Chapter 1

Introduction

1.1 Schizophrenia

Schizophrenia is a highly complex, common and debilitating psychiatric illness characterised by a breakdown of how a person perceives and responds to the reality around them. The clinical symptoms for this disorder have changed and evolved since its first description as *dementia praecox* by Emil Kraepelin in 1887, but its most striking, and perhaps defining, features remain its positive symptoms, comprising hallucinations (false perceptions), delusions (irrational beliefs), and disorganised speech and behaviour. This contrasts with schizophrenia's negative symptoms where there is an absence of normal social function, typically in the form of social withdrawal and lack of motivational drive. The prognosis of individuals with schizophrenia varies dramatically: approximately half of patients have poor outcomes one year after their first episode [1, 2], and around 20% suffer from chronic relapses and severe symptoms for the remainder of their lives [3, 4]. Despite its severe symptoms and varied prognosis, schizophrenia is common in the general population with a lifetime risk of $\sim$0.7%, and it is not surprisingly that the disorder has substantial societal and personal costs [5]. Patients with schizophrenia rarely fulfil their full occupational potential, with over 80% of affected individuals permanently unemployed [6, 7]. From reasons ranging from suicides to metabolic disease from antipsychotic use, people with schizophrenia have a decreased life expectancy of 12 to 15 years when compared to the general population [8]. Furthermore, individuals with schizophrenia are perceived with unwarranted social stigma and are unfairly described as unpredictable and dangerous [9]. This, combined with already difficult clinical outcomes, contributes to the isolation and distress faced by people with mental illnesses.

Substantial progress has been made by large-scale epidemiological, imaging, functional, and genetic studies to elucidate the nature of schizophrenia in the past few decades. In this

Chapter, I first lay out the diagnostic criteria for schizophrenia, and describe the clinical heterogeneity in its presentation. I then describe how symptoms are currently managed, the prognosis facing people with schizophrenia, and the current prevalence, incidence, and burden of disease. I briefly discuss environmental exposures that have been shown to increase risk of schizophrenia, before describing in detail the varied and complex contributions to schizophrenia's genetic architecture. I then discuss the arrival of sequencing as a means of studying rare variants, the first results from these studies, and the biological insights that have emerged. Finally, I briefly outline the aspects of schizophrenia genetics that I attempt to address in this Thesis.

1.1.1 Diagnostic criteria and clinical heterogeneity in presentation

The operational diagnostic criteria for schizophrenia are defined in the five editions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) [10]. These definitions are built on a number of historical descriptions of schizophrenia, incorporating Kraepelin's focus on its relapsing and deteriorating course, Bleuler's emphasis on negative symptoms such as social withdrawal and detachment from reality, and Schneider's first-rank symptoms that laid out the core features of psychotic manifestation [11]. As clinical research in psychiatric disorders advanced, the characteristics by which schizophrenia was defined also evolved, highlighting different aspects of these historical descriptions [12]. The most recent version, DSM-V, defines five core symptoms for schizophrenia: hallucinations, delusions, disorganised speech, grossly disorganised or catatonic behaviour, and negative symptoms [13]. For a full diagnosis of schizophrenia, the DSM-V requires the presence of at least two of these core symptoms over a period of six months with at least one month of active symptoms, and at least one of these symptoms must be psychotic (e.g. hallucinations, delusions, or disorganised speech). Cognitive deficits are regarded as a characteristic feature of schizophrenia, with 3.7% to 5.2% of schizophrenia patients given an additional diagnosis of intellectual disability [14]. However, cognitive impairment was not included as a diagnostic criterion in DSM-V, as it did not sufficiently distinguish between schizophrenia and other psychiatric disorders [13]. Because there are no diagnostic biomarkers or physiological tests for schizophrenia, diagnoses are only made by a psychiatrist with careful examination of the individual's behaviour and recent history.

From this definition, it is clear that a diagnosis of schizophrenia represents a wide range of possible symptoms occurring with varying duration and severity resulting in different long-term outcomes. Because of the breadth of its diagnostic criteria, schizophrenia can be perceived as a syndromic concept, one that could even encompass some number of biological disorders of brain with different underlying etiologies but sharing similar symptomatic

manifestation [4, 15]. In addition to being broadly defined and heterogeneous, many of its core symptoms are not unique and are observed in a number of other psychiatric disorders. The Schneiderian first-rank symptoms, frequently used to describe the primary presentation of schizophrenia, have also been observed in patients with bipolar disorders [16]. Psychotic symptoms are also present, albeit less frequently, in bipolar disorders and major depression [17]. Major depressive disorder with severe psychotic symptoms is diagnosed as psychotic depression [18], and schizophrenia with prominent mood symptoms is diagnosed as schizoaffective disorder. In addition, differential diagnoses like schizophreniform, psychotic and delusional disorders may be given instead when a full diagnosis of schizophrenia is not satisfied, despite these conditions sharing a number of symptoms characteristic of schizophrenia [10]. Indeed, psychotic symptoms may not even originate from underlying psychiatric illness: hallucinations and delusions can be induced by substance abuse and other general medical conditions [17], and are observed at a sub-clinical level in 5% of individuals without a psychiatric diagnosis [19]. Finally, individuals with schizophrenia often have additional symptoms that generally define other psychiatric disorders, including depression, anxiety, substance abuse, obsessive-compulsive disorder, panic disorder, and post-morbid cognitive impairment [20]. These observations suggest that while the current categorical classification for schizophrenia may be a clinically convenient and useful concept, it overlooks the symptomatic and possible etiological overlap with other psychiatric conditions.

1.1.2 Disease management and prognosis

Following a clinical diagnosis of schizophrenia, patients are generally prescribed antipsychotic medication to control positive symptoms. Despite many iterations of these drugs over the years, they are designed to target a single biological mechanism - the blocking of dopamine D2 receptor (D2R) activity [21]. The first generation of antipsychotics, such as chlorpromazine (low potency), fluphenazine and haloperidol (high potency), are effective in addressing positive symptoms like hallucinations and delusions, but can cause severe extra-pyramidal or movement-related side-effects, including tremors, rigidity, and spasms [22]. The second generation of antipsychotics, such as aripiprazole, olanzapine and risperidone, were developed in the 1980s to target D2R with lower affinity and also disrupt other neuronal receptors (e.g. serotonin, epinephrine). While these have reduced motor side-effects, second-generation antipsychotics have significant metabolic side-effects, including increased rates of weight gain, dyslipidemia, and diabetes [21, 23]. A first-generation antipsychotic, clozapine, is prescribed in the case of treatment-resistant schizophrenia, but its use is limited by its severe side-effects, one of which is agranulocytosis, or lowered white blood count, that can be potentially fatal [24, 22].

Antipsychotic drugs generally have some efficacy in treating the core psychotic symptoms, but a number of key issues emerge from their use in schizophrenia. First, both generations of antipsychotic drugs appear to have limited effectiveness in addressing the negative and cognitive symptoms of schizophrenia [21]. Even if positive symptoms are treated, many patients still suffer from a lack of motivation and social withdrawal, preventing them from resuming normal lives. In addition, for reasons ranging from severe side-effects, limited perceived efficacy, and social stigma, antipsychotic use in chronic schizophrenia suffers from substantial drop-out rates, with reports stating that 74% of patients discontinued their assigned treatment before the end of an 18-month study [24]. These high drop rates contribute to a higher risk of relapse of psychotic symptoms.

There is substantial patient heterogeneity in the prognosis of schizophrenia, with some patients showing signs of recovery while others following a chronic and deteriorating course. A five-year follow-up study of schizophrenia patients after the first psychotic episode demonstrated that around half showed some signs of symptom remission, and another quarter had adequate social functioning during this time [2]. Only 13.7% met the full criteria for a prolonged recovery. A long-term study following patients for 15 to 25-years supported this result, and similarly found that about 50% of cases have reasonable outcomes while only 16% achieve a late-phase recovery [1]. The increased mortality in schizophrenia has been attributed to a number of causes of death: individuals with schizophrenia are at a greater risk of dying from a large range of natural causes (cardiovascular diseases, digestive diseases, endocrine diseases, infectious diseases, and respiratory diseases), and strikingly, have a 1.73-fold higher risk of accidents and a 12-fold higher risk of suicide [8]. A number of these may not be mechanistically related to the biology of schizophrenia, but rather due to an inability or aversion to accessing health care, or unhealthy lifestyle choices that generally increase risk of cardiovascular disease [25]. Despite better outcomes than previously thought, broad progress in therapeutic development and societal support is needed to improve the prognosis of individuals with schizophrenia.

1.1.3 Epidemiology and global burden of disease

The lifetime prevalence for schizophrenia, or the proportion of people who had schizophrenia in a study population, is estimated to be four in every one thousand individuals [5]. The lifetime morbid risk, or the proportion of people who had or will eventually develop schizophrenia, is 7.2 per 1,000 individuals. In layman's terms, around seven in every thousand individuals will be diagnosed with schizophrenia in their lives. Interestingly, there is substantial variability in these estimates from different studies: a meta-analysis found that the first and third quartile of estimates of lifetime morbid risk is 4.7 and 17.2 per 1,000

respectively [5]. This variation is observed between countries, and even between regional sites and neighbourhoods [26]. Estimates of incidence and lifetime risk also exclude other psychotic disorders, which are relatively common in the general population. In a population survey in Finland, the following lifetime prevalences are observed: 0.32% for schizoaffective disorder, 0.07% for schizophreniform disorder, 0.18% for delusional disorder, and 0.42% for substance induced disorders [17]. Combined, the lifetime prevalence of all psychotic disorders including schizophrenia is over 3% in this nationally representative sample.

Schizophrenia symptoms begin to appear in the late teens with a peak between 20 and 30 years of age [27]. Schizophrenia at an earlier age is extremely rare, and has a prevalence of about 1 per 10,000 in children [28]. In a representative sample of schizophrenia patients from Germany, the mean age of onset for the earliest sign of a mental disorder, first psychotic symptom, and first hospitalisation is 25.4, 27.9, and 30.0 years of age respectively [29]. A number of factors appear to influence the mean age of onset, with pre-morbid functioning and gender among the most significant. Individuals with earlier, youth-onset schizophrenia have more severe cognitive deficits on executive function, IQ, and verbal memory while individuals with much later onset have a more specific and limited pattern of cognitive deficits [28]. The mean age of onset occurs three to five years earlier in men, and the age of onset distributions when stratified by gender also have visibly different distributions [27]: age-of-onset for men reaches a maximum at an earlier age, while a secondary peak is observed in females after the age of 40. Finally, schizophrenia is more commonly observed in men, with a male-to-female rate ratio of 1.4 (1.3 - 1.6, 95% CI) [5].

The Global Burden of Disease study use disability-adjusted life years (DALYs), defined as the sum of years of life lost (YLLs) and years lived with disability (YLDs), to measure disease and injury burden in the world [30]. Even though schizophrenia occurs less frequently ($< 1\%$) than other major causes of disability and mortality, such as cardiovascular diseases, cancers, and neurodegenerative diseases in developed nations and infectious disease in developing nations, it is ranked as the 43rd leading cause of disability-adjusted life years globally, and unlike many other conditions, affects both developing or developed countries to a very similar extent. Notably, from 1990 to 2010, schizophrenia's per-capita DALYs increased by 10.5% while the burden of disease in mental and behavioural disorders as a group increased by 5.9%, a trend that runs counter to the progress made in common infectious diseases (-59.9%), maternal disorders (-42.6%), cancer (-2.1%), and cardiovascular diseases (-5.7%). Globally, we see that profile of disease burden is shifting from infectious diseases affecting neonates and children to cancers, heart diseases, and mental illnesses like schizophrenia.

1.1.4 Environmental risk factors

Large-scale epidemiological studies have demonstrated that the a number of environmental exposures are strongly associated with schizophrenia, each with substantial effects (odds ratio [OR] > 2) on risk. First, childhood adversity and trauma, encompassing neglect, sexual, physical, and emotional abuse, are significantly linked with the risk of psychosis, with an overall odds ratio of 2.79 ($2.34 - 3.31$, 95% CI) [31–34]. Furthermore, individuals suffering from extreme stress in early life, such as growing up in a time of persistent and extreme famine, have increased rates of brain abnormalities and psychiatric disorders [35–38]. Adverse prenatal outcomes, such as obstetrical complications, low birth weight, and shortened gestation period, are also significant predictors of schizophrenia [39, 40]. In the pharmacological space, a number of studies have suggested that long-term cannabis use increases the risk of general psychotic disorders and schizophrenia. In a study of 45,570 Swedish conscripts, the odds ratio for schizophrenia in chronic, heavy users of cannabis was $\sim$2.1 when compared to individuals who did not use cannabis, and this result remained significant even after controlling for other psychiatric illnesses and social background [41]. Subsequent analyses in New Zealand, Germany, and U.K. replicated these results with very similar effects [42]. However, no study has definitively shown that cannabis use is causally linked to schizophrenia; it is also known that individuals with psychotic disorders are generally prone to higher rates of substance abuse, and cannabis use may be an outcome rather than a cause of schizophrenia. Another robust, though broadly defined, environment exposure for schizophrenia is urbanicity. People born or brought up in cities experience higher rates of psychosis [43], and this result remains significant even after controlling for socio-economic status and ethnic composition [44]. The association with urbanicity is independent of the metric by which urbanicity is defined (urban-rural [binary] or population density [quantitative]). However, the mechanism underlying this association remains unclear; it is possible that urbanicity is a proxy for more specific environmental exposures like substance use, social isolation, and pollution. Finally, migration and minority status has been linked in increased rates of schizophrenia. Two studies based in London and The Hague have identified a dose-response relationship between the proportion of non-white ethnic minorities in a neighbourhood and the incidence of schizophrenia, finding higher rates of schizophrenia in minority groups when they are a smaller proportion of the regional population [45, 46]. In summary, a number of environmental factors are robustly linked with schizophrenia. However, because of the high levels of correlation between these exposures (e.g. urbanicity, drug use, minority status) and the obvious fact that significant associations certainly do not imply causation, great care must be taken when extrapolating notions of causality from these results.

1.2 The genetic architecture of schizophrenia

1.2.1 Family studies find substantial genetic component to risk

Since the early days of psychiatry, it was believed that schizophrenia, and psychiatric traits in general, had a substantial genetic component. Family studies have consistently shown that relatives of schizophrenia patients are at greater risk than the general population, with the lifetime risk nearly ten-fold higher in siblings or offspring of individuals with schizophrenia [47]. Furthermore, a person's risk for schizophrenia increases with the number of affected family members, with the lifetime risk increasing to 16% when both a parent and a sibling are affected, and 46% in the offspring of two parents with schizophrenia [47]. However, familial clustering does not prove the existence of a genetic component as it can be confounded by shared environmental factors, which is why scientists turned to twin and adoption studies to estimate schizophrenia's true genetic component. Monozygotic (MZ) and dizygotic (DZ) twin pairs enable the estimation of the broad-sense and narrow-sense (additive) genetic heritability along with the variance explained by shared environmental influences. The twin study approach uses the following properties of MZ and DZ twins: that MZ twins share the entire additive genetic component, DZ twins share approximately half, and MZ and DZ twins have the same shared environmental component. These studies found a strikingly high concordance in monozygotic twins that was vastly greater than the concordance observed in dizygotic twins: with a DSM-III definition of schizophrenia, MZ twin pairs showed a concordance of 47.6% while DZ twin pairs showed a concordance of 9.5%, with an estimated heritability or h^2 of 0.85 [48]. One of the most cited estimates of the genetic component of schizophrenia comes from a meta-analysis of twelve twin studies, which refined the point estimate of the broad-sense heritability of schizophrenia to 81% ($73-90$%, 95% CI), with consistent evidence for a shared environmental contribution of 11% ($3-19$% CI) [49]. While substantial heterogeneity was observed among the point estimates from the twelve twin studies, together, these studies show consistent support for a large genetic component in the etiology of schizophrenia.

However, the twin study method has been criticised for a number of its core assumptions, including the possibility that monozygotic twins are more likely to share the same environmental exposures when compared to dizygotic twins. To address this, scientists turned to studies investigating the rates of psychiatric illness in children with biological parents who developed schizophrenia but who were given up for adoption. These adoption studies compared the incidence of schizophrenia in adopted children from parents with schizophrenia to the incidence in adopted children from non-schizophrenia parents. The first of such efforts in 1966 found that 10.6% of 47 adopted children with affected mothers

developed schizophrenia, and found none in 50 control children [50]. A number of subsequent adoption studies replicated and extended these results [51]. Some of these showed genetic liability for schizophrenia conferred risk for broader psychiatric phenotypes (e.g. schizoid personality disorders). Other studies followed adopted children of affected fathers, enabling them to exclude intra-uterine influences as a potential environmental confounder. Furthermore, nation-wide adoption studies demonstrated that the biological relatives of an adopted individual with schizophrenia have higher than expected incidences of schizophrenia, while the adopted family have incidences no different than baseline [51]. Together, the results from family studies spanning nearly eighty years and multiple designs conclusively demonstrate that a substantial genetic contribution exists in the etiology of schizophrenia.

1.2.2 Genome-wide association studies implicate common polygenic variation

Subsequent studies sought to clarify the nature of the genetic contributions to risk of schizophrenia with the ultimate goal of identifying the number, frequencies, and effect sizes of risk alleles in the human population. The existence of a monogenic architecture was immediately excluded due to the absence of clear segregation patterns in families. The search for individual variants of substantial effect continued in linkage and candidate gene studies, but these efforts were largely unsuccessful in identifying risk factors that explain schizophrenia's genetic liability. A more likely hypothesis describing a polygenic architecture akin to other complex traits was proposed by Gottesman and Shield as early as 1967 [52]. This model suggests that a very large number of loci of modest effect together contribute to the liability of developing schizophrenia. This hypothesis can explain the high concordance in twin studies, and the increased risk when more relatives of an individual are affected. It also provides an explanation for the surprisingly high incidence of schizophrenia in general population despite its negative prognosis, since selection cannot effectively eliminate so many common variants with such modest effects of fitness. Despite the polygenic model's plausibility, it was not until the arrival of array-based genotyping at a population scale that this theory is proven true.

The completion of the Human Genome Project and the HapMap Project helped create comprehensive catalogues of millions of common variants in the human population [53]. The maturation of DNA microarray technologies at around the same time enabled the multiplex genotyping of hundreds of thousands of single nucleotide polymorphisms in a single individual. Finally, statistical methods were developed to robustly test individual markers for association with different human traits, controlling for systematic biases from

sample ascertainment, genotyping error, and multiple testing. The convergence of these milestones paved the way for a new era of genetic mapping in human disease. Since the mid-2000s, genome-wide association studies (GWAS) have made significant progress in advancing our understanding of the genetic architecture of complex diseases, confirming that many human traits and disorders indeed have a polygenic component [54]. The early results for psychiatric disorders were less compelling; although a multi-stage GWAS for schizophrenia in 2008 identified a single SNP near *ZNF804A* [55], smaller studies investigating common variants in Crohn's disease and Type 1 Diabetes identified many more loci at genome-wide significance.

Instead of simply identifying individual loci using the GWAS approach, a landmark study combined the additive effects of nominally significant loci into quantitative scores, and computed and tested these scores for association to schizophrenia in an independent sample [56]. The scores generated on SNPs with $P < 0.5$ was highly correlated with schizophrenia risk ($P = 9 \times 10^{-19}$), and explained around 3% of the variance. This result was replicated in several other independent data sets and at varying P-value thresholds. Notably, the polygenic score for schizophrenia was specific to psychiatric disease, having no association with cardiovascular and autoimmune diseases. The limited success in identifying individual loci originated in part due to the breadth of schizophrenia's diagnostic criteria and differences in its genetic architecture when compared to other complex traits. Schizophrenia likely encompassed a number of disorders with different underlying etiologies. Therefore, individuals recruited into schizophrenia studies represented a highly heterogeneous clinical sample, which resulted in reduced statistical power when detecting variants which conferred risk for a subset of individuals. Second, schizophrenia appeared to have fewer loci of individually large effect when compared to other complex traits. For instance, autoimmune disorders had common risk variants with odds ratios of greater than 1.5 [54], which enabled robust associations with only a few hundred cases. While early association studies of schizophrenia did not have sufficient power to identify many individual risk loci, they strongly confirmed a polygenic component to schizophrenia involving common variants. Reassuringly, a subsequent genome-wide association study of 36,989 cases and 113,075 controls identified 128 independent common variant associations (minor allele frequency [MAF] > 2%) that remained significant after multiple testing correction (Figure 5.1) [57]. Combined, these 128 loci explained 3.4% of variation in the liability-threshold model. Variance components analysis on the same sample determined that more than 71% of all one Megabase (Mb) regions in the genome contained at least one common risk allele, and estimated the additive heritability from common variants to be 27.4% [58]. Therefore, the modest effects of common variants (median odds ratio [OR] = 1.08) are combined to produce

a polygenic contribution estimated to explain a notable fraction of the overall genetic liability in schizophrenia.

1.2.3 Recurrent copy number events confer substantial risk

Copy number variants (CNVs) are a type of structural variation that either deletes or duplicates a segment of DNA greater than 1 Kilobase in size. As a class of variation, CNVs account for the largest proportion of bases that vary between individuals [59], and confer risk for a number of Mendelian and complex diseases [60]. Early results from cytogenetic studies, such the identification of trisomy 21 as the cause of Down syndrome and a burden of chromosomal abnormalities in children with autism, suggested that structural variants may explain at least a portion of the genetic liability in brain disorders [61]. This hypothesis was validated when a copy number variant - the 22q11.2 deletion - was implicated as the first genetic risk locus for schizophrenia. The 22q11.2 deletion is highly recurrent, and has two common breakpoints resulting in a removal of 3 Mb or 1.5 Mb of sequence and a single copy loss in 30 to 40 genes. While this deletion causes a broader syndrome characterised by cognitive impairment and physical abnormalities, nearly 24% of carriers have psychiatric symptoms satisfying the full diagnostic criteria for schizophrenia [62].

With the arrival of array-based genotyping technologies, these early results were generalised when individuals with schizophrenia were shown to have a greater genome-wide burden of rare copy number variants compared to controls [63]. A genome-wide analysis comprising of 3,391 cases and 3,181 controls demonstrated that a 1.15-fold enrichment of rare and large CNVs (MAF $< 1\%$ and > 100kb) existed in individuals with schizophrenia. The enrichment was even greater at 1.32-fold for deletions with a length of least 500 Kilobases. Subsequent studies began to implicate individual loci at genome-wide significance, beginning with the 1q21.1, 15q11.2, and Neurexin 1 loci [64]. More recently, follow-up of putative risk loci in tens of thousands of individuals identified 11 rare CNVs that individually conferred substantial risk for schizophrenia (ORs $2 - 60$, Figure 5.1) [63, 65–67]. These risk CNVs appeared to be highly recurrent and shared nearly the same breakpoints within each locus. This was due to the mechanism by which these CNVs were formed: they are flanked by segmental duplications, which enable higher rates of non-allelic homologous recombination and increased mutability at these regions. Because they are subject to strong negative selection, the 11 risk CNVs remain very rare events in the general population and explain only a small fraction of the genetic liability for schizophrenia [68]. Together, these findings established that both common variants and rare structural variation contribute to the complex genetic architecture of schizophrenia.

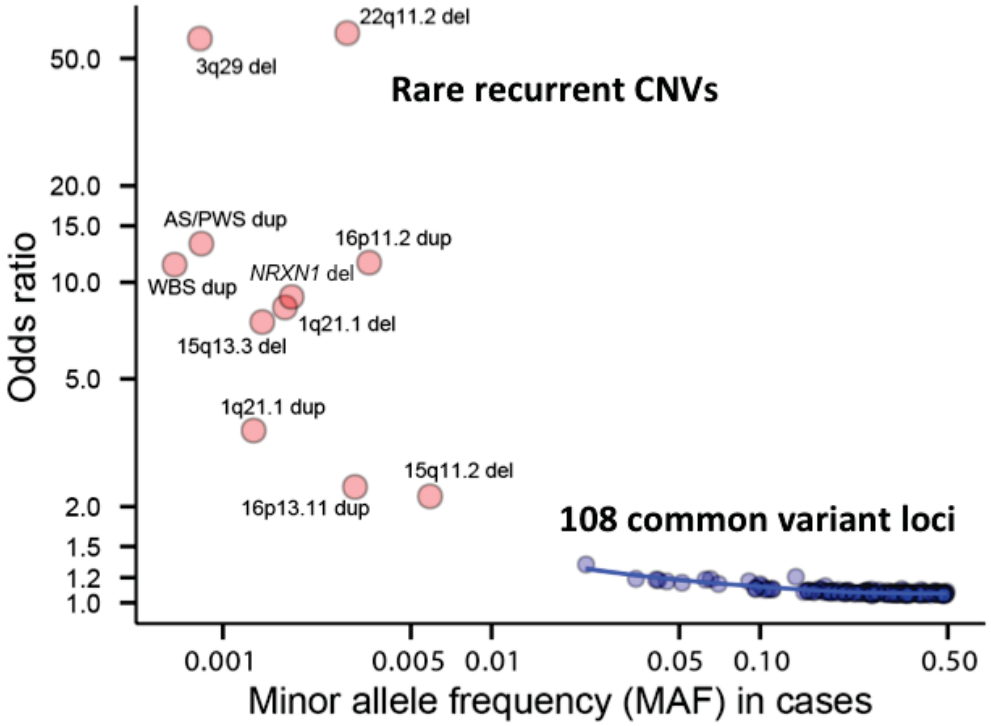

Fig. 1.1 **Risk variants for schizophrenia.** The effect size of each genome-wide significant risk variant for schizophrenia, as described in Ripke *et al.* and Rees *et al.*, were plotted against its allele frequency in cases [57, 67].

1.2.4 Shared common risk variants across psychiatric disorders

The categorical symptoms used to define psychiatric disorders are generally not exclusive to a single diagnosis. For instance, the Schneiderian First Rank symptoms for schizophrenia are observed in individuals with bipolar disorders, and schizophrenia patients often have symptoms characteristic of depression, anxiety, and obsessive-compulsive disorder [16, 20]. Furthermore, relatives of patients with bipolar disorders are 4.9-fold more likely to have schizophrenia than relatives of control individuals [69]. These observations provide early evidence that genetic risk may be shared between psychiatric disorders.

Polygenic risk scores generated from schizophrenia case-control data were significantly associated with bipolar disorder ($P < 7 \times 10^{-9}$) and explained up to 1.9% of the variance, demonstrating that schizophrenia and bipolar disorder shared at least some common risk variants [56]. To further explore the shared genetic etiology between psychiatric disorders, array-based genotype data from case-control studies of schizophrenia, bipolar disorder, major depressive disorder, attention-deficit hyperactivity disorders, and autism were used to estimate the narrow-sense heritability of each disorder and the genetic correlation between each pair of disorders [70, 71]. Remarkably, common risk variants were significantly shared across all these conditions. The strongest correlation was observed between schizophrenia and bipolar disorder (0.68; $0.62 - 0.72$, 95% CI), an expected result when considering clinical and epidemiological evidence for overlapping symptoms. The weakest correlation was observed between schizophrenia and autism (0.16; $0.1 - 0.22$, 95% CI), which was also expected since autism is believed to have more of a neurodevelopmental etiology. Reassuringly, no correlation was observed with Crohn's disease, demonstrating that the sharing of common variants was specific to psychiatric disorders. The significant genetic correlation between psychiatric disorders is likely driven by a number of pleiotrophic risk alleles tagged by common variants, and suggests that some number of biological mechanisms of disease may too be shared between brain disorders previously thought to be largely distinct.

1.3 Whole-exome sequencing as a means of studying rare variants

By accumulating sufficiently large sample sizes for GWAS needed for discovery and replication, large consortia have been particularly successful in identifying common variants of small effects even in highly heterogeneous traits [72]. The many thousands of common genetic variants associated with increased risk in complex diseases have opened up unprecedented opportunities for the elucidation of disease pathways, mechanisms, and genetic architecture.

Despite the success of the GWAS approach, the biological mechanisms of the vast majority of common risk loci remain largely unknown, with many interspersed in intergenic regions or in linkage with multiple variants in close proximity to a number of genes [73]. The nature of linkage disequilibrium (LD) in association studies has made it difficult to pinpoint the precise functional variant, gene, and pathway implicated, and while functional annotation and network connectivity can help prioritize genes, the causal variant may not have been typed at all, and large LD blocks may mask multiple independent causal variants that account for additional genetic risk. Furthermore, the risk factors discovered by GWAS are generally low-effect common variants that explain only a fraction of genetic heritability. For instance, the genome-wide significant common risk loci for schizophrenia identified from 36,989 cases explain only 3.4% of variation in the liability-threshold model [57]. Even if all common risk alleles were identified, they only explain around 27.4% of the broad-sense heritability previously estimated to be around 81% [49, 58]. The genetic architecture of schizophrenia is far from being fully ascertained, with the number of loci, effect sizes, frequencies, and interactions yet to be determined, and ultimately, existing genotyping arrays only assay a subset of all variants that may confer disease risk.

It is almost certain that genetic variation other than common SNPs are associated with complex disease risk. Already, rare structural variation has been demonstrated to play a non-trivial role in the manifestation of a range of psychiatric disorders including autism, Alzheimer's disease, and schizophrenia [74]. Rare CNVs contribute to increased risk for schizophrenia and bipolar disorder, with at least 11 large CNVs conferring substantial risk for severe psychiatric outcomes [67]. Furthermore, genome-wide association studies are unable to investigate SNPs and indels that are rare in the population (MAF $< 1\%$) or unique to a single individual. Because negative selection acts most strongly on variants with large fitness coefficients, the variants that confer the most risk for disease necessarily reside in the lower end of allele frequency spectrum. As technologies mature, approaches that characterize this rarer subset of risk-conferring variation are rapidly scaled up to complement existing genotyping efforts, in hopes of completing the genetic picture on complex disorders.

While *de novo* assembly of the first human genome required the capillary sequencing of long reads that took nearly a decade to complete, next-generation whole-genome sequencing (WGS) instead generates and aligns short reads ($<$ 100bp) from a single individual to the human reference genome to build a variation map [75]. If the genome is sequenced at reasonable coverage ($30-60\times$), we can identify nearly all common and rare single nucleotide variants and the vast majority of large structural variants with reasonable accuracy. This technology is sufficient in generating the high-resolution datasets required to fine-map existing risk loci, uncover population-specific variants, and identify ultra-rare exonic

variants with clear functional consequences. However, whole-genome sequencing remains prohibitively expensive for sequencing many thousands of individuals at the coverage required to accurately call variants. Data production, processing, and storage of high-coverage whole-genome data remain costly and time-consuming: uncompressed reads of a single genome at standard 30× coverage is approximately 250 Gigabytes in size, with the compressed BAM file reaching nearly 300 Gb in size [76].

As a cost-effective alternative, whole exome sequencing (WES) selectively sequences only coding regions of a genome at very high-coverage using a target enrichment strategy [77]. The target region is usually between 35 and 65 Megabases, representing at most 2% of the human genome. Not only are sequencing costs much lower, but production, storage, and analyses of exome sequences are not as computationally intensive. Furthermore, coding variants are much easier to functionally interpret, classify, and annotate than those in non-coding regions, which prove valuable in downstream analyses [78].

1.3.1 Common study designs for sequencing studies

Parent-proband trio studies investigating the role of *de novo* mutations

Already, whole-exome sequencing has been successful in identifying causal variants for Mendelian traits. The technology is particularly effective in resolving severe disorders where cases are rare and sporadic and the causal variant is likely *de novo* in origin. Every individual has an average of 74 germline *de novo* mutations, of which one resides in the protein-coding region [79, 80]. These *de novo* events are systematically identified by comparing the exomes of the biological parents and the proband and looking for variants that violate principles of Mendelian inheritance. *De novo* mutations are more enriched for alleles conferring substantial risk for disease compared to inherited rare and common variants because they have not undergone post-zygotic negative selection. Compounded by the absolute rarity of these events, *de novo* mutations with highly damaging functional consequences (e.g. putative loss-of-function) have a high prior of being pathogenic for disease when compared to inherited variation. Since observing multiple damaging *de novo* events in a single protein-coding gene is extremely unlikely, sequencing a small number of cases and identifying genes with multiple *de novo* hits is often sufficient in the discovery of the causal variant in sporadic Mendelian disorders. Because of the relative straightforwardness of this analysis, whole-exome sequencing has been extremely successful in discovering genes underlying unsolved monogenic disorders. The first wave of whole-exome sequencing analyses identified the causal genes for Miller syndrome (*DHODH*), Kabuki syndrome (*KMT2D*), and Bohring-Opitz syndrome (*ASXL1*) [81–83] by sequencing fewer than 15 affected individuals. These

early results motivated the formation of clinical cohorts composed of many thousands of individual with sporadic and severe disorders likely of monogenic etiology to enable large-scale gene discovery.

In addition to diagnosing severe monogenic disorders, whole-exome sequencing has revealed a major role of rare variation in psychiatric and neurodevelopmental disorders, implicating individual genes, gene sets, and biological processes. Even in highly heterogeneous and complex human disorders, the rarity of damaging *de novo* events makes it possible to observe statistically significant recurrence of mutations in individual genes with smaller sample sizes than would be required in a case-control design. Two early whole-exome sequencing study identified *de novo* mutations in 151 patients with intellectual disability to better understand its genetic etiology [84, 85]. One study found diagnostic variants in 16% of patients [84], while the other found a 3.9-fold excess of *de novo* LoF mutations in cases compared to controls [85]. These results confirmed that *de novo* mutations are an important cause of intellectual disability, and that whole-exome sequencing is a highly useful diagnostic tool despite the disorder's substantial clinical and genetic heterogeneity. Trio studies investigating autism spectrum disorders similarly found that damaging *de novo* mutations are elevated in simplex cases compared to controls [80, 86, 87]. However, the rate of *de novo* events in individuals with autism was less than the rate observed in intellectual disability, suggesting that *de novo* mutations play an important but more limited role in the genetic architecture of autism. However, sufficient numbers of *de novo* mutations in $\sim$600 probands were observed in the same genes to implicate novel autism risk genes, including *CHD8* and *KATNAL2*.

Rare variant association analyses using case-control data sets

Whole-exome sequencing has been less successful in identifying genomic regions in which the burden of rare variants differ between cases and controls. While the methodology behind common variant association testing is now well established, rare variants cannot be individually tested due to their absolute rarity in the human population, and must be aggregated into sets in order to be analysed [88]. Purifying selection strongly reduces the allele frequencies of highly damaging variants, and thus, variants with the strongest effects are likely to be much rarer in the population. Because of this, neutral variants vastly outnumber damaging rare alleles, and increase the baseline level of noise in collapsing tests [88]. Rare variant analyses enrich for risk alleles by aggregating only variants below a particular allele frequency threshold (i.e. $< 0.1\%$) and with a likely damaging coding consequence (e.g. missense or loss-of-function). In order to reduce costs, the first analyses attempting to identify rare risk variants for complex diseases used targeted sequencing in a small number

of genes. This approach achieved mixed success: the sequencing of 25 candidate genes in 24,892 cases with autoimmune diseases and 41,911 matched controls demonstrated a limited role of rare coding variants [89] while the targeted sequencing of 1,326 genes in 9,946 psoriasis cases and matched controls did not identify any gene with a burden of rare variants [90]. On the other hand, targeted sequencing in 63 known prostate cancer risk regions in 9,237 individuals did not identify novel genes, but found that rare SNPs explained a notable fraction of prostate cancer risk [91], and another study sequencing four candidate genes in 438 cases and 327 controls identified a burden of rare variants for hypertriglyceridemia [92]. Only a small number of rare variant association studies identified individual risk genes, and this required the sequencing of tens of thousands of individuals. For example, the targeted sequencing of 56 genes in 28,207 individuals with inflammatory bowel disease and 17,575 healthy controls identified rare risk variants in *NOD2*, *IL18RAP*, *CUL2*, *C1orf106*, *PTPN22* and *MUC19*, and protective variants in *IL23R* and *CARD9* [93]. These results suggest that extremely large samples would be needed to identify rare variants with only a moderate effect on risk.

Several large-scale efforts have tried to expand targeted sequencing approaches to test the entire exome. For instance, the NHLBI exome-sequencing project attempted to identify rare risk variants for cardiometabolic traits and cardiovascular disease using the exomes of 6,500 individuals [78]. The analysis of these exomes did not identify any novel genes for any of these traits. These data were then combined with imputed genotypes of 64,132 individuals, array-based genotyping of rare variants (Exomechip) in 15,936 individuals, targeted sequences in 6,721 cases and 6,711 controls, and exome sequences in 9,793 individuals. Only then did the study identify rare alleles in *LDLR* and *APOA5* as conferring risk for myocardial infarction [94]. It is clear that very few case-control whole-exome sequencing studies at this time have sufficient power to identify risk genes. Thus, rare variant association testing likely will require much larger sample sizes in the tens of thousands in order to successfully identify risk genes at exome-wide significance.

1.4 Early results from sequencing in schizophrenia

Whole-exome sequencing studies investigating *de novo* mutations and case-control burden have demonstrated that rare variation plays an important role in the genetic architecture of schizophrenia. A number of early studies found that *de novo* missense and loss-of-function mutations were elevated in cases compared to controls [95–97], and proposed a number of possible candidate genes based on one or two *de novo* events. The largest study of schizophrenia *de novo* mutations so far whole-exome sequenced 617 parent-proband trios,

and found intriguing patterns in groups of synaptic proteins and gene targets of *FMRP* [98]. The study further found that individuals with school grades below the median had a higher enrichment of *de novo* mutations, suggesting that there was a link between these more damaging variants and cognition. While it nominated *TAF13* as a possible candidate gene, the study did not have sufficient power to identify a single risk gene despite having similar sample sizes as early autism and intellectual disability trio data sets. In total, seven studies have studied *de novo* mutations in 1,077 schizophrenia probands, and identified thirty-eight genes with two or more *de novo* nonsynonymous mutations [66, 98, 99, 95, 97, 100–102]. These studies have found suggestive evidence for candidate genes, including *EHMT1*, *DLG2*, *TAF13* and *SETD1A* [66, 98, 99], but much larger data sets are required to robustly demonstrate these are true schizophrenia genes achieving genome-wide significance.

A recent case-control exome sequencing study with 2,543 schizophrenia cases and 2,543 matched controls compared the rate of rare variants in individual genes between cases and controls using a one-sided burden test and the SNP-set (sequence) kernel association test (SKAT) [103, 104]. To enrich for risk variants, the authors stratified their analyses by allele frequency and functional class (missense or missense and loss-of-function). Unfortunately, they did not identify any individual gene at a Bonferroni *P*-value of 1.25×10^{-6}. Instead, the study tested for a rare variant signal in biologically meaningful gene sets, and found a significant burden of rare disruptive variants across a set of 2,546 genes selected on the basis of a variety of biological hypotheses about schizophrenia risk and previous genome-wide screens, including GWAS, copy number variation (CNV) and *de novo* mutation studies [103]. Furthermore, an enrichment in the targets of *FMRP* and synaptic density proteins was also observed, similar to observations in the analysis of *de novo* mutations. Despite not having sufficient power to identify individual genes, these analyses demonstrate that rare variants contribute to the genetic architecture of schizophrenia, and risk genes will eventually be identified with sufficiently large data sets.

1.5 Biological insights from genetic studies of schizophrenia

A number of biological insights have emerged from these early genetic results in schizophrenia. First, gene set enrichment analyses of *de novo* CNVs from 662 trios provided evidence that these events disproportionately disrupted genes that were components of the post-synaptic density proteome [66]. This observation was partially explained by a strong enrichment in genes of the *N*-methyl-D-aspartate receptor (NMDAR) and neuron activity-regulated

cytoskeleton-associated (ARC) protein postsynaptic density signalling complexes, and further supported the hypothesis that synaptic processes were dysregulated in schizophrenia. Enrichment analyses of *de novo* single nucleotide polymorphisms in the same trios replicated these results, and found that large effect SNVs and indels also clustered in genes in the NMDAR and ARC complex [98]. Furthermore, schizophrenia *de novo* mutations were enriched in voltage-gated calcium channels and transcriptional targets of the Fragile X mental retardation protein (*FMRP*), a result also observed in recent analyses of rare variants in autism [105]. This study also found a nominal overlap with *de novo* LoF variants from probands with intellectual disability ($P = 0.019$, uncorrected), but this result was based on the observation of a single *de novo* event in the schizophrenia probands. A large case-control analysis of whole-exome sequencing data further strengthened these observations by demonstrating a burden of damaging variants in genes in the NMDAR and ARC components of the post-synaptic density, calcium signaling genes, and translational targets of *FMRP* [103], and similarly, a case-control study of copy number variants in 4,719 schizophrenia cases and 5,917 controls also implicated components of the post-synaptic density, calcium channel genes and targets of *FMRP* [106]. Together, analyses from multiple study designs analysing different forms of rare variation suggest an overlapping set of biological processes, such as transmission at glutamatergic synapses, are perturbed in schizophrenia.

Genetic risk loci identified in genome-wide association studies provide additional insights into the pathogenesis of schizophrenia. A number of intriguing genome-wide hits have been identified in the largest GWAS to date, one of which is a common variant near the dopamine receptor D2 gene [57]. First- and second-generation antipsychotic drugs work by inhibiting D2R activity, and furthermore, abnormal pre-synaptic dopaminergic activity is a major hypothesis of schizophrenia pathogenesis [107]. The discovery of this single genetic signal suggests that other common variant loci may highlight novel biological processes and valuable therapeutic targets warranting functional follow-up. Gene set analyses of common risk variants found enrichment for brain and immune enhancers, but no specific pathways appeared significant [57]. A study investigating biological pathways using common variant data from individuals with schizophrenia, major depression, and bipolar disorder found evidence that risk variants aggregate in a number of core biological processes, including histone methylation, neuronal signalling pathways, and components of the post-synaptic density [108]. Therefore, overlapping results from common and rare variant are reaffirming previous hypotheses of disease pathogenesis and identifying novel and specific mechanisms in the etiology of schizophrenia.

1.6 Goals of this Thesis

Recent studies have demonstrated that the genetic architecture of common disorders are highly polygenic and involves a combination of common, rare, and *de novo* risk variants distributed across the genome. Furthermore, analyses of rare variation support a complex and heterogeneous architecture involving many thousands of risk alleles and hundreds of genes, suggesting that very large sample sizes will be required to convincingly identify individual risk genes using only rare coding alleles. This polygenicity is best exemplified in studies of neurodevelopmental disorders, such as autism spectrum disorder and intellectual disability, which required many thousands of exome sequences to identify genes at genome-wide significance [105, 109]. Despite several whole-exome sequencing studies investigating rare variants in schizophrenia, no individual gene had been significantly implicated using rare coding SNVs. Because of these promising results, multiple large consortia have been established to generate large sequencing data sets that will enable researchers to understand the link between rare variants and human traits and disorders. Three such efforts include the UK10K study, the Deciphering Developmental Disorder (DDD) study, and the Autism Sequencing Consortium (ASC). Initiated in 2010, the UK10K project has sequenced the whole-exomes of 5,296 individuals, including those diagnosed with autism, schizophrenia, obesity, and a number of rare diseases suspected to have a monogenic etiology. The goal of the project is to characterize rare variants in the UK population, and determine the contribution of these variants to a broad spectrum of traits and disorders with very different genetic architectures. On the other hand, the Deciphering Development Disorders study aims to use exome sequencing to help identify potential genetic causes of severe, undiagnosed developmental disorders. Over 4,000 trios have been sequenced thus far, and over the next few years, the project aims to sequence a total of 12,000 trios. Finally, the Autism Sequencing Consortium sought to generate large whole-genome and whole-exome sequencing data sets to identify loci associated with increased risk of autism across the allele frequency spectrum through a combination of *de novo*, dominant, and recessive analyses of rare variants. Hopefully, after the completion of these large projects, much more will be understood about the role that rare variants have in the genetic architecture of complex disorders.

In my dissertation, I processed and analysed high-coverage sequence data sets from the UK10K project and the DDD study in an attempt to identify risk genes containing rare variants with large effects that contribute to increased risk in psychiatric disorders, with a primary focus on schizophrenia. I first conducted rare variation association analyses using data generated in the UK10K study, which included 1,488 UK schizophrenia and 399 Finnish cases. We then aggregated *de novo* and case-control data from other published schizophrenia data sets in order to increase power for gene discovery. In the process, I improved the

procedures used in generating high-quality sequencing data (variant calling, filtering, and annotation), and refined and applied existing statistical procedures used in common and rare variant association testing. To increase statistical power, I combined signal from multiple whole-exome data sets, applied various filters on allele frequency, variant annotation, and predicted functional impact, and meta-analysed rare variants across family and case-control designs. Furthermore, rare variants identified in individuals with schizophrenia were compared and contrasted with those from probands in the DDD and ASC studies in order to understand the genetic connections between psychiatric and neurodevelopmental disorders. This included performing quality control and analysing a large independent copy number variant data set to increase power. After combining rare SNVs and CNVs data sets, I tested a number of biological hypotheses related to schizophrenia risk, and showed that the rare variants supported a neurodevelopmental etiology to schizophrenia. In summary, I sought to contribute to the understanding of the genetic architecture of complex psychiatric disorders through a comprehensive analysis of available high-coverage sequencing data.