when using realistic levels of misclassification due to the improvement in overall sample size. Examination of the cohorts for which phenotype information was available showed that there were often rates of TRS higher than would be expected based on estimates of TRS prevalence. To reduce the impact of misclassification on the results, and taking advantage of the boosted sample size, the decision was made to remove as many of the 'unknown' phenotype samples as possible. This led to a total loss of 2179 non-TRS cases from the meta-analysis. It was not possible to remove all 14 of these cohorts at this time, as the removal of the remaining 7 would have led to almost half of the TRS cases having no matching non-TRS controls, and thus remain in the analysis currently. It is hoped that phenotypic information will be collected for these cohorts prior to a future data freeze, allowing all samples to be utilised and accurately classified.

SNP-Based Heritability and Genetic Correlation

SNP-based heritability on the observed and liability scales were calculated via LD score regression using the LDSC software V1.0.1 [134]. In all instances, the LD reference was the European ancestry-specific data from phase 3 of 1000 Genomes [82]. Prior to analysis, SNPs with an INFO score < 0.9 were excluded, and the datasets were trimmed to contain only those that are present in the third phase of the International HapMap Project [138]. This is a reference set of 1,440,616 SNPs genotyped in 1,184 individuals from 11 global populations. For the liability scale, a population prevalence for TRS of 0.3 (in relation to the controls being schizophrenia cases in this context) and a sample prevalence of 0.5 were used. The genetic correlation was also calculated using the ldsc software, using pre-computed LD Scores based on the 1000 genomes dataset described above. The summary statistics from this analysis were compared to both the most recent GWAS from the schizophrenia working group of the PGC [98] and the interaction analysis [114].

LD-Based Clumping of GWAS Results

The LD-based clumping of loci was conducted via PLINK V1.9 [171]. The LD reference used was the European ancestry-specific dataset available from phase 3 of 1000 genomes [82], with a gene locations list based on GRCh37 (accessed and downloaded via: https://www.cog-genomics.org/static/bin/plink/glist-hg19). The physical distance threshold for clumping was set to 3000kb, the significance threshold for index SNPs was set to $p < 1e\text{-}04$ and the LD threshold for clumping was $r^2=0.1$. A locus was considered to be significant if the p-value was less than 5E-08.

Results

The final meta-analysis was conducted on a total of 34 datasets (FinnGen, CLOZUK3, CLOZNOR and 31 PGC European datasets). Phenotypic information was used to classify 18,979 TRS cases (see Table 12), and 20,344 non-TRS cases. Currently, the non-TRS samples contain 2,432 individuals of 'unknown' phenotype. The total final sample size was 39,323 cases with schizophrenia (Figure 5).

The genomic inflation factor for the final meta-analysis was calculated to be 1.009. Table 13 presents the lambda for each input GWAS separately. The QC procedure conducted during this work appears to have effectively controlled for population stratification, as well as other common causes of genomic inflation, signified by the low lambda of the meta-analysis, and each of the input GWAS. The LDSC intercept was calculated to be 1.0218.

GWAS	LAMBDA
GSA	1.038
PSYC	1.055
I550	1.014
I650	1.016
A600	1.017
A500	1.037
OMEXA	1.026
OMEXB	1.037
FINNGEN	1.047
TOTAL	1.009

Table 13 The genomic inflation factor (lambda) of each input GWAS separately, plus for the final meta-analysis (TOTAL)

In total, there were five SNPs that surpassed genome-wide significance (Table 14), all of which were located within the same region of chromosome 1 (Figure 6). The most significant SNP was rs7549089, with an OR of 1.1415 and P-value of 3.45E-08.

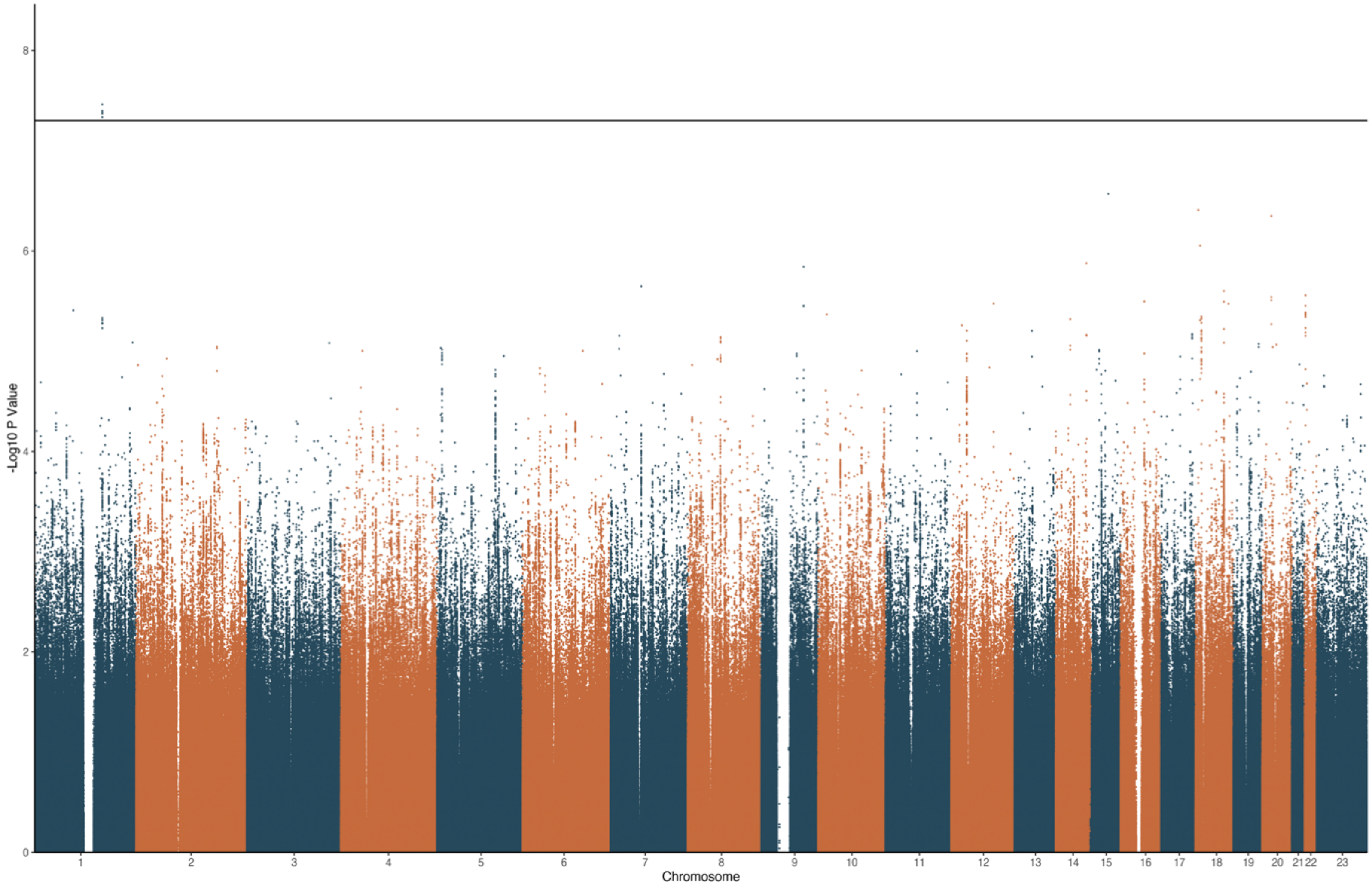

Figure 5: Manhattan plot of TRS versus non-TRS case-case GWAS, N= 39,323

CHR	BP	SNP	A1	A2	N	P	OR
1	166431020	rs7549089	C	T	9	3.45E-08	1.1415
1	166430484	rs10918471	G	A	9	4.00E-08	1.1408
1	166432463	rs1908312	T	A	9	4.17E-08	1.1408
1	166436454	rs977513	G	C	9	4.26E-08	1.1408
1	166431463	rs10753733	A	G	9	4.63E-08	1.1401

Table 14 GWAS results for the five GWS SNPs. N= The number of input GWAS that the SNP was present in (out of a possible maximum of 9), A1=Effect Allele that the OR/P were calculated in relation to, A2=Reference Allele)

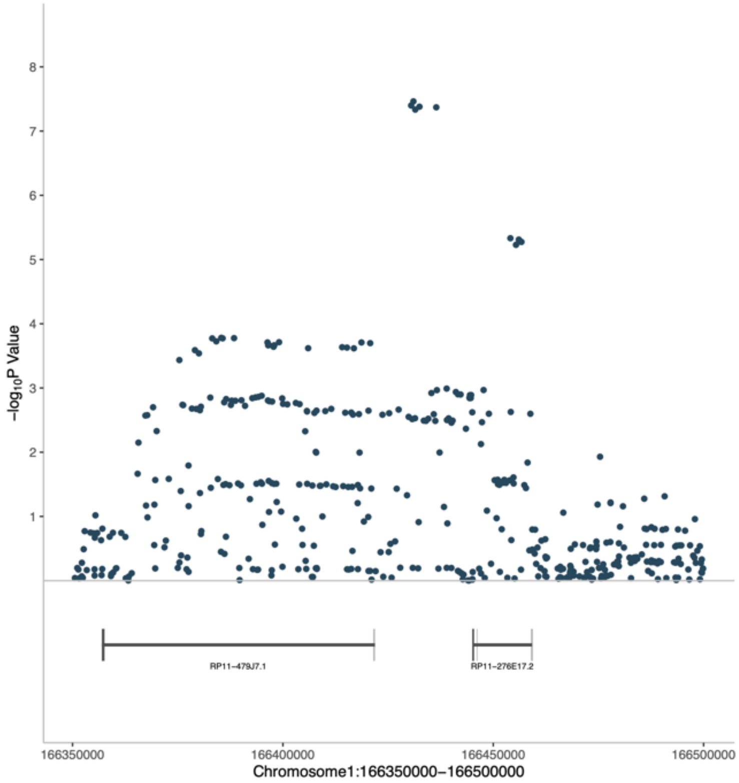

Figure 6: Regional Association plot for the genome-wide significant locus

LD-Based Clumping Procedure

LD-Based clumping of the results identified 454 loci with a P<1E-04. The top ten loci can be seen in Table 15. For four of these, the BP boundaries of the locus contain a single protein-coding gene, and a further four are intergenic.

CHR	SNP	P	POS START	POS END	GENES WITHIN LOCUS
1	rs7549089	3.45E-08	166430484	166456704	-
15	rs3784351	2.69E-07	68642220	68642220	ITGA11
18	rs148108347	3.90E-07	5956042	5956042	L3MBTL4
20	rs11907443	4.49E-07	19394130	19404850	SLC24A3
18	rs10502392	8.83E-07	9762049	9767615	RAB31
14	rs885845	1.33E-06	96728916	96732345	ATG2B, BDKRB1, BDKRB2
9	rs140994521	1.44E-06	105042198	105307824	LINC00587
18	rs12232766	2.51E-06	58445895	58670832	-
22	rs4819826	2.77E-06	19640282	19642645	-
16	rs10221167	3.19E-06	52947414	52969512	-

Table 15: LD-based clumping results for the top ten loci, including gene information

SNP-Based Heritability and Genetic Correlation Analysis

The SNP-based heritability of the TRS GWAS on the observed scale was calculated to be -0.007, and -0.0102 on the liability scale. Due to the heritability being calculated as negative, genetic correlation analysis cannot be conducted in ldsc. The negative heritability can be indicative of a lack of genetic signal but can also occur as a result of a trait having an oligogenic pattern of inheritance, with a much smaller number of genes contributing to the phenotype versus a polygenic trait. It is not unfeasible that the TRS-specific genetic signal, which this GWAS was attempting to capture, is contained within a small number of genomic loci.

Discussion

In this chapter, a major collaboration with the schizophrenia working group of the PGC was undertaken to collect the phenotypic and genetic data required to perform a direct case-case GWAS of TRS versus non-TRS. Such an analysis remains the gold standard option for subtypes of disorders with very high genetic correlation, but to date has been difficult to perform to date because of limited data availability. Here, it was possible to collect genetic and phenotypic information for over 40,000 schizophrenia cases, resulting in a final refined sample size of 18,979 cases and 20,344 controls. A quality control procedure was developed in the process of this work, utilising two previously published QC pipelines and additional work in plink2 and R. All genetic datasets underwent 3 rounds of QC and were merged together based on genotyping array to form case-case cohorts containing individuals with both TRS and non-TRS. The datasets were then imputed against the HRC1.1 reference panel before undergoing association testing and a final meta-analysis. There was no evidence of genomic inflation of the results, in the whole meta-analysis or any of the separate GWAS, signalling that the QC procedure developed in this chapter was effective at controlling for common sources of genomic inflation. Here, for the first time, a genome wide significant association with TRS was identified, as were a small number of sub genome-wide significant loci of potential biological interest. These will be discussed below.

Associated Loci

The biological relevance of the genome wide significant locus identified in this analysis, an intergenic region on chromosome 1 containing five variants surpassing P<5E-08, is difficult to elucidate. It does not overlap with any of the loci from the most recent GWAS from the PGC and it is ~ 292,000 base pairs from the closest protein-coding gene (*FAM7B*). There is however a pseudogene contained within this area (*FMO7P*) in addition to a long intergenic non-protein coding RNA (*LINC01675*). Flavin Containing Dimethylaniline Monooxygenase 7 is what is known as an unprocessed pseudogene, meaning that it originated as a result of the aggregation of mutations during gene duplication ultimately rendering the gene untranslatable. Humans are thought to have five 'functional' FMO genes, which play a key role in the metabolism of a wide range of medications, including antipsychotics [187]. For example, *FMO3* has been implicated in the N-oxygenation of clozapine [188,189], olanzapine [190] and loxapine [191] into their respective N-oxide metabolites. FMO3 is also thought to be involved in the N-oxygenation of nicotine [192]. *FMO7P*, in conjunction with *FAM78B*, was attributed to a genome-wide significant locus in a GWAS of unipolar depression of East Asian individuals [193], as well as a genome-wide significant locus in a GWAS of educational attainment in combination with *LINC01675* [194].

Historically considered as 'junk DNA' with limited biological relevance, there is growing evidence to support the role of pseudogenes in a range of disorders and diseases [195-197]. A review by Cheetham and colleagues outlined a number of ways in which pseudogenes could have biological relevance, including their translation into functional full-length or truncated proteins, their actions as inhibitors of the translation of their parental genes, their manipulation of 3D chromatin interactions to regulate parental gene expression, and their transfer of pathogenic alleles to their parental genes via the process of gene conversion [197]. Further analysis will be required to elucidate the functional relevance of this locus to treatment resistance but given the documented role of FMO's in the metabolism of antipsychotics, and the suspected biological relevance of pseudogenes in disorder states, this finding is potentially very interesting, and warrants continued investigation.

Very little is known about the function of the protein encoded by family with sequence similarity 87 member B (*FAM78B*). The gene itself has been implicated in GWAS of a number of potentially relevant phenotypes. For example, it has been implicated in multiple large-scale GWAS of cortical thickness [198,199], and was also located within a genome-wide significant locus in a combined MTAG [122] and GWAS study of educational attainment and math ability in 1.1 million individuals [200]. Additionally, it was within one of three loci that surpassed genome-wide significance in a GWAS study of time to cocaine dependency from first use [201]. Each minor allele of the lead variant was associated with 0.57 fewer years to dependency from first use. Finally, the most significant association containing this gene identified to date was found in GWAS of white blood cell count and neutrophil count in a multi-ancestry GWAS of 64,784 individuals from the PAGE study [202]. The locus containing this gene was negatively associated with both neutrophils and whole white blood cell count.

Although only one locus surpassed genome-wide significance, it is expected that further loci will be identified as sample sizes increase over subsequent data freezes and the definitions of TRS and non-TRS become more refined. The most likely loci to benefit from the resulting increase in power are those just below genome-wide significance in the current study. Hence, the four loci with GWAS P<1E-06 are commented upon below, given they have all also been implicated in studies of relevant phenotypes. The lead variant of the next most significant locus, rs3784351 (OR= 1.191, P=2.69E-07), is an intronic variant in the *ITGA11* gene on chromosome 15. The protein product of gene *ITGA11*, Integrin alpha-11, is a receptor for collagen which is ubiquitously expressed throughout the body, including in brain tissue. Although not previously implicated in the study of schizophrenia or TRS, it has been investigated previously in a GWAS meta-analysis of antidepressant efficacy in major depressive disorder, where it was nominally significantly associated with 2-week outcomes following initiation of SSRI's [203]. It has also been implicated in GWAS of cortical surface area [198] and sulcal depth [199], where the locus did surpass genome-wide significance. Interestingly, ITGA11 has also been implicated in studies of treatment outcome in cancer, and has been associated with poorer prognosis [204] and multidrug resistance [205] in cancer research.

The lead variant of the next locus, rs148108347 (OR= 0.663, P= 3.89E-07), is a non-coding transcript variant located within the *L3MBTL4* gene. Lethal(3)Malignant Brain Tumour-Like Protein 4 is what is known as a putative polycomb group (PcG) protein. PcG proteins play a key role in the repression of gene transcription, predicted to be via modification of chromatin or histones. Again, this gene has not been identified in GWAS of schizophrenia or TRS, but it has been implicated in studies of treatment outcome. For example, in a study of ACE inhibitors, used primarily for the treatment of hypertension, a locus attributed to this gene was found to significantly positive associated with ACE inhibitor discontinuation due to adverse drug reactions [206]. Another study examined treatment response in irritable bowel syndrome (IBS), and identified a significant association between *L3MBTL4* and the frequency of episodes of pain during treatment for IBS [207]. This is the only locus discussed here that is negatively associated with treatment resistance.

The lead variant of the next locus, rs11907443 (OR= 1.232, P= 4.49E-07), is an intronic variant in *SLC24A3.* Solute carrier family 24 member 3 has been associated with a number of phenotypes including, again, measures of treatment response. For example, in a study of almost 195,000 people, rs143934587 (not present in this current analysis) was nominally significantly associated with response to citalopram or escitalopram, with an OR of 6.71 and a p-value of 7E-07 [208]. This is therefore the second of the top five loci in this analysis that have been previously implicated in treatment response in psychiatric disorders, specifically depression, albeit not at genome-wide significance. *SLC24A3* was also found to be significantly associated with educational attainment [194] and multiple measures of smoking behaviour [209-211].

Finally, rs10502392 (OR=1.33, P=8.83E-07) is an intronic variant in *RAB31*. Ras-Related Protein Rab-31, and other members of the same protein family, are key regulators of intracellular membrane trafficking, from the formation of transport vesicles to their fusion with membranes. When Rab proteins become activated, they bind to GTP and become capable of recruiting different sets of downstream effectors directly responsible for vesicle formation, movement, tethering and fusion to the membrane. They also play a key role in the normal function of the Golgi apparatus. *RAB31* has not been implicated in TRS or schizophrenia more widely at any point previously, but it has been found to be associated with basophil count at genome wide significance in two large-scale GWAS [212,213]. Basophils are the least common form of granulocytes involved primarily in inflammatory responses and secrete compounds that are involved in the co-ordination of the immune response such as histamine and serotonin.

Limitations

There are a number of limitations that should be considered. Firstly, despite making no exclusions based on ancestry and developing a QC pipeline that could allow for cross-ancestry analyses, the final meta-analysis was approximately 98% European. Whilst two of the GWAS, those arrayed on the Infinium Psych Array and Global Screening Array (GSA) chips, were more diverse at approximately 90% Europeans, due to issues with data sharing, it was only possible to include PGC samples from the core European set at this time. The analysis does not include the 14 East Asian cohorts, or the African American and Latino samples that were ascertained from the Genomic Psychiatry Cohort (GPC) by the PGC for the latest schizophrenia GWAS. For subsequent data freezes, an emphasis must be placed on collecting non-European samples.

The next limitation is the need to include individuals without phenotype information at this time to facilitate the use of all available TRS cases. Matching by genotyping platform when merging cohorts is necessary due to low levels of direct overlap between certain chips, but in the case of cohorts genotyped on Infinium OmniExpress arrays, the vast majority of non-TRS cases were lacking phenotype information. Whilst as many unscreened non-TRS cases were removed from the analysis as possible, removal of all of them would have led to a loss of almost two thirds of the available TRS cases, due to them no longer having sufficient numbers of corresponding non-TRS cases. As a result, it is almost certain that misclassified individuals remain in the non-TRS cases at this time. Misclassification may also be occurring in the TRS cases, as information was not available to determine treatment adherence, and thus rule out pseudo-resistance where the observed non-response to treatment is being caused by clinical or pharmacokinetic factors, rather than issues of a pharmacodynamic nature. A detailed overview of the impact of misclassification and future work to overcome it can be found below and in the general discussion chapter of this thesis.

Finally, due to the time required to collect the phenotypic information that made this work possible and perform the QC on over 40 genotype datasets, only limited downstream analysis has been conducted to date. There are numerous potential avenues for further work based on the results of this chapter, which will be discussed here.

Planned Future Work

Immediate next steps include efforts to better quantify the genetic correlation of these results with the interaction analysis results and PGC3 SCZ using MiXeR to conduct the analysis [135]. MiXeR estimates the total number of causal variants that are shared between traits, as well as the number of trait specific causal variants, and quantifies the genetic overlap of phenotypes regardless of their genetic correlation. It also takes into account information regarding LD, MAF, sample size and cryptic relatedness. Heritability estimates will also be calculated using the LDAK software package [214], which differs from the heritability model utilised in the LDSC software because SNP heritability is expected to vary with both LD and MAF.

The calculation of PRS from the TRS GWAS results in an independent dataset containing both TRS and non-TRS individuals (expected to be NCMH) will also be completed, and an assessment of their association with treatment resistance and other schizophrenia-associated phenotypes conducted. This could include treatment resistance itself, definable within the sample by clozapine prescription, as well as phenotypes that have displayed associations with treatment resistance in previous literature, for example age at onset [215] and measures of educational attainment [114]

A replication of the interaction analysis [114] is also planned for completion, with the hypothesis that the improvements in phenotypic quality facilitated by the collection of further information, as well as an increased TRS sample size, will lead to a reduction in the genetic correlation observed between the two GWAS. This will require the reorganisation of the available samples into separate case-control GWAS of TRS and non-TRS. Because of this, the corresponding healthy controls for each PGC cohort were retained throughout the QC and removed only once imputation had been completed. There are several, large cohorts of just TRS (multiple phases of the Cardiff University cohort CLOZUK), which could quite readily be meta-analysed together to form the TRS case-control GWAS. The non-TRS GWAS will require more work to prepare, with TRS individuals, as best as they can be defined, being excluded from the cohorts. The accurate definition of TRS and non-TRS will be key for this analysis, which for reasons that will be discussed further in the general discussion, is not currently possible for all samples.

Future work will also focus on the better elucidation of the biological relevance of the genome wide significant locus, through procedures such as fine mapping to determine the most likely causal variant, and the summary statistics more widely. For example, gene ontology (GO) classifications, such as those available from the GO database (http://geneontology.org/), or developed by the SynGO consortium [216] can be tested for their association with TRS. Gene set enrichment analysis, where RNA-seq data is used to determine if genes of interest are differentially expressed in specific tissue and cell types, is also a potential avenue for further research.

In addition, further refinement of the non-TRS samples, both through the collection of phenotypic information for the cohorts for which there is currently none available, and the utilisation of the wider phenotypic information collected, will be necessary. Conversations are ongoing with the PIs of all but two of the 14 cohorts for which there is currently no phenotypic information available, and it should be feasible to collect clozapine / OPCRIT89 information for the majority of these cohorts prior to a subsequent data freeze. In addition, an examination of the antipsychotic medication information and duration of illness variables should allow for individuals with a high likelihood of being treatment resistant who have not yet been prescribed clozapine to be reclassified, or at least excluded from the non-TRS cases. It will be most appropriate to complete this portion of the work with the active involvement of a psychiatrist, due to the complex, non-standardised nature of the information available, and the variability in international prescribing practices (Hálfdánarson et al., 2017). An in-depth discussion of how to more accurately define both TRS and non-TRS, and the limitations of the current definitions, can be found in the general discussion chapter.

The final avenue of future research that will be discussed here is to expand the investigation of treatment resistance to additional psychiatric disorders. This is a primary aim of a new Horizon Europe grant psych STRATA, which plans to investigate treatment response and outcome across schizophrenia, bipolar disorder, and major depressive disorder. The analysis outlined here is an output of one of the work packages of this grant, and so will form the basis of a significant amount of work going forward.

Conclusion

People with TRS can often face a difficult path to therapeutic benefit, with only one licensed treatment option that can take years to be prescribed. One of the factors limiting improvements in treatment options for TRS at this time is that neurobiology specific to TRS remains poorly understood. The amalgamation and inclusion of individuals with TRS into broad schizophrenia research cohorts was paramount to the success of recent large-scale case-control GWAS, but it has come at the cost of phenotypic heterogeneity. In addition, the data required to conduct a direct comparison of TRS and non-TRS has until now proven incredibly difficult to collate, as both phenotypic information to define TRS and non-TRS accurately and corresponding individual level genotypes were required. In this chapter, through international collaboration with the schizophrenia working group of the PGC, and a small number of external analysts, it was finally possible to collect the data necessary to conduct a direct case-case GWAS of TRS versus non-TRS. A QC procedure was developed to allow for the combination of previously genotyped cohorts together into TRS versus non-TRS case-case cohorts, and with a final sample size of close to 40,000, a genome-wide significant locus was identified. Whilst post-GWAS interrogation of these results are just beginning, this GWAS represents a crucial first step in better elucidating the biological pathways and mechanisms underlying treatment resistance in schizophrenia.

Research Chapter 3: Association of Clozapine Metabolism with Absolute Neutrophil Count in Treatment Resistant Schizophrenia Cases

Chapter Summary

Clozapine remains the singular licensed treatment option for TRS, with demonstrable therapeutic benefit in approximately 60% of users and a significant association with reduced suicidal ideation. However, it remains globally under prescribed, in part due to concerns related to its expansive adverse effect profile. Of particular concern are the potential effects of clozapine on the immune system, with agranulocytosis and neutropenia remaining rare but serious side effects, necessitating regular haematological monitoring for all clozapine users. A recent small-scale study reported that clozapine plasma concentrations were inversely correlated with neutrophil counts in 41 individuals, most of whom were within their first year of treatment with clozapine. Attempting to replicate and further investigate this finding, metabolic, haematological and genetic data from a UK cohort of long-term clozapine users linked to a clozapine monitoring service, CLOZUK2 (N = 208) was extracted for investigation. Multiple linear regressions accounting for several potential confounding factors such as clozapine dose and time on clozapine, demonstrated a significant decrease in absolute neutrophil count (ANC), approximately 141 cells/mm3 for every 0.1 mg/L increase in clozapine concentration. Further regression models demonstrated that this relationship was diminished by the inclusion of the metabolic ratio of clozapine and norclozapine, its primary active metabolite, as a covariate. This metabolic ratio was negatively associated

with neutrophil concentrations, and further analysis revealed that three SNPs previously associated with norclozapine plasma concentrations and the metabolic ratio (rs61750900, rs2011425 and rs1126545) were also significantly associated with ANC. These SNPs all reside within CYP* and UGT* genes involved in the metabolism of clozapine, and these results highlight the need for continued investigation of pharmacogenomic variants and their role in the development of adverse side effects.

Introduction

Clozapine

First synthesised in the mid 1950's by Swiss pharmaceutical company Wander AG, clozapine was the first atypical anti-psychotic medication to be developed. Clozapine displays a relatively low affinity for type 2 dopamine receptors (D2) as compared to other typical anti-psychotic medications, displaying stronger antagonistic properties against D4 and 5-HT2A receptors [217]. Clozapine is considered the gold-standard treatment for individuals with TRS, and has been found to be associated with higher rates of occupational activity and living independently, as well as decreased hospitalisation rates and levels of compulsory treatment in this patient subgroup [218]. It has also been shown to significantly reduce the risk of suicide in individuals with TRS, as well as significantly lower rates of parkinsonism and tardive dyskinesia as compared to typical anti-psychotic medications such as haloperidol [219]. However, despite this, clozapine remains underutilised, not just in the UK, but across the globe [220,221]. According to the National Institute for Health and Care Excellence (NICE), in 2002, 63,000 individuals had TRS but only 21% of them were receiving Clozapine. Rates of clozapine prescribing have risen over time, increasing to 30% of those with TRS by 2007 [222] and 54% by 2010 [147]. However, the medication remains under prescribed, with psychiatrist attitudes often being cited as a significant factor in this. Studies have found that psychiatrists can be reluctant to initiate clozapine treatment, preferring to utilise polypharmacy in cases of inadequate treatment response [223], or preferring to delay clozapine initiation until after three or more unsuccessful antipsychotic treatments [224]. There are also well documented ethnic inequalities in clozapine prescription rates, with a systematic review of literature concluding that Black and Hispanic individuals accessing health care services in the UK and the USA were significantly less likely to receive clozapine than White service users [225]. Data from 10,512 individuals across England and Wales found that black service users had only 62% of the odds of receiving clozapine versus white service users, and were more likely to receive injectable / depot antipsychotic medications than other ethnic groups [226].

Reluctancy surrounding clozapine prescription can also in part be attributed to the risk of a small number of serious adverse side effects. Whilst many of clozapine's common side effects are analogous to other antipsychotic medications, such as constipation, headaches and nausea, clozapine is associated with relatively higher rates of metabolic side effects, such as the dysregulation of insulin and glucose [227]. In addition, the FDA has issued five so-called 'black box' warnings for clozapine, the highest level of warning for medications. These include myocarditis, seizures, risk of cardiovascular events in individuals with dementia, orthostatic hypotension, and perhaps the biggest cause for concern, blood dyscrasias.

Blood Dyscrasias

Clozapine has been associated with a range of blood dyscrasias [228], but the two most common are neutropenia and agranulocytosis, with prevalence's of 3.8% and 0.9% respectively [229,230]. Neutropenia is a condition that is characterised by a lower-than-normal number of neutrophils, typically below 1500 cells/mm^3 of blood. Neutrophils are a subtype of granulocytes that account for up to 70% of all white blood cells in an individual's immune system [231], and perform the important task of engulfing and destroying pathogens, a process known as phagocytosis. Whilst neutropenia can be short-lived, for example in cases of drug-induced neutropenia, there are also chronic forms of the condition that occur without clinical cause. This includes Benign Ethnic Neutropenia (BEN), where neutrophil count remains chronically low with no increased risk of infection. This condition most frequently occurs in individuals of African ancestries, with prevalence estimates of 25-50% [232]. This condition has recently been linked to the atypical chemokine receptor 1 (*ACKR1*) gene, referred to in previous literature as the Duffy antigen receptor complex (*DARC*) [180]. This GWAS of lowest absolute neutrophil count (ANC) in 552 individuals of African ancestry taking clozapine demonstrated that individuals who were homozygous for the C allele of rs2814778 (the Duffy-null genotype), were 20 times more likely to be classified as having neutropenia.

In contrast, agranulocytosis is an acute medical emergency, characterised by a significant reduction of granulocytes (again, the most affected being neutrophils) to dangerously low levels, typically below 500/mm^3. The result of this is a severely compromised immune system, leaving the affected individual highly vulnerable to infections. It has also been documented that unlike neutropenia where neutrophil count recovers fairly rapidly after discontinuation of the medication inducing the condition, neutrophils remain dangerously low for several days, or even weeks [233]. Agranulocytosis therefore has the potential to be life threatening; a review in 2006 reported a case fatality rate of clozapine- induced agranulocytosis of 4-16%, dependent on whether treatment with granulocyte colony stimulating factor (G-CSF), a blood growth factor that induces rapid neutrophil proliferation in the bone marrow was given [234]. The risk of agranulocytosis during Clozapine treatment first came to attention in 1975, when a letter was published in the Lancet reporting sixteen cases of clozapine-induced agranulocytosis, resulting in eight deaths [235]. It was noted in this letter that haematological monitoring could potentially allow for clozapine treatment to be continued in a safe manner, but it ultimately led to the medication being withdrawn from healthcare for nearly twenty years. It was at this point that two clinical trials, published a year apart, highlighted the potential role of clozapine in the subgroup of patients who had failed to respond to other anti-psychotic medications [236,237]. These studies, collectively involving just under 420 patients, demonstrated that clozapine had greater therapeutic benefits than chlorpromazine in individuals with TRS, and both concluded that frequent haematological monitoring, particularly at the beginning of the treatment regimen, would allow for the safe administration of the drug.

Clozapine Metabolism and the Relationship Between Clozapine and Neutrophils

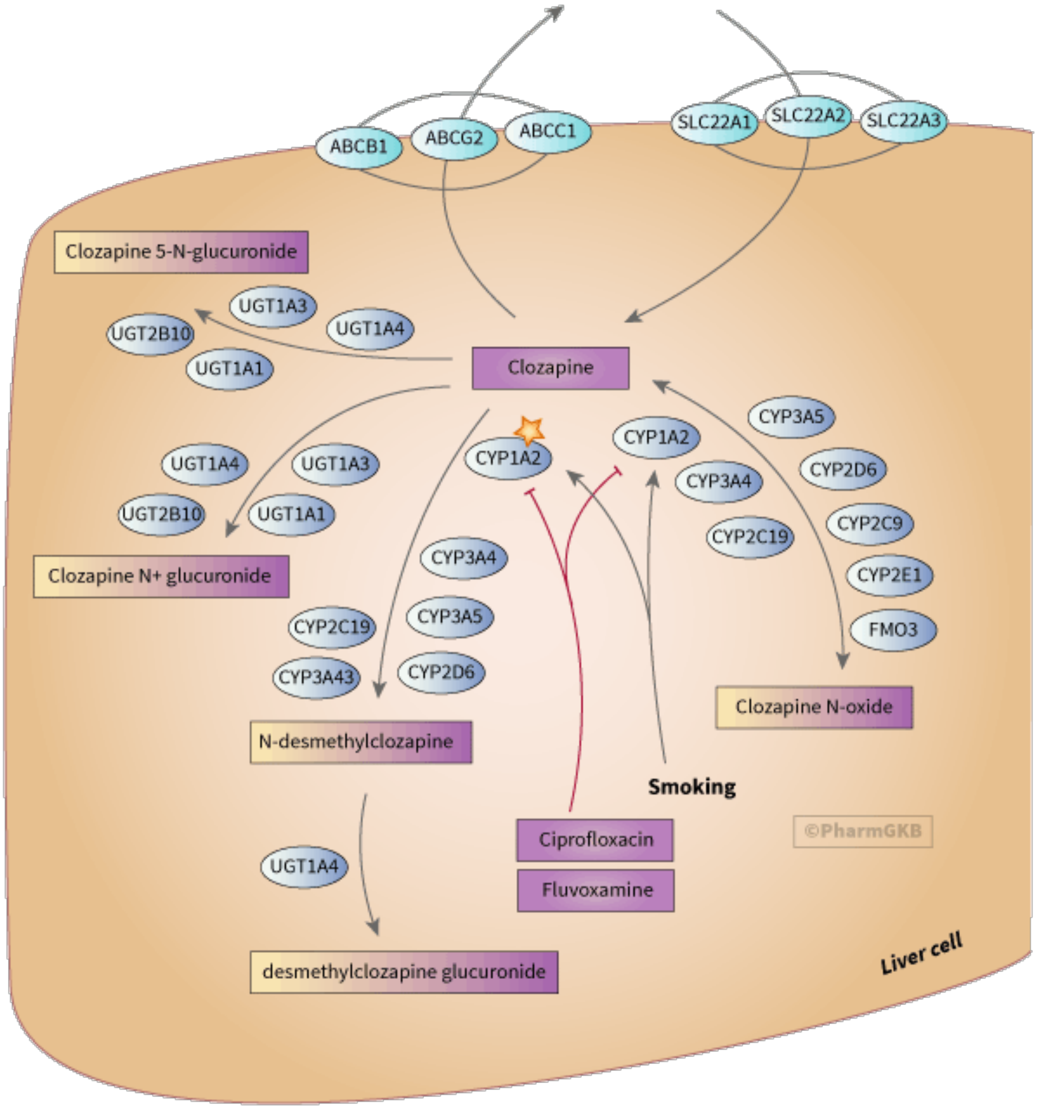

Figure 7: A diagram of a liver cell with an overview of clozapine metabolism and transport. Light blue denotes transporters gene, darker blue generic genes, solid purple drugs, and purple-yellow gradient metabolites. The star denotes significance. Figure accessed via https://www.pharmgkb.org/pathway/PA166163661, and first published in the following article [238]. Available for use under a Creative Commons BY-SA 4.0 license

An overview of clozapine metabolism can be seen in Figure 7. Clozapine is metabolised extensively in the liver, primarily through demethylation to produce N-desmethylclozapine (also referred to as norclozapine) or oxidation to clozapine N-oxide [239]. Cytochrome P450 3A4 (CYP3A4) is thought to be responsible for approximately 70% of clozapine clearance, with more minor involvement from several other members of the cytochrome P450 family (CYP1A2, CYP2C8, CYP2C19) and flavin containing dimethylaniline monooxygenase 3 (FMO3) [240]. Although numerous enzymes are capable of forming norclozapine and clozapine N-oxide, CYP3A4 and CYP1A2 are considered to be the major catalysts [241]. A small number of other metabolites have been identified in the urine of patients receiving clozapine, but their clinical significance is more poorly understood [242]. Clozapine is almost completely metabolised prior to excretion.

The precise mechanism by which clozapine is affecting neutrophil levels remains contested. One of the prevailing theories is that clozapine, or one its active metabolites, is bioactivated into a highly chemically reactive nitrenium ion, which has the capacity to induce apoptosis in neutrophils [243]. It was also noted in this study that the neutrophils that underwent apoptosis induced by the nitrenium ion displayed cell surface haptenation, with the nitrenium ion acting as the hapten. A hapten is a small molecule that is capable of eliciting an immune response, but only when it is attached to a larger protein, and it is possible that the nitrenium ion is able to induce an immune response against the neutrophils that they are bound to, leading to the observed apoptosis. A recent study demonstrated that clozapine, norclozapine and clozapine N-oxide are all capable of conversion into nitrenium ions [244]. It has been posited that the bioactivation of clozapine and its metabolites to this nitrenium ion is being done by the neutrophils themselves through the production of hypochlorous acid, a very strong oxidising agent, when the neutrophils become activated [245].

Previous Studies of Blood Dyscrasias and Clozapine

There have been a small number of studies investigating the relationship between plasma concentrations of clozapine metabolites and blood dyscrasias to date, but consensus has not been reached. This is likely due to the relatively small sample sizes of each study. For example, in an investigation of 5 patients with confirmed agranulocytosis, plasma clozapine and norclozapine levels were with the therapeutic window [246]. In addition, when compared to the serum levels of 59 patients on clozapine who had not developed agranulocytosis, there was no significant difference observed. Another study, this time with a cohort of 37 schizophrenia patients taking clozapine observed no significant associations between neutrophil count and plasma clozapine and norclozapine levels [247]. A more recent study, with a significantly increased sample size of 129 patients, did report a significant positive association between serum concentrations of the metabolite norclozapine and neutrophil count, as well as a positive association with clozapine / norclozapine ratio [248].

Finally, a study of 41 patients based in Mexico found a negative association between plasma clozapine concentration and neutrophil count [249]. In multiple regression analyses, the study observed significant negative associations between neutrophil count and multiple clozapine measures, including plasma levels, dose and time on clozapine. The same relationship was observed with leucocyte count, the name given to refer to all forms of white blood cells, including neutrophils. No significant associations were observed between neutrophil counts and norclozapine concentration, leading the authors to conclude that the effect of clozapine on neutrophils was being driven by clozapine itself, and not its metabolite.

Chapter Aims

The aims of this chapter were to attempt a replication of the findings put forward by Vaquero-Baez and colleagues in a larger cohort, utilising a UK-based cohort of individuals with TRS, CLOZUK2 [96]. The use of this cohort allowed for the incorporation of genetic data, specifically a set of common genetic variants that had recently been found to be associated with clozapine concentration, norclozapine concentration and clozapine:norclozapine ratio [250].

Methods

All statistical analysis, data curation and data visualisations presented here were completed using the programming language Rv4.0.2 through the GUI RStudio (2021.09.0 Build 351). All the work that is about to be presented was completed by myself independently under the supervision of Dr. Pardiñas and Professor Walters, unless otherwise indicated.

Cohort Description

Absolute Neutrophil Count data, clozapine plasma concentration data and genetic data were all collected as part of the CLOZUK2 study [96]. Whole blood samples and phenotypic information were obtained via a collaboration with Leyden Delta (Nijmegen, Netherlands), one of the major companies involved in the haematological monitoring and supply of clozapine in the UK. The ANC and genetic ancestry data were curated as part of a GWAS of ANC in individuals of African ancestry [180]. Ancestry was determined through the use of Ancestry Informative Markers (AIMs), a small group of SNPs that display highly divergent allele frequency distributions in individuals of different genetic ancestries [251]. ANC data was curated to generate a lowest ANC variable, where the data was subset to include only the lowest neutrophil count reading on record for each individual. Genotyping was completed by deCODE Genetics (Reykjavik, Iceland), following standard pre-imputation quality control procedures [252], and imputation was performed on the Michigan Imputation Server, utilising the Haplotype Reference Consortium panel [83,183]. Clozapine and norclozapine plasma concentration data was curated as part of a GWAS of clozapine plasma concentrations [250]. The CLOZUK2 dataset contained assays taken from just under 4000 individuals at 15,504 time points. Concentrations of < 0.05mg/L of either metabolite (indicative of treatment non-adherence) were excluded, as were assays for which the blood sampling took place either < 6 hours or > 24 hours after the patients most recent clozapine dose; this is the recommended window for clozapine blood monitoring, to ensure drug absorption is adequate at the time of sampling [253].

Inclusion Criteria

The dataset was curated to include only individuals who had a plasma clozapine concentration, measured in mg/L, and an ANC measurement, expressed in cells/mm^3, taken within a 21-day window of each other. A metabolic ratio variable was computed as the ratio of clozapine plasma concentration: norclozapine plasma concentration for each individual, who were then excluded if the ratio was > 3 or < 0.5 (n=19). Whilst a ratio of < 0.5 (indicative of high norclozapine in relation to clozapine) could be the result of genuine rapid metabolism, it is more likely to be indicative of a patient not adhering to their treatment regimen, due to the half-life of norclozapine being longer than that of clozapine. [254] Ratios of above 3 suggest that something is inhibiting the normal metabolism of clozapine, for example a concomitant medication [254,255]. It can also be indicative of treatment non-adherence, with high/normal levels of clozapine with an absence of norclozapine pointing to a patient only taking their medication on the morning of their blood monitoring appointment, in an attempt to conceal non-adherence. A total of 19 individuals were found to have ratios outside these boundaries and were subsequently excluded. All remaining individuals were then checked to determine if they were being prescribed clozapine for the control of drug-induced psychotic symptoms in Parkinson's Disease, the only other currently licensed use case for clozapine (n=1). Due to insufficient numbers of non-European individuals remaining in the sample following the application of these criteria's, the decision was made to restrict the cohort to individuals of European ancestry at this time, leaving a final total sample size of 208 individuals.

Statistical Analysis of Metabolite Data

All statistical analysis for this chapter were conducted in R V4.02. Initially, the bivariate relationships between ANC and four key variables (clozapine concentration, norclozapine concentration, time on clozapine treatment and daily clozapine dose) was assessed via spearman correlations. Spearman's correlation was selected due to the non-normal distribution of these variables, as assessed via visual inspection of histograms of the variables and Shapiro-Wilk's tests of normality. Following on from this, a series of multivariate linear regression models were developed to further assess the relationship between ANC (cells/mm^3) and plasma clozapine concentration (mg/L), with an additional eight covariates: plasma norclozapine concentration (mg/L), daily clozapine dose (mg), time on clozapine treatment (days), time between clozapine dose and blood sampling on the day of sampling (hours), sex (male/female), age (years), and age^2. Both age and age^2 were included in the analysis to account for a possible non-linear relationship between ANC and age [256]. A further two models were then run with additional covariates. First, the clozapine: norclozapine ratio was introduced to assess the relationship between clozapine metabolism and ANC, and then four pharmacogenomic SNPs were added in an attempt to identify potential mediators. These SNPs will be discussed in the below section.

Pharmacogenomic Analysis

Four SNPs were selected based on the results of clozapine, norclozapine and metabolic ratio GWAS conducted by Pardiñas and colleagues [250], details of which are presented in Table 16. Rs2472297 is an intergenic SNP located between *CYP1A1* and *CYP1A2* on chromosome 15. It was found to be significantly negatively associated with clozapine plasma concentration and carrying one copy of the minor allele of this variant was found to be associated with a reduction in clozapine daily dose of 50mg, and 100mg for homozygous carriers. It was hypothesised by the authors of this work that this may be linked to the location of this SNP in an area with a high density of aryl hydrocarbon receptor (AHR) protein binding sites. AHR binding in these areas has been shown to induce expression of CYP enzymes in hepatocytes, the main site of clozapine metabolism.

PHENOTYPE	SNP	MINOR ALLELE	BETA	SE	MAF
CLOZAPINE	**RS2472297**	T	–0.089	0.013	27.94
NORCLOZAPINE	**RS61750900**	T	–0.149	0.018	9.9
NORCLOZAPINE	**RS2011425**	G	–0.112	0.019	8.65
RATIO	**RS61750900**	T	0.212	0.012	9.9
RATIO	**RS1126545**	T	0.078	0.01	14.22

Table 16: Pharmacogenomic SNPs used in the regression analyses, and their association to clozapine metabolite concentrations in the GWAS from Pardiñas et al. (2019). The MAF has

been calculated based on the European only subset of CLOZUK2. SE= standard error of the beta

The next SNP was rs61750900. This was significantly associated with two of the investigated phenotypes; negatively with plasma norclozapine concentration, and positively with the metabolic ratio. This is a missense variant in the *UGT2B10* gene. The protein product of this gene is a glucuronosyltransferase, a type of enzyme that catalyses glucuronidation reactions. This is the process by which a glucuronic acid is covalently affixed to lipophilic compounds, as a method of elimination. The enzyme encoded for by this particular gene plays a major role in the glucuronidation of clozapine, as well as its major active metabolite, *N*-desmethylclozapine (norclozapine) [257]. Rs2011425, located on chromosome two, was also a missense variant located within a UGT gene, in this case *UGT1A4*.

Finally, rs1126545 is a missense variant located with the *CYP2C18* gene. The functional relevance of this gene is less clear, as it is not currently included in the canonical metabolic pathway for clozapine [238]. However, that being said, in vitro study has shown that is it capable of bioactivating clozapine [258]. In the Pardiñas paper, this SNP was positively associated with clozapine metabolic ratio.

A fifth SNP, rs2879954, was considered for analysis, after it was found to be negatively associated with clozapine serum concentration in a different paper, which conducted a GWAS of clozapine serum concentration adjusted for individually assessed smoking habits [259]. Unfortunately, the imputation quality of this SNP in the CLOZUK2 cohort was relatively low, with the SNP missing in 78 / 208 individuals in the cohort. As such, it was not further investigated at this time.

Results

The dataset contained 208 individuals following the application of the inclusion criteria, 147 males and 61 females. Descriptive statistics regarding this cohort are provided in Table 17. None of the individuals in this cohort had ANC levels at the time of blood sampling that would have been indicative of neutropenia or agranulocytosis, and on average, they had been taking clozapine for over 3 years at the time of sampling. 17 individuals were within the first 18 weeks of clozapine treatment, and the shortest duration of treatment was just under 11 weeks, with a longest duration of almost 20 years. They can therefore be generally thought of as a cohort of long-term clozapine users who are able to tolerate clozapine treatment and the necessitated haematological monitoring relatively well (the implications of this will be discussed further below). The average clozapine daily dose was 357mg (SD=136) for males and 322mg (SD=134) for females.

VARIABLE	**AVERAGE +/- SD**	
	Male, N=147	**Female, N=61**
CLOZAPINE (MG/L)	0.47 (0.28)	0.51 (0.32)
NORCLOZAPINE (MG/L)	0.27 (0.17)	0.29 (0.17)
DAILY DOSE (MG/DAY)	357 (136)	322.2 (134)
AGE (YEARS)	40.3 (13.2)	42.7 (14.2)
TIME ON TREATMENT (YEARS)	3.26 (0.8)	3.21 (1.1)
ABSOLUTE NEUTROPHIL COUNT (1000 CELLS/MM3)	3.1 (1.2)	3.1 (1.2)

Table 17: Covariates used in the correlation and regression analyses, and their distribution in the CLOZUK2 sample described in this study.

Correlation Analysis

Spearman's correlation results are reported in Table 18. None of the four variables were significantly associated with lowest ANC in these bivariate analyses.

OUTCOME	LOWEST ANC	
	rho	p
CLZ	−0.125	0.062
DMC	−0.033	0.629
TIME ON CLOZAPINE	−0.073	0.289
DAILY DOSE	−0.033	0.629

Table 18: Spearman's rank correlations of several variables with ANC, reproducing the approach of Vaquero-Baez et al. (2019)

Regression Analysis

In total, three linear regression models were conducted to assess the relationship between clozapine metabolism and lowest ANC. Model 1 contained clozapine plasma concentration, with the eight covariates discussed above. In model 2, the clozapine metabolic ratio was added, in line with the approach taken in previous clozapine TDM studies [260,261]. Finally in model 3, the pharmacogenomic variants discussed above were added to construct a final regression model. The results for each of these models can be seen in Table 19.

Model 1

Both clozapine and norclozapine level were significantly associated with ANC; clozapine level was negatively associated ($\beta = -1.41$, $p = 0.009$) and a positive association was found for norclozapine ($\beta = 1.77$, $p = 0.049$). For each 0.1 mg/L increase in plasma clozapine concentration, there was an associated decrease in ANC of 141 cells/mm3, and for each 0.1

mg/L increase in norclozapine concentration, there was an increase in ANC of 177 cells/mm3. In addition, time on clozapine was negatively associated with lowest ANC, as was age squared. Age was positively associated.

Model 2

With the addition of metabolic ratio in the model, the effect sizes of clozapine and norclozapine shifted toward zero, becoming nonsignificant. Thus, a significant proportion of their initial association appeared to be explained by their ratio, which was negatively associated with ANC (β = −0.69, p = 0.021). For every unit increase in the clozapine/norclozapine ratio there was an estimated associated decrease of 690 cells/mm3 in ANC. The relationship between lowest ANC and age, age squared and time on clozapine was once again observed. The effect size for age and age squared remained extremely similar, whilst the relationship between time on clozapine and lowest ANC was significantly stronger in model 2 versus model 1 (from -0.001 to -0.150).

VARIABLE	**MODEL 1**		**MODEL 2**		**MODEL 3**	
	Estimates	p	Estimates	p	Estimates	p
CLOZAPINE CONCENTRATION	**−1.410**	**0.009**	0.54	0.585	0.06	0.95
NORCLOZAPINE CONCENTRATION	**1.77**	**0.049**	−1.450	0.385	−0.570	0.738
DAILY DOSE	0	0.887	0	0.931	0	0.948
GENDER (MALE)	0.03	0.886	0.04	0.829	0.08	0.626
DAYS	−0.010	0.113	−0.010	0.124	−0.010	0.105
AGE	**0.12**	**0.001**	**0.11**	**0.001**	**0.12**	**0.001**
AGE SQUARED	**−0.001**	**0.004**	**−0.001**	**0.004**	**−0.001**	**0.004**
TIME ON CLOZAPINE	**−0.001**	**0.036**	**−0.150**	**0.044**	**−0.001**	**0.02**
TIME BETWEEN DOSE AND SAMPLE	−0.050	0.105	0.06	0.08	−0.050	0.122
RATIO			**−0.690**	**0.021**	**−0.540**	**0.035**
RS2472297_T					−0.130	0.324
RS61750900_T					**−0.410**	**0.048**
RS2011425_G					**0.45**	**0.026**
RS1126545_T					**0.33**	**0.039**

Table 19: Results of the three regression analyses with ANC as outcome. Bold highlight indicates statistically significant effect sizes ($p < 0.05$). Rs2472297 was associated with clozapine plasma concentration, rs61750900 was associated with norclozapine plasma concentration and ratio, rs2011425 was associated with norclozapine plasma concentration and rs1126545 was associated with ratio

Model 3

Of the four SNPs included in the analysis, three were significantly associated with ANC; rs61750900_T (β = −0.41, p = 0.048), rs2011425_G (β = 0.45, p = 0.026) and rs1126545_T (β = 0.33, p = 0.039). Each minor allele was associated with a decrease in ANC of 410 cells/mm3 and increases of 450 cells/mm3 and 330 cells/mm3, respectively. In this model, the metabolic ratio remained significantly associated with a decrease in ANC (β = −0.54, p = 0.035). The effect size of the metabolic ratio was reduced between model 2 and model 3, following the inclusion of the pharmacogenomic variants.

Discussion

General Discussion

In this analysis of 208 longstanding clozapine users, a relationship between clozapine plasma concentration and ANC was observed. However, this association was found to be mediated by the metabolic ratio of clozapine with one of its primary metabolites, norclozapine. For every unit increase of clozapine:norclozapine ratio, indicative of either more rapid clearance of norclozapine or relatively slower metabolism of clozapine, ANC decreased by 690 cells/mm^3. The addition of four pharmacogenomic variants previously associated with measures of clozapine levels led to three further significant associations with ANC, both negative (rs61750900) and positive (rs2011425, rs1126545). Based on the results of the regression models, these associations do not appear to be clozapine dose dependent.

Replication of Previous Findings

The effect sizes observed in this analysis compared to the findings reported by Vaquero-Baez and colleagues in 2019 were considerably smaller. Besides the effects of the so-called winner's curse [262], and the use of a substantially larger cohort (208 vs. 41), there are multiple other factors to consider. Firstly, the original study was conducted in Mexico City, Mexico, and although no formal information regarding the ancestry of the individuals within this cohort is given within the paper, they are likely to be of diverse ancestries [263]. Furthermore, it is likely that the cohort in the 2019 paper had significantly higher levels of non-European admixture than the CLOZUK2 sample, that was recruited in the UK and made up only of those of European ancestry, as inferred from genetic analyses. It therefore cannot be ruled out that the work of Vaquero-Baez has uncovered a population- or ancestry-specific effect, although no literature exists at this time to support this. As previously discussed, an increased prevalence of neutropenia has been found in individuals with schizophrenia of African ethnicity [264], and "benign" (constitutional) neutropenia rates vary widely based on genetic ancestry [180,265]. Indeed, the rate of clozapine-induced neutropenia in Finland seems to be 20 times higher than that of other European countries with similar clozapine prescribing rates [266], which may have attributed to the study that ultimately led to all but total withdrawal of clozapine from use in 1975 [235]. However, to date, no specific risk has been found in Mexican, Latino or Native American people.

Another noteworthy difference is that the daily clozapine dose prescribed is substantially different between the two cohorts, as inferred from a comparison of the study descriptive statistics. In this study, average clozapine doses were 348 mg/day for males and 313 mg/day for females, while Vaquero-Baez et al. (2019) reported average doses of 223 mg/day for males and 105 mg/day for females (t-test p male = 1.31 × 10–10; p female = 1.43 × 10–6). There was also a significant difference in the average time that the cohorts had been on treatment with clozapine, upwards of three years in the present study with 3.26 years for males and 3.21 years for females, and less than one year in Vaquero-Baez et al. (2019) with 10 months for males and 6.5 months for females (t-test p male = 5.9 × 10–3; p female = 6.18 × 10–16). This might have contributed to the smaller effect size observed between clozapine metabolites and ANC, as our sample represents individuals who are longer-term clozapine users who are able to tolerate clozapine reasonably well, and therefore likely excludes the approximately 10% of people who, after a year of treatment, might go on to exhibit neutropenia or other immune-related adverse effects [229]. The cohort in the 2019 study also contained individuals just initiating clozapine, meaning it contained a larger proportion of individuals who were within the 18-week period of highest risk for neutropenia and agranulocytosis than the cohort analysed here [267]. However, the main associations found across the successive regressions in this analysis suggest that clozapine might have sustained effects on ANC even in individuals without obvious haematological adverse effects. This is consistent with the rationale for the current practice of continued haematological monitoring of people being prescribed clozapine throughout their course of treatment, and supports that additional measures to lower their risk of infections might indeed be warranted even in long term clozapine users [268], for example ensuring their timely access to annual influenza vaccines [269].

Limitations

There are a few limitations in this work that must be considered when interpreting the results of this project. Although several covariates of suspected relevance to neutrophil count / clozapine metabolism were used in the regression analyses, no data were available for the CLOZUK2 cohort for several factors that are known to affect clozapine metabolism, including concomitant medications taken by the patients[270], the use of tobacco products[271] and the consumption of coffee and other caffeine-containing substances[272]. As well as this, the cross-sectional nature of the ANC data used in this study means there is no information regarding neutrophil trajectories over the course of treatment. Replicating this analysis in a longitudinal dataset would help to overcome this, and would allow for better estimations of the scale of the detected associations and whether ANC varies in particular patterns over time. Finally, due to a lack of samples with alternative genetic ancestries, this analysis was conducted in a sample of exclusively Europeans.

Future Work

Since the conclusion of this project, a third phase of the CLOZUK cohort (CLOZUK3) has been collected, first described in detail elsewhere [185]. This cohort of roughly 1400 samples has longitudinal measures of both clozapine levels and neutrophil counts, meaning that the work presented here can now be replicated in a larger, longitudinal sample of clozapine users. This would address two of the primary limitations of this analysis.

Chapter Conclusion

Neutropenia and agranulocytosis are significant barriers in the wider use of clozapine. The mechanisms underlying clozapine's effects on neutrophils are still not fully understood, and previous literature on the topic has been variable, owing to small sample sizes and heterogeneity of study design. In this chapter, a cohort of 208 clozapine users were curated with the metabolic, haematological and genetic data required to investigate the relationship between clozapine and ANC in finer detail. Clozapine metabolic ratio and three pharmacogenomic variants were demonstrated to be significantly associated with ANC, in a cohort of individuals who had been taking clozapine for an average of several years, and at the time of sampling did not have an ANC indicative of neutropenia or agranulocytosis. Whilst the results outlined here cannot provide specific insights into the biological processes underlying clozapine's impact on neutrophils, it does highlight a number of potential avenues for future research, including replication of the analysis in a longitudinal sample to better clarify the relationship between clozapine and neutrophil count across the course of treatment. The results are also in line with current guidelines surrounding ongoing haematological monitoring in clozapine users past the initial 18 weeks of treatment when agranulocytosis risk is highest.

Discussion and Conclusion

Summary of Results

Schizophrenia has a complex, polygenic genetic architecture, and is highly heterogeneous in terms of symptoms, response to treatment and prognosis. Large-scale collaborative GWAS have been successful in identifying common genetic variation associated with the disorder, but more work is required to understand how individuals with schizophrenia differ genetically from one another, and whether they can be stratified into subgroups. The ability to define more homogeneous cohorts of those with schizophrenia could advance the field of schizophrenia genetics research and diagnosis in psychiatry, and could be a first step to a more personalised medicine approach to the treatment of the disorder. This could be through the identification of novel therapeutic targets, leading ultimately to the development of new medications, or through identifying genes and pathways that will allow for the better utilisation of current treatment options. It could also allow for the identification of biomarkers that are specific to subtypes of the disorder, allowing for treatment regimens to be tailored to patients more individually from the outset. For example, if a biomarker of TRS could be confidently identified, individuals with TRS could more rapidly be prescribed treatments that have an increased chance of being effective, for example earlier access to clozapine. Currently, this is not the case; a recent review of the literature [273] found that the duration of delay from when a patient satisfies the eligibility criteria for clozapine treatment to the time of clozapine commencement ranged from 19.3 weeks to 5.5 years, and that the duration of illness prior to clozapine initiation ranged anywhere from 1.1 to 9.7 years.

In research chapter 1, I conducted an analysis on the largest available GWAS for bipolar disorder and schizophrenia, to identify disorder specific genes and genomic loci that could provide key insights into the neurobiological differences between the two disorders. Using the CC-GWAS method, I identified 27 loci that had significantly different allele frequencies between the two disorders, 26 of which can be considered schizophrenia unique. The summary statistics were significantly genetically correlated with schizophrenia, but not bipolar disorder, as well as a number of phenotypes previously associated with schizophrenia, including intelligence. Polygenic risk score analysis in an independent prevalence cohort, Cardiff COGS, showed a significant negative association between the CC-GWAS PRS and age at onset of psychosis, and positive associations with both negative and disorganised symptom domains. In each of these instances, the CC-GWAS PRS was associated with the phenotypes with the same direction of effect that the schizophrenia PRS displayed, with a similar effect size.

In research chapter 2, the focus shifted from differentiating schizophrenia from bipolar disorder, to investigating the genetic differences between treatment-resistant schizophrenia TRS) and non-TRS. Due to the availability of a large collection of previously genotyped schizophrenia cohorts, a direct case-case GWAS remained the gold standard, the CC-GWAS method not being appropriate for use anyway as a consequence of the high genetic correlation between TRS and non-TRS. Through a major collaboration with the schizophrenia working group of the PGC, I defined 40,000 schizophrenia cases as being resistant or responsive to treatment and conducted the largest case-case analysis of treatment resistance in schizophrenia to date. A genome-wide significant locus was, for the first time, identified, as were a number of loci, albeit below genome-wide significance, that had previously been implicated in GWAS studies of treatment response in other disorders, including major depressive disorder.

In research chapter 3, the research question was focussed within TRS, aiming to better elucidate the relationship between clozapine and absolute neutrophil count. Utilising the pharmacokinetic and haematological data collected as part of the mandatory blood monitoring whilst taking clozapine, I identified a significant association between the ratio of clozapine and its main metabolite norclozapine and ANC. This association was found to be in part mediated by a small number of variants that had previously been identified in GWAS of norclozapine concentration and clozapine:norclozapine ratio. This provided support for ongoing blood monitoring even in patients who have been maintained successfully on clozapine for many years and lent tentative evidence to the theory that clozapine's impact on neutrophils is not restricted to early-stage treatment. This reinforces that additional measures to lower risk of infections might indeed be warranted even in long term clozapine users [268], for example ensuring timely access to annual influenza vaccine [269] and COVID-19 booster vaccinations [274].

Key Contributions and Novel Findings

There are a number of key research findings that this thesis represents. Research chapter 1 acted as a validation of the CC-GWAS method and identified its potential utility in investigating both cross-disorder and within disorder phenotypic associations. It also identified a number of loci that appear to be specific to schizophrenia versus bipolar disorder, offering potential future insights into neurobiology specific to schizophrenia. The amalgamation of the phenotypic information required to undertake the work of research chapter 2, through extensive collaboration, and the development of a QC pipeline that allows for the combination of large numbers of previously genotyped cohorts without evidence of genomic inflation, are both key contributions, and ultimately resulted in the identification of the first genome-wide significant locus in TRS. The work will also undoubtedly prove valuable to ongoing efforts to understand the genetics of treatment resistance across the field of psychiatry, for example as part of the Horizon EU grant Psych-STRATA. Finally, research chapter 3 identified an association between clozapine metabolism and absolute neutrophil count in individuals who had been stably taking clozapine for several years. Whilst more work is required to both replicate this finding in a larger cohort and better elucidate the pathways / mechanisms underlying this association, this finding does have potential clinical implications.

Strengths and Limitations

Whilst the work of this thesis does provide a number of key contributions to the literature surrounding schizophrenia and its stratification into more genetically homogenous subgroups, it is not without its limitations, which will be discussed below.

General Limitations

One limitation shared across all three chapters is the lack of ancestral diversity in the samples. In research chapter 3 it was necessary to exclude non-European individuals completely, due to there not being sufficient numbers of people of other genetic ancestries to analyse. The schizophrenia cohort amalgamated as part of research chapter 2 ultimately ended up being approximately 98% Europeans, due to the majority of the cohorts being drawn from the core European datasets of the schizophrenia working group of the PGC. Even the most ancestrally diverse batch, those genotyped on the GSA array, was 90.2% from European ancestries. In research chapter 1, the iteration of the schizophrenia GWAS used for the analysis was approximately 19% East Asian and 91% European and did not contain the 14,394 African American and Latino samples that were later added to the meta-analysis. The version of the bipolar disorder GWAS used in the analysis was exclusively European. Schizophrenia is present globally, and the prevalence of schizophrenia does not deviate significantly by country [275]. However, prevalence has been demonstrated to significantly differ by ethnicity, with ethnic groups from outside Western Europe having significantly increased rates of schizophrenia and psychoses related diagnoses [276]. It is possible that this is related to environmental factors known to be associated with increased schizophrenia risk. For example, it could be a result of migration [277], with the majority of genetics research taking place in countries where non-European ancestries are minorities. The explanation for the link between migration and schizophrenia risk is not certain, however it is possible that the chronic experience of social defeat, racial discrimination and adverse living conditions could be contributing factors. It is also possible that there are ancestry-specific genetic variants associated with schizophrenia risk which are not currently being identified due to the Eurocentric nature of much of the GWAS literature. It is therefore paramount that non-European individuals are included in schizophrenia research, and more work is required to ensure that they are adequately represented.

Strengths and Limitations: Research Chapter One

These results demonstrate the utility of the CC-GWAS method for identifying potential genetic differences between pairs of disorders. As long as the limitations of the method are considered (discussed below), and the loci are suitably assessed through follow up analysis and review of the literature, CC-GWAS can effectively identify loci with divergent effects in disorders, as well as loci that are specific to only one disorder. Local genetic correlation analysis, for example through LAVA, can be used in conjunction to provide additional information about any identified loci, and statistical fine mapping techniques can be utilised to better characterise the identified loci. It therefore presents the opportunity to generate sets of genes that could offer insights into the aetiology of specific disorders, and in turn future potential drug targets. This would allow for treatment options to become more specific for each disorder, which currently tend to be treated with the same set of medications. In addition to the identification of disorder-specific genes, the CC-GWAS summary statistics can be subjected to much the same post-hoc analysis as case-control GWAS summary statistics. Here, it was demonstrated that CC-GWAS results could be used for genetic correlation analysis, and the generation of PRS, but many other potential applications exist. CC-GWAS summary statistics may therefore have extensive applications in cross-disorder research.

The CC-GWAS method is not without its limitations. First, it important not to generalise the results too much beyond the input case control GWAS, as you cannot guarantee that a variant that displayed no association in the selected input GWAS has not displayed a significant association in other GWAS of the disorders, or that a gene identified via the CC-GWAS has not previously been implicated in a disorder tagged by a different variant, as was found to be the case for *WSCD2*, discussed previously. This can be mitigated somewhat by using the largest, most highly powered GWAS as the input, but again, the onus is on the user to investigate each observed association. In this way, CC-GWAS summary statistics are analogous to GWAS, with CC-GWAS acting as an initial step in genomic discovery. Another limitation of relevance in this work is the effect of the power of the input GWAS on the results. In this case, the schizophrenia GWAS had significantly higher statistical power than the bipolar disorder GWAS, and so it is possibly not surprising that the loci that were identified here were schizophrenia specific. If a GWAS of more comparable statistical power was available for bipolar disorder, it is possible that loci specific to bipolar disorder could also have been identified, as it seems to be highly unlikely that the genetics of bipolar disorder consist exclusively of loci shared with schizophrenia. After all, the genetic correlation between schizophrenia and bipolar disorder, whilst high, is discernible from 1, at 0.68 [278].The final limitation of relevance to the work of this thesis is the fact that the method can only be applied to disorders, or subtypes of disorders, with a genetic correlation of < 0.8. This negates its use, currently, in sex stratified analyses of disorders where the genetic correlation is normally very close to 1, but also of highly related disorder subtypes, for example TRS and non-TRS. The 'delta' CC-GWAS method was found to be analogous to conducting a test for interaction between two case-control GWAS of TRS and non-TRS, as was performed by Pardiñas and colleagues in a previous PGC secondary analysis [114].

Strengths and Limitations: Research Chapter Two

The primary strength of this analysis was the sample size that it was possible to collect. For this analysis, it was necessary to amalgamate both the phenotypic information needed to define treatment resistance with a relatively high degree of accuracy, and the individual level genotypic data required to perform a direct case-case GWAS. With a final, harmonised sample size of just over 39000 individuals, this is the largest case-case analysis of TRS versus non-TRS ever conducted. In addition, a QC pipeline that allows for the combination and joint analysis of large numbers of previously genotyped cohorts together without evidence of genomic inflation has been developed, which may be useful as work continues as part of the Psych-STRATA grant, where analogous work will be conducted with a focus on bipolar disorder and major depressive disorder.

The primary limitation of the work presented in research chapter two is the known inclusion of misclassified individuals. [279]. Whilst approximately half of the individuals for which there was no available phenotypic information were removed from the analysis, approximately 2,500 individuals of 'unknown phenotype' are still present. This remains necessary currently; the exclusion of these people leads to an almost complete loss of non-TRS controls from the OMEX GWAS batches, which contain close to two thirds of the available TRS cases. Their complete removal would therefore lead to a significant reduction in the overall sample size. Based on the most current estimates of the prevalence of TRS, 20-30% of the individuals in these cohorts would be expected to be treatment resistant, but there is currently no way to define them. In addition, whilst clozapine prescription is considered a relatively effective proxy for treatment resistance [96,280,281], due to its use solely as a medication in individuals who have failed to respond to at least two (in practice, often more) antipsychotic medications, the definition of the treatment responsive individuals is more uncertain. Clozapine is widely underutilised internationally, with delays in initiation and premature discontinuation both contributing to this [220]. It is likely, therefore, that even in the cohorts for whom clozapine information was available, people who are treatment resistant (or are likely to become characterised as such with time) remain in the non-TRS side of the analysis at this time. The analyst who provided the FinnGen summary statistics attempted to quantify this, utilising the > 20 years of data available for the individuals in the dataset. When examining ~1500 individuals for whom there was at least 20 years of information available, if a cut-off of 2005 was applied, 486 individuals had received a clozapine prescription (32.1%). By 2010, 633 individuals had been prescribed clozapine (41.8%), and by 2015, 697 individuals had been prescribed clozapine (46.0%). Finally, by 2020, 746 individuals were receiving clozapine (49.2%). This means that at the 2005 time point, 33.8% of the non-TRS controls would have been wrongly classified. The majority of the PGC datasets used here cover a much shorter time course than FinnGen, and so it is likely that TRS individuals not far enough into their disease course to be prescribed clozapine are present in the controls. Finland also has one of the highest rates of clozapine prescription in the world [221], with their prescribing guidelines calling for a reduction in dose of clozapine in instances of neutropenia, rather than a cessation of the drug. As such, the levels of misclassification in samples from other countries may in fact be higher than the rate observed in FinnGen. Work is ongoing to address this issue, as discussed further below.

Strengths and Limitations: Research Chapter Three

The analysis conducted in research chapter 3 represented an improvement on the previously published work on this subject [246-249] in a number of ways, primarily in terms of sample size. The sample size (208) was five times larger than the work by Vaquero-Baez and colleagues (n=41) and approaching twice the size of the next largest analysis conducted by Smith and colleagues (n=129). The individually small sample sizes of these studies no doubt contributed to the variability of results that have been observed to date, and the lack of reproducibility across the literature.

The primary limitation of research chapter 3 is the cross-sectional nature of the available data. There was no information regarding patient's ANC over the course of their treatment, or at baseline before treatment with clozapine had begun. As such, whilst the research outlined in this chapter does point to a negative association between clozapine metabolism and ANC that is not limited to the first 18 weeks of clozapine treatment, this limitation should be taken into consideration when interpreting the results. In addition, there are a number of variables that are known to affect clozapine metabolism and ANC, which are missing from the analysis due to a lack of available data. For example, there is a well-established link between clozapine metabolism and smoking [282], and smokers are commonly found to have significantly lower clozapine: norclozapine ratios than non-smokers. There is also a well-documented association between clozapine metabolism and caffeine [283], with caffeine acting as an inhibitor of clozapine via its inhibition of the CYP1A2 receptor. There was also no available data on concomitant medication besides clozapine, a variable that could have potentially affected clozapine metabolism [284] and ANC [285]. Whilst the original study that inspired this analysis did have concomitant medication information and found no significant effect [249], it is important to consider that the effect observed here could be at least partially explained by other medications being taken by the patients. Finally, whilst it is the largest analysis of clozapine and ANC conducted to date, the sample size is still relatively small, at 208 individuals, and replication in a larger sample will be required for validation.

Potential Research / Clinical Applications

Potential utility for differential diagnosis

Rates of misdiagnosis in the field of psychiatry remain high [286], with differential diagnosis between schizophrenia and bipolar disorder continuing to be a challenge in some circumstances [287]. There is significant overlap between the disorders not just in terms of genetics, but also symptoms, patterns of inheritance in families, outcome, and treatment response [288]. Delays in correct diagnosis can have meaningful, long-term impacts on a patient's wellbeing and outcome, as it can lead to unnecessary or ineffective treatments being trialled, as well as frustration and upset for the patient themselves when their symptoms do not respond [289]. Although the predictive power of CC-GWAS derived PRS will be limited by the statistical power of the method and the input case-control GWAS, PRS could potentially be used as a tool for differential diagnosis. If an individual who first presents to medical services had a high PRS derived from this analysis, this would perhaps be indicative of a schizophrenia diagnosis. This will however require the input case-control GWAS to be not only large, but well characterised, made up of more homogenous case cohorts. More immediately, it could also have a number of other applications in research. For example, CC-GWAS PRS could be used to stratify previously collected research cohorts into more homogenous subgroups, such as a 'core' group of individuals who could be considered the closest to the archetypal presentation of schizophrenia. This could allow for the better elucidation of within disorder phenotypic associations, which may currently be obscured by the use of broader, heterogenous schizophrenia cohorts.

Potential biomarkers for neurobiological insights and novel drug targets

The disorder specific loci identified using the CC-GWAS method have the capacity to inform and improve current understanding regarding disorder specific pathology, and with time, could inform potential novel drug targets. Current psychiatric medications, whilst effective for many individuals, provide only symptomatic relief and do not fully address the full symptom profile of a significant proportion of patients. Genes that are identified through downstream analysis of a CC-GWAS analysis could represent pathways and mechanisms that are particularly important to the aetiology of a specific disorder, and as such could be good candidates for future drug targets. Evidence from human genetics research has been shown to be positively correlated with the success of drug development [290], and the CC-GWAS could potentially provide another layer of evidence for a novel drug target.

Potential biomarkers of treatment resistance in schizophrenia

To the best of knowledge, this is the first time that common genetic variants have been identified that have been associated specifically with TRS at genome-wide significance. Genes identified through case-case GWAS as being significantly associated with TRS could ultimately become novel drug targets, improving treatment options for a subgroup of individuals who historically have had only limited choices. During the period of time between diagnosis and clozapine commencement, patients are not experiencing adequate therapeutic benefits from antipsychotic medication. The duration of this period of untreated illness has been associated with poorer prognosis and treatment outcomes in a number of studies [291,292]. There is also evidence to suggest that delays in clozapine initiation reduce the efficacy of clozapine itself [293]. The identification of new drug targets, either informed by improved understanding of neurobiology specific to TRS or otherwise, has the potential to meaningfully impact the quality of treatments received by individuals with TRS. In addition, whilst probably not clinically useful, a TRS PRS could be utilised in research to stratify schizophrenia cohorts and identify individuals outside of those prescribed clozapine who have a high likelihood of having TRS. This would have been an incredibly useful tool to have in the preparation of the cohorts used in research chapter 2, both in the cohorts for whom no phenotype information was available, and in those where clozapine information was but did not necessarily capture all TRS individuals within the sample.

Potential link with treatment resistance of other psychiatric disorders

The post-GWAS analysis required to implicate specific genes, as well as replicate the findings in independent samples has not been completed. Nevertheless, the genes most implicated by lead SNP position within the nominally significant loci examined as part of research chapter 2 had all been implicated in previous GWAS studies of treatment response in other disorders, including two that were associated with treatment response in major depression. Treatment resistance is an ongoing issue across the field of psychiatry, with prevalence estimates varying widely from 20-60% for major psychiatric disorders [294]. It is also a topic of increasing interest within the research community, with estimates made that just under half of papers focussing on psychiatric disorders published in 2019 examining aspects of treatment resistance, an increase of approximately 20% from 2000 [294]. The concept that genomic loci associated with treatment resistance could be shared across major psychiatric disorders is potentially very interesting and is one of the key areas of interest of the Psych STRATA grant, a Horizon EU grant, which will examine treatment resistance in schizophrenia, bipolar disorder and major depression, as well as trans-diagnostically.

Accurate Definitions of TRS and non-TRS will be key to future research

Individuals with TRS are present in significant numbers within schizophrenia research cohorts. The final sample was 48.2% TRS, and whilst this will have been affected by the inclusion of the CLOZUK datasets which contain only TRS individuals, intra-cohort rates of treatment resistance were found to be significantly above the 20-30% rate that would be expected based on estimates of TRS prevalence in clinical practise. This was despite the fact that individuals with TRS cannot currently be perfectly called within the cohorts, with individuals remaining in the non-TRS controls at this time (discussed in detail previously). High rates of TRS in these cohorts is perhaps not surprising, due to a large proportion of PGC cohorts recruiting samples from within in-patient psychiatric units or blood monitoring services for clozapine, but it is a major consideration for future work. To fully define TRS in research samples phenotypic information beyond clozapine is required, to allow for the identification of people not yet prescribed clozapine, or who cannot tolerate it. The matter is even more complex with regards to accurately defining individuals with non-TRS, which at this point in time is based only on an absence of clozapine. This of course only has limited accuracy, and the collected medication information has the potential to improve the homogeneity of the non-TRS cohorts, through the identification of people who are likely to be treatment resistant in the absence of a clozapine prescription.

Genetic Variants Associated with Clozapine Metabolism also associated with ANC

This study demonstrated for the first time a potential link between pharmacogenetic variants associated with clozapine metabolism and absolute neutrophil count. Three variants were found to be significant when added as covariates in model 3 (rs61750900, rs2011425 and rs1126545), and were associated with a decrease in ANC of approximately 400 cells/mm^3 per effect allele. This is the first time, to the best of knowledge, that genetic determinants of clozapine pharmacokinetics were shown to have a potential effect on ANC, and more work is required to better understand the potential relationship observed here.

Clozapine Metabolism Associated with ANC even in Long-Term Clozapine Users

The cohort studied in research chapter 3 had on average been taking clozapine for several years and had not developed neutropenia or agranulocytosis during their treatment. However, a negative association between clozapine metabolism and ANC was still observed in these individuals, suggesting that clozapine affects ANC even in those who seem to tolerate the medication reasonably well. This would suggest that the current practice of continued regular haematological monitoring past the initial 18 week period associated with the highest risk of blood dyscrasias [267] is necessary, and that long-term clozapine users should be given priority access to measures that will lower their risk of developing infections.

Future Work

The work of research chapter 1 is complete, and in the process of being prepared for publication as part of a collaboration with another research group. A potential avenue for further work would be to implement the CC-GWAS+ method, discussed within the supplementary note of the CC-GWAS paper [115] and available as part of the 'CCGWAS' R package (https://github.com/wouterpeyrot/CCGWAS), where a direct case-case GWAS of schizophrenia and bipolar disorder is completed and added to the analysis to boost the statistical power. It is possible if this was completed that more bipolar disorder specific loci could be identified. In addition, the final published form of both the schizophrenia and bipolar disorder GWAS are larger with a greater degree of ancestral diversity than the versions used in this analysis, and it could therefore be prudent to repeat the analysis with the now available, larger GWAS. The CC-GWAS method is appropriate for use in the comparison of multi-ancestry GWAS, although it is noted by the authors that false positive associations can arise due to differential tagging of a causal stress test SNP caused by differences in ancestries between the GWAS [115].

The work of research chapter 2 is still ongoing. Future work will focus on the better elucidation of the biological relevance of the genome wide significant locus, and will also entail further refinement of the non-TRS samples, both through the collection of phenotypic information for the cohorts for which there is currently none available, and the utilisation of the wider phenotypic information collected. Through an examination of the antipsychotic medication information and duration of illness variables, individuals with a high likelihood of being treatment resistant who have not yet been prescribed clozapine can be reclassified, or at least excluded from the non-TRS controls. This work will likely have to be completed with the involvement of a psychiatrist, due to the complex, non-standardised nature of the information available, and the variability in international prescribing practices [295].There are also analytical techniques that can be implemented to account for misclassification in the GWAS, such as proposed here [296]. In addition, a series of downstream analyses will be undertaken, including the calculation of PRS and genetic correlation analysis, as well as significant work regarding gene sets and express. This is discussed in more detail in the discussion section of research chapter 2. In addition, a replication of the interaction analysis, where two case-control GWAS of TRS and non-TRS are analysed together [114] is planned for completion. With this in mind, the corresponding healthy controls for all of the cohorts were retained throughout the QC procedure and were removed only after imputation had been completed. Whilst the number of non-TRS samples is unlikely to increase significantly, the TRS case-control GWAS should be significantly larger. It would likely be prudent, however, to wait until all available samples have been fully classified as treatment resistant or responsive before conducting this analysis, as this analytical method will also be affected by treatment misclassification.

Since the completion of research chapter 3, a new wave of CLOZUK data has become available. This new phase of CLOZUK has a number of key advantages over previous iterations. Firstly, it will allow for a replication of the analysis in a larger sample size, and secondly, it has longitudinal measures of clozapine concentration and ANC. The new CLOZUK cohort therefore has the potential to overcome the two primary limitations of the work discussed here. The new phase of CLOZUK should therefore allow for a better understanding of the relationship between clozapine metabolism and ANC. This work is currently underway by others in the department, and the results are awaited with interest.

Conclusions

The heterogeneity of schizophrenia has been a significant barrier in the development of more personalised medicine approaches to treatment. The accurate identification of more homogenous subgroups of individuals with schizophrenia will be of significant importance to treatment regimens becoming more specific and could also lead to the identification of disorder/subtype distinct neurobiology being better understood, and with time, novel drug targets. This thesis used a range of methods to investigate common genetic variation specific to schizophrenia as a broad phenotype outside of its pleiotropy with bipolar disorder, identifying 27 loci that may hold key biological insights for the aetiology of schizophrenia. Through major collaboration with an international consortium, it then focused on treatment resistance, performing a direct case-case GWAS of TRS versus non-TRS and ultimately culminating in the identification of the first genome wide significant association with TRS. It also acted as an important proof of principle for future work that will attempt to investigate treatment resistance across major psychiatric disorders as part of the Psych-STRATA EU grant. Finally, the largest investigation to date of the association between clozapine metabolism and absolute neutrophil count was conducted, identifying novel associations with three common genetic variants associated with norclozapine concentration / metabolic ratio and ANC, and also demonstrating that clozapine negatively affects neutrophil count even in long term clozapine users. Whilst future work is required to fully elucidate the biological and mechanistic insights that these findings may represent, the work presented here was a crucial first step, and may ultimately contribute to the developments of novel treatment options for individuals with schizophrenia. This could either be through the identification of specific genes and drug targets, or through the better characterisation of research cohorts allowing for subgroup specific associations to be revealed.

References

1. Kendler. Tracing the Roots of Dementia Praecox: The Emergence of Verrücktheit as a Primary Delusional-Hallucinatory Psychosis in German Psychiatry From 1860 to 1880. *Schizophrenia bulletin* **46**, 765-773 (2020).
2. Kraepelin, E. *Psychiatrie; ein Lehrbuch für Studierende und Aerzte: Klinische Psychiatrie*, (Barth, 1913).
3. Blueler, E. *Dementia Praecox, Oder, Gruppe der Schizophrenien*, (Deuticke, Liepzig, Germany, 1911).
4. Kolker, R. *Hidden Valley Road*, (Hachette UK, 2020).
5. Solmi, M. *et al.* Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry* (2021).
6. Loebel, A.D. *et al.* Duration of psychosis and outcome in first-episode schizophrenia. *Am J Psychiatry* **149**, 1183-8 (1992).
7. Larson, M.K., Walker, E.F. & Compton, M.T. Early signs, diagnosis and therapeutics of the prodromal phase of schizophrenia and related psychotic disorders. *Expert review of neurotherapeutics* **10**, 1347-1359 (2010).
8. Cardno, A.G. *et al.* Factor analysis of schizophrenic symptoms using the OPCRIT checklist. *Schizophrenia Research* **22**, 233-239 (1996).
9. Marder, S.R., Davis, J.M. & Chouinard, G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. *J Clin Psychiatry* **58**, 538-46 (1997).
10. Smith, Mar, C.M. & Turoff, B.K. The structure of schizophrenic symptoms: a meta-analytic confirmatory factor analysis. *Schizophr Res* **31**, 57-70 (1998).
11. Legge *et al.* Associations Between Schizophrenia Polygenic Liability, Symptom Dimensions, and Cognitive Ability in Schizophrenia. *JAMA Psychiatry* **78**, 1143-1151 (2021).
12. *Diagnostic and statistical manual of mental disorders : DSM-5™*, (American Psychiatric Publishing, a division of American Psychiatric Association, Washington, DC ;, 2013).
13. Marwaha, S. *et al.* Rates and correlates of employment in people with schizophrenia in the UK, France and Germany. *British Journal of Psychiatry* **191**, 30-37 (2007).
14. Evensen, S. *et al.* Prevalence, Employment Rate, and Cost of Schizophrenia in a High-Income Welfare Society: A Population-Based Study Using Comprehensive Health and Welfare Registers. *Schizophrenia Bulletin* **42**, 476-483 (2016).
15. Jonsdottir, A. & Waghorn, G. Psychiatric disorders and labour force activity. *Mental health review journal* (2015).
16. Onaolapo, O.J. & Onaolapo, A.Y. Nutrition, nutritional deficiencies, and schizophrenia: An association worthy of constant reassessment. *World J Clin Cases* **9**, 8295-8311 (2021).
17. Stubbs, B. *et al.* How much physical activity do people with schizophrenia engage in? A systematic review, comparative meta-analysis and meta-regression. *Schizophrenia Research* **176**, 431-440 (2016).
18. Ziedonis, D. *et al.* Tobacco use and cessation in psychiatric disorders: National Institute of Mental Health report. *Nicotine & Tobacco Research* **10**, 1691-1715 (2008).

19. Quigley, H. & MacCabe, J.H. The relationship between nicotine and psychosis. *Therapeutic Advances in Psychopharmacology* **9**, 2045125319859969 (2019).
20. Hjorthøj, C., Stürup, A.E., McGrath, J.J. & Nordentoft, M. Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. *The Lancet Psychiatry* **4**, 295-301 (2017).
21. Hor, K. & Taylor, M. Suicide and schizophrenia: a systematic review of rates and risk factors. *Journal of psychopharmacology (Oxford, England)* **24**, 81-90 (2010).
22. Palmer, B.A., Pankratz, V.S. & Bostwick, J.M. The lifetime risk of suicide in schizophrenia: a reexamination. *Archives of general psychiatry* **62**, 247-253 (2005).
23. Suokas, J.T. *et al.* Epidemiology of suicide attempts among persons with psychotic disorder in the general population. *Schizophrenia Research* **124**, 22-28 (2010).
24. Sher, L. & Kahn, R.S. Suicide in Schizophrenia: An Educational Overview. *Medicina (Kaunas)* **55**(2019).
25. Knapp, M., Mangalore, R. & Simon, J. The Global Costs of Schizophrenia. *Schizophrenia Bulletin* **30**, 279-293 (2004).
26. Ride, J., Kasteridis, P., Gutacker, N., Aragon Aragon, M.J. & Jacobs, R. Healthcare Costs for People with Serious Mental Illness in England: An Analysis of Costs Across Primary Care, Hospital Care, and Specialist Mental Healthcare. *Applied health economics and health policy* **18**, 177-188 (2020).
27. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. *Epidemiologic Reviews* **30**, 67-76 (2008).
28. Shenton, M.E., Whitford, T.J. & Kubicki, M. Structural neuroimaging in schizophrenia from methods to insights to treatments. *Dialogues in Clinical Neuroscience* **12**, 317-332 (2010).
29. Sommer, I.E. *et al.* How Frequent Are Radiological Abnormalities in Patients With Psychosis? A Review of 1379 MRI Scans. *Schizophrenia Bulletin* **39**, 815-819 (2013).
30. Wheeler & Voineskos, A.N. A review of structural neuroimaging in schizophrenia: from connectivity to connectomics. *Front Hum Neurosci* **8**, 653 (2014).
31. Dabiri, M. *et al.* Neuroimaging in schizophrenia: A review article. *Front Neurosci* **16**, 1042814 (2022).
32. van Erp, T.G.M. *et al.* Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry* **21**, 547-553 (2016).
33. Carlsson, A. & Lindqvist, M. Effect of Chlorpromazine or Haloperidol on Formation of 3-Methoxytyramine and Normetanephrine in Mouse Brain. *Acta Pharmacologica et Toxicologica* **20**, 140-144 (1963).
34. Seeman, P., Chau-Wong, M., Tedesco, J. & Wong, K. Brain receptors for antipsychotic drugs and dopamine: direct binding assays. *Proc Natl Acad Sci U S A* **72**, 4376-80 (1975).
35. O'Donnell, P. & Grace, A.A. Dysfunctions in Multiple Interrelated Systems as the Neurobiological Bases of Schizophrenic Symptom Clusters. *Schizophrenia Bulletin* **24**, 267-283 (1998).
36. Krystal, J.H. *et al.* Subanesthetic Effects of the Noncompetitive NMDA Antagonist, Ketamine, in Humans: Psychotomimetic, Perceptual, Cognitive, and Neuroendocrine Responses. *Archives of General Psychiatry* **51**, 199-214 (1994).
37. Stahl. Beyond the dopamine hypothesis of schizophrenia to three neural networks of psychosis: dopamine, serotonin, and glutamate. *CNS Spectrums* **23**, 187-191 (2018).

38. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a Complex Trait: Evidence From a Meta-analysis of Twin Studies. *Archives of General Psychiatry* **60**, 1187-1192 (2003).
39. Polderman, T.J.C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* **47**, 702-709 (2015).
40. Pettersson, E. *et al.* Genetic influences on eight psychiatric disorders based on family data of 4 408 646 full and half-siblings, and genetic data of 333 748 cases and controls. *Psychol Med* **49**, 1166-1173 (2019).
41. Smeland, O.B., Frei, O., Dale, A.M. & Andreassen, O.A. The polygenic architecture of schizophrenia — rethinking pathogenesis and nosology. *Nature Reviews Neurology* **16**, 366-379 (2020).
42. Robinson, N. & Bergen, S.E. Environmental Risk Factors for Schizophrenia and Bipolar Disorder and Their Relationship to Genetic Risk: Current Knowledge and Future Directions. *Frontiers in Genetics* **12**(2021).
43. Kendler & Zerbin-Rüdin, E. Abstract and review of “studien über vererbung und entstehung geistiger störungen. I. Zur vererbung und neuentstehung der dementia praecox.” (Studies on the inheritance and origin of mental illness: I. To the problem of the inheritance and primary origin of dementia praecox.). *American Journal of Medical Genetics* **67**, 338-342 (1996).
44. Kallmann, F.J. The genetics of schizophrenia. (1938).
45. Kendler, K.S. & Klee, A. The place of Franz Kallmann's 1938 "the genetics of schizophrenia" in the history of psychiatric genetics. *Am J Med Genet B Neuropsychiatr Genet* **189**, 26-36 (2022).
46. Luxenburger, H. Vorläufiger bericht über psychiatrische Serienuntersuchungen an zwillingen. *Zeitschrift für die gesamte Neurologie und Psychiatrie* **116**, 297-326 (1928).
47. Kallmann, F.J. The genetic theory of schizophrenia: An analysis of 691 schizophrenic twin index families. *American Journal of Psychiatry* **103**, 309-322 (1946).
48. Rosenthal, D. Problems of sampling and diagnosis in the major twin studies of schizophrenia. *Journal of Psychiatric Research* **1**, 116-134 (1962).
49. Gottesman, I.I. & Shields, J. *Schizophrenia and genetics: A twin study vantage point*, xviii, 433-xviii, 433 (Academic Press, Oxford, England, 1972).
50. Kringlen, E. *Heredity and environment in the functional psychoses: An epidemiological–clinical twin study*, (Elsevier, 2013).
51. Gottesman, I.I. *Schizophrenia genesis: The origins of madness*, (WH Freeman/Times Books/Henry Holt & Co, 1991).
52. Cardno, A.G. & Gottesman, I.I. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *American journal of medical genetics* **97**, 12-17 (2000).
53. Heston, L.L. Psychiatric Disorders in Foster Home Reared Children of Schizophrenic Mothers. *British Journal of Psychiatry* **112**, 819-825 (1966).
54. Tienari, P. *et al.* Genetic boundaries of the schizophrenia spectrum: evidence from the Finnish Adoptive Family Study of Schizophrenia. *Am J Psychiatry* **160**, 1587-94 (2003).
55. Lowing, P.A., Mirsky, A.F. & Pereira, R. The inheritance of schizophrenia spectrum disorders: a reanalysis of the Danish adoptee study data. *Am J Psychiatry* **140**, 1167-71 (1983).

56. Kety, S.S. *et al.* Mental illness in the biological and adoptive relatives of schizophrenic adoptees. Replication of the Copenhagen Study in the rest of Denmark. *Arch Gen Psychiatry* **51**, 442-55 (1994).
57. Wender, P.H., Rosenthal, D., Kety, S.S., Schulsinger, F. & Welner, J. Crossfostering: A research strategy for clarifying the role of genetic and experiential factors in the etiology of schizophrenia. *Archives of General Psychiatry* **30**, 121-128 (1974).
58. Hilker, R. *et al.* Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biological Psychiatry* **83**, 492-498 (2018).
59. Riley, B. Linkage studies of schizophrenia. *Neurotoxicity Research* **6**, 17-34 (2004).
60. Pulver, A.E. *et al.* Sequential strategy to identify a susceptibility gene for schizophrenia: Report of potential linkage on chromosome 22q12-q13.1: Part 1. *American Journal of Medical Genetics* **54**, 36-43 (1994).
61. Polymeropoulos, M.H. *et al.* Search for a schizophrenia susceptibility locus on human chromosome 22. *American Journal of Medical Genetics* **54**, 93-99 (1994).
62. Pulver, A.E. *et al.* Schizophrenia: A genome scan targets chromosomes 3p and 8p as potential sites of susceptibility genes. *American Journal of Medical Genetics* **60**, 252-260 (1995).
63. Wildenauer, D.B. *et al.* Additional support for schizophrenia linkage on chromosomes 6 and 8: A multicenter study. *American Journal of Medical Genetics* **67**, 580-594 (1996).
64. Straub, R.E. *et al.* Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am J Hum Genet* **71**, 337-48 (2002).
65. Schwab, S.G. *et al.* Evaluation of a susceptibility gene for schizophrenia on chromosome 6p by multipoint affected sib-pair linkage analysis. *Nat Genet* **11**, 325-7 (1995).
66. Pulver, A.E. *et al.* Follow-up of a report of a potential linkage for schizophrenia on chromosome 22q12-q13.1: Part 2. *American Journal of Medical Genetics* **54**, 44-50 (1994).
67. Ng, M.Y. *et al.* Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry* **14**, 774-85 (2009).
68. Mansournia, M.A., Hernán, M.A. & Greenland, S. Matched designs and causal diagrams. *International Journal of Epidemiology* **42**, 860-869 (2013).
69. Sullivan, P.F. How Good Were Candidate Gene Guesses in Schizophrenia Genetics? *Biological Psychiatry* **82**, 696-697 (2017).
70. Allen, N.C. *et al.* Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genetics* **40**, 827-834 (2008).
71. Johnson, E.C. *et al.* No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes. *Biological Psychiatry* **82**, 702-708 (2017).
72. Farrell, M.S. *et al.* Evaluating historical candidate genes for schizophrenia. *Molecular Psychiatry* **20**, 555-562 (2015).
73. Riley, J.H., Allan, C.J., Lai, E. & Roses, A. The use of single nucleotide polymorphisms in the isolation of common disease genes. *Pharmacogenomics* **1**, 39-47 (2000).
74. Liu, H. *et al.* Genetic variation at the 22q11 PRODH2/DGCR6 locus presents an unusual pattern and increases susceptibility to schizophrenia. *Proc Natl Acad Sci U S A* **99**, 3717-22 (2002).

75. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet* **71**, 877-92 (2002).
76. Chumakov, I. *et al.* Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc Natl Acad Sci U S A* **99**, 13675-80 (2002).
77. Risch, N. & Merikangas, K. The Future of Genetic Studies of Complex Human Diseases. *Science* **273**, 1516-1517 (1996).
78. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
79. Wang, D.G. *et al.* Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**, 1077-1082 (1998).
80. Ma, C., Blackwell, T., Boehnke, M. & Scott, L.J. Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. *Genetic Epidemiology* **37**, 539-550 (2013).
81. Gibbs, R.A. *et al.* The International HapMap Project. *Nature* **426**, 789-796 (2003).
82. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
83. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279 (2016).
84. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).
85. Klein, R.J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385-389 (2005).
86. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**, D896-D901 (2017).
87. O'Donovan, M.C. *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics* **40**, 1053-1055 (2008).
88. Deans, P.J.M. *et al.* Psychosis Risk Candidate ZNF804A Localizes to Synapses and Regulates Neurite Formation and Dendritic Spine Structure. *Biological Psychiatry* **82**, 49-61 (2017).
89. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757 (2009).
90. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
91. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747 (2009).
92. Mesman, S., Bakker, R. & Smidt, M.P. Tcf4 is required for correct brain development during embryogenesis. *Molecular and Cellular Neuroscience* **106**, 103502 (2020).
93. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969-976 (2011).
94. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics* **45**, 1150-1159 (2013).
95. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
96. Pardiñas, A.F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics* **50**, 381-389 (2018).

97. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature genetics* **51**, 1670-1678 (2019).

98. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502-508 (2022).

99. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509-516 (2022).

100. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).

101. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481-487 (2016).

102. Priya, A., Johar, K. & Wong-Riley, M.T.T. Specificity protein 4 functionally regulates the transcription of NMDA receptor subunits GluN1, GluN2A, and GluN2B. *Biochimica et biophysica acta* **1833**, 2745-2756 (2013).

103. Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590 (2018).

104. Ruderfer, D.M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry* **19**, 1017-1024 (2014).

105. Mullins, N. *et al.* GWAS of Suicide Attempt in Psychiatric Disorders and Association With Major Depression Polygenic Risk Scores. *Am J Psychiatry* **176**, 651-660 (2019).

106. Cleynen, I. *et al.* Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. *Molecular Psychiatry* **26**, 4496-4510 (2021).

107. Zhang, J.-P. *et al.* Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis. *American Journal of Psychiatry* **176**, 21-28 (2018).

108. Jonas, K.G. *et al.* Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders. *Transl Psychiatry* **9**, 300 (2019).

109. Wray, N.R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *J Child Psychol Psychiatry* **55**, 1068-87 (2014).

110. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**(2018).

111. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).

112. Wu, Y., Zheng, Z., Visscher, P.M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol* **18**, 86 (2017).

113. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nature Genetics* **52**, 437-447 (2020).

114. Pardiñas, A.F. *et al.* Interaction Testing and Polygenic Risk Scoring to Estimate the Association of Common Genetic Variants With Treatment Resistance in Schizophrenia. *JAMA Psychiatry* **79**, 260-269 (2022).

115. Peyrot, W.J. & Price, A.L. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. *Nature Genetics* **53**, 445-454 (2021).

116. Smoller, J.W. & Finn, C.T. Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **123C**, 48-58 (2003).

117. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics* **53**, 817-829 (2021).

118. Consortium, T.S.P.G.-W.A.S.G. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381**, 1371-1379 (2013).

119. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address, p.m.h.e. & Cross-Disorder Group of the Psychiatric Genomics, C. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482.e11 (2019).

120. Curtis, D. *et al.* Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. *Psychiatr Genet* **21**, 1-4 (2011).

121. Ruderfer, D.M. *et al.* Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715.e16 (2018).

122. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* **50**, 229-237 (2018).

123. Qi, G. & Chatterjee, N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLOS Genetics* **14**, e1007549 (2018).

124. van der Meer, D. *et al.* Understanding the genetic determinants of the brain with MOSTest. *Nat Commun* **11**, 3512 (2020).

125. Grotzinger, A.D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour* **3**, 513-525 (2019).

126. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018).

127. Byrne, E.M. *et al.* Conditional GWAS analysis to identify disorder-specific SNPs for psychiatric disorders. *Molecular Psychiatry* **26**, 2070-2081 (2021).

128. Lloyd-Jones, L.R., Robinson, M.R., Yang, J. & Visscher, P.M. Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. *Genetics* **208**, 1397-1408 (2018).

129. Peyrot, W.J. & Price, A.L. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. *bioRxiv*, 2020.03.04.977389 (2020).

130. Stahl, E.A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature Genetics* **51**, 793-803 (2019).

131. Moore, D.L., Apara, A. & Goldberg, J.L. Krüppel-like transcription factors in the nervous system: novel players in neurite outgrowth and axon regeneration. *Molecular and cellular neurosciences* **47**, 233-243 (2011).

132. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183 (2016).

133. The Schizophrenia Working Group of the Psychiatric Genomics, C., Ripke, S., Walters, J.T.R. & O'Donovan, M.C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv*, 2020.09.12.20192922 (2020).

134. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291-295 (2015).
135. Frei, O. *et al.* Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nature Communications* **10**, 2417 (2019).
136. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics (Oxford, England)* **36**, 930-933 (2020).
137. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).
138. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).
139. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics (Oxford, England)* **33**, 272-279 (2017).
140. Psychiatric, G.C.B.D.W.G. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature genetics* **43**, 977-983 (2011).
141. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C.A. An integrated framework for local genetic correlation analysis. *Nature Genetics* **54**, 274-282 (2022).
142. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Molecular Psychiatry* **19**, 37-40 (2014).
143. Choi, S.W. & O'Reilly, P.F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, giz082 (2019).
144. Demontis, D. *et al.* Genome-wide association study implicates CHRNA2 in cannabis use disorder. *Nature neuroscience* **22**, 1066-1074 (2019).
145. Scheper, G.C. *et al.* Mitochondrial aspartyl-tRNA synthetase deficiency causes leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation. *Nature Genetics* **39**, 534-539 (2007).
146. Hamshere, M.L. *et al.* 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**, 708-712 (2013).
147. Mortimer, A.M., Singh, P., Shepherd, C.J. & Puthiryackal, J. Clozapine for treatment-resistant schizophrenia: National Institute of Clinical Excellence (NICE) guidance in the real world. *Clin Schizophr Relat Psychoses* **4**, 49-55 (2010).
148. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nature Communications* **12**, 1164 (2021).
149. Kane, J.M. *et al.* Clinical Guidance on the Identification and Management of Treatment-Resistant Schizophrenia. *J Clin Psychiatry* **80**(2019).
150. Meltzer, H.Y. Treatment-Resistant Schizophrenia - The Role of Clozapine. *Current Medical Research and Opinion* **14**, 1-20 (1997).
151. Siskind, D. *et al.* Rates of treatment-resistant schizophrenia from first-episode cohorts: systematic review and meta-analysis. *The British Journal of Psychiatry* **220**, 115-120 (2022).

152. Lally, J. *et al.* Two distinct patterns of treatment resistance: clinical predictors of treatment resistance in first-episode schizophrenia spectrum psychoses. *Psychol Med* **46**, 3231-3240 (2016).
153. Correll, C.U., Brevig, T. & Brain, C. Patient characteristics, burden and pharmacotherapy of treatment-resistant schizophrenia: results from a survey of 204 US psychiatrists. *BMC Psychiatry* **19**, 362 (2019).
154. Kennedy, J.L., Altar, C.A., Taylor, D.L., Degtiar, I. & Hornberger, J.C. The social and economic burden of treatment-resistant schizophrenia: a systematic literature review. *International Clinical Psychopharmacology* **29**(2014).
155. Legge, S.E. *et al.* Reasons for discontinuing clozapine: A cohort study of patients commencing treatment. *Schizophrenia Research* **174**, 113-119 (2016).
156. Luykx, J.J., Stam, N., Tanskanen, A., Tiihonen, J. & Taipale, H. In the aftermath of clozapine discontinuation: comparative effectiveness and safety of antipsychotics in patients with schizophrenia who discontinue clozapine. *The British Journal of Psychiatry* **217**, 498-505 (2020).
157. Wagner, E. *et al.* Clozapine augmentation strategies – a systematic meta-review of available evidence. Treatment options for clozapine resistance. *Journal of Psychopharmacology* **33**, 423-435 (2019).
158. Gillespie, A.L., Samanaite, R., Mill, J., Egerton, A. & MacCabe, J.H. Is treatment-resistant schizophrenia categorically distinct from treatment-responsive schizophrenia? a systematic review. *BMC Psychiatry* **17**, 12 (2017).
159. Vita, A. *et al.* Treatment-Resistant Schizophrenia: Genetic and Neuroimaging Correlates. *Frontiers in Pharmacology* **10**(2019).
160. Howes, O.D. & Kapur, S. A neurobiological hypothesis for the classification of schizophrenia: type a (hyperdopaminergic) and type B (normodopaminergic). *British Journal of Psychiatry* **205**, 1-3 (2014).
161. Joober, R. *et al.* Increased prevalence of schizophrenia spectrum disorders in relatives of neuroleptic-nonresponsive schizophrenic patients. *Schizophrenia Research* **77**, 35-41 (2005).
162. Krebs, M.O. *et al.* Brain derived neurotrophic factor (BDNF) gene variants association with age at onset and therapeutic response in schizophrenia. *Mol Psychiatry* **5**, 558-62 (2000).
163. Anttila, S. *et al.* Association between 5-HT2A, TPH1 and GNB3 genotypes and response to typical neuroleptics: a serotonergic approach. *BMC Psychiatry* **7**, 22 (2007).
164. Krebs, M.O. *et al.* Dopamine D3 receptor gene variants and substance abuse in schizophrenia. *Mol Psychiatry* **3**, 337-41 (1998).
165. Need, A.C. *et al.* Pharmacogenetics of antipsychotic response in the CATIE trial: a candidate gene analysis. *European Journal of Human Genetics* **17**, 946-957 (2009).
166. Koga, A. *et al.* GWAS analysis of treatment resistant schizophrenia: interaction effect of childhood trauma. *Pharmacogenomics* **18**, 663-671 (2017).
167. Liou, Y.J. *et al.* Genome-wide association study of treatment refractory schizophrenia in Han Chinese. *PLoS One* **7**, e33598 (2012).
168. Li, J. & Meltzer, H.Y. A genetic locus in 7p12.2 associated with treatment resistant schizophrenia. *Schizophr Res* **159**, 333-9 (2014).
169. Altman, D.G. & Bland, J.M. Interaction revisited: the difference between two estimates. *BMJ* **326**, 219 (2003).

170. Mitchell, B. *et al.* Using previously genotyped controls in genome-wide association studies (GWAS): application to the Stroke Genetics Network (SiGN). *Frontiers in Genetics* **5**(2014).
171. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
172. Hubbard, L. *et al.* DRAGON-Data: A platform and protocol for integrating genomic and phenotypic data across large psychiatric cohorts. *medRxiv*, 2022.01.18.22269463 (2022).
173. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Research Notes* **7**, 901 (2014).
174. Sariya, S. *et al.* Rare Variants Imputation in Admixed Populations: Comparison Across Reference Panels and Bioinformatics Tools. *Front Genet* **10**, 239 (2019).
175. Panoutsopoulou, K. & Walter, K. Quality Control of Common and Rare Variants. in *Genetic Epidemiology: Methods and Protocols* (ed. Evangelou, E.) 25-36 (Springer New York, New York, NY, 2018).
176. Graffelman, J. & Moreno, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. **12**, 433-448 (2013).
177. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
178. Conomos, M.P., Miller, M.B. & Thornton, T.A. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* **39**, 276-293 (2015).
179. Gogarten, S.M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346-5348 (2019).
180. Legge, S.E. *et al.* A genome-wide association study in individuals of African ancestry reveals the importance of the Duffy-null genotype in the assessment of clozapine-related neutropenia. *Molecular Psychiatry* **24**, 328-337 (2019).
181. Huddart, R. *et al.* Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research. *Clinical Pharmacology & Therapeutics* **105**, 1256-1262 (2019).
182. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
183. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284-1287 (2016).
184. Stanaway, I.B. *et al.* The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol* **43**, 63-81 (2019).
185. Kappel, D.B. *et al.* Genomic Stratification of Clozapine Prescription Patterns Using Schizophrenia Polygenic Scores. *Biol Psychiatry* (2022).
186. Gelman, A., Hill, J. & Vehtari, A. *Regression and Other Stories*, (Cambridge University Press, Cambridge, 2020).
187. Phillips, I.R. & Shephard, E.A. Drug metabolism by flavin-containing monooxygenases of human and mouse. *Expert Opin Drug Metab Toxicol* **13**, 167-181 (2017).
188. Tugnait, M. *et al.* N-oxygenation of clozapine by flavin-containing monooxygenase. *Drug Metab Dispos* **25**, 524-7 (1997).

189. Fang, J. Metabolism of clozapine by rat brain: the role of flavin-containing monooxygenase (FMO) and cytochrome P450 enzymes. *Eur J Drug Metab Pharmacokinet* **25**, 109-14 (2000).
190. Ring, B.J. *et al.* Identification of the human cytochromes P450 responsible for the in vitro formation of the major oxidative metabolites of the antipsychotic agent olanzapine. *J Pharmacol Exp Ther* **276**, 658-66 (1996).
191. Luo, J.P. *et al.* In vitro identification of the human cytochrome p450 enzymes involved in the oxidative metabolism of loxapine. *Biopharm Drug Dispos* **32**, 398-407 (2011).
192. Park, S.B., Jacob, P., Benowitz, N.L. & Cashman, J.R. Stereoselective metabolism of (S)-(-)-nicotine in humans: formation of trans-(S)-(-)-nicotine N-1'-oxide. *Chem Res Toxicol* **6**, 880-8 (1993).
193. Giannakopoulou, O. *et al.* The Genetic Architecture of Depression in Individuals of East Asian Ancestry: A Genome-Wide Association Study. *JAMA Psychiatry* **78**, 1258-1269 (2021).
194. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat Genet* **54**, 437-449 (2022).
195. Sen, K. Relevance of Pseudogenes to Human Genetic Disease. in *eLS* (2013).
196. Chen, X. *et al.* Re-recognition of pseudogenes: From molecular to clinical applications. *Theranostics* **10**, 1479-1499 (2020).
197. Cheetham, S.W., Faulkner, G.J. & Dinger, M.E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat Rev Genet* **21**, 191-201 (2020).
198. Shadrin, A.A. *et al.* Vertex-wise multivariate genome-wide association study identifies 780 unique genetic loci associated with cortical morphology. *Neuroimage* **244**, 118603 (2021).
199. van der Meer, D. *et al.* The genetic architecture of human cortical folding. *Sci Adv* **7**, eabj9446 (2021).
200. Lee, J.J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**, 1112-1121 (2018).
201. Sherva, R. *et al.* Genome-wide association study of phenotypes measuring progression from first cocaine or opioid use to dependence reveals novel risk genes. *Exploration of medicine* **2**, 60-73 (2021).
202. Hu, Y. *et al.* Multi-ethnic genome-wide association analyses of white blood cell and platelet traits in the Population Architecture using Genomics and Epidemiology (PAGE) study. in *BMC genomics* Vol. 22 432 (2021).
203. Uher, R. *et al.* Common Genetic Variation and Antidepressant Efficacy in Major Depressive Disorder: A Meta-Analysis of Three Genome-Wide Pharmacogenetic Studies. *American Journal of Psychiatry* **170**, 207-217 (2013).
204. Pan, Y. *et al.* Analysis of differential gene expression profile identifies novel biomarkers for breast cancer. *Oncotarget; Vol 8, No 70* (2017).
205. Seguin, L., Desgrosellier, J.S., Weis, S.M. & Cheresh, D.A. Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance. *Trends in Cell Biology* **25**, 234-240 (2015).
206. Ghouse, J. *et al.* Polygenic risk score for ACE-inhibitor-associated cough based on the discovery of new genetic loci. *Eur Heart J* (2022).

207. Vollert, J. *et al.* Genotypes of Pain and Analgesia in a Randomized Trial of Irritable Bowel Syndrome. *Frontiers in Psychiatry* **13**(2022).

208. Li, Q.S., Tian, C., Seabrook, G.R., Drevets, W.C. & Narayan, V.A. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. *Translational Psychiatry* **6**, e889-e889 (2016).

209. Pasman, J.A. *et al.* Genetic Risk for Smoking: Disentangling Interplay Between Genes and Socioeconomic Status. *Behav Genet* **52**, 92-107 (2022).

210. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet* **104**, 65-75 (2019).

211. Wootton, R.E. *et al.* Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychol Med* **50**, 2435-2443 (2020).

212. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231.e11 (2020).

213. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

214. Speed, D., Holmes, J. & Balding, D.J. Evaluating and improving heritability models using summary statistics. *Nat Genet* **52**, 458-462 (2020).

215. Iasevoli, F. *et al.* Relationships between early age at onset of psychotic symptoms and treatment resistant schizophrenia. *Early Interv Psychiatry* **16**, 352-362 (2022).

216. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217-234.e4 (2019).

217. Meltzer, H.Y. An overview of the mechanism of action of clozapine. *J Clin Psychiatry* **55 Suppl B**, 47-52 (1994).

218. Wheeler, A., Humberstone, V. & Robinson, G. Outcomes for schizophrenia patients with clozapine treatment: how good does it get? *J Psychopharmacol* **23**, 957-65 (2009).

219. Meltzer, H.Y. Clozapine: balancing safety with superior antipsychotic efficacy. *Clin Schizophr Relat Psychoses* **6**, 134-44 (2012).

220. Mistry, H. & Osborn, D. Underuse of clozapine in treatment-resistant schizophrenia. *Advances in Psychiatric Treatment* **17**, 250-255 (2011).

221. Bachmann, C.J. *et al.* International trends in clozapine use: a study in 17 countries. *Acta Psychiatrica Scandinavica* **136**, 37-51 (2017).

222. Downs, J. & Zinkler, M. Clozapine: national review of postcode prescribing. *Psychiatric Bulletin* **31**, 384-387 (2007).

223. Nielsen, J., Dahm, M., Lublin, H. & Taylor, D. Psychiatrists' attitude towards and knowledge of clozapine treatment. *J Psychopharmacol* **24**, 965-71 (2010).

224. Daod, E. *et al.* Psychiatrists' attitude towards the use of clozapine in the treatment of refractory schizophrenia: A nationwide survey. *Psychiatry Res* **275**, 155-161 (2019).

225. Ventura, A.M.B., Hayes, R.D. & Fonseca de Freitas, D. Ethnic disparities in clozapine prescription for service-users with schizophrenia-spectrum disorders: a systematic review. *Psychological Medicine* **52**, 2212-2223 (2022).

226. Das-Munshi, J., Bhugra, D. & Crawford, M.J. Ethnic minority inequalities in access to treatments for schizophrenia and schizoaffective disorders: findings from a nationally representative cross-sectional study. *BMC Medicine* **16**, 55 (2018).

227. Pillinger, T. *et al.* Comparative effects of 18 antipsychotics on metabolic function in patients with schizophrenia, predictors of metabolic dysregulation, and association

with psychopathology: a systematic review and network meta-analysis. *Lancet Psychiatry* **7**, 64-77 (2020).

228. Latif, Z., Jabbar, F. & Kelly, B.D. Clozapine and blood dyscrasia. *The Psychiatrist* **35**, 27-29 (2011).
229. Myles, N. *et al.* Meta-analysis examining the epidemiology of clozapine-associated neutropenia. *Acta psychiatrica Scandinavica.* **138**, 101-109 (2018).
230. Alvir, J.M.J., Lieberman, J.A., Safferman, A.Z., Schwimmer, J.L. & Schaaf, J.A. Clozapine-Induced Agranulocytosis -- Incidence and Risk Factors in the United States. *The New England journal of medicine.* **329**, 162-167 (1993).
231. Actor, J.K. 2 - Cells and Organs of the Immune System. in *Elsevier's Integrated Review Immunology and Microbiology (Second Edition)* (ed. Actor, J.K.) 7-16 (W.B. Saunders, Philadelphia, 2012).
232. Whiskey, E., Olofinjana, O. & Taylor, D. The importance of the recognition of benign ethnic neutropenia in black patients during treatment with clozapine: case reports and database study. *Journal of psychopharmacology* **25**, 842-845 (2011).
233. Taylor, D., Vallianatou, K., Whiskey, E., Dzahini, O. & MacCabe, J. Distinctive pattern of neutrophil count change in clozapine-associated, life-threatening agranulocytosis. *Schizophrenia* **8**, 21 (2022).
234. Schulte, P. Risk of clozapine-associated agranulocytosis and mandatory white blood cell monitoring. in *Ann Pharmacother*, Vol. 40 683-8 (United States, 2006).
235. Idänpään-Heikkilä, J., Alhava, E., Olkinuora, M. & Palva, I. Letter: Clozapine and agranulocytosis. *Lancet* **2**, 611 (1975).
236. Claghorn, J. *et al.* The Risks and Benefits of Clozapine versus Chlorpromazine. *Journal of clinical psychopharmacology.* **7**, 377???384-384 (1987).
237. Kane, J., Honigfeld, G., Singer, J. & Meltzer, H. Clozapine for the Treatment-Resistant Schizophrenic: A Double-blind Comparison With Chlorpromazine. *Archives of General Psychiatry* **45**, 789-796 (1988).
238. Thorn, C.F., Müller, D.J., Altman, R.B. & Klein, T.E. PharmGKB summary: clozapine pathway, pharmacokinetics. *Pharmacogenetics and Genomics* **28**(2018).
239. Pirmohamed, M., Williams, D., Madden, S., Templeton, E. & Park, B.K. Metabolism and bioactivation of clozapine by human liver in vitro. *J Pharmacol Exp Ther* **272**, 984-90 (1995).
240. Wagmann, L., Meyer, M.R. & Maurer, H.H. What is the contribution of human FMO3 in the N-oxygenation of selected therapeutic drugs and drugs of abuse? *Toxicology Letters* **258**, 55-70 (2016).
241. Zhang, W.V., Esposito, F., Edwards, R.J., Ramzan, I. & Murray, M. Interindividual Variation in Relative CYP1A2/3A4 Phenotype Influences Susceptibility of Clozapine Oxidation to Cytochrome P450-Specific Inhibition in Human Hepatic Microsomes. *Drug Metabolism and Disposition* **36**, 2547 (2008).
242. Schaber, G., Stevens, I., Gaertner, H.J., Dietz, K. & Breyer-Pfaff, U. Pharmacokinetics of clozapine and its metabolites in psychiatric patients: plasma protein binding and renal clearance. *Br J Clin Pharmacol* **46**, 453-9 (1998).
243. Williams, D.P., Pirmohamed, M., Naisbitt, D.J., Uetrecht, J.P. & Park, B.K. Induction of Metabolism-Dependent and -Independent Neutrophil Apoptosis by Clozapine. *Molecular Pharmacology* **58**, 207 (2000).
244. Geib, T., Thulasingam, M., Haeggström, J.Z. & Sleno, L. Investigation of Clozapine and Olanzapine Reactive Metabolite Formation and Protein Binding by Liquid

Chromatography-Tandem Mass Spectrometry. *Chemical Research in Toxicology* **33**, 2420-2431 (2020).

245. Liu, Z.C. & Uetrecht, J.P. Clozapine is oxidized by activated human neutrophils to a reactive nitrenium ion that irreversibly binds to the cells. *Journal of Pharmacology and Experimental Therapeutics* **275**, 1476 (1995).

246. Hasegawa, M., Cola, P.A. & Meltzer, H.Y. Plasma clozapine and desmethylclozapine levels in clozapine-induced agranulocytosis. *Neuropsychopharmacology* **11**, 45-7 (1994).

247. Oyewumi, L.K., Cernovsky, Z.Z., Freeman, D.J. & Streiner, D.L. Relation of blood counts during clozapine treatment to serum concentrations of clozapine and nor-clozapine. *Can J Psychiatry* **47**, 257-61 (2002).

248. Smith, R.L. *et al.* Correlation Between Serum Concentrations of N-Desmethylclozapine and Granulocyte Levels in Patients with Schizophrenia: A Retrospective Observational Study. *CNS Drugs* **31**, 991-997 (2017).

249. Vaquero-Baez, M. *et al.* Clozapine and desmethylclozapine: correlation with neutrophils and leucocytes counting in Mexican patients with schizophrenia. *BMC Psychiatry* **19**, 295 (2019).

250. Pardiñas, A.F. *et al.* Pharmacogenomic Variants and Drug Interactions Identified Through the Genetic Analysis of Clozapine Metabolism. *American Journal of Psychiatry* **176**, 477-486 (2019).

251. Phillips, C. *et al.* Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics* **1**, 273-280 (2007).

252. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nature protocols* **5**, 1564-1573 (2010).

253. RJ, F. A practical approach to clozapine therapeutic drug monitoring. *CMHP Bulletin* **2**, 4-5 (2010).

254. Ellison, J.C. & Dufresne, R.L. A review of the clinical utility of serum clozapine and norclozapine levels. *Mental Health Clinician* **5**, 68-73 (2015).

255. Couchman, L., Morgan, P.E., Spencer, E.P. & Flanagan, R.J. Plasma Clozapine, Norclozapine, and the Clozapine:Norclozapine Ratio in Relation to Prescribed Dose and Other Factors: Data From a Therapeutic Drug Monitoring Service, 1993–2007. *Therapeutic Drug Monitoring* **32**, 438-447 (2010).

256. Prabhakar, M., Ershler, W.B. & Longo, D.L. Bone marrow, thymus and blood: changes across the lifespan. *Aging Health* **5**, 385-393 (2009).

257. Lu, D., Xie, Q. & Wu, B. N-glucuronidation catalyzed by UGT1A4 and UGT2B10 in human liver microsomes: Assay optimization and substrate identification. *Journal of Pharmaceutical and Biomedical Analysis* **145**, 692-703 (2017).

258. Dragovic, S., Gunness, P., Ingelman-Sundberg, M., Vermeulen, N.P.E. & Commandeur, J.N.M. Characterization of Human Cytochrome P450s Involved in the Bioactivation of Clozapine. *Drug Metabolism and Disposition* **41**, 651 (2013).

259. Smith, R.L. *et al.* Identification of a novel polymorphism associated with reduced clozapine concentration in schizophrenia patients—a genome-wide association study adjusting for smoking habits. *Translational Psychiatry* **10**, 198 (2020).

260. Rostami-Hodjegan, A. *et al.* Influence of dose, cigarette smoking, age, sex, and metabolic activity on plasma clozapine concentrations: a predictive model and nomograms to aid clozapine dose adjustment and to assess compliance in individual patients. *J Clin Psychopharmacol* **24**, 70-8 (2004).

261. Couchman, L., Bowskill, S.V., Handley, S., Patel, M.X. & Flanagan, R.J. Plasma clozapine and norclozapine in relation to prescribed dose and other factors in patients aged <18 years: data from a therapeutic drug monitoring service, 1994-2010. *Early Interv Psychiatry* **7**, 122-30 (2013).
262. Kraft, P. Curses--winner's and otherwise--in genetic epidemiology. *Epidemiology* **19**, 649-51; discussion 657-8 (2008).
263. Ruiz-Linares, A. *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet* **10**, e1004572 (2014).
264. Kelly, D.L. *et al.* Clozapine underutilization and discontinuation in African Americans due to leucopenia. *Schizophr Bull* **33**, 1221-4 (2007).
265. Haddy, T.B., Rana, S.R. & Castro, O. Benign ethnic neutropenia: what is a normal absolute neutrophil count? *J Lab Clin Med* **133**, 15-22 (1999).
266. Griffith, R.W. & Saameli, K. Letter: Clozapine and agranulocytosis. *Lancet* **2**, 657 (1975).
267. Atkin, K. *et al.* Neutropenia and agranulocytosis in patients receiving clozapine in the UK and Ireland. *Br J Psychiatry* **169**, 483-8 (1996).
268. Siskind, D. *et al.* Consensus statement on the use of clozapine during the COVID-19 pandemic. *J Psychiatry Neurosci* **45**, 222-223 (2020).
269. Pandarakalam, j.p. & Paul, J. Revisiting Clozapine in a Setting of COVID-19. *American Journal of Psychiatry and Neuroscience* **8**, 50-58 (2020).
270. Singh, H., Dubin, W.R. & Kaur, S. Drug interactions affecting clozapine levels. *Journal of Psychiatric Intensive Care* **11**, 52-65 (2015).
271. Mican, L.M. What to do when your patient who takes clozapine enters a smoke-free facility. *Current psychiatry* **13**, 47-48 (2014).
272. Raaska, K., Raitasuo, V., Laitila, J. & Neuvonen, P.J. Effect of caffeine-containing versus decaffeinated coffee on serum clozapine concentrations in hospitalised patients. *Basic Clin Pharmacol Toxicol* **94**, 13-8 (2004).
273. Thien, K. & O'Donoghue, B. Delays and barriers to the commencement of clozapine in eligible people with a psychotic disorder: A literature review. *Early Intervention in Psychiatry* **13**, 18-23 (2019).
274. Govind, R., Fonseca de Freitas, D., Pritchard, M., Hayes, R.D. & MacCabe, J.H. Clozapine treatment and risk of COVID-19 infection: retrospective cohort study. *Br J Psychiatry* **219**, 368-374 (2021).
275. Charlson, F.J. *et al.* Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophrenia Bulletin* **44**, 1195-1203 (2018).
276. Halvorsrud, K., Nazroo, J., Otis, M., Brown Hajdukova, E. & Bhui, K. Ethnic inequalities in the incidence of diagnosis of severe mental illness in England: a systematic review and new meta-analyses for non-affective and affective psychoses. *Social Psychiatry and Psychiatric Epidemiology* **54**, 1311-1323 (2019).
277. Selten, J.-P., Cantor-Graae, E. & Kahn, R.S. Migration and schizophrenia. *Current Opinion in Psychiatry* **20**(2007).
278. Cardno, A.G. & Owen, M.J. Genetic Relationships Between Schizophrenia, Bipolar Disorder, and Schizoaffective Disorder. *Schizophrenia Bulletin* **40**, 504-515 (2014).

279. Wray, N.R., Lee, S.H. & Kendler, K.S. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur J Hum Genet* **20**, 668-74 (2012).

280. Suzuki, T. *et al.* Defining treatment-resistant schizophrenia and response to antipsychotics: A review and recommendation. *Psychiatry Research* **197**, 1-6 (2012).

281. Üçok, A. *et al.* Correlates of Clozapine Use after a First Episode of Schizophrenia: Results From a Long-term Prospective Study. *CNS Drugs* **30**, 997-1006 (2016).

282. Wagner, E., McMahon, L., Falkai, P., Hasan, A. & Siskind, D. Impact of smoking behavior on clozapine blood levels – a systematic review and meta-analysis. *Acta Psychiatrica Scandinavica* **142**, 456-466 (2020).

283. Yartsev, A. & Peisah, C. Caffeine-clozapine interaction associated with severe toxicity and multiorgan system failure: a case report. *BMC Psychiatry* **21**, 192 (2021).

284. Taylor, D. Pharmacokinetic interactions involving clozapine. *British Journal of Psychiatry* **171**, 109-112 (1997).

285. Andersohn, F., Konzen, C. & Garbe, E. Systematic Review: Agranulocytosis Induced by Nonchemotherapy Drugs. *Annals of Internal Medicine* **146**, 657-665 (2007).

286. Ayano, G. *et al.* Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in specialized psychiatry centers in Ethiopia. *Annals of General Psychiatry* **20**, 10 (2021).

287. Singh, T. & Rajput, M. Misdiagnosis of bipolar disorder. *Psychiatry (Edgmont)* **3**, 57-63 (2006).

288. Yamada, Y., Matsumoto, M., Iijima, K. & Sumiyoshi, T. Specificity and Continuity of Schizophrenia and Bipolar Disorder: Relation to Biomarkers. *Current Pharmaceutical Design* **26**, 191-200 (2020).

289. Nasrallah, H.A. Consequences of misdiagnosis: inaccurate treatment and poor patient outcomes in bipolar disorder. *J Clin Psychiatry* **76**, e1328 (2015).

290. Nelson, M.R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856-60 (2015).

291. Shah, P. *et al.* The impact of delay in clozapine initiation on treatment outcomes in patients with treatment-resistant schizophrenia: A systematic review. *Psychiatry Res* **268**, 114-122 (2018).

292. John, A.P., Ko, E.K.F. & Dominic, A. Delayed Initiation of Clozapine Continues to Be a Substantial Clinical Concern. *Can J Psychiatry* **63**, 526-531 (2018).

293. Üçok, A. *et al.* Delayed initiation of clozapine may be related to poor response in treatment-resistant schizophrenia. *Int Clin Psychopharmacol* **30**, 290-5 (2015).

294. Howes, O.D., Thase, M.E. & Pillinger, T. Treatment resistance in psychiatry: state of the art and new directions. *Mol Psychiatry* **27**, 58-72 (2022).

295. Hálfdánarson, Ó. *et al.* International trends in antipsychotic use: A study in 16 countries, 2005–2014. *European Neuropsychopharmacology* **27**, 1064-1076 (2017).

296. Smith, S., Hay, e.H., Farhat, N. & Rekaya, R. Genome wide association studies in presence of misclassified binary responses. *BMC Genet* **14**, 124 (2013).

297. Boycott, K.M. *et al.* Autosomal-Recessive Intellectual Disability with Cerebellar Atrophy Syndrome Caused by Mutation of the Manganese and Zinc Transporter Gene SLC39A8. *Am J Hum Genet* **97**, 886-93 (2015).

298. Wu, Y. *et al.* Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Transl Psychiatry* **10**, 209 (2020).

299. Cheour, M., Zribi, H., Abdelhak, S., Drira, S. & Ben Osman, A. [Darier's disease: an evaluation of its neuropsychiatric component]. *Encephale* **35**, 32-5 (2009).

300. Goes, F.S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet* **168**, 649-59 (2015).

301. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat Genet* **49**, 1576-1583 (2017).

302. Ikeda, M. *et al.* Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. *Schizophr Bull* **45**, 824-834 (2019).

303. Popovic, D. *et al.* Rab GTPase-activating proteins in autophagy: regulation of endocytic and autophagy pathways by direct binding to human ATG8 modifiers. *Mol Cell Biol* **32**, 1733-44 (2012).

304. Smith, S.M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat Neurosci* **24**, 737-745 (2021).

305. Nagel, M., Speed, D., van der Sluis, S. & Østergaard, S.D. Genome-wide association study of the sensitivity to environmental stress and adversity neuroticism cluster. *Acta Psychiatrica Scandinavica* **141**, 476-478 (2020).

306. Baselmans, B. *et al.* The Genetic and Neural Substrates of Externalizing Behavior. *Biol Psychiatry Glob Open Sci* **2**, 389-399 (2022).

307. Bigdeli, T.B. *et al.* Genome-Wide Association Studies of Schizophrenia and Bipolar Disorder in a Diverse Cohort of US Veterans. *Schizophrenia bulletin* **47**, 517-529 (2021).

308. Lam, M. *et al.* Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. *Am J Hum Genet* **105**, 334-350 (2019).

309. GrandPré, T., Nakamura, F., Vartanian, T. & Strittmatter, S.M. Identification of the Nogo inhibitor of axon regeneration as a Reticulon protein. *Nature* **403**, 439-44 (2000).

310. Howard, D.M. *et al.* Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat Commun* **9**, 1470 (2018).

311. Coleman, J.R.I. *et al.* Genome-wide gene-environment analyses of major depressive disorder and reported lifetime traumatic experiences in UK Biobank. *Mol Psychiatry* **25**, 1430-1446 (2020).

Appendix

Appendix 1: Discussion of Top Genes from Research Chapter 1

Identified Genes

In total, 27 loci were identified in this chapter. However, only a subsection will be discussed. For three of the CC-GWAS loci, the full posterior probability of the corresponding FINEMAP locus from PGC3 SCZ is accounted for by CC-GWAS significant SNPs and will be discussed first. For a further four, over 50% of the posterior probability is accounted for CC-GWAS significant SNPs, and the full credible SNP set of the corresponding FINEMAP locus falls within the base pair boundaries of a CC-GWAS locus and will also be discussed. With the exception of two (STAG1 and TBC1D5), each of these genes are in the final prioritised gene list generated as part of the work PGC3 SCZ. Finally, the locus for which there was no overlap with a PGC3 FINEMAP locus will also be discussed.

SLC39A8

Solute Carrier Family 39 Member 8 is a longstanding schizophrenia associated gene, having been implicated in multiple GWAS of schizophrenia prior to PGC3 SCZ, where genome wide significance was retained [95-97]. The SLC39A8 protein is an electroneutral transporter protein, located within the plasma membrane and integral to the uptake of zinc and manganese. It has also been hypothesised that SLC39A8 may mediate the extracellular uptake of manganese by the cells of the blood-brain-barrier [297]. This protein has been extensively implicated in GWAS of neuroimaging measures, measures of brain volume and cortical thickness, with reported association counts of 48, 46 and 35 on GWAS catalog respectively, as well as a wide range of other phenotypes. SLC39A8 was also a significant association within the original CC-GWAS paper, in both the schizophrenia versus anorexia nervosa and schizophrenia versus ADHD analyses [115]. It has not previously been implicated in bipolar disorder specifically, and follow-up analysis in studies assessing the pleiotropy of psychiatric disorders concluded that the association was specific to schizophrenia [119,298]

R3HDM2

R3H Domain Containing 2 is another well-documented schizophrenia risk gene [96,97]. Limited information is available regarding the function of the protein product of this gene, although it is predicted to enable RNA binding activity. Similarly to SLC39A8, R3HDM2 has been reported in a wide range of phenotypes, most commonly studies of traits related to cholesterol measures. It has not previously been reported in a GWAS of bipolar disorder, or other psychiatric phenotypes.

ATP2A2

ATPase Sarcoplasmic/Endoplasmic Reticulum Ca2+ Transporting 2 is an intracellular pump that catalyses the hydrolysis of ATP in conjunction with the transport of calcium from the cytosol to the endo- / sarcoplasmic reticulum. It is the causative gene in the autosomal dominant disorder Darier disease, which is characterised by wart-like blemishes on the body, and there is longstanding literature surrounding the presence of neuropsychiatric symptoms and disorders in cases with Darier disease [299]. It is a well-established schizophrenia risk gene, with genome wide significant associations reported in multiple GWAS [95,96,300-302], and was also located within a genome wide significant locus in both the schizophrenia versus anorexia nervosa and schizophrenia versus autism spectrum disorder

CC-GWAS analyses [115]. The gene was implicated in a previous bipolar disorder GWAS when the sample was restricted to just bipolar disorder type 1 [130], but was not genome wide significant in PGC3 BD or other previous bipolar disorder GWAS.

TBC1D5

TBC1 Domain Family Member 5 plays a key role in autophagy [303], a process by which misfolded / dysfunctional cellular components are broken down and removed from the cell. Damaged cellular components are packaged into an autophagosome, which then merges with a lysosome (a cell organelle containing hydrolytic enzymes), to degrade the contents and release the broken-down components back into the cytosol. TBC1D5 has been implicated in several previous schizophrenia GWAS in addition to PGC3 SCZ [95,96,300-302], and was also a genome wide significant gene in the CC-GWAS analysis of schizophrenia versus Tourette's and other tic disorders [115]. In addition, it has been implicated in GWAS of educational attainment [194,200]. It has not been previously implicated in any GWAS of bipolar disorder.

KLF6

KLF Transcription Factor 6 is a member of the Kruppel-like family of proteins and functions as a transcription factor, integral to the process of converting DNA into messenger RNA by binding to the upstream regulatory elements of genes. Previous to PGC3 SCZ, this gene was only reported to be associated with schizophrenia risk in one GWAS [97], and has not previously been reported in the bipolar disorder literature.

CRHR1

Corticotropin Releasing Hormone (CRH) Receptor 1 is a G-protein coupled receptor that binds to CRH, the primary hormone involved in stress response. CRH stimulates the synthesis of adrenocorticotropic hormone (ACTH) in the pituitary gland, as part of the hypothalamic-pituitary-adrenal (HPA) axis. It is primarily secreted by the paraventricular nucleus (PVN) of the hypothalamus in response to adverse stress. CRHR1 has been implicated in a wide range of neuropsychiatric phenotypes, including multiple measures of white matter integrity [304], neuroticism [305] and aggressive behaviour in ADHD [306]. It has been implicated in schizophrenia in one GWAS prior to PGC3 SCZ [307], and has not been previously implicated in bipolar disorder in any published GWAS.

STAG1

Stromal Antigen 1 encodes for a component of the protein complex cohesin, which is integral to the process of cell division and replication. Cohesin ensures that sister chromatids remain connected to each other throughout metaphase and facilitates the attachment of cytoskeletal structures called spindles to the chromatids. Both of these functions are crucial to the proper and full segregation of chromatids to opposite poles of the cell undergoing cell division, resulting in both newly formed daughter cells having a full complement of chromosomes at the completion of telophase. STAG1 has been implicated in GWAS of schizophrenia multiple times prior to PGC3 SCZ [95-97,300-302]. It was also implicated in the CC-GWAS analysis of schizophrenia versus autism spectrum disorder [115], as well as MTAG analysis of schizophrenia [298] and a study that utilised a method known as ASSET (Association analysis based on subsets) to investigate the pleiotropy between cognition, educational attainment and schizophrenia [308]. A variant within STAG1 surpassed genome wide significance in a study of pleiotropy shared between disorders, but it was concluded in follow-up analysis that this variant was specific to schizophrenia and was not associated with any of the other seven disorders under study [119].

RTN4

Reticulon 4 was the closest gene by SNP position for the locus that displayed no overlap with a PGC3 SCZ FINEMAP locus and was not genome wide significant in either input GWAS. Reticulon 4 is a neurite outgrowth inhibitor and is thought to be one of the reasons CNS axons are unable to regenerate as readily as axons in the peripheral nervous system [309]. It has not been implicated in GWAS of psychiatric disorders previously, with the exception of nominal associations in GWAS of major depression [310,311]

Appendix 2: Full PRS results in Cardiff COGS, Research Chapter 1

Table 20: Full PRS results from research chapter 1, age at onset of psychosis

Age at Onset of Psychosis	Schizophrenia PRS			CC-GWAS PRS			Bipolar Disorder PRS		
(Intercept)	9.705***	9.628***	9.858***	9.854***	9.772***	9.931***	9.891***	9.899***	9.961***
Polygenic Risk Scores									
Pt_1	-0.977**			-0.896**			-0.289		
Pt_0.05		-1.054**			-0.829**			-0.132	
Pt_5e.08			-0.459			0.002			-0.328
Covariates									
PC1	3.795	4.834	2.116	2.031	3.692	-1.4	-0.696	-1.074	-0.227
PC2	30.944	41.325	52.419	21.306	33.37	25.087	21.898	24.391	36.519
PC3	90.847	88.575	83.058	89.844	91.134	84.616	89.236	87.828	76.859
PC4	-3.83	-4.279	-3.887	-4.337	-4.85	-4.26	-4.517	-4.467	-3.963
PC5	24.734*	26.161**	24.284*	26.716**	27.482**	23.547*	23.188*	23.242*	23.856*
male_sex	-1.298*	-1.293*	-1.379*	-1.387*	-1.387*	-1.424*	-1.413*	-1.414*	-1.412*
Age_at_Interview	0.373***	0.374***	0.369***	0.372***	0.372***	0.370***	0.370***	0.370***	0.368***
R2	0.248	0.249	0.24	0.246	0.245	0.238	0.238	0.238	0.239
Adj. R2	0.239	0.241	0.231	0.238	0.236	0.229	0.23	0.229	0.23
Num. obs.	720	720	720	720	720	720	720	720	720
*** = p < 0.001; ** = p < 0.01; * = p < 0.05									

Table 21: Full PRS results from research chapter 1, negative symptoms of diminished expressivity, based upon the symptoms of affective flattening (a loss or lack of emotional expressiveness) and alogia (difficulty speaking)

Negative Symptoms of Diminished Expressivity	Schizophrenia PRS			CC-GWAS PRS			Bipolar Disorder PRS		
(Intercept)	**-0.416***	**-0.401***	**-0.426***	**-0.428****	**-0.414***	**-0.436****	**-0.432****	**-0.430****	**-0.442****
Polygenic Risk Scores									
Pt_1	**0.098****			**0.090***			0.033		
Pt_0.05		**0.125*****			**0.116****			0.028	
Pt_5e.08			0.07			0.028			0.059
Covariates									
PC1	3.443	3.204	3.417	3.601	3.195	3.768	3.856	3.874	3.729
PC2	36.056	34.55	32.47	36.935	35.084	35.397	36.84	36.675	34.481
PC3	11.892	12.294	11.986	12.089	11.587	12.554	11.652	11.477	13.417
PC4	-0.281	-0.236	-0.283	-0.221	-0.148	-0.193	-0.2	-0.185	-0.28
PC5	-1.221	-1.411	-1.204	-1.417	-1.646	-1.083	-1.053	-1.029	-1.149
male_sex	**0.185***	**0.182***	**0.190***	**0.194****	**0.191****	**0.196****	**0.195****	**0.195****	**0.194****
Age_at_Interview	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.005
R2	0.025	0.03	0.02	0.023	0.028	0.017	0.017	0.017	0.019
Adj. R2	0.015	0.02	0.01	0.013	0.018	0.006	0.006	0.006	0.009
Num. obs.	755	755	755	755	755	755	755	755	755
*** = p < 0.001; ** = p < 0.01; * = p < 0.05									

Table 22: Full PRS results from research chapter 1, disorganised symptoms, made up of positive thought disorder (derailment, pressure of speech) and inappropriate affect (display of reactions that do not match the situation)

Disorganised Symptoms	**Schizophrenia PRS**			**CC-GWAS PRS**			**Bipolar Disorder PRS**		
(Intercept)	-0.079	-0.057	-0.099	-0.101	-0.085	-0.114	-0.103	-0.092	-0.12
Polygenic Risk Scores									
Pt_1	**0.170*****			**0.132*****			**0.076***		
Pt_0.05		**0.200*****			**0.151*****			**0.096****	
Pt_5e.08			**0.096***			0.041			0.066
Covariates									
PC1	5.472	5.154	5.624	5.842	5.37	6.085	6.15	6.118	6.099
PC2	45.363	42.998	40.63	46.77	44.295	44.511	46.911	46.736	43.85
PC3	-7.245	-6.584	-7.074	-6.925	-7.591	-6.24	-7.852	-8.845	-5.449
PC4	-1.891	-1.813	-1.874	-1.788	-1.695	-1.748	-1.737	-1.655	-1.857
PC5	0.011	-0.277	0.082	-0.24	-0.485	0.249	0.325	0.451	0.171
male_sex	0.073	0.07	0.084	0.089	0.087	0.093	0.091	0.087	0.091
Age_at_Interview	-0.002	-0.003	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002
R2	0.036	0.046	0.017	0.025	0.03	0.011	0.014	0.018	0.013
Adj. R2	0.026	0.036	0.007	0.015	0.02	0	0.004	0.007	0.003
Num. obs.	755	755	755	755	755	755	755	755	755
*** = p < 0.001; ** = p < 0.01; * = p < 0.05									

Table 23:Full PRS results from research chapter 1, Use of other non-prescription drugs. According to the ratings guide, 'Other Non-Prescription Drugs include: amphetamine, cocaine, heroin, LSD, solvents, benzodiazepine and ecstasy. As with cannabis use, 'regular use' refers to use that is persistent for one month or repeatedly within one year (i.e., at least once a week for at least 6 months of the year)

Non-Prescription Drugs	Schizophrenia PRS			CC-GWAS PRS			Bipolar Disorder PRS		
(Intercept)	20.992	19.983	17.432	23.618	22.711	20.071	20.964	24.181	20.539
Polygenic Risk Scores									
Pt_1	0.001			0.049			-0.048		
Pt_0.05		-0.034			0.05			-0.086	
Pt_5e.08			-0.1			-0.177*			-0.034
Covariates									
PC1	14.402	14.516	15.019	14.378	14.24	15.313	14.502	14.511	14.52
PC2	108.985	108.797	113.607	110.277	109.517	114.567	108.172	107.542	110.011
PC3	36.004	36.291	36.813	36.076	36.039	34.497	36.893	38.151	34.983
PC4	-1.282	-1.27	-1.181	-1.192	-1.158	-1.389	-1.272	-1.341	-1.255
PC5	0.795	0.864	0.954	0.646	0.597	0.712	0.736	0.584	0.844
male_sex	0.926***	0.931***	0.940***	0.925***	0.924***	0.940***	0.928***	0.932***	0.928***
Age_at_Interview	-0.074	-0.074	-0.073	-0.076	-0.075	-0.075	-0.074	-0.076	-0.074
YOB	-0.01	-0.01	-0.008	-0.011	-0.011	-0.01	-0.01	-0.012	-0.01
AIC	789.752	789.616	788.538	789.478	789.449	785.837	789.486	788.884	789.606
BIC	835.628	835.492	834.414	835.353	835.325	831.713	835.361	834.76	835.481
Log Likelihood	-384.876	-384.808	-384.269	-384.739	-384.725	-382.919	-384.743	-384.442	-384.803
Deviance	769.752	769.616	768.538	769.478	769.449	765.837	769.486	768.884	769.606
Num. obs.	726	726	726	726	726	726	726	726	726
*** = p < 0.001; ** = p < 0.01; * = p < 0.05									