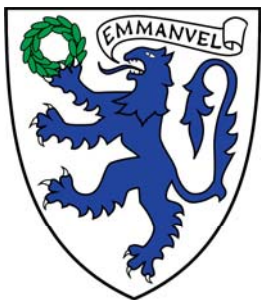
The contribution of rare variants to risk of schizophrenia and neurodevelopmental disorders



Tarjinder Singh

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

January 2017

Declaration

I hereby declare that I carried out the work described in this Thesis between September 2012 and August 2016 under the supervision of Dr. Jeffrey C. Barrett at the Wellcome Trust Sanger Institute. The contents of this Thesis has not been submitted in whole or in part for any other degree or qualification at the University of Cambridge, or any other University. This Thesis does not exceed the specified length limit, and is formatted according to the requirements set by the Biology Degree Committee and the Board of Graduate Studies.

Tarjinder Singh
January 2017

Acknowledgements

I would first like to thank my supervisor Jeff for giving me the opportunity to write my thesis in his group and for his support and encouragement in the past four years. I arrived in Cambridge in 2012 with plans to complete a one-year MPhil degree in the genetics of blood traits before attending medical school the following year. Little did I know that I would be fortunate enough to stay at the Sanger and explore an area of genetics and medicine that I had never previously encountered. Now, four years later, I am more motivated than ever to delve deeper into the world of statistical genetics and use its advances to understand the fundamental causes of mental illnesses. This formative experience would not be possible if not for the patience, mentorship, and guidance that Jeff has shown me. I sincerely hope we keep in touch, and perhaps find an opportunity to work together again in the future.

The research in this Thesis requires the coordinated efforts of numerous collaborators in the UK and around the world. I thank the many clinicians and scientists in the UK10K consortium who designed the initial study that laid the foundation for this Thesis. In particular, I would like to thank Mike Owen, Mick O'Donovan, Dave Curtis, and Matthew Hurles for productive discussions, advice, and contributions to this work. I want to express my gratitude to the Wellcome Trust and Williams College for their generous financial support. Most importantly, I want to thank the tens of thousands of patients and participants who enrolled in the studies described in this Thesis, without whom none of this work is possible and for which I am indescribably grateful.

On a more personal note, I want to extend my gratitude to the members of the Barrett team, past and present, for engaging team meetings, memorable retreats, and entertaining lunch discussions. I also want to thank my friends at Emmanuel College for all the enjoyable Formal dinners, pub crawls, and European travels that have made life in Cambridge so entertaining. In particular, I want to thank Albert and Uttara for their unwavering friendship, support, and honesty. I am grateful to the PhD students at the Sanger Institute, especially the Class of 2012, for all the enjoyable times we've had, from the scenic drives through the Peak District to the adventures in Warsaw. Lastly, I want to thank my parents and sister for their patience, encouragement, love and support throughout the years.

Abstract

In recent years, whole-exome sequencing has successfully identified genes in which rare variants confer substantial risk for neurodevelopmental disorders, such as autism spectrum disorders and intellectual disability. In many of these studies, the same gene is implicated in a wide variety of diagnoses and presentations. Despite a number of rare variant studies in schizophrenia, no gene has been significantly implicated using rare coding variants. In this Thesis, I compiled the largest rare variant data set in schizophrenia to date, and meta-analysed the whole-exome sequences of 1,077 trios, 4,268 cases, and 9,343 matched controls. With these data, I identified a genome-wide significant association between rare loss-of-function (LoF) variants in *SETD1A* and risk for schizophrenia. I additionally found that *SETD1A* is substantially depleted of LoF variants in the general population, and that LoF variants in this gene increased risk for a range of neurodevelopmental disorders. Combined, our results implicate epigenetic regulation, specifically histone modification, as a mechanism in the pathogenesis of schizophrenia, and suggest that rare risk alleles may potentially be shared between schizophrenia and other neurodevelopmental disorders.

To better understand if *SETD1A* finding can be generalized to a larger number of rare schizophrenia risk variants, I jointly analysed the trio and case-control exome data with array-based copy number variant calls from 6,882 cases and 11,255 controls. I found that individuals with schizophrenia carried a significantly higher burden of rare damaging variants in 3,488 "highly constrained" genes with a near-complete depletion of truncating variants. Rare variant enrichment analyses demonstrated that the rare schizophrenia risk variants were most strongly enriched in autism risk genes, and genes diagnostic of severe developmental disorders. I further showed that schizophrenia patients with intellectual disability had a greater enrichment of rare damaging variants in highly constrained genes, but that a weaker but significant enrichment existed throughout the larger schizophrenia population. Combined, these results demonstrate that schizophrenia risk loci of large effect across a range of variant types implicate a common set of genes shared with broader neurodevelopmental disorders, suggesting a path forward in identifying additional risk genes in psychiatric disorders and further supporting a neurodevelopmental etiology to the pathogenesis of schizophrenia.

Table of contents

List of figures

List of tables

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.

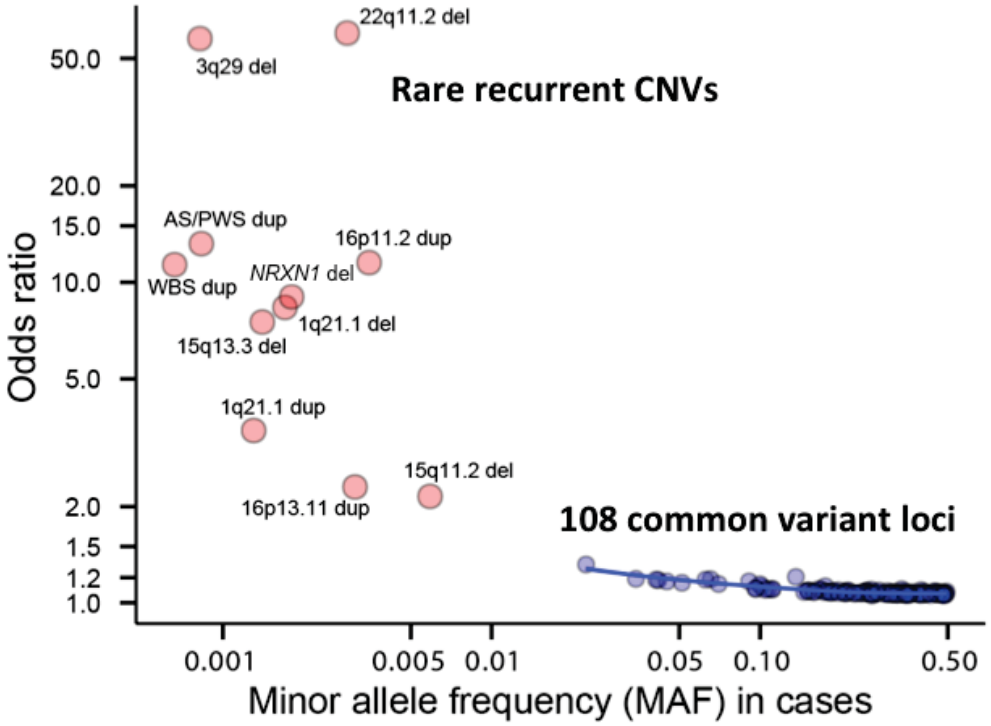


Fig. 1.1 **Risk variants for schizophrenia.** The effect size of each genome-wide significant risk variant for schizophrenia, as described in Ripke *et al.* and Rees *et al.*, were plotted against its allele frequency in cases [57, 67].

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.

Chapter 2

A protocol for the quality control of whole-exome sequencing data sets

2.1 Challenges behind the production and analysis of sequencing data

Whole-exome sequencing has emerged as the technology of choice in investigating the contribution of rare variation in the genetic basis of complex disorders. It has been most successful in identifying genes underlying rare Mendelian disorders, in which only a small number of samples are needed to reveal causal variants of large effect [110, 111]. Early results from complex diseases have demonstrated that a genome-wide burden of disruptive variants exists in cases compared to controls. However, the identification of individual risk genes remains elusive because a large number of genes appear to underlie many complex traits and our ability to differentiate pathogenic variants from neutral polymorphisms remains limited [78, 103, 88]. Much larger sample sizes, possibly in the tens of thousands, are required to identify sufficient numbers of rare variants to implicate individual risk genes [88, 105]. While studies have individually analysed a small number of exomes, in aggregate tens of thousands of whole-exome sequences have been generated to date [112]. Meta-analyses leveraging published data sets are beginning to have sufficient power for gene discovery.

Standardized protocols currently exist for performing variant discovery on whole-exome sequence data [113–117]. Raw reads in a FASTQ files are first mapped to a genome reference, duplicated reads are marked in the resulting BAM file to reduce amplification bias, and base quality scores are empirically adjusted for systematic errors. A variant caller such as Samtools mpileup or GATK HaplotypeCaller identifies sites at which a potential variant exists relative to the reference, and calculates the probabilities of each possible genotype

at that site [113, 114]. For very large data sets, samples are called individually and merged before variant calling occurs in aggregate. This enables the incorporation of variant-level information across samples when determining the appropriate genotype. Subsequently, a variant classifier, such as GATK Variant Quality Score Recalibration (VQSR), filters out mapping and sequencing artefacts. The remaining variants are annotated with predicted biological consequences and analysed. These best practices were successfully applied in Mendelian disorder studies and parent-proband trio studies analysing *de novo* mutations [110, 118, 98].

As we begin to jointly analyse thousands of samples aggregated from published studies, additional complexities in the preparation and production of whole-exome sequencing data begin to emerge. First, sequencing technologies have a higher genotyping error rate than array-based calls, and unlike common variant association studies, genotype refinement using a reference panel is unlikely to improve the quality of variant calls at the lowest end of the allele frequency spectrum [116]. To partially address this, each sample is sequenced to sufficiently high depth to ensure reasonable coverage ($40\times$ or greater) over the entire exome [116]. However, the enrichment of coding sequences using DNA hybridization inherently leads to uneven coverage: certain regions are captured to much greater affinity due to sequence context (high GC content), while other baits fail when overlapping polymorphisms modify its annealment affinity. Baits targeting low complexity regions capture reads from other repetitive sequences, leading to a even greater disparity of coverage across the exome. These limitations are further exacerbated by the substantial batch effects that appear from combining data from different exome sequencing studies. Depending on study design, researchers sequence samples to different mean coverage, which result in higher quality calls in some samples over others. Furthermore, a number of commercial captures are available for target enrichment, and each have systematic biases in its regional coverage of the exome. Finally, sequencing centres have different protocols for sample preparation, sequencing, and data production that are subject to change as technology progresses, all of which introduces additional variability between groups of samples. To aggregate and meta-analyse published sequencing data sets, we must first address these sources of systematic bias which often confound the results of rare variant association tests.

In this Chapter, I first describe the whole-exome sequencing data generated in the UK10K project, the Deciphering Developmental Disorders (DDD) study, INTERVAL study, Swedish Schizophrenia study, and the Sequencing Initiative Suomi (SiSU) project, all of which are analysed in Chapters 3 and 4. I then highlight the steps taken to prepare these data for analysis, and detail best practices to harmonize sequence production, variant calling, and variant- and sample-level QC across many thousands of whole-exome sequences. Useful

metrics for comparing variant quality between data sets are shown and discussed. Using diagnostic *de novo* mutations from the DDD study, I determine which *in silico* annotation tool best differentiated pathogenic from benign variants, which I then use to classify missense variants in subsequent analyses. Finally, I describe published whole-exome sequencing data sets of schizophrenia parent-proband trios, and extend a method of modelling the recurrence of *de novo* mutations for gene discovery.

2.1.1 Publication note and contributions

The results described in this chapter was peer-reviewed and published earlier this year [119]. I briefly summarise the various contributions to this project. The neuro group within the UK10K study recruited and whole-exome sequenced schizophrenia cases. This initiative was led by Aarno Palotie, Michael J. Owen, Jeffrey C. Barrett, and Daniel Geschwind. The sequencing team at the Wellcome Trust Sanger Institute performed exome capture, sequencing, and alignment for the UK10K and INTERVAL studies. I received the raw VCF for the Finnish case-control data set from Mitja I. Kurki and Aarno Palotie. I performed all subsequent production, and QC steps for these data under the supervision of Jeffrey C. Barrett. Unless explicitly stated, the parts of the peer-reviewed publication reproduced in this Chapter are my original work.

2.2 Materials and methods

2.2.1 Sample collections

Individuals clinically diagnosed with schizophrenia were recruited and exome sequenced as part of eight neurodevelopmental collections (Aberdeen, Collier, Edinburgh, Gurling, Muir, UK-SCZ, Finnish-SCZ, and Kuusamo) in the UK10K sequencing project. Matched population controls were selected from non-psychiatric arms of the UK10K project, healthy blood donors from the INTERVAL project, and five Finnish population studies (ENGAGE, Familial dyslipidemia, FINRISK, Health 2000, and METSIM). Additional details on the UK10K dataset are described in Table 2.1 and 2.2, and the sequence data have been deposited into the European Genome-phenome Archive (EGA) under study accession EGAO00000000079. The Swedish schizophrenia case-control study had been described in an earlier publication [103], and I acquired processed VCFs for this data set via dbGaP authorized access (Accession: phs000473.v1.p1). A total of 2,536 schizophrenia cases and 2,543 controls were available for analysis. The DDD study was designed to further our understanding of broader developmental disorders while advancing clinical genetics practice in the UK. 4,281 children

Collection	Sample size	Population	Description
ABERDEEN	391	UK	Schizophrenia cases with cognitive measurements recruited in Aberdeen, Scotland.
COLLIER	172	UK	Subjects recruited from three different studies: the Genetics and Psychosis (GAP) study (early-onset schizophrenia), the Maudsley twin series, and the Maudsley family study (families with a history of schizophrenia or bipolar disorder).
EDINBURGH	234	UK	Subjects recruited from psychiatric facilities in Scotland with IQ > 70. 138 are familial cases, and 100 have deep neuroimaging information.
GURLING	45	UK	Subjects from multiply affected families all of which are unilineal for transmission of schizophrenia.
MUIR	103	UK	Subjects with autism, schizophrenia or some sort of psychoses with diagnoses of mental retardation. Only individuals with schizophrenia were included in our analysis.
UKSCZ	542	UK	UK and Irish subjects selected for a positive family history of schizophrenia (collected as sib-pairs or from multiplex kindreds), or are systematically recruited from South Wales and have undergone detailed cognitive testing.
KUUSAMO	120	Finland	Subjects from the Finnish Kuusamo internal isolate where there is a three-fold lifetime risk for schizophrenia.
Finnish SCZ	281	Finland	Subjects from a population cohort recruited from national registers and have two affected siblings.

Table 2.1 Description of samples collections included as cases in the UK10K schizophrenia analysis. 1,353 cases remained after sample quality control.

with diverse, severe undiagnosed developmental disorders and their parents were exome sequenced to identify novel risk genes carrying variants of large effect. Patient recruitment, sample collection, sequencing production, and initial analysis of the dataset were described in detail in a previous publication [118]. The sequence data had been deposited into the EGA under study accession EGAS00001000775.

The SiSU project is an international collaboration generating whole genome and whole-exome sequence data from Finnish samples, and consists of a number of prospective and case-control cohorts, including the ENGAGE, FINRISK, Health 2000, and METSIM studies (http://www.sisuproject.fi/content/cohorts). The Northern Finnish 1966 Birth Cohort (NFBC) is a geographically based representative birth cohort including 96% (N = 12,068) of all live births in the two most northern provinces of Finland in 1966. The Northern Finnish Intellectual Disability Cohort (NFID) is an ongoing sample collection of individuals who have been diagnosed with ICD-10 diagnosis of intellectual disability or specific developmental disorder of speech and language of unknown etiology (ICD-10 codes: F70-F79 and F80-F89). The current sample includes 324 patients and their first-degree family members ($N = 631$, 92 full trios) with GWAS and WES data available. Combined, 5,720 Finnish exomes from the SiSU project were available for analysis.

Collection	Sample size	Population	Description
UK10K Obesity TwinsUK	67	UK	Consists of individuals from the TwinsUK study with a BMI > 40.
UK10K Obesity Generation Scotland	422	UK	Subjects belong to a family-based genetic study from across Scotland, and consists of individuals with extreme obesity. Only unrelated individuals are included.
UK10K Rare Severe Insulin Resistance	119	UK	Trios in which the probands have been diagnosed with severe insulin resistance. Only unaffected parents are included as controls.
UK10K Rare Neuromuscular	116	UK	Trios in which the probands have congenital muscle dystrophies or myopathies, neurogenic conditions, mitochondrial disorders, or periodic paralysis. Only unaffected parents are included as controls.
UK10K Rare Thyroid	123	UK	Trios in which the probands have congenital hypothyroidism due to either dysgenesis or dyshormonogenesis. Only unaffected parents are included as controls.
UK10K Rare Hypercholesteromia	123	UK	Trios in which the probands have a consistently high level of LDL, and do not contain common APOB and PCSK9 mutations, or detectable LDLR mutations. Only unaffected parents are included as controls.
INTERVAL Sequencing Project	4499	UK	A cohort of healthy blood donors collected from 25 donation centres across England.
ENGAGE	283	Finland	A collection of individuals selected from Health 2000 and FINRISK cohorts based on properties of their metabolic trait profiles.
Health 2000 Survey	277	Finland	A study based on a nationally representative sample of persons aged 30 and over, with a goal of obtaining general public health information on the working-aged and aged population.
Familial dyslipidemia study	84	Finland	Individuals from families with dyslipidemia and are of Finnish origins. Only unrelated individuals are included.
FINRISK study controls	769	Finland	The FINRISK study is a large population survey investigating risk factors of chronic, non-communicable diseases. We include samples that are controls in an on-going inflammatory bowel disease exome sequencing study.
METSIM study	984	Finland	The cross-sectional METSIM study investigates genetic and non-genetic risk factors related to Type II Diabetes, cardiovascular disease, and insulin resistance. The controls included were sequenced to investigate rare variation related to these phenotypes.

Table 2.2 Description of samples collections included as controls in the UK10K schizophrenia analysis. 4,769 controls remained after sample quality control.

Informed consent was obtained for all samples. Further information is available at http://uk10k.org/, http://www.ddduk.org, http://www.intervalstudy.org.uk/, and http://www.sisuproject.fi/.

2.3 Sequence data production

2.3.1 Sample preparation

DNA samples in the UK10K, DDD, and INTERVAL studies were sequenced at the Wellcome Trust Sanger Institute (Hinxton, Cambridge). One to three micrograms of DNA was sheared to $\sim$100 to 400 base pairs using either a Covaris E210 or LE220 machine (Covaris, Woburn, MA, USA), and processed using Illumina paired-end DNA library preparation. Three different captures were used to capture targeted coding regions: an expanded custom Agilent SureSelect Human All Exon v.3 capture with custom ELID C0338371 in the UK10K project, the Agilent SureSelect Human All Exon v.3 Kit (ELID S02972011) in the DDD study, and the Agilent SureSelect Human All Exon v.5 kit in INTERVAL study. All libraries were subsequently sequenced on Illumina HiSeq 2000 with 75 base paired-end reads in multiple batches according manufacturer's protocol over the duration of each project.

2.3.2 Alignment and BAM processing

Sequencing reads that failed quality control (QC) were first removed using the Illumina GA pipeline. Remaining raw reads were mapped to the reference genome (GRCh37 in UK10K; GRCh37_hs37d5 in DDD and INTERVAL studies) using BWA (v0.5.9-r16 in UK10K; v0.5.10 in DDD and INTERVAL) [113], and duplicate fragments were marked using Picard (v1.36 in UK10K; v1.98 in DDD; v1.114 in INTERVAL) [120]. GATK (version 1.1-5-g6f43284 in UK10K; version 3.1-1-g07a4bf8 in DDD; version 3.2-2-gec30cee in INTERVAL) was used to perform local realignment around indels [115], and recalibrate base qualities in each sample BAM. I applied VerifyBamID (v1.0) to estimate the Freemix value, which is representative of the contamination fraction in our sequence data [121]. I used the recommended thresholds for contamination, and removed samples if they had Freemix score ≥ 0.03. 31 samples or 2% of the UK10K data set were excluded, while 201 samples or 4.5% of the INTERVAL data set were excluded. We were unsure if the excess contamination in the INTERVAL study occurred during sample extraction, preparation, or sequencing.

2.3.3 Variant calling

I first called variants in individual samples using GATK Haplotype Caller (version 3.2-2-gec30cee). All samples were merged into random batches of 200 using CombineGVCFs, and then joint-called using GenotypeGVCFs at default settings [115, 122]. Because three different exome captures had been used, variant calling was performed on the union of Agilent v.3 and v.5 baits with 100 base pairs of flanking sequence. I subsequently ran the GATK VQSR on all GENCODE coding variants using default settings. This joint calling protocol was suggested by the GATK development team for the production of large sequencing data sets.

2.4 Variant calling and quality control across capture and batch

2.4.1 Adjusting for differences between capture and batch

The sequence data for individuals of UK ancestry was generated at the Wellcome Trust Sanger Institute using the same Illumina sequencing platform and some version of the SureSelect Human All Exon v.3 or v.5 captures. However, substantial differences exist between the exome captures, and this must be carefully adjusted for if samples were to be jointly analysed in a case-control framework. The v.5 capture improved coverage across the entire exome by shifting problematic coding baits into the intronic region and excluding a small percentage of repetitive and problematic genes. Because of this, the v.3 and v.5 captures shared only 77% of their targeted regions, and a simple intersection could not be used to prioritise genomic regions for a joint analysis. To best harmonize calls across projects, I first re-called samples together using a common calling pipeline at the union of both Agilent captures with 100 bp of flanking sequence. Instead of calculating coverage at v.3 and v.5 captures, I calculated per-sample read depth at all coding exons defined by GENCODE version 19 to evaluate differences in coverage and sequence quality [123]. From these data, I identified a set of well-behaved coding regions with sufficient coverage across batches and captures for subsequent QC and analysis.

In Figure 2.1, the v.5-captured samples (INTERVAL) had lower read depth across the entire exome, but covered a larger percentage of coding regions than in earlier v.3 captures (DDD and UK10K). The samples in the UK10K study were divided into two batches, reflecting a known chemistry change that occurred early in the project. DDD exomes more closely resembled the UK10K v.3 samples in regional coverage but clear differences still

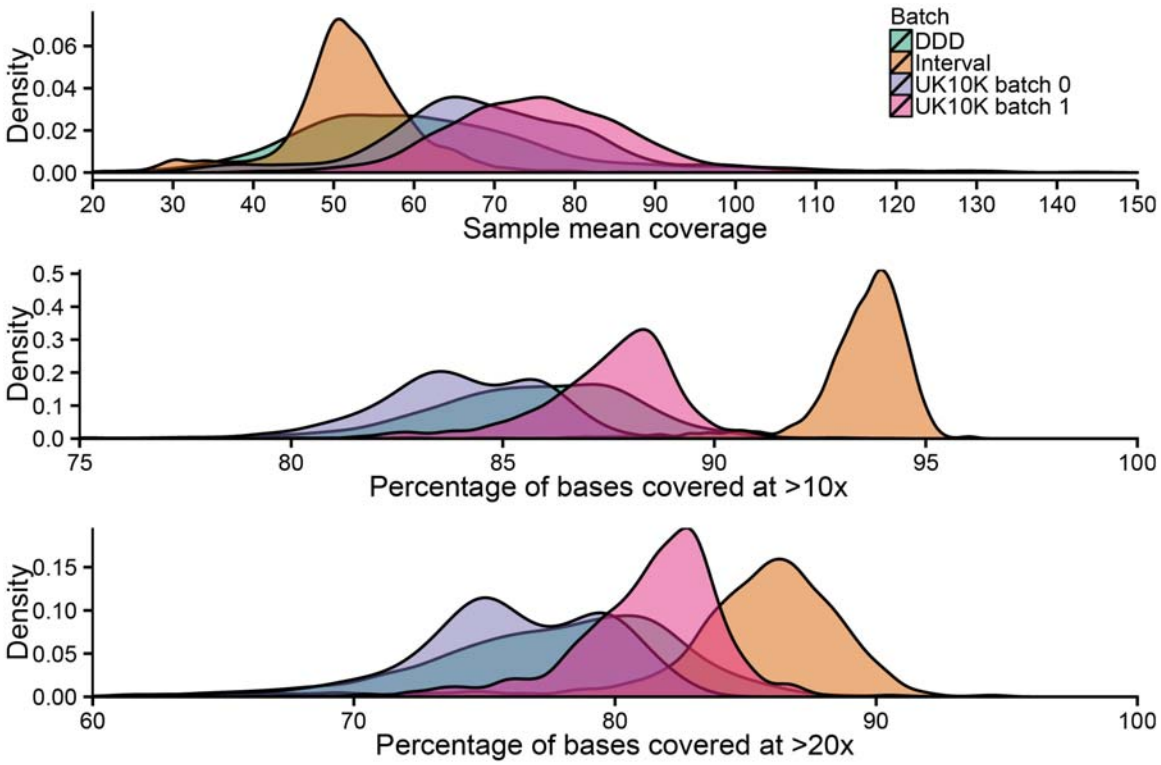


Fig. 2.1 **Density plots of sequence coverage in the UK10K, INTERVAL, and DDD datasets.** Per-sample sequence coverage was calculated and summarised from exome sequencing data generated in the UK10K (N = 4,734 in batch 0, and N = 562 in batch 1), INTERVAL (N = 4,502), and DDD (N = 1,972) datasets. The UK10K dataset was separated into two sequencing batches. Top: sample mean coverage; Middle: percentage of GENCODE v19 coding bases covered at 10× or more in each sample; Bottom: percentage of GENCODE v.19 coding bases covered at 20x or more in each sample.

existed between the v.3 and custom v.3 capture. Since all schizophrenia cases were sequenced using the v.3 capture, I have less power to detect rare variant associations in regions where this capture has limited coverage. I restricted our analysis to variants with a read depth of $7\times$ or more in at least 80% of samples in each of the four batches (UK10K batch 0, UK10K batch 1, DDD, and INTERVAL). For a more stringent filter, I identified exons that were covered at $10\times$ or more in at least 80% of samples in each batch for a total of 28.5 Mbs. By applying these filters, I excluded regions that were not covered with sufficiently high depth in our v.3-captured cases, or were not targeted in our v.5-captured controls by design.

2.5 Sample-level quality control for case-control analysis

The combined case-control data set consisted of individuals recruited from three countries: the UK, Sweden, and Finland. The UK and Finnish cases were recruited as part of the UK10K project, and the Swedish individuals were recruited in an independent study. While cases were called with nationality-matched controls, each subgroup was processed and sequenced at a different location with different reagents, and had to be analysed separately to reduce the effects of possible confounders like population stratification. Because of this, I performed sample-level and variant-level quality control steps on each nationality separately, and describe these steps in detail below.

2.5.1 Sample-level QC in the UK10K-INTERVAL case-control data set

In the UK10K data set, we sequenced the exomes of 1,488 UK individuals with schizophrenia and 5,469 matched controls without a known neuropsychiatric diagnosis. After per-sample depth analysis, I removed 22 samples with low coverage ($\leq 75\%$ of the GENCODE v.19 coding region covered at $\geq 10\times$). I next identified high-quality LD-pruned SNPs to investigate familial relatedness, non-European population ancestry, and outlying heterozygosity rates in our data set. To acquire these variants, I extracted common SNPs (MAF $> 5\%$) that passed a stringent VQSR threshold (tranche sensitivity 99.0%), had missingness $< 3\%$, and Hardy-Weinberg equilibrium χ^2 P-values $> 1 \times 10^{-3}$ in the UK10K and INTERVAL sequencing batches. I merged this subset of samples and variants with the 1000 Genomes Phase III release, and retained 43,837 SNPs with MAF $> 5\%$ and missingness $< 3\%$ in the combined dataset. These variants were LD-pruned on PLINK v1.9 with parameters –indep-pairwise 50 5 0.2 while excluding extended regions of high LD (chr 6: 25,000,000-35,000,000, and chr 8: 7,000,000-13,000,000) [124]. After filtering, a total of 19,554 high-quality LD-pruned SNPs were available for analysis.

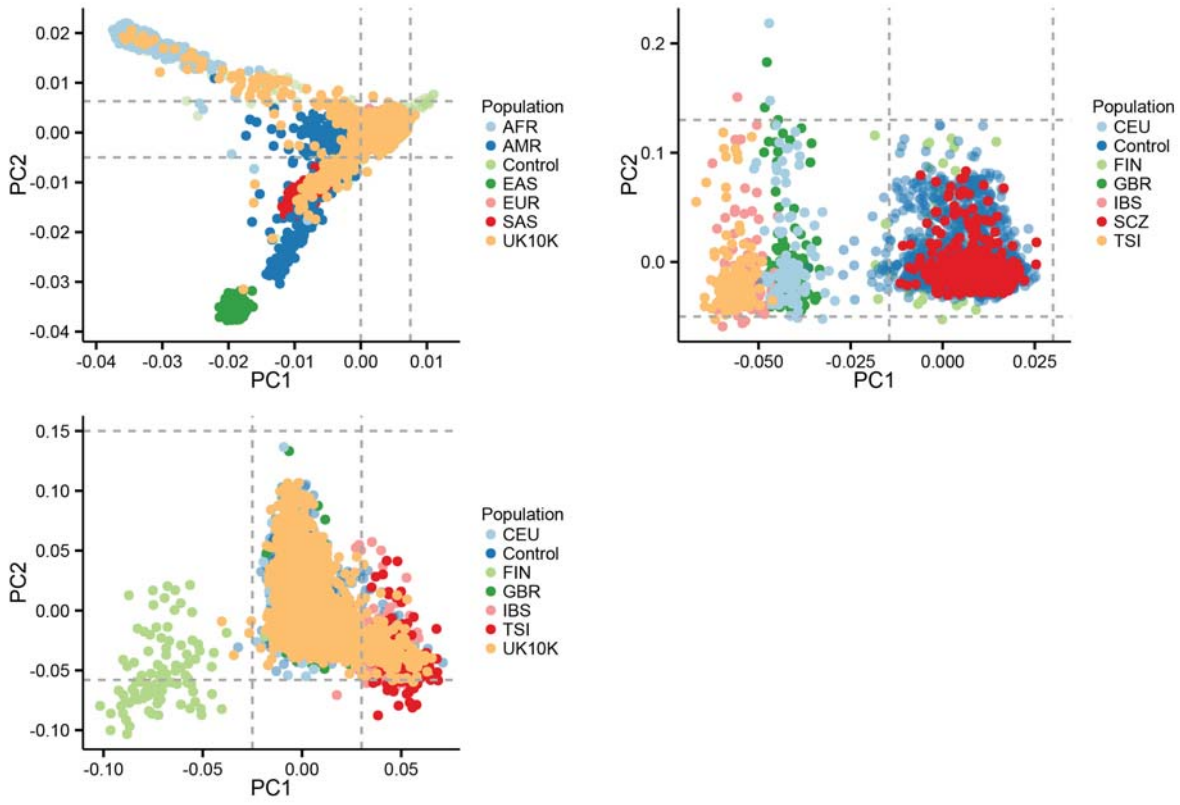


Fig. 2.2 **Principal components analysis of UK and Finnish samples in our UK10K schizophrenia dataset.** Principal components were estimated using 1000 Genomes samples, onto which I projected our cases and controls. I verified if samples had the same population ancestry (UK or Finnish) as reported in the sample manifests, and excluded individuals who were of non-European ancestry. Thresholds for sample inclusion and exclusion are shown as dashed lines in each plot. **Top left**: Population structure of all UK10K samples, with 1000 Genomes populations used as bases. Samples bracketed by the dotted lines are of European ancestry; **Bottom left**: PCA plot of individuals of non-Finnish European ancestry in the UK10K dataset with 1000 Genomes European populations used as bases. Samples not within the UK cluster (bracketed by the dotted lines) were excluded from analysis; **Top right**: PCA plot of individuals of Finnish ancestry in the UK10K dataset. Samples not in the Finnish cluster (bracketed by the dotted lines) were excluded from analysis. The three-letter symbols describing each population originate from nomenclature in the 1000 Genomes Project. UK10K: samples in our case-control study; SCZ: schizophrenia cases; Control: controls from our study.

Principal components analysis (PCA) was performed using PLINK v1.947 with 1000 Genomes Phase III samples as reference populations. I observed that 407 individuals were of non-European ancestry (Figure 2.2). In a second PCA using only European populations as reference, I observed that our samples were predominantly of UK or North European ancestry, with a small number of cases more related to 1000 Genomes individuals from the Iberian peninsula (Figure 2.2). I retained these individuals, but noted that they may have to be grouped into a separate batch or excluded in later analyses. I estimated kinship coefficients between each sample pair using KING v1.448 [125], and removed 39 duplicate samples and 68 samples with abnormal values likely due to some level of contamination. Individuals in first, second, and third-degree relationships were identified, and 190 samples were selectively removed until the maximum pairwise kinship coefficient within the cohort is 0.09375. In all, 826 samples were removed during QC, resulting in a final cohort of 6,122 UK samples (1,353 cases and 4,769 controls).

2.5.2 Sample-level QC in the Finnish and Swedish case-control data sets

In the UK10K data set, we sequenced the exomes of 399 Finnish individuals with schizophrenia and 2,116 matched controls, and performed variant calling using the GATK pipeline at the Broad Institute (Cambridge, MA). After obtaining unprocessed VCFs containing these samples, I excluded 16 samples with lower-than-expected coverage, and determined that all samples within the Finnish data set were of either non-Finnish European or Finnish ancestry (Figure 2.2). A more detailed projection using 1000 Genomes European individuals revealed that 27 samples were more closely related to non-Finnish Europeans in ancestry, and I excluded these 27 individuals from further analysis. From relatedness analysis, I excluded 67 samples. In all, 103 samples were removed during QC, resulting in a final data set of 2,412 samples (392 cases and 2,020 controls). A similar analysis within the Swedish case-control data set determined that all samples were of non-Finnish European or Finnish ancestry. I excluded 17 samples due to relatedness, resulting in a final data set of 5,073 individuals (2,519 cases and 2,554 controls).

2.6 Variant filtering in case-control data sets

2.6.1 Variant filtering in the UK10K-INTERVAL data set

Standard protocol for variant filtering recommends the use of GATK VQSR for calculating the probability that a variant is real, and selecting a threshold that maintains a desired sensitivity for true variants. The VQSR model trains on the annotation metrics (mapping quality, strand bias, quality by depth) of validated variants from the HapMap project and the 1000 Genomes Project to classify the remaining variants. However, recent studies have suggested that VQSR is less effective in filtering ultra-rare variants, especially those that are seen only once (singletons) or twice (doubletons) in the data set [126]. Notably, VQSR does not filter individual genotypes, which allows low-quality calls to be inaccurately retained if that site on average passes VQSR filtering. The inability to remove these low-quality genotypes within variable sites adds unnecessary noise in downstream analyses. However, recommended thresholds for filtering individual genotypes have not been established.

To complement the GATK filtering step, I empirically derived site and genotype filters by evaluating the sensitivity and specificity of different thresholds using a training set consisting of real rare variants and sequencing artefacts. First, I assumed that rare and singleton ExomeChip genotype calls in 295 UK10K cases (83 in batch 0, 212 in batch 1) represented real variants, and evaluated concordance with corresponding calls in our sequence data to assess sensitivity. Second, I identified inherited variants unique to parent-proband pair (inherited doubletons) and Mendelian inheritance inconsistent variants within DDD parent-proband trios to evaluate SNP and indel filtering thresholds. I computed the percentage of inherited variants retained and putative *de novo* variants removed at various thresholds to evaluate the effectiveness of our variant filtering. Using these data, I explored genotype thresholds across a number of variant and genotype-level metrics, including VQSLOD score, reference allele read depth (DP0), alternate allele read depth (DP1), allelic balance (AB), genotype quality (GQ), and mean genotype quality (GQ_MEAN). Variant thresholds were determined for SNPs and indels separately. In summary, I used rare array-based variants and rare Mendelian inheritance consistent (truth sets) and inconsistent variants from trios (false set) to calibrate variant filtering thresholds.

Variant filtering thresholds for SNPs

Applying the following filters achieved a reasonable compromise between sensitivity and specificity within our case-control data set (Figure 2.3):

- Exclude variants outside the VQSR tranche with 99.75% sensitivity

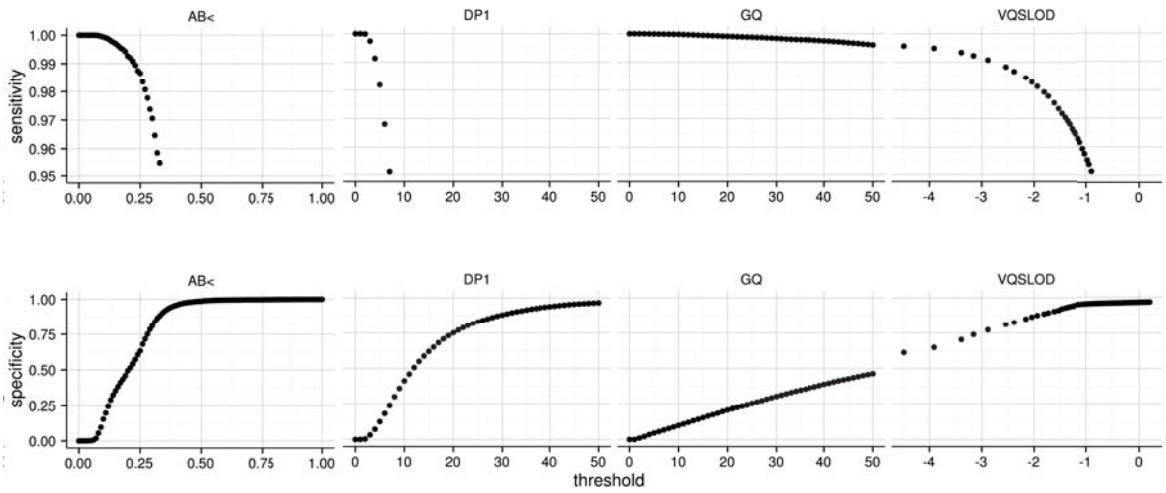


Fig. 2.3 **The evaluation of different variant filtering thresholds using rare DDD inherited variants and Mendelian inconsistent variants as a testing set.** I evaluated the sensitivity and specificity across a range of thresholds for each variant and genotype-metric. AB<: retain variants with allelic balance greater than this threshold; DP1: retain variants with alternate allele read depth greater than this threshold; GQ: retain variants with genotype quality greater than this threshold; VQSR: retain variants with a GATK variant recalibration scores greater than this threshold.

- Exclude variants with mean GQ < 30
- Exclude genotype calls with GQ < 30
- Exclude genotype calls with DP1 < 2
- Exclude genotype calls with AB < 0.2 and AB > 0.8

Using these thresholds, I removed 95.63% of all Mendelian inconsistent genotype calls while retaining 98.38% of all doubleton inherited variants. In the ExomeChip data set, I retained 99.45% of variants seen only once in the UK10K samples, and 99.62% of all heterozygote calls. While GATK recommended a more conservative VQSLOD score threshold (either VQSRTranche99.50 or VQSRTranche99.0), I found that a less stringent VQSR filter combined with genotype-level thresholds retained a larger percentage of rare inherited variants while attaining reasonable specificity. If VQSR were applied without genotype-level filters, only 40.8% of all Mendelian inconsistent genotype calls would be excluded were I to maintain a comparable sensitivity of 98% for doubleton inherited variants. I also removed SNPs with missingness greater than 20%, and tested SNPs for deviation from Hardy-Weinberg equilibrium within each sequencing batch (UK10K batch 0, UK10K batch 1, and INTERVAL) and within the entire data set. The Hardy-Weinberg filter addressed mapping issues that arose from differences in exome baits or decoy sequences used during alignment:

mismapped variants often are seen only as heterozygotes in one batch and homozygotes in another. Any variant that deviated from Hardy-Weinberg equilibrium with χ^2 *P*-values of $< 1 \times 10^{-8}$ in any batch or in the entire data set was excluded. Finally, I excluded variants that resided in low-complexity regions, the 2% of the genome highly enriched for repetitive sequences in which alignment and variant calling is more difficult (see Heng *et al.* [117] for a more precise definition and motivation for its use). At each stage of filtering, I reported the per-sample transition-transversion rate (TiTv), the number of heterozygote calls, the number of non-reference homozygous calls, and the number of variants observed only once within the UK10K-INTERVAL call set (Figure 2.4, 2.5). The variant metrics appeared comparable across the four batches, and the mean sample TiTv was ∼3.26, the expected rate for coding SNPs in European populations.

Variant filtering thresholds for indels

Using the same approach described above for SNPs, I found that the following filters achieved a reasonable compromise between sensitivity and specificity for indel discovery within our case-control data set:

- Exclude variants outside the VQSR tranche with 99.50% sensitivity
- Exclude variants with mean GQ < 90
- Exclude genotype calls with GQ < 90
- Exclude genotype calls with DP1 < 2
- Exclude genotype calls with AB < 0.25 and AB > 0.8

Using these variant and genotype-level thresholds, I removed 92.35% of all unfiltered Mendelian inconsistent indel calls while retaining 93.60% of all doubleton inherited indels. Applying VQSR alone was not sufficient to acquiring a clean indel set: even at a stringent VQSLOD threshold of -0.3151 (VQSRTrancheINDEL0.00to99.00), I only achieved specificity of 40.72% for Mendelian inconsistent indels. I also removed indels with missingness greater than 20%, and those that deviated from Hardy-Weinberg equilibrium with χ^2 *P*-values of $< 1 \times 10^{-8}$. I removed indels that resided in low-complexity, highly repetitive regions (defined in the previous section) that could not be appropriately aligned using short-read technology. Lastly, I excluded all indels that have more than two alternate alleles, or were clustered within 3 bp of another indel. Following these indel filtering steps, the number of indels and frameshift:inframe ratio appeared comparable across all batches (Figure 2.6).

From previous studies of parent-proband trio studies, we expected to find one coding *de novo* mutation per proband [79, 80]. In our DDD trio data set, we observed 92 *de novo* SNVs

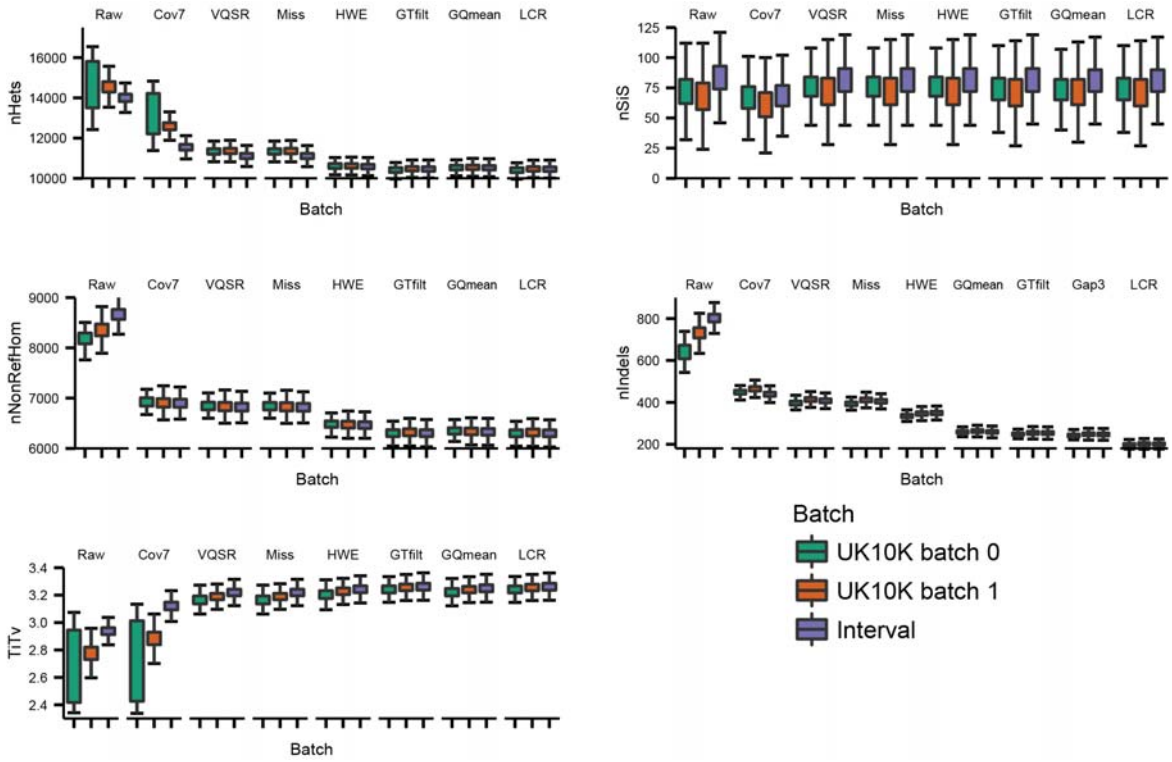


Fig. 2.4 **Variant metrics in the UK10K and INTERVAL datasets after each variant filtering step.** Box plots of per-sample heterozygote count (nHets), non-reference homozygote count (nNonRefHom), TiTv (TiTv), number of singletons (nSiS), and number of indels (nIndels) following each variant QC step. Variant metrics were summarised across all samples in the UK10K and INTERVAL datasets. Raw: no variant QC steps applied; Cov7: restricting to variants with at least 7× mean coverage; VQSR: GATK variant calibration using default parameters; Miss: filter for excess missingness; HWE: filter for deviation from Hardy-Weinberg equilibrium; GTfilt: filter for low alternate allele read depth, and abnormal allelic balance; GQmean: filter for low genotype quality; LCR: exclude variants in low-complexity regions.

and 12 *de novo* indels per proband prior to variant filtering, and 4 *de novo* SNVs and 0.92 *de novo* indels per proband after variant filtering. The observed *de novo* mutation rate in our data set still exceeded the expected rate of mutation described in previous studies, suggesting that our variant QC was not sufficiently strict to over-filter genuine *de novo* events while vastly reducing the number of false positives.

2.6.2 Variant filtering in the Finnish and Swedish data sets

In the Finnish data set, SNPs and individual genotype calls were excluded according to the following criteria: VQSLOD < -2.6557 (VQSRTrancheSNP99.75), GQ < 30, or GQ_MEAN < 30. Indel sites and genotypes were excluded according to the following criteria: VQSLOD < -0.2731 (VQSRTrancheIndel99.50), GQ < 90, or GQ_MEAN < 30. In addition, I removed variants with missingness greater than 20%, or if they deviated from Hardy-Weinberg equilibrium with χ^2 *P*-values of $< 1 \times 10^{-8}$. All variants within low-complexity regions were excluded. I also removed all indels that have more than two alternate alleles, or were located within 3 base pairs of another indel. After variant and genotype-level QC, the sample TiTv and frameshift:inframe ratio was $\sim$3.29 and $\sim$1.01 respectively, which was comparable across batches of the Finnish data set and with the UK10K-INTERVAL call set (Figure 2.5, 2.6).

I was unable to acquire raw BAMs for the Swedish data set to re-call and perform QC from scratch. However, the Swedish data set as provided already had very stringent filters applied during a previous analysis, and I analysed the dataset with little additional QC. Variant sites and genotypes were filtered out if the Hardy-Weinberg equilibrium χ^2 *P*-values $< 1 \times 10^{-8}$, missingness > 20%, or if they reside within in low-complexity regions. After variant and site QC, the sample TiTv and frameshift:inframe ratio was $\sim$3.28 and $\sim$1.15 respectively, which was comparable across batches and with the UK10K-INTERVAL call set.

2.7 Comparison of population genetics metrics across data sets

Following sample and variant QC, 6,122 UK samples (1,353 cases and 4,769 controls), 2,412 Finnish samples (392 cases and 2,020 controls), and 5,073 Swedish samples (2,519 cases and 2,554 controls) were available for analysis. Variant counts and population genetic metrics between data sets and sequencing batches were harmonized: the sample TiTv (mean $\sim$3.25) and the frameshift:inframe ratio were comparable across all populations and batches

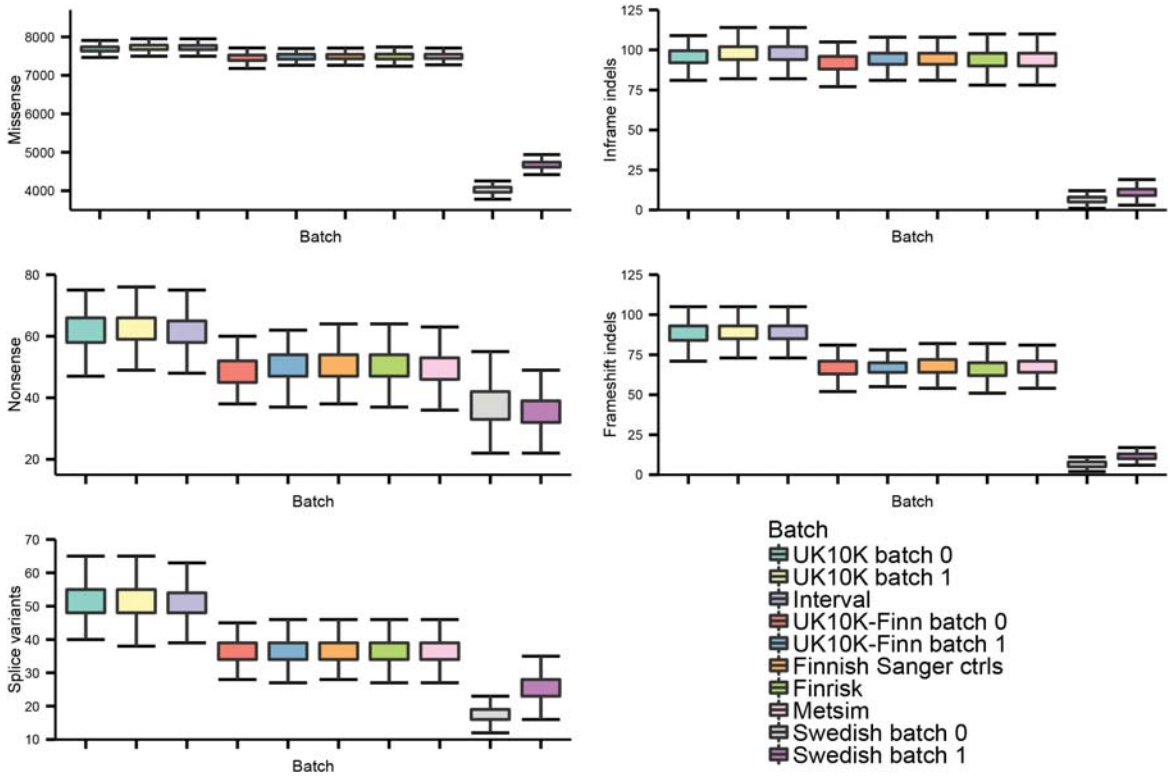


Fig. 2.5 **Variant counts summarised according to variant class and sequencing batch in the UK10K, INTERVAL, Finnish, and Swedish datasets.** Box plots of per-sample variant counts in the UK10K, INTERVAL, Finnish, and Swedish datasets. All samples included in our meta-analysis are represented in the figure. The UK10K datasets was sub-divided according to sequencing batches (batch 0 and batch 1), and sample ancestry (UK and Finnish). The Finnish control datasets was separated by study of origin (Metsim, Finrisk, and Sanger controls). The Swedish case-control dataset was separated into two sequencing batches. Differences exist in total variant counts between the UK, Finnish, and Swedish collections, likely reflecting differences in sequencing depth, capture reagents, sequencing protocol, read alignment, and variant calling. However, variant counts and population genetics metrics were consistent between cases and controls within each population group.

(Figure 2.6). However, I still observed some differences between variant counts between the UK, Finnish, and Swedish data sets (Figure 2.5). The UK, Finnish, and Swedish samples were independently produced and called at different sequencing centres, and the discrepancy in variant counts likely reflected differences in capture, sequencing batch, calling procedure, and quality control. In particular, the Swedish data set we acquired from dbGAP underwent extremely stringent variant filtering, and had per-sample variant counts nearly half of that observed in the UK10K-INTERVAL data set and the 1000 Genomes Phase III data set. These differences would confound rare variant tests and need to be explicitly corrected for. In subsequent analyses, I adjusted for between-population differences by treating them as separate analytical groups. More importantly, cases and controls within each population group appeared to be well-matched, and this was reflected in the null statistics of subsequent variant and gene-based analyses.

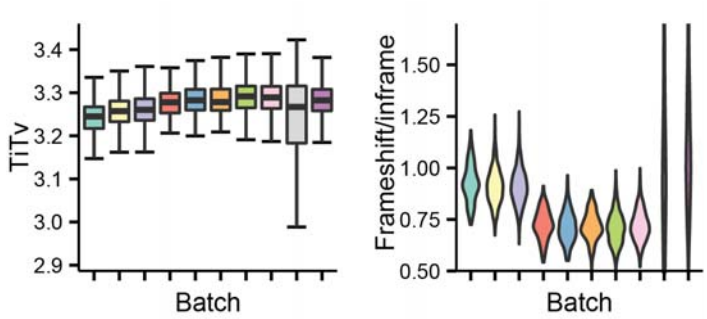


Fig. 2.6 **Distributions of TiTv and frameshift-inframe ratios in the UK10K, INTERVAL, Finnish, and Swedish datasets.** Here, I have a box plot of sample TiTv (left) and violin plot of sample frameshift-to-inframe ratio (right) in the UK10K, INTERVAL, Finnish, and Swedish datasets. All samples included in our meta-analysis are represented in the figure. See 2.5 for the legend, and a description of each batch and sub-study. Following sample and variant QC, the per-sample transition-to-transversion ratio was comparable between all populations (mean $\sim$ 3.25).

2.8 Systematic annotation of coding variants

I used the Ensembl Variant Effect Predictor (VEP) version 75 to annotate coding variants with GENCODE version 19 transcripts as reference [127]. VEP plugins were used to apply *in silico* classifiers to missense variants, such as PolyPhen, SIFT, and CADD [128–130]. For each variant, I assigned a functional consequence on a per-gene basis, aggregating all transcript-level annotations and retaining only the most severe consequence. Coding variants were assigned into the following functional categories:

1. Loss-of-function or disruptive (LoF) variants
    - Frameshift
    - Stop-gained
    - Splice acceptor and donor variants
2. Initiator codon variants
3. Inframe deletion or insertions
4. Missense variants (mis)
5. Synonymous variants

Following other rare variant studies [94, 103, 105], I stratified our analyses into two functional classes: 1. LoF variants, 2. missense or initiator codon variants.

2.9 Evaluating the effectiveness of existing *in silico* predictors of pathogenicity

The use of variant annotation tools to prioritise coding variation has helped increase statistical power for gene discovery [88, 94]. Most variants identified in the coding region reside in the rarer end (MAF $< 0.1\%$) of the allele frequency spectrum (AFS) [78]. If the predicted functional consequence of variants were disregarded, a simple comparison of allele counts between cases and controls would be diluted by large numbers of non-functional variants [88]. Functional annotation tools intend to accurately distinguish the pathogenic, disease-causing variants from neutral polymorphisms, thus enriching our analyses on causal risk variants while decreasing the rate of background noise. However, over-filtering and removing true signals can have a detrimental effect on our power, especially when the allele counts of rare damaging variants are already low due to purifying selection. A delicate balance between specificity and sensitivity in annotating disease-causing variants is required to maximize our power in detecting true associations.

2.9.1 The interpretation of protein-coding consequences

In the simplest case, a variant is annotated as functional based on its effect on a protein product. A true loss-of-function (LoF) variant either drastically reduces levels of the gene product or disrupts a protein's ability to carry out key functions. This can be through truncations, aberrant splicing, shifts in coding sequences, and pre-mature stop codons. A

missense variant causes an amino acid substitution, which can lead to change in protein functionality. Many of these changes are benign and would not be subject to strong selection: missense variants may substitute amino acids without affecting charge and folding or disrupt a domain or peptide that is irrelevant to protein function. On other hand, some missense variants eliminate protein function by disrupt protein folding or modifying the charge of an active site. Thus, even if a variant labelled as missense or loss-of-function by VEP, additional information is needed to properly evaluate its pathogenicity.

2.9.2 A description of existing annotation tools

Because of this, a series of statistical tools have been developed to predict the pathogenicity of missense variants. These missense classifiers primarily differ in the statistical approach applied, the features inputed into the model, and the training and testing data set used for calibration and evaluation (Table 2.3). For instance, PolyPhen2 uses a Bayesian classifier to characterize missense variants based on structural information about the binding site, protein domains, contact with ligands, and subunit interactions [129]. All the calculated features are trained using a Bayes classifier on the HumDiv data set, a curated list of variants causing Mendelian disorders, and the Humvar data set, a more comprehensive list of risk alleles from UniProt [131]. SIFT, another popular tool, models function using a multiple sequence alignment of proteins, and determines which base mutations was most tolerated across similar proteins [128]. A SIFT score of 0.05 indicates the alternate allele was observed in 5% of all alignments and could be considered not as damaging. Other missense classifiers include LRT, MutationTaster, MutationAssessor, FATHMM, Radial SVM, MetaLR, GERP++, PhyloP, Condel2, and SiPhy [132, 133]. Some of these, like GERP++, and PhyloP, classify variants based on the degree of sequence conservation between species, while others, like Radial SVM and Condel2, are ensemble classifiers that integrate results from other tools to annotate variants. CADD differs in its approach completely by simulating its training set and comparing these randomly generated alleles to the set of derived alleles common between the human-chimpanzee ancestral genome [130]. A support vector machine with a large set of features, including SIFT and PolyPhen, was used to model the relative deleteriousness of all possible alleles across the genome. Unlike the other tools, CADD could annotate both coding and non-coding variants.

When evaluating these models, differences in statistical approach, input features, and training and testing data must be carefully considered to prevent issues of circularity and bias (Table 2.3). For example, MutationalTaster incorporates frequency information from 1000 Genomes when determining pathogenicity; a testing set consisting of rare damaging variants as pathogenic and common variants as benign would inflate the classifier's effectiveness.

Ensemble classifiers like CONDEL and Radial SVM incorporated MutationTaster, SIFT, and PolyPhen as features, and indirectly incorporated frequency information. Furthermore, only a few robust variant data sets exist for evaluating the effectiveness of each classifier, and many of these classifiers already use them for training. For instance, PolyPhen, CONDEL, Radial SVM, and MetaLR trained on the same set of curated coding variants provided by Uniprot, while others trained on the Human Gene Mutation Database (HGMD) database. Therefore, a new and wholly independent dataset is best suited for evaluating the performance of these *in silico* classifiers.

2.9.3 Strategy for evaluating variant annotation tools

I evaluated the effectiveness of available annotation tools for LoF and missense variants using a series of novel variant sets previously not used for classifier training. First, I used a clinician-curated set of variants from the DDD study. *De novo* and inherited variants were identified and validated in 1,133 affected probands, and variants disrupting known developmental disorder genes were manually curated to determine if these variants were pathogenic relative to the patient's phenotype. I used all clinically reportable variants as a truth set, and all rejected variants as a false set. For an additional truth set, I accumulated *de novo* mutations from 2,263 trios sequenced as part of the Autism Sequencing Consortium, and 2,500 trios sequenced in the Simon Simplex Collection. I identified all *de novo* missense variants disrupting autism risk genes from Sanders *et al.* [109] as another truth set.

The ExAC database contained coding variants from 60,706 unrelated individuals without severe paediatric diseases joint-called in a single pipeline [112]. It is important to note that this release of ExAC contained a number of individuals with psychiatric phenotypes. I assumed that the fraction of pathogenic variants in this data set was substantially lower than in the DDD and ASD studies, and used missense variants with MAF $< 1\%$ in ExAC as a false set. Finally, I re-annotated a large set of functional, protein-coding variants manually curated by Uniprot for an additional training set. This truth set consisted of variants described by Uniprot as disease-causing and our negative truth set were variants described as general polymorphisms. I applied the following *in silico* classifiers to the missense variants: PolyPhen2, SIFT, LRT, MutationTaster, MutationAssessor, FATHMM, Radial SVM, MetaLR, GERP++, PhyloP, Condel2, CADD, and SiPhy. Using causal variants identified in the DDD and ASD studies, I determined guidelines for prioritising variants in trio and case-control analyses.

Name	Method	Features	Training set	Classifies
CADD	Support vector machine.	Conservation metrics, functional genomic data, transcript information, and protein-level scores. 63 annotations in total.	Derived alleles common between the human-chimpanzee ancestral genome and simulated alleles.	All variants (SNPs and indels of all classes)
CONDEL2	Weighted average of other classifier scores.	Other variant classifiers, including PolyPhen2, Mutation Assessor, and SIFT.	HumVar and HumDiv variant databases.	Missense variants
FATHMM	Hidden Markov Model.	Alignment of homologous sequences across species, along with an overlaying of protein domain information.	HGMD and UniProt databases.	Missense variants
GERP++ RS	Maximum likelihood estimation and dynamic programming to define constrained elements.	Multiple alignment to study conservation across species. A comparative genomics approach.	Genomes of humans and other species. Primarily mammalian.	Missense variants
LR	Logistic regression.	11 other classifiers, including PolyPhen, SIFT, MutationAssessor, FATHMM.	UniProt database.	Missense variants
LRT	Likelihood ratio test.	Multiple alignment of human genes with other species to identify conserved regions.	Genomes of 32 vertebrate species.	Missense variants
MutationAssessor	Computing a relative conservation score, and calibrating on the training data.	Multiple alignment of gene and protein families within and between species.	UniProt database.	Missense variants
PhyloP	Unsupervised phylogenetic methods to estimate probabilities.	Sequence alignment across species to identify departures from neutral rate of substitution.	Genome-wide alignment of36 species.	Missense variants
PolyPhen2 HDIV	Naive Bayes clasifer.	Eight sequence-based and three structure-based predictive features from multiple sequence alignment.	HumDiv database.	Missense variants
Polyphen2 HVAR	Naive Bayes clasifer.	Eight sequence-based and three structure-based predictive features from multiple sequence alignment.	HumVar database.	Missense variants
RadialSVM score	Support vector machine.	11 other classifiers, including PolyPhen, SIFT, MutationAssessor, FATHMM.	UniProt database.	Missense variants
SIFT	Scores computed from sequencing alignment conservative. A scaled probability calculated for each position.	Sequence alignment across species to identify evolutionary conservation of amino acids.	Genomes of humans and other species.	Missense variants
SiPhy	Hidden Markov Model.	Inferring site-specific substitution biases directly from sequence alignments. Conservation-based method.	Genomes of humans and other species.	Missense variants
VEST3	Random forest.	86 features, including amino acid properties and conservation scores.	HGMD database and common variants in the NHLBI exome sequencing project.	Missense variants

Table 2.3 Description and summary of statistical tools developed to predict the pathogenicity of coding variants. The statistical method, features, and training set of each missense classifier were described. More information on these tools could be found in the annotation database dbNSFP [132, 133].

2.9.4 Preparation of annotation files

I used the Annovar tool and the dbNSFP v2.7 database [132, 133] to annotate all missense variants with the following classifiers: PolyPhen2, SIFT, LRT, MutationAssessor, FATHMM, Radial SVM, MetaLR, GERP++, PhyloP, CADD, and SiPhy. I used VEP to annotate variants with CADD and Condel2 scores. Condel2 scores were separately downloaded from FannsDB and parsed to be compatible with VEP.

2.9.5 Classifiers display variable performance depending on test data

First, I tested the effectiveness of the 14 classifiers in identifying pathogenic and benign variants in the UniProt data set. I found that ensemble classifiers, such as LR score, Radial SVM, VEST3, and CONDEL, had the greatest area under the curve (AUC), and reached a sensitivity and specificity of just under 90% (Figure 2.7). These classifiers used PolyPhen, SIFT and conservation scores as features to train more flexible statistical methods like the random forest and support vector machine. This was followed by CADD and PolyPhen that reached a sensitivity and specificity of just under 80%. SIFT and annotation methods based on conservation did not perform as well as the other classifiers.

I next evaluated the classifiers using pathogenic *de novo* mutations from the DDD and ASD studies as positive testing data. I first used UniProt benign polymorphisms as negative testing data. As seen in Figure 2.8, I found that the missense classifiers performed substantially worse when classifying *de novo* mutations when compared to UniProt pathogenic variants. None of the classifiers had a discrimination threshold that simultaneously achieved a sensitivity and specificity of greater than 82%. Ensemble classifiers like LR score and Radial SVM still outperformed the remaining classifiers. Along with CADD, these more flexible methods outperformed PolyPhen, SIFT, and other conservation-based annotations. Finally, I used ExAC missense variants with MAF $< 1\%$ as an alternate negative testing data set, while still using diagnostic *de novo* mutations as the positive testing set. Again, the ensemble classifiers massively outperformed the remaining annotation tools, with LR score and Radial SVM leading with the highest AUC (Figure 2.9).

I attempted to identify optimal discrimination thresholds for each missense classifier using Youden's J-statistic. Surprisingly, the optimal discrimination threshold for each annotation tool was highly sensitive to the testing data set used. For LR score, the optimal threshold for the Uniprot testing data set was 0.28, and resulted in a sensitivity and specificity of 0.92 and 0.87 respectively. However, this same threshold resulted in a sensitivity and specificity of 0.68 and 0.98 when classifying *de novo* mutations and ExAC common variants. The optimal threshold for this data set was instead 0.037, which yielded a sensitivity and specificity of

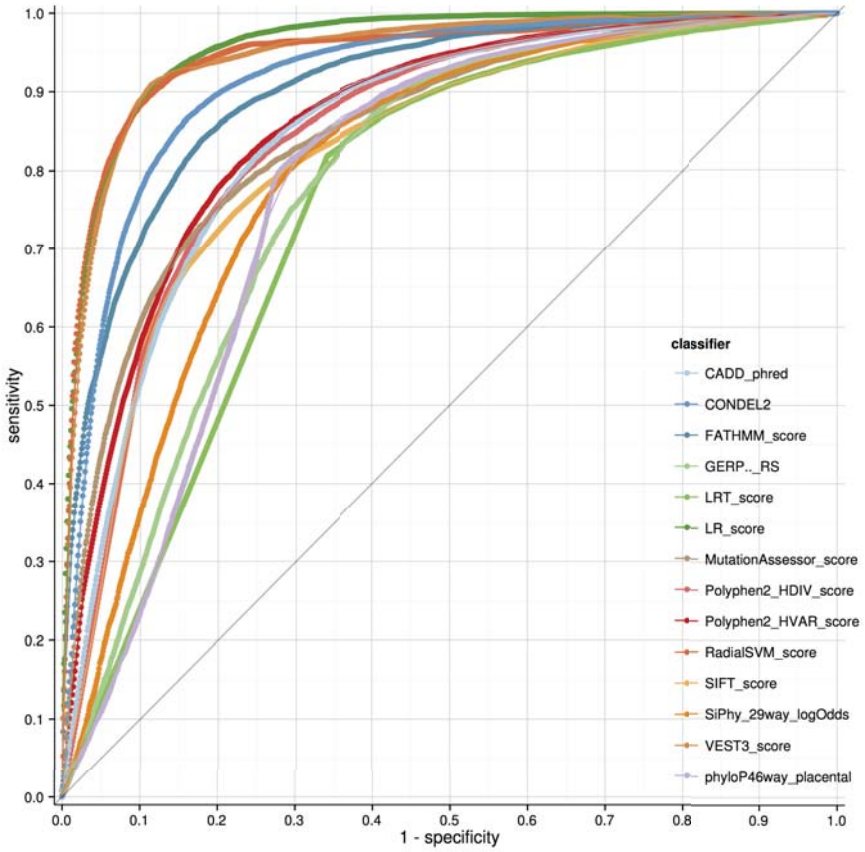


Fig. 2.7 **ROC curve evaluating the performance of missense classifiers on UniProt pathogenic-benign variants.** UniProt pathogenic variants were used as the positive testing set, while UniProt polymorphic (benign) variants were used as the negative testing set. The sensitivity and 1− specificity was plotted at various threshold settings for each classifier.

0.96 and 0.94 respectively. Unfortunately, this pattern was also observed for Radial SVM and VEST3. While the ensemble classifiers appear to outperform the other classifiers, identifying the discrimination thresholds at which this generally occurs is not at all straightforward. While it is difficult to explain this variability in optimal thresholds, these ensemble classifiers directly or indirectly incorporate allele frequency as a feature in their models, and this may lead to biases in evaluation depending on the proportion of common and rare variants in the testing data sets.

Ultimately, I selected CADD > 15 to classify missense variants as damaging in our case-control analysis. While CADD had an AUC lower than LR pred and Radial SVM, it had optimal discrimination thresholds that were highly comparable across the different testing data sets. The sensitivity and specificity at these optimal thresholds did not vary significantly between different testing data sets. For *de novo* — ExAC common variant data set, the optimal threshold was 14.1, the sensitivity was 0.84, and specificity was 0.86; for the *de novo* — Uniprot benign data set, the optimal threshold was 16.3, with a sensitivity of 0.76, and specificity of 0.79; for the Uniprot pathogenic-benign data set, the threshold was 15.4, with a sensitivity of 0.82 and specificity of 0.75. CADD performed robustly across each of our testing data sets, and its performance was superior to both PolyPhen, SIFT, and the other conservation scores. Finally, its continuous score extended to synonymous, splice, LoF, intronic, and intergenic variants, which may be useful for analyses that extended beyond missense variants.

2.9.6 A comparison of annotation approach with other whole-exome sequencing studies

While the annotation approach described here does not differ drastically with approaches used by other whole-exome sequencing studies, it does differ in some notable aspects, which I discuss here. Nearly all studies grouped functional coding variants into two categories for analysis: loss-of-function variants (defined as nonsense, essential splice, and frameshift variants), and nonsynonymous variants (defined as missense and inframe indels) [98, 103, 105, 118, 94, 112, 134, 135]. Variants were annotated based on the most severe consequence on any transcript. Where studies generally differed was in the tool used to annotate variants, the transcript reference database, and the *in silico* classifiers used to prioritise pathogenic missense variants. For instance, Purcell *et al.* and Fromer *et al.* used PLINK/SEQ to annotate variants according to the RefSeq transcript definitions; Do *et al.* and De Rubeis *et al.* used SnpEff and also according to RefSeq transcripts; the DDD and ExAC studies used VEP according to Ensembl GENCODE transcript definitions; Genovese *et al.* used SnpEff

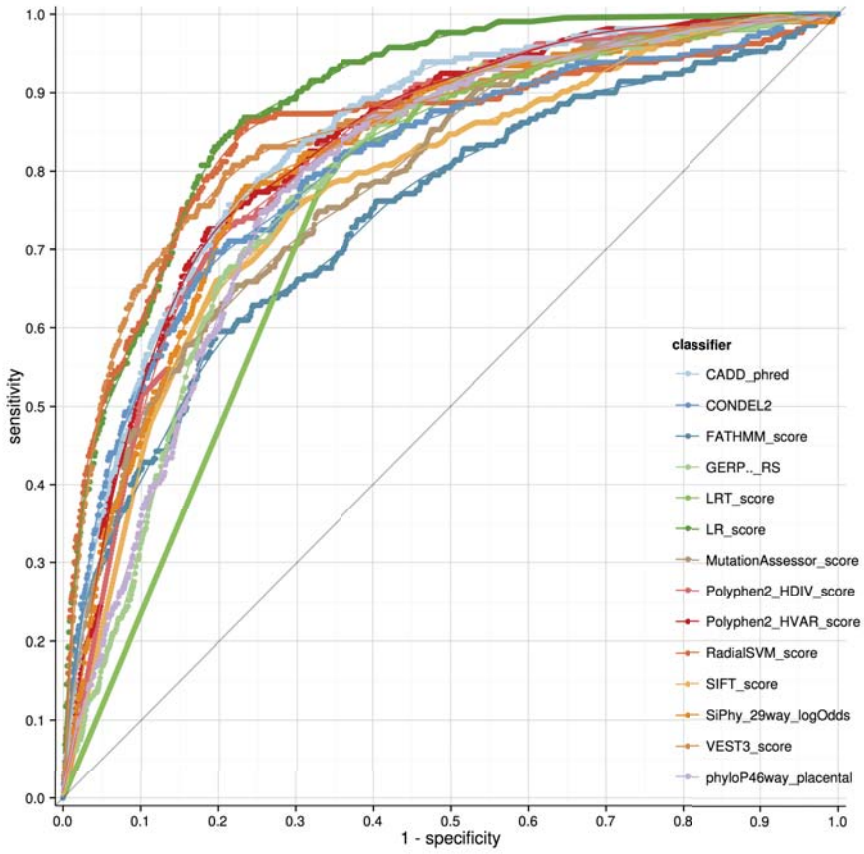


Fig. 2.8 **ROC curve evaluating the performance of missense classifiers on pathogenic *de novo* mutations and benign variants from UniProt.** Pathogenic *de novo mutations* from the DDD and autism studies were used as the positive testing set, while UniProt polymorphic (benign) variants were used as the negative testing set. The sensitivity and 1 - specificity was plotted at various threshold settings for each classifier.

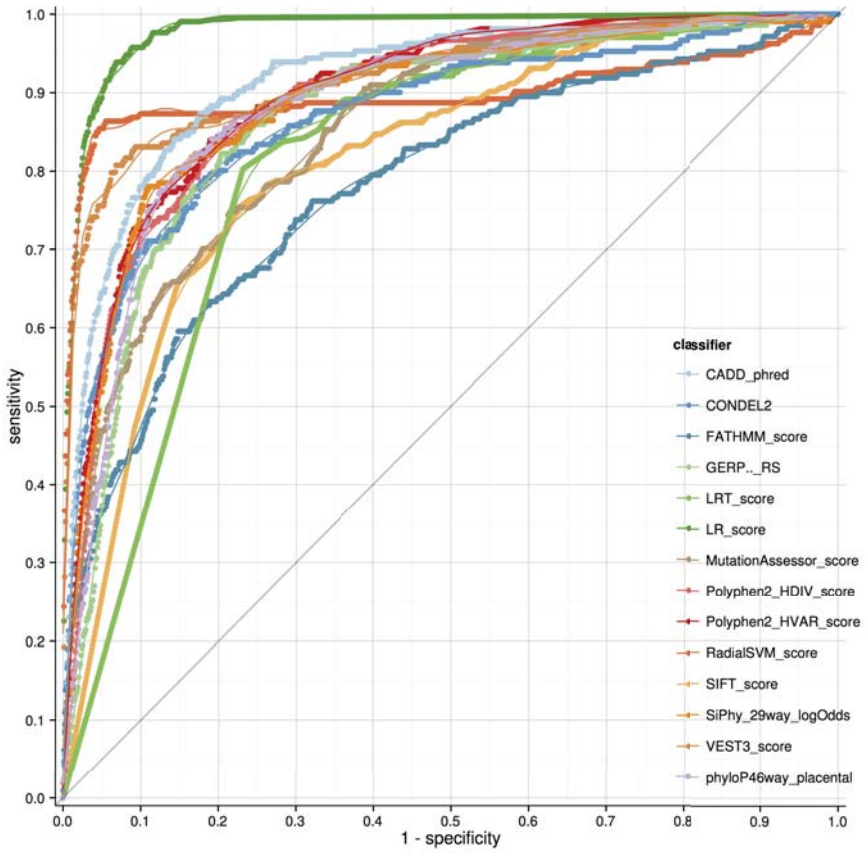


Fig. 2.9 **ROC curve evaluating the performance of missense classifiers on pathogenic *de novo* mutations and ExAC missense variants with MAF** $> 1\%$**.** Pathogenic *de novo mutations* from the DDD and autism studies were used as the positive testing set, while ExAC missense variants with MAF $> 1\%$ variants were used as the negative testing set. The sensitivity and 1 - specificity was plotted at various threshold settings for each classifier.

according to GENCODE transcripts, and Fuchsberger *et al.* used a combination of annotation tools that included SnpEff and ANNOVAR according to GENCODE transcripts. In many of these studies, little justification was provided for the choice of annotation tool and transcript database. Notably, when 80 million variants were annotated using multiple approaches, only a 83% agreement was observed in the annotation for exonic variants when using the RefSeq or Ensembl GENCODE transcript sets as references [136]. In the end, I decided to annotate variants using VEP with GENCODE transcripts as reference because a number of data sets and resources used in our analyses, such as ExAC database, GTeX database, and the DDD study, followed this approach. In addition, as seen in the following section, the GENCODE transcript reference contained a more complete set of coding genes, which permitted the analysis of an additional 1,067 protein-coding genes. However, as discussed in [136], variant annotation remained an unsolved problem, and no single annotation software or transcript set was identified as directly superior to the others.

Whole-exome sequencing studies also differed in the tools used for classifying missense variants as pathogenic and benign. Fromer *et al.*, Do *et al.* and De Rubeis *et al.* used PolyPhen-2, while Purcell *et al.*, Fuchsberger *et al.*, and Genovese *et al.* used an ensemble approach in which missense variants classified as damaging by multiple tools were defined as pathogenic. In the previous sections, I demonstrated that a number of the classifiers, including CADD, outperformed PolyPhen-2 and SIFT. On the other hand, the ensemble approach incorporated a number of tools that did not perform well in our evaluation (such as LRT), or was not very robust and had very different optimal discrimination thresholds depending on the testing set used . Furthermore, in Table 2.3, I described complicated interdependencies between the different annotation tools, in which the same data sets were used for training and evaluation, and some tools even incorporated SIFT and PolyPhen as features during training. Thus, I decided to use CADD to annotate missense variants in our analysis, which achieved reasonable sensitivities and specificities while robust to the choice of the testing data set. I did not apply LOFTEE, as no other case-control or trio study performed additional filtering on loss-of-function variants. However, this remains an unsolved problem, and no single approach could be suggested as directly superior to the others.

2.10 A meta-analysis of published schizophrenia parent-proband trio studies

Recent studies have leveraged whole-exome sequencing to identify *de novo* mutations in parent-proband trios. These mutations are very rare germline events that arose in a single

generation, and their unlikely occurrence in individual genes have been used to implicate risk genes for severe Mendelian disorders. In these disorders, gene discovery did not require a well-calibrated statistical model: for instance, five of the six probands sequenced with Wiedemann-Steiner syndrome had LoF mutations in *KMT2A* [137], while nine of the ten probands sequenced with Kabuki syndrome had *de novo* truncating events in *KMT2D* [82]. However, for more complex and heterogeneous disorders, the burden of *de novo* mutations was likely spread over many genes. Early sequencing studies of hundreds of schizophrenia and autism probands successfully demonstrated that a genome-wide excess of *de novo* mutations existed in cases compared to controls [80, 96], but were underpowered to identify individual genes.

Because recent studies have suggested that case-control and *de novo* data appeared to implicate an overlapping set of genes [105], I aggregated validated de novo mutations identified in schizophrenia trios from seven published studies for analysis with our case-control cohort [98, 99, 95, 97, 100–102]. I ensured that all *de novo* mutations included our analysis had been validated with Sanger sequencing, and that each parent-proband trio was included only once in our analysis (Table 2.4). For example, the Xu *et al.* 2011 and 2012 studies and the Takata et al. 2014 study analysed trios from the same underlying cohort. After excluding sample duplicates, I identified 118 LoF and 662 missense *de novo* mutations in 1,077 schizophrenia probands for subsequent analysis.

2.11 Gene-specific mutation rates based on GENCODE transcripts

To implicate individual genes using *de novo* mutations, a robust method of evaluating the excess of *de novo* events is needed. One approach to evaluating the excess of *de novo* mutations is to first estimate the expected per-generation rate of new mutations in gene g (μ_g). Given this gene-specific rate, the probability of observing X new mutations in gene g as observed in N trios can be modelled using the following Poisson distribution:

$$X \sim Pois(2N\mu_g)$$
$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} P(X = i)$$

where X is number of *de novo* mutations in gene g, μ_g is the gene-specific mutation rate, and N is the number of trios in our study. However, establishing robust gene-specific mutation

First author	Year	Journal	Sample size	Capture	Sequencer	Validation	PMID
Guipponi	14	PLOS ONE	53	Agilent SureSelect Human ALL Exon kits	HiSeq	Yes (Sanger)	25420024
Takata	14	Neuron	231	Agilent SureSelect v2 (n = 85 trios), NimbleGen SeqCap EZ v2 (n = 180 trios)	HiSeq	Yes (Sanger)	24853937
McCarthy	14	Mol Psychiatry	57	NimbleGen's SeqCap EZ Human Exome Library v2.0 probes	HiSeq (101 bp PE reads)	Yes (Sanger)	24776741
Fromer	14	Nature	617	Agilent SureSelect Human All Exon v.2, Nimblegen SeqCap EZ Human Exome Library v2.0, Agilent SureSelect Human All Exon 50MB	HiSeq (76 bp, 101 bp, 76 bp PE reads)	Yes (Sanger)	24463507
Gulsuner	13	Cell	105	NimbleGen SeqCap EZ Human Exome Library v2.0	HiSeq (101 bp PE reads)	Yes (Sanger)	23911319
Xu	12	Nature Genetics	231	Agilent SureSelect v2 (n = 85 trios), NimbleGen SeqCap EZ v2 (n = 180 trios)	HiSeq	Yes (Sanger)	23042115
Xu	11	Nature Genetics	53	Agilent SureSelect Human All Exon Target Enrichment System	HiSeq (50 bp PE reads)	Yes (Sanger)	21822266
Girard	12	Nature Genetics	14	Agilent SureSelect All Exome Kit v.1	HiSeq (76 bp PE reads)	Yes (Sanger)	21743468

Table 2.4 Published studies identifying *de novo* mutations in schizophrenia parent-proband trios using whole-exome sequencing. The Xu *et al.* and Takata *et al.* studies analysed the trios from the same underlying cohort. After excluding sample duplicates, 1,077 schizophrenia trios were available for analysis.

rates is challenging: genes differ significantly in both total coding length and local sequence context, resulting in substantial differences in their mutability.

A recent study generated gene-specific mutation rates by considering the tri-nucleotide context of each base change, and integrating these locally adjusted rates across an entire gene [138]. The probabilities of each of the 192 possible mutational changes were described as constant values in a mutation rate table. To calculate a gene-specific mutation rate for different types of mutations (LoF, missense, synonymous), the authors determined all possible mutational changes in the gene that would introduce a change of that particular class, and added the tri-nucleotide probabilities of all of these theoretical events. As a robustness check, the study showed that the correlation between the number of rare synonymous variants in each gene and the probability of a synonymous mutation as defined by the mutational rate model was 0.94.

Because of the reliability of this model as demonstrated in its use in previous studies of autism and developmental disorders [105, 118], I chose to incorporate it in our analysis of schizophrenia trios, with a few minor adjustments. First, the gene-specific mutation rates in Samocha *et al.* were calculated based on canonical transcripts as defined by an older version of NCBI RefSeq database (pre-2014), which described fewer protein-coding genes and transcripts per gene than the GENCODE database [123]. Second, the missense mutation rates did not incorporate *in silico* annotations to prioritise more damaging events, and

restricting our analysis to only CADD $\geq$ 15 missense variants further reduces the mutational target of each gene and improves power [130]. To address these limitations, I identified the canonical GENCODE v.19 coding transcript of each gene as defined by the APPRIS annotation pipeline. APPRIS incorporated information from protein structure, functional information, and evolutionary evidence to identify one transcript per gene as the principal functional isoform. In the case of multiple principal transcripts, I conservatively selected the longest APPRIS principal transcript. Gene-specific mutation rates for LoF, missense, and synonymous variants for each GENCODE transcript were computed using the tri-nucleotide mutation rates and method previously described in Samocha *et al*, by adding the probabilities of all theoretical mutational events. I then annotated all possible missense mutations with CADD scores, and calculated a gene-specific mutation rate for missense variants with CADD PHRED score $\geq$ 15.

For genes that existed in both transcript references (RefSeq and GENCODE), our mutation rates based on GENCODE transcripts correlated well with those described in Samocha *et al.*, with a correlation coefficient of 0.97 and 0.98 for missense and LoF mutations respectively (Figure 2.10). Notably, our gene mutation rates were on average greater than the published rates since I conservatively selected for longer transcripts when multiple principal isoforms are available. By using GENCODE over RefSeq, I generated rates for an additional 1,067 protein-coding genes, enabling statistical tests on a more comprehensive set of genes. I also found that only 44% of all possible missense variants had CADD $\geq$ 15, resulting in a substantial reduction in the mutational target for most genes in the genome (Figure 2.11). Interestingly, there was substantial variability in the fraction of CADD damaging sites in different genes: I found that missense damaging sites were nearly completely absent in around $\sim$1,500 genes, while in other genes, more than 75% of all missense sites can be prioritised as damaging. This variability appears to be a property of gene function, since olfactory receptors as a class appear to have the lowest proportion of missense damaging sites. As later shown in Section 4.3.2, these classifier-adjusted rates increased our power to distinguish patterns in *de novo* burden across neurodevelopmental and psychiatric disorders.

2.12 Discussion

Using whole-exome sequence data from the UK10K study, INTERVAL study, Swedish Schizophrenia project, and the SiSU project, I generated a discovery data set of 4,264 schizophrenia cases and 9,343 controls. Despite following standard protocol for alignment and joint calling all samples at the same time, I still observed substantial batch effects from different exome captures used at different time points of the experiment. To address this, I

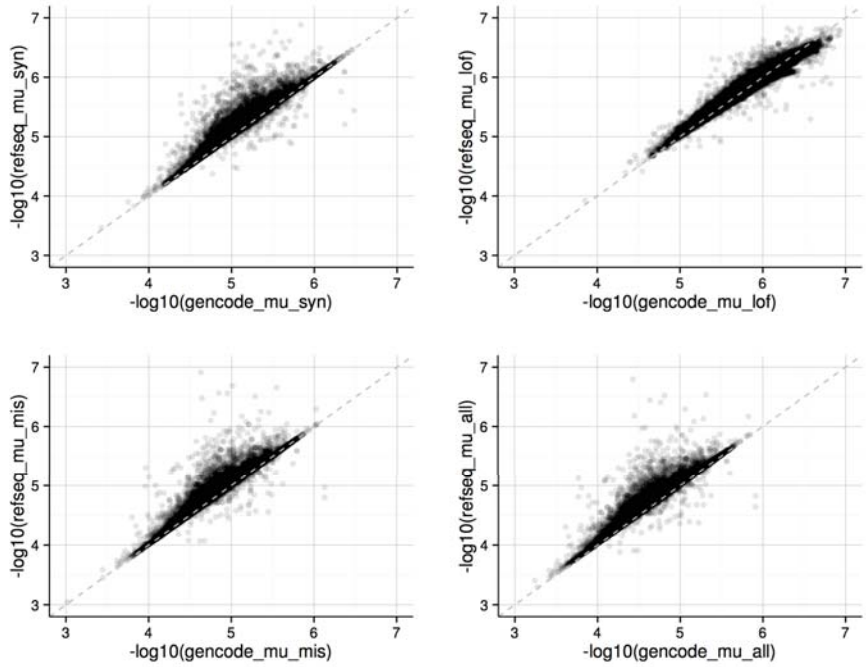


Fig. 2.10 **Correlation between mutation rates generated using GENCODE and RefSeq transcript databases** I compared the LoF, missense, synonymous, and total mutation rates generated using the two different transcript references. Each dot represented a different gene, and mutation rate μ calculated from RefSeq was plotted along the Y-axis, while the rate from GENCODE was plotted along the X-axis.

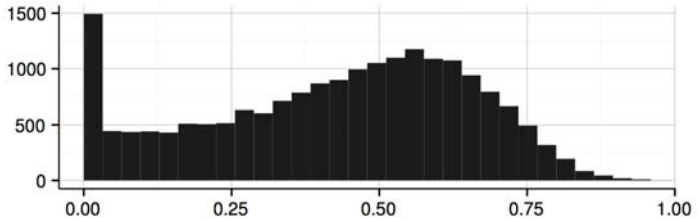


Fig. 2.11 **The ratio of the damaging missense mutation rate to the missense mutation rate of each GENCODE coding gene.** The ratio between missense rate using only CADD damaging sites to the rate from all missense sites was displayed using a histogram. The mean of the bi-modal distribution was 0.44.

restricted our analysis to regions with reasonable coverage in all samples (7× or greater in 80% of each sequencing batch), and then identified appropriate variant- and genotype-level filters using rare inherited and Mendelian inconsistent calls from the DDD study. I found that well-calibrated threshold filters on variant- and genotype-level quality metrics (GQ, DP, and AB) complemented well with a supervised method like GATK VQSR to produce reasonable sensitivity and specificity for rare variant calls. A small number of common coding SNPs was sufficient for sample-level QC aimed at reducing potential biases from ancestry, relatedness, and contamination. Following sample and variant QC, I observed no genome-wide inflation in rare variant tests in subsequent analyses (Section 3.3.1, 3.3.6).

To increase power of collapsing tests of missense variants, I tested the effectiveness of a number of available *in silico* classifiers on a set of *de novo* mutations from the DDD study that were reported back to patients and their families as clinically significant. Ensemble classifiers (LR pred and Radial SVM) performed well when compared to commonly used tools like SIFT and PolyPhen, but a fixed discrimination threshold could not be reliably determined. As a second best option, I decided to annotate missense variants with a CADD score ≥ 15 as damaging, excluding up to 80% of all benign polymorphisms while retaining up to 80% of all diagnostic missense variants. I restricted our subsequent analyses to damaging missense and LoF variants. Lastly, I extended the tri-nucleotide *de novo* model to all canonical GENCODE transcripts, and generated mutation rates for damaging missense variants in addition to all other functional classes. Taken together, the steps highlighted in this Chapter lay the framework for analyses of rare variant data that should also be applicable in future exome sequencing studies.

2.13 Consortia

I would like to acknowledge the following consortia for providing data for the analyses described in this thesis.

2.13.1 UK10K consortium

Richard Anney, Mohammad Ayub, Anthony Bailey, Gillian Baird, Jeff Barrett, Douglas Blackwood, Patrick Bolton, Gerome Breen, David Collier, Paul Cormican, Nick Craddock, Lucy Crooks, Sarah Curran, Petr Danecek, Richard Durbin, Louise Gallagher, Jonathan Green, Hugh Gurling, Richard Holt, Chris Joyce, Ann LeCouteur, Irene Lee, Jouko Lönnqvist, Shane McCarthy, Peter McGuffin, Andrew McIntosh, Andrew McQuillin, Alison Merikangas, Anthony Monaco, Dawn Muddyman, Michael O'Donovan, Michael Owen, Aarno Palotie,

Jeremy Parr, Tiina Paunio, Olli Pietilainen, Karola Rehnström, Tarjinder Singh, David Skuse, Jim Stalker, David St. Clair, Jaana Suvisaari, Hywel Williams

2.13.2 DDD Study

Nadia Akawi, Saeed Al-Turki, Kirsty Ambridge, Jeffrey Barrett, Daniel Barrett, Tanya Bayzetinova, Nigel Carter, Stephen Clayton, Eve Coomber, Helen Firth, Tomas Fitzgerald, David FitzPatrick, Sebastian Gerety, Susan Gribble, Matthew Hurles, Philip Jones, Wendy Jones, Daniel King, Netravathi Krishnappa, Laura Mason, Jeremy McRae, Parker Michael, Anna Middleton, Ray Miller, Katherine Morley, Vijaya Parthiban, Elena Prigmore, Diana Rajan, Alejandro Sifrim, Tarjinder Singh, Adrian Tivery, Margriet van Kogelenberg, Caroline Wright

2.13.3 Swedish Schizophrenia Study

Sarah Bergen, Kimberly Chambert, Menachem Fromer, Christina M. Hultman, Anna K. Kähler, Steve McCarroll, Jennifer L. Moran, Shaun Purcell, Stephan Ripke, Douglas Ruderfer, Edward Scolnick, Pamela Sklar, Patrick F. Sullivan

2.13.4 INTERVAL study

Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and the NIHR Cambridge Biomedical Research Centre (www.cambridge-brc.org.uk). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre.

A complete list of the investigators and contributors to the INTERVAL trial is provided in reference [139], and http://www.intervalstudy.org.uk/about-the-study/whos-involved/interval-contributors/.

2.13.5 Sequencing Initiative Suomi project

The Sequencing Initiative Suomi (SISu) project is an international collaboration between research groups aiming to build tools for genomic medicine. These groups are generating

whole genome and whole exome sequence data from Finnish samples and provide data resources for the research community. Key groups of the project are from Universities of Eastern Finland, Oulu and Helsinki and The Institute for Health and Welfare, Finland, Lund University, The Wellcome Trust Sanger Institute, University of Oxford, The Broad Institute of Harvard and MIT, University of Michigan, Washington University in St. Louis, and University of California, Los Angeles (UCLA). The project is coordinated in the Institute for Molecular Medicine Finland at the University of Helsinki.

Chapter 3

SETD1A is associated with schizophrenia and neurodevelopmental disorders

3.1 Introduction

3.1.1 Motivation behind rare variant analyses in psychiatric disorders

Recent genome-wide association studies have demonstrated that a large proportion of the genetic liability of psychiatric disorders resides in thousands of common alleles each with modest effect [56, 140, 70]. This realization motivated global efforts to aggregate studies with ever-larger sample sizes, and ultimately resulted in the discovery of over 108 common risk loci for schizophrenia [57]. Concurrent analyses similarly suggested that common polygenic variation explained most of the genetic risk in autism spectrum disorders, a condition considered neurodevelopmental in origin. Combining genotyping data sets further showed that schizophrenia, bipolar disorder, and anxiety shared common risk variants, successfully recapitulating the overlap in clinical symptoms across psychiatric disorders [70].

Despite the successes of common variant analyses, studies investigating rare coding variation (minor allele frequency $< 0.1\%$) provide an unique opportunity to extend our understanding of the genetic architecture of psychiatric disorders. First, alleles that confer substantial risk for human disease are expected to reside in the rare end of the allele frequency spectrum. These variants are subject to strong negative selection, and thus, are depleted in the general population. In addition, because coding alleles cause changes at the mRNA and protein level, they are easier to fine-map than common intergenic variants, and are more likely to cause obvious cellular changes in human carriers. Both these properties increase the success and interpretability of subsequent functional studies, which are critical for elucidating the biological mechanisms underlying human disorders. Furthermore, while

ultra-rare variants explain only a modest fraction of the broad-sense heritability of complex disorders, they contribute substantially to individual liability, and are immediately useful in clinical practice for identifying patients with higher risk for disease [141]. At the moment, genetic counselling and genetic testing are limited to fully penetrant alleles for Mendelian traits (e.g. *HTT* repeat length for Huntington's disease or *HBB* allele for sickle cell anaemia) or rare variants of large effect (e.g. *BRCA1/2* alleles for breast cancer or *APOE* alleles for cardiovascular disease), and many more of these clinically relevant variants remain to be discovered. Fortunately, the decreasing costs of whole-exome sequencing has enabled the identification of very rare, often private, protein-coding variants in sufficiently large populations, and well-designed studies leveraging this technology can advance our limited understanding of the rare variant contribution to complex disorders.

3.1.2 Early studies of rare variants in psychiatric disorders

The first results that suggested an important role for rare variants in psychiatric disorders came from karyotyping and cytogenetic studies of large structural variation. These early studies demonstrated that individuals with autism had elevated rates of chromosomal abnormalities, with large rearrangements observed in 5 to 7% of cases [61]. Because many of these risk copy number variant (CNVs) were recurrent, highly penetrant in cases and nearly absent in controls, even very small studies had sufficient power to identify putative risk loci, such as the 15q11.13 duplication in autism [142, 143]. The 22q11.2 deletion was the first structural variant to be significantly implicated in schizophrenia [144], and nearly 24% of carriers had psychiatric symptoms that fulfilled the full diagnostic criteria for schizophrenia [62].

With the arrival of array-based genotyping technologies, these early results were generalized across psychiatric and neurodevelopmental disorders when individuals with schizophrenia, bipolar disorders, and autism were shown to have a greater genome-wide burden of copy number variants compared to controls [63, 145, 146]. In particular, schizophrenia cases had a 3.6-fold enrichment of rare deletions (>500 Kilobases), while between 5 and 10% of individuals with autism carried large structural variants [63, 147]. Family-based studies further identified a 2.3-fold and 5.6-fold excess of *de novo* CNV events were observed in probands with schizophrenia and autism respectively [148, 63, 147]. Follow-up of putative risk loci in many thousands of individuals identified 11 rare recurrent CNVs that individually conferred substantial risk for schizophrenia (ORs $2-60$) [63, 65–67], and an analysis of *de novo* structural variants in 1,124 families identified six risk CNVs for autism [149]. Together, these findings firmly established that rare structural variation contributed to the complex genetic architecture of psychiatric disorders.

However, there is great difficulty in translating these discoveries to an improved understanding of the biological mechanisms underlying schizophrenia. Many of the 11 schizophrenia risk CNVs (e.g.. 22q11.2 deletion) disrupt hundreds of Kilobases and the function of numerous genes; despite thorough functional studies in *in vivo* and *in vitro* systems, the identification of precise genes underlying the relevant psychiatric symptoms remained difficult and time-consuming for most of these loci [150]. On the other hand, whole-exome sequencing has enabled the identification of ultra-rare, disruptive variants at single base resolution, and multiple studies leveraging this technology have shown that many of the observations from analyses of structural variants also extend to this better-resolved class of rare variants. Three studies that whole-exome sequenced $\sim$600 autism parent-proband trios demonstrated that autism cases had an excess of damaging *de novo* SNVs compared to controls, and identified a number of novel risk genes using gene-level rare variant tests (i.e. *ANK2*, *CHD8*, *DYRK1A*, *GRIN2B*, *KATNAL2*, *POGZ*, *SCN2A*) [80, 86, 87]. Schizophrenia individuals also had an excess of rare LoF variants compared to controls [96, 97, 103, 98], but these studies did not have sufficient power to implicate individual risk genes using the reoccurrence of *de novo* mutations or case-control burden.

3.1.3 Emerging results from sequencing studies of neurodevelopmental disorders

Meta-analyses of *de novo* mutations identified in autism probands

The successes of early whole-exome sequencing studies motivated the Autism Sequencing Consortium (ASC) to aggregate even larger sequencing data sets in hopes of identifying additional risk genes. As a follow-up of the smaller trio studies from 2012, the ASC meta-analysed whole-exome sequence data for 2,270 trios, and used a robust mutation rate framework to identify genes with a statistically elevated rate of *de novo* events [105]. The study also compiled a independent cohort of 1,601 cases and 5,397 ancestry-matched controls in which *de novo* mutations could not be identified. Because case-control and *de novo* data appeared to implicate an overlapping set of genes, the authors developed a novel statistical framework that tested for disease association for each gene by combining information from *de novo* mutations, inherited variants, and case-control burden [151]. The model was calibrated such that *de novo* mutations carried the most weight followed by inherited and case-control data. The framework also modelled LoF variants and PolyPhen-damaging missense variants separately but integrated the two sources of information into a single test. Using this more sophisticated hierarchical Bayesian model, the study identified 22 genes at FDR $< 5\%$ and 107 genes at FDR $< 30\%$ in which a disruptive variant conferred substantial risk for autism.

Pathway analyses of these genes implicated synaptic formation, transcriptional regulation and chromatin remodelling as core biological processes in the development of autism. Together, these results suggest that many thousands of exome sequences (over 14,000 in this analysis) are required identify risk genes at genome-wide significance, and that the integration of *de novo* mutations with case-control burden of rare variants can result in a substantive increase in statistical power.

Insights into neurodevelopment from the Deciphering Developmental Disorders study

The same methods and technologies were concurrently applied to study the genetic contributions to developmental disorders. As part of the Deciphering Developmental Disorders (DDD) study, 1,113 children were recruited from regional genetic services across the UK and Ireland with clinical features including intellectual disability (87% of individuals), cranial abnormalities (30%), seizures (24%), and autism (12%) [118]. 1,618 validated *de novo* mutations were identified in this data set, nearly a three-fold excess when compared to expectation in the general population. 317, or 28%, of these children carried a likely pathogenic *de novo* mutation in the DECIPHER DDG2P database, a curated set of 1,129 genes previously demonstrated to carry variants causing developmental disorders. Using gene-specific mutation rates, the study identified 12 new genes associated with developmental disorders. Surprisingly, seven of the ten most significant genes in the ASC meta-analysis (FDR < 0.1%) were also implicated as risk genes for severe developmental disorders. This suggests that autism and broader neurodevelopmental disorders have at least some genetic overlap, and that leveraging this shared genetics may be useful for identifying additional risk genes in future studies.

3.1.4 Goal and aims

Despite the several whole-exome sequencing studies investigating rare variants in schizophrenia, no individual gene had been significantly implicated using rare coding SNVs. Motivated by the new statistical methods and emerging results from the ASC and the DDD study, I aggregated existing family-based and case-control sequencing data sets in schizophrenia, and combined *de novo* recurrence and case-control burden to identify novel risk genes. By meta-analyzing the whole-exome sequences of 4,264 schizophrenia cases, 9,343 controls and 1,077 trios, I hoped to attain sufficient power to identify novel genes that carry alleles conferring substantial risk for schizophrenia.

3.1.5 Publication note and contributions

The results described in this chapter was peer-reviewed and published earlier this year [119]. I briefly summarise the various contributions to this project. The sources of the data used were provided in Chapter 2. I performed all the production, and QC steps for these data, and designed the statistical approach to integrate the case-control data with *de novo* mutation recurrence. Other than the DDD proband phenotypic similarity analysis and *SETD1A* splice reporter assay, I performed all the analysis described in this Chapter, as well as generated all the Figures and Tables. Olli Pietiläinen, Moira Blyth, Trevor Cole, Shelagh Joss, David Collier, and Mandy Johnstone kindly provided phenotypic details for the UK10K, DDD, and SiSU *SETD1A* carriers. I wrote the first draft of the manuscript, and received very helpful corrections, comments, and suggestions from my supervisor Jeffrey C. Barrett. The manuscript was further improved after receiving useful comments from Dave Curtis, Patrick Sullivan, Michael Owen, Michael O'Donovan. Unless explicitly stated, the parts of the peer-reviewed publication reproduced in this chapter are my original work.

3.2 Materials and methods

3.2.1 Gene-based analysis in the case-control data set

A description of the study collections that compose of the schizophrenia case-control data set was provided in Section 2.2.1. There, I also highlighted the key steps taken to align, call, and prepare the sequence data. In total, rare variants from 4,264 cases and 9,343 controls, and *de novo* mutations from 1,077 trios were available for analysis. To identify genes with a significant burden of rare, damaging variants, I first applied the Fisher's exact test as implemented in PLINK/SEQ [104, 152]. I collapsed all rare variants identified in the coding region of each gene as defined by GENCODE v.19, and tested for an excess of LoF variants and LoF combined with damaging missense variants in cases compared to controls. Because I analysed only variants with MAF < 0.5%, the probability of an individual carrying more than one LoF or damaging missense variant was low. Therefore, I coded an individual as 1 if they carry a rare allele in a gene, and 0 otherwise. I applied the Fisher's exact test at three different minor allele frequency (MAF) thresholds (singletons, $\geq 0.1\%$ and $\geq 0.5\%$), as was performed in previous rare variant analyses of schizophrenia [103]. To evaluate significance, I performed two million case-control permutations within each population (UK, Finnish, and Swedish) to control for ancestry and batch-specific differences.

I also tried to replicate previous results which found a polygenic burden of rare variants in schizophrenia cases compared to controls. The approach used for gene set enrichment

analysis broadly followed the methodology described in Purcell *et al.* and implemented in PLINK/SEQ and the SMP utility [103]. This method of gene set enrichment testing featured more prominently and is elaborated further in Section 4.2.2. Briefly, the gene set enrichment statistic was calculated as the sum of single gene burden test-statistics corrected for exome-wide differences between cases and controls. Statistical significance was determined using two million case-control permutations as described above. The reported odds ratios and confidence intervals from the enrichment analyses were calculated from raw counts without taking into account ancestry and batch-specific differences in cases and controls.

3.2.2 Meta-analysis of *de novo* mutations and case-control burden

3.2.3 Frequentist method of meta-analysis using Fisher's method

I aggregated validated *de novo* mutations identified in 1,077 schizophrenia trios from seven published studies for analysis with our case-control cohort. Recurrence of *de novo* mutations was modelled as the Poisson probability of observing N or more *de novo* variants in a gene given a baseline gene-specific mutation rate obtained from the method described in Samocha *et al.*, modified to produce LoF and damaging missense rates for each canonical GENCODE v.19 gene (see Section 2.11) [138, 123]:

$$X \sim Pois(2N_t\mu)$$
$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} P(X = i)$$

where N_t was the number of schizophrenia trios in our analysis (1,077), X was number of observed *de novo* mutations within the trio data set, and μ was the gene-specific mutation rate. A one-sided Fisher's exact test (described above, in Section 3.2.1) was used to model the difference in rare LoF (MAF $< 0.1\%$) burden between cases and controls. Previous case-control whole-exome sequencing studies similarly used one-sided tests for gene discovery [103, 105]. In particular, Purcell *et al.* suggested that the one-sided test was appropriate since current case-control studies for schizophrenia would not have sufficient power to detect rare protective alleles, and that prior work on the burden of copy number variants and *de novo* mutations suggested a predominantly one-sided model in which rare alleles increase risk for disease. However, any significant result I report would remain significant regardless of a one-sided or two-sided model. Subsequently, *de novo* and case-control burden P-values

were meta-analysed using Fisher's combined probability method:

$$X^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$$

where p_i was the P-value for the ith test, $k = 2$ was the number of tests being combined, and X^2 followed a χ-square distribution with $2k = 4$ degrees of freedom. To calculate an odds ratio for LoF variants in each gene, we treated the 1,077 probands as additional cases in our case-control data set. For the schizophrenia discovery data set, the per-gene odds ratios were calculated from observed LoF variants in 5,341 cases and 9,343 controls. Because the number of observed LoF variants in each gene were often quite small, the odds ratio calculation was corrected using penalized maximum likelihood logistic regression model (Firth's method, implemented in the logistf R package).

3.2.4 Bayesian modeling of *de novo* and case-control variants using TADA

In addition to the frequentist method of meta-analysis, I also applied the Transmission and Disequilibrium Association (TADA) method as described in He *et al.* [151] and implemented in De Rubeis *et al.* [105]. TADA is a hierarchical Bayesian statistical method for the joint analysis of case-control and family studies, in which information from the recurrence of *de novo* mutations was integrated with inherited and case-control burden in a single statistical test. Briefly, variants in N_t trios were classified as *de novo*, transmitted, or non-transmitted, and all variants of each category were collapsed to a single count per gene. Counts of *de novo* mutations were modelled using a Poisson distribution with two exogenous parameters: μ, the gene-specific mutation rate for the specific variant class, and γ, the relative risk of disease-associated variants. Case-control counts were similarly modelled using a Poisson distribution, but instead of μ, the rate parameter depended on the general frequency of rare variants in the population (q), scaled by sample size. A Bayesian approach was used to test if a gene conferred disease risk, with the null and alternate hypotheses defined as $H_0 : \gamma = 1$ and $H_1 : \gamma > 1$ respectively. Within this framework, different relative risk parameters γ were used to model LoF and missense variants, and *de novo* and case-control variants. These γ parameters were crucial for weighting the importance of different types of information when the joint statistic was computed. Generally, $\bar{\gamma}_d > \bar{\gamma}$ and $\bar{\gamma}_{\text{LoF}} > \bar{\gamma}_{\text{mis}}$. Finally, per-gene Bayes factors were calculated for LoF and missense variants separately, and then combined.

The robustness of results from TADA depended heavily on the specification of its hyperparameters, which were dependent on the (unknown) genetic architecture of the trait.

These include the relative risks for *de novo* and case-control variants (parametrized by $\bar{\gamma}_d$ and $\bar{\gamma}$), and the number of true risk genes in schizophrenia (k). To apply TADA, I needed to first define hyperparameters that reasonably represent schizophrenia's true genetic architecture. However, the estimation of $\bar{\gamma}$ required the identification of a small set of true risk genes, but no risk genes have yet been discovered in schizophrenia. Using estimates from the autism analysis would be incorrect, since autism has a greater excess of *de novo* LoF and missense mutations than schizophrenia. To use TADA in a robust manner, I ran the model across a range of reasonable parameters to determine if any signal appeared significant throughout:

- $\bar{\gamma}_d \in \{2,4,6,8,10,12,15,20\}$ for LoF variants
- $\bar{\gamma} \in \{1,2,4\}$ for LoF inherited and case-control variants
- $\bar{\gamma}_d \in \{1,2,4\}$ for missense variants
- $\bar{\gamma} = 1$ for missense inherited and case-control variants
- $k \in \{100,500,1000,2000\}$

I used the default values for the remaining parameters, and applied the following restrictions: $\bar{\gamma}_d > \bar{\gamma}$ and $\bar{\gamma}_{\text{LoF}} > \bar{\gamma}_{\text{mis}}$.

3.2.5 Validation of variants of interest

The experimental validation of individual variants of interest was performed by Elena Prigmore of the DDD study. Primers were designed using Primer 3 to produce products between 400 and 600 bp in length centred on the site of interest. Using genomic DNA from all trio members as templates, PCR reactions were carried out using Thermo-Start Taq DNA Polymerase (Thermo Scientific) following the manufacturer's protocol, and successful PCR products were capillary sequenced. Traces from all trio members were aligned, viewed, and scored for the presence or absence of the variant.

3.2.6 Functional consequence of the exon 16 splice acceptor deletion

The functional assay described here was performed by Sebastian S. Gerety of the DDD study to assess the impact of the exon 16 splice acceptor site variant. A custom minigene construct was first created by cloning the entire 696 bp genomic region encompassing exons 15, 16, 17 and intervening introns of human *SETD1A*, fused in-frame to a C-terminal GFP. We flanked the cassette with a strong upstream promoter and a downstream polyadenylation sequence. We transfected plasmids containing either the reference or deletion-containing forms into HELA cells, and these cells were grown for 2 days under standard conditions. The RNA was

extracted (RNEasy, Qiagen) from the transfected cells and cDNA was synthesized (Superscript III, Invitrogen). Minigene-specific primers were designed to avoid amplification of endogenous HELA derived transcripts. The first pair of primers spanned all three exons, thus allowing us to detect overall splicing changes (Pair 1, Forward 2: TCGAAGAGTCATAAACACTGCCATG, Reverse 9: GTGAACAGCTCCTCGCCCTTG). We also designed pairs of exonic, intron-spanning primers to distinguish splicing events upstream (Pair 2, Forward 1: TTTGCAGGATCCCATCGAAGAGTC, exon 16 reverse: CACTGTCCATGATGGCGGAGGTA) and downstream (Pair 3, exon16 forward: CTGCTGAGCGCCATCGGTAC, exon17 reverse: CTGAACTTGTGGCCGTTTACGTC) of exon 16. We performed PCR on the cDNA from two transfection replicates of each sample. Agarose gels were used identify PCR product size differences (DNA ladder: 2-log ladder, New England Biolabs), which were further analysed by capillary sequencing.

3.2.7 Phenotype clustering in DDD probands

The phenotypic clustering analysis of DDD probands was performed by Jeremy McRae. Clinical geneticists as part of the DDD study systematically recorded phenotypes of probands with severe developmental disorders using the Human Phenotype Ontology (HPO) [153]. The Human Phenotype Ontology version 2013-11-30 was used to record phenotypes of these individuals. We leveraged this systematic phenotypic data to assess the probability that the probands shared more similar clinical features than expected by chance. For each pair of terms, we calculated the information content (defined as the negative logarithm of the probability of the terms' usage within 4,295 DDD probands) for the most informative common ancestor. We estimated the similarity of HPO terms between two individuals as the maximum information content (maxIC) from pairwise comparisons of the HPO terms for the two individuals. We then estimated the phenotype similarity for a set of N probands as the sum of all the pairwise maxIC scores. A null distribution of similarity scores was simulated from randomly sampled sets of N DDD probands, and the P-value was calculated as the proportion of simulated scores greater than or equal to the observed score.

3.3 Results

3.3.1 Study design

The case-control data set consisted of 357,088 damaging missense and 55,955 LoF variants called in 4,264 cases and 9,343 controls (Figure 3.1). I restricted our analyses to rare variants, stratified by allele frequency (singletons, $< 0.1\%$, and $< 0.5\%$) and function (LoF and

damaging missense variants). I first replicated the enrichment of rare LoF variants in the previously implicated set of 2,456 genes [103] in our UK and Finnish schizophrenia data sets ($P = 7 \times 10^{-4}$). Having confirmed that rare disruptive variants spread among many genes are associated with schizophrenia risk, I tested for an excess of disruptive variants within each of 18,271 genes in cases compared to controls using the Fisher's exact test. Despite our large sample size, the per-gene statistics followed a null distribution in all tests, and I was unable to implicate any gene via case-control burden of disruptive variants (Figures 3.2, 3.3).

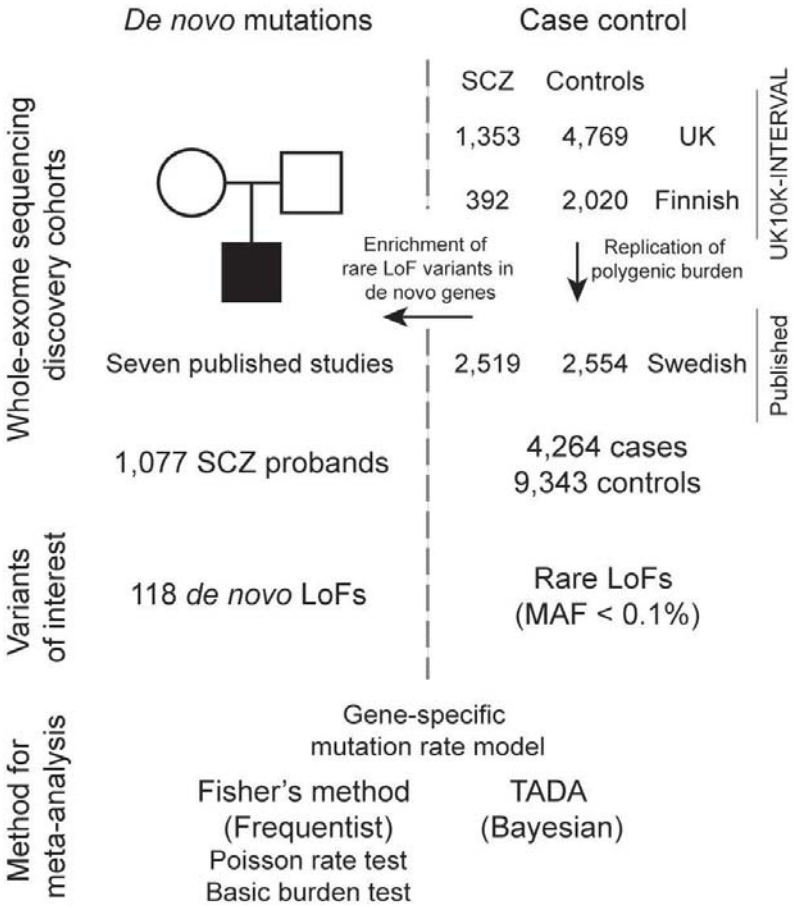


Fig. 3.1 **Study design for the schizophrenia exome meta-analysis.** The source of sequencing data, sample sizes, variant classes, and analytical methods are described. Details on case-control samples are shown on the right, while parent-proband trios are described on the left.

3.3.2 LoF variants in *SETD1A* are associated with schizophrenia

To determine whether the integration of *de novo* mutations with case-control burden might succeed in discovering risk genes in schizophrenia, I aggregated, processed, and re-annotated

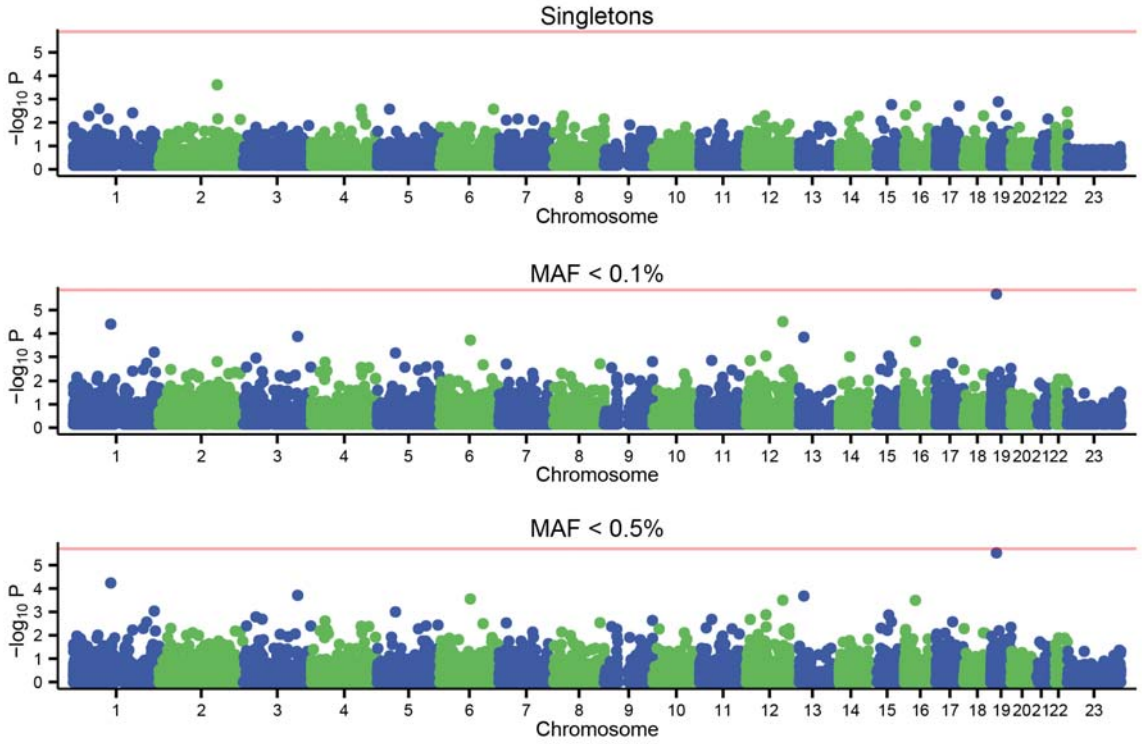


Fig. 3.2 **Manhattan plot of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.** I tested for an excess of LoF variants within 18,271 genes using Fisher's exact test. $-\log_{10}$ P-values were plotted against the chromosomal location (mid-point) of each gene. I showed results from three allele frequency thresholds (singletons, $< 0.1\%$ and $< 0.5\%$) for aggregating rare variants. No gene exceeded the exome-wide significant threshold of $P = 1.25 \times 10^{-6}$ (red line).

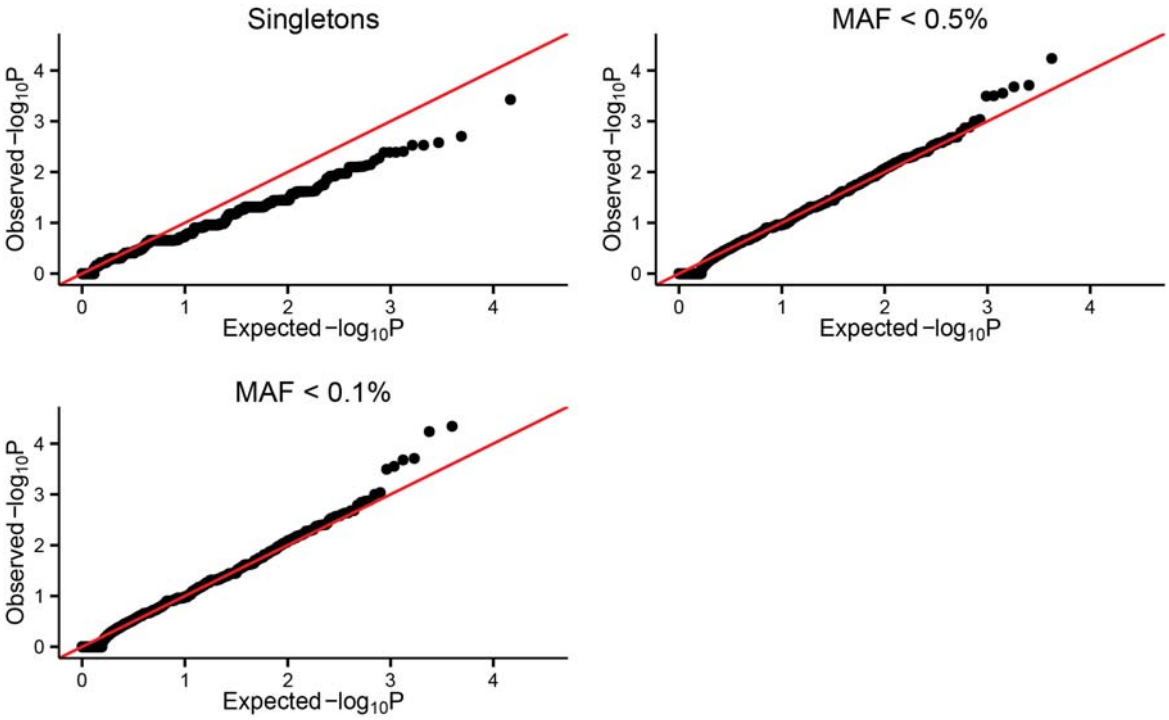


Fig. 3.3 **QQ plots of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.** I tested for an excess of LoF variants within 18,271 genes using Fisher's exact test, and plotted the ordered $-\log_{10}$ *P*-values against transformed *P*-values sampled from the uniform distribution. The QQ plots for gene burden tests with minor allele frequency cut-offs of 0.1% and 0.5% followed an expected null distribution. The QQ plot for the burden test of singleton variants still showed deflation because the per-gene counts are too low and the data does not meet the asymptotic requirements of the statistical test. I included *P*-values from informative tests in which genes have at least one case LoF count.

de novo mutations in 1,077 schizophrenia probands from seven published studies, and found 118 LoF and 662 missense variants [98, 99, 95, 97, 100–102]. Thirty-eight genes had two or more *de novo* nonsynonymous mutations, two of which (*SETD1A* and *TAF13*) had been previously suggested as candidate schizophrenia genes [98, 99]. I found that the 754 genes with *de novo* mutations were significantly enriched in rare LoF variants in cases compared to controls from our main dataset. In these 754 genes, the most significant case-control enrichment across allele frequency thresholds and functional class was for the test of LoF variants with MAF $< 0.1\%$ ($P = 2.1 \times 10^{-4}$; OR 1.08, $1.02 - 1.14$, 95% CI), which I focused on for subsequent analysis.

Motivated by this overlap of genes with *de novo* mutations and excess case-control burden, I meta-analysed *de novo* variants in the 1,077 published schizophrenia trios with rare LoF variants (MAF < 0.1%) in 4,264 cases and 9,343 controls. I used two analytical approaches, one based on Fisher's method to combine *de novo* and case-control *P*-values, and the other using the transmission and *de novo* association (TADA) model to integrate *de novo*, transmitted, and case-control variation using a hierarchical Bayesian framework [105, 151] (Figure 3.1). I focused on results that were significant in both analyses, and which did not depend on the choice of parameters in TADA (Figure 3.8). In both methods, loss-of-function mutations in a single gene, *SETD1A*, were significantly associated with schizophrenia risk (Table 3.2, Fisher's combined $P = 3.3 \times 10^{-9}$). I observed three *de novo* mutations and seven case LoF variants in our discovery cohort, and none in our controls (Figure 3.6). In one of the seven case carriers, direct genotyping in parents confirmed that the LoF variant (c.518-2A>G) was a *de novo* event, but genotypes were not available for the other parents. I looked for additional *SETD1A* LoF variants in unpublished whole exomes from 2,435 unrelated schizophrenia cases and 3,685 controls [154], but none were identified (Table 3.2). Thus, in more than 20,000 exomes, I observed ten case and zero control LoF variants (corrected OR 35.2, $4.5 - 4528$, 95% CI). Although the confidence intervals are wide, rare LoF variants in *SETD1A* conferred substantial risk for schizophrenia. No other gene approached genome-wide significance (Table 3.1, Figures 3.4, 3.5).

3.3.3 Robustness of the *SETD1A* association

Previous large sequencing analyses such as the Swedish schizophrenia, DDD and NHLBI myocardial infarction studies [103, 118, 94] had defined genome-wide significance for gene burden tests using a Bonferroni correction for the number of genes and the number of functional and frequency cut-offs tested. For example, $P < 1.25 \times 10^{-6}$ is 0.05 corrected for 20,000 genes tested for nonsynonymous and LoF variants, and a further correction for two frequency thresholds would require the even more stringent cut-off of $P < 6.25 \times 10^{-7}$). For

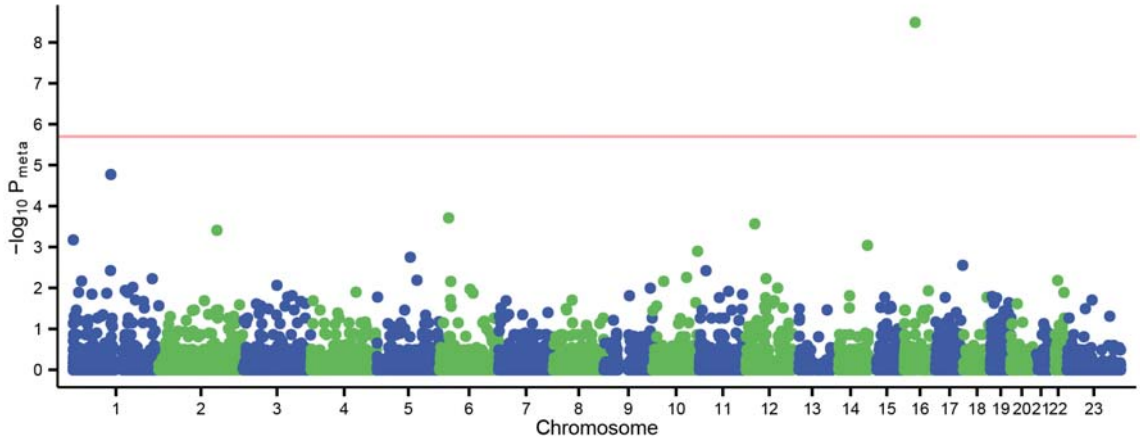


Fig. 3.4 **Manhattan plot of the meta-analysis of *de novo* mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.** *De novo* and case-control burden *P*-values were meta-analysed using Fisher's combined probability method. $-\log_{10}$ *P*-values were plotted against the chromosomal location (mid-point) of each gene. A total of 18,271 genes were tested. Only *SETD1A* exceeded exome-wide significance, with $P = 3.3 \times 10^{-9}$. Red line: $P = 1.25 \times 10^{-6}$.

Gene name	μ_{LoF}	$N_{de\ novo}$	N_{case}	N_{control}	$P_{de\ novo}$	P_{burden}	P_{meta}
SETD1A	6.6e-06	3	7	0	4.6e-07	0.0003	3.3e-09
TAF13	1.3e-06	2	1	0	3.7e-06	0.31	1.7e-05
HIST1H1E	2.4e-07	1	3	0	0.00053	0.031	0.00019
BCAT1	1.9e-06	1	8	3	0.004	0.0058	0.00027
XIRP2	3.3e-06	0	41	35	1	3.5e-05	0.00039
KLHL17	3e-06	1	4	0	0.0065	0.0096	0.00067
HSP90AA1	3.1e-06	1	5	1	0.0066	0.013	0.00091
MKI67	1e-05	2	5	10	0.00024	0.53	0.0013
CAST	3.1e-06	0	15	6	1	0.00019	0.0018
ENDOV	2.2e-06	0	10	2	1	0.00031	0.0028

Table 3.1 Meta-analysis results for 1,077 trios, 4,264 cases and 9,343 controls. Only *SETD1A* reached exome-wide significance.

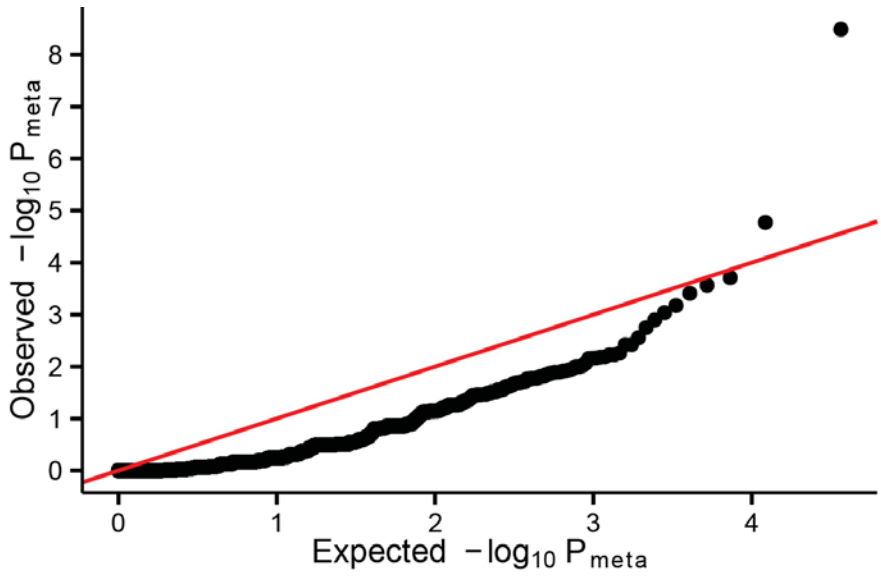


Fig. 3.5 **QQ plot of the meta-analysis of *de novo* mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.** *De novo* and case-control burden *P*-values were meta-analysed using Fisher's combined probability method, and the $\log_{10}$ *P*-values plotted against transformed *P*-values sampled from the uniform distribution. Because only a subset of genes had *de novo* LoF variants, Fisher's method deflated the combined *P*-value of genes without any *de novo* information.

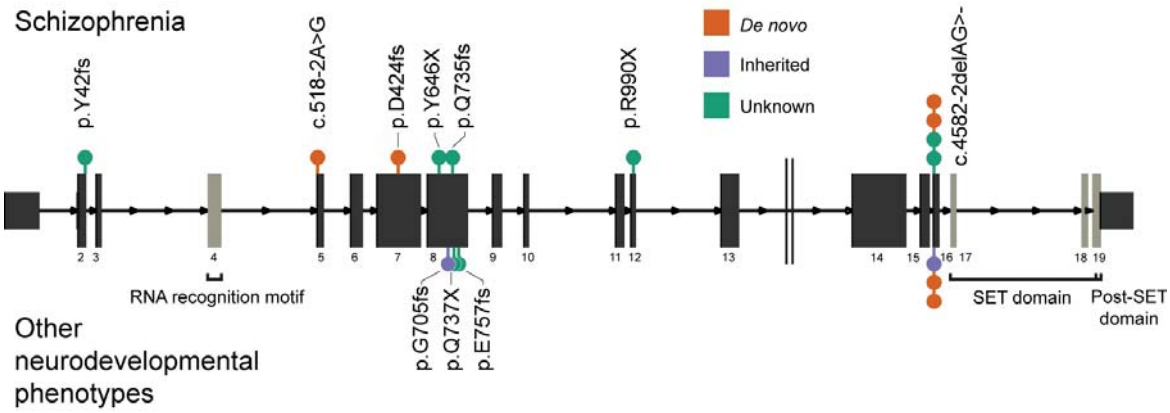


Fig. 3.6 **The genomic position and coding consequences of 16 *SETD1A* LoF variants observed in the schizophrenia exome meta-analysis, the DDD study, and the SiSU project.** Variants discovered in patients with schizophrenia are plotted above the gene, and those discovered in individuals with other neurodevelopmental disorders (from DDD and SiSU) are plotted below. Each variant is coloured according to its mode of inheritance. All LoF variants appear before the conserved SET domain, which is responsible for catalysing methylation. Seven LoF variants occur at the same two-base deletion at the exon 16 splice acceptor (c.4582-2delAG>-).

Phenotype	Data set	*De novo*	Case	Control	Test	*P* value
Schizophrenia	UK10K-INTERVAL		2 of 1,353	0 of 4,769		
	UK10K Finnish		2 of 392	0 of 2,020		
	Swedish (published)		3 of 2,519	0 of 2,554		
	All case-control		7 of 4,264	0 of 9,343	Fisher's exact[a]	0.0003
	Schizophrenia parent-proband trios	3 of 1,077			Poisson exact[b]	4.6×10^{-7}
	Case-control + *de novo* (discovery)	3 of 1,077	7 of 4,264	0 of 9,343	Fisher's combined[c]	3.3×10^{-9}
	Swedish (replication)		0 of 2,435	0 of 3,685		
	All schizophrenia samples	3 of 1,077	7 of 6,699	0 of 13,028	Fisher's combined[c]	5.6×10^{-9}
Other neurodevelopmental phenotypes	DDD study	2 of 4,281	2 of 4,281	See note[d]	Fisher's combined[c]	0.003
	ASD trios	0 of 2,297				
	ID trios	0 of 151				
Combined	All samples	5 of 7,806	9 of 10,980	0 of 13,028	Fisher's combined[c]	3.2×10^{-8}

Table 3.2 **Results from statistical tests associating disruptive variants in *SETD1A* to schizophrenia and developmental delay.** None of these tests incorporated exomes from the ExAC database. The number of *SETD1A* LoF variants and the sample size of each dataset are indicated in each cell. The statistical tests were performed as follows: *a*: a one-sided burden test of case-control LoF variants using Fisher's exact test, *b*: the Poisson probability of observing *N de novo* variants in *SETD1A* given a calibrated baseline gene-specific mutation rate, *c*: meta-analysis of de novo and case-control burden *P*-values using Fisher's combined probability test, *d*: the INTERVAL dataset (n = 4,769) were used as matched controls.

these thresholds to control false positives, however, the test being used must produce well-calibrated *P*-values. This had been shown to be true for standard approaches in a case-control setting, such as the basic burden test, Fisher's exact test, and the sequence kernel association test (SKAT), as long as the cases and controls were well-matched and residual differences are corrected for [103, 94]. On the other hand, parent-proband trio studies used a Poisson or Binomial model parametrised by gene-specific mutation rates and the discovery sample size to test for an elevated rate of *de novo* mutations. While this approach was powerful, it was less robust than the approaches described above. *De novo* test statistics were highly sensitive to the specification of gene-specific mutation rates, which were well-established for SNVs but not small indels. Furthermore, the low counts in *de novo* studies made results sensitive to the size of the discovery dataset.

Depletion of *SETD1A* LoF variants in the ExAC database

I performed five analyses to ensure our *SETD1A* association was robust to possible confounders of rare variant association testing. First, to validate our observation of the rarity of disruptive variants in *SETD1A* in unaffected individuals, I examined the Exome Aggregation Consortium (ExAC) v0.3 for the LoF variants in *SETD1A* [112]. All exomes in ExAC were joint-called using the GATK v3.2 pipeline, and included other public exome datasets, such as the 1000 Genomes Project and NHLBI-GO Exome Sequencing Project, with additional quality control compared to their original releases. In 60,706 unrelated exomes, I observed seven LoF variants in *SETD1A*. Since the v0.3 release aggregated studies of psychiatric disor-

ders including the Swedish schizophrenia study, I excluded all samples from these data sets, leaving only four LoF variants in 45,376 exomes without a known neuropsychiatric diagnosis. I next applied the same stringent QC metrics used in our analysis to ExAC data. I found that the 16:30976302-GC/G indel observed in two individuals was located at the same position as a high-quality SNP, and occurred at a homopolymer run of cytosines. At the genotype level, both calls had a genotype quality (GQ) Phred probability of < 40, far lower than used in our study in which I required indels to have a GQ > 90. In addition, the variant has poor allelic balance (AB < 0.15), and the BAM alignment reflected these low-quality metrics [112]. Given this evidence, I excluded the putative indel. Two high-quality *SETD1A* LoF variants in 45,376 unaffected ExAC exomes remained. Following the approach in Samocha *et al.*, I determined the significance of the depletion of *SETD1A* LoF variants in ExAC using a signed Z-score of the χ-squared deviation between observed and expected counts [138]. I scaled the expected LoF counts provided by ExAC (43 in 60,706) to 45,376 exomes (expected 32.5), and calculated the one-tailed *P*-value of the signed Z-score assuming two observed LoF variants. Observing only two LoF variants when expecting 32.5 variants represented a substantial depletion compared to chance expectation ($P = 4.4 \times 10^{-8}$). According to its pLI score, a measure of constraint relative to other coding genes calculated using the ExAC data, *SETD1A* is among the 3% most constrained genes in the human genome [112]; LoF variants in *SETD1A* were almost totally absent in the general population.

Dependence of results on specification of mutation rate

Second, four of the ten *SETD1A* carriers with schizophrenia had the same two-base deletion at the exon 16 splice acceptor (c.4582-2delAG>-), at least two of which occurred as *de novo* mutations (Figure 3.6). Since this variant underpinned the statistical significance of our observation, I investigated it further in several ways. First, to rule out sequencing artefacts, I confirmed a clean call where I had access to the raw sequencing reads (n = 2), and noted that both published *de novo* mutations at this position had been validated with Sanger sequencing [99, 101]. Second, our model, and therefore the test statistic that I report, was dependent on a gene-specific mutation rate. To address the possibility that the recurrent mutation occurred at a hypermutable site (and thus our model was not well calibrated), I determined that our observations would be exome-wide significant ($P < 1.25 \times 10^{-6}$) even if the mutation rate at this position were up to ten-fold higher (7×10^{-5}) than the cumulative LoF rate for all other positions in *SETD1A* (6.6×10^{-6}). If the two-base deletion mutation rate were truly this high (e.g. greater than 99.99% of all per-gene LoF mutation rates), I would expect to find 6.4 observations in 45,376 non-schizophrenia exomes in ExAC, but I observed only 1 (Fisher's exact test $P = 0.013$).

Functional assay evaluating the function of recurrent deletion of the exon 16 splice acceptor

Third, we used a minigene construct to show that this two-base deletion resulted in the retention of the upstream intron. As expected, strong GFP expression was detected from the reference sequence construct. This suggested correct splicing occurred between exons, leading to in-frame GFP translation. The mutant form displayed dramatically weaker GFP expression. mRNA was extracted from the transfected cells, and PCR reactions spanning all three exons revealed an increased transcript size in the mutant form compared to reference (Figure 3.7a). A PCR reaction spanning just the first 2 exons (15/16) revealed a similar shift in size, suggesting that the splice site deletion/mutation was causing intron retention between exons 15 and 16 (Figure 3.7b). Sanger sequencing of the PCR products confirmed this aberrant splicing outcome (Figure 3.7c). The predicted translation product would therefore include translation of exon 15, the subsequent intron, and out-of-frame translation of exon 16, resulting in a premature stop within this exon. The downstream splicing event to exon 17 was not affected. These data indicated that in a human *in vitro* system, the recurrent indel we observe in probands resulted in a premature stop codon and a truncated *SETD1A* protein.

Independence of results on parameterization in TADA

Fourth, to ensure our results were robust when applying TADA, I generated Bayes factor across a set of reasonable hyperparameters, and the results largely agreed with those obtained from the Fisher's combined probability method: only one gene, *SETD1A*, had reached genome-wide significance (Figure 3.8). I found that the most influential parameters were $\bar{\gamma}_d$ (mean relative risk of *de novo* LoF variants), $\bar{\gamma}$ (mean relative risk for case-control LoF variants), and π_0 (fraction of true risk genes). While holding these parameters constant, the Bayes factors did not vary to any appreciable degree across the remaining hyperparameters. I found that our signal in *SETD1A* had a q-value < 0.01 as long as $\gamma_d > 4$, $\gamma > 4$, and $k > 100$. If I assumed a greater mean relative risk for LoF variants in *SETD1A* ($\bar{\gamma} > 8$ and $\gamma > 8$) as expected for strong risk alleles in a constrained gene, *SETD1A* was exome-wide significant for any reasonable specification of k. No other gene has q-value < 0.01 under any tested parametrization, including the parametrization used in the previous autism meta-analysis (Table 3.3). Thus, the *SETD1A* result from the Bayesian analyses were robust at all reasonable specifications of the model's hyperparameters.

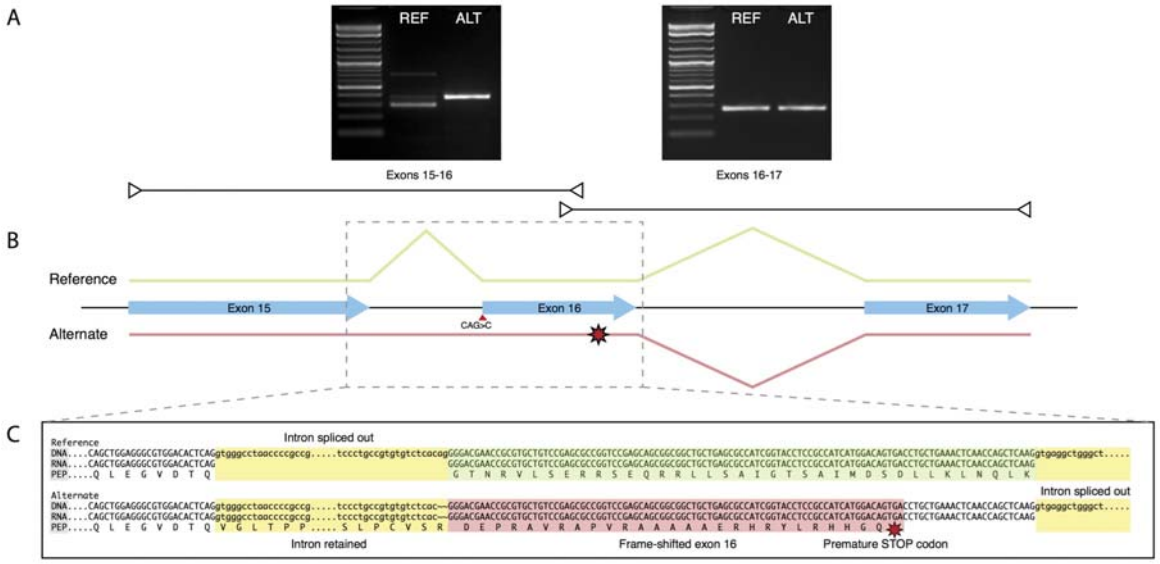


Fig. 3.7 **Results from the minigene experiment assessing the impact of the exon 16 splice acceptor site variant.** This figure and the data contained within were generated and provided by Sebastian S. Garety. **A.** Minigene constructs driving expression of exons 15, 16 (Ref and Alt), and 17 fused to GFP were transfected into HELA cells. RT-PCR analysis of cell lysates using primer pair 2, spanning exons 15, 16, and the intervening intron revealed a change in size of PCR products suggesting retention of the intervening intron in the construct containing the splice-acceptor deletion (panel A, Exons 15-16, REF versus ALT). PCRs with primer pair 3, spanning the intron downstream of exon 16 showed no change in band sizes (panel A, Exons 16-17, REF versus ALT), suggesting this intron was correctly spliced out in both reference and alternate forms. **B.** Depiction of genomic locus surrounding the exon 16 splice acceptor deletion. The predicted structure of reference (green) and deletion containing (red) transcripts were shown above and below genomic map. The red star indicated a predicted premature stop codon due to intron retention and resulting frame-shifted translation. **C.** Results from capillary sequencing of PCR products from panel A confirmed intron retention in the splice acceptor deletion construct (panel C, RNA, yellow box). This resulted in a predicted frame-shifted translation of exon 16 (panel C, PEP, red box), and a premature truncation of the protein 28 amino acids into exon 16 (red star). Downstream intron splicing was confirmed by capillary sequencing to be intact in both constructs.

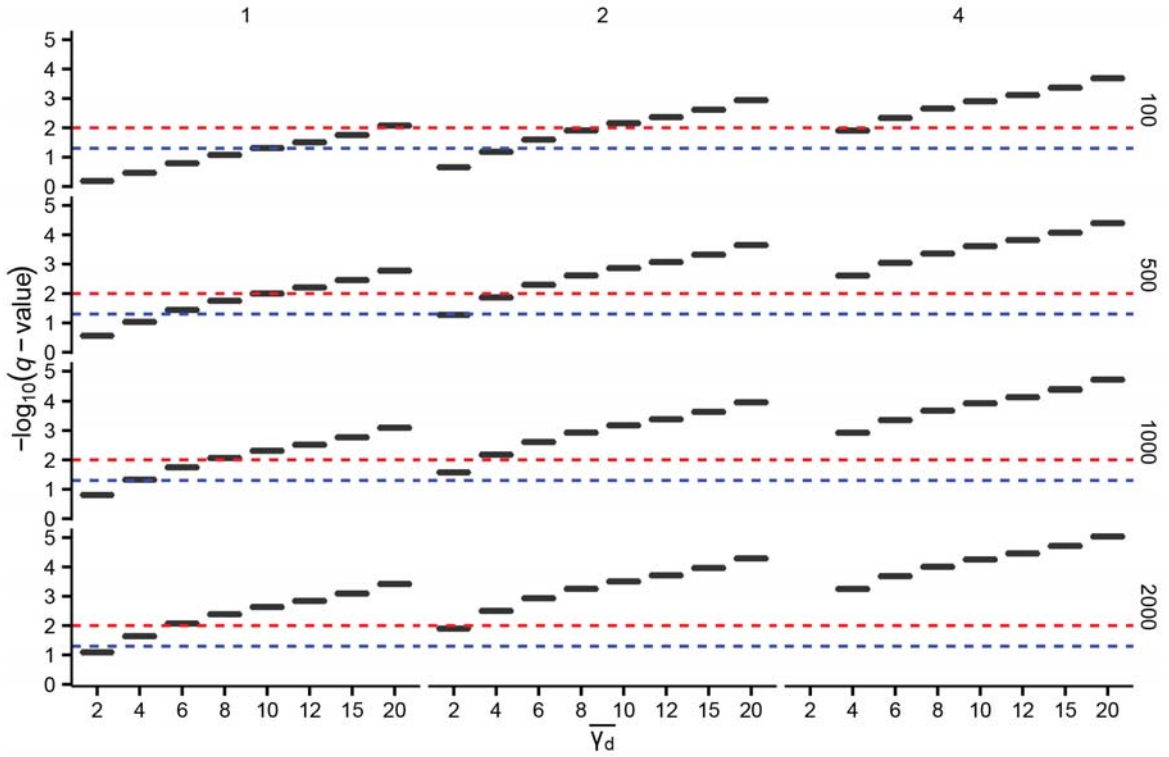


Fig. 3.8 **The robustness of the *SETD1A* result across reasonable parameters in the TADA model.** Because the TADA model depended heavily on the specification of its hyperparameters, I calculated the log q-value of *SETD1A* across different mean relative risk of *de novo* variants ($\bar{\gamma}_d$), mean relative risk of case-control variants ($\bar{\gamma}$), and numbers of true schizophrenia risk genes (k). Each vertical column is a different value for $\bar{\gamma}$, and each horizontal facet is a different value for k. Our signal in *SETD1A* had a q-value < 0.01 as long as $\gamma_d > 4$, $\gamma > 4$, and $k > 100$. Blue line: $P = 0.05$; red line: $P = 0.01$.

Gene name	DNM LoF	Case LoF	Ctrl LoF	DNM Mis15	Case Mis	Ctrl Mis	BF	q-value
SETD1A	3	7	0	0	24	50	2e+05	7.7e-05
XIRP2	0	41	35	0	81	145	7.4e+02	0.01
TAF13	2	1	0	0	4	9	5.9e+02	0.016
HSPA8	1	0	1	2	5	12	2.7e+02	0.025
BCAT1	1	8	3	0	10	25	2.7e+02	0.031
CAST	0	15	6	0	33	50	1.6e+02	0.041
NIPAL3	1	2	1	1	8	22	1.3e+02	0.05
HSP90AA1	1	5	1	0	20	48	1.2e+02	0.059
SSBP3	1	1	0	1	4	11	1.1e+02	0.066
KLHL17	1	4	0	0	28	58	1e+02	0.073
MKI67	2	5	10	0	27	75	1e+02	0.078
SLC25A24	0	14	6	0	8	22	92	0.084
PIK3C2B	1	3	2	1	54	83	89	0.089
DPYD	1	6	7	1	34	92	88	0.093
HIST1H1E	1	3	0	0	14	22	77	0.099
IGSF22	0	13	5	0	23	73	69	0.1
RYR3	0	11	6	2	120	242	68	0.11
ENDOV	0	10	2	0	5	10	66	0.11
LPHN2	1	2	3	1	28	49	45	0.12
PHF7	1	0	0	1	4	16	43	0.13
ORC3	0	16	10	0	25	41	42	0.14
BLNK	1	2	0	0	6	19	40	0.14
URB2	1	12	13	0	20	36	37	0.15
ZEB1	1	2	0	0	21	37	36	0.15
NUP214	1	4	2	0	74	146	33	0.16
CRYBG3	1	1	4	1	20	48	31	0.17
BTNL2	1	2	1	0	3	7	31	0.17
INHBC	1	2	1	0	9	18	30	0.18
POGZ	1	2	0	0	29	44	29	0.19
STAC2	0	3	3	2	13	30	29	0.19
DLG2	1	4	3	0	22	58	28	0.2
PRRC2A	1	3	1	0	10	37	27	0.2
ST3GAL6	1	1	0	0	7	15	27	0.21
KRT15	0	4	0	1	18	28	27	0.21
RB1CC1	1	3	2	0	24	39	23	0.22
ZDHHC5	1	1	0	0	29	59	23	0.22
SMARCC2	1	3	2	0	19	38	23	0.23
OR2T2	0	12	7	0	0	1	23	0.23
ATG12	1	2	2	0	6	24	22	0.24
XPR1	1	1	0	0	5	22	22	0.24
AOX1	0	9	6	1	36	81	22	0.25
CDKL1	0	8	2	0	10	23	21	0.25
SPDYC	0	5	2	1	17	21	21	0.25
RECK	0	7	4	1	21	52	20	0.26
RTTN	1	9	10	0	42	82	19	0.26
XIRP1	0	13	8	0	27	88	18	0.27
SLC12A7	0	9	3	0	37	55	18	0.27
SYNGAP1	1	1	0	0	20	25	18	0.28
SCLT1	0	7	1	0	8	11	18	0.28
EPHA2	1	6	7	0	43	84	18	0.28
PYCARD	0	3	0	1	1	6	17	0.29
GTPBP3	1	1	1	0	16	23	17	0.29
SHANK1	1	1	0	0	10	15	17	0.29
KDM5C	1	1	0	0	3	6	17	0.3

Table 3.3 TADA results using the hyperparameters in the De Rubeis *et al.* autism meta-analysis. Only *SETD1A* has a q-value < 0.01.

Phenotype	Data set	Case	Control	Test	*P* value
Schizophrenia	All schizophrenia case-control samples (ignoring *de novo* status)	10 of 7,776	0 of 13,028		
	Non-schizophrenia ExAC exomes		2 of 45,376		
	All samples	10 of 7,776	2 of 58,404	Fisher's exact	2.6×10^{-8}
Neurodevelopmental disorders	DDD study	4 of 4,281	See note[a]	Fisher's exact	2.9×10^{-4}
	ASD trios	0 of 2,297			
	ID trios	0 of 151			
Combined	All samples	14 of 14,505	2 of 58,404	Fisher's exact	1.2×10^{-8}

Table 3.4 **Burden tests associating disruptive variants in *SETD1A* to schizophrenia and developmental delay.** *De novo* status of variants was ignored and non-schizophrenia exomes from the ExAC database were incorporated as controls. The number of *SETD1A* LoF variants and the sample size of each dataset were indicated in each cell. *a*: the full control dataset (n = 58,404) was used to calculate the *P*-value.

Burden testing with non-psychiatric ExAC exomes as additional controls

Finally, to demonstrate that our result was significant independent of mutation rate specification, I ignored the *de novo* status of variants in our discovery and replication datasets, creating a combined dataset of 7,776 cases and 13,028 controls. I then included unaffected ExAC exomes as additional controls, and observed ten LoF variants in 7,776 cases and two LoF variants in 58,404 controls. Using a basic test of case-control burden (Table 3.4), I found that LoF variants in *SETD1A* were significantly associated with schizophrenia (Fisher's exact test: $P = 2.6 \times 10^{-8}$; OR 37.6, $8.0 - 353$, 95% CI). This result was driven by ten very rare variants in our schizophrenia cases: six were observed in only one individual each, and the seventh, the two-base recurrent deletion at the exon 16 splice acceptor (c.4582-2delAG>-), was observed in four individuals. Two of the four recurrent indels were *de novo*, and the other two were found in unrelated individuals of different ancestry (one from Sweden and one from the UK). Similarly, of the two LoF variants in ExAC, one was observed in only one individual and the other was the recurrent indel in an individual of African ancestry. Thus, our burden test of very rare variants in *SETD1A* would not be confounded by systematic differences between sub-populations in the ExAC exomes and our dataset. Taken together, these five analyses excluded many possible artefacts, and provided confidence in our conclusion that LoF variants in *SETD1A* conferred substantial risk for schizophrenia.

3.3.4 *SETD1A* is associated with severe developmental disorders

All heterozygous carriers of *SETD1A* LoF variants satisfied the full diagnostic criteria for schizophrenia, including classic positive symptoms such as hallucinations, prominent disorganization, and paranoid delusions (Table 3.5). Six of these individuals were male and four were female. Eight patients had evidence of chronic illness, requiring long-term input from psychiatric services. Notably, of the seven *SETD1A* LoF carriers for whom

Variant	Data set	Mode	Clinical features	Intellectual functioning
16:30970178_T/T GATG frameshift	UK10K-Finns	Case	Psychotic episodes with hallucinations and prominent disorganization, requiring psychiatric hospitalization. Chronic illness with deterioration.	Probable mild intellectual disability. Completed compulsory education, but repeated several grades.
16:30974752_A/G splice acceptor	UK10K-Finns	*De novo*	Disorganized schizophrenia with severe positive and negative symptoms with hallucinations, delusions and aggression. Chronic, severe symptoms requiring long psychiatric hospitalization. Early onset at age 10. Has mild facial dysmorphology.	Severe learning difficulties, diagnosed with minimal brain damage, abnormal EEG; mild mental retardation. Unable to complete compulsory education. Developmental delay.
16:30976334_AC/A frameshift	Takata *et al.*[13]	*De novo*	Psychotic with persecutory delusions and thought disorder in addition to obsessional thoughts, compulsive behaviors and rituals. Persistent negative symptoms, disorganized behavior and delusional thinking. First psychotic break at age 21. As a child (age <10 years), displayed social isolation, excessive fears, inattentiveness, learning difficulties and obsessive-compulsive disorder–like rituals. Moderately deteriorating course.	Learning difficulties noted as a child. Delayed milestones. School performance declined from age 16. Worked as security officer.
16:30977140_C/G stop gained	UK10K	Case	Chronic hallucinations and delusions, partially controlled by depot medication.	Minor problems with memory or understanding. No secondary school diploma.
16:30977405_CAG/C frameshift	Swedish	Case	Two brief admissions, no record of antipsychotic treatment. No immediate family history of psychiatric disorders.	No information on intellectual functioning or educational attainment.
16:30980962_C/T stop gained	Swedish	Case	Multiple hospitalizations, with 8 years of antipsychotic medication. No immediate family history of psychiatric disorders.	No information on intellectual functioning or educational attainment.
16:30992057_CAG/C splice acceptor	UK10K	Case	Breech delivery. Epilepsy with seizures from ages 2 to 18. Socially isolated and dependent on parents till age 40, when presented with bizarre somatic delusions, paranoid delusions and auditory hallucinations including running commentary. Developed negative symptoms alongside ongoing psychotic symptoms and required long-term institutional care. Symptoms were persistent and unresponsive to antipsychotic medication.	Borderline intelligence. Attended mainstream school and left age 17 without a secondary school diploma. Worked as warehouseman.
16:30992057_CAG/C splice acceptor	Swedish	Case	Multiple hospitalizations, with 8 years of antipsychotic medication. No immediate family history of psychiatric disorders.	No information on intellectual functioning or educational attainment.
16:30992057_CAG/C splice acceptor	Takata *et al.*[13]	*De novo*	Developed schizophrenia aged 18 with delusions, disorganized behavior, poor motivation, flattened affect and social isolation. Compulsive behaviors since 4th grade. Since first episode of psychosis, did not return to previous level of functioning.	Finished high school, but slow learner and inattentive. Delayed developmental milestones.
16:30992057_CAG/C splice acceptor	Guiponni *et al.*[20]	*De novo*	Undifferentiated schizophrenia.	Developmental delay.

Table 3.5 **Phenotypes of individuals in the schizophrenia exome meta-analysis who carry LoF variants in *SETD1A*.** For each individual, I provide the genomic coordinates of the variant, its mode of inheritance, and the study from which each patient was first recruited. "Clinical features" describes notable neuropsychiatric or neurodevelopmental symptoms in each individual, and "Intellectual functioning" provides additional information on reported cognitive phenotypes.

any information on intellectual functioning was available, one was noted to have severe learning difficulties while the six appeared to have mild to moderate learning difficulties. Four patients were noted to have achieved developmental milestones with clinically salient delays (Table 3.5). I was unable to confirm if the three Swedish carriers had any form of cognitive impairment. This was consistent with previous reports that individuals with autism or schizophrenia who have *de novo* LoF mutations have a higher rate of cognitive impairment [98, 105].

To investigate whether *SETD1A* might play a role in other neurodevelopmental disorders, I looked for *de novo* LoF mutations in *SETD1A* in 3,581 published trios with autism, severe developmental disorders, or intellectual disability [105, 118, 85, 84], but found none. I

Variant	Data set	Mode	Clinical features	Intellectual functioning
16:30977316_G/GC frameshift	DDD	Maternally inherited	Capillary hemangiomas, abnormality of the eyebrow, broad nasal tip, wide mouth, thick lower lip vermilion, short philtrum, overgrowth, renal duplication. 5.29 years old.	Delayed speech and language development.
16:30992057_CAG/C splice acceptor	DDD	Maternally inherited	Infantile axial hypotonia, delayed gross motor development, midfrontal capillary hemangioma. 0.55 years old.	Not detailed due to age
16:30992057_CAG/C splice acceptor	DDD	*De novo*	Mild global developmental delay, hypertelorism, wide nasal bridge, hydrocele testis. 3.14 years old.	Aggressive behavior, autoaggression. First words spoken between 2 to 2.5 years of age.
16:30992057_CAG/C splice acceptor	DDD	*De novo*	Global developmental delay, macrocephaly, nevus flammeus of the forehead, wide and flat nose, mandibular prognathia, hypopigmentation of the skin, wide intermammillary distance, truncal obesity. Has breath-holding attacks and night terrors. 6.09 years old.	Delayed speech and language development.
16:30977411_C/T stop gained	NFID	Case	Short stature, mild facial morphology, EEG abnormalities, delusional disorder, has psychosis.	Mental retardation
16:30977473_G/GC frameshift	NFBC	Case	Epilepsy during childhood (grand mal status epilepticus), diagnosed with personality disorder.	Not detailed

Table 3.6 **Phenotypes of individuals in the DDD study and SiSU project who carry LoF variants in *SETD1A***. For each individual, I provide the genomic coordinates of the variant, its mode of inheritance, and the study from which each patient was first recruited. "Clinical features" describes notable neuropsychiatric or neurodevelopmental symptoms in each individual, and "Intellectual functioning" provides additional information on reported cognitive phenotypes. NFID: Northern Finnish Intellectual Disability study; NFBC: Northern Finnish Birth Cohort.

next turned to an additional 3,148 children with diverse, severe, developmental disorders recruited as part of the DDD study, and discovered four probands with LoF variants in *SETD1A* (Table 3.6). Three of these occurred at the recurrent exon 16 splice junction indel described above (two *de novo*, one maternally inherited), and the fourth was a maternally inherited frameshift insertion (Figure 3.6). We validated all four LoF variants using Sanger sequencing. All four probands have developmental delay with additional phenotypes that cluster within the larger DDD study using the HPO clustering analysis (empirical $P = 0.042$). I additionally observed a *de novo* CNV deleting 650 Kilobases encompassing *SETD1A* (chr16:30,964,376−31,614,891, Figure 3.9) in a DDD proband. CNV calling and quality control in the DDD study was described in a previous publication [118], and the call was supported by signal from 156 probes. The proband had global developmental delay, absent speech, motor delay, sleep disturbance, developmental regression, feeding difficulties in infancy, and generalized myoclonic seizures. *SETD1A* did not reach exome-wide significance as a developmental disorder gene within the DDD study alone ($P = 3.0 \times 10^{-3}$), but when I jointly analysed all samples using the frequentist meta-analysis approach, the association was clear to both severe developmental disorders and schizophrenia ($P = 3.1 \times 10^{-8}$, Table 3.2). Because all of the DDD *SETD1A* carriers were under 12 years old at recruitment and as schizophrenia rarely manifests at this age [28], it remains unknown if these individuals will develop schizophrenia.

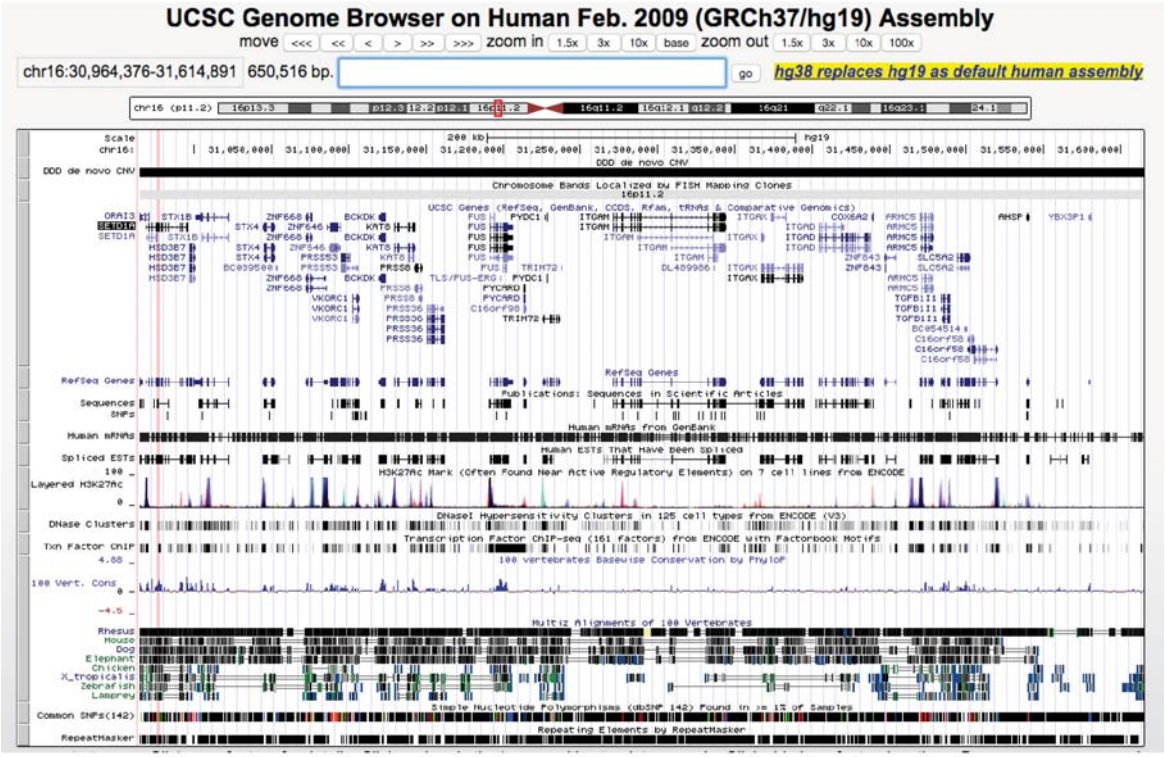


Fig. 3.9 ***De novo* microdeletion of a single copy of *SETD1A* identified in the DDD study.** A proband was identified to have a 650 kb deletion encompassed *SETD1A* and 29 other genes. The figure showing the deletion was generated using the UCSC Genome Browser (https://genome.ucsc.edu/).

In 5,720 unrelated Finnish individuals exome sequenced as part of the Sequencing Initiative Suomi project, I identified two additional heterozygous LoF variants in *SETD1A*. One individual with a stop-gain variant was recruited as part of the Northern Finnish Intellectual Disability cohort with a diagnosis of mental retardation, short stature, mild facial dysmorphology, and EEG abnormalities (Table 3.6). Notably, this individual was also diagnosed with delusional disorder and unspecified psychosis at 15 years of age. The second *SETD1A* LoF carrier belonged to the Northern Finnish 1966 Birth Cohort (NFBC), a representative, geographically based population cohort. This individual had epileptic episodes at 7 years of age, and was diagnosed with an unspecified personality disorder by a psychiatrist. Thus, in an additional search for *SETD1A* LoF carriers, only two were found, both in individuals affected by neuropsychiatric disorders.

3.3.5 Power calculations to show co-morbid cognitive impairment in schizophrenia *SETD1A* carriers

While I found an association between *SETD1A* and schizophrenia and developmental disorders, I was unable to demonstrate whether LoF variants in this gene specifically decreased cognitive ability in individuals with schizophrenia. I performed a power calculation to determine the sample size required to show additional cognitive impairment in *SETD1A* LoF carriers with schizophrenia. I assumed that pre-morbid IQ in individuals diagnosed with schizophrenia followed a Gaussian distribution with mean μ_0 and standard deviation σ. I further assumed that the distribution of pre-morbid IQ in carriers of *SETD1A* LoF variants was also Gaussian, shared the same standard deviation σ, but had a shifted mean μ_1. To calculate the sample size needed to show that μ_0 and μ_1 were statistically different, I performed power calculations using a one-sided t-test of means with a range of parameters for the effect size and frequency of *SETD1A* LoF variants.

I defined the following:

- N = sample size (individuals diagnosed with schizophrenia)
- $d = \frac{|\mu_0 - \mu_1|}{\sigma}$, or the effect size (in s.d. units) of *SETD1A* LoF variants on pre-morbid IQ
- $\alpha = 0.05$, Type I error probability
- p = frequency of LoF variants in *SETD1A* in schizophrenia cases

Figure 3.10 showed power to detect this effect across the following parameter combinations:

- $N \in \{5000, 10000, \ldots 100000\}$

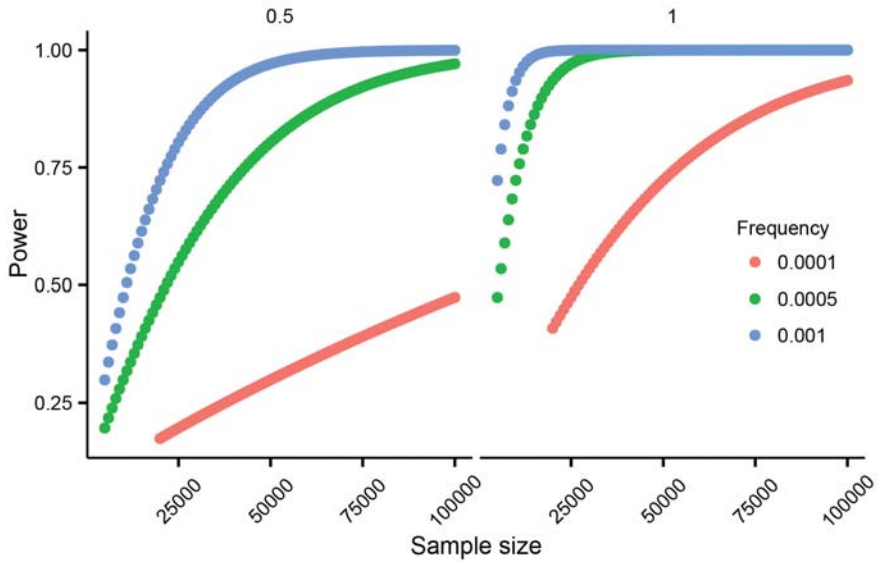


Fig. 3.10 **Sample size curves for detecting an increased risk of pre-morbid cognitive impairment in schizophrenia *SETD1A* LoF carriers.** I performed power calculations using a simple one-sided *t*-test to identify sample sizes required to show possible cognitive impairment in *SETD1A* schizophrenia carriers. Effect sizes *d* (0.5, 1), and allele frequencies (0.0001, 0.0005, 0.001) are varied to show their influence on statistical power. I assumed a Type I error probability of 0.05. For these effect sizes and frequencies, a sample of tens of thousands of cases would be needed.

- $d \in \{0.5, 1\}$, or $\mu_1 = \mu_0 - \sigma \times d$
- $p \in \{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$

Assuming a modest effect on cognition ($d = 0.5$) and that only one in 10,000 schizophrenia patients carried a LoF variant in *SETD1A*, a sample size of over 100,000 individuals would be required for 50% power to detect the effect on cognition. If this effect was greater ($d = 1$) and the true frequency was similar to the 0.1% observed in our study, a sample size of over 10,000 individuals would have $> 50\%$ power.

3.3.6 *De novo* burden in neurodevelopmental disorders

Even though our study had an overall sample size comparable to recent ASD and DD studies that identified 7 ASD genes and 32 DD genes [105, 118], I was only able to implicate a single schizophrenia gene at genome-wide significance. To investigate this further, I aggregated and analysed *de novo* mutations from four different studies: 1,113 probands with developmental disorders [118], 2,297 ASD probands [105], and 566 control probands [155, 80]. Using this data set, I compared the rates of *de novo* events in each group relative to baseline exome-wide mutation rates. Briefly, *de novo* mutations (x_d) in each neurodevelopmental condition were modelled as $x_d \sim \text{Pois}(2N_t\mu_G)$, where N_t is the number of trios, μ_G is the genome-wide mutation rate for a particular functional class, and x_d is the observed number of *de novo* mutations in N_t trios. The genome-wide mutation rate of each variant class was calculated as the sum of all gene-specific mutation rates in Samocha *et al.* [138] ($\mu_{\text{syn}} = 0.137$, $\mu_{\text{damaging mis}} = 0.165$, $\mu_{\text{LoF}} = 0.043$). I modelled *de novo* mutations in control trios to ensure that the genome-wide mutation rates were well calibrated. I reported the probability of observing x_d or more mutations in N_t trios given the genome-wide mutation rate, and used the Poisson exact test to determine if pairwise differences in *de novo* rates existed between control, schizophrenia, autism, and developmental disorder trios. I reported the two-sided *P*-values and rate ratios, and Bonferroni correction was used to adjust for multiple testing.

The rates of *de novo* mutations across damaging missense and LoF variants were significantly higher in DD than in ASD, and higher in ASD than in schizophrenia (Figure 3.11). Indeed, the rate of damaging missense variants in schizophrenia was not different from baseline rates ($P = 0.45$) and only nominally higher than in controls ($P = 0.029$), and the rates of LoF variants were only slightly elevated ($P = 5.7 \times 10^{-3}$). In ASD, by contrast, missense ($P = 9.4 \times 10^{-10}$) and LoF ($P = 3.7 \times 10^{-15}$) rates were significantly greater than expectation. In developmental disorders, the rates were even higher (missense: $P = 2.5 \times 10^{-17}$; LoF: $P = 1.3 \times 10^{-31}$) (Figure 3.11). Across all genes in the genome, the rate of disruptive

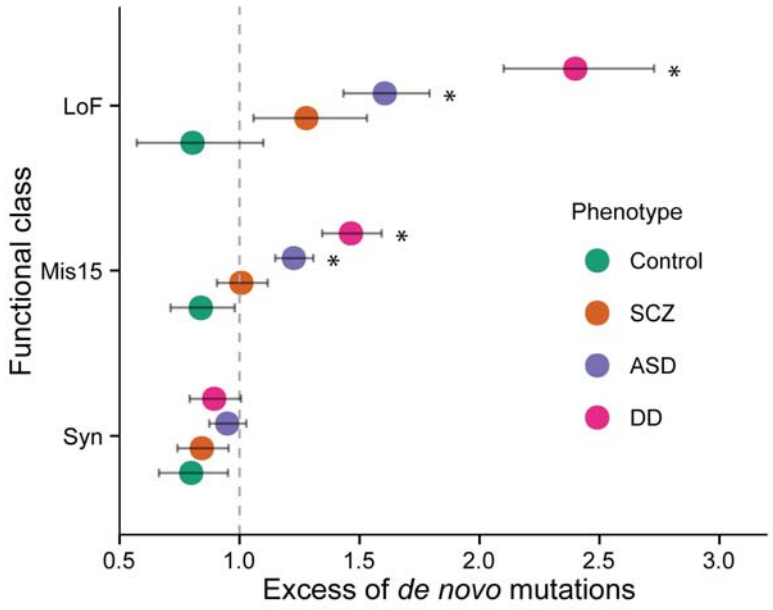


Fig. 3.11 **A comparison of genome-wide *de novo* mutation rates in probands with autism, developmental disorders, schizophrenia, and controls.** Rates were modelled using calibrated genome-wide mutation rates. Significant excess of *de novo* mutations when compared to the baseline model was marked with an asterisk ($P < 4 \times 10^{-3}$, Bonferroni correction for 12 tests). Nominal significance could be inferred from the error bars (95% CI).

de novo variants differed dramatically across these disorders. Because the recurrence of *de novo* mutations is a particularly powerful way to identify risk genes, the weak excess of *de novo* variants in schizophrenia provides at least a partial explanation for the limited success of this strategy to date in identifying genes for this disorder.

3.4 Discussion

In one of the largest exome-sequencing studies of complex disease to date, I identified an association between rare LoF variants in *SETD1A* and risk of schizophrenia and other severe neurodevelopmental phenotypes. A previous report [99] suggested *SETD1A* as a candidate schizophrenia gene based on two of the *de novo* mutations included in our analysis. Our study establishes the *SETD1A* association at a significance exceeding a Bonferroni corrected P-value of 1.25×10^{-6} independent of any specification of gene mutation rate. Indeed, in keeping with observations in other neurodevelopmental disorder sequencing studies, even larger meta-analyses of schizophrenia exomes will be required to define the phenotypic spectrum of *SETD1A* LoF variant carriers, and to identify new risk genes.

SETD1A, also known as *KMT2F*, encodes one of the methyltransferases that catalyse the methylation of lysine residues in histone H3. Loss-of-function variants in at least five other genes within this family result in dominant Mendelian disorders characterized by severe developmental phenotypes including intellectual disability [156]. These include Wiedemann-Steiner syndrome (*KMT2A*), Kleefstra syndrome (*EHMT1*), and Kabuki syndrome (*KMT2D*) (Figure 3.12). Moreover, rare *de novo* LoF mutations and copy number variants in *KMT2C*, *KMT2E*, *KDM5B*, and *KDM6B* have been recently associated with autism risk [109]. The developmental and cognitive phenotypes of *SETD1A* carriers are consistent with these other Mendelian conditions of epigenetic machinery; however, among all genes associated with developmental disorders and intellectual disability, *SETD1A* is the first shown to definitively predispose to schizophrenia, offering insights into the biological differences underlying these conditions [118, 157]. As with other risk genes for severe neurodevelopmental phenotypes, it is possible that an allelic series of LoF variants exists in *SETD1A*, where different variants increase risk for different clinical features. However, seven of the 16 LoF variant carriers (Figure 3.12) have the same two base deletion at the splice acceptor of exon-16 (c.4582-2delAG>-): four in individuals with schizophrenia and three in individuals diagnosed with other developmental disorders. Thus, the same variant is associated with both schizophrenia and developmental disorders.

Detailed phenotypes from the DDD and SISu studies suggest that *SETD1A* carriers may have distinctive features, including delayed speech and language development, epilepsy,

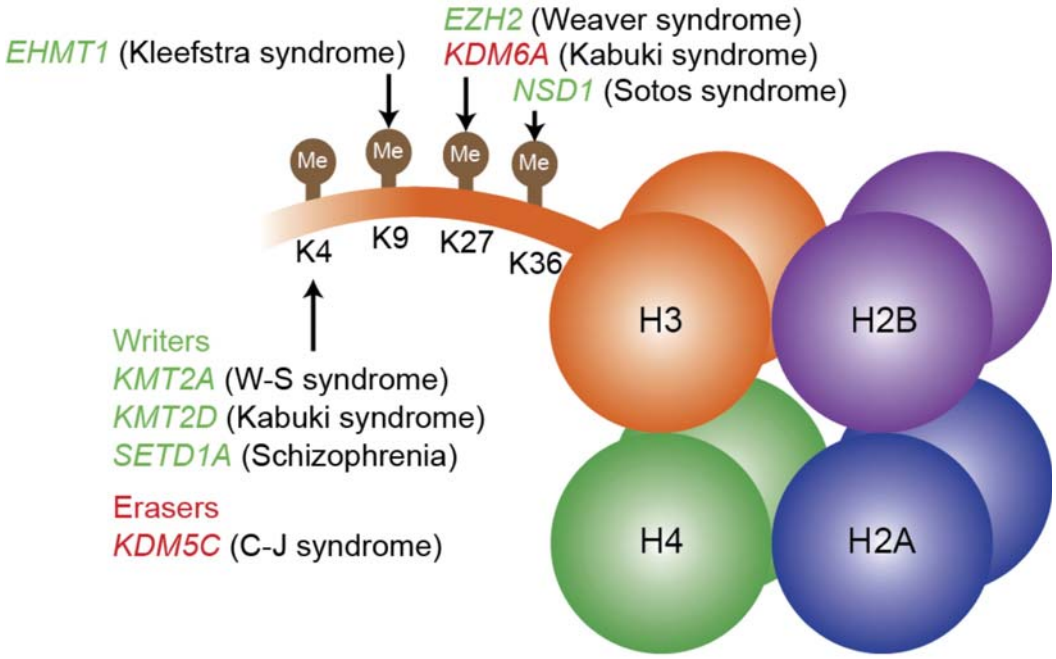


Fig. 3.12 **Mendelian disorders of epigenetic machinery at histone H3.** Writers (in green) add methyl groups at the specified residue of the histone tail, while erasers (in red) perform targeted demethylation. Disrupting variants in writers and erasers described in the figure result in well-known examples of dominant, highly penetrant disorders characterised by developmental delay and intellectual disability. Only the tail of histone H3 and its four key lysine residues are illustrated here. Alternate nomenclature: *EHMT1* (also known as *KMT1D*), *EZH2* (*KMT6A*), *NSD1* (*KMT3B*), *SETD1A* (*KMT2F*).

personality disorder, and facial dysmorphology (Table 3.6). While cognitive and developmental phenotypes in schizophrenia patients are sparser, four individuals had delayed developmental milestones, one is noted as having mild facial dysmorphology and minimal brain damage, and another had epileptic seizures during childhood (Table 3.5). However, impairment of cognitive function is now generally regarded, along with positive and negative symptoms, as an integral feature of schizophrenia rather than a co-morbidity, and our study, as designed, cannot address whether variants in *SETD1A* are specifically associated with the cognitive features of the disorder. Indeed, it would require a re-sequencing study with detailed cognitive measurements on tens of thousands of patients (Figure 3.10) to decisively answer this question.

The clinical heterogeneity observed in carriers of *SETD1A* LoF variants is reminiscent of at least 11 large copy number variant syndromes (one of which, 16p11.2 is nearby, but not overlapping *SETD1A*), which cause schizophrenia in addition to many other developmental defects [67, 158]. A canonical example is the 22q11.2 deletion syndrome, which is characterised by schizophrenia in 22.6% of adult carriers [159], highly variable intellectual impairment [160], and numerous severe neurological and physical defects [161]. A considerably larger cohort (such as the hundreds of cases of 22q11.2 deletion syndrome studied to date) will be needed to accurately estimate the relative penetrance of *SETD1A* LoF variants for schizophrenia, developmental disorders, and other clinical features.

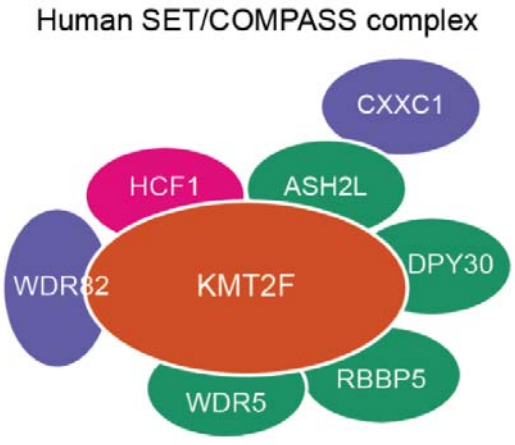


Fig. 3.13 **SET1/COMPASS complex** A highly conserved protein complex that methylates the tail of histone H3. *SETD1A* or *KMT2F* is one of the catalytic cores of this complex.

While disruptions of *SETD1A* are very rare events and occur in only a small fraction of schizophrenia cases (0.13% in our meta-analysis; 0.062% − 0.24% 95% CI), several lines of evidence suggest that histone H3 methylation is more broadly relevant to schizophrenia.

The H3K4 methylation gene ontology category (GO:51568) showed the strongest statistical enrichment among 4,939 biological pathways in GWAS data of psychiatric disorders [108]. This category contains 20 genes, including *SETD1A* and six others (*ASH2L*, *CXXC1*, *RBBP5*, *WDR5*, *DPY30*, and *WDR82*) [162–164] that together form the SET1/COMPASS complex, through which *SETD1A* regulates transcription by targeted methylation (Figure 3.13). Indeed, two of the genes in GO:51568 (*WDR82* and *KMT2E*) are near genome-wide significant associations to schizophrenia [57]. A previous study of *de novo* CNVs in schizophrenia trios identified one deletion and one duplication overlapping *EHMT1*, another histone methyltransferase [66] implicated in developmental delay, and a range of congenital abnormalities [164]. While no gene in the H3K4 category reached exome-wide significance, we observed a *de novo* mutation and one case LoF variant in *KMT2D*, one case LoF variant in *KMT2A* and *KMT2B*, and two case LoF variants in *KMT2C* and *KMT2E*. These highly constrained genes were in the same methyltransferase family as *SETD1A*, and in which LoF variants also caused severe developmental disorders. Finally, conserved H3K4me3 peaks identified in pre-frontal cortical neurons co-localise with genes related to biological mechanisms in schizophrenia including glutamatergic and dopaminergic signalling [165]. Our implication of *SETD1A* therefore contributes to the growing body of evidence that chromatin modification, specifically histone H3 methylation, is an important mechanism in the pathogenesis of schizophrenia.

Chapter 4

Schizophrenia risk genes are shared with neurodevelopmental disorders

4.1 Introduction

4.1.1 Early evidence for a neurodevelopmental etiology to schizophrenia

While the precise causes of schizophrenia remain unknown, the neurodevelopmental hypothesis postulates that certain genetic or environmental insults early in brain development ultimately manifest in adolescence and adulthood. Since its formulation by Weinberger, Murray and Lewis in 1987 [166, 167], evidence from clinical, epidemiological, imaging, and genetic studies has emerged to support this model of schizophrenia pathogenesis. First, through CT, MRI, and histochemistry staining techniques, neuroimaging studies identified gross brain abnormalities in schizophrenia patients prior to and at the onset of illness, including structural differences in the dorsolateral prefrontal cortex, hippocampus, cingulate cortex, and superior temporal gyrus [168–170]. Individuals with schizophrenia also had a general reduction in cortical gray matter, or a loss of nerve cell bodies and branching dendrites, when compared to unaffected siblings [171, 172]. Additional imaging studies also identified widespread white matter abnormalities, suggesting neuron connectivity may be impaired due to dysfunctional myelination [173]. Together, these results indicated that brain morphology and function was systematically altered in schizophrenia, with many changes present prior to the onset of disease.

Adverse pre-natal outcomes and lower childhood cognitive ability were linked to the development of schizophrenia in large-scale epidemiological studies. Developmental delay and obstetrical complications were associated with up to a 4.6-fold increase in the schizophre-

nia risk [39], and on average, individuals with schizophrenia displayed deficits in cognitive and motor function during childhood preceding the onset of illness [40]. Pre-term births, defined as low birth weight and a shortened gestation period, also increased risk for a range of childhood and psychiatric disorders, including schizophrenia [174].

In addition, a number of early environmental exposures have been associated with schizophrenia risk. First, children born during times of extreme and persistent famine in the Netherlands and China sustained increased rates of psychiatric disorders and brain abnormalities in later life [35–38]. Second, infections during the neonatal period, in particular with *Toxoplasma gondii*, were associated with increased risk for schizophrenia [175, 176]. Third, early childhood traumas, especially sexual abuse, were linked to a 3.16-fold increase in reported psychotic symptoms [31–34]. Finally, individuals who migrated between the ages of 0 and 4 years were more frequently diagnosed with psychotic disorders (rate ratio = 2.96), and this risk decreased with older age at migration [177]. Combined, these epidemiological and clinical studies suggested that early environmental exposures and pre-morbid symptoms in childhood were strong predictors of development of schizophrenia in adolescence and adulthood.

Evidence for a neurodevelopmental etiology to schizophrenia was further supported by recent results from genetic analyses of common variants. By comparing array-based genotype data across disorders, these studies demonstrated that common risk variants are shared, to varying degrees, between individuals with schizophrenia, bipolar disorder, major depressive disorder, attention-deficit hyperactivity disorder, and autism spectrum disorders (ASD) [70, 71]. The strongest correlation was observed between schizophrenia and bipolar disorder (0.64 ± 0.04, 95% CI), with the weakest between schizophrenia and autism, a neurodevelopmental disorder (0.16 ± 0.06, 95% CI). The genetic correlation between many of these psychiatric and neurodevelopmental disorders is likely driven by a number of pleiotrophic common variants; however, the biological processes that underlie these variants have not yet been identified. Combined, results across imaging, epidemiological, clinical, and genetic studies suggest that certain neurodevelopmental processes, when dysregulated, could result in increased risk for adult-onset psychiatric disorders.

4.1.2 Sharing of rare variants between autism spectrum disorders and intellectual disability

Recent sequencing studies demonstrated that the sharing of genetic risk in brain disorders extended to rare coding variants, with most of this evidence coming from analyses of autism, intellectual disability (ID), and developmental disorders. The largest sequencing study of

autism to date meta-analysed multiple sources of rare variant data, including *de novo* SNVs, *de novo* small CNVs, and inherited rare variants, and implicated 46 genes and 6 CNV regions at a FDR of 5% [109]. I intersected these autism risk genes with the Developmental Disorder Genotype-Phenotype (DDG2P) database to determine if they were additionally associated with broader syndromic features [157, 118]. This database was developed as a tool for identifying likely causal variants for severe developmental disorders in the Deciphering Developmental Disorders (DDD) study. While the original list identified developmental disorder genes using information from OMIM, UniProt, and a systematic screen of journal publications since 2005, it had since incorporated robust gene discoveries from the DDD study. Intriguingly, 20 of the 46 autism genes and all six risk CNVs had previously been described as dominant causes of severe developmental disorders (Figure 4.1). Some of these, such as *ADNP*, *ARID1B*, the 1q21.1 and 22q11.2 locus, defined well-known clinical syndromes characterized by intellectual disability and distinctive facial features [157, 118]. Further support for this shared overlap came from phenotypic analyses of probands with mutations in these genes. Autistic individuals with an IQ below the median (89) had a 1.7-fold higher rate of *de novo* CNVs and SNVs when compared to probands with an IQ above the study median [149, 155, 109]. However, an excess burden of *de novo* mutations was still observed in cases even at an IQ of above 130, suggesting that while these rare variants were strongly associated with cognitive impairment, they also contributed to risk in the full range of individuals with autism. Together, these genetic analyses showed that a shared genetic etiology existed across neurodevelopmental disorders, with a particularly strong rare variant overlap between autism spectrum disorders and intellectual disability.

4.1.3 Individual loci increasing risk for schizophrenia and neurodevelopmental disorders

However, the evidence from rare variants for a broader shared genetic etiology between schizophrenia and neurodevelopmental disorders is more mixed. An analysis of *de novo* mutations from schizophrenia probands 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 [98]. A whole-exome sequencing study of 2,536 schizophrenia cases and 2,543 controls tested for a burden of rare LoF and nonsynonymous variants in candidate gene sets for autism and intellectual disability, including genes hit by *de novo* mutations in intellectual disability and autism, but did not observe any overlap [103]. Evidence at individual rare schizophrenia risk loci suggested that a partial, perhaps weaker overlap may exist between psychiatric and neurodevelopmental disorders. First,

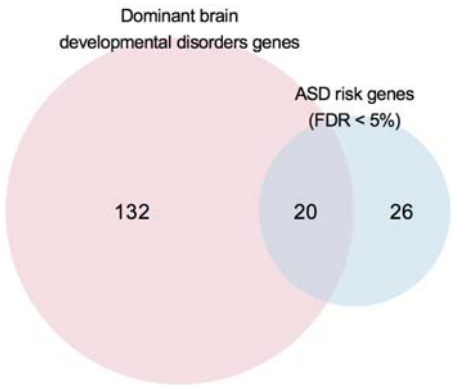


Fig. 4.1 **The overlap between autism risk genes and dominant developmental disorder genes.** This Venn Diagram illustrates the overlap between the autism risk genes implicated by Sanders *et al.* at FDR $< 5\%$ (46 genes) and dominant brain developmental disorder risk genes described in the DDG2P database (152 genes).

all 11 recurrent rare copy number variants shown to substantially increased the risk for schizophrenia (OR > 2) also increase risk for developmental disorders and congenital malformations [67, 158]. Notably, the penetrance of these CNVs was at least several fold higher for the development of a childhood-onset disorder, such as ID and ASD, than for schizophrenia. In our meta-analysis of 16,000 whole exomes, I showed that LoF variants in *SETD1A* conferred substantial risk for both schizophrenia and developmental disorders [119]. Seven of the ten carriers with schizophrenia had pre-morbid additional learning difficulties, and four additional carriers were identified among 4,281 children with severe developmental disorders sequenced as part of the DDD study. Therefore, emerging results from these individual risk loci showing pleiotropic effects offer the possibility that a larger number of developmental disorder genes could additionally confer substantial risk for schizophrenia.

4.1.4 Genes with near-complete depletion of protein-truncating variants

Insights into the rare variant architecture of autism and developmental disorders also emerged from a large-scale analysis that identified individual genes intolerant to mutational change. This effort was led by the Exome Aggregation Consortium (ExAC), a global effort to compile publicly available exome sequence data, and aimed to find the set of genes most enriched for variants that individually confer substantial risk for human disease. They calculated the selective constraint for every gene in the genome by comparing the observed number of rare

loss-of-function variants in exomes from 60,706 unrelated individuals without severe, early-onset disorders to the number predicted by a gene-specific mutation rate model [138]. Using a Gaussian mixture model, each gene was assigned a probability of being loss-of-function intolerant (pLI) score, which separated genes with sufficient observations into LoF intolerant (pLI > 0.9) and LoF tolerant (pLI < 0.1). From these analyses, 3,230 genes were identified with near-complete depletion of such truncating variants [109, 112, 138], which I refer to as the "highly constrained" gene set. The pLI score correlated well with other approaches that also aimed to identify genes under purifying selection, and as expected, pLI > 0.9 genes were over-represented in OMIM as having variants causing autosomal and X-linked dominant Mendelian diseases [138]. When applied to sequencing studies of autism trios, constrained genes were found to contain a 2.3-fold enrichment of *de novo* LoF variants compared to expectation in the mutational rate model [109, 112, 138]. It was not too surprising then, that autism risk genes identified in De Rubeis and Sanders *et al.* were overwhelmingly genes that were under selective constraint. Furthermore, the targets of key neural regulatory genes previously implicated in autism, such as translational targets of *FMRP*, promoter targets of *CHD8*, and splice targets of *RBFOX* [105, 178], also showed significant overlap with the constrained gene set. Finally, the *de novo* LoF mutations identified in probands with severe developmental disorders and intellectual disability also resided disproportionately in genes with more extreme constraint values ($P < 1 \times 10^{-6}$) [138]. Given this evidence, it is possible that the variants conferring substantial risk in psychiatric disorders, including schizophrenia and bipolar disorders, also resided within these highly constrained genes.

4.1.5 Aims and goals

Here, I describe a series of analyses integrating large-scale genetic datasets to explore the potential overlap of genetic risk between schizophrenia and broader developmental disorders. I jointly analysed data from whole-exome sequences from 1,077 schizophrenia trios, 4,264 cases and 9,343 controls, and array-based CNV calls from 6,882 cases and 11,255 controls. While the identification of individual genes remained difficult, I performed enrichment analyses testing for a higher burden of rare, disruptive SNVs and CNVs in 1,766 gene sets, including the highly constrained gene set and other groups of genes previously implicated in intellectual disability and autism. I also obtained cognitive measures for a subset of schizophrenia cases, including 279 patients with pre-morbid intellectual disability, and 1,165 cases who do not have intellectual disability. I compared the enrichment of rare variants in each of these clinical subsets to determine if there was a link between LoF burden and additional cognitive impairment. Combined, I present a detailed analysis of one of the largest accumulation of rare variant data for schizophrenia to date to better understand which genes

are implicated by this class of variants, and how they relate to neurodevelopment more generally.

4.1.6 Publication note and contributions

The results described in this Chapter has been submitted to BiorXiv and is currently undergoing peer-review. I designed the study, aggregated the required data, performed all of the analysis, and generated all the Figures and Tables described in this Chapter. This work was completed under the supervision of Jeffrey C. Barrett. Elliot Rees kindly provided the ClozUK CNV calls from his previous publication [67]. James T. R. Walters provided detailed phenotypic information for the Cardiff data set. Mandy Johnstone provided clinical details for the MUIR data set. Robin M. Murray, Marta Di Forti, Elvira Bramon, and Conrad Iyegbe provided cognitive measures for the London cohort. Jaana Suvisaari and Minna Tornianen provided cognitive measures for the Finnish cohort. Patrick Sullivan provided data on educational attainment on the Swedish individuals. I wrote the first draft of the manuscript, and received very helpful corrections, comments, and suggestions from my supervisor Jeffrey C. Barrett. The manuscript was further improved after receiving useful comments from Dave Curtis, Michael J. Owen, and Michael C. O'Donovan. Unless explicitly stated, the parts of the peer-reviewed publication reproduced in this chapter are my original work.

4.2 Methods

4.2.1 Sample collections

The data production, and quality control of the schizophrenia case-control whole-exome sequencing data set were described in detail in Section 2.4 and in a previous publication [119]. Briefly, I jointly called each case data set with its nationality-matched controls, and excluded samples based on contamination, coverage, non-European ancestry, and excess relatedness. I applied a number of empirically derived variant- and genotype-level filters, including filters on GATK VQSR, genotype quality, read depth, allele balance, missingness, and Hardy-Weinberg disequilibrium. The per-sample metrics were comparable between batches following QC. In total, 4,264 cases and 9,343 controls were available for analysis.

The data production and quality control of the array-based CNV case-control data set were described in an earlier publication [179]. The schizophrenia cases were recruited as part of the CLOZUK and CardiffCOGS studies, which consisted of both schizophrenia individuals taking the antipsychotic clozapine and a general sample of cases from the UK. Matched controls were selected from four publicly available non-psychiatric data sets. All

samples were genotyped using Illumina arrays at the Broad Institute, and processed and called under the same protocol. The log R ratios and B-allele frequencies were generated using the Illumina Genome Studio software, and CNVs were called with PennCNV using a consensus set of 520,766 probes shared across arrays. Individuals with outlying values in raw CNV metrics (log R ratio and B-allele frequencies) and per-sample CNV counts were excluded. I further excluded samples based on non-European ancestry, excess relatedness, and contamination. Only CNVs supported by more than 10 probes and greater than 10 Kilobases in size were retained to ensure high quality calls. In total, 6,882 cases and 11,255 controls were available for analysis. Finally, Sanger-validated *de novo* mutations identified through whole exome-sequencing of 1,077 schizophrenia parent-proband trios were aggregated and re-annotated for enrichment analyses [98, 101, 95, 102, 99, 96, 97]. A full description of each trio study, including sequencing and capture technology and sample recruitment was provided in Section 3.2.3.

The Ensembl Variant Effect Predictor (VEP) version 75 was used to annotate all variants (SNVs and CNVs) according to GENCODE v.19 coding transcripts. I defined frameshift, stop gained, splice acceptor and donor variants as loss-of-function (LoF), and missense or initiator codon variants with a CADD Phred score ≥ 15 as damaging missense. A deletion was annotated as disrupting a gene if the deletion overlapped a part of the gene's coding sequence. I more conservatively defined genes as duplicated only if the entire canonical transcript of the gene overlapped with the duplication event.

4.2.2 Rare variant gene set enrichment analyses

Case-control enrichment burden tests

For the case-control SNV data set, I performed permutation-based gene set enrichment tests using an extension of the variant threshold method described in Price *et al.* [180]. The method assumed that variants with a minor allele frequency (MAF) below an unknown threshold T were more likely to be damaging than variants with a MAF above T, and this threshold was allowed to differ for every gene or pathway tested. To consider different possible values for threshold T, a gene or gene set test statistic $t(T)$ was calculated for every allowable T, and the maximum test-statistic, or $t_{\max}$, was selected. The statistical significance of $t_{\max}$ was evaluated by permuting phenotypic labels, and calculating $t_{\max}$ from the permuted data such that different values of T could be selected following each permutation. In Price *et al.*, $t(T)$ was defined as the z-score calculated from regressing the phenotype on the sum of the allele counts of variants in a gene with MAF $< T$. I extended this method to test for enrichment in gene sets by regressing schizophrenia status on the total number of damaging

alleles in the gene set of interest with MAF $< T$ ($X_{in,T}$) while correcting for the total number of damaging alleles genome-wide with MAF $< T$ ($X_{all,T}$). $X_{all,T}$ was added as a covariate to control for any exome-wide differences between schizophrenia cases and controls, ensuring any significant gene set result was significant beyond baseline differences. $t(T)$ was defined as the t-statistic testing if the regression coefficient of $X_{in,T}$ deviated from 0. I then calculated $t(T)$ for all thresholds below a minor allele frequency of 0.1%, and selected the maximum value for the $t_{\max}$ based on the observed data. To calculate a null distribution for $t_{\max}$, I performed two million case-control permutations within each population (UK, Finnish, and Swedish) to control for batch and ancestry, and calculated $t_{\max}$ for each permuted sample while allowing T to vary. The P-value for each gene set was calculated as the fraction of the two million permuted samples that had a greater $t_{\max}$ than what was observed in the unpermuted data. The odds ratio and 95% confidence interval of each gene set was calculated using a logistic regression model, regressing schizophrenia status on X_{in} while controlling for total number of variants genome-wide (X_{all}) and population (UK, Finnish, and Swedish). Unlike gene set P-values which were calculated using permutation across multiple frequency thresholds, the odds ratios and 95% CI were calculated using only variants observed once in our data set (allele count of 1) to ensure they were comparable between tested gene sets.

CNV logistic regression

For enrichment analyses using the case-control CNV data set, I adapted the logistic regression framework described in Raychaudhuri *et al.* and implemented in PLINK to compare the case-control differences in the rate of CNVs overlapping a specific gene set [181]. Importantly, this method corrected for differences in CNV size and total genes disrupted [182, 106, 181]. I first restricted our analyses to coding deletions and duplications, and tested for enrichment using the following model:

$$\log \frac{P_{i,\text{case}}}{1-P_{i,\text{case}}} = \beta_0 + \beta_1 s_i + \beta_2 g_{\text{all}} + \beta_3 g_{\text{in}} + \varepsilon \tag{4.1}$$

where for individual i, p_i is the probability they have schizophrenia, s_i is the total length of CNVs, g_{all} is the total number of genes overlapping CNVs, and g_{in} is the number of genes within the gene set of interest overlapping CNVs. It has been shown that β_1 and β_2 sufficiently controlled for the genome-wide differences in the rate and size of CNVs between schizophrenia cases and controls, while β_3 captured the true gene set enrichment above this background rate [182, 106, 181]. For each gene set, I reported the one-sided P-value, odds ratio, and 95% confidence interval of β_3.

Weighted permutation-based sampling of *de novo* mutations

For each variant class of interest (LoF, missense, and synonymous as control), I tabulated the total number of *de novo* mutations observed in the 1,077 schizophrenia trios (N_{obs}). I then generated 2 million random samples of N_{obs} *de novo* mutations of the variant class of interest. To ensure the mutations were reasonably distributed across the genome, I weighted the probability of observing a *de novo* event in a gene by its estimated mutation rate. These baseline gene-specific mutation rates were calculated using the method described in Samocha *et al.* and extended to produce LoF and damaging missense rates for each GENCODE v.19 gene [138]. I then calculated one-sided enrichment *P*-values for each gene set as the fraction of the two million random samples that had a greater or equal number of *de novo* mutations in the gene set of interest than what is observed in the 1,077 trios:

$$P_{\text{gene set}} = \frac{\text{number of times } N_i \geq N_{\text{obs}}}{N_{\text{perm}}} \tag{4.2}$$

where N_i is the number of *de novo* mutations in random sample i that hit a gene in the gene set of interest, and N_{perm} is the total number of random samples (2×10^6). The effect size of the enrichment was calculated as the ratio between the number of observed mutations in the gene set of interest and the average number of mutations in the gene set across the two million random samples, or $\frac{N_{\text{obs}}}{E(N_i)}$. I adapted a method in Fromer *et al.* to calculate 95% credible intervals for the enrichment statistic [98]. I first generated a list of one thousand evenly spaced values between 0 and ten times the point estimate of the enrichment. For each value, the mutation rates of genes in the gene set of interest were multiplied by that amount, and 50,000 random samples of *de novo* mutations were generated using these weighted rates. The probability of observing the number of mutations in the gene set of interest given each effect size multiplier was calculated as the fraction of samples in which the number of mutations in the gene set was the same as the observed number in the 1,077 trios. I normalised the probabilities across the 1,000 values to generate a posterior distribution of the effect size, and calculated the 95% credible interval using this empirical distribution.

4.2.3 Combined joint analysis

Gene set *P*-values calculated using the case-control SNV, case-control CNV, and *de novo* data were meta-analysed using Fisher's combined probability method to provide a single test

statistic for each gene set:

$$X^2_{2k} \sim -2(\ln(p_{\text{DNM}}) + \ln(p_{\text{SNV}}) + \ln(p_{\text{CNV}}))$$

where p_{DNM}, p_{SNV}, and p_{CNV} are the gene set *P*-values for the corresponding test, $k = 3$ is the number of tests being combined, and X^2 followed a χ-square distribution with $2k = 6$ degrees of freedom. I corrected for the number of gene sets tested in the discovery analysis (N = 1,766) by controlling the false discovery rate (FDR) using the Benjamini-Hochberg approach. The p.adjust() function in R was used to calculate FDR-corrected *p*-values, or *q*-values, for each gene set. I reported only results with a *q*-value of less than 5%.

4.2.4 Description of gene sets

Public gene set databases

When aggregating different gene sets from various sources, I re-mapped all gene identifiers to the GENCODE v.19 release, and excluded all non-coding genes from further analysis. First, I accessed and combined gene sets from five public databases: Gene Ontology (release 146; June 22, 2015 release), KEGG (July 1, 2011 release), PANTHER (May 18, 2015 release), REACTOME (March 23, 2015 release), and the Molecular Signatures Database (MSigDB) hallmark processes (version 4, March 26, 2015 release). Given our focus on very rare (MAF $< 0.1\%$ or singleton variants) and *de novo* variants, I had limited power to detect enrichment in small gene sets, as evident in previous studies of schizophrenia and autism rare variation in which the strongest signals came from aggregating hundreds of genes [98, 103, 105]. Therefore, I restricted our analyses to 1,687 gene sets from the five public databases with more one hundred genes.

Schizophrenia candidate gene sets

I additionally tested gene sets selected based on biological hypotheses about schizophrenia risk, and genome-wide screens investigating rare variants in broader neurodevelopmental disorders. These included gene sets described in previous enrichment analyses of schizophrenia rare variants [66, 103]: translational targets of *FMRP* [183, 184], components of the post-synaptic density [66, 103], ion channel proteins [103], components of the ARC, mGluR5, and NMDAR complexes [103], proteins at cortical inhibitory synapses [182, 185], targets of mir-137 [103], and genes near schizophrenia common risk loci [57, 103].

Constrained genes

To extend results from autism and intellectual disability, I tested if the burden of rare variants in individuals with schizophrenia was similarly concentrated in genes intolerant of protein-truncating variants. I used the pLI metric described in the ExAC v0.3.1 database as a measure of gene-level selective constraint [112]. Since the full v0.3.1 release contained the Swedish schizophrenia study, I used the subset of the ExAC database that excluded data sets that included individuals with a psychiatric diagnosis for all analyses in this study. The pLI metric was computed from non-psychiatric release of 45,376 exomes. I defined all genes annotated with pLI > 0.9 as "highly constrained", and genes annotated with pLI < 0.9 were described as "ExAC unconstrained". The "highly constrained" gene set was composed of 3,488 genes, while the "ExAC unconstrained" gene set was composed of 14,753 genes. To provide a higher resolution test of how damaging variants were distributed at different levels of constraint, I further ranked and grouped genes into deciles and bideciles according to the pLI metric (top 10%, top 20%, etc.), and tested for rare variant enrichment using these smaller gene sets.

Risk genes for autism and neurodevelopmental disorders

The DECIPHER Developmental Disorder Genotype-Phenotype (DDG2P) database (April 13, 2015 release) was used to define genes diagnostic of developmental disorders [157, 118]. For a high confidence list as used for clinical reporting in the DDD study, I included genes with a monoallelic or a X-linked dominant mode of inheritance and robust evidence in the literature ("Confirmed DD Genes", "Probable DD gene", "Both DD and IF"). From these genes, I created four lists based on mechanism (LoF or LoF/missense) and affected organ system (brain/cognition or any organ system). I further extended these list with novel genes for severe developmental disorders identified in 4,293 parent-proband trios exome sequenced in the DDD study [186]. The 94 genome-wide significant genes were described in Supplementary Table 3 in McRae *et al.*. Significant genes with *de novo* LoF mutations were appended to the LoF and LoF/missense lists, while genes with only *de novo* missense mutations were only added to the LoF/missense lists. To define a list of high-quality autism risk genes, I used the genome-wide results from the largest meta-analysis of ASD whole-exome sequences to date [109]. ASD risk genes were defined as genes with a FDR $< 10\%$ or $< 30\%$ in Sanders *et al.* For a less stringent list of candidate neurodevelopmental and autism risk genes, I separately defined ASD and developmental disorder *de novo* genes as genes hit by a LoF or a LoF/missense *de novo* variant in the Sanders *et al.* and the DDD study [109, 118]. I additionally incorporated gene sets previously shown to be enriched for

de novo mutations in autism probands: targets of *CHD8* [105, 187, 178], splice targets of *RBFOX* [105, 188, 189], hippocampal gene expression networks [190], and neuronal gene lists from the Gene2cognition database (http://www.genes2cognition.org) [105].

Brain expression gene sets

Finally, as background gene sets, I defined cerebellar and cortical genes as those that expressed in at least 80% of the corresponding human brain samples in the Brainspan RNA-seq dataset [191]. I defined a gene as expressed in a sample if the exon and whole gene read counts were greater than 10 counts, and the Cufflinks lower-bound FPKM estimate was greater than 0 [192]. For brain-enriched genes or genes preferentially expressed in the brain, I compared the differential expression of individual genes in the brain against all other tissues in the GTEx dataset [193], and identified a subset that is 2-fold enriched with a FDR $< 5\%$.

4.2.5 Conditional analyses

A number of gene sets previously implicated in neurodevelopmental disorders, such as the translational targets of *FMRP*, were enriched for constrained genes and brain-expressed genes [138]. However, both these larger gene sets contained a disproportionate number of *de novo* mutations in autism probands, making it difficult to determine if our results for smaller gene sets were significant beyond the enrichment in brain-expressed and highly constrained genes. To address this, I extended each of three methods used for gene set enrichment to condition on different gene set backgrounds. I first restricted all variants analysed to those that reside in the background gene list (B) before testing for an excess of rare variants in genes shared between the gene set of interest (K) and the background list. I focused on two background gene sets: brain-enriched genes from GTEx, and the ExAC constrained gene list (pLI > 0.9) (described above). In the enrichment analyses of the case-control SNV data, I modified the variant threshold method to regress schizophrenia status on the total number of damaging alleles in genes present in both the gene set of interest and the background gene set ($K \cap B$), while correcting for the total number of damaging alleles in the set of all background genes (B). The logistic regression model for the case-control CNV data was modified to:

$$\log \frac{P_{i,\text{case}}}{1 - P_{i,\text{case}}} = \beta_0 + \beta_1 s_i + \beta_2 g_B + \beta_3 g_{K \cap B} + \varepsilon \tag{4.3}$$

where g_B is the total number of background genes overlapping a CNV, and $g_{K \cap B}$ is the number of genes in the intersection of the gene set of interest and the background list overlapping

a CNV. Finally, I determined the total number of *de novo* mutations observed in the 1,077 schizophrenia trios that hit a gene in the background gene list. I then generated 2 million random samples with the same number of *de novo* mutations. For each gene set, one-sided enrichment *P*-values were calculated as the fraction of two million random samples that had a greater or equal number of *de novo* mutations in genes in $K \cap B$ than what was observed in the 1,077 trios. Gene set *P*-values were combined using Fisher's method. I restricted our conditional enrichment analysis to gene sets with q-value $< 1\%$ in the discovery analysis, and adjusted for multiple testing using Bonferroni correction.

4.2.6 Rare variants and cognition in schizophrenia

Within the UK10K study, 97 individuals from the MUIR collection were given discharge diagnoses of mild learning disability and schizophrenia (ICD-8 and -9). The recruitment guidelines of the MUIR collection were described in detail in a previous publication [194]. In brief, evidence of remedial education was a prerequisite to inclusion, and individuals with pre-morbid IQs below 50 or above 70, severe learning disabilities, or were unable to give consent were excluded. The Schizophrenia and Affective Disorders Schedule-Lifetime version (SADS-L) in people with mild learning disability, PANSS, RDC, and DSM-III-R, and St. Louis Criterion were applied to all individuals to ensure that any diagnosis of schizophrenia was robust. In the clinical information provided alongside the Swedish and Finnish case-control data sets, I identified 182 schizophrenia individuals who were similarly diagnosed with intellectual disability. Combined, I identified 279 individuals with a diagnosis of schizophrenia and intellectual disability.

I used cognitive testing and educational attainment in the remaining samples to identify schizophrenia individuals without intellectual disability. For 502 individuals from the Cardiff collection in the UK10K study, I acquired their pre-morbid IQ as extrapolated from National Adult Reading Test (NART), and identified 412 individuals for analysis after excluding all individuals with predicted pre-morbid IQ of less than 85 (or below one standard deviation of the population distribution for IQ). I additionally acquired information on educational attainment in 54 schizophrenia individuals in the UK10K London collection, and retained 27 individuals who completed at least 13 years of schooling. These individuals completed additional schooling following compulsory education. Lastly, the California Verbal Learning Test was conducted on 124 Finnish schizophrenia individuals sequenced as part of UK10K, and a composite score was generated from measures of verbal and visual working memory, verbal abilities, visuoconstructive abilities, and processing speed. All individuals with intellectual disability had been excluded from cognitive testing. Within this set of samples, I additionally excluded any individuals who ranked in the lowest decile in CVLT composite

score, and retained 92 individuals for analysis. According to these criteria, I identified 531 of 697 schizophrenia individuals from the UK and Finnish data sets with cognitive data as not having intellectual disability. I additionally acquired data on educational attainment for the Swedish schizophrenia cases and controls from the Swedish National Registry. After excluding individuals with intellectual disability, I identified 751 schizophrenia individuals who did not attend secondary school (less than 9 years of schooling), 776 schizophrenia individuals who completed compulsory schooling but did not complete secondary schooling (less than 12 years of schooling), and 634 schizophrenia individuals who completed at least compulsory and upper secondary schooling (at least 12 years of schooling). I defined the subset of 634 schizophrenia individuals as cases without intellectual disability. In total, combining the UK, Finnish, and Swedish data, I identified 1,165 schizophrenia individuals without cognitive impairment.

Using the case-control SNV enrichment method, I tested for differences in rare variant burden between the following samples: 279 schizophrenia individuals with ID and 9,270 matched controls, and 1,165 schizophrenia individuals without ID and 9,270 matched controls. I also tested for differences in rare variant burden between 279 schizophrenia individuals with ID and the 1,165 schizophrenia individuals without ID. These analyses were restricted to two gene sets of interest: the constrained gene set (pLI $>$ 0.9) and diagnostic developmental disorder genes with brain abnormalities as described in DECIPHER DDG2P database (Figure 4.11, 4.12). Because we performed three pairwise tests of LoF burden across two gene sets, I controlled for multiple testing using Bonferroni correction, and required any result to have a p-value of less than 0.0083 (0.05/6) to be significant.

4.3 Results

4.3.1 Study design

To maximize our power to detect signals of enrichment of damaging variants in groups of genes, I performed a meta-analysis of three different types of rare coding variant studies. Previous results from these data gave us confidence to proceed with gene set enrichment analyses. Statistical tests of the case-control exome data used case-control permutations within each population (UK, Finnish, Swedish) to generate empirical *P*-values to test hypotheses. When applying this method, I observed no genome-wide inflation was observed in burden tests of individual genes (Section 3.3.1). In the curated set of *de novo* mutations, I observed the expected exome-wide number of synonymous mutations given gene mutation rates from previously validated models [138], suggesting variant calling was generally unbiased across

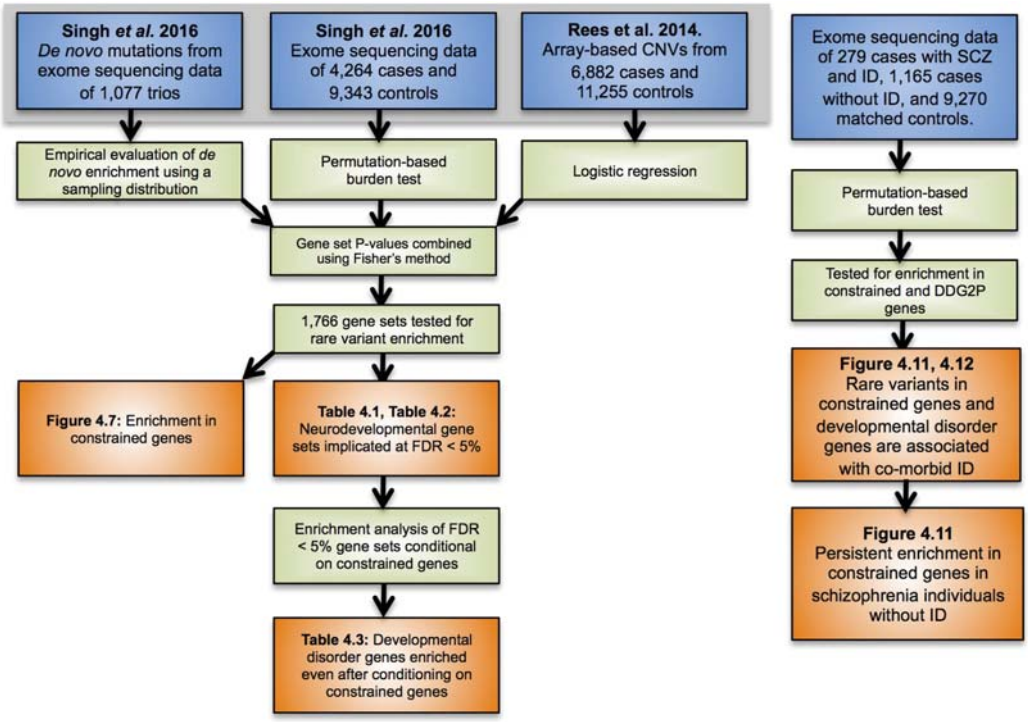


Fig. 4.2 **Analysis workflow.** Data sets are shown in blue, statistical methods and analysis steps are shown in green, and results (figures and tables) from the analysis are shown in orange. The left chart describes analyses testing for enrichment in 1,766 gene sets using the entire data set. The right chart describes analyses testing for enrichment in constrained and developmental disorder genes in the subset of cases with cognitive information.

GENCODE v.19 coding genes (Section 3.3.6). Lastly, the case-control CNV data set had been previously analysed for burden of CNVs affecting individual genes and enrichment analyses in targeted gene sets [182, 179]. Because I had limited power to implicate individual genes, I focused our analyses on testing for an excess of rare damaging variants in schizophrenia patients in a number of gene sets. For each data type (case-control SNV, CNV, and *de novo* mutations), I used previously described methods appropriate to each data set to test for an excess of rare variants (Figure 4.2). Gene set *P*-values computed using the three methods were meta-analysed using Fisher's Method to provide a single *P*-value for each gene set. Because I weighted the information from each data type equally, gene sets achieving significance typically show at least some signal in all three types of data.

4.3.2 Selection of allele frequency thresholds and consequence severity

For the case-control whole-exome data, I applied an extension of the variant threshold model for gene set enrichment analyses. With this method, I did not need to select an *a priori* MAF cut-off, and was able to test damaging variants at a number of frequency thresholds. All thresholds below a MAF of 0.1% were tested, and statistical significance was assessed by permutation testing. For all the whole-exome data (case-control and trio data), I restricted gene set analyses to loss-of-function variants, since these variants had been demonstrated to show the strongest enrichment for truly damaging variants compared to other functional classes. In total, 118 LoF *de novo* variants were observed in the 1,077 parent-proband trios.

For the case-control CNV data, I compared the CNV burden at four MAF thresholds ($< 1\%$, $< 0.5\%$, $< 0.1\%$, singleton), and three variant classes (deletions, duplications, and both). When conducting additional robustness checks (Section 4.3.3), I found that the gene set *P*-values for CNV burden were dramatically inflated even when testing for enrichment in a large number of random gene sets (Figure 4.3). After stratifying by CNV size, frequency, type (deletion and duplications), and quality and testing for burden, I determined that this inflation was driven in part by very large (overlapping more than 10 genes), common (MAF between 0.1% and 1%) CNVs observed mainly in either cases or controls. Excluding this highly influential class of CNVs greatly reduced the genomic inflation (Figure 4.4). Unfortunately, some of these were the 11 recurrent schizophrenia CNVs, and likely harboured true risk genes. However, because these CNVs were highly recurrent in cases, depleted in controls and disrupted a large number of genes, any gene set that included even a single gene within these CNVs would appear to be significant, even after controlling for total CNV length and genes overlapped. To ensure our model was well-calibrated and its *P*-values followed a null distribution for random gene sets, I conservatively restricted our analysis to rare and small copy number events (Figure 4.4). In summary, I restricted our analysis to case-control

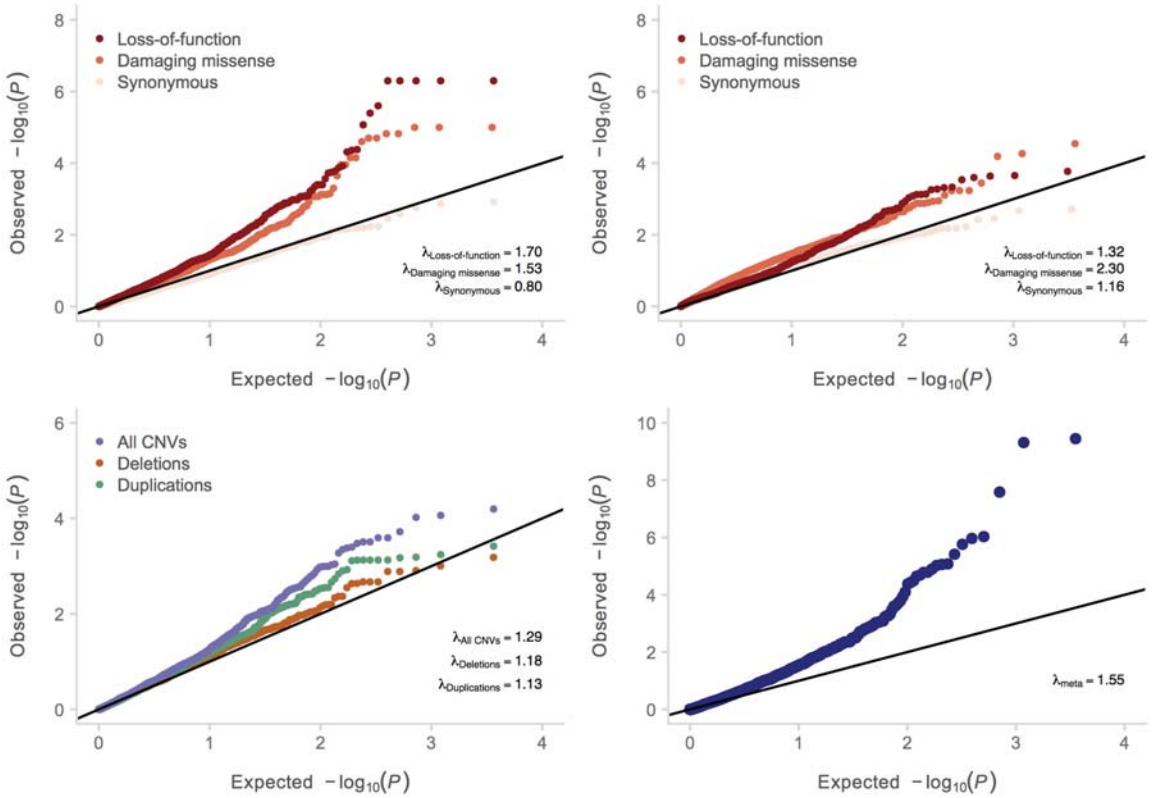


Fig. 4.3 **Q-Q plots of *P*-values from enrichment tests of 1,766 gene sets. Top left:** case-control SNVs from whole-exome sequence data; **Top right:** *de novo* mutations from 1,077 trios; **Bottom left:** case-control CNVs; **Bottom right:** meta-analysed *P*-values from Fisher's method (dark blue). Calibrated MAF cut-offs and a tailored enrichment test were applied to each variant type. Each dot represented a different gene set. General inflation of *P*-values from tests of disruptive variants (loss-of-function in *de novo* tests, and CNVs) was observed. The genomic inflation parameter λ was provided for each distribution. Damaging missense: missense variants with CADD Phred > 15.

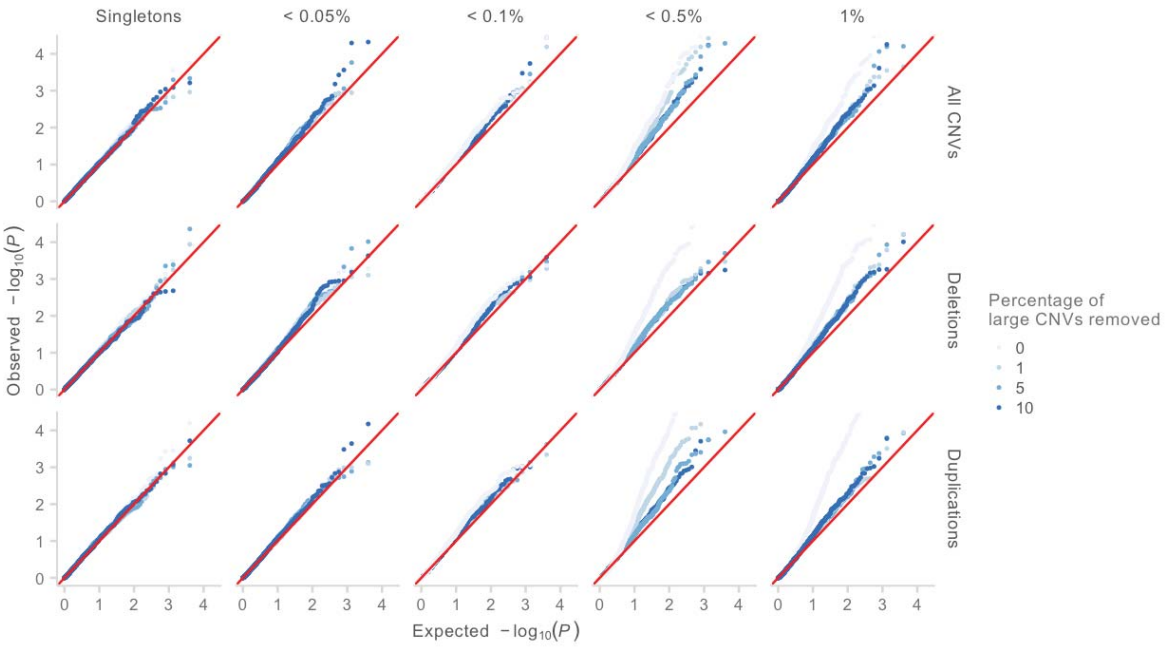


Fig. 4.4 **The use of frequency and size cut-offs in CNV gene sets enrichment tests to reduce genomic inflation.** Q-Q plots were generated based on *P*-values from CNV enrichment tests of random gene sets, using different MAF cut-offs (Singletons, $< 0.05\%$, $< 0.1\%$, 1%) and CNV size cut-offs (removing the top 1%, 5%, and 10% of CNVs overlapping the most genes). Each dot represented a different gene set. Inflation followed the expected null distribution when more stringent MAF thresholds and size cut-offs were applied (see MAF $< 0.1\%$, and removing the 10% of CNVs overlapping the most genes). Singletons: CNVs observed to occur once in our data set.

loss-of-function (LoF) variants, small deletions and duplications overlapping fewer than seven genes (excluding the largest 10% of CNVs) with MAF $< 0.1\%$ (Figure 4.4), and *de novo* mutations annotated as LoF.

4.3.3 Robustness of enrichment analyses

I tested for an excess of rare damaging variants in schizophrenia patients in 1,766 gene sets. However, I observed an inflation in the quantile-quantile (Q-Q) plot of gene set *P*-values (Figure 4.3), so I took several steps to ensure our results were not biased due to methodological or technical artefacts in our data. First, biases related to analytical method or data QC should systematically affect all classes of variants, including synonymous variants. Using the same data and methods, I observed no inflation of *P*-values when testing for enrichment of synonymous variants in our case-control and *de novo* analyses (Figure 4.3). Second, I uniformly sampled genes from the genome (as defined by GENCODE v.19) to generate random gene sets with the same size distribution as the 1,766 gene sets in our discovery analysis. For each random set, I calculated gene set *P*-values for the case-control SNV data, case-control CNV data, and *de novo* data using the appropriate method and frequency cut-offs across all variant classes. Reassuringly, I observed null distributions in all such Q-Q plots regardless of variant class and analytical method (Figure 4.5). These findings suggested that our methods sufficiently corrected for known genome-wide differences in LoF and CNV burden between cases and controls, and other technical confounders like batch and ancestry. I then tabulated the number of gene sets each gene was found in, and discovered that certain genes were over-represented in pathways from the four gene set databases compared to a random sampling of genes from the genome (Figure 4.6). Furthermore, the top 1000 over-represented genes were generally more enriched for rare disruptive variants in schizophrenia cases compared to controls ($P = 0.005$, Figure 4.6b) while no enrichment was observed after excluding the top 5000 most frequent genes. This observation would partially explain the inflation in our Q-Q plots, but there was not an obvious reason for why certain genes were over-represented in these public databases. I hypothesized that an ascertainment bias may partially explain this: some genes, like *p53*, *TNF*, *NFKB*, and *APOE*, are much more thoroughly investigated in the literature because disruption in these genes across species result in striking biological consequences. It could also be that these over-represented genes have multiple core functions impacting a number of biological processes. A pathway analysis of common variants in psychiatric disorders also displayed similar inflation of *P*-values when testing for enrichment in gene sets from GO, KEGG, and Reactome, suggesting that gene sets from these public databases were also enriched for common variant signal in schizophrenia. [108]. Together, these results indicated that interpretation of pathway analyses requires

careful attention to potential sources of bias, but that our data, analytic methods, and main results are robust.

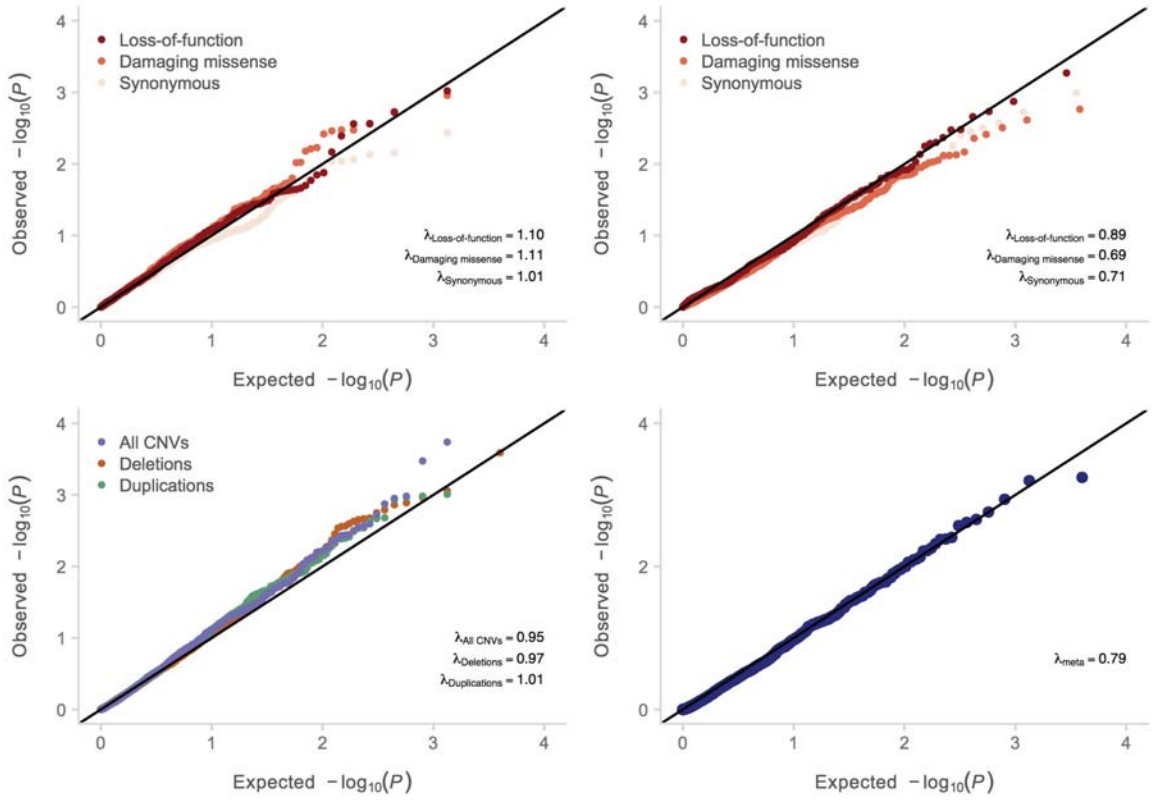


Fig. 4.5 **Q-Q plots of *P*-values from enrichment tests of random gene sets. Top left**: case-control SNVs from whole-exome sequence data; **Top right**: *de novo* mutations from 1,077 trios; **Bottom left**: case-control CNVs. Genes were randomly sampled from the genome to create gene sets with the same size distribution as the 1,766 tested gene sets. Each dot represented a different gene set. Calibrated MAF cut-offs and a tailored enrichment test were applied to each variant type. The genomic inflation parameter λ was provided for each distribution. No inflation of test statistics was observed across all variant types. Damaging missense: missense variants with CADD Phred > 15.

4.3.4 Rare, damaging schizophrenia variants are concentrated in constrained genes

Recent studies have demonstrated that recurrent *de novo* LoF and missense mutations identified in probands with autism or developmental disorders were overwhelmingly concentrated in the set of highly constrained genes [109, 112, 138], suggesting that at least some of the constraint was driven by severe neurodevelopmental consequences of having only one

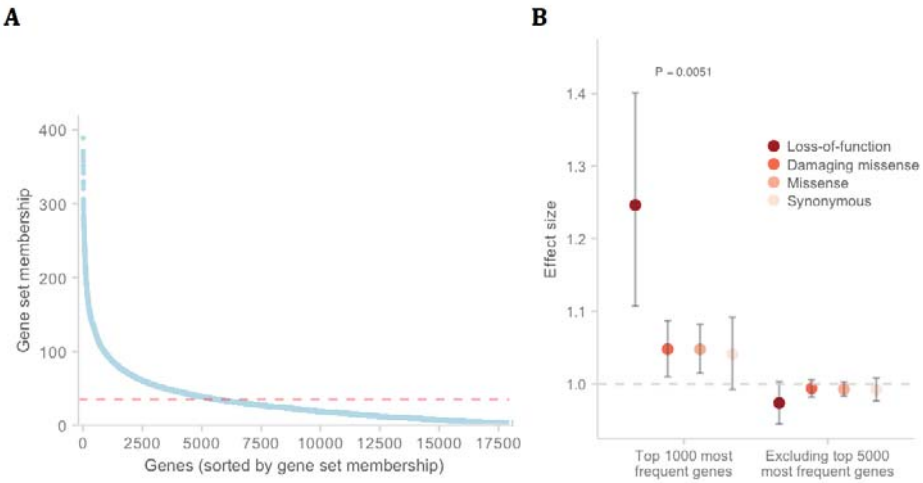


Fig. 4.6 **Non-random sampling of genes in the 1,766 gene sets resulted in non-null enrichment of disruptive variants. A**: Genes were ranked and plotted based on the number of gene sets they belonged in. The top 1000 genes were massively over-represented in gene sets from public databases, and genes outside the top 5000 genes were under-represented. **B**: Case-control SNV burden tests of genes over-represented and under-represented in the 1,766 gene sets. The top 1000 most over-represented genes showed a significant enrichment of LoF variants, while no enrichment was observed for genes outside the top 5000 genes. Plotted *P*-values were from burden tests of LoF variants, and error bars described the 95% confidence interval of the burden estimate. Damaging missense: missense variants with CADD Phred > 15.

functioning copy of these genes. I found that rare damaging variants in schizophrenia cases were also enriched in the highly constrained gene set ($P < 3.6 \times 10^{-10}$, Table 4.1, Figure 4.7), with support in case-control SNVs ($P < 5 \times 10^{-7}$; OR 1.24, $1.16 - 1.31$, 95% CI), case-control CNVs ($P = 2.6 \times 10^{-4}$; OR 1.21, $1.15 - 1.28$, 95% CI), and *de novo* mutations ($P = 6.7 \times 10^{-3}$; OR 1.36, $1.1 - 1.68$, 95% CI). The constrained genes signal in schizophrenia was distributed across many genes: if I ranked genes by decreasing significance, the enrichment disappeared in the case-control SNV analysis ($P > 0.05$) only after the exclusion of the top 50 genes, suggesting that many genes contributed to this observation, rather than just a handful of genes with very large burden.

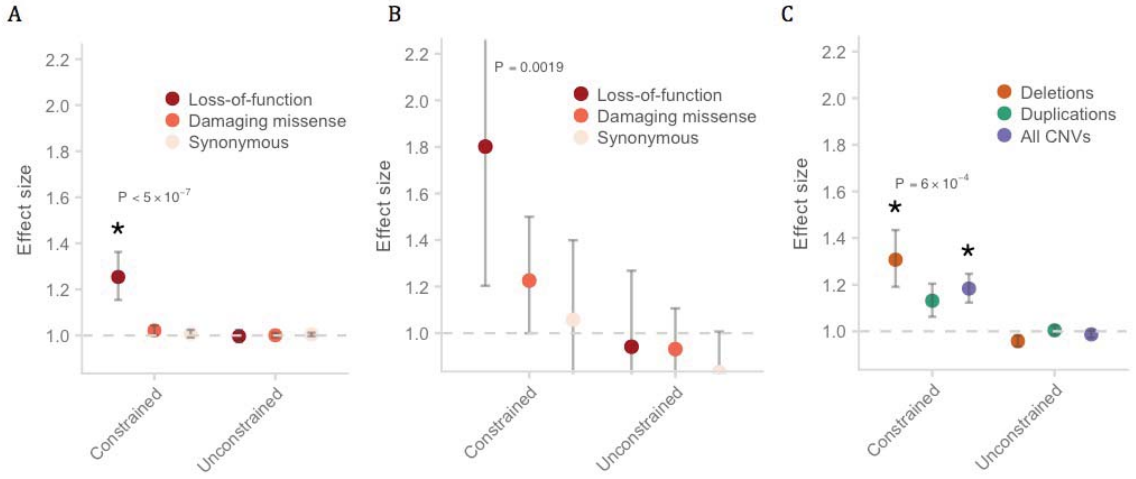


Fig. 4.7 **Enrichment of schizophrenia rare variants in constrained genes. A:** Schizophrenia cases compared to controls for rare SNVs and indels; **B:** Rates of *de novo* mutations in schizophrenia probands compared to control probands; **C:** Case-control CNVs. *P*-values shown were from the test of LoF enrichment in **A**, LoF and damaging missense enrichment in **B**, and all CNVs enrichment in **C**. Error bars represent the 95% CI of the point estimate. Constrained: 3,488 genes with near-complete depletion of truncating variants in the ExAC database; Unconstrained: genes not under genic constraint; Damaging missense: missense variants with CADD Phred > 15. Asterisk: $P < 1 \times 10^{-3}$.

4.3.5 Comparing the enrichment in constrained genes across neurodevelopmental disorders

I next contrasted the degree of enrichment of *de novo* mutations in constrained genes between probands with developmental disorders, autism, and schizophrenia. First, I aggregated and re-annotated *de novo* mutations from four studies (1,113 probands with developmental disorders [118], 4,038 probands with ASD [109, 105], and 2,134 control probands [155, 105]), and

used the Poisson exact test to compare the *de novo* rates in constrained genes between affected probands and matched controls. I tested for differences in counts in each functional class (synonymous, missense, damaging missense, and LoF) separately, and displayed the one-sided *P*-value, rate ratio, and 95% CI of each comparison in Figure 4.8 and 4.9. Overall, while the enrichment in schizophrenia was consistent with observations in developmental disorders and autism [138, 105], the absolute effect size was smaller (Figure 4.8, 4.9). Finally, in the remaining 14,753 genes in the genome, I observed no excess burden of rare damaging variants in schizophrenia, autism, and severe developmental disorders, suggesting dominant alleles conferring substantial risk for brain disorders are concentrated in the constrained gene set (Figure 4.7, 4.9, 4.10).

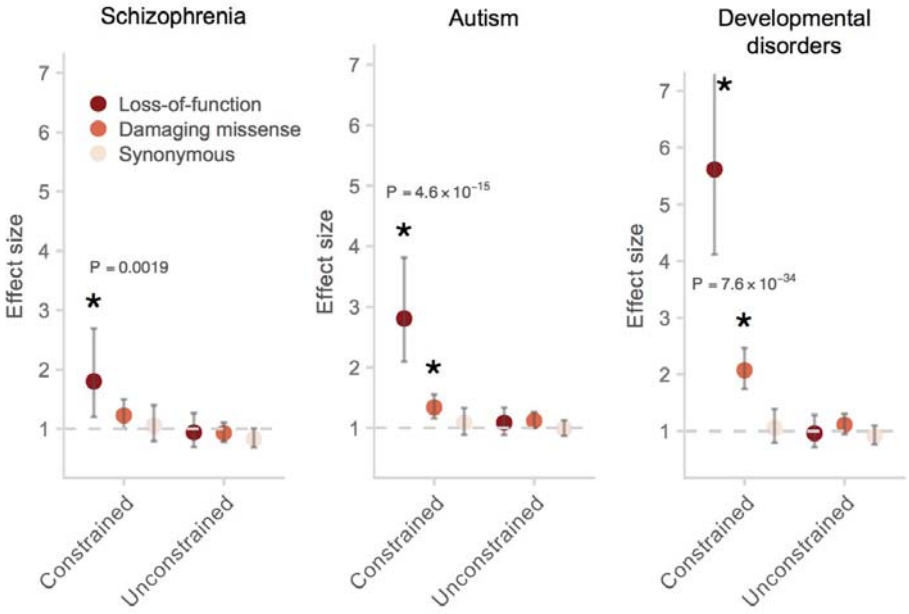


Fig. 4.8 **Enrichment of *de novo* mutations in genes with near-complete depletion of truncating variants across schizophrenia and neurodevelopmental disorders.** In autism, schizophrenia, and severe neurodevelopmental disorders, *de novo* mutations were enriched in a subset of genes under genic constraint, with no excess of polygenic burden in the remaining genes. To generate 95% CI and *P*-values, the rate of *de novo* mutations in affected trios (1,077 schizophrenia trios, 1,133 trios with severe neurodevelopmental disorders, and 4,038 trios with autism) was compared against the rate in unaffected control trios (2,038 trios) using a Poisson exact test. Plotted *P*-values were from the Poisson test of LoF mutations. Damaging missense: missense variants with CADD Phred > 15.

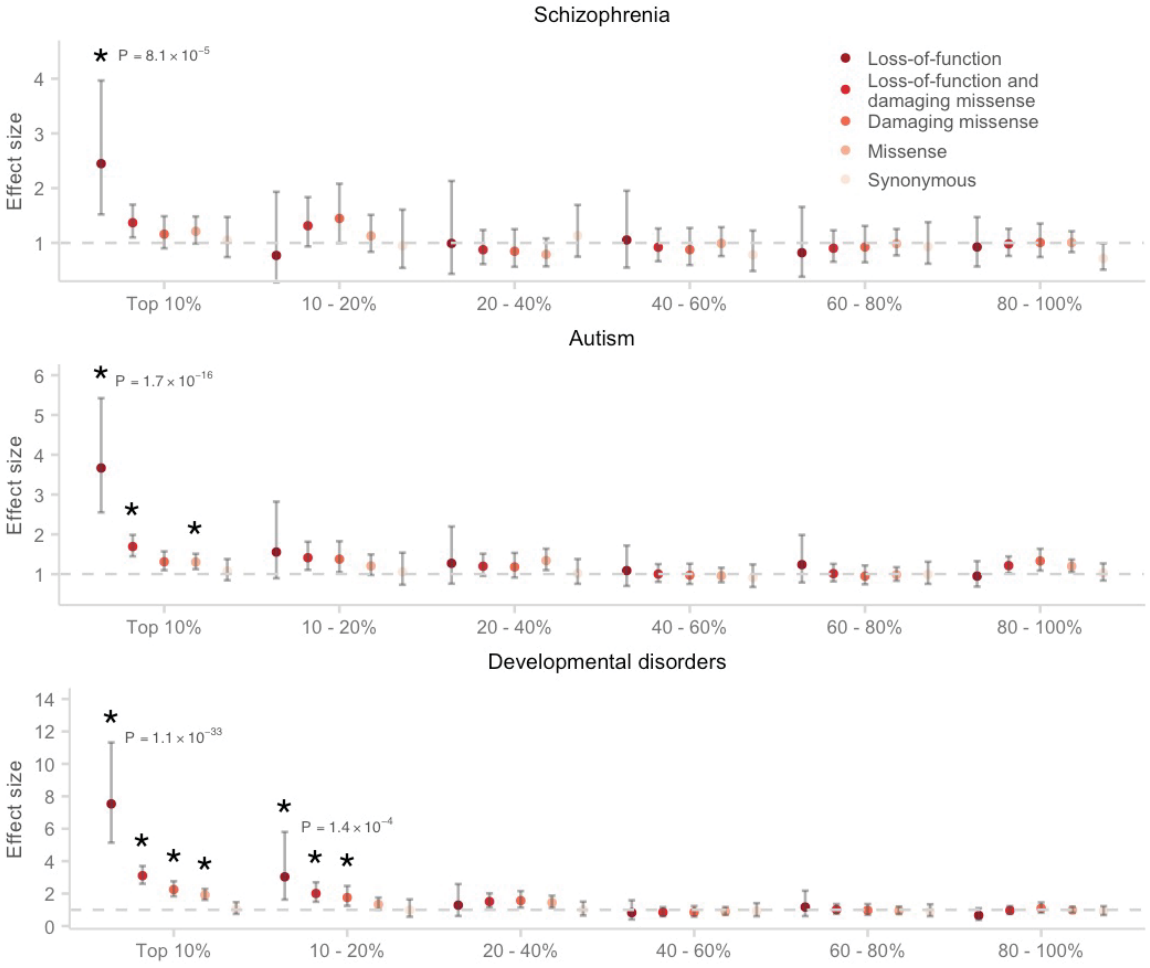


Fig. 4.9 **Enrichment of *de novo* mutations in genes ordered and grouped by genic constraint across schizophrenia and neurodevelopmental disorders.** Genes were ordered by their degree of constraint (pLI score), and grouped into six categories: the 10% most constrained, 10 – 20% most constrained, 20 – 40% most constrained, and so on. The rate of *de novo* mutations in affected trios (1,077 schizophrenia trios, 1,133 trios with severe neurodevelopmental disorders, and 4,038 trios with autism) was compared against the rate in unaffected control trios (2,038 trios) using a Poisson exact test. A significant enrichment of rare LoF and damaging missense variants was only observed in the 20% most constrained genes, while no signal was observed in less constrained genes. Error bars were 95% CI of the estimate. Plotted *P*-values were from the Poisson test of LoF mutations. Damaging missense: missense variants with CADD Phred $>$ 15.

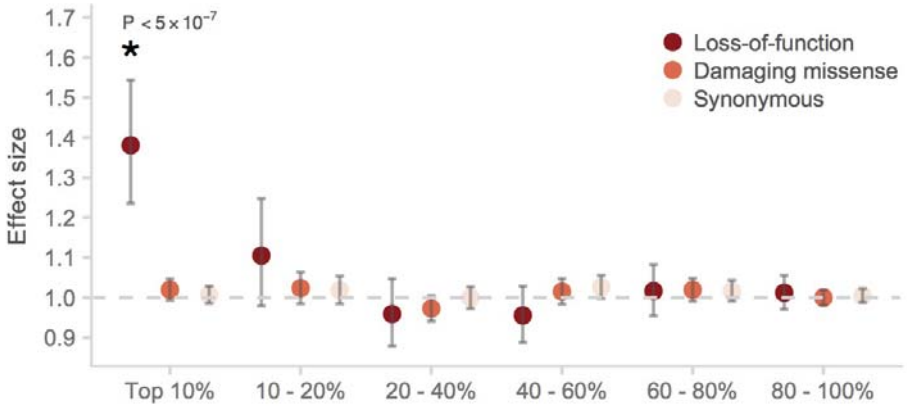


Fig. 4.10 **Enrichment of case-control SNVs in genes ordered and grouped by genic constraint.** Genes were ordered by their degree of constraint (pLI score), and grouped into six categories: the 10% most constrained, $10-20$% most constrained, $20-40$% most constrained, and so on. A significant enrichment of rare LoF and damaging missense variants was only observed in the 10% most constrained genes, while no signal was observed in less constrained genes. Synonymous variants followed an expected null distribution. Error bars were 95% CI of the estimate. The asterisk indicated that $P < 1 \times 10^{-3}$. Damaging missense: missense variants with CADD phred > 15.

4.3.6 Schizophrenia risk genes are shared with other neurodevelopmental disorders

Given the consistent enrichment of rare damaging variants in constrained genes in schizophrenia, autism, and neurodevelopmental disorders, I next determined whether these variants affected the same genes. I found that both autism risk genes identified from exome sequencing analyses [109] and genes in which LoF variants are known causes of severe developmental disorders [157] were significantly enriched for rare variants in individuals with schizophrenia ($P_{\text{ASD}} = 9.5 \times 10^{-6}$; $P_{\text{DD}} = 2.3 \times 10^{-6}$; Table 4.1). Previous studies had shown an enrichment of rare damaging variants in mRNA targets of *FMRP* in both schizophrenia and autism [155, 103, 105], which I confirmed (Table 4.1). I sought to identify further shared biology by testing targets of neural regulatory genes previously implicated in autism [105, 178], and observed similar enrichment of promoter targets of *CHD8* ($P = 1.1 \times 10^{-6}$) and splice targets of *RBFOX* ($P = 1.3 \times 10^{-5}$).

I tested an additional 1,759 gene sets, and observed a total of 35 with an enrichment at FDR $q < 0.05$ (Table 4.2). I replicated previously implicated gene sets, like glutamatergic synaptic density proteins comprising the NMDAR and ARC complexes [98, 66, 103, 183], and identified novel gene sets, such as regulation of transmembrane transport (GO:0034762) and cytoskeleton organisation (GO:0007010). Notably, the gene sets most significantly

Name	N_{genes}	Est_{SNV}	P_{SNV}	Est_{DNM}	P_{DNM}	Est_{CNV}	P_{CNV}	P_{meta}	Q_{meta}
ExAC constrained genes (pLI > 0.9)	3488	1.24 (1.16-1.31)	$< 5.0 \times 10^{-7}$	1.36 (1.1-1.68)	0.0067	1.21 (1.15-1.28)	0.00026	3.60×10^{-10}	4.30×10^{-7}
Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders with brain abnormalities	156	1.42 (1.07-1.88)	0.011	4.18 (2.21-8.03)	0.00073	1.92 (1.54-2.39)	0.0016	2.30×10^{-6}	0.00067
Sanders *et al.* autism risk genes (FDR < 10%)	66	1.28 (0.97-1.69)	0.0095	3.96 (1.65-9.94)	0.019	2.21 (1.75-2.79)	0.00033	9.50×10^{-6}	0.0017
Darnell *et al.* targets of FMRP	790	1.24 (1.13-1.36)	8.5×10^{-6}	1.31 (0.83-2.09)	0.17	1.32 (1.2-1.47)	0.0032	9.30×10^{-7}	0.00038
Cotney *et al.* CHD8-targeted promoters (hNSC and human brain tissue)	2920	1.09 (1.02-1.16)	0.0008	1.77 (1.36-2.31)	0.00025	1.11 (1.05-1.18)	0.027	1.10×10^{-6}	0.00038
G2CDB: mouse cortex post-synaptic density consensus	1527	1.2 (1.11-1.3)	2.5×10^{-6}	1.57 (1.06-2.33)	0.028	1.04 (0.96-1.11)	0.32	3.90×10^{-6}	0.00097
Weynvanhentenryck *et al.* CLIP targets of RBFOX	967	1.21 (1.11-1.33)	4.8×10^{-5}	1.84 (1.21-2.8)	0.0085	1.07 (0.98-1.17)	0.2	1.30×10^{-5}	0.002
NMDAR network (defined in Purcell *et al.)*	61	1.66 (1.09-2.54)	0.0061	5.6 (2.06-16.09)	0.017	2.46 (1.78-3.4)	0.0028	3.70×10^{-5}	0.0044
GOBP: chromatin modification (GO:0016568)	519	1.29 (1.13-1.49)	0.00018	2.26 (1.32-3.94)	0.0099	1.12 (0.99-1.28)	0.18	4.20×10^{-5}	0.0046

Table 4.1 **Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR $< 1\%$.** The effect sizes and corresponding *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) are shown for each gene set, along with the Fisher's combined *P*-value (P_{meta}) and the FDR-corrected *Q*-value (Q_{meta}). I only show the most significant gene set if there are multiple ones from the same data set or biological process. All gene sets displayed had been previously implicated in ASD and ID. N_{genes}: number of genes in the gene set; Est: effect size estimate and its lower and upper bound assuming a 95% CI; DNM: *de novo* mutations.

enriched (FDR $q < 0.01$) for schizophrenia rare variants (Table 4.1) were all neurodevelopmental gene sets previously implicated in autism and intellectual disability (mRNA targets of *FMRP*, chromatin modification, organization, and binding [GO], promoter targets of *CHD8* [157, 105, 178, 183]) as well as the large and generic set of cerebellum expressed and brain-enriched genes. A number of these gene sets, such as the translational targets of *FMRP* and risk genes for autism and developmental disorders, significantly overlapped with brain-expressed genes and constrained genes, both of which also carried a disproportionate burden of rare variants in schizophrenia. I extended previous methods to allow for conditional analyses using different gene set backgrounds, and found that the FDR $< 5\%$ neurodevelopmental gene sets were significant even after controlling for baseline enrichment in brain-enriched genes, demonstrating that they were biologically meaningful beyond brain expression (Table 4.3). Strikingly, only two gene sets, known ASD risk genes ($P = 4 \times 10^{-4}$) and diagnostic DD genes ($P = 3 \times 10^{-5}$), had an excess of rare coding variants above the enrichment already observed in constrained genes (Table 4.3). Thus, in addition to biological pathways implicated specifically in schizophrenia, at least a portion of the schizophrenia risk conferred by rare variants of large effect is shared with childhood onset disorders of neurodevelopment.

4.3.7 Schizophrenia rare variants are associated with intellectual disability

In the autism spectrum disorders, the observed excess of rare damaging variants was much greater in individuals with intellectual disability than those with normal levels of cognitive function [155]. A similar reduction in cognitive function was observed in schizophrenia carriers of *SETD1A* LoF variants and the 22q11.2 deletion syndrome [159, 119]. Motivated by these observations, I next sought to explore whether this pattern is consistent in schizophrenia in a wider set of genes. 279 individuals in the whole-exome data set had pre-morbid intellectual disability in addition to fulfilling the full diagnostic criteria for schizophrenia. I also accumulated cognitive phenotype data for the remaining samples, and identified 1,165 individuals with schizophrenia who I could confirm do not have intellectual disability (after excluding pre-morbid $IQ < 85$, fewer than 12 years of schooling or lowest decile of composite cognitive measures, depending on available data). When stratifying into these two groups (cases with intellectual disability, unknown cognitive status, no intellectual disability), I observed that the burden of damaging rare variants in constrained genes was significantly greater in the small set of cases with confirmed intellectual disability than in both the remaining schizophrenia cases and matched controls (Figure 4.11). Schizophrenia individuals

Name	N_genes	P_SNV	P_DNM	P_CNV	P_meta	Q-value
ExAC constrained genes (pLI > 0.9)	3488	5.00E-07	0.0067	0.00026	3.60E-10	4.30E-07
Top 10% of genes ranked by genic constraint	1824	5.00E-07	0.00083	0.0029	4.90E-10	4.30E-07
Top 3% of genes ranked by genic constraint	548	5.00E-07	0.0021	0.086	2.60E-08	1.50E-05
Darnell et al. targets of FMRP	790	8.50E-06	0.17	0.0032	9.30E-07	0.00038
Cotney et al. CHD8-targeted promoters (hNSC and human brain tissue)	2920	0.0008	0.00025	0.027	1.10E-06	0.00038
Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders with brain abnormalities	156	0.011	0.00073	0.0016	2.30E-06	0.00067
G2CDB: mouse cortex post-synaptic density consensus	1527	2.50E-06	0.028	0.32	3.90E-06	0.00097
Cotney et al. CHD8-targeted promoters (hNSC, mouse, human brain tissue)	2073	0.00012	0.0033	0.14	8.70E-06	0.0017
Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders	266	0.0039	0.0018	0.0087	9.40E-06	0.0017
Sanders et al. autism risk genes (FDR < 10%)	66	0.0095	0.019	0.00033	9.50E-06	0.0017
Weynvanhentenryck et al. CLIP targets of RBFOX	967	4.80E-05	0.0085	0.2	1.30E-05	0.002
Cerebellum-expressed genes (Brainspan)	15976	0.00085	0.13	0.0011	1.60E-05	0.0024
Cotney et al. CHD8-targeted promoters (human brain tissue)	2663	0.0055	0.00048	0.055	2.00E-05	0.0028
Sanders et al. autism risk genes (FDR < 30%)	180	0.0035	0.16	0.00045	3.20E-05	0.0041
NMDAR network (defined in Purcell et al.)	61	0.0061	0.017	0.0028	3.70E-05	0.0044
GOBP: chromatin modification (GO:0016568)	519	0.00018	0.0099	0.18	4.20E-05	0.0046
Dominant, diagnostic DDG2P genes, in which LoF and missense variants results in developmental disorders with brain abnormalities	191	0.047	0.002	0.0044	5.10E-05	0.0053
Dominant, diagnostic DDG2P genes, in which LoF and missense variants results in developmental disorders	340	0.011	0.0022	0.022	6.40E-05	0.0063
Cotney et al. CHD8-targeted promoters (hNSC)	9111	0.0043	0.047	0.0034	8.10E-05	0.0075

Post-synaptic density genes (as defined in Purcell et al.)	690	0.0004	0.23	0.011	0.00012	0.01
GOBP: chromatin organization (GO:0006325)	642	0.0008	0.0031	0.64	0.00016	0.014
GOCC: postsynaptic density (GO:0014069)	177	0.00017	0.28	0.042	0.00019	0.016
G2CDB cortex post-synaptic density consensus	748	0.00027	0.31	0.029	0.00024	0.018
GOCC: cell projection part (GO:0044463)	861	0.0016	0.19	0.0085	0.00025	0.019
GOBP: cytoskeleton organization (GO:0007010)	835	4.40E-05	0.88	0.095	0.00034	0.024
G2CDB: human PSP	1096	0.0004	0.15	0.093	0.00048	0.032
G2CDB mouse cortex post-synaptic density full list	967	0.00014	0.2	0.23	0.00054	0.034
G2CDB: TAP-PSD-95 pull-down core list	118	0.0011	0.18	0.032	0.00054	0.034
G2CDB: human cortex biopsy post-synaptic density genes	1056	0.0006	0.12	0.096	0.0006	0.036
G2CDB human orthologues of mouse NRC	184	0.0088	0.02	0.061	0.00082	0.047
Sanders et al. genes with damaging de novo mutations (LoF and mis3)	1702	0.48	0.0079	0.0028	0.00083	0.047
Brain-biased expression, using GTeX data	9349	0.0065	0.051	0.034	0.00087	0.047
GOMF: chromatin binding (GO:0003682)	446	0.01	0.0013	0.9	0.00092	0.047
GOBP: regulation of transmembrane transport (GO:0034762)	384	0.087	0.34	0.0004	0.00092	0.047
GOBP: chromosome organization (GO:0051276)	882	0.0012	0.014	0.72	0.00093	0.047

Table 4.2 **Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR $< 5\%$.** The effect sizes and corresponding *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) are shown for each gene set, along with the Fisher's combined *P*-value (P_{meta}) and the FDR-corrected *Q*-value (Q_{meta}). N_{genes}: number of genes in the gene set; Est: effect size estimate and its lower and upper bound assuming a 95% CI; SNV: single nucleotide variants from whole-exome data; DNM: *de novo* mutations.

Background	Name	P_SNV	P_DNM	P_CNV	P_meta
Brain-biased expression, using GTeX data	Top 10% of genes ranked by genic constraint	5.00E-07	0.0043	0.023	1.50E-08
Brain-biased expression, using GTeX data	ExAC constrained genes (pLI > 0.9)	5.00E-07	0.017	0.02	4.90E-08
Brain-biased expression, using GTeX data	Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders with brain abnormalities	0.015	0.00038	0.002	2.10E-06
Brain-biased expression, using GTeX data	Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders	0.016	0.00095	0.0019	4.90E-06
Brain-biased expression, using GTeX data	Top 3% of genes ranked by genic constraint	4.00E-05	0.004	0.21	5.70E-06
Brain-biased expression, using GTeX data	Sanders et al. autism risk genes (FDR < 30%)	0.0003	0.093	0.0022	9.70E-06
Brain-biased expression, using GTeX data	Sanders et al. autism risk genes (FDR < 10%)	0.0065	0.016	0.001	0.000016
Brain-biased expression, using GTeX data	GOBP: chromatin organization (GO:0006325)	0.011	0.00045	0.036	0.000024
Brain-biased expression, using GTeX data	GOBP: chromatin modification (GO:0016568)	0.0076	0.00084	0.029	0.000025
Brain-biased expression, using GTeX data	Post-synaptic density genes (as defined in Purcell et al.)	0.0004	0.25	0.0022	0.000029
Brain-biased expression, using GTeX data	Dominant, diagnostic DDG2P genes, in which LoF and missense variants results in developmental disorders with brain abnormalities	0.074	0.0011	0.0036	0.000037
Brain-biased expression, using GTeX data	G2CDB: mouse cortex post-synaptic density consensus	0.000054	0.11	0.056	0.000043
Brain-biased expression, using GTeX data	NMDAR network (defined in Purcell et al.)	0.003	0.014	0.0082	0.000043
Brain-biased expression, using GTeX data	Dominant, diagnostic DDG2P genes, in which LoF and missense variants results in developmental disorders	0.049	0.0032	0.0034	0.000064
Brain-biased expression, using GTeX data	G2CDB cortex post-synaptic density consensus	0.0008	0.32	0.0045	0.00013
Brain-biased expression, using GTeX data	GOBP: chromosome organization (GO:0051276)	0.029	0.00038	0.26	0.00027
Brain-biased expression, using GTeX data	Darnell et al. targets of FMRP	0.001	0.27	0.012	0.0003
Brain-biased expression, using GTeX data	GOCC: postsynaptic density (GO:0014069)	0.0014	0.31	0.018	0.00063
Brain-biased expression, using GTeX data	G2CDB: human PSP	0.00036	0.52	0.044	0.00066
Brain-biased expression, using GTeX data	Cotney et al. CHD8-targeted promoters (hNSC and human brain tissue)	0.27	0.00039	0.083	0.0007
Brain-biased expression, using GTeX data	Cotney et al. CHD8-targeted promoters (hNSC, mouse, human brain tissue)	0.059	0.00097	0.19	0.00083
Brain-biased expression, using GTeX data	G2CDB: human cortex biopsy post-synaptic density genes	0.00055	0.5	0.042	0.00088
Brain-biased expression, using GTeX data	Cotney et al. CHD8-targeted promoters (human brain tissue)	0.35	0.00076	0.059	0.0011
Brain-biased expression, using GTeX data	G2CDB mouse cortex post-synaptic density full list	0.0025	0.19	0.06	0.0019
Brain-biased expression, using GTeX data	GOMF: chromatin binding (GO:0003682)	0.13	0.00038	0.6	0.002
Brain-biased expression, using GTeX data	GOCC: cell projection part (GO:0044463)	0.0039	0.24	0.036	0.0022
Brain-biased expression, using GTeX data	G2CDB human orthologues of mouse NRC	0.0046	0.14	0.068	0.0026
Brain-biased expression, using GTeX data	GOBP: cytoskeleton organization (GO:0007010)	0.00065	0.69	0.12	0.0031
Brain-biased expression, using GTeX data	Sanders et al. genes with damaging de novo mutations (LoF and mis3)	0.31	0.0033	0.067	0.0038
Brain-biased expression, using GTeX data	G2CDB: TAP-PSD-95 pull-down core list	0.026	0.19	0.021	0.0053
Brain-biased expression, using GTeX data	Weynvanhentenryck et al. CLIP targets of RBFOX	0.0031	0.082	0.47	0.0061
Brain-biased expression, using GTeX data	GOBP: regulation of transmembrane transport (GO:0034762)	0.21	0.59	0.014	0.048
Brain-biased expression, using GTeX data	Cotney et al. CHD8-targeted promoters (hNSC)	0.19	0.22	0.43	0.23

ExAC constrained genes (pLI > 0.9)	Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders with brain abnormalities	0.02	0.0019	0.0059	3.00E-05
ExAC constrained genes (pLI > 0.9)	Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders	0.016	0.0028	0.014	0.000076
ExAC constrained genes (pLI > 0.9)	Top 10% of genes ranked by genic constraint	0.00015	0.036	0.36	0.00019
ExAC constrained genes (pLI > 0.9)	Dominant, diagnostic DDG2P genes, in which LoF and missense variants results in developmental disorders	0.035	0.0022	0.064	0.00043
ExAC constrained genes (pLI > 0.9)	Sanders et al. autism risk genes (FDR < 10%)	0.032	0.024	0.0064	0.00044
ExAC constrained genes (pLI > 0.9)	Sanders et al. genes with damaging de novo mutations (LoF and mis3)	0.081	0.043	0.0015	0.00046
ExAC constrained genes (pLI > 0.9)	Sanders et al. autism risk genes (FDR < 30%)	0.011	0.14	0.0034	0.00047
ExAC constrained genes (pLI > 0.9)	Dominant, diagnostic DDG2P genes, in which LoF and missense variants results in developmental disorders with brain abnormalities	0.067	0.0043	0.021	0.0005
ExAC constrained genes (pLI > 0.9)	GOMF: chromatin binding (GO:0003682)	0.0048	0.0053	0.62	0.0012
ExAC constrained genes (pLI > 0.9)	GOBP: chromatin organization (GO:0006325)	0.014	0.0084	0.45	0.0031
ExAC constrained genes (pLI > 0.9)	G2CDB: TAP-PSD-95 pull-down core list	0.0096	0.16	0.08	0.0062
ExAC constrained genes (pLI > 0.9)	NMDAR network (defined in Purcell et al.)	0.074	0.024	0.072	0.0063
ExAC constrained genes (pLI > 0.9)	GOBP: chromatin modification (GO:0016568)	0.071	0.005	0.4	0.0069
ExAC constrained genes (pLI > 0.9)	GOBP: chromosome organization (GO:0051276)	0.032	0.0064	0.71	0.007
ExAC constrained genes (pLI > 0.9)	Top 3% of genes ranked by genic constraint	0.0076	0.04	0.49	0.0072
ExAC constrained genes (pLI > 0.9)	Cotney et al. CHD8-targeted promoters (hNSC and human brain tissue)	0.095	0.013	0.3	0.015
ExAC constrained genes (pLI > 0.9)	Cotney et al. CHD8-targeted promoters (hNSC, mouse, human brain tissue)	0.19	0.031	0.12	0.025
ExAC constrained genes (pLI > 0.9)	Cotney et al. CHD8-targeted promoters (human brain tissue)	0.057	0.026	0.48	0.025
ExAC constrained genes (pLI > 0.9)	GOBP: regulation of transmembrane transport (GO:0034762)	0.048	0.63	0.025	0.026
ExAC constrained genes (pLI > 0.9)	G2CDB human orthologues of mouse NRC	0.087	0.068	0.21	0.038
ExAC constrained genes (pLI > 0.9)	GOCC: postsynaptic density (GO:0014069)	0.016	0.53	0.21	0.047
ExAC constrained genes (pLI > 0.9)	Weynvanhentenryck et al. CLIP targets of RBFOX	0.059	0.072	0.55	0.06
ExAC constrained genes (pLI > 0.9)	Post-synaptic density genes (as defined in Purcell et al.)	0.071	0.4	0.087	0.062
ExAC constrained genes (pLI > 0.9)	G2CDB cortex post-synaptic density consensus	0.096	0.48	0.063	0.07
ExAC constrained genes (pLI > 0.9)	Darnell et al. targets of FMRP	0.17	0.4	0.046	0.073
ExAC constrained genes (pLI > 0.9)	Cotney et al. CHD8-targeted promoters (hNSC)	0.72	0.092	0.057	0.084
ExAC constrained genes (pLI > 0.9)	Brain-biased expression, using GTeX data	0.1	0.11	0.5	0.11
ExAC constrained genes (pLI > 0.9)	G2CDB: mouse cortex post-synaptic density consensus	0.26	0.14	0.28	0.17
ExAC constrained genes (pLI > 0.9)	G2CDB mouse cortex post-synaptic density full list	0.12	0.28	0.34	0.18
ExAC constrained genes (pLI > 0.9)	G2CDB: human cortex biopsy post-synaptic density genes	0.13	0.41	0.24	0.19
ExAC constrained genes (pLI > 0.9)	G2CDB: human PSP	0.12	0.45	0.24	0.19
ExAC constrained genes (pLI > 0.9)	GOCC: cell projection part (GO:0044463)	0.23	0.62	0.11	0.22
ExAC constrained genes (pLI > 0.9)	GOBP: cytoskeleton organization (GO:0007010)	0.38	0.9	0.21	0.5

Table 4.3 **Results from enrichment analyses of FDR < 5% gene sets, conditional on brain-expressed and ExAC constrained genes.** I restricted enrichment analyses to genes that resided in two different background gene sets (brain-enriched expression in GTeX, and ExAC-constrained genes), and determined if gene sets with FDR $< 5\%$ in the meta-analysis still had significance above the specific background. The *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) were shown for each gene set, along with the Fisher's combined *P*-value (P_{meta}). N_{genes}: number of genes in the gene set; SNV: single nucleotide variants from whole-exome data; DNM: *de novo* mutations.

with ID had a significantly elevated number of variants in diagnostic developmental disorder genes compared to the remaining cases and controls (Figure 4.12), and two additionally carried LoF variants in *KMT2A* and *KMT2D*. These two genes are from the same family of lysine methyltransferases as *SETD1A*, also known as *KMT2F*, shown previously as a schizophrenia risk gene [119].

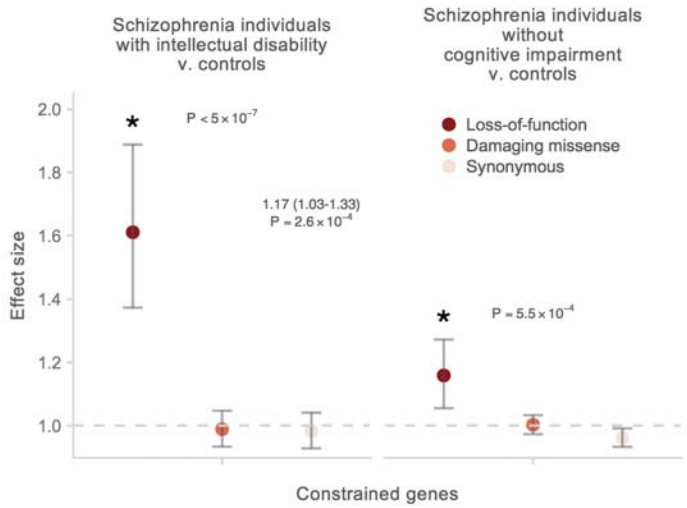


Fig. 4.11 **Enrichment of rare variants in constrained genes between schizophrenia (SCZ) individuals with ID, schizophrenia individuals without ID, and matched controls.** The *P*-values shown were calculated from the burden test of LoF variants between the corresponding cases and matched controls. The enrichment of LoF variants in constrained genes between SCZ individuals with ID and SCZ individuals without ID was displayed as effect sizes and *P*-values above the case-control comparisons. Error bars represent the 95% CI of the point estimate. Damaging missense: missense variants with CADD phred > 15.

While the damaging rare variants in constrained genes were most strongly enriched in the subset of schizophrenia patients with intellectual disability, I still observed a significant burden in the individuals who did not have intellectual disability ($P < 5.5 \times 10^{-4}$) (Figure 4.11). I additionally identified twelve schizophrenia cases without ID carrying LoF variants in developmental disorder genes from the DDG2P database. These individuals satisfied the full diagnostic criteria for schizophrenia without signs of pre-morbid intellectual disability (Table 4.4). Combined, I show that rare damaging variants in constrained genes in schizophrenia follow the pattern previously described in autism: concentrated in individuals with intellectual disability, but not exclusive to that group.

Variant	Gene	nLoF in ExAC	pLI	Expected syndrome based on DDG2P	Clinical features	Educational attainment	Predicted pre-morbid IQ
1:151377686_GA/G frameshift variant	POGZ	2	1.000	Intellectual Disability	Acute onset at age 20, treatment-resistant schizophrenia with severe depressive and negative symptoms.	Attended mainstream school, and achieved A level exams.	105
1:151400550_C/T stop gained	POGZ	2	1.000	Intellectual Disability	Paranoid schizophrenia, moderate negative symptoms, alcohol dependence.	Attended mainstream school, and achieved A level exams.	108
11:102076807_T/C splice donor variant	YAP1	0	0.999	Coloboma, ocular, with or without hearing impairment, cleft lip/palate, and/or mental retardation	Schizoaffective depression diagnosis, prominent negative symptoms.	Attended mainstream school, but left with no qualification.	86
16:89367335_C/A stop gained	ANKRD11	2	1.000	KBG syndrome	Late age of onset at age > 35, severe psychosis with first rank symptoms, multiple admissions, marked negative symptoms, and depression.	Attended mainstream school, but left with no qualification.	102

Table 4.4 **Phenotypes of schizophrenia individuals with cognitive information carrying LoF variants in developmental disorder genes.** Of the 531 UK10K schizophrenia individuals without intellectual disability, I acquired detailed clinical information for four out of the eight carriers of LoF variants in severe developmental disorders genes. These variants were observed only once in our data set and absent in the ExAC database. For each LoF variant, I provide its genomic coordinates (hg19) and the gene disrupted, the number of high-quality LoF variants within this gene identified in 60,706 ExAC individuals and the corresponding pLI score, and the expected developmental disorder syndrome according to DECIPHER. For each carrier, I describe notable neuropsychiatric symptoms (Clinical features), the level of education achieved (Education attainment), and the predicted pre-morbid IQ as extrapolated from National Adult Reading Test (NART). These four carriers satisfy the full diagnostic criteria for schizophrenia, and do not appear to be outliers in the expected cognitive range of schizophrenia patients. To identify high-quality ExAC LoF variants, I retained only variants in the canonical transcript and were called as homozygote (and not missing) in at least 85% of the ExAC data set (accessed on July 4th, 2016).

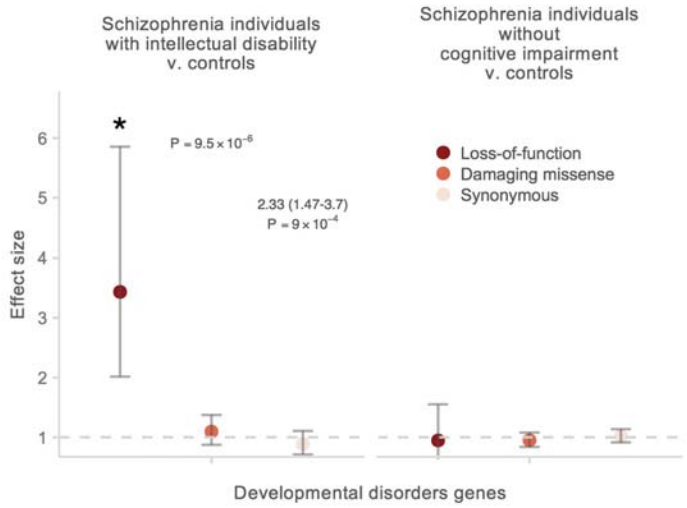


Fig. 4.12 **Enrichment of rare variants in diagnostic developmental disorder genes between schizophrenia (SCZ) individuals with ID, schizophrenia individuals without ID, and matched controls.** The *P*-values shown were calculated from the burden test of LoF variants between the corresponding cases and matched controls. The enrichment of LoF variants in constrained genes between SCZ individuals with ID and SCZ individuals without ID were displayed as effect sizes and *P*-values above the case-control comparisons. Error bars represent the 95% CI of the point estimate. Damaging missense: missense variants with CADD Phred > 15.

4.4 Discussion

My integrated analysis of rare variants from thousands of whole-exome sequences provides evidence for a partially shared genetic etiology between schizophrenia and other neurodevelopmental disorders. While the identification of individual genes remains difficult at current samples sizes, I demonstrate that the burden of *de novo* mutations, rare SNVs and CNVs in schizophrenia is primarily concentrated in a subset of 3,488 genes under genic constraint, an observation shared with autism and intellectual disability. Furthermore, enrichment analyses in a large number of gene sets demonstrate that the most robust burden of rare variants in schizophrenia resides in genes in which LoF variants are diagnostic for severe developmental disorders and in known autism risk genes. These results were supported by a recently published whole-exome sequencing study of Swedish schizophrenia cases and controls [134]. In so far as the genes responsible for intellectual disability necessarily have effects during central nervous system development, and those that influence ASD must exert their effects in infancy at the very latest, the findings demonstrate that genetic perturbations adversely affecting nervous system development also increase risk for schizophrenia. My findings therefore support the hypothesis that severe, psychiatric illnesses manifesting in adulthood can have origins early in development.

I additionally show that some of these perturbations have clear manifestations in childhood, and that risk variants of large effect in schizophrenia are associated with pre-morbid intellectual disability. Our observations are consistent with results in autism in which individuals carrying LoF *de novo* mutations are more likely to also have cognitive impairment [71, 109, 155]. Notably, I found that a weaker but still significant rare variant burden was observed in schizophrenia patients without intellectual disability, showing that variants of large effect do not simply confer risk for a small subset of schizophrenia patients but are relevant to disease pathogenesis more broadly.

My data support the general observation that genetic risk factors for psychiatric and neurodevelopmental disorders do not follow clear diagnostic boundaries, and that the variants disrupting the same genes, and quite possibly, the same biological processes, result in a wide range of phenotypic manifestation. For instance, a number of schizophrenia patients without intellectual disability carry LoF variants in developmental disorder genes that are purified of damaging mutations in the general population. This clinical pleiotrophy is reminiscent of LoF variants in *SETD1A* and 11 large copy number variant syndromes, previously shown to confer risk for schizophrenia in addition to other prominent developmental defects [67, 119]. I do not preclude the possibility that allelic series of LoF variants exist in these genes; however, the most common deletion in the 22q11.2 locus and a recurrent two base deletion in *SETD1A* are associated with both schizophrenia and more severe neurodevelopmental disorders,

suggesting the same variants confer risk for a range of clinical features [119, 195, 196]. Ultimately, it may prove difficult to clearly partition patients genetically into subgroups with similar clinical features, especially if genes and variants previously thought to cause well-characterized Mendelian disorders can have such varied outcomes. This pattern is consistent with the hypothesis that LoF variants in constrained genes result in a spectrum of neurodevelopmental outcomes with the burden of mutations highest in intellectual disability and least in schizophrenia, corresponding to a gradient of neurodevelopmental pathology indexed by cognitive impairment [15].

Despite the complex nature of genetic contributions to risk of schizophrenia, it is notable that across study designs (trio or case-control) and variant class (SNVs or CNVs), risk loci of large effect are concentrated in a small subset of genes. Previous rare variant analyses in other neurodevelopmental disorders, such as autism, have successfully integrated information across *de novo* SNVs and CNVs to identify novel risk loci [109]. As sample sizes increase, meta-analyses leveraging the shared genetic risk across study designs and variant types will be similarly well powered to identify additional risk genes in schizophrenia.

Chapter 5

Discussion and future directions

5.1 Summary of findings

In recent years, whole-exome sequencing has successfully identified individual genes in which rare variants or *de novo* mutations confer substantial risk for autism, intellectual disability, and severe developmental disorders. Indeed, these studies of broader neurodevelopmental disorders have independently revealed that many of the same genes are disrupted in patients with a wide range of diagnoses and presentations. In this Thesis, I compiled the largest rare variant data set in schizophrenia to date, meta-analysing the whole-exome sequences of 1,077 schizophrenia trios, 4,268 cases, and 9,343 matched controls. Using these data, I implicated at genome-wide significance the first gene, *SETD1A*, for which loss of function (LoF) variants conferred substantial risk for schizophrenia (OR $>$ 4), an adult-onset neuropsychiatric disorder (Figure 5.1). Intriguingly, the ten schizophrenia individuals with *SETD1A* disrupted had some degree of cognitive impairment, and LoF variants in the same gene were also found to confer risk for severe developmental disorders with highly variable presentation. *SETD1A* encodes a histone methyltransferase that catalysed the mono-, di-, and trimethylation of histone H3-K4, and loss-of-function mutations in the family of H3-K4 histone methyltransferase cause Mendelian conditions characterized by intellectual disability and developmental delay (e.g.. *KMT2A* and *KMT2D* are highly penetrant for Wiedenmann-Steiner Syndrome and Kabuki's syndrome, respectively). These results implicate epigenetic regulation, specifically histone modification, as a mechanism in the pathogenesis of schizophrenia, and suggest that rare risk alleles may potentially be shared between schizophrenia and broader neurodevelopmental disorders.

To better understand if the findings relating to *SETD1A* can be extended and generalized to a larger number of rare schizophrenia risk variants, I performed a series of analyses that explored the potential overlap of genetic risk between schizophrenia and broader develop-

mental disorders. I jointly analysed the trio and case-control exome data set with array-based CNV calls from 6,882 cases and 11,255 controls, and found that individuals with schizophrenia carried a significantly higher burden of rare damaging variants in 3,488 genes with a near-complete depletion of truncating variants across all variant types. This concentration of risk alleles in constrained genes was previously observed in autism, intellectual disability, and severe developmental disorders. I then performed rare variant enrichment analyses in 1,766 gene sets, and found that the rare variant burden was most strongly enriched in known autism risk genes, and genes diagnostic of severe developmental disorders. This result was significant even after controlling for the baseline enrichment in genes depleted of truncating variants. Finally, in a subset of schizophrenia patients with intellectual disability, I showed that this burden is even stronger than in the general schizophrenia population, mirroring previous results comparing autism individuals with and without cognitive impairment. Combined, these results demonstrate that schizophrenia risk loci of large effect across a range of variant types implicate a common set of genes shared with broader neurodevelopmental disorders, suggesting a path forward in identifying additional risk genes in psychiatric disorders and further supporting a neurodevelopmental etiology to the pathogenesis of schizophrenia.

5.2 Limitations of results described in this Thesis

5.2.1 Limitations in the interpretation of protein-coding consequences

Here, I discuss a number of limitations and caveats that are important when discussing the generalisability of the results in my Thesis, and are helpful in placing my assertions in context. First, the protocol used to prioritise rare coding variation in this Thesis is not optimal, and likely has a detrimental effect on power for gene discovery. During the process of variant annotation, I applied the Variant Effect Predictor tool to assign coding consequences to each variant while using the GENCODE transcript database as reference. I then annotated variants based on the most severe consequence on any transcript. However, most genes have more than one transcript or isoform, and these transcripts can be tissue-specific, expressed at particular time-points, and perform different functions. Despite the emergence of large gene expression studies such as GTeX and BrainSpan [191, 193], our catalog of gene transcripts remain incomplete. Most experiments still perform transcript quantification on bulk tissue, limiting our understanding of transcript abundance in different cell types. Furthermore, short-read RNA-seq technology has severe limitations when used to reconstruct full-length transcripts, and because of this, relevant transcripts remain missing or others are falsely included in public databases. Lastly, certain tissues, like the developing human brain, cannot

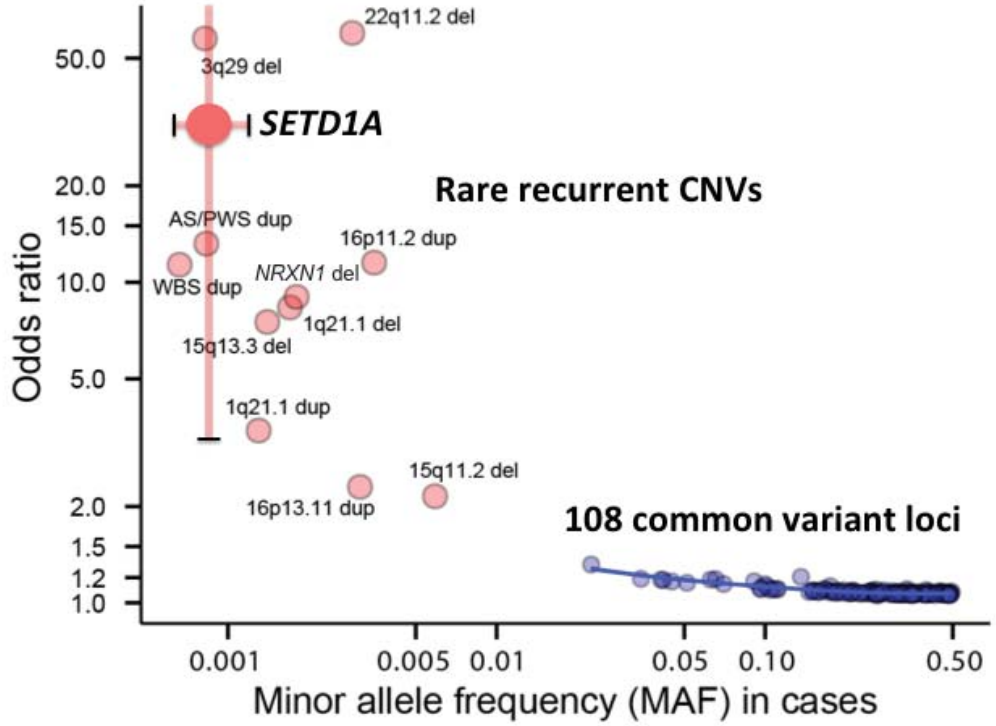


Fig. 5.1 **Risk variants for schizophrenia, with *SETD1A* included.** The effect size of each genome-wide significant risk variant for schizophrenia, as described in Ripke *et al.* and Rees *et al.*, were plotted against its allele frequency in cases [57, 67].

be easily or ethically accessed, which further limits our understanding of spatial-temporal abundance of each gene transcript.

Beyond the limitations of existing transcript references, current annotation protocols are not well-suited for handling variants that could have a specific consequence for one transcript and a conflicting consequence for another transcript. A recent study compared the concordance in annotation when different transcript set and software packages were used to predict the coding consequence of 80 million variants [136]. Surprisingly, there was only a 44% concordance in annotations for putative LoF variants when using the RefSeq and Ensembl transcript sets, and a concordance of 65% was observed when comparing LoF predictions from the VEP and ANNOVAR pipelines. Clearly, the choice of transcript reference and pipeline has a significant influence on the downstream analysis of whole-exome sequencing data. These limitations lead to the exclusion of real pathogenic variants, and further dilute our case-control analyses with large numbers of non-functional variants, all of which affect the power for gene discovery. Ultimately, we need to improve the quality of transcript reference databases, and include both abundance and spatial-temporal information of individual isoforms when annotating variants in future studies.

5.2.2 Insufficient standardisation of clinical data

Second, the limited and variable quality of the clinical data in the studies discussed in this Thesis prevented me from drawing robust connections between rare variation in schizophrenia patients and specific clinical features, such as cognitive impairment and congenital malformations. For instance, *SETD1A* belonged to a family of methyltransferases that when disrupted resulted in severe developmental disorders with a range of cognitive and physical co-morbidities. However, I could not acquire cognitive data for the vast majority of the 4,264 schizophrenia cases, and very variable clinical data was available for the ten *SETD1A* LoF carriers. While the ten carriers appeared to have some degree of cognitive impairment, this is purely a descriptive statement; insufficient clinical data was available to statistically compare this observation with the remainder of the schizophrenia data set. Furthermore, a number of these carriers only had information related to schizophrenia status, and it is quite possible that these individuals had additional co-morbidities such as seizures, facial dysmorphology and developmental delay.

On a similar thread, I identified an enrichment of rare damaging variants in developmental disorder and autism genes, but was unable to acquire the appropriate phenotypic information to determine if carriers of these LoF variants represented a distinct population of patients. Schizophrenia carriers of LoF variants in *CHD8*, an autism risk gene, could potentially have autistic features in addition to psychosis [105]. Furthermore, the disruption of specific

developmental disorder genes, such as *KMT2A* and *KMT2D*, causes characteristic facial and physical dysmorphology [157, 118], and our data did not allow us to determine if this were true of the carriers in the schizophrenia data set. Ultimately, the lack of high-quality and standardised clinical data accompanying large-scale genetic data is a severe limitation in our current study, and of association studies of psychiatric traits moving forward. Comprehensive phenotyping would be required to investigate whether carriers of rare LoF variants in constrained genes represented a distinct population of patients when compared to the remaining cases, or if these variants were associated with patterns in age-of-onset, pre-morbid impairment, neurological co-morbidities, relapse, and severity.

5.2.3 Limitations in the definition of the constrained gene list

The enrichment of rare risk variants in constrained genes is among the most striking results in early whole-exome sequencing studies of psychiatric and neurodevelopmental disorders. However, the definition of genic constraint, or the probability of loss-of-function intolerant (pLI), has caveats that need considered when interpreting the significance of our results. First, the pLI score was calculated using an expectation-maximization algorithm that assigned genes to one of three categories: null (in which LoF variants is completely tolerated), recessive (in which homozygous LoF variants is not tolerated), and haploinsufficient (in which a single copy loss is not tolerated). Genes above an arbitrary probability threshold of 0.9 were described as loss-of-function intolerant. From this definition, it is clear that the power to assign a gene to one of these categories is highly dependent on gene length; longer genes would have a greater number of expected loss-of-function variants, enabling more robust estimates of LoF depletion. Despite that notable size of the ExAC study, there may not be sufficient observations of rare LoF variants in smaller genes to detect a deviation from expectation, and these genes may have a pLI score less than 0.9 and defined as unconstrained for this reason. For example, *ARX* is a gene in which LoF variants cause severe mental retardation [157, 118], but its pLI score was estimated to be 0.74 because a 4:0 expected-to-observed LoF variant ratio was insufficient for estimating genic constraint. Therefore, a gene's ranking along the distribution of pLI score is highly dependent on statistical power that is a property of the gene sequence, and the metric itself is not a valid proxy for the strength of selection or degree of constraint. Furthermore, as described earlier in this Section, most genes have a number of transcripts that are sometimes regulated in a tissue or time-specific manner with varying functionality, and this is ignored during the modelling of genic constraint. LoF variants across all transcripts of a gene were aggregated during the calculation of pLI, and it is conceivable that LoF variants in certain transcripts are benign while in others, it is severely pathogenic. Over-simplifying the question of annotation could result in the

assignment of a gene to the unconstrained category when it is actually haploinsufficient in one isoform. Finally, the 45,376 exomes used to estimate the depletion of LoF variants were aggregated from studies that include cases diagnosed with complex disorders. While exomes of individuals with psychiatric disorders were explicitly excluded, the ExAC study had an increased incidence of autoimmune disorders, metabolic syndrome, Type II diabetes, and cardiovascular disorders [112], and risk genes for these conditions would have biased pLI scores that would be lower than expected for the general population . Given the many limitations of the ExAC constraint metric, the list of constrained genes is incomplete and imperfect, and should be treated as so. It is a rough but relatively effective tool in identifying a set of genes that are likely to be haploinsufficient in the genome. The significant enrichment of rare variants in constrained genes should be interpreted as simply an indication that there exists a large number of genes that carry rare LoF variants that substantially increase risk for psychiatric and neurodevelopmental disorders in the genome. Given this enrichment, the next step is to identify these genes with a scale-up in sample size of whole-exome sequencing studies.

5.2.4 Interpretation and generalisability of gene set results

A core set of biological processes had been implicated from gene set enrichment analyses of rare risk variants, including histone methylation, neuronal signalling pathways, and components of the post-synaptic density. However, these results come with limitations and caveats, and must be interpreted in context. First, the gene sets described in public databases originated from a variety of sources with varying methods of ascertainment. For example, the Gene Ontology database curated information from over 100,000 peer-reviewed papers that modelled biological function in a range of cell types, tissues, developmental time-points, and model organisms. The biological assay, method of sample extraction, and threshold for statistical significance likely varied between these studies. These sources of variability influenced the list of genes assigned to a single biological process, which then affects the interpretation of a gene set enrichment result. One example of this is the definition of *FMRP* targets used in autism, schizophrenia, and intellectual disability studies [103, 98, 105]. *FMRP* is a protein believed to be involved in synaptic plasticity through translational regulation, inhibiting protein synthesis through binding to mRNA. Two studies had identified the translational targets of *FMRP* in independent experiments [183, 184], and surprisingly, there was little overlap between the two gene lists. Only one of the lists from Darnell *et al.* showed a significant signal in schizophrenia and autism analyses, while no signal was observed using the Ascano list [103, 98, 105]. The precise reason for the discrepancy between the two studies remained unknown, but it was suspected the choice of

cell type may be the source of the difference: the Darnell study looked for targets in mouse brain tissue, while the Ascano study identified targets in a human embryonic kidney cell line. However, this also meant that the enrichment of rare variants in *FMRP* targets from the Darnell study could originate from an over-representation of brain genes. These issues make it difficult for us to generalise the insights of gene set enrichment analyses to something biologically relevant for schizophrenia.

Furthermore, I observed substantial overlap between the gene sets enriched for schizophrenia risk variants, which also make it difficult to draw specific insights from our burden results. The 1,766 gene sets used in our analysis, and the 35 FDR < 5% gene sets were notably enriched with constrained genes when compared to compared to a random sampling of genes from the genome (Figure 5.2, 5.3). For example, 67% of the Darnell *et al. FMRP* gene targets and 74% of the DDG2P developmental disorder genes were constrained. After restricting our analyses to constrained genes, only developmental disorder and autism risk genes remained significantly enriched for schizophrenia risk variants. I could not differentiate if the other results were biologically significant, and not due to an statistical over-sampling of constrained genes. Therefore, given the size of the tested gene sets and the substantial overlap between them, it is difficult to draw conclusions about specific pathways and mechanisms in the pathogenesis of schizophrenia. To gain meaningful insight into the neurobiology of schizophrenia, we ought to move beyond gene set analyses and focus on identifying individual genes such as *SETD1A* at genome-wide significance, and follow-up each one of those genes to elucidate the mechanisms underlying schizophrenia pathogenesis.

5.3 Future directions

5.3.1 Whole-genome sequencing at the population scale

Recent studies have made significant progress in advancing our understanding of the genetics of schizophrenia. These results come from independent studies investigating select aspects of schizophrenia's genetic architecture, with SNP genotyping identifying large numbers of common variants, array-based CNVs implicating large effect CNVs, and whole-exome sequencing demonstrating a burden of rare variants. Based on the results from the past decade, the path forward for identifying risk alleles for schizophrenia is clear. Additional samples will be genotyped using arrays in ever larger numbers, and imputation using the Haplotype Reference Consortium panel will enable the identification of risk variants with minor allele frequencies as low as 0.1% [197]. Already, the Psychiatric Genetics Consortium has plans to massively scale up its GWAS efforts for a number of psychiatric disorders [198].

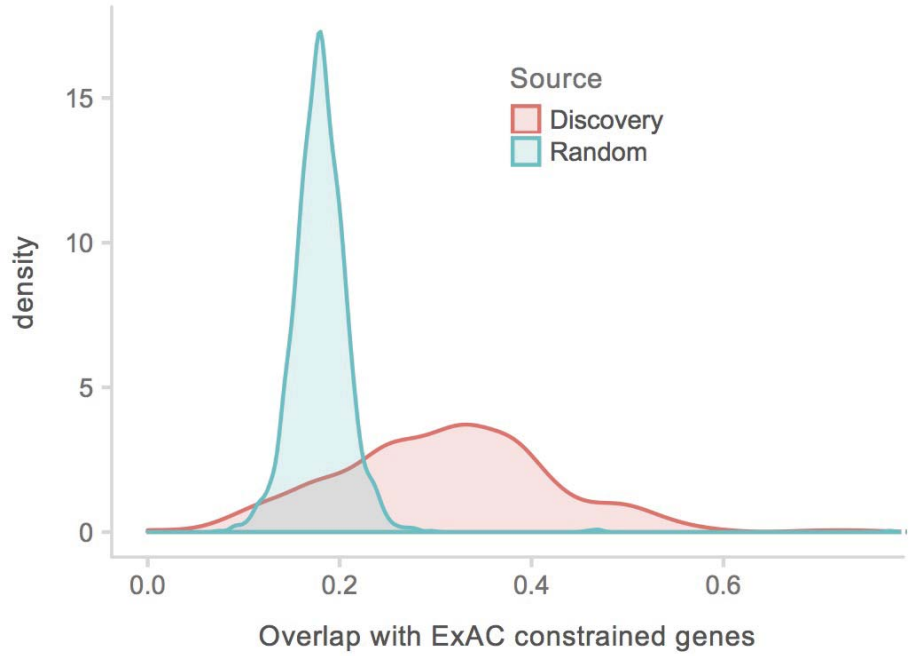


Fig. 5.2 **Distribution of overlap coefficients with the constrained gene set.** The overlap coefficients between each of the 1,766 discovery gene sets described in Chapter 4 and the constrained gene set were calculated. Random gene sets were sampled from the genome with the same size distribution as the discovery gene sets, and their overlap coefficients with the constrained gene set were also computed. I plotted these values as a density plot. The overlap coefficient is a similarity measure defined as $\frac{|X \cap Y|}{\min(|X|,|Y|)}$, where X and Y are sets of genes.

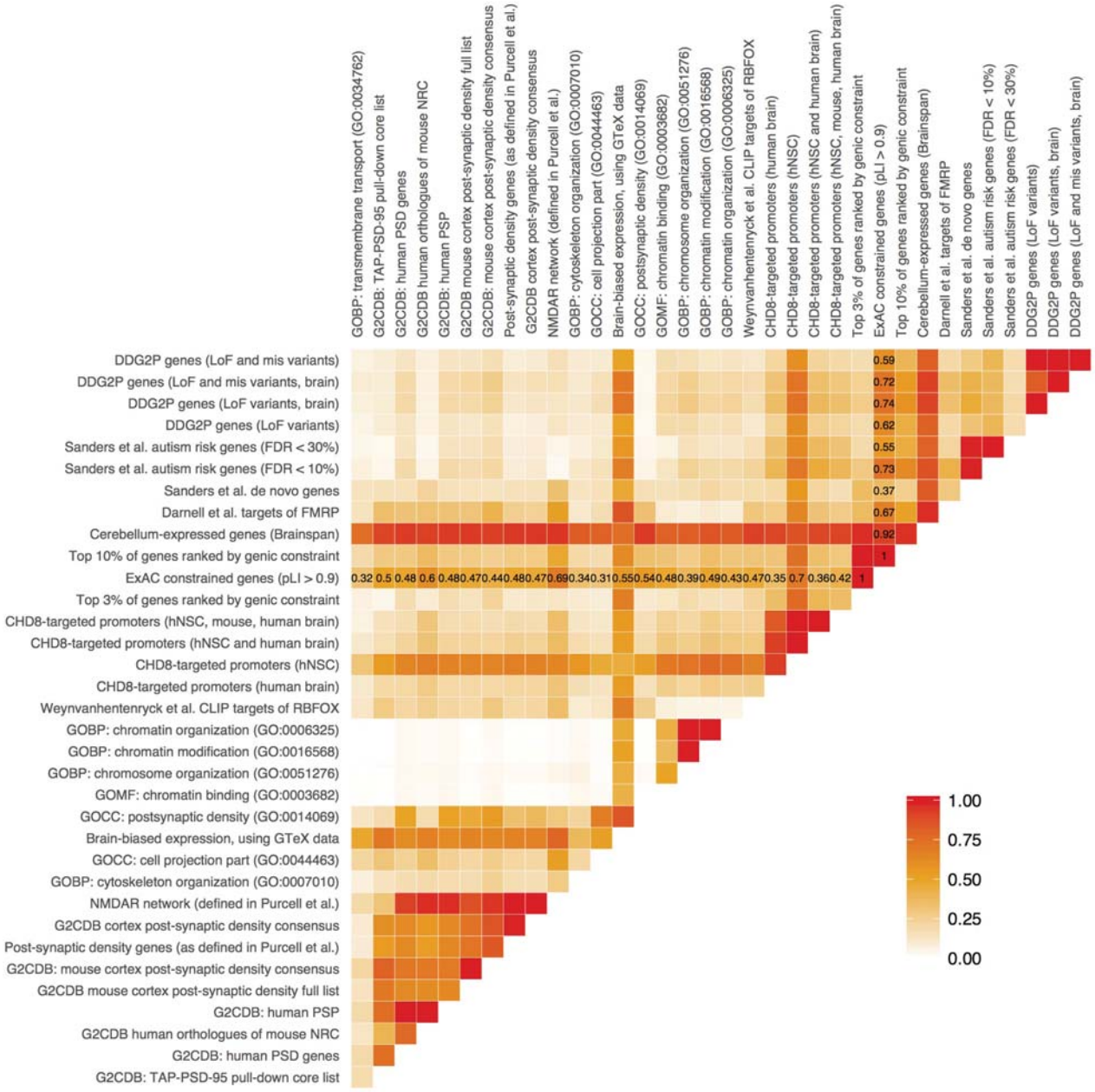


Fig. 5.3 **Heatmap of overlap coefficients calculated between FDR $< 5\%$ gene sets.** The overlap coefficients of gene sets enriched for rare coding variants conferring risk for schizophrenia were computed, clustered, and displayed as a heatmap. The overlap coefficient is a similarity measure defined as $\frac{|X \cap Y|}{\min(|X|,|Y|)}$, where X and Y are sets of genes. I also provided the overlap coefficients between each gene set and the constrained gene set as a rounded decimal in the Figure.

Separately, to identify novel risk genes based on rare coding variants, tens of thousands whole-exome sequences will be produced and analysed, leveraging both *de novo* mutations from a trio design and inherited variants from a case-control design. A *de novo* approach for gene discovery will be less helpful in discovering risk variants that can be inherited without a substantial decrease in fitness. New methods will be needed to identify groups of risk alleles that have moderate penetrance, and this may include leveraging genetic resources such as the Exome Aggregation Consortium to exclude neutral variants using allele frequency estimates from hundreds of thousands of exomes to increase power. Despite the clear overlap between the variant types, there is little integration in current studies of common SNPs, and rare CNVs, and SNVs for gene discovery. To produce a complete picture of the genetics of schizophrenia, whole-genome sequencing is best positioned to study the interplay of common and rare variants in the same individual. Whole-genome sequences from many thousands of schizophrenia individuals will be integrated with whole-exome sequencing data and array-based data to produce a complete picture of schizophrenia's genetic architecture, improve risk stratification, and refine clinical diagnoses.

5.3.2 Specificity of shared risk alleles for individual psychiatric disorders

An overlapping set of genes appear to be disrupted by *de novo* mutations in autism, severe developmental disorders, intellectual disability, and now schizophrenia. A number of these shared risk genes have been identified, and all of them are depleted of protein-truncating variants in the general population. A single-copy loss of these neurodevelopmental disorder genes, including *ARID1B*, *CHD8*, and *POGZ*, increases risk for a range of syndromic features in addition to cognitive impairment and autism [105, 118]. While cognitive impairment is co-morbid with schizophrenia and autism to varying degrees, the relative risk of a disruptive variant in these genes for each clinical diagnosis has not been robustly estimated, and it remains unclear if these genes preferentially confer risk for a subset of neurodevelopmental phenotypes. Determining the relative penetrance of these shared risk alleles is important for refining clinical diagnoses and inferring meaningful and specific biology for individual neurodevelopmental and psychiatric disorders. To model the relative risks of genes for neurodevelopmental disorders, we will need to compare and contrast the tens of thousands of whole-exomes generated by different consortia, including the Autism Sequencing Consortium, the DDD study, and other schizophrenia sequencing efforts. Since these variants are extremely rare in the population, very large data sets will be required to identify sufficient numbers of carriers to make robust inferences on individual phenotypes. For instance, only 16

SETD1A carriers were observed in over 30,000 exomes analysed in our study, and screening tens of thousands of schizophrenia patients will be necessarily to accurately estimate its penetrance for cognitive and neurodevelopmental outcomes [119]. However, comprehensive and comparable clinical data across all these data sets, and not absolute sample size, will be the limited factor for this type of analysis. Existing whole-exome sequencing data sets jointly analysed a number of smaller, highly heterogeneous clinical cohorts with incomplete phenotypic data, which prevented a comprehensive analysis of co-morbid symptoms. Furthermore, existing autism and intellectual disability studies have very specific ascertainment criteria, with many focusing on simplex, sporadic cases that are generally more severe than other individuals that may share diagnoses of autism and cognitive impairment, and this limits our ability to generalise estimates of relative risk for the larger population of potential carriers. Despite these challenges, elucidating the disease specificity of highly penetrant syndromic variants remains an important task, as we ultimately want to characterise the phenotypic spectrum of these genes for clinical diagnosis, genetic counselling, and the discovery of new disease biology.

5.3.3 *In vitro* and *in vivo* modeling of risk genes for neurodevelopmental disorders

Because *SETD1A* is involved in chromatin modification and regulates the transcription of a number of unknown genes, the precise biological consequences of haploinsufficiency in this gene remain difficult to predict without well-designed functional assays. This is reminiscent of the autism risk genes identified in trio studies, which are also involved in global processes such as chromatin modification and global transcriptional regulation. One example of such a gene is *CHD8*, an ATP-dependent chromatin remodeler that increases risk for intellectual disability, autism, gastrointestinal abnormalities, and other syndromic features [199]. A single-copy loss of *CHD8* was predicted to dysregulate critical pathways and networks of genes associated with neurodevelopment. Two functional studies used RNA-seq and ChIP-seq to identify binding sites for *CHD8* and the downstream genes it directly and indirectly regulates [178, 187]. Genes downregulated by *CHD8* implicated pathways involved in synapse formation, neuron differentiation, and axon guidance. Furthermore, *CHD8*-bound and *CHD8*-downregulated genes were strongly enriched for autism risk genes. An *in vivo* zebrafish model of *CHD8* recapitulated physical features present in human carriers, including macrocephaly and impairment of gastrointestinal motility [199]. Similarly, mice with a single copy loss of *SHANK3*, another high-penetrant autism gene, exhibited repetitive grooming habits, deficits in social interaction, and defects in striatal synapses [200].

Therefore, functional studies will prove to be an invaluable tool in elucidating the mechanism by which these genes increase risk for neurodevelopmental disorders.

Here, I briefly discuss functional experiments designed to elucidate the biological processes disrupted by a single copy of *SETD1A*. A comprehensive discussion of the technical details of these experiments are beyond the scope of this Thesis, and admittedly, there are many caveats and limitations when modelling disease in model organisms and cellular systems. I would also like to emphasise that the experiments described in this Section will be performed in collaboration with research groups who have expertise in addressing the many technical challenges in play. First, as a proof-of-concept study, we have developed a mouse model of *SETD1A* in which the entirety of exon-2 is deleted to recapitulate the heterozygous loss-of-function genotype observed in human carriers. We plan on conducting three categories of experiments to understand the precise function of this genes. First, we will deeply phenotype the *SETD1A*-heterozygous mouse to understand differences in behaviour and deficits in the cognitive dimension. Abnormal behaviour in schizophrenia is highly complex and heterogeneous, and include symptoms in the positive, negative, and cognitive dimensions. A number of assays have been developed to determine the severity of each cluster of symptoms [201]; however, other than the 22q11.2 deletion mouse model, no valid genetic model for schizophrenia exists [202, 203]. We will compare and contrast observations of the *SETD1A* mice with existing mouse models of schizophrenia to determine if there is a consistent pattern of behavioural abnormalities that emerge. Second, schizophrenia is associated with a number of morphological differences in the human brain. Using histology and MRI, we hope to pinpoint neuroanatomical abnormalities to specific regions in order to determine differences in brain development that arise from *SETD1A* haploinsufficiency. Neuroanatomical abnormalities in mice can serve as a good first step to narrow down relevant cell types and tissues for *in vitro* experiments in human cells. Finally, we will extract brain tissue and leverage RNA-seq and ChIP-seq to identify *SETD1A*-bound regions and *SETD1A*-targeted genes. H3-K4 methyltransferases like *SETD1A* open previously closed chromatin and are responsible for transcriptional activation across the genome. *SETD1A* haploinsufficiency likely results in differential methylation and consequently differential transcription at specific regions across the genome. This dysregulation of downstream genes might be linked to the disease phenotype. We will additionally profile the transcriptomic and epigenetics dynamics in different parts of the brain (e.g. hippocampus or the prefrontal cortex) to identify tissue-specific consequences. Using these data, we hope to find biological processes and co-expression networks that may be relevant to schizophrenia or neurodevelopment.

We have also engineered two LoF variants in *SETD1A* into human induced pluripotent stem cell (iPSC) lines. We will use a combination of RNA-seq and ChIP-seq to characterize

transcriptomic and epigenetic changes in neuronal progenitors. This analysis will provide a complementary set of differentially methylated regions and differentially expressed genes, and results from the mouse model and the iPSC line will be compared and contrasted. We will test these gene sets for enrichment in common schizophrenia risk loci, intellectual disability and autism risk genes. In summary, *in vitro* and *in vivo* models of highly penetrant rare variants have proven useful in studying genes for neurodevelopmental disorders, and will likely be applied to many more risk genes in the future to advance our understanding of the disease mechanisms underlying autism, intellectual disability, and schizophrenia.

5.4 Concluding remarks

It is truly an exciting time for the field of psychiatric genetics. The past two decades have seen the identification of the first robust genetic risk factors, the validation of the polygenic model, the demonstration of genetic sharing between neurodevelopmental and psychiatric disorders, and increasing support for a number of hypotheses on disease mechanism. The path forward to uncovering the varied and complex genetic contributions is clearer than ever. In time, whole-genome sequencing will discover an ever-increasing number of common and rare genetic risk factors and provide a complete picture of the genetic architecture of psychiatric disorders. This comprehensive map of genetic risk factors will serve as the foundation of functional studies that seek to elucidate the mechanisms underlying disease pathogenesis, and reveal valid and meaningful therapeutic targets that may lead to more effective treatments. Furthermore, robust genetic markers will improve clinical practice by increasing diagnostic accuracy and informing more useful diagnostic categories and dimensions for these heterogeneous conditions. These advances, along with societal efforts to provide increased support and reduce social stigma, will hopefully improve the quality of lives of the many people profoundly affected by mental illness.

References

[1] G. Harrison, K. Hopper, T. Craig, E. Laska, C. Siegel, J. Wanderling, K. C. Dube, K. Ganev, R. Giel, W. an der Heiden, S. K. Holmberg, A. Janca, P. W. Lee, C. A. León, S. Malhotra, A. J. Marsella, Y. Nakane, N. Sartorius, Y. Shen, C. Skoda, R. Thara, S. J. Tsirkin, V. K. Varma, D. Walsh, and D. Wiersma, "Recovery from psychotic illness: a 15- and 25-year international follow-up study.," *The British journal of psychiatry : the journal of mental science*, vol. 178, pp. 506–17, jun 2001.

[2] D. G. Robinson, M. G. Woerner, M. McMeniman, A. Mendelowitz, and R. M. Bilder, "Symptomatic and functional recovery from a first episode of schizophrenia or schizoaffective disorder.," *The American journal of psychiatry*, vol. 161, pp. 473–9, mar 2004.

[3] A. Barbato, "Psychiatry in transition: outcomes of mental health policy shift in Italy.," *The Australian and New Zealand journal of psychiatry*, vol. 32, pp. 673–9, oct 1998.

[4] M. J. Owen, A. Sawa, and P. B. Mortensen, "Schizophrenia," *The Lancet*, vol. 6736, no. 15, pp. 1–12, 2016.

[5] J. McGrath, S. Saha, D. Chant, and J. Welham, "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality," *Epidemiologic Reviews*, vol. 30, pp. 67–76, may 2008.

[6] S. Marwaha, S. Johnson, P. Bebbington, M. Stafford, M. C. Angermeyer, T. Brugha, J.-M. Azorin, R. Kilian, K. Hansen, and M. Toumi, "Rates and correlates of employment in people with schizophrenia in the UK, France and Germany.," *The British journal of psychiatry : the journal of mental science*, vol. 191, pp. 30–7, jul 2007.

[7] T. Burns, J. Catty, T. Becker, R. E. Drake, A. Fioritti, M. Knapp, C. Lauber, W. Rössler, T. Tomov, J. van Busschbach, S. White, D. Wiersma, and EQOLISE Group, "The effectiveness of supported employment for people with severe mental illness: a randomised controlled trial.," *Lancet (London, England)*, vol. 370, pp. 1146–52, sep 2007.

[8] S. Saha, D. Chant, and J. McGrath, "A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time?," *Archives of general psychiatry*, vol. 64, pp. 1123–31, oct 2007.

[9] A. H. Crisp, M. G. Gelder, S. Rix, H. I. Meltzer, and O. J. Rowlands, "Stigmatisation of people with mental illnesses.," *The British journal of psychiatry : the journal of mental science*, vol. 177, pp. 4–7, jul 2000.

[10] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (DSM-5{®})*. Arlington, VA: American Psychiatric Publishing, 2013.

[11] Adityanjee, Y. A. Aderibigbe, D. Theodoridis, and V. R. Vieweg, "Dementia praecox to schizophrenia: the first 100 years.," *Psychiatry and clinical neurosciences*, vol. 53, pp. 437–48, aug 1999.

[12] A. Jablensky, "The 100-year epidemiology of schizophrenia," *Schizophrenia Research*, vol. 28, no. 2-3, pp. 111–125, 1997.

[13] R. Tandon, W. Gaebel, D. M. Barch, J. Bustillo, R. E. Gur, S. Heckers, D. Malaspina, M. J. Owen, S. Schultz, M. Tsuang, J. Van Os, and W. Carpenter, "Definition and description of schizophrenia in the DSM-5," *Schizophrenia Research*, vol. 150, no. 1, pp. 3–10, 2013.

[14] V. a. Morgan, H. Leonard, J. Bourke, and A. Jablensky, "Intellectual disability co-occurring with schizophrenia and other psychiatric illness: population-based study.," *The British journal of psychiatry : the journal of mental science*, vol. 193, pp. 364–72, nov 2008.

[15] M. J. Owen, "New approaches to psychiatric diagnostic classification," *Neuron*, vol. 84, no. 3, pp. 564–571, 2014.

[16] J. Nordgaard, S. M. Arnfred, P. Handest, and J. Parnas, "The diagnostic status of first-rank symptoms," *Schizophrenia Bulletin*, vol. 34, no. 1, pp. 137–154, 2008.

[17] J. Perälä, J. Suvisaari, S. I. Saarni, K. Kuoppasalmi, E. Isometsä, S. Pirkola, T. Partonen, A. Tuulio-Henriksson, J. Hintikka, T. Kieseppä, T. Härkänen, S. Koskinen, and J. Lönnqvist, "Lifetime prevalence of psychotic and bipolar I disorders in a general population.," *Archives of general psychiatry*, vol. 64, pp. 19–28, jan 2007.

[18] A. J. Rothschild, "Challenges in the treatment of major depressive disorder with psychotic features," *Schizophrenia Bulletin*, vol. 39, no. 4, pp. 787–796, 2013.

[19] J. van Os, R. J. Linscott, I. Myin-Germeys, P. Delespaul, and L. Krabbendam, "A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness-persistence-impairment model of psychotic disorder.," *Psychological medicine*, vol. 39, no. 2, pp. 179–95, 2009.

[20] P. F. Buckley, B. J. Miller, D. S. Lehrer, and D. J. Castle, "Psychiatric comorbidities and schizophrenia.," *Schizophrenia bulletin*, vol. 35, pp. 383–402, mar 2009.

[21] S. Leucht, C. Corves, D. Arbter, R. R. Engel, C. Li, and J. M. Davis, "Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis.," *Lancet (London, England)*, vol. 373, pp. 31–41, jan 2009.

[22] H. Y. Meltzer, "Update on Typical and Atypical Antipsychotic Drugs," *Annual Review of Medicine*, vol. 64, no. 1, p. 120928131129008, 2012.

[23] C. Rummel-Kluge, K. Komossa, S. Schwarz, H. Hunger, F. Schmid, C. A. Lobos, W. Kissling, J. M. Davis, and S. Leucht, "Head-to-head comparisons of metabolic side effects of second generation antipsychotics in the treatment of schizophrenia: A systematic review and meta-analysis," *Schizophrenia Research*, vol. 123, no. 2-3, pp. 225–233, 2010.

[24] J. A. Lieberman, T. S. Stroup, J. P. McEvoy, M. S. Swartz, R. A. Rosenheck, D. O. Perkins, R. S. E. Keefe, S. M. Davis, C. E. Davis, B. D. Lebowitz, J. Severe, J. K. Hsiao, and Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Investigators, "Effectiveness of antipsychotic drugs in patients with chronic schizophrenia.," *The New England journal of medicine*, vol. 353, pp. 1209–23, sep 2005.

[25] M. DE Hert, V. Schreurs, D. Vancampfort, and R. VAN Winkel, "Metabolic syndrome in people with schizophrenia: a review.," *World psychiatry : official journal of the World Psychiatric Association (WPA)*, vol. 8, no. 1, pp. 15–22, 2009.

[26] J. B. Kirkbride, P. Fearon, C. Morgan, P. Dazzan, K. Morgan, R. M. Murray, and P. B. Jones, "Neighbourhood variation in the incidence of psychotic disorders in Southeast London," *Social Psychiatry and Psychiatric Epidemiology*, vol. 42, no. 6, pp. 438–445, 2007.

[27] D. Castle, P. Sham, and R. Murray, "Differences in distribution of ages of onset in males and females with schizophrenia," *Schizophrenia Research*, vol. 33, no. 3, pp. 179–183, 1998.

[28] T. K. Rajji, Z. Ismail, and B. H. Mulsant, "Age at onset and cognition in schizophrenia: meta-analysis.," *The British journal of psychiatry : the journal of mental science*, vol. 195, pp. 286–93, oct 2009.

[29] H. Hafner, K. Maurer, W. Loffler, and A. Riecher-Rossler, "The influence of age and sex on the onset and early course of schizophrenia," *Br.J Psychiatry*, vol. 162, no. MAY, pp. 80–86, 1993.

[30] C. J. L. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. a. Salomon, S. Abdalla, V. Aboyans, J. Abraham, I. Ackerman, R. Aggarwal, S. Y. Ahn, M. K. Ali, M. Alvarado, H. R. Anderson, L. M. Anderson, K. G. Andrews, C. Atkinson, L. M. Baddour, A. N. Bahalim, S. Barker-Collo, L. H. Barrero, D. H. Bartels, M.-G. Basáñez, A. Baxter, M. L. Bell, E. J. Benjamin, D. Bennett, E. Bernabé, K. Bhalla, B. Bhandari, B. Bikbov, A. Bin Abdulhak, G. Birbeck, J. a. Black, H. Blencowe, J. D. Blore, F. Blyth, I. Bolliger, A. Bonaventure, S. Boufous, R. Bourne, M. Boussinesq, T. Braithwaite, C. Brayne, L. Bridgett, S. Brooker, P. Brooks, T. S. Brugha, C. Bryan-Hancock, C. Bucello, R. Buchbinder, G. Buckle, C. M. Budke, M. Burch, P. Burney, R. Burstein, B. Calabria, B. Campbell, C. E. Canter, H. Carabin, J. Carapetis, L. Carmona, C. Cella, F. Charlson, H. Chen, A. T.-A. Cheng, D. Chou, S. S. Chugh, L. E. Coffeng, S. D. Colan, S. Colquhoun, K. E. Colson, J. Condon, M. D. Connor, L. T. Cooper, M. Corriere, M. Cortinovis, K. C. de Vaccaro, W. Couser, B. C. Cowie, M. H. Criqui, M. Cross, K. C. Dabhadkar, M. Dahiya, N. Dahodwala, J. Damsere-Derry, G. Danaei, A. Davis, D. De Leo, L. Degenhardt, R. Dellavalle, A. Delossantos, J. Denenberg, S. Derrett, D. C. Des Jarlais, S. D. Dharmaratne, M. Dherani, C. Diaz-Torne, H. Dolk, E. R. Dorsey,

T. Driscoll, H. Duber, B. Ebel, K. Edmond, A. Elbaz, S. E. Ali, H. Erskine, P. J. Erwin, P. Espindola, S. E. Ewoigbokhan, F. Farzadfar, V. Feigin, D. T. Felson, A. Ferrari, C. P. Ferri, E. M. Fèvre, M. M. Finucane, S. Flaxman, L. Flood, K. Foreman, M. H. Forouzanfar, F. G. R. Fowkes, M. Fransen, M. K. Freeman, B. J. Gabbe, S. E. Gabriel, E. Gakidou, H. a. Ganatra, B. Garcia, F. Gaspari, R. F. Gillum, G. Gmel, D. Gonzalez-Medina, R. Gosselin, R. Grainger, B. Grant, J. Groeger, F. Guillemin, D. Gunnell, R. Gupta, J. Haagsma, H. Hagan, Y. a. Halasa, W. Hall, D. Haring, J. M. Haro, J. E. Harrison, R. Havmoeller, R. J. Hay, H. Higashi, C. Hill, B. Hoen, H. Hoffman, P. J. Hotez, D. Hoy, J. J. Huang, S. E. Ibeanusi, K. H. Jacobsen, S. L. James, D. Jarvis, R. Jasrasaria, S. Jayaraman, N. Johns, J. B. Jonas, G. Karthikeyan, N. Kassebaum, N. Kawakami, A. Keren, J.-P. Khoo, C. H. King, L. M. Knowlton, O. Kobusingye, A. Koranteng, R. Krishnamurthi, F. Laden, R. Lalloo, L. L. Laslett, T. Lathlean, J. L. Leasher, Y. Y. Lee, J. Leigh, D. Levinson, S. S. Lim, E. Limb, J. K. Lin, M. Lipnick, S. E. Lipshultz, W. Liu, M. Loane, S. L. Ohno, R. Lyons, J. Mabweijano, M. F. MacIntyre, R. Malekzadeh, L. Mallinger, S. Manivannan, W. Marcenes, L. March, D. J. Margolis, G. B. Marks, R. Marks, A. Matsumori, R. Matzopoulos, B. M. Mayosi, J. H. McAnulty, M. M. McDermott, N. McGill, J. McGrath, M. E. Medina-Mora, M. Meltzer, G. a. Mensah, T. R. Merriman, A.-C. Meyer, V. Miglioli, M. Miller, T. R. Miller, P. B. Mitchell, C. Mock, A. O. Mocumbi, T. E. Moffitt, A. a. Mokdad, L. Monasta, M. Montico, M. Moradi-Lakeh, A. Moran, L. Morawska, R. Mori, M. E. Murdoch, M. K. Mwaniki, K. Naidoo, M. N. Nair, L. Naldi, K. M. V. Narayan, P. K. Nelson, R. G. Nelson, M. C. Nevitt, C. R. Newton, S. Nolte, P. Norman, R. Norman, M. O'Donnell, S. O'Hanlon, C. Olives, S. B. Omer, K. Ortblad, R. Osborne, D. Ozgediz, A. Page, B. Pahari, J. D. Pandian, A. P. Rivero, S. B. Patten, N. Pearce, R. P. Padilla, F. Perez-Ruiz, N. Perico, K. Pesudovs, D. Phillips, M. R. Phillips, K. Pierce, S. Pion, G. V. Polanczyk, S. Polinder, C. A. Pope, S. Popova, E. Porrini, F. Pourmalek, M. Prince, R. L. Pullan, K. D. Ramaiah, D. Ranganathan, H. Razavi, M. Regan, J. T. Rehm, D. B. Rein, G. Remuzzi, K. Richardson, F. P. Rivara, T. Roberts, C. Robinson, F. R. De Leòn, L. Ronfani, R. Room, L. C. Rosenfeld, L. Rushton, R. L. Sacco, S. Saha, U. Sampson, L. Sanchez-Riera, E. Sanman, D. C. Schwebel, J. G. Scott, M. Segui-Gomez, S. Shahraz, D. S. Shepard, H. Shin, R. Shivakoti, D. Singh, G. M. Singh, J. a. Singh, J. Singleton, D. a. Sleet, K. Sliwa, E. Smith, J. L. Smith, N. J. C. Stapelberg, A. Steer, T. Steiner, W. a. Stolk, L. J. Stovner, C. Sudfeld, S. Syed, G. Tamburlini, M. Tavakkoli, H. R. Taylor, J. a. Taylor, W. J. Taylor, B. Thomas, W. M. Thomson, G. D. Thurston, I. M. Tleyjeh, M. Tonelli, J. a. Towbin, T. Truelsen, M. K. Tsilimbaris, C. Ubeda, E. a. Undurraga, M. J. van der Werf, J. van Os, M. S. Vavilala, N. Venketasubramanian, M. Wang, W. Wang, K. Watt, D. J. Weatherall, M. a. Weinstock, R. Weintraub, M. G. Weisskopf, M. M. Weissman, R. a. White, H. Whiteford, N. Wiebe, S. T. Wiersma, J. D. Wilkinson, H. C. Williams, S. R. M. Williams, E. Witt, F. Wolfe, A. D. Woolf, S. Wulf, P.-H. Yeh, A. K. M. Zaidi, Z.-J. Zheng, D. Zonies, A. D. Lopez, M. a. AlMazroa, and Z. a. Memish, "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010.," *Lancet*, vol. 380, pp. 2197–223, dec 2012.

[31] J. Read, J. Van Os, A. P. Morrison, and C. A. Ross, "Childhood trauma, psychosis and schizophrenia: A literature review with theoretical and clinical implications," *Acta Psychiatrica Scandinavica*, vol. 112, no. 5, pp. 330–350, 2005.

[32] C. Morgan and H. Fisher, "Environment and schizophrenia: Environmental factors in schizophrenia: Childhood trauma - A critical review," *Schizophrenia Bulletin*, vol. 33, no. 1, pp. 3–10, 2007.

[33] M. C. Cutajar, P. E. Mullen, J. R. P. Ogloff, S. D. Thomas, D. L. Wells, and J. Spataro, "Schizophrenia and other psychotic disorders in a cohort of sexually abused children.," *Archives of general psychiatry*, vol. 67, no. 11, pp. 1114–9, 2010.

[34] L. Arseneault, M. Cannon, H. L. Fisher, G. Polanczyk, T. E. Moffi, and A. Caspi, "Childhood trauma and children ' s emerging psychotic symptoms : A genetically sensitive longitudinal cohort study," *American Journal of Psychiatry*, vol. 168, no. January, pp. 65–72, 2011.

[35] E. Susser and S. Lin, "Schizophrenia after prenatal exposure to the Dutch hunger winter of 1944–1945," *Schizophrenia Research*, vol. 9, p. 140, apr 1993.

[36] H. W. Hoek, A. S. Brown, and E. Susser, "The Dutch Famine and schizophrenia spectrum disorders," *Social Psychiatry and Psychiatric Epidemiology*, vol. 33, pp. 373–379, jul 1998.

[37] D. St Clair, "Rates of Adult Schizophrenia Following Prenatal Exposure to the Chinese Famine of 1959-1961," *JAMA*, vol. 294, p. 557, aug 2005.

[38] M.-Q. Xu, W.-S. Sun, B.-X. Liu, G.-Y. Feng, L. Yu, L. Yang, G. He, P. Sham, E. Susser, D. St. Clair, and L. He, "Prenatal Malnutrition and Adult Schizophrenia: Further Evidence From the 1959-1961 Chinese Famine," *Schizophrenia Bulletin*, vol. 35, pp. 568–576, mar 2009.

[39] M. C. Clarke, A. Tanskanen, M. Huttunen, D. A. Leon, R. M. Murray, P. B. Jones, and M. Cannon, "Increased risk of schizophrenia from additive interaction between infant motor developmental delay and obstetric complications: Evidence from a population-based longitudinal study," *American Journal of Psychiatry*, vol. 168, no. 12, pp. 1295–1302, 2011.

[40] H. Dickson, K. R. Laurens, a. E. Cullen, and S. Hodgins, "Meta-analyses of cognitive and motor function in youth aged 16 years and younger who subsequently develop schizophrenia," *Psychological Medicine*, vol. 42, no. SEPTEMBER 2011, pp. 743–755, 2012.

[41] S. Andréasson, P. Allebeck, A. Engström, and U. Rydberg, "Cannabis and schizophrenia. A longitudinal study of Swedish conscripts.," *Lancet (London, England)*, vol. 2, pp. 1483–6, dec 1987.

[42] R. M. Murray, P. D. Morrison, C. Henquet, and M. Di Forti, "Cannabis, the mind and society: the hash realities.," *Nature reviews. Neuroscience*, vol. 8, pp. 885–95, nov 2007.

[43] B. D. Kelly, E. O'Callaghan, J. L. Waddington, L. Feeney, S. Browne, P. J. Scully, M. Clarke, J. F. Quinn, O. McTigue, M. G. Morgan, A. Kinsella, and C. Larkin, "Schizophrenia and the city: A review of literature and prospective study of psychosis and urbanicity in Ireland," *Schizophrenia Research*, vol. 116, no. 1, pp. 75–89, 2010.

[44] L. Krabbendam and J. Van Os, "Schizophrenia and urbanicity: A major environmental influence - Conditional on genetic risk," *Schizophrenia Bulletin*, vol. 31, no. 4, pp. 795–799, 2005.

[45] J. Boydell, J. van Os, K. McKenzie, J. Allardyce, R. Goel, R. G. McCreadie, and R. M. Murray, "Incidence of schizophrenia in ethnic minorities in London: ecological study into interactions with environment.," *BMJ (Clinical research ed.)*, vol. 323, pp. 1336–8, dec 2001.

[46] W. Veling, E. Susser, J. Van Os, J. P. Mackenbach, J. P. Selten, and H. W. Hoek, "Ethnic density of neighborhoods and incidence of psychotic disorders among immigrants," *American Journal of Psychiatry*, vol. 165, no. 1, pp. 66–73, 2008.

[47] P. McGuffin, M. Owen, and A. Farmer, "Genetic basis of schizophrenia," *The Lancet*, vol. 346, no. 8976, pp. 678–682, 1995.

[48] a. E. Farmer, P. McGuffin, and I. I. Gottesman, "Twin concordance for DSM-III schizophrenia. Scrutinizing the validity of the definition.," *Archives of general psychiatry*, vol. 44, pp. 634–641, jul 1987.

[49] P. F. Sullivan, K. S. Kendler, and M. C. Neale, "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies.," *Archives of general psychiatry*, vol. 60, pp. 1187–92, dec 2003.

[50] L. L. Heston, "Psychiatric disorders in foster home reared children of schizophrenic mothers.," *British Journal of Psychiatry*, vol. 112, no. 489, pp. 819–825, 1966.

[51] L. J. Ingraham and S. S. Kety, "Adoption studies of schizophrenia.," *American journal of medical genetics*, vol. 97, pp. 18–22, aug 2000.

[52] I. I. Gottesman and J. Shields, "A polygenic theory of schizophrenia.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 58, no. 1, pp. 199–205, 1967.

[53] The International Haplotype Consortium, "A haplotype map of the human genome.," *Nature*, vol. 437, pp. 1299–320, oct 2005.

[54] WTCCC, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.," *Nature*, vol. 447, no. 7145, pp. 661–78, 2007.

[55] M. C. O'Donovan, N. Craddock, N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. A. Spencer, B. Howie, H.-T. Leung, A. M. Hartmann, H.-J. Möller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. M. Williams, I. Giegling, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, P. V. Gejman, S. Cichon, M. M. Nöthen, M. Gill, A. Corvin, D. Rujescu, G. Kirov, M. J. Owen, N. G. Buccola, B. J. Mowry, R. Freedman, F. Amin, D. W. Black, J. M. Silverman, W. F. Byerley, and C. R. Cloninger, "Identification of loci associated with schizophrenia by genome-wide association and follow-up.," *Nature genetics*, vol. 40, no. 9, pp. 1053–5, 2008.

[56] S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.," *Nature*, vol. 460, pp. 748–52, aug 2009.

[57] Schizophrenia Working Group of the Psychiatric Genomics Consortium, "Biological insights from 108 schizophrenia-associated genetic loci.," *Nature*, vol. 511, pp. 421–7, jul 2014.

[58] P.-R. Loh, G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler, M. C. O'Donovan, B. M. Neale, N. Patterson, and A. L. Price, "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis," *Nature Genetics*, vol. 47, pp. 1385–1392, nov 2015.

[59] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, Wellcome Trust Case Control Consortium, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome.," *Nature*, vol. 464, pp. 704–12, apr 2010.

[60] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy number variation in human health, disease, and evolution.," *Annual review of genomics and human genetics*, vol. 10, pp. 451–81, jan 2009.

[61] S. E. Folstein and B. Rosen-Sheidley, "Genetics of autism: complex aetiology for a heterogeneous disorder.," *Nature reviews. Genetics*, vol. 2, pp. 943–55, dec 2001.

[62] K. C. Murphy, L. a. Jones, and M. J. Owen, "High rates of schizophrenia in adults with velo-cardio-facial syndrome.," *Archives of general psychiatry*, vol. 56, no. 10, pp. 940–945, 1999.

[63] The International Schizophrenia Consortium, "Rare chromosomal deletions and duplications increase risk of schizophrenia.," *Nature*, vol. 455, pp. 237–41, sep 2008.

[64] D. St Clair, "Copy number variation and schizophrenia," *Schizophrenia Bulletin*, vol. 35, no. 1, pp. 9–12, 2009.

[65] D. Malhotra and J. Sebat, "CNVs: harbingers of a rare variant revolution in psychiatric genetics.," *Cell*, vol. 148, pp. 1223–41, mar 2012.

[66] G. Kirov, a. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, K. Chambert, D. Toncheva, L. Georgieva, D. Grozeva, M. Fjodorova, R. Wollerton, E. Rees, I. Nikolov, L. N. van de Lagemaat, A. Bayés, E. Fernandez, P. I. Olason, Y. Böttcher, N. H. Komiyama, M. O. Collins, J. Choudhary, K. Stefansson, H. Stefansson, S. G. N. Grant, S. Purcell, P. Sklar, M. C. O'Donovan, and M. J. Owen, "De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia.," *Molecular psychiatry*, vol. 17, pp. 142–53, feb 2012.

[67] E. Rees, J. T. R. Walters, L. Georgieva, A. R. Isles, K. D. Chambert, A. L. Richards, G. Mahoney-Davies, S. E. Legge, J. L. Moran, S. a. McCarroll, M. C. O'Donovan, M. J. Owen, and G. Kirov, "Analysis of copy number variations at 15 schizophrenia-associated loci.," *The British journal of psychiatry : the journal of mental science*, vol. 204, pp. 108–14, feb 2014.

[68] E. Rees, V. Moskvina, M. J. Owen, M. C. O'Donovan, and G. Kirov, "De novo rates and selection of schizophrenia-associated copy number variants.," *Biological psychiatry*, vol. 70, pp. 1109–14, dec 2011.

[69] V. Vallès, J. Van Os, R. Guillamat, B. Gutiérrez, M. Campillo, P. Gento, and L. Fañanás, "Increased morbid risk for schizophrenia in families of in-patients with bipolar illness.," *Schizophrenia research*, vol. 42, pp. 83–90, apr 2000.

[70] S. H. Lee, S. Ripke, B. M. Neale, S. V. Faraone, S. M. Purcell, R. H. Perlis, B. J. Mowry, A. Thapar, M. E. Goddard, J. S. Witte, D. Absher, I. Agartz, H. Akil, F. Amin, O. a. Andreassen, A. Anjorin, R. Anney, V. Anttila, D. E. Arking, P. Asherson, M. H. Azevedo, L. Backlund, J. a. Badner, A. J. Bailey, T. Banaschewski, J. D. Barchas, M. R. Barnes, T. B. Barrett, N. Bass, A. Battaglia, M. Bauer, M. Bayés, F. Bellivier, S. E. Bergen, W. Berrettini, C. Betancur, T. Bettecken, J. Biederman, E. B. Binder, D. W. Black, D. H. R. Blackwood, C. S. Bloss, M. Boehnke, D. I. Boomsma, G. Breen, R. Breuer, R. Bruggeman, P. Cormican, N. G. Buccola, J. K. Buitelaar, W. E. Bunney, J. D. Buxbaum, W. F. Byerley, E. M. Byrne, S. Caesar, W. Cahn, R. M. Cantor, M. Casas, A. Chakravarti, K. Chambert, K. Choudhury, S. Cichon, C. R. Cloninger, D. a. Collier, E. H. Cook, H. Coon, B. Cormand, A. Corvin, W. H. Coryell, D. W. Craig, I. W. Craig, J. Crosbie, M. L. Cuccaro, D. Curtis, D. Czamara, S. Datta, G. Dawson, R. Day, E. J. De Geus, F. Degenhardt, S. Djurovic, G. J. Donohoe, A. E. Doyle, J. Duan, F. Dudbridge, E. Duketis, R. P. Ebstein, H. J. Edenberg, J. Elia, S. Ennis, B. Etain, A. Fanous, A. E. Farmer, I. N. Ferrier, M. Flickinger, E. Fombonne, T. Foroud, J. Frank, B. Franke, C. Fraser, R. Freedman, N. B. Freimer, C. M. Freitag, M. Friedl, L. Frisén, L. Gallagher, P. V. Gejman, L. Georgieva, E. S. Gershon, D. H. Geschwind, I. Giegling, M. Gill, S. D. Gordon, K. Gordon-Smith, E. K. Green, T. a. Greenwood, D. E. Grice, M. Gross, D. Grozeva, W. Guan, H. Gurling, L. De Haan, J. L. Haines, H. Hakonarson, J. Hallmayer, S. P. Hamilton, M. L. Hamshere, T. F. Hansen, A. M. Hartmann, M. Hautzinger, A. C. Heath, A. K. Henders, S. Herms, I. B. Hickie, M. Hipolito, S. Hoefels, P. a. Holmans, F. Holsboer, W. J. Hoogendijk, J.-J. Hottenga, C. M. Hultman, V. Hus, A. Ingason, M. Ising, S. Jamain, E. G. Jones, I. Jones, L. Jones, J.-Y. Tzeng, A. K. Kähler, R. S. Kahn, R. Kandaswamy, M. C. Keller, J. L. Kennedy, E. Kenny, L. Kent, Y. Kim, G. K. Kirov, S. M. Klauck, L. Klei, J. a. Knowles, M. a. Kohli, D. L. Koller, B. Konte, A. Korszun, L. Krabbendam, R. Krasucki, J. Kuntsi, P. Kwan, M. Landén, N. Långström, M. Lathrop, J. Lawrence, W. B. Lawson, M. Leboyer, D. H. Ledbetter, P. H. Lee, T. Lencz, K.-P. Lesch, D. F. Levinson, C. M. Lewis, J. Li, P. Lichtenstein, J. a. Lieberman, D.-Y. Lin, D. H. Linszen, C. Liu, F. W. Lohoff, S. K. Loo, C. Lord, J. K. Lowe, S. Lucae, D. J. MacIntyre, P. a. F. Madden, E. Maestrini, P. K. E. Magnusson, P. B. Mahon, W. Maier, A. K. Malhotra, S. M. Mane, C. L. Martin, N. G. Martin, M. Mattheisen, K. Matthews, M. Mattingsdal, S. a. McCarroll, K. a. McGhee, J. J. McGough, P. J. McGrath, P. McGuffin, M. G. McInnis, A. McIntosh, R. McKinney, A. W. McLean, F. J. McMahon, W. M. McMahon, A. McQuillin, H. Medeiros, S. E. Medland, S. Meier, I. Melle, F. Meng,

J. Meyer, C. M. Middeldorp, L. Middleton, V. Milanova, A. Miranda, A. P. Monaco, G. W. Montgomery, J. L. Moran, D. Moreno-De-Luca, G. Morken, D. W. Morris, E. M. Morrow, V. Moskvina, P. Muglia, T. W. Mühleisen, W. J. Muir, B. Müller-Myhsok, M. Murtha, R. M. Myers, I. Myin-Germeys, M. C. Neale, S. F. Nelson, C. M. Nievergelt, I. Nikolov, V. Nimgaonkar, W. a. Nolen, M. M. Nöthen, J. I. Nurnberger, E. a. Nwulia, D. R. Nyholt, C. O'Dushlaine, R. D. Oades, A. Olincy, G. Oliveira, L. Olsen, R. a. Ophoff, U. Osby, M. J. Owen, A. Palotie, J. R. Parr, A. D. Paterson, C. N. Pato, M. T. Pato, B. W. Penninx, M. L. Pergadia, M. a. Pericak-Vance, B. S. Pickard, J. Pimm, J. Piven, D. Posthuma, J. B. Potash, F. Poustka, P. Propping, V. Puri, D. J. Quested, E. M. Quinn, J. A. Ramos-Quiroga, H. B. Rasmussen, S. Raychaudhuri, K. Rehnström, A. Reif, M. Ribasés, J. P. Rice, M. Rietschel, K. Roeder, H. Roeyers, L. Rossin, A. Rothenberger, G. Rouleau, D. Ruderfer, D. Rujescu, A. R. Sanders, S. J. Sanders, S. L. Santangelo, J. a. Sergeant, R. Schachar, M. Schalling, A. F. Schatzberg, W. a. Scheftner, G. D. Schellenberg, S. W. Scherer, N. J. Schork, T. G. Schulze, J. Schumacher, M. Schwarz, E. Scolnick, L. J. Scott, J. Shi, P. D. Shilling, S. I. Shyn, J. M. Silverman, S. L. Slager, S. L. Smalley, J. H. Smit, E. N. Smith, E. J. S. Sonuga-Barke, D. St Clair, M. State, M. Steffens, H.-C. Steinhausen, J. S. Strauss, J. Strohmaier, T. S. Stroup, J. S. Sutcliffe, P. Szatmari, S. Szelinger, S. Thirumalai, R. C. Thompson, A. a. Todorov, F. Tozzi, J. Treutlein, M. Uhr, E. J. C. G. van den Oord, G. Van Grootheest, J. Van Os, A. M. Vicente, V. J. Vieland, J. B. Vincent, P. M. Visscher, C. a. Walsh, T. H. Wassink, S. J. Watson, M. M. Weissman, T. Werge, T. F. Wienker, E. M. Wijsman, G. Willemsen, N. Williams, a. J. Willsey, S. H. Witt, W. Xu, A. H. Young, T. W. Yu, S. Zammit, P. P. Zandi, P. Zhang, F. G. Zitman, S. Zöllner, B. Devlin, J. R. Kelsoe, P. Sklar, M. J. Daly, M. C. O'Donovan, N. Craddock, P. F. Sullivan, J. W. Smoller, K. S. Kendler, and N. R. Wray, "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.," *Nature genetics*, vol. 45, pp. 984–94, sep 2013.

[71] E. B. Robinson, B. St Pourcain, V. Anttila, J. A. Kosmicki, B. Bulik-Sullivan, J. Grove, J. Maller, K. E. Samocha, S. J. Sanders, S. Ripke, J. Martin, M. V. Hollegaard, T. Werge, D. M. Hougaard, T. D. Als, M. Baekvad-Hansen, R. Belliveau, D. Demontis, A. Dumont, J. Goldstein, J. Grauholm, C. S. Hansen, T. F. Hansen, D. Howrigan, F. Lescai, M. Mattheisen, J. Moran, O. Mors, M. Nordentoft, B. Norgaard-Pedersen, T. Poterba, J. Poulsen, C. Stevens, R. Walters, B. M. Neale, D. M. Evans, D. Skuse, P. B. Mortensen, A. D. Børglum, A. Ronald, G. D. Smith, and M. J. Daly, "Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population," *Nature Genetics*, vol. 48, pp. 552–555, mar 2016.

[72] P. M. Visscher, M. a. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery.," *American journal of human genetics*, vol. 90, pp. 7–24, jan 2012.

[73] J. P. a. Ioannidis, G. Thomas, and M. J. Daly, "Validating, augmenting and refining genome-wide association signals.," *Nature reviews. Genetics*, vol. 10, pp. 318–29, may 2009.

[74] A. L. Collins and P. F. Sullivan, "Genome-wide association studies in psychiatry: what have we learned?," *The British journal of psychiatry : the journal of mental science*, vol. 202, pp. 1–4, jan 2013.

[75] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant,

R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. a. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. a. Baybayan, V. a. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. a. Bridgham, R. C. Brown, A. a. Brown, D. H. Buermann, A. a. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. a. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. a. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. a. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. a. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. a. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. a. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry. - Supplement," *Nature*, vol. 456, no. 7218, pp. 53–9, 2008.

[76] A. Sboner, X. J. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein, "The real cost of sequencing: higher than you think!," *Genome Biology*, vol. 12, no. 8, p. 125, 2011.

[77] J. K. Teer and J. C. Mullikin, "Exome sequencing: the sweet spot before whole genomes.," *Human molecular genetics*, vol. 19, pp. R145–51, oct 2010.

[78] J. a. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes.," *Science (New York, N.Y.)*, vol. 337, pp. 64–9, jul 2012.

[79] D. F. Conrad, J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright,

G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, and P. Awadalla, "Variation in genome-wide mutation rates within and between human families.," *Nature genetics*, vol. 43, no. 7, pp. 712–4, 2011.

[80] S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. a. Meyer, K. Bilguvar, S. M. Mane, N. Sestan, R. P. Lifton, M. Günel, K. Roeder, D. H. Geschwind, B. Devlin, and M. W. State, "De novo mutations revealed by whole-exome sequencing are strongly associated with autism.," *Nature*, vol. 485, pp. 237–41, may 2012.

[81] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. a. Nickerson, J. Shendure, and M. J. Bamshad, "Exome sequencing identifies the cause of a mendelian disorder.," *Nature genetics*, vol. 42, no. 1, pp. 30–35, 2010.

[82] S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K.-I. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad, and J. Shendure, "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome.," *Nature genetics*, vol. 42, pp. 790–3, sep 2010.

[83] A. Hoischen, B. W. M. van Bon, B. Rodríguez-Santiago, C. Gilissen, L. E. L. M. Vissers, P. de Vries, I. Janssen, B. van Lier, R. Hastings, S. F. Smithson, R. Newbury-Ecob, S. Kjaergaard, J. Goodship, R. McGowan, D. Bartholdi, A. Rauch, M. Peippo, J. M. Cobben, D. Wieczorek, G. Gillessen-Kaesbach, J. a. Veltman, H. G. Brunner, and B. B. B. a. de Vries, "De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome.," *Nature genetics*, vol. 43, no. 8, pp. 729–31, 2011.

[84] J. de Ligt, M. H. Willemsen, B. W. M. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. a. Koolen, P. de Vries, C. Gilissen, M. del Rosario, A. Hoischen, H. Scheffer, B. B. a. de Vries, H. G. Brunner, J. a. Veltman, and L. E. L. M. Vissers, "Diagnostic exome sequencing in persons with severe intellectual disability.," *The New England journal of medicine*, vol. 367, pp. 1921–9, nov 2012.

[85] A. Rauch, D. Wieczorek, E. Graf, T. Wieland, S. Endele, T. Schwarzmayr, B. Albrecht, D. Bartholdi, J. Beygo, N. Di Donato, A. Dufke, K. Cremer, M. Hempel, D. Horn, J. Hoyer, P. Joset, A. Röpke, U. Moog, A. Riess, C. T. Thiel, A. Tzschach, A. Wiesener, E. Wohlleber, C. Zweier, A. B. Ekici, A. M. Zink, A. Rump, C. Meisinger, H. Grallert, H. Sticht, A. Schenck, H. Engels, G. Rappold, E. Schröck, P. Wieacker, O. Riess, T. Meitinger, A. Reis, and T. M. Strom, "Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study.," *Lancet*, vol. 380, pp. 1674–82, nov 2012.

[86] B. M. Neale, Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C.-F. Lin, C. Stevens, L.-S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid,

I. Newsham, Y. Wu, L. Lewis, Y. Han, B. F. Voight, E. Lim, E. Rossin, A. Kirby, J. Flannick, M. Fromer, K. Shakir, T. Fennell, K. Garimella, E. Banks, R. Poplin, S. Gabriel, M. DePristo, J. R. Wimbish, B. E. Boone, S. E. Levy, C. Betancur, S. Sunyaev, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, B. Devlin, R. a. Gibbs, K. Roeder, G. D. Schellenberg, J. S. Sutcliffe, and M. J. Daly, "Patterns and rates of exonic de novo mutations in autism spectrum disorders.," *Nature*, vol. 485, pp. 242–5, may 2012.

[87] B. J. O'Roak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler, "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations.," *Nature*, vol. 485, pp. 246–50, may 2012.

[88] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander, "Searching for missing heritability: Designing rare variant association studies," *Proceedings of the National Academy of Sciences*, jan 2014.

[89] K. A. Hunt, V. Mistry, N. A. Bockett, T. Ahmad, M. Ban, J. N. Barker, J. C. Barrett, H. Blackburn, O. Brand, O. Burren, F. Capon, A. Compston, S. C. Gough, L. Jostins, Y. Kong, J. C. Lee, M. Lek, D. G. Macarthur, J. C. Mansfield, C. G. Mathew, C. A. Mein, M. Mirza, S. Nutland, S. Onengut-Gumuscu, E. Papouli, M. Parkes, S. S. Rich, S. Sawcer, J. Satsangi, M. J. Simmonds, R. C. Trembath, N. M. Walker, E. Wozniak, J. A. Todd, M. A. Simpson, V. Plagnol, and D. A. van Heel, "Negligible impact of rare autoimmune-locus coding-region variants on missing heritability," *Nature*, vol. 498, no. 7453, pp. 232–235, 2013.

[90] H. Tang, X. Jin, Y. Li, H. Jiang, X. Tang, X. Yang, H. Cheng, Y. Qiu, G. Chen, J. Mei, F. Zhou, R. Wu, X. Zuo, Y. Zhang, X. Zheng, Q. Cai, X. Yin, C. Quan, H. Shao, Y. Cui, F. Tian, X. Zhao, H. Liu, F. Xiao, F. Xu, J. Han, D. Shi, A. Zhang, C. Zhou, Q. Li, X. Fan, L. Lin, H. Tian, Z. Wang, H. Fu, F. Wang, B. Yang, S. Huang, B. Liang, X. Xie, Y. Ren, Q. Gu, G. Wen, Y. Sun, X. Wu, L. Dang, M. Xia, J. Shan, T. Li, L. Yang, X. Zhang, Y. Li, C. He, A. Xu, L. Wei, X. Zhao, X. Gao, J. Xu, F. Zhang, J. Zhang, Y. Li, L. Sun, J. Liu, R. Chen, S. Yang, J. Wang, and X. Zhang, "A large-scale screen for coding variants predisposing to psoriasis.," *Nature genetics*, vol. 46, no. 1, pp. 45–50, 2014.

[91] N. Mancuso, N. Rohland, K. a. Rand, A. Tandon, A. Allen, D. Quinque, S. Mallick, H. Li, A. Stram, X. Sheng, Z. Kote-Jarai, D. F. Easton, R. a. Eeles, P. consortium, L. Le Marchand, A. Lubwama, D. Stram, S. Watya, D. V. Conti, B. Henderson, C. a. Haiman, B. Pasaniuc, and D. Reich, "The contribution of rare variation to prostate cancer heritability.," *Nature genetics*, vol. 48, pp. 30–5, jan 2016.

[92] C. T. Johansen, J. Wang, M. B. Lanktree, H. Cao, A. D. McIntyre, M. R. Ban, R. A. Martins, B. A. Kennedy, R. G. Hassell, M. E. Visser, S. M. Schwartz, B. F. Voight, R. Elosua, V. Salomaa, C. J. O'Donnell, G. M. Dallinga-Thie, S. S. Anand, S. Yusuf, M. W. Huff, S. Kathiresan, and R. A. Hegele, "Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia.," *Nature genetics*, vol. 42, no. 8, pp. 684–7, 2010.

[93] M. a. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagacé, B. Neale, K. S. Lo, P. Schumm, L. Törkvist, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, G. Lettre, A. Franke, M. D'Amato, D. P. B. McGovern, J. H. Cho, J. D. Rioux, R. J. Xavier, and M. J. Daly, "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.," *Nature genetics*, vol. 43, pp. 1066–73, nov 2011.

[94] R. Do, N. O. Stitziel, H.-H. Won, A. B. Jørgensen, S. Duga, P. Angelica Merlini, A. Kiezun, M. Farrall, A. Goel, O. Zuk, I. Guella, R. Asselta, L. A. Lange, G. M. Peloso, P. L. Auer, D. Girelli, N. Martinelli, D. N. Farlow, M. A. DePristo, R. Roberts, A. F. R. Stewart, D. Saleheen, J. Danesh, S. E. Epstein, S. Sivapalaratnam, G. Kees Hovingh, J. J. Kastelein, N. J. Samani, H. Schunkert, J. Erdmann, S. H. Shah, W. E. Kraus, R. Davies, M. Nikpay, C. T. Johansen, J. Wang, R. A. Hegele, E. Hechter, W. Marz, M. E. Kleber, J. Huang, A. D. Johnson, M. Li, G. L. Burke, M. Gross, Y. Liu, T. L. Assimes, G. Heiss, E. M. Lange, A. R. Folsom, H. A. Taylor, O. Olivieri, A. Hamsten, R. Clarke, D. F. Reilly, W. Yin, M. A. Rivas, P. Donnelly, J. E. Rossouw, B. M. Psaty, D. M. Herrington, J. G. Wilson, S. S. Rich, M. J. Bamshad, R. P. Tracy, L. Adrienne Cupples, D. J. Rader, M. P. Reilly, J. A. Spertus, S. Cresci, J. Hartiala, W. H. Wilson Tang, S. L. Hazen, H. Allayee, A. P. Reiner, C. S. Carlson, C. Kooperberg, R. D. Jackson, E. Boerwinkle, E. S. Lander, S. M. Schwartz, D. S. Siscovick, R. McPherson, A. Tybjaerg-Hansen, G. R. Abecasis, H. Watkins, D. A. Nickerson, D. Ardissino, S. R. Sunyaev, C. J. O'Donnell, D. Altshuler, S. Gabriel, and S. Kathiresan, "Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction," *Nature*, vol. 518, pp. 102–106, dec 2014.

[95] S. L. Girard, J. Gauthier, A. Noreau, L. Xiong, S. Zhou, L. Jouan, A. Dionne-Laporte, D. Spiegelman, E. Henrion, O. Diallo, P. Thibodeau, I. Bachand, J. Y. J. Bao, A. H. Y. Tong, C.-H. Lin, B. Millet, N. Jaafari, R. Joober, P. a. Dion, S. Lok, M.-O. Krebs, and G. a. Rouleau, "Increased exonic de novo mutation rate in individuals with schizophrenia.," *Nature genetics*, vol. 43, pp. 860–3, sep 2011.

[96] B. Xu, J. L. Roos, P. Dexheimer, B. Boone, B. Plummer, S. Levy, J. A. Gogos, and M. Karayiorgou, "Exome sequencing supports a de novo mutational paradigm for schizophrenia.," *Nature genetics*, vol. 43, pp. 864–8, sep 2011.

[97] B. Xu, I. Ionita-Laza, J. L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J. a. Gogos, and M. Karayiorgou, "De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia.," *Nature genetics*, vol. 44, pp. 1365–9, dec 2012.

[98] M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, M. C. O'Donovan, and M. C. O'Donovan, "De novo mutations in schizophrenia implicate synaptic networks.," *Nature*, vol. 506, pp. 179–184, feb 2014.

[99] A. Takata, B. Xu, I. Ionita-Laza, J. L. Roos, J. a. Gogos, and M. Karayiorgou, "Loss-of-function variants in schizophrenia risk and SETD1A as a candidate susceptibility gene.," *Neuron*, vol. 82, pp. 773–80, may 2014.

[100] S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton, S. Casadei, C. Rippey, H. Shahin, V. L. Nimgaonkar, R. C. P. Go, R. M. Savage, N. R. Swerdlow, R. E. Gur, D. L. Braff, M.-C. King, and J. M. McClellan, "Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network.," *Cell*, vol. 154, pp. 518–29, aug 2013.

[101] M. Guipponi, F. A. Santoni, V. Setola, C. Gehrig, M. Rotharmel, M. Cuenca, O. Guillin, D. Dikeos, G. Georgantopoulos, G. Papadimitriou, L. Curtis, A. Méary, F. Schürhoff, S. Jamain, D. Avramopoulos, M. Leboyer, D. Rujescu, A. Pulver, D. Campion, D. P. Siderovski, and S. E. Antonarakis, "Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes.," *PloS one*, vol. 9, p. e112745, jan 2014.

[102] S. E. McCarthy, J. Gillis, M. Kramer, J. Lihm, S. Yoon, Y. Berstein, M. Mistry, P. Pavlidis, R. Solomon, E. Ghiban, E. Antoniou, E. Kelleher, C. O'Brien, G. Donohoe, M. Gill, D. W. Morris, W. R. McCombie, and A. Corvin, "De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability.," *Molecular psychiatry*, vol. 19, pp. 652–8, jun 2014.

[103] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kähler, L. Duncan, E. Stahl, G. Genovese, E. Fernández, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. E. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. N. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. a. McCarroll, and P. Sklar, "A polygenic burden of rare disruptive mutations in schizophrenia.," *Nature*, vol. 506, pp. 185–90, feb 2014.

[104] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test.," *American journal of human genetics*, vol. 89, pp. 82–93, jul 2011.

[105] S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, K. Samocha, A. E. Cicek, Y. Kou, L. Liu, M. Fromer, S. Walker, T. Singh, L. Klei, J. Kosmicki, F. Shih-Chen, B. Aleksic, M. Biscaldi, P. F. Bolton, J. M. Brownfeld, J. Cai, N. G. Campbell, A. Carracedo, M. H. Chahrour, A. G. Chiocchetti, H. Coon, E. L. Crawford, S. R. Curran, G. Dawson, E. Duketis, B. a. Fernandez, L. Gallagher, E. Geller, S. J. Guter, R. S. Hill, J. Ionita-Laza, P. Jimenz Gonzalez, H. Kilpinen, S. M. Klauck, A. Kolevzon, I. Lee, I. Lei, J. Lei, T. Lehtimäki, C.-F. Lin, A. Ma'ayan, C. R. Marshall, A. L. McInnes, B. Neale, M. J. Owen, N. Ozaki, M. Parellada, J. R. Parr, S. Purcell, K. Puura, D. Rajagopalan, K. Rehnström, A. Reichenberg, A. Sabo, M. Sachse, S. J. Sanders, C. Schafer, M. Schulte-Rüther, D. Skuse, C. Stevens, P. Szatmari, K. Tammimies, O. Valladares, A. Voran, W. Li-San, L. a. Weiss, A. J. Willsey, T. W. Yu, R. K. C. Yuen, E. H. Cook, C. M. Freitag, M. Gill, C. M. Hultman, T. Lehner, A. Palotie, G. D. Schellenberg, P. Sklar, M. W. State, J. S. Sutcliffe, C. a. Walsh, S. W. Scherer, M. E. Zwick, J. C. Barett, D. J. Cutler, K. Roeder, B. Devlin, M. J. Daly, and J. D. Buxbaum,

"Synaptic, transcriptional and chromatin genes disrupted in autism.," *Nature*, vol. 515, pp. 209–15, nov 2014.

[106] J. P. Szatkiewicz, C. O'Dushlaine, G. Chen, K. Chambert, J. L. Moran, B. M. Neale, M. Fromer, D. Ruderfer, S. Akterin, S. E. Bergen, A. Kähler, P. K. E. Magnusson, Y. Kim, J. J. Crowley, E. Rees, G. Kirov, M. C. O'Donovan, M. J. Owen, J. Walters, E. Scolnick, P. Sklar, S. Purcell, C. M. Hultman, S. a. McCarroll, and P. F. Sullivan, "Copy number variation in schizophrenia in Sweden," *Molecular Psychiatry*, vol. 19, pp. 762–773, jul 2014.

[107] O. D. Howes and S. Kapur, "The dopamine hypothesis of schizophrenia: version III–the final common pathway.," *Schizophrenia bulletin*, vol. 35, pp. 549–62, may 2009.

[108] Psychiatric Genetics Consortium, "Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways.," *Nature neuroscience*, jan 2015.

[109] S. Sanders, X. He, A. Willsey, A. Ercan-Sencicek, K. Samocha, A. Cicek, M. Murtha, V. Bal, S. Bishop, S. Dong, A. Goldberg, C. Jinlu, J. Keaney, L. Klei, J. Mandell, D. Moreno-De-Luca, C. Poultney, E. Robinson, L. Smith, T. Solli-Nowlan, M. Su, N. Teran, M. Walker, D. Werling, A. Beaudet, R. Cantor, E. Fombonne, D. Geschwind, D. Grice, C. Lord, J. Lowe, S. Mane, D. Martin, E. Morrow, M. Talkowski, J. Sutcliffe, C. Walsh, T. Yu, D. Ledbetter, C. Martin, E. Cook, J. Buxbaum, M. Daly, B. Devlin, K. Roeder, and M. State, "Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci," *Neuron*, vol. 87, pp. 1215–1233, sep 2015.

[110] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad, "Exome sequencing identifies the cause of a mendelian disorder.," *Nature genetics*, vol. 42, pp. 30–5, jan 2010.

[111] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for Mendelian disease gene discovery.," *Nature reviews. Genetics*, vol. 12, pp. 745–55, nov 2011.

[112] Exome Aggregation Consortium, "Analysis of protein-coding genetic variation in 60,706 humans," tech. rep., oct 2015.

[113] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2078–9, aug 2009.

[114] A. Mckenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. a. DePristo, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.," *Genome research*, vol. 20, pp. 1297–303, sep 2010.

[115] M. a. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. a. Philippakis, G. del Angel, M. a. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data.," *Nature genetics*, vol. 43, pp. 491–8, may 2011.

[116] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 2987–93, nov 2011.

[117] H. Li, "Toward better understanding of artifacts in variant calling from high-coverage samples.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 2843–51, oct 2014.

[118] The Deciphering Developmental Disorders Study, "Large-scale discovery of novel genetic causes of developmental disorders.," *Nature*, vol. 519, pp. 223–8, mar 2015.

[119] T. Singh, M. I. Kurki, D. Curtis, S. M. Purcell, L. Crooks, J. McRae, J. Suvisaari, H. Chheda, D. Blackwood, G. Breen, O. Pietiläinen, S. S. Gerety, M. Ayub, M. Blyth, T. Cole, D. Collier, E. L. Coomber, N. Craddock, M. J. Daly, J. Danesh, M. DiForti, A. Foster, N. B. Freimer, D. Geschwind, M. Johnstone, S. Joss, G. Kirov, J. Körkkö, O. Kuismin, P. Holmans, C. M. Hultman, C. Iyegbe, J. Lönnqvist, M. Männikkö, S. A. McCarroll, P. McGuffin, A. M. McIntosh, A. McQuillin, J. S. Moilanen, C. Moore, R. M. Murray, R. Newbury-Ecob, W. Ouwehand, T. Paunio, E. Prigmore, E. Rees, D. Roberts, J. Sambrook, P. Sklar, D. S. Clair, J. Veijola, J. T. R. Walters, H. Williams, P. F. Sullivan, M. E. Hurles, M. C. O'Donovan, A. Palotie, M. J. Owen, and J. C. Barrett, "Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders," *Nature Neuroscience*, vol. 19, pp. 571–577, mar 2016.

[120] Picard, "Picard."

[121] G. Jun, M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke, and H. M. Kang, "Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.," *American journal of human genetics*, vol. 91, pp. 839–48, nov 2012.

[122] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.," *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. 11, pp. 11.10.1–11.10.33, oct 2013.

[123] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project.," *Genome research*, vol. 22, pp. 1760–74, sep 2012.

[124] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. a. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses.," *American journal of human genetics*, vol. 81, pp. 559–75, sep 2007.

[125] T. Thornton, H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan, and N. Risch, "Estimating kinship in admixed populations.," *American journal of human genetics*, vol. 91, pp. 122–38, jul 2012.

[126] J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon, "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.," *Genome medicine*, vol. 5, no. 3, p. 28, 2013.

[127] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2069–70, aug 2010.

[128] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.," *Nature protocols*, vol. 4, pp. 1073–81, jan 2009.

[129] I. a. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations.," *Nature methods*, vol. 7, pp. 248–9, apr 2010.

[130] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants.," *Nature genetics*, vol. 46, pp. 310–5, mar 2014.

[131] T. U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research*, vol. 43, pp. 204–212, oct 2014.

[132] C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, and X. Liu, "Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies.," *Human molecular genetics*, pp. 1–13, dec 2014.

[133] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations," *Human Mutation*, vol. 34, no. 9, pp. 2393–2402, 2013.

[134] G. Genovese, M. Fromer, E. A. Stahl, D. M. Ruderfer, K. Chambert, M. Landén, J. L. Moran, S. M. Purcell, P. Sklar, P. F. Sullivan, C. M. Hultman, and S. A. McCarroll, "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia," *Nature Neuroscience*, no. October, 2016.

[135] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U.

Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Müller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikram, B.-J. Kim, Y. Kim, Y. J. Kim, M.-S. Kwon, J. Lee, S. Lee, K.-H. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B.-G. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stančáková, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S.-H. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, W. Y. So, C. H. T. Tam, U. Afzal, D. Aguilar, R. Arya, T. Aung, E. Chan, C. Navarro, C.-Y. Cheng, D. Palli, A. Correa, J. E. Curran, D. Rybin, V. S. Farook, S. P. Fowler, B. I. Freedman, M. Griswold, D. E. Hale, P. J. Hicks, C.-C. Khor, S. Kumar, B. Lehne, D. Thuillier, W. Y. Lim, J. Liu, Y. T. van der Schouw, M. Loh, S. K. Musani, S. Puppala, W. R. Scott, L. Yengo, S.-T. Tan, H. A. Taylor Jr., F. Thameem, G. Wilson, T. Y. Wong, P. R. Njølstad, J. C. Levy, M. Mangino, L. L. Bonnycastle, T. Schwarzmayr, J. Fadista, G. L. Surdulescu, C. Herder, C. J. Groves, T. Wieland, J. Bork-Jensen, I. Brandslund, C. Christensen, H. A. Koistinen, A. S. F. Doney, L. Kinnunen, T. Esko, A. J. Farmer, L. Hakaste, D. Hodgkiss, J. Kravic, V. Lyssenko, M. Hollensted, M. E. Jørgensen, T. Jørgensen, C. Ladenvall, J. M. Justesen, A. Käräjämäki, J. Kriebel, W. Rathmann, L. Lannfelt, T. Lauritzen, N. Narisu, A. Linneberg, O. Melander, L. Milani, M. Neville, M. Orho-Melander, L. Qi, Q. Qi, M. Roden, O. Rolandsson, A. Swift, A. H. Rosengren, K. Stirrups, A. R. Wood, E. Mihailov, C. Blancher, M. O. Carneiro, J. Maguire, R. Poplin, K. Shakir, T. Fennell, M. DePristo, M. Hrabé de Angelis, P. Deloukas, A. P. Gjesing, G. Jun, P. Nilsson, J. Murphy, R. Onofrio, B. Thorand, T. Hansen, C. Meisinger, F. B. Hu, B. Isomaa, F. Karpe, L. Liang, A. Peters, C. Huth, S. P. O'Rahilly, C. N. A. Palmer, O. Pedersen, R. Rauramaa, J. Tuomilehto, V. Salomaa, R. M. Watanabe, A.-C. Syvänen, R. N. Bergman, D. Bharadwaj, E. P. Bottinger, Y. S. Cho, G. R. Chandak, J. C. N. Chan, K. S. Chia, M. J. Daly, S. B. Ebrahim, C. Langenberg, P. Elliott, K. A. Jablonski, D. M. Lehman, W. Jia, R. C. W. Ma, T. I. Pollin, M. Sandhu, N. Tandon, P. Froguel, I. Barroso, Y. Y. Teo, E. Zeggini, R. J. F. Loos, K. S. Small, J. S. Ried, R. A. DeFronzo, H. Grallert, B. Glaser, A. Metspalu, N. J. Wareham, M. Walker, E. Banks, C. Gieger, E. Ingelsson, H. K. Im, T. Illig, P. W. Franks, G. Buck, J. Trakalo, D. Buck, I. Prokopenko, R. Mägi, L. Lind, Y. Farjoun, K. R. Owen, A. L. Gloyn, K. Strauch, T. Tuomi, J. S. Kooner, J.-Y. Lee, T. Park, P. Donnelly, A. D. Morris, A. T. Hattersley, D. W. Bowden, F. S. Collins, G. Atzmon, J. C. Chambers, T. D. Spector, M. Laakso, T. M. Strom, G. I. Bell, J. Blangero, R. Duggirala, E. S. Tai, G. McVean, C. L. Hanis, J. G. Wilson, M. Seielstad, T. M. Frayling, J. B. Meigs, N. J. Cox, R. Sladek, E. S. Lander, S. Gabriel, N. P. Burtt, K. L. Mohlke, T. Meitinger, L. Groop, G. Abecasis, J. C. Florez, L. J. Scott, A. P. Morris, H. M. Kang, M. Boehnke, D. Altshuler, and M. I. McCarthy, "The genetic architecture of type 2 diabetes," *Nature*, vol. 536, pp. 41–47, jul 2016.

[136] D. J. McCarthy, P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J.-b. Cazier, and P. Donnelly, "Choice of transcripts and software has a large effect on variant annotation.," *Genome medicine*, vol. 6, no. 3, p. 26, 2014.

[137] W. D. Jones, D. Dafou, M. McEntagart, W. J. Woollard, F. V. Elmslie, M. Holder-Espinasse, M. Irving, A. K. Saggar, S. Smithson, R. C. Trembath, C. Deshpande, and M. A. Simpson, "De novo mutations in MLL cause Wiedemann-Steiner syndrome," *American Journal of Human Genetics*, vol. 91, no. 2, pp. 358–364, 2012.

[138] K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. a. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, R. a. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, and M. J. Daly, "A framework for the interpretation of de novo mutation in human disease," *Nature Genetics*, vol. 46, pp. 944–950, aug 2014.

[139] C. Moore, J. Sambrook, M. Walker, Z. Tolkien, S. Kaptoge, D. Allen, S. Mehenny, J. Mant, E. Di Angelantonio, S. G. Thompson, W. Ouwehand, D. J. Roberts, and J. Danesh, "The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial.," *Trials*, vol. 15, p. 363, 2014.

[140] J. Yang, T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher, "Genome partitioning of genetic variation for complex traits using common SNPs.," *Nature genetics*, vol. 43, no. 6, pp. 519–25, 2011.

[141] N. Chatterjee, J. Shi, and M. García-Closas, "Developing and evaluating polygenic risk prediction models for stratified disease prevention.," *Nature reviews. Genetics*, vol. 17, pp. 392–406, jul 2016.

[142] C. Gillberg, S. Steffenburg, J. Wahlström, I. C. Gillberg, A. Sjöstedt, T. Martinsson, S. Liedgren, and O. Eeg-Olofsson, "Autism associated with marker chromosome.," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 30, pp. 489–94, may 1991.

[143] C. Gillberg, "Chromosomal disorders and autism," *Journal of Autism and Developmental Disorders*, vol. 28, no. 5, pp. 415–425, 1998.

[144] M. Karayiorgou, M. a. Morris, B. Morrow, R. J. Shprintzen, R. Goldberg, J. Borrow, a. Gos, G. Nestadt, P. S. Wolyniec, and V. K. Lasseter, "Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 17, pp. 7612–7616, 1995.

[145] J. Sebat, D. L. Levy, and S. E. McCarthy, "Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders," *Trends in Genetics*, vol. 25, no. 12, pp. 528–535, 2009.

[146] L. Georgieva, E. Rees, J. L. Moran, K. D. Chambert, V. Milanova, N. Craddock, S. Purcell, P. Sklar, S. McCarroll, P. Holmans, M. C. O'Donovan, M. J. Owen, and G. Kirov, "De novo CNVs in bipolar affective disorder and schizophrenia.," *Human molecular genetics*, pp. 1–7, jul 2014.

[147] D. Levy, M. Ronemus, B. Yamrom, Y.-h. Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, A. Buja, A. Krieger, S. Yoon, J. Troge, L. Rodgers, I. Iossifov, and M. Wigler, "Rare de novo and transmitted copy-number variation in autistic spectrum disorders.," *Neuron*, vol. 70, pp. 886–97, jun 2011.

[148] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.-H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.-C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, "Strong Association of De Novo Copy Number Mutations with Autism," *Science*, vol. 316, pp. 445–449, apr 2007.

[149] S. J. Sanders, a. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. a. Thomson, C. E. Mason, K. Bilguvar, P. B. S. Celestino-Soper, M. Choi, E. L. Crawford, L. Davis, N. R. D. Wright, R. M. Dhodapkar, M. DiCola, N. M. DiLullo, T. V. Fernandez, V. Fielding-Singh, D. O. Fishman, S. Frahm, R. Garagaloyan, G. S. Goh, S. Kammela, L. Klei, J. K. Lowe, S. C. Lund, A. D. McGrew, K. a. Meyer, W. J. Moffat, J. D. Murdoch, B. J. O'Roak, G. T. Ober, R. S. Pottenger, M. J. Raubeson, Y. Song, Q. Wang, B. L. Yaspan, T. W. Yu, I. R. Yurkiewicz, A. L. Beaudet, R. M. Cantor, M. Curland, D. E. Grice, M. Günel, R. P. Lifton, S. M. Mane, D. M. Martin, C. a. Shaw, M. Sheldon, J. a. Tischfield, C. a. Walsh, E. M. Morrow, D. H. Ledbetter, E. Fombonne, C. Lord, C. L. Martin, A. I. Brooks, J. S. Sutcliffe, E. H. Cook, D. Geschwind, K. Roeder, B. Devlin, and M. W. State, "Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism.," *Neuron*, vol. 70, pp. 863–85, jun 2011.

[150] C. Golzio, J. Willer, M. E. Talkowski, E. C. Oh, Y. Taniguchi, S. Jacquemont, A. Reymond, M. Sun, A. Sawa, J. F. Gusella, A. Kamiya, J. S. Beckmann, and N. Katsanis, "KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant," *Nature*, vol. 485, no. 7398, pp. 363–367, 2012.

[151] X. He, S. J. Sanders, L. Liu, S. De Rubeis, E. T. Lim, J. S. Sutcliffe, G. D. Schellenberg, R. a. Gibbs, M. J. Daly, J. D. Buxbaum, M. W. State, B. Devlin, and K. Roeder, "Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes.," *PLoS genetics*, vol. 9, p. e1003671, jan 2013.

[152] PLINKSEQ, "PLINKSEQ."

[153] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. M. Black, D. L. Brown, M. Brudno, J. Campbell, D. R. FitzPatrick, J. T. Eppig, A. P. Jackson, K. Freson, M. Girdea, I. Helbig, J. a. Hurst, J. Jähn, L. G. Jackson, A. M. Kelly, D. H. Ledbetter, S. Mansour, C. L. Martin, C. Moss, A. Mumford, W. H. Ouwehand, S.-M. Park, E. R. Riggs, R. H. Scott, S. Sisodiya, S. Van Vooren, R. J. Wapner, A. O. M. Wilkie, C. F. Wright, A. T. Vulto-van Silfhout, N. de Leeuw, B. B. a. de Vries, N. L. Washingthon, C. L. Smith, M. Westerfield, P. Schofield, B. J. Ruef, G. V. Gkoutos, M. Haendel, D. Smedley, S. E. Lewis, and P. N. Robinson, "The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data.," *Nucleic acids research*, vol. 42, pp. D966–74, jan 2014.

[154] G. Genovese, A. K. Kähler, R. E. Handsaker, J. Lindberg, S. a. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. Landén, M. Höglund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H. Grönberg, C. M. Hultman, and S. a. McCarroll, "Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence.," *The New England journal of medicine*, vol. 371, pp. 2477–87, dec 2014.

[155] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. a. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, J. D. Smith, B. Paeper, D. a. Nickerson, J. Dea, S. Dong, L. E. Gonzalez, J. D. Mandell, S. M. Mane, M. T. Murtha, C. a. Sullivan, M. F. Walker, Z. Waqar, L. Wei, a. J. Willsey, B. Yamrom, Y.-h. Lee, E. Grabowska, E. Dalkic, Z. Wang, S. Marks, P. Andrews, A. Leotta, J. Kendall, I. Hakker, J. Rosenbaum, B. Ma, L. Rodgers, J. Troge, G. Narzisi, S. Yoon, M. C. Schatz, K. Ye, W. R. McCombie, J. Shendure, E. E. Eichler, M. W. State, and M. Wigler, "The contribution of de novo coding mutations to autism spectrum disorder.," *Nature*, vol. 515, pp. 216–21, nov 2014.

[156] J. a. Fahrner and H. T. Bjornsson, "Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states.," *Annual review of genomics and human genetics*, vol. 15, pp. 269–93, jan 2014.

[157] H. V. Firth, S. M. Richards, a. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter, "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.," *American journal of human genetics*, vol. 84, pp. 524–33, apr 2009.

[158] G. Kirov, E. Rees, J. T. R. Walters, V. Escott-Price, L. Georgieva, A. L. Richards, K. D. Chambert, G. Davies, S. E. Legge, J. L. Moran, S. a. McCarroll, M. C. O'Donovan, and M. J. Owen, "The penetrance of copy number variations for schizophrenia and developmental delay.," *Biological psychiatry*, vol. 75, pp. 378–85, mar 2014.

[159] A. S. Bassett, E. W. C. Chow, J. Husted, R. Weksberg, O. Caluseriu, G. D. Webb, and M. a. Gatzoulis, "Clinical features of 78 adults with 22q11 Deletion Syndrome.," *American journal of medical genetics.*, vol. 138, pp. 307–13, nov 2005.

[160] N. J. Butcher, E. W. C. Chow, G. Costain, D. Karas, A. Ho, and A. S. Bassett, "Functional outcomes of adults with 22q11.2 deletion syndrome.," *Genetics in medicine*, vol. 14, pp. 836–43, oct 2012.

[161] A. K. Ryan, J. A. Goodship, D. I. Wilson, N. Philip, A. Levy, H. Seidel, S. Schuffenhauer, H. Oechsler, B. Belohradsky, M. Prieur, A. Aurias, F. L. Raymond, J. Clayton-Smith, E. Hatchwell, C. McKeown, F. A. Beemer, B. Dallapiccola, G. Novelli, J. A. Hurst, J. Ignatius, A. J. Green, R. M. Winter, L. Brueton, K. Brøndum-Nielsen, and P. J. Scambler, "Spectrum of clinical features associated with interstitial chromosome 22q11 deletions: a European collaborative study.," *Journal of medical genetics*, vol. 34, pp. 798–804, oct 1997.

[162] J.-H. Lee and D. G. Skalnik, "CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex.," *The Journal of biological chemistry*, vol. 280, pp. 41725–31, dec 2005.

[163] J.-H. Lee, C. M. Tate, J.-S. You, and D. G. Skalnik, "Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex.," *The Journal of biological chemistry*, vol. 282, pp. 13419–28, may 2007.

[164] T. Kleefstra, W. a. van Zelst-Stams, W. M. Nillesen, V. Cormier-Daire, G. Houge, N. Foulds, M. van Dooren, M. H. Willemsen, R. Pfundt, a. Turner, M. Wilson, J. McGaughran, a. Rauch, M. Zenker, M. P. Adam, M. Innes, C. Davies, a. G.-M. López, R. Casalone, a. Weber, L. a. Brueton, a. D. Navarro, M. P. Bralo, H. Venselaar, S. P. a. Stegmann, H. G. Yntema, H. van Bokhoven, and H. G. Brunner, "Further clinical and molecular delineation of the 9q subtelomeric deletion syndrome supports a major contribution of EHMT1 haploinsufficiency to the core phenotype.," *Journal of medical genetics*, vol. 46, pp. 598–606, sep 2009.

[165] A. Dincer, D. P. Gavin, K. Xu, B. Zhang, J. T. Dudley, E. E. Schadt, and S. Akbarian, "Deciphering H3K4me3 broad domains associated with gene-regulatory networks and conserved epigenomic landscapes in the human brain.," *Translational psychiatry*, vol. 5, no. February, p. e679, 2015.

[166] D. R. Weinberger, "Implications of normal brain development for the pathogenesis of schizophrenia.," *Archives of general psychiatry*, vol. 44, no. 7, pp. 660–669, 1987.

[167] R. M. Murray and S. W. Lewis, "Is schizophrenia a neurodevelopmental disorder?," *British medical journal (Clinical research ed.)*, vol. 295, pp. 681–2, sep 1987.

[168] I. C. Wright, S. Rabe-Hesketh, P. W. Woodruff, A. S. David, R. M. Murray, and E. T. Bullmore, "Meta-analysis of regional brain volumes in schizophrenia.," *The American journal of psychiatry*, vol. 157, pp. 16–25, jan 2000.

[169] S. H. Fatemi and T. D. Folsom, "The neurodevelopmental hypothesis of schizophrenia, revisited.," *Schizophrenia bulletin*, vol. 35, pp. 528–48, may 2009.

[170] S. V. Haijma, N. Van Haren, W. Cahn, P. C. M. P. Koolschijn, H. E. Hulshoff Pol, and R. S. Kahn, "Brain volumes in schizophrenia: A meta-analysis in over 18 000 subjects," *Schizophrenia Bulletin*, vol. 39, no. 5, pp. 1129–1138, 2013.

[171] H. B. M. Boos, A. Aleman, W. Cahn, H. Hulshoff Pol, and R. S. Kahn, "Brain volumes in relatives of patients with schizophrenia: a meta-analysis.," *Archives of general psychiatry*, vol. 64, pp. 297–304, mar 2007.

[172] D. C. Glahn, A. R. Laird, I. Ellison-Wright, S. M. Thelen, J. L. Robinson, J. L. Lancaster, E. Bullmore, and P. T. Fox, "Meta-Analysis of Gray Matter Anomalies in Schizophrenia: Application of Anatomic Likelihood Estimation and Network Analysis," *Biological Psychiatry*, vol. 64, no. 9, pp. 774–781, 2008.

[173] I. Ellison-Wright and E. Bullmore, "Meta-analysis of diffusion tensor imaging studies in schizophrenia," *Schizophrenia Research*, vol. 108, no. 1-3, pp. 3–10, 2009.

[174] C. Nosarti, A. Reichenberg, R. M. Murray, S. Cnattingius, M. P. Lambe, L. Yin, J. MacCabe, L. Rifkin, and C. M. Hultman, "Preterm birth and psychiatric disorders in young adult life," *Arch.Gen.Psychiatry*, vol. 69, no. 1538-3636 (Electronic), pp. E1—-E8, 2012.

[175] P. B. Mortensen, B. N??rgaard-Pedersen, B. L. Waltoft, T. L. S??rensen, D. Hougaard, E. F. Torrey, and R. H. Yolken, "Toxoplasma gondii as a Risk Factor for Early-Onset Schizophrenia: Analysis of Filter Paper Blood Samples Obtained at Birth," *Biological Psychiatry*, vol. 61, no. 5, pp. 688–693, 2007.

[176] M. G. Pedersen, H. Stevens, C. B. Pedersen, B. Nørgaard-Pedersen, and P. B. Mortensen, "Toxoplasma infection and later development of schizophrenia in mothers," *American Journal of Psychiatry*, vol. 168, no. 8, pp. 814–821, 2011.

[177] W. Veling, H. W. Hoek, J. P. Selten, and E. Susser, "Age at migration and future risk of psychotic disorders among immigrants in the Netherlands: A 7-year incidence study," *American Journal of Psychiatry*, vol. 168, no. 12, pp. 1278–1285, 2011.

[178] J. Cotney, R. a. Muhle, S. J. Sanders, L. Liu, a. J. Willsey, W. Niu, W. Liu, L. Klei, J. Lei, J. Yin, S. K. Reilly, A. T. Tebbenkamp, C. Bichsel, M. Pletikos, N. Sestan, K. Roeder, M. W. State, B. Devlin, and J. P. Noonan, "The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment.," *Nature communications*, vol. 6, p. 6404, 2015.

[179] E. Rees, J. T. R. Walters, K. D. Chambert, C. O'Dushlaine, J. Szatkiewicz, A. L. Richards, L. Georgieva, G. Mahoney-Davies, S. E. Legge, J. L. Moran, G. Genovese, D. Levinson, D. W. Morris, P. Cormican, K. S. Kendler, F. a. O'Neill, B. Riley, M. Gill, A. Corvin, P. Sklar, C. Hultman, C. Pato, M. Pato, P. F. Sullivan, P. V. Gejman, S. a. McCarroll, M. C. O'Donovan, M. J. Owen, and G. Kirov, "CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1.," *Human molecular genetics*, vol. 23, pp. 1669–76, mar 2014.

[180] A. L. Price, G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev, "Pooled Association Tests for Rare Variants in Exon-Resequencing Studies," *American Journal of Human Genetics*, vol. 86, no. 6, pp. 832–838, 2010.

[181] S. Raychaudhuri, J. M. Korn, S. A. McCarroll, D. Altshuler, P. Sklar, S. Purcell, and M. J. Daly, "Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function," *PLoS Genetics*, vol. 6, no. 9, 2010.

[182] A. J. Pocklington, E. Rees, J. T. R. Walters, J. Han, D. H. Kavanagh, K. D. Chambert, P. Holmans, J. L. Moran, S. A. McCarroll, G. Kirov, M. C. O'Donovan, and M. J. Owen, "Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia," *Neuron*, vol. 86, no. 5, pp. 1203–1214, 2015.

[183] J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. Chi, D. D. Licatalosi, J. D. Richter, and R. B. Darnell, "FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism.," *Cell*, vol. 146, pp. 247–61, jul 2011.

[184] M. Ascano, N. Mukherjee, P. Bandaru, J. B. Miller, J. D. Nusbaum, D. L. Corcoran, C. Langlois, M. Munschauer, S. Dewell, M. Hafner, Z. Williams, U. Ohler, and T. Tuschl, "FMRP targets distinct mRNA sequence elements to regulate protein expression," *Nature*, vol. 492, no. 7429, pp. 382–386, 2012.

[185] E. A. Heller, W. Zhang, F. Selimi, J. C. Earnheart, M. A. ??limak, J. Santos-Torres, I. Iba??ez-Tallon, C. Aoki, B. T. Chait, and N. Heintz, "The biochemical anatomy of cortical inhibitory synapses," *PLoS ONE*, vol. 7, no. 6, 2012.

[186] J. F. McRae, S. Clayton, T. W. Fitzgerald, J. Kaplanis, E. Prigmore, D. Rajan, A. Sifrim, S. Aitken, N. Akawi, M. Alvi, K. Ambridge, D. M. Barrett, T. Bayzetinova, P. Jones, W. D. Jones, D. King, N. Krishnappa, L. E. Mason, T. Singh, A. R. Tivey, M. Ahmed, U. Anjum, H. Archer, R. Armstrong, J. Awada, M. Balasubramanian, S. Banka, D. Baralle, A. Barnicoat, P. Batstone, D. Baty, C. Bennett, J. Berg, B. Bernhard, A. P. Bevan, M. Bitner-Glindzicz, E. Blair, M. Blyth, D. Bohanna, L. Bourdon, D. Bourn, L. Bradley, A. Brady, S. Brent, C. Brewer, K. Brunstrom, D. J. Bunyan, J. Burn, N. Canham, B. Castle, K. Chandler, E. Chatzimichali, D. Cilliers, A. Clarke, S. Clasper, J. Clayton-Smith, V. Clowes, A. Coates, T. Cole, I. Colgiu, A. Collins, M. N. Collinson, F. Connell, N. Cooper, H. Cox, L. Cresswell, G. Cross, Y. Crow, M. D'Alessandro, T. Dabir, R. Davidson, S. Davies, D. de Vries, J. Dean, C. Deshpande, G. Devlin, A. Dixit, A. Dobbie, A. Donaldson, D. Donnai, D. Donnelly, C. Donnelly, A. Douglas, S. Douzgou, A. Duncan, J. Eason, S. Ellard, I. Ellis, F. Elmslie, K. Evans, S. Everest, T. Fendick, R. Fisher, F. Flinter, N. Foulds, A. Fry, A. Fryer, C. Gardiner, L. Gaunt, N. Ghali, R. Gibbons, H. Gill, J. Goodship, D. Goudie, E. Gray, A. Green, P. Greene, L. Greenhalgh, S. Gribble, R. Harrison, L. Harrison, V. Harrison, R. Hawkins, L. He, S. Hellens, A. Henderson, S. Hewitt, L. Hildyard, E. Hobson, S. Holden, M. Holder, S. Holder, G. Hollingsworth, T. Homfray, M. Humphreys, J. Hurst, B. Hutton, S. Ingram, M. Irving, L. Islam, A. Jackson, J. Jarvis, L. Jenkins, D. Johnson, E. Jones, D. Josifova, S. Joss, B. Kaemba, S. Kazembe, R. Kelsell, B. Kerr, H. Kingston, U. Kini, E. Kinning, G. Kirby, C. Kirk, E. Kivuva, A. Kraus, D. Kumar, V. A. Kumar, K. Lachlan, W. Lam, A. Lampe, C. Langman, M. Lees, D. Lim, C. Longman, G. Lowther, S. A. Lynch, A. Magee, E. Maher, A. Male, S. Mansour, K. Marks, K. Martin, U. Maye, E. McCann, V. McConnell, M. McEntagart, R. McGowan, K. McKay, S. McKee, D. J. McMullan, S. McNerlan, C. McWilliam, S. Mehta, K. Metcalfe, A. Middleton, Z. Miedzybrodzka, E. Miles, S. Mohammed, T. Montgomery, D. Moore, S. Morgan, J. Morton, H. Mugalaasi, V. Murday, H. Murphy, S. Naik, A. Nemeth, L. Nevitt, R. Newbury-Ecob, A. Norman, R. O'Shea, C. Ogilvie, K.-R. Ong, S.-M. Park, M. J. Parker, C. Patel, J. Paterson, S. Payne, D. Perrett, J. Phipps, D. T. Pilz, M. Pollard, C. Pottinger, J. Poulton, N. Pratt, K. Prescott, S. Price, A. Pridham, A. Procter, H. Purnell, O. Quarrell, N. Ragge, R. Rahbari, J. Randall, J. Rankin, L. Raymond, D. Rice, L. Robert, E. Roberts, J. Roberts, P. Roberts, G. Roberts, A. Ross, E. Rosser, A. Saggar, S. Samant, J. Sampson, R. Sandford, A. Sarkar, S. Schweiger, R. Scott, I. Scurr, A. Selby, A. Seller, C. Sequeira, N. Shannon, S. Sharif, C. Shaw-Smith, E. Shearing, D. Shears, E. Sheridan, I. Simonic, R. Singzon, Z. Skitt, A. Smith, K. Smith, S. Smithson, L. Sneddon, M. Splitt, M. Squires, F. Stewart, H. Stewart, V. Straub, M. Suri, V. Sutton, G. J. Swaminathan, E. Sweeney, K. Tatton-Brown, C. Taylor, R. Taylor, M. Tein, I. K. Temple, J. Thomson, M. Tischkowitz, S. Tomkins, A. Torokwa, B. Treacy, C. Turner, P. Turnpenny, C. Tysoe, A. Vandersteen, V. Varghese, P. Vasudevan, P. Vijayarangakannan, J. Vogt, E. Wakeling, S. Wallwark, J. Waters, A. Weber, D. Wellesley, M. Whiteford, S. Widaa, S. Wilcox, E. Wilkinson, D. Williams, N. Williams, L. Wilson, G. Woods, C. Wragg, M. Wright, L. Yates, M. Yau, C. Nellaker, M. J. Parker, H. V. Firth, C. F. Wright, D. R. FitzPatrick, J. C. Barrett, and M. E. Hurles, "Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation," tech. rep., apr 2016.

[187] A. Sugathan, M. Biagioli, C. Golzio, S. Erdin, I. Blumenthal, P. Manavalan, A. Ragavendran, H. Brand, D. Lucente, J. Miles, S. D. Sheridan, A. Stortchevoi, M. Kellis, S. J. Haggarty, N. Katsanis, J. F. Gusella, and M. E. Talkowski, "CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors," *Proceedings of the National Academy of Sciences*, vol. 111, no. 42, pp. E4468–E4477, 2014.

[188] S. Weyn-Vanhentenryck, A. Mele, Q. Yan, S. Sun, N. Farny, Z. Zhang, C. Xue, M. Herre, P. Silver, M. Zhang, A. Krainer, R. Darnell, and C. Zhang, "HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism," *Cell Reports*, vol. 6, pp. 1139–1152, mar 2014.

[189] B. L. Fogel, E. Wexler, A. Wahnich, T. Friedrich, C. Vijayendran, F. Gao, N. Parikshak, G. Konopka, and D. H. Geschwind, "RBFOX1 regulates both splicing and transcriptional networks in human neuronal development," *Human Molecular Genetics*, vol. 21, no. 19, pp. 4171–4186, 2012.

[190] M. Johnson, K. Shkura, S. Langley, A. Delahaye-Duriez, P. Srivastava, and E. Al., "Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease," *Nature Neuroscience*, vol. 19, no. 2, pp. 1–10, 2015.

[191] H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. M. Sousa, M. Pletikos, K. a. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman, and N. Sestan, "Spatio-temporal transcriptome of the human brain.," *Nature*, vol. 478, pp. 483–9, oct 2011.

[192] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.," *Nature protocols*, vol. 7, no. 3, pp. 562–78, 2012.

[193] The GTEx Consortium, "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, pp. 648–660, 2015.

[194] G. a. Doody, E. C. Johnstone, T. L. Sanderson, D. G. Owens, and W. J. Muir, "'Pfropfschizophrenie' revisited. Schizophrenia in people with mild learning disability," *The British Journal of Psychiatry*, vol. 173, pp. 145–153, aug 1998.

[195] S. Ben-Shachar, Z. Ou, C. A. Shaw, J. W. Belmont, M. S. Patel, M. Hummel, S. Amato, N. Tartaglia, J. Berg, V. R. Sutton, S. R. Lalani, A. C. Chinault, S. W. Cheung, J. R. Lupski, and A. Patel, "22q11.2 Distal Deletion: A Recurrent Genomic Disorder Distinct from DiGeorge Syndrome and Velocardiofacial Syndrome," *American Journal of Human Genetics*, vol. 82, no. 1, pp. 214–221, 2008.

[196] E. Michaelovsky, A. Frisch, M. Carmel, M. Patya, O. Zarchi, T. Green, L. Basel-Vanagaite, A. Weizman, and D. Gothelf, "Genotype-phenotype correlation in 22q11.2 deletion syndrome," *BMC Medical Genetics*, vol. 13, no. 1, p. 122, 2012.

[197] S. McCarthy, S. Das, W. Kretzschmar, R. Durbin, G. Abecasis, and J. Marchini, "A reference panel of 64,976 haplotypes for genotype imputation," tech. rep., dec 2015.

[198] A. Corvin and P. F. Sullivan, "What Next in Schizophrenia Genetics for the Psychiatric Genomics Consortium?," *Schizophrenia Bulletin*, vol. 42, no. 3, p. sbw014, 2016.

[199] R. Bernier, C. Golzio, B. Xiong, H. a. Stessman, B. P. Coe, O. Penn, K. Witherspoon, J. Gerdts, C. Baker, A. T. Vulto-van Silfhout, J. H. Schuurs-Hoeijmakers, M. Fichera, P. Bosco, S. Buono, A. Alberti, P. Failla, H. Peeters, J. Steyaert, L. E. L. M. Vissers, L. Francescatto, H. C. Mefford, J. a. Rosenfeld, T. Bakken, B. J. O'Roak, M. Pawlus, R. Moon, J. Shendure, D. G. Amaral, E. Lein, J. Rankin, C. Romano, B. B. a. de Vries, N. Katsanis, and E. E. Eichler, "Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development.," *Cell*, vol. 158, pp. 263–276, jul 2014.

[200] J. Peça, C. Feliciano, J. T. Ting, W. Wang, M. F. Wells, T. N. Venkatraman, C. D. Lascola, Z. Fu, and G. Feng, "Shank3 mutant mice display autistic-like behaviours and striatal dysfunction.," *Nature*, vol. 472, no. 7344, pp. 437–42, 2011.

[201] J. L. Brigman, C. Graybeal, and A. Holmes, "Predictably irrational: assaying cognitive inflexibility in mouse models of schizophrenia.," *Frontiers in neuroscience*, vol. 4, pp. 19–28, jan 2010.

[202] K. L. Stark, B. Xu, A. Bagchi, W.-S. Lai, H. Liu, R. Hsu, X. Wan, P. Pavlidis, A. a. Mills, M. Karayiorgou, and J. a. Gogos, "Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model.," *Nature genetics*, vol. 40, pp. 751–60, jun 2008.

[203] J. Ellegood, S. Markx, J. P. Lerch, P. E. Steadman, C. Genç, F. Provenzano, S. a. Kushner, R. M. Henkelman, M. Karayiorgou, and J. a. Gogos, "Neuroanatomical phenotypes in a mouse model of the 22q11.2 microdeletion.," *Molecular psychiatry*, vol. 19, pp. 99–107, jan 2014.