Using whole exome sequencing data to elucidate the role of structural variation in schizophrenia

A thesis submitted for the degree of Doctor of Philosophy at Cardiff University
by

Jack Bakewell
May 2023

Supervisors: Dr Elliott Rees
Professor George Kirov

Centre for Neuropsychiatric Genetics and Genomics

Acknowledgements

I'd first like to give my thanks to Elliott Rees. His competency as a researcher is nothing short of extraordinary and words fail to express how much I have learned and benefited from his guidance over the last four years, in all aspects of my work. While conducting research during the pandemic was often difficult, Elliott's remarkable dedication to my progress never faltered, and he was always able to aid in my challenges with great kindness and patience. I honestly could not have asked for a better supervisor and am glad to have the opportunity to continue working with Elliott on a new project.

I also thank George Kirov for his guidance throughout my PhD and sharing his impressive knowledge of CNV work, as well as his great sense of humour. He was also very dedicated to my progress and helped immensely with the planning of my projects and my writing. His clarity and precision was often a much need counter to my waffling tendencies, and this thesis is all the better for it.

I thank Andrew Pocklington for humouring me during my brief foray into protein-protein interaction work and his encouragement throughout my studentship. I'd also like to thank Nick Bray for his kindness and support, particularly in the early stages of my time at Cardiff, and for giving me the opportunity to undertake my PhD. The broader psychosis research group at Cardiff have also been of great benefit to my work, in addition to being fine company at The Flora on Wednesdays!

I'm extremely grateful for the encouragement of my partner Vira, and for her great patience while I completed my write up. Finally, I owe this thesis to my family: Ben, Sam, Alex, Charlotte, Ray, Pat, and my dear parents Kim and Ian. Without their unwavering love I could not have completed this work.

Many thanks to the UKRI Medical Research Council for entirely funding my PhD.

Summary

Large, rare structural variants (SVs) have consistently been shown to confer liability for schizophrenia. However, almost all previous studies have been based on data derived from genotyping microarrays, which can only be used to detect a small number of SV types and have limited utility for identifying variants at the smaller end of the size spectrum (<100kb). Therefore, I assessed whether data derived from whole exome sequencing (WES) can be used to identify SVs in schizophrenia that have hitherto gone undetected. To do this, I applied two structural variant callers, CLAMMS and InDelible, to the WES data of two in-house samples for which SVs had previously been called using array data. As each caller mines a different aspect of WES data, they are sensitive to different types and sizes of SVs.

The first WES dataset I applied these methods to is derived from a trios sample consisting of 616 schizophrenia probands and their parents. Both callers identified *de novo* SVs that were not detected in the array data, some of which overlapped genes that have been implicated in previous studies of schizophrenia or are plausible candidate risk genes. The second dataset was generated from 927 schizophrenia cases who have been extensively tested for cognitive ability. Subsets of small (<100kb), rare SVs generated by both callers were found to be associated with cognitive deficits, indicating that SVs previously undetected in the array data are implicated in schizophrenia symptomology. My thesis therefore provides evidence that WES data can be used to detect SVs under-reported in the literature that may have a role in schizophrenia.

Structure of thesis

This thesis has seven chapters: An introduction, a methods chapter, four results chapters and a final discussion. The introduction gives an overview of schizophrenia as a clinical entity and describes the most up-to-date findings regarding its genetic risk factors. It also describes genomic structural variation, including the different subtypes and mechanisms of formation, and closes with a section on the aims and objectives of my research. The methods chapter details the algorithms of the two structural variant callers used to conduct my PhD projects. The four results chapters report findings obtained from the application of the structural variant callers to the two in-house datasets, and each have their own introduction, methods, results and discission sections. The final chapter discusses how my research meets my thesis aims and objectives, and the implications of my findings for future studies of structural variation in schizophrenia.

Table of Contents

Chapter 1: Introduction

1. Schizophrenia

Psychiatric disorders are clinical conditions whose symptoms involve abnormalities in thought, perception, emotion, and behaviour, cause significant distress, and negatively impact social and occupational functioning (Kaplan & Sadock, 2015). Among the most severe is schizophrenia, a disorder which typically onsets in late adolescence or early adulthood and tends to follow a chronic course of illness (Tandon et al., 2009a). Schizophrenia was initially proposed as a separate clinical entity in the late 19th century by psychiatrist Emil Kraepelin. Kraepelin observed that there were several distinct syndromes associated with psychosis with differential patterns of onset, symptomology, and outcome (Kraepelin, 1896). He recognized that one such condition onset in youth and was characterized by a progressive deterioration of cognitive and social abilities, thus proposing the term ‘dementia praecox’ (‘premature dementia’) as its diagnostic label (Kraepelin, 1919). In the early 20th century, psychiatrist Eugen Bleuler proposed a name change, as he observed that the condition had a more heterogeneous presentation than was implied by Kraepelin’s label. Bleuler chose the term ‘schizophrenia’ (‘split psyche’), as he thought that the disorder’s hallmark feature was the disintegration of mental faculties (Bleuler, 1911). While Bleuler’s label is still in use, the diagnostic criteria for the condition underwent significant revisions over the course of the 20th century (Tandon et al., 2009a).

1.1 Symptoms

In current clinical practice, the two classification systems that are mostly commonly used to diagnose schizophrenia (and other psychiatric disorders) are the Diagnostic and Statistical Manual of Mental Disorders version 5 (Association, 2013) and the International Classification of Diseases version 10 (World Health Organization, 2019). According to both, schizophrenia is a syndromic disorder, primarily characterized by ‘positive’ and ‘negative’ symptoms (described below). While the criteria do not require that onset occurs in adolescence or early adulthood, symptoms must have been present for at least six months to warrant a diagnosis. Individuals typically do not present with all symptoms, but at least two must be present for a significant amount of time during a one-month period to meet the full criteria. When an individual presents with symptoms for the first time, it is referred to as their ‘first episode’ of the disorder, as schizophrenia is not necessarily chronic.

1.1.1 Positive symptoms

Positive symptoms are so-called because they indicate the presence of psychological characteristics that are typically not observed in healthy individuals. They are grouped into four categories: delusions, hallucinations, disorganized thinking/speech, and disorganized movement (Tandon et al., 2009a). Delusions are defined as strongly held beliefs that are maintained despite a lack of confirming evidence and/or the presence of contradicting evidence (Garety & Freeman, 1999). In the context of psychiatric diagnosis, they must also be inconsistent with an individual’s socio-cultural norms and educational background. Hallucinations are the perception of phenomena in the absence of external stimuli (Aleman & Larøi, 2008). While auditory hallucinations (typical in the form of voices) are the most common type observed in schizophrenia, visual, olfactory, and gustatory hallucinations have also been reported (Larøi, 2012). Disorganized thinking and speech typically present as difficulty staying on a topic of conversation, use of neologisms, and incoherent sentence structures (‘word salad’) (Andreasen, 1979). Disorganized movements, frequently referred to as ‘catatonic’ symptoms, are relatively uncommon (~5% prevalence; (Usman et al., 2011)). Individuals will display purposeless movements or be motionless for extended periods (Fink & Taylor, 2006). The terms ‘positive’ and ‘psychotic’ are often used interchangeably to denote the whole symptom cluster.

1.1.2 Negative symptoms

Negative symptoms are so-called because they indicate the absence of psychological characteristics that are typically observed in healthy individuals. They can be divided into two categories: deficits of volition and deficits of affect (Tandon et al., 2009b). Loss of interest in usually enjoyable activities and a lack of motivation to pursue goals are common in schizophrenic individuals (Foussias & Remington, 2010). These symptoms are usually observed alongside a blunting of emotions (affective flattening) and an inability to experience pleasure (anhedonia) (Kirkpatrick & Galderisi, 2008). Social withdrawal is typical but can be a behavioural consequence of positive symptoms (Blanchard et al., 2011). Negative symptoms are usually present before the onset of positive symptoms and are less responsive to therapeutic intervention (Möller, 2007). Although cognitive impairments can technically be defined as a negative symptom for diagnostic purposes, in clinical research they are usually studied and discussed as a separate aspect of disorder symptomology.

1.1.3 Cognitive impairments

Progressive impairments of cognitive ability were thought to be a hallmark feature of schizophrenia by Kraepelin. While contemporary studies have confirmed that schizophrenic individuals on average perform worse than healthy individuals on battery tests designed to assess a broad spectrum of cognitive domains, they also suggest that deficits are not present in all cases. In a meta-analytic review of existing literature encompassing 2,204 individuals with first-episode schizophrenia and 2,775 age- and gender-matched controls, (Mesholam-Gately et al., 2009) found that 80% of cases performed worse than controls in at least one of 10 cognitive domains: immediate verbal memory, delayed verbal memory, visual memory, processing speed, language function, visuo-spatial awareness, working memory, executive functioning, vigilance, motor coordination, social cognition and general cognitive ability (IQ). Standardised mean effect sizes (MES) ranged from -0.64 to -1.20, indicating that impairments are not restricted to a subset of domains, though were most severe in immediate verbal memory (MES = -1.20) and processing speed (MES = -0.96).

These and similar findings have led the development of cognitive test batteries specifically designed to test cognitive impairments in schizophrenia. One such battery is the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Consensus Cognitive Battery (Marder & Fenton, 2004), which incorporates 10 tests designed to measure 7 domains: processing speed, working memory, attention/vigilance, verbal learning, visual learning, reasoning, and social cognition. By assessing cognitive ability in a schizophrenia case cohort, (Lynham et al., 2018) found that cases performed ~2 standard deviations lower than healthy controls across all domains, with verbal learning and processing speed most strongly impacted.

Case/control studies of cognitive impairment may underestimate the prevalence of cognitive impairments in schizophrenia, as they do not account for cases who perform at the same or higher level than controls, but whose cognitive abilities have still been impacted by the disorder (Keefe & Harvey, 2012). (Keefe et al., 2005) investigated the proportion of 107 schizophrenia cases whose current level of cognitive function falls below their expected level of cognitive ability, based on the antecedent factors of parental education levels and the results of vocabulary/word pronunciation tests that are designed to estimate premorbid IQ. Composite scores for current cognition were generated from the results of a test battery designed to measure 7 cognitive domains. To control for demographic factors, the same analysis was conducted on 50 healthy subjects whose parental education, age, and ethnic background did not differ significantly from cases. 98.1% of cases had current cognitive abilities that fell below expectation, compared with 40% of controls, indicating that the rate of cognitive impairment in schizophrenia may indeed be higher than is suggested by case/control studies.

The onset and course of cognitive deficits in schizophrenia have been investigated by longitudinal studies. A meta-analysis by (Woodberry et al., 2008) found that by age 16, individuals who would subsequently develop schizophrenia had significantly lower IQ and motor function than those who did not, suggesting neurodevelopmental risk factors. While cognitive impairments worsen from the premorbid into the prodromal stages of the disorder, it is unclear whether cognition continues to deteriorate following the first episode. (Szöke et al., 2008) found that cognitive deficits were stable in 261 patients over a 10-year follow-up period, with no association between cognitive ability, the course of other symptoms, and therapeutic intervention. Individuals who achieved stable remission of symptoms in the first year had an improved cognitive baseline over other patients, however. In a 13-year follow-up study of 15 patients who developed the disorder between the ages 12-18 (early-onset), (Smith et al., 2009) reported a significant decline in cognitive function, particularly in the domains of verbal memory and attention. Collectively, these results suggest that the course of cognitive deficits in schizophrenia is partly mediated by adolescent neurodevelopmental changes following disorder onset (Rapoport et al., 2005), but then stabilise in adulthood.

1.1.4 Schizoaffective disorder

Schizoaffective disorder is a distinct diagnostic entity from schizophrenia. However, the symptom overlap between the disorders is such that they are thought to share many of the same underlying neurobiological dysfunctions (Malaspina et al., 2013). Schizoaffective disorder cases are therefore often combined with schizophrenia cases in research to increase statistical power at the expense of phenotypic homogeneity. Schizoaffective disorder is also characterized by positive and negative symptoms but differs from schizophrenia insofar as symptoms typically associated with affective disorders, such as major depressive disorder and bipolar disorder, are also present, however psychotic symptoms need to be present outside affective episodes for the diagnosis to be made (Malaspina et al., 2013). Individuals diagnosed with schizophrenia may later receive a diagnosis of schizoaffective disorder if affective symptoms develop in later episodes. Conversely, first-episode schizoaffective disorder may later be diagnosed as schizophrenia if affective symptoms are no longer prominent (Keshavan et al., 2011).

1.2 Epidemiology

Approximately 0.3-0.7% of the global population are thought to be affected by schizophrenia (Saha et al., 2005; Tandon et al., 2009a), and lifetime prevalence is ~0.4% (Messias et al., 2007). Reported prevalence rates vary widely across countries and ethnic groups (Jablensky, 2000), though this is thought to be primarily a consequence of variations in diagnostic criteria, healthcare access, cultural attitudes, and birth rates (higher prevalence in fast-growing populations) (Charlson et al., 2018). Lifetime incidence is associated with sex (1.42x higher in males (Aleman et al., 2003)) and urbanicity (March et al., 2008). The median age-at-onset is 25 years (Solmi et al., 2022), and stratifies by sex. Males tend to be 21-25 years at first-episode diagnosis and females 25-30 years (Li et al., 2016). Late-onset cases are more likely to be female, who have a second peak age-at-onset of ~45 years (Li et al., 2016). Despite these differences, schizophrenia has equal prevalence across both sexes (Saha et al., 2005). However, there is evidence that cognitive and negative symptoms are more severe in males (Nawka et al., 2013), while females are more likely to present with affective symptoms and therefore constitute a majority of schizoaffective disorder cases (Thara & Kamath, 2015).

1.3 Economic and social costs

The World Health Organization ranks schizophrenia as the 8th leading cause of years lived with disability (YLD), accounting for 1.1% of all YLDs in 2019 (Organization, 2019). The direct economic cost of the disorder is estimated to be ~1.5% of healthcare budgets globally (Chong et al., 2016). Indirect costs are also substantial, as ~90% of schizophrenia cases are not employed (Marwaha & Johnson, 2004) and typically require intensive support from relatives/caregivers during the more severe stages of illness (Awad & Voruganti, 2008). Individuals with schizophrenia are also at high-risk of homelessness, substance abuse, and interactions with the criminal justice system (Fazel & Grann, 2006), though the majority do not exhibit violent or anti-social behaviours (Fazel et al., 2009). They are also around 8 times more likely than non-schizophrenic individuals to live alone (Hakulinen et al., 2019) and have significantly reduced fecundity compared to healthy controls (Power et al., 2013). The significant stigma associated with severe psychiatric illness often precludes gainful employment prospects even among recovered individuals (Thornicroft et al., 2009).

1.4 Outcomes

Schizophrenia is associated with a 13- to 15-year reduction in life expectancy (Hjorthøj et al., 2017) and a lifetime risk of suicide among cases is approximately 5% (Hor & Taylor, 2010). A longitudinal meta-analysis by (Jääskeläinen et al., 2013) found that 13% of schizophrenic individuals achieved recovery, defined as clinical remission and improved social functioning that have persisted for at least 2 years. Median recovery rates were found to be fairly stable across studies (6.0% to 18.4%), with no significant stratification by sex, first-episode status, or origin of sample. The annual recovery rate was estimated to be 1.4%, and no evidence was found that recovery rates have increased over the last 50 years, despite radical improvements in psychiatric healthcare services. Around 60% of patients have been found to experience a reduction in symptoms following antipsychotic treatments (Kinon et al., 2010; Leucht et al., 2017; Suzuki et al., 2015), but rates of relapse are high (Alvarez-Jimenez et al., 2012), and improvements are largely limited to positive symptoms. A significant proportion of clinical research is dedicated to establishing biomarkers that can predict recovery (Koutsouleris et al., 2016).

1.5 Environmental risk factors

Several environmental exposures have been identified that increase schizophrenia liability, including prenatal infection and malnutrition, childhood adversity, substance abuse, urbanicity and migrant status.

1.5.1 Prenatal risk factors

Prenatal complications, typically involving prenatal infection or malnutrition, appear to play a role in schizophrenia. (Brown et al., 2004) noted a three-fold increase in schizophrenia risk, compared to controls, following exposure to influenza in early gestation. Similarly, (Babulas et al., 2006) reported a five-fold increased risk in offspring exposed to maternal genital and reproductive infections in the periconceptional period (but not later in gestation) compared to controls. In one of the most famous studies of psychiatric risk factors, (Hoek et al., 1998) found a two-fold increased schizophrenia liability in the offspring of mothers who were pregnant during the 1944-45 'Dutch Winter Hunger' - a famine imposed on the Netherlands by a German trade embargo during the Second World War - compared to Dutch reference cohort born during the same period. These findings were replicated in Chinese individuals born during the 1959–1961 Chinese famine, which resulted from disruptions to agriculture during the 'Great Leap Forward' (St Clair, 2005), suggesting that parental malnutrition is a risk factor shared across ethnic groups.

1.5.2 Childhood adversity

Childhood adversity (CA) has been reported by several studies to increase schizophrenia liability. In a meta-analysis of 12 case/control studies,(Varese et al., 2012) showed that individuals with a psychotic disorder, including schizophrenia, were 2.72 times more likely to have experienced any type of CA than controls. The largest impact was for emotional abuse (OR: 3.40). In a study of ~200 schizophrenia cases, (Larsson et al., 2013) found that 82% had experienced some form of CA, the most common subtype being emotional neglect (65%), though they did not compare with a healthy control group. Assessing the impact of CA on symptom dimensions, (Lee et al., 2018) reported associations with negative symptoms and cognitive impairments, but not with positive symptoms. In contrast to other findings (Bendall et al., 2008; Morgan & Fisher, 2007), CA was not found to be correlated with symptom severity. CA was negatively correlated with global functioning (r = -0.109) in a meta-analysis (Trotta et al., 2015), with emotional neglect producing the largest effect (r = -0.250). Correlations were reported for social functioning dimensions tested separately, but not occupational functioning.

1.5.3 Substance abuse

Cannabis use has been consistently reported as one of the strongest environmental risk factors for schizophrenia, particularly during adolescence. A meta-analysis by (Marconi et al., 2016) found that individuals who had ever used cannabis had a two-fold increased risk compared to controls, and an accelerated rate of onset. The association is dose-dependent, with up to a five-fold higher risk reported in individuals using high-potency medicinal tetrahydrocannabinol (THC), the psychoactive constituent of cannabis (Di Forti et al., 2015). Among schizophrenia cases, cannabis use has been associated with more severe positive symptoms, but not with negative symptoms. Some studies have reported an association between frequent alcohol use and schizophrenia (Jones et al., 2011), though these findings have not always been replicated (Hartz et al., 2014). There is also evidence that alcohol abuse can worsen the course of the disorder (Regier et al., 1990). It is possible that there is also an association between nicotine use and schizophrenia risk; a systematic review by (Mustonen et al., 2018) identified 6 longitudinal studies that reported increased nicotine use among individuals that would later develop schizophrenia. However, as schizophrenia tends to have a long prodromal phase and an extremely high proportion of cases self-medicate with nicotine (de Leon & Diaz, 2005), causal relations are difficult to assess.

1.5.4 Urbanicity and migrant status

Degree of urbanicity, measured in terms of population density, is associated with an approximately 1.5- to 4-fold increase schizophrenia (Kelly et al., 2010; Kirkbride et al., 2017; March et al., 2008). These findings have been consistent across geographical locations and ethnic groups. In a meta-analysis of 4 population studies, (Vassos et al., 2012) found that incidence of schizophrenia was 2.27 higher in the most urban location compared with the most rural location. Moreover, migrant/ethnic minority status confers a 2- to 3-fold higher risk for schizophrenia than native-born status and is consistent across different migrant groups and their countries of destination (Anderson et al., 2015; Bourque et al., 2011; Cantor-Graae & Selten, 2005). There is some evidence that risk is stratified by ethnic group, however, with migrants from African and Caribbean backgrounds having the highest incidence rates (Coid et al., 2008), though these differences may reflect access to care and diagnostic inaccuracies resulting from cultural and linguistic barriers (McGrath et al., 2004).

1.6 Neurobiology

The neurobiology of schizophrenia is highly complex, involving the dysfunction of multiple brain regions and neural pathways. Several approaches have been used to investigate it, including neuroimaging, post-mortem studies, pharmacological studies, animal models, cellular models, and genetic studies. In this section I summarise the most important findings and explain how they relate to current aetiological hypotheses.

1.6.1 Brain structure abnormalities

No specific brain structure abnormality has been observed in all schizophrenic study participants, reflecting an underlying heterogeneity to the disorder. However, neuroimaging and post-mortem neuropathology studies have detected several abnormalities that have proven to be replicable.

1.6.1.1 Enlarged lateral ventricles

Enlarged lateral ventricles are among the most robust findings in schizophrenia. Based on structural magnetic resonance imaging (MRI) data, (Nakamura et al., 2007) found that first episode cases have marginally larger ventricles at baseline compared to healthy controls, and a 10.4% ventricular enlargement after 1.5 years. A meta-analysis by (Erp et al., 2016) reported an 18% ventricular enlargement compared to healthy controls, in cases whose mean duration of illness was 10 years. Together, these results suggest a progressive loss of brain tissue following disorder onset. Enlarged ventricles have been associated with more severe negative symptoms and poor treatment response (Lieberman et al., 2005), though other studies have contradicted these findings (Gur et al., 2000; Ho et al., 2003).

1.6.1.2 Grey and white matter volume

A volumetric reduction in both grey and white matter volume is another frequently reported brain structure abnormality. A meta-analysis of longitudinal MRI studies by (Vita et al., 2012) found significant volumetric reductions over time of whole brain grey matter in chronic schizophrenia cases. Compared to controls, the annualized percentage change (APC) was −0.59%. Significant reductions were also found in frontal grey (APC = −0.74%) and white (APC = −0.32%) matter, parietal white matter (APC = −0.32%), and temporal white matter (APC = −0.39%). Another meta-analysis of longitudinal MRI studies reported a progressive volumetric reduction only in the right anterior cingulate grey matter in high-risk individuals through transition into first-episode schizophrenia (Liloia et al., 2021), suggesting that some structural abnormalities may be more prevalent at different stages of illness. It has been suggested that these findings may be confounded by the impact of antipsychotic medications. However, stability or increases in grey matter volume following antipsychotic treatments have been reported by multiple studies (Haren et al., 2011; Lieberman et al., 2005; Navari & Dazzan, 2009), indicating that volumetric reductions are indeed due to disorder progression and not psychiatric interventions.

1.6.1.3 White matter integrity

White matter integrity (WMI) can be investigated using diffusion tensor imaging (DTI) and serves as a proxy for the connectivity of different brain regions (D. K. Jones et al., 2013). DTI studies have revealed widespread significant reductions in WMI, involving the whole cortex and many sub-cortical structures, compared to controls. Corpus callosum, superior longitudinal fasciculus, cingulate, and thalamic radiations are consistently noted as the most severely affected tracts (Ellison-Wright & Bullmore, 2009; Friston, 2011; Karlsgodt et al., 2008) and have been found to be impacted in both high-risk individuals and first-episode schizophrenia (Peters et al., 2010; Samartzis et al., 2014). As in the case of volumetric reductions, antipsychotic treatment has been associated with increased WMI in the cingulate and superior longitudinal fasciculus between cases (Wang et al., 2013). WMI correlates with cognitive impairments in schizophrenia, particularly processing speed and working memory (Kochunov et al., 2017), while disruption of thalamic radiations is thought to be significant in the development of positive symptoms (Jiang, Patten, Zarakenho, 2021).

1.6.1.4 Subcortical abnormalities

The impacts of schizophrenia on sub-cortical brain structures have also been studied. A meta-analysis of MRI studies noted significant volumetric reductions in the hippocampus (-4.10%), amygdala (-3.80%), thalamus (-2.74%), and nucleus accumbens (-3.69%) in chronic cases compared to controls (Haijma et al., 2013). A significant volumetric increase in the pallidum (2.28%) was also noted and found to be associated with age and illness duration, which may reflect the impact of antipsychotic treatment. In first-episode cases, reductions have been reported in the volumes of the hippocampus, amygdala, and thalamus, compared to controls (Takahashi et al., 2006), suggesting that these regions are among the first to be impacted. No association has been found between any measure of volumetric reduction and antipsychotic treatment (Haijma et al., 2013).

1.6.1.5 Aetiology of brain structure abnormalities

There has been much debate over whether structural brain abnormalities in schizophrenia represent neurodevelopmental or neurodegenerative processes. Many abnormalities are evident prior to symptom onset, indicating a neurodevelopmental component, while progressive ventricular enlargements and grey matter reductions in chronic cases suggest a neurodegenerative pathology. However, symptoms tend to be episodic or stable, inconsistent with the progressive deterioration observed in neurodegenerative disorders. Post-mortem histological studies have also found no evidence of gliosis in schizophrenia cases (Harrison, 1999), a neuropathology thought to be the hallmark of neurodegeneration (Garden & Campbell, 2016). Thus, structural abnormalities likely have a neurodevelopmental aetiology (Rapoport et al., 2012) that is exacerbated by the behavioural/functional consequences of symptoms (social withdrawal, unemployment, substance abuse, etc.) (Li et al., 2017; Zhang et al., 2015).

1.6.2 Neurochemical dysfunctions

While neurochemical dysfunctions in schizophrenia are also heterogenous, there is broad support for the involvement of dopamine, glutamate and GABA neurotransmitter systems in the aetiology, progression, and treatment of the disorder.

1.6.2.1 Dopamine

All currently prescribed antipsychotic medications are dopamine D2 receptor antagonists (Howes & Kapur, 2009), and drugs that increase transmission at dopaminergic neurons (e.g., amphetamine and L-DOPA) can worsen positive symptoms in cases (Lieberman et al., 1990) and induce schizophrenia-like symptoms in healthy controls (Angrist et al., 1980). Genetic studies have consistently reported strong associations with variants intersecting the D2 receptor gene DRD2 gene (Ripke et al., 2014), which are thought to increase its expression (Y. Zhang et al., 2007). These receptors are preferentially expressed in the mesolimbic pathway, projecting from the midbrain ventral tegmental area to the striatum (Grace, 2016) . Positron emission tomography (PET) studies have reported elevated activity of this pathway in schizophrenia cases, compared to controls, which correlates primarily with the severity of positive symptoms (Howes et al., 2012). There is also evidence that hypo-activity of D1 dopaminergic neurons in the frontal cortex is involved in schizophrenia. PET studies have shown reduced activity in frontal cortical regions in cases compared to controls that correlates with low dopamine metabolite levels in cerebrospinal fluid (CSF) ((Goldman-Rakic et al., 2000) and is associated with negative symptom severity and cognitive impairment (Abi-Dargham et al., 2002). Animal studies have revealed that lesions in the frontal cortex produce elevated activity in the striatum (Peters et al., 2016), providing evidence that cortical and subcortical dopamine dysfunction observed in schizophrenia are causally related. Injection of THC has been associated with higher mesolimbic dopamine activity and reduced striatal dopamine reuptake in both imaging and animal studies (Bloomfield et al., 2014; Bossong et al., 2009), suggesting that dopamine dysfunctions may also account for the relation between cannabis use and schizophrenia.

1.6.2.2 Glutamate

The role of glutamate in schizophrenia was hypothesized from the observation that the NMDA receptor antagonists ketamine and PCP induce psychological and behavioural states in healthy controls that closely resemble both positive and negative schizophrenia symptoms (Krystal et al., 1994). (Kim et al., 1980) noted a reduction of CSF glutamate metabolites in cases, though these findings have not been replicated (Goff & Wine, 2008). Systemic injections of NMDA receptor antagonists have been shown to increase cortical glutamate levels in animal models

(Jeevakumar & Kroener, 2016; Moghaddam et al., 1997), which correlates with abnormal motor function, and cognitive and social impairments, suggesting that hyperactivity of glutamatergic neurons could underlie schizophrenia symptoms. NMDA receptors are an essential modulator of most forms of synaptic plasticity, which is thought to be crucial for learning and memory formation (Lau & Zukin, 2007), and therefore can likely explain most forms of cognitive impairment in schizophrenia (Snyder & Gao, 2013). Genetic studies have shown a strong association between variants impacting *GRIN2A*, which codes for an NMDA receptor subunit, and schizophrenia case status (Pardiñas et al., 2018). As ~90% of neurons in the human brain are glutamatergic, it is still unknown which regions and pathways are most affected in schizophrenia, though it is probable that glutamate dysfunction accounts for the global abnormalities in brain structure observed in cases (Moghaddam & Javitt, 2012).

1.6.2.3 GABA

Dysfunction of GABA-ergic inhibitory interneurons has also been implicated in schizophrenia, and are thought to exacerbate aberrant glutamate activity. Post-mortem transcriptional studies have consistently noted a reduction of GAD67 mRNA in the frontal cortex of schizophrenia individuals (Curley et al., 2011; Hashimoto et al., 2008). The GAD67 enzyme metabolizes glutamate into GABA, which in turn inhibits the activity of glutamatergic neurons, thereby serving as an important modulator of glutamate neurotransmission (Möhler, 2012). Lower levels of GABA membrane transporter GAT1 and a lower density of fronto-temporal GABA-ergic interneurons have also been observed (Glausier & Lewis, 2011; Katsel et al., 2011). A reduction in CSF GABA concentrations has been reported in first-episode cases, compared to controls (Ongür et al., 2010), and drugs that increase GABA activity, such as benzodiazepines, can relieve negative and cognitive symptoms that are typically refractory to typical antipsychotic treatments (Lavoie et al., 2007).

1.6.2.4 Aetiology of neurochemical dysfunctions

It is still unclear if or how dopamine and glutamate-GABA dysfunctions are causally related, with some studies suggesting that they underlie different symptom clusters or may even be associated with distinct subtypes of schizophrenia (Howes & Kapur, 2009; Javitt, 2010). While subcortical dopamine abnormalities account for positive symptoms, they do not seem to directly explain the occurrence of negative symptoms or the most prevalent cognitive impairments (verbal memory and processing speed). Glutamate-GABA dysfunctions can account for a wider range of symptoms, particularly in relation to synaptic plasticity, but there is no evidence they are correlated with mesolimbic pathway activation in cases (Stone, 2011), and they are not targeted by any approved antipsychotic treatments (Moghaddam & Javitt, 2012). However, there are bidirectional projections between glutamate and GABA neurons in the frontal cortex and the midbrain VTA (Morales & Margolis, 2017), suggesting that primary dysfunctions could occur at either region before propagating to the other in a manner that is heterogeneous between cases (Lisman et al., 2008). It is also possible that glutamate-GABA dysfunctions represent the more progressive aspects of disease aetiology underlying chronic illness, while dopamine dysfunctions occur more acutely and periodically, thus explaining the typically episodic trajectory of positive symptoms (Carlsson & Carlsson, 2006).

1.6.3 Neuroinflammation

Recent studies have found increasing support for the role of neuroinflammation in schizophrenia. The disorder has been associated with increased risk of autoimmune diseases, including type 1 diabetes and multiple sclerosis (Eaton et al., 2006). Treatments that reduce immune activity have been found to reduce symptom severity (Frydecka et al., 2018; Kroken et al., 2018; Sommer et al., 2014), though these findings have proven difficult to replicate (Müller et al., 2015). Elevated levels of neuroinflammatory markers, such as interleukin-6, have also been detected in blood of both first episode and chronic schizophrenia cases, compared to controls (Goldsmith et al., 2016; Miller et al., 2011). In genetic studies, some of the strongest common variant associations to date occur within the major-histocompatibility-complex coding region on chromosome 6 and are thought to increase expression of genes involved in complement system activation (particularly C4) (Sekar et al., 2016). Neuroinflammation has been associated with many psychiatric disorders, and therefore likely underlies or exacerbates a wide range of symptoms not limited to schizophrenia (Khandaker et al., 2015).

1.7 Genetics

Family, twin and adoption studies have shown that schizophrenia is a highly heritable disorder (Kendall et al., 2017). Early family studies from showed that risk increases with closer familial relationship to probands (Gottesman & Shields, 1972) This has been confirmed more recently in large population-based studies, e.g. (Lichtenstein et al., 2009). Estimates of disorder heritability can be derived from comparisons of phenotypic concordance between sets of monozygotic and dizygotic twin pairs, under the assumption that each pair are exposed to similar environmental risk factors. If a disorder has a strong genetic component (high heritability), concordance will be higher among monozygotic than dizygotic twin pairs, given that they share an additional 50% of their DNA. A meta-analyses 12 twin studies reported that genetic variance accounts for an estimated 81% of schizophrenia liability (95% CI: 73%-90%) (Sullivan et al., 2003). Heritability can also be estimated from the phenotypic concordance of diagnosed parents and biological offspring adopted into a different household soon after birth. In this case, genetic overlap is 50%, but environmental exposures can be dissimilar. Disorders with high heritability will again show high rates of concordance, given that the variance contributed by environmental risk is low. (Tienari et al., 2004) noted a 10-fold increased risk of schizophrenia and related psychotic disorders in the adopted-away offspring of 145 diagnosed mothers, compared with adopted-away children of mothers with no psychiatric disturbance, again demonstrating that schizophrenia has a strong genetic basis.

1.8 Genetic architecture

The discovery that schizophrenia is largely shaped by genetic factors has led to an explosion of research investigating its molecular genetic basis, an effort that has now been ongoing for several decades (Sullivan et al., 2012). Recent findings have been driven by rapid advancements in DNA processing technologies, including the development of genotyping microarrays and high-throughput sequencing platforms (Mardis, 2008a). Consistent with early hypotheses based on symptom and demographic heterogeneity (Gottesman & Shields, 1972), both common and rare variant studies have revealed that the disorder is highly polygenic; resulting from the additive effects of likely thousands of genetic risk factors (Owen et al., 2016). Every individual in the general population has some genetic liability to schizophrenia but will only develop the disorder if the quantity and/or effect sizes of risk factors they carry (in addition to environmental exposures) are large enough to result in clinically significant neurobiological dysfunction. This is known as the liability threshold model of disease, first proposed for complex traits in the 1960s (Falconer, 1965) (Figure 1.1). Rare variants are more likely to have larger effects than commons ones, due to the impact of selection pressures (Figure 1.2) (Sullivan et al., 2012). Fewer rare variants are therefore required to cause illness, and burden of rare variants is associated with illness severity (Zoghbi et al., 2021). Several methods/study designs have been used to elucidate schizophrenia genetics, though in contemporary research, the predominant and most successful approach is the case/control association study (Visscher et al., 2017).

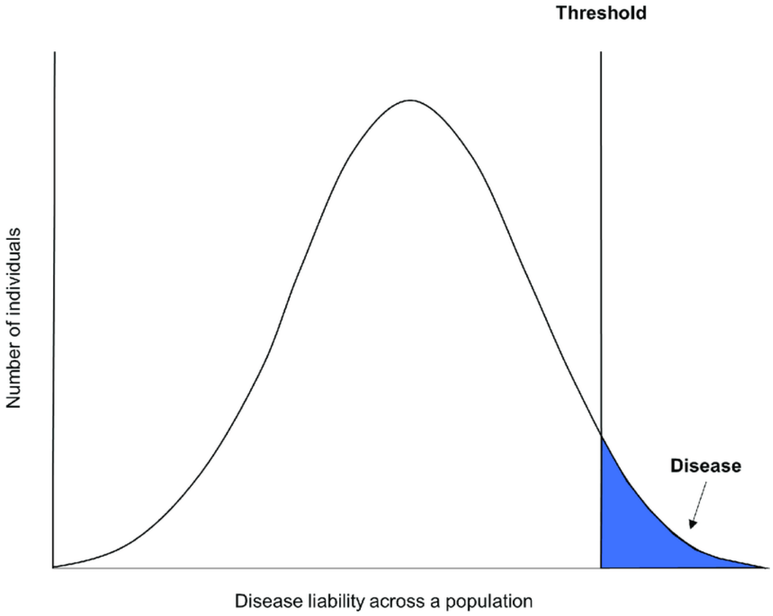


Figure 1.1. The liability threshold model of disease. Adapted from (Howe et al., 2018)

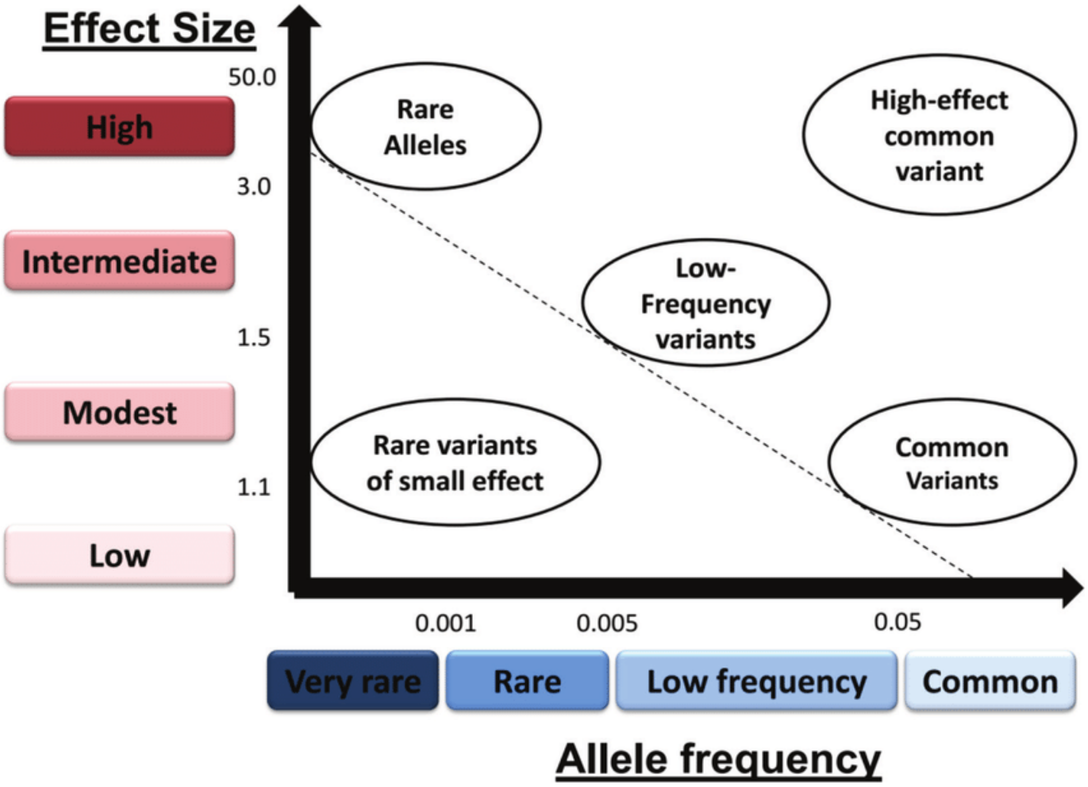


Figure 1.2. Association between allele frequency and effect size. Adapted from Junior et al. (2017)

1.8.1 Common risk variants

1.8.1.1 Genome-wide association studies

Genome-wide association studies (GWAS) assess the relative incidence of single nucleotide polymorphisms (SNP) in cases and controls to determine disease association and have been extremely successful for the detection of schizophrenia risk variants with population frequencies > 1% (Ripke et al., 2014). Typically, SNP genotypes for each study participant are derived from genotyping microarrays and filtered according to minor allele frequency (MAF) to isolate common variants (Manolio et al., 2009). Imputation is then applied to genotype data using known haplotype structures to increase the number of variants available for analysis (Marchini & Howie, 2010). For each SNP, a logistic regression model is constructed to estimate the probability of schizophrenia case status as a function of genotype and relevant covariates (usually age, sex, and ancestry) (Sullivan et al., 2012). As >1 million SNPs are tested simultaneously, the Bonferroni-corrected p-value threshold for statistical significance is set to 5 x 10^{-8} (Sullivan et al., 2012). This threshold has generated replicable genome-wide significant hits in independent studies and in meta-analyses (Ripke et al., 2014).

1.8.1.2 PGC3 Schizophrenia GWAS

GWAS sample sizes have increased substantially over the last decade, increasing the power of studies to detect schizophrenia-associated variants. The latest large-scale schizophrenia GWAS was conducted by the Psychiatric Genetics Consortium (PGC) and included 76,755 cases and 243,649 controls of predominantly European ancestry (Trubetskoy et al., 2022). Testing SNPs with MAF > 0.05, 287 genomic-wide significant hits were reported at independent (separately heritable) loci, concentrated in genes that are most highly expressed in glutamatergic neurons, particularly in frontal cortex and hippocampus, and in cortical and striatal inhibitory interneurons. These results are consistent with both the dopamine and glutamate-GABA hypotheses of schizophrenia neurobiology.

Likely causal variants were ascertained from the raw GWAS findings using fine-mapping, transcriptomic and functional genomic analyses. Fine-mapping aims to isolate causal variants by separate testing of SNPs within an associated locus while accounting for the effects of linkage disequilibrium (LD), i.e. the co-segregation of variants in populations based on haplotypes (Wang & Huang, 2022). SNPs that occur in multiple, overlapping blocks of LD variants in the locus, each of which are independently associated with the phenotype in question, are more likely to be causal than SNPs that occur in just one. Seventy genes that contained variants refined by fine mapping were prioritised for further investigation.

Putative causal SNPs were also determined based on their occurrence within expression quantitative trait loci (eQTLs), i.e., regulatory regions that dictate mRNA expression levels (Albert & Kruglyak, 2015). Summary-based Mendelian randomization (SMR) (Zhu et al., 2016) was used to ascertain 55 GWAS hits that co-localize with eQTLs for genes expressed in adult or fetal brain, or in whole blood, giving a total of 120 unique prioritized genes impacted by putatively causal variants, of which 106 were protein-coding. The prioritized genes were enriched for genes expressed at the synapse, including voltage-gated calcium and chloride channel subunits, NMDA and metabotropic receptors, and genes that play a role in endocytosis and synaptic organization.

1.8.1.3 SNP-based heritability and polygenic risk scores

The most strongly associated risk SNP identified by the PGC3 schizophrenia GWAS has an odds ratio of 1.23. Such modest effect sizes are not unexpected, given that all genotypes of tested SNPs have a population frequency of at least 0.05. However, when effect sizes of risk SNPs are combined, they explain ~25% of schizophrenia heritability. This value will continue to increase as studies become more highly powered but will always fall short of the actual heritability conferred by SNPs as GWAS do not account for the effects of SNP interactions. Moreover, additional heritability will be explained by rare SNPs and non-SNP genetic factors, such as indels and structural variants. The difference between heritability that can be currently explained by all known genetic risk factors and that approximated by twin studies (~80% for schizophrenia) is known as 'missing heritability' (Owen & Williams, 2017).

An individual’s genetic liability for schizophrenia, known as a polygenic risk score (PRS), can be calculated by combining the effect sizes of all risk SNP alleles they carry (Purcell et al., 2009). Variance explained by PRS will differ according to the GWAS p-value threshold used for SNP inclusion. PGC3 GWAS found that including SNPs with $p < 0.05$ produces PRS that can explain an average of 0.073 of variance in schizophrenia liability across test samples, while PRS based only on genome-wide significant SNPs explained an average of 0.024. This indicates that there are many SNPs contributing to schizophrenia heritability that are yet to meet the genome-wide significance threshold. PRS was able to explain most variance in liability in samples of European ancestry. This is expected given the ancestry make-up of the GWAS but does show that common schizophrenia variants differ between ancestries.

1.8.2 Rare risk variants

1.8.2.1 Next generation sequencing

Due to cost and technical limitations, genotyping arrays are not designed to capture rare SNPs (Brady & Vermeesch, 2012). Moreover, the rarity of SNPs negatively correlates with their occurrence in haplotype blocks, such that imputation cannot be used to infer their presence. The advent of high-throughput, next-generation sequencing (NGS), however, has enabled the rapid and efficient detection of rare

SNPs (and other variant types, including indels and structural variants) without the need for imputation (Mardis, 2008b). In the context of genomics, NGS has two subtypes: whole exome (WES) and whole genome (WGS) sequencing. To date, most rare variant studies have been based on WES, as many rare schizophrenia risk variants are thought to occur in protein-coding regions (Sullivan et al., 2012). The exome constitutes only 2% of the human genome (Ng et al., 2010) making WES a cost-effective approach in terms of data storage and computation. As all my PhD research projects used data generated by WES, I describe the technical aspects of a typical WES run in section 2.4. The use of WGS to investigate genetic risk factors is still in its infancy as the quantity of data required for adequately powered studies is in most cases so large that the associated costs still outweigh potential benefits (Lappalainen et al., 2019). In this section, I describe findings regarding the contribution of rare SNPs and indels to schizophrenia risk. Structural variants have also been strongly implicated and I describe their role in section 2 of the current chapter.

1.8.2.2 Single-nucleotide variants and indels

GWAS are currently underpowered to detect rare schizophrenia risk SNPs and indels at whole-genome significance level. Nevertheless, case-control association analysis can still be used to detect rare risk variants if studies are appropriately designed to maximise variant effect sizes. To this end, researchers use two main approaches, often in conjunction: 1) test association with overall burden of variants, rather than each variant individually; 2) test only those variants that exhibit features known to increase deleteriousness. Approach 1 combines the effect sizes of individual variants while 2 ensures individual variants are only tested if there is reason to believe their functional impact will be sufficiently deleterious. Both approaches carry the additional advantage of limiting multiple testing burden.

In one of the largest WES-based study of rare schizophrenia risk variants to date, the schizophrenia exome meta-analysis (SCHEMA) consortium implemented both approaches successfully (Singh et al., 2022). SNPs and indels with a minor allele count ≤ 5 were identified in 24,248 cases (~1/3 the number included in the PGC3 schizophrenia GWAS) and 97,322 controls of predominantly European ancestry, and only retained for testing if they produced a premature stop codon (i.e. protein-truncating, PTV), or an amino acid substitution that is likely to disrupt protein function (damaging missense). Schizophrenia cases were found to have a significantly higher burden of PTVs across 3,063 genes that are highly intolerant to loss-of-function (LoFi; evidenced by a lower PTV rate than would be expected by chance) ($p = 7.6 \times 10^{-35}$; odds ratio = 1.26). Moreover, PTV and damaging missense burden for 18,321 genes was tested for association with schizophrenia case status. Although 5.6 million variants were included in this analysis, only 23,321 independent tests were carried out, giving a Bonferroni corrected p-value threshold of 2.14×10^{-6}.

Variant burdens for 10 genes were found to be statistically significant at this threshold. All are highly intolerant to loss-of-function, evidenced by a much lower PTV rate than would be expected by chance. The highest reported odds ratio was 44.2, for 11 PTV and damaging missense variants in *CUL1,* demonstrating the large effect size of rare variants that alter protein coding. The 10 genes coded for calcium and NMDA receptor subunits, regulators of neuronal migration growth, transcriptional regulators, nuclear transport proteins, and ubiquitination proteins. Two had also been implicated by PCG3 GWAS loci, indicating a partial convergence of rare and common risk factors.

In an independent sample of 11,580 schizophrenia cases and 10,555 controls, (Liu et al., 2023) also found that schizophrenia cases carried a significantly higher burden of PTVs, but limited variants to those occurring in 80 LoFi genes produced by a data-driven algorithm that prioritises genes previously implicated in schizophrenia ($p = 5.4 \times 10^{-6}$, odds ratio = 1.48). They argued that the higher effect size compared to the equivalent SCHEMA analysis demonstrates the efficacy of this variant prioritisation approach. The study also meta-analysed their own sample with that of SCHEMA to assess schizophrenia PTV burden in the 80 prioritised genes across ancestry groups and found that PTV burden was largely consistent across 5 diverse groups. This suggests that largely the same genes may be implicated by rare schizophrenia risk variants across ancestries, in contrast with the stratified genetic architecture of common risk variants.

1.8.2.3 De novo variants

Some rare variants arise spontaneously in the germline and are therefore not inherited from either parent. These *'de novo'* variants are more likely to be deleterious than transmitted ones, as they are yet to undergo selection pressure (Veltman & Brunner, 2012), and therefore present an opportunity to elucidate the genomic factors underlying disease. *De novo* variant studies typically have a ‘trio’ design, involving the comparison of variants called in schizophrenic probands and both parents to determine their transmission status. In WES data for 617 schizophrenia trios and 713 control trios, (Fromer et al., 2014) found that *de novo* SNPs and indels were not significantly enriched in proband cases compared to controls. However, genes that had previously been implicated in schizophrenia were significantly enriched for non-synonymous (protein altering) *de novo* mutations ($p = 7 \times 10^{-4}$) in cases, as were genes expressed at the post-synapse of glutamatergic neurons ($p = 0.019$), and specifically those encoding NDMA receptor ($p = 0.025$) and the ARC complexes ($p = 4.8 \times 10^{-5}$).

Among 606 coding *de novo* variants called from WES data for 613 probands, (Rees et al., 2020) reported a significant excess of *de novo* PTVs affecting 3,471 LoFi genes ($p = 2.3 \times 10^{-3}$), but not loss-of-function tolerant genes. These data were combined with previously studied WES data for 2,831 trios to investigate the gene burden of *de novo* PTVs. While none were found to be significant after multiple testing correction, the most strongly associated gene (*SETD1A*) had been previously implicated in schizophrenia (Singh et al., 2016). The second most strongly associated gene (*CUL1*) was novel in this study but would later be confirmed to harbour rare schizophrenia risk variants by (Singh et al., 2022). This demonstrates that genes most strongly affected by *de novo* mutations in schizophrenia probands are likely to contribute to disorder risk.

1.8.3 Pleiotropy

Both common and rare genetic risk factors for schizophrenia also confer liability to other psychiatric and neurological disorders. (Anttila et al., 2018) estimated common variant correlations (rg) between 42 traits by comparing the extent of LD between SNPs that are associated with different traits and found that schizophrenia is most correlated with bipolar disorder (rg = 0.68), major depressive disorder (rg = 0.34) and obsessive-compulsive disorder (rg = 0.33). (Howrigan et al., 2020) analysed *de novo* variants in ~3,000 schizophrenia probands and found that they are significantly enriched for genes that that have been shown to harbour *de novo* variants in other neurodevelopmental disorders (NDD). (Rees et al., 2021) showed that genes associated with *de novo* variants in NDDs were enriched for *de novo* PTVs in ~3,500 schizophrenia probands ($p = 3.3 \times 10^{-11}$), and that the set of *de novo* PTVs and damaging missense variants identified in NDD probands were significantly enriched in the schizophrenia probands ($p = 5.0 \times 10^{-6}$). Thus, not only are the same genes implicated by rare *de novo* variants in both disorders, but those genes also tend to be disrupted in the same ways.

These findings imply that the current diagnostic boundaries for schizophrenia do not reflect the underlying neurobiology of the disorder, which is largely shared with that of other disorders (Owen, 2012). While this may not impact the clinical efficacy of current diagnostic systems (at least in the short term), it does suggest that a more dimensional approach to psychiatric disorders may improve the validity of genetic research. This may entail the grouping of cases according to symptoms that are shared between disorders (e.g. delusions in schizophrenia and bipolar disorder), so that the mechanisms underlying them may be more precisely investigated, which could then be used to facilitate the development of more finely targeted therapeutics (Cuthbert & Insel, 2013).

1.9 Summary

In this section I have given an overview of schizophrenia, including descriptions of its symptoms, epidemiology, neurobiology, and known environmental and genetic risk factors. To summarise, schizophrenia is severe psychiatric disorder characterised by positive, negative, and cognitive symptoms. It has a global prevalence of ~0.5% and tends to follow a chronic course, though 30% of cases achieve lasting recovery. Studies of schizophrenia neurobiology have converged on glutamate and dopamine pathway dysfunctions as primary aetiological mechanisms, though much is still unknown about relation between these dysfunctions and symptomology. Environmental risk factors include prenatal complications, childhood adversity, substance abuse and urbanicity. However, as twin studies have suggested that schizophrenia’s heritability is ~80%, most of the disorder’s liability is driven by genetic factors. Schizophrenia is highly polygenic, with a genetic architecture shaped by potentially thousands of common and rare genetic risk factors. Associated variants are enriched for genes expressed in the postsynaptic density of glutamate neurons, and whose roles include synaptic organisation and plasticity, neurotransmission, transcriptional regulation, and ubiquitination. *De novo* variants have also been implicated. In the next section I describe structural variants and their role in schizophrenia aetiology, as all my PhD research involved calling structural variants in schizophrenia cases.

2. Structural variants

Structural variants (SVs) are alterations to the genome that are typically defined as larger than 50 base pairs (bps) (Feuk et al., 2006). In recent years, however, this class has been thought to apply to all variants larger than SNPs, Indels, and small tandem repeats (> ~10bp). They can be grouped into two broad types: 'balanced' and 'unbalanced'. Unbalanced SVs involve deviations from the normal diploid allele quantity, and therefore also known as copy number variants (CNVs) (Redon et al., 2006). Heterozygous deletions and duplications are the most common CNVs (Conrad et al., 2010), though homozygous events and triplications have been identified (Itsara et al., 2009). Balanced SVs are changes in the location (translocation) or orientation (inversion) of a sequence, relative to a reference genome, that do not affect allele quantity. Some sequences known as retrotransposons can be copied or extracted from the genome and inserted at a different locus by cellular machinery; such events are called 'retrotranspositions' and can be unbalanced or balanced (Burns & Boeke, 2012). SVs are not distributed randomly across the genome and are generally 'recurrent' or 'nonrecurrent'. Recurrent SVs constitute ~60% of all events and have approximately the same start and end bases (breakpoints) across unrelated individuals, regardless of allele frequency (Sharp et al., 2006). Breakpoints of non-recurrent SVs, on the other hand, occur at different loci across unrelated individuals and are often unique (Weckselblatt & Rudd, 2015).

2.1 Mechanisms of formation

The mechanisms of SV formation vary by SV type and recurrence. CNVs and inversions most commonly arise through non-allelic homologous recombination during cell division (Stankiewicz & Lupski, 2010), while translocations are typically formed through DNA repair mechanism know as non-homologous end-joining (Nambiar & Raghavan, 2011). DNA Replication-based mechanisms, such as fork stalling and template switching and microhomology-mediated break-induced replication can produce complex sequence rearrangements that involve combinations of SV types (Burssed et al., 2022), while retrotranspositions arise exclusively from the activity of retrotransposon machinery (Kazazian Jr., 2004). All mechanisms of SV formation have been observed in somatic and germline cells, though are more likely to occur in the latter as meiotic cell division is a more complex process than mitosis (Abyzov et al., 2016).

2.1.1 Non-allelic homologous recombination

Non-allelic homologous recombination (NAHR) is the exchange of highly similar, but non-allelic sequences between during cell division, and can occur within or between chromatids. Sequences that are most prone to NAHR are found between low-copy repeats (LCRs), defined as DNA stretches >1kb in size that are constituted by successive, highly homologous (>90%) sequences. LCRs make up ~5% of the human genome (Stankiewicz & Lupski, 2010). NAHR can also result from Alu short-interspersed nuclear elements (SINEs), a type of retrotransposon that is highly abundant (~11% of human genome), typically around 300bp in length and have >80% sequence homology (Deininger, 2011). Long-interspersed nuclear elements (LINES) are another type of retrotransposon that can underlie NAHR but are less prone due to their length (~6kb) and lower sequence homology (~70%) (Cordaux & Batzer, 2009). Generally, shorter repeat elements with higher sequence similarity align/mispair more easily, increasing the chance of NAHR.

SV size is determined by the relative distance between the mis-paired repeated sequence and its corresponding allele, while type is primarily determined by chromatid orientation (Chen et al., 2014). When non-allelic homologous sequences on sister chromatids are positively orientated (i.e in the same direction), recombination simultaneously produces deletion and duplication events (Figure 1.3, A). When they are negatively oriented, a deletion and inverted duplication occur. Intra-chromatid recombination results in a deletion if sequences are positively orientated and an inversion if they are negative orientated (Figure 1.3, A). Homologous sequences can be exchanged between chromatids on different chromosomes, producing translocation events that may be balanced or unbalanced. Balanced (or ‘reciprocal’) translocations involve an equal exchange of DNA between chromosomes, with no net loss or gain, while unbalanced translocations involve an unequal exchange, producing a net loss of DNA in one chromosome and a net gain in the other.

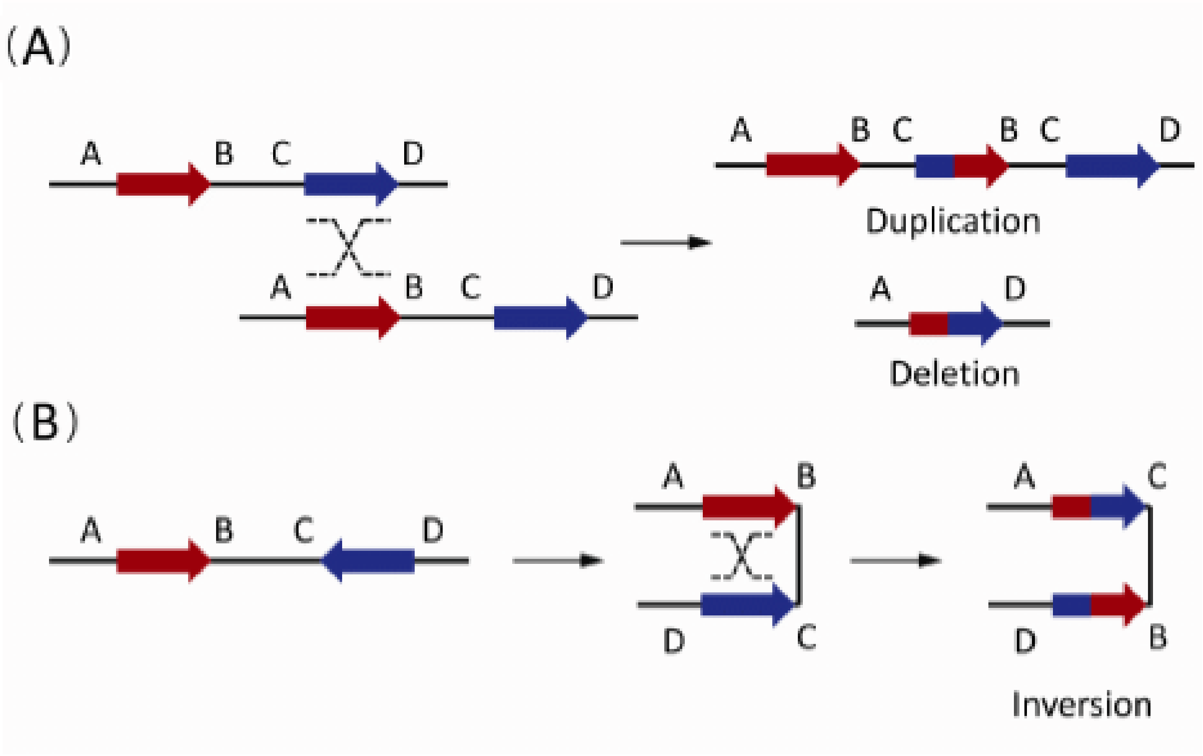


Figure 1.3. Non-allelic homologous recombination forming reciprocal a deletion and duplication (A), and an inversion (B). Adapted from (Chen et al., 2014)

2.1.2. Non-homologous end-joining

When double-strand DNA breaks (DSBs) occur during cell division, repair machinery typically uses the allelic sister chromatid sequence as a homologous template for repair. When a sister chromatid is not available, however, broken ends of DSBs are directly ligated through non-homologous end-joining (NHEJ) (Figure 1.4) (Chang et al., 2017). Nucleases and polymerases may resect damaged bases and add new bases before NHEJ occurs, resulting in small (<10bp) indels (Chang et al., 2017). More extensive damage involving several DSBs at adjacent loci can confound the repair machinery, leading to NHEJ of disparate strands, thereby forming deletions and intra-chromosomal translocation events (Chang et al., 2017). Moreover, an intervening sequence between DSBs may change its orientation prior to NHEJ, resulting in an inversion. In cases where multiple chromosomes are damaged simultaneously (as in the presence of ionizing radiation), NHEJ can produce inter-chromosomal translocations. As NHEJ does not depend on genomic features (such as LCRs), it typically produces non-recurrent SVs. However, NHEJ-based SVs are more likely to occur in regions prone to DNA damage, such as highly repetitive sequences and transcriptionally active sites (Barlow et al., 2013)

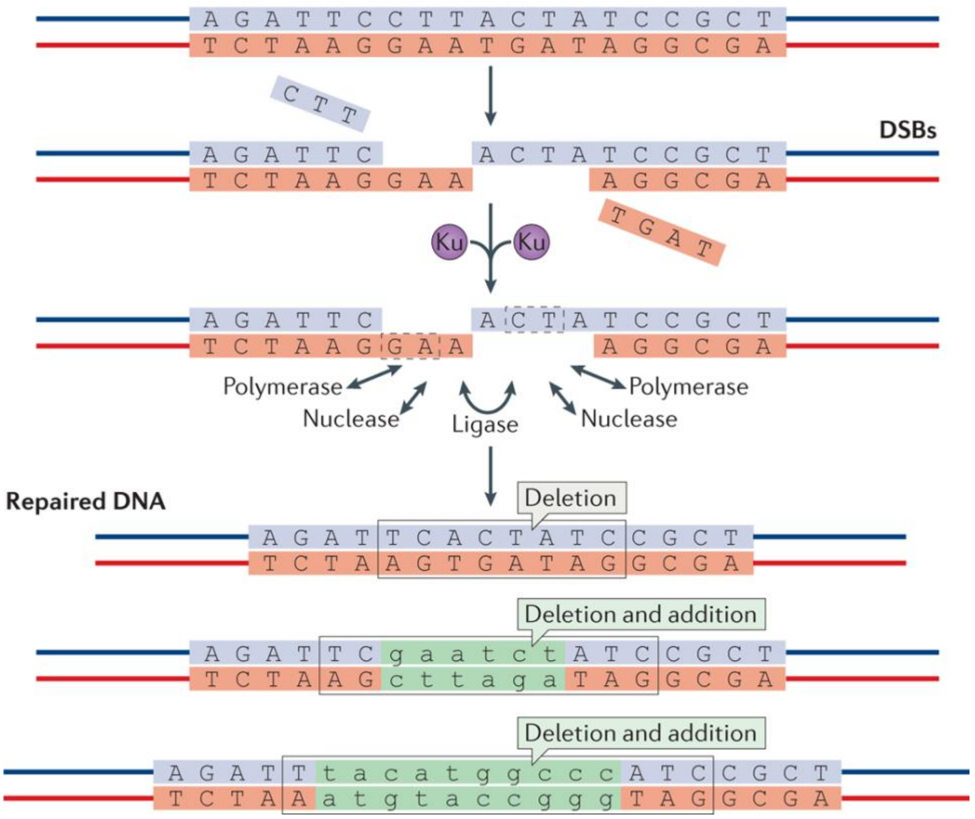


Figure 1.4. Non-homologous end joining, resulting in small deletions and insertions (additions). Adapted from (Chang et al., 2017)

2.1.3 Replication-based mechanisms

Lesions or obstacles encountered during DNA replication, such as DNA damage, secondary DNA structures and transcriptional machinery, can cause a replication fork to temporarily stall and disengage its lagging strand (Burssed et al., 2022). If an adjacent fork is replicating an homologous sequence, the free lagging strand may anneal to its template and restart synthesis at a new position (Burssed et al., 2022). Depending on whether the new template is upstream or downstream of the original, duplications and deletions will occur, respectively. If the new replication fork is moving in the opposite direction to the original, the replicated sequence will be inverted (Burssed et al., 2022) In rare instances, a lagging strand can switch to a template on a different chromosome, leading to a translocation event (Burssed et al., 2022). Fork stalling and template switching (FoSTes) is most likely to occur in sequences with high microhomology, defined as short, stretches of highly similar bases that can be contiguous or interspersed (Hastings et al., 2009). In such regions, a lagging strand can switch templates many times in succession (Figure 1.5, A), giving rise to complex SVs that combine several subtypes.

The breakage or collapse of a replication fork results in a single-ended DSB (seDSB), in which only the new synthesized double helix is broken, leaving the original intact. To repair the seDSB before DNA synthesis can continue, nucleases resect bases on the 5’ end, leaving an exposed 3’ overhang (Burssed et al., 2022). If this overhang contains a microhomology, it can invade a different region of the genome from the original strand, resulting in structural variation when DNA synthesis restarts (Figure 1.5, B). This mechanism is called microhomology-mediated break-induced replication (MMBIR). As in FoSTeS, if the position on the new strand is upstream or downstream of the original position, a duplication or deletion will occur, respectively. If the new strand is in the opposite orientation, there will be an inversion. If it is on a different chromosome, a translocation will be the result. MMBIR can also be repeated if additional fork breakages occur after DNA synthesis restarts, giving rise to complex events.

FoSTeS and MMBIR are more likely to occur at the beginning of the replication process, as polymerases are more likely to dissociate from the replication fork complex as it is still being formed (Hastings et al., 2009). As polymerases switch between different complexes at this stage, they tend to carry a small number of bases from their original template strand to their new template strand, which in turn produces small insertions at each resulting SV breakpoint (Hastings et al., 2009). It has thus been estimated that 35% of all SVs created by these mechanisms contain these short insertions (Hastings et al., 2009).

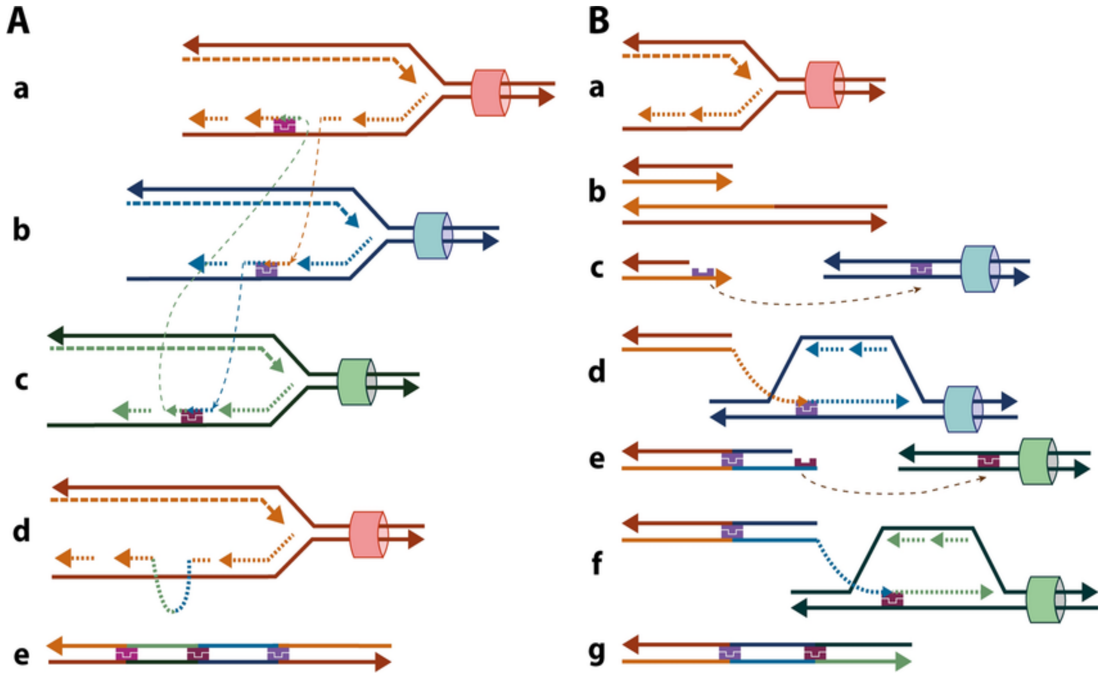


Figure 1.5. Replication-based mechanisms for complex SV formation: fork stalling and template switching (A) and microhomology-mediated break-induced replication (B). In A, a halt occurs at the replication fork (a), leading the lagging strand to detach from its initial template (b). Because of existing microhomology (shown in purple), the lagging strand shifts to a different template (indicated by the dashed line) at another active replication fork, and restarts the process of DNA synthesis (c). Ultimately, the strand goes back to its original template (d), and the newly formed DNA now includes adjacent sequences that were initially located in separate areas of the genome (e). In B, a replication fork collapses when it encounters a DNA lesion (a), resulting in a single-ended double-strand break (b). A resection of the 5′–3′ break creates a 3′ overhang with an exposed microhomology (c) (shown in purple), which acts as a template for a lagging strand from different region of the genome, where DNA synthesis is restarted (d). If the replication fork collapses again, the process can be repeated (e,f), resulting in a complex SV that again unites previously distant parts of the genome. Adapted from (Burssed et al., 2022).

In regions of extended contiguous microhomology (also known as ‘microsatellites’), the replication complex can slip, causing the template and newly synthesized DNA strands to slip out of alignment, resulting in the formation of a DNA ‘loop’ on the synthesized strand whose bases correspond to one or more copies of the repeated sequence (Mirkin, 2007). This loop may be resolved in different ways. There is a specialized type of repair machinery that can identify and excise them (Mirkin, 2007). In some cases, however, they are not recognized, and in future rounds of replication are integrated into the genome as the replication complex itself cannot recognise previous slippages (Figure 1.6). The resulting event is known as a small tandem repeat expansion and are usually defined relative to the number of repeats in a reference sequence. While these events are not typically defined as structural variants, they can be detected by one of the SV callers I used in my research.

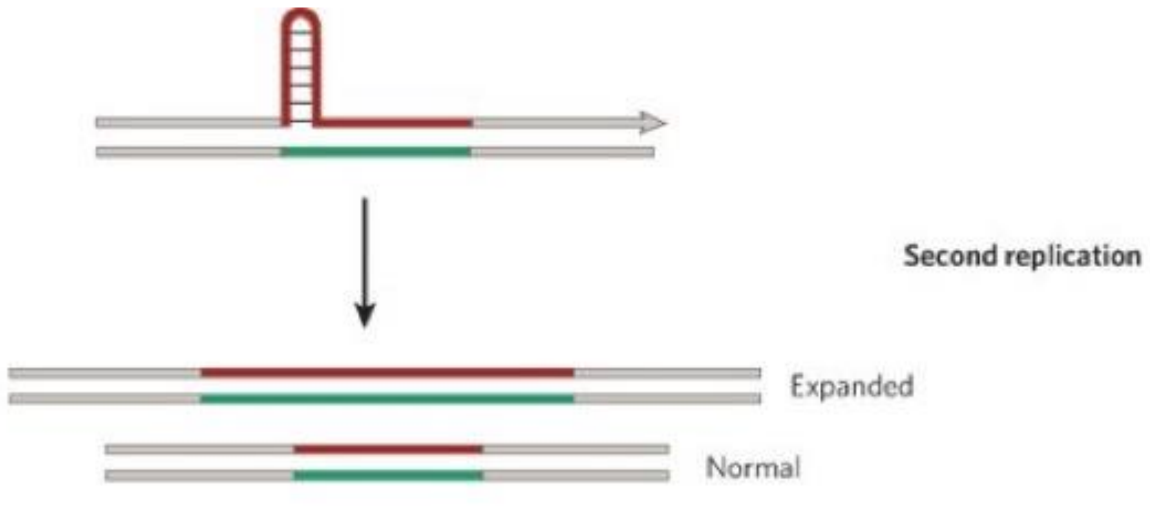


Figure 1.6. Formation of a small tandem repeat expansion. A DNA loop formed by slippage of the replication complex (top), is integrated into the genome in a second round of replication (bottom). Adapted from (Mirkin, 2007)

2.1.4 Retrotransposition

Retrotranspositions are generated by transcriptional retrotransposon machinery, typically acting on mobile elements (MEs), or retrotransposons, that are replicated throughout the genome (Lander et al., 2001). The machinery transcribes MEs into RNA, then reverse transcribes the RNA into cDNA, whereupon it is inserted at a different locus (Kazazian Jr., 2004) (Figure 1.7). Retrotransposons are thought to share an evolutionary origin with retroviruses but differ in that transcribed retrotransposons cannot leave their host cells (Lander et al., 2001). The most abundant type in the human genome is Alu, a ~300bp SINE mentioned in section 2.2.1 as a possible substrate for NAHR. It has been estimated that there are > 1 million Alu copies interspersed throughout the genome (Deininger, 2011). 6kb L1

LINEs are also common, with ~500,000 copies, though only a small fraction of these are still active and capable of retrotransposition (Brouha et al., 2003).

L1 retrotransposon machinery can also generate pseudogenes, defined as sequences that bear close resemblance to functional genes at different locations but cannot be transcribed due to modifications in their structure (Esnault et al., 2000). This occurs through the ‘hijacking’ of a typical transcription process, whereby the retrotransposon machinery binds and reverse transcribes processed mRNA (Figure 1.7) (Esnault et al., 2000). The number of exons in a pseudogene will vary according to how many had been transcribed when the disruption occurred, but also the number that are reverse transcribed by the L1 machinery (Wei et al., 2001). A pseudogene can also be inserted in reverse orientation to the original gene (Zhang et al., 2004). Pseudogene retrotranspositions are quite common; according to some estimates there are ~20,000 instances in the typical human genome (Zhang et al., 2004).

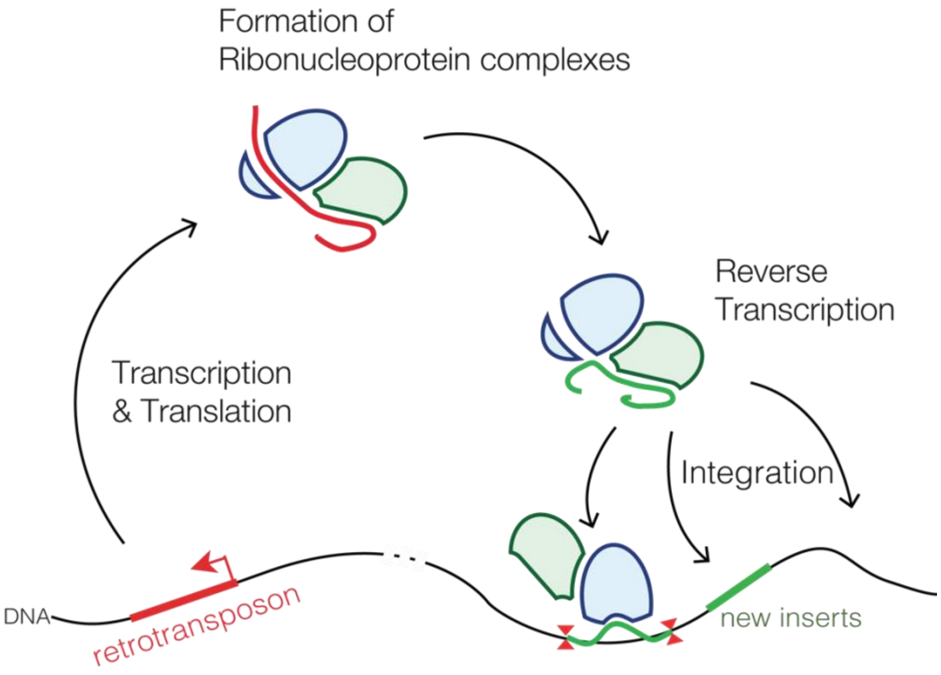


Figure 1.7. Mechanism of retrotransposition. Adapted from https://en.wikipedia.org/wiki/Retrotransposon

2.2 Structural variants in the human genome

SVs tend to be more deleterious than SNPs/indels and therefore occur less frequently in the general population (Collins et al., 2003). The precise number of SVs that can be identified in a human genome differs according to the technologies used for detection, but current estimates based on NGS data range from 4,000-30,000 per genome (Chaisson et al., 2019; Mills et al., 2011; Sudmant et al., 2015). This variability is primarily a function of sensitivity to smaller (<1kb) variants (Mills et al., 2011). Despite being less frequent than SNPs/indels, they affect significantly more individual bases and therefore account for a larger proportion of the genomic differences between individuals. Moreover, they cause a much broader range of functional consequences than the smaller variant classes. At larger size ranges, hundreds of genes and inter-genic functional sequences can be affected at once, resulting in the extensive disruption of thousands of interconnected biological processes. It is therefore challenging to determine the precise causal relations between SVs and clinical phenotypes.

However, there are studies that have used predictive tools to estimate the impacts of SVs. (Abel et al., 2020) used WGS data from 17,795 individuals to investigate the impact of SVs across the human genome, using tools that can predict the functional consequences of variants. They found that individuals carry, on average, 2.9 rare (MAF < 0.01) SVs that alter coding regions, affecting the dosage or structure of 4.2 genes and accounting for 4.0-11.2% of rare high-impact coding alleles. The majority of these were deletions (54.5%), with fewer duplications (42.2%) and a small number of inversions and complex events that disrupt exons. They also estimated that a typical genome carries 19.1 rare noncoding deletions that are as deleterious as PTVs, suggesting that further characterisation of non-coding regions will reveal additional variants of high-effect size in case-control analyses.

The first SV to be recognized as the cause of a disorder was chromosome 21 trisomy, which is so large it can be detected through standard light microscopy and gives rise to the set of physical and intellectual symptoms known as Down's syndrome (Asim et al., 2015). Large (>1mb) SVs that cause well-defined syndromes are known as genomic disorders and are typically associated with developmental abnormalities (Lupski, 1998). Other examples are Angelman syndrome, resulting from the deletion of 15q11-q13 (Kishino et al., 1997), and DiGeorge syndrome, which is associated with the deletion of 22q11.2 (McDonald-McGinn, 2015a). While genomic disorder SVs are usually highly recurrent, there can be significant symptom heterogeneity between individuals (Lupski, 1998). This is likely a function of the large number of genes impacted and the modifying effects of other genomic and environmental factors (Lupski, 2007). Given that a large proportion of genomic disorder SVs arise through *de novo* mutation (in the form of NAHR), prevalence is similar between ancestry groups except in cases where the typical substrate for mutation varies between ancestries (Lupski, 1998).

SVs have also been shown to have significant pleiotropy, conferring risk for several traits simultaneously. (Auwerx et al., 2022) tested the association of the copy numbers of CNVs called in 331,522 participants of the UK Biobank and 57 continuous clinically relevant traits and identified 131 hits across 47 traits. In addition to confirming previously known associations, such as the negative impact of 1q21.1–1q21.2 deletion on height, they found 26 traits that were associated with 16p11.2 BP4-BP5, and 16 traits with 22q11.21. Thirty-eight percent of the autosomal CNV associations considered also harboured a SNP signal for the same trait in previous studies, demonstrating that some of the same disease mechanism underlie different types of variants. In another striking finding, deletion and duplication of the same loci conferred opposite effects on many traits. For example, 16p13.11 duplication was associated with decreased age at menopause, whereas its deletion was associated with increased age. This phenomenon is known as ‘mirror phenotypes’. (Auwerx et al., 2022) also tested the association between burden of deletions and duplications, in terms of the number of affected megabases, and the same 57 traits. Thirty-five of them (61%) were significantly associated with at least one burden metric, including increased levels of adiposity, liver/kidney damage biomarkers, leukocytes, glycemic values, anxiety, decreased global physical capacity or intelligence.

2.3 Structural variants in schizophrenia

There have been several studies to date that have investigated the role of SVs in schizophrenia. While there is no evidence that inversions, translocations, and retrotranspositions are associated with the disorder, rare CNVs confer significant risk. Eleven rare, mostly recurrent CNVs have been consistently shown to be highly enriched in schizophrenia cases, compared to controls (Kirov et al., 2007; Levinson et al., 2011; Malhotra & Sebat, 2012; Rees et al., 2016; Walsh et al., 2008). These are shown in Table 1.1 and described below, along with their odds ratios and p values in a large meta-analysis conducted by (Rees et al., 2016), consisting of 6,882 schizophrenia cases and 6,316 controls, in addition to case-control data used in previous analyses.

Locus	**OR (95% CI)**	**P**
1q21.1 del	8.35 (4.65-14.99)	4.1×10^{-13}
1q21.1 dup	3.45 (1.92-6.20)	9.9×10^{-5}
2p16.3 *(NRXN1)* del	9.01 (4.44-18.29)	1.3×10^{-11}
3q29 del	57.65 (7.58-438.44)	1.5×10^{-9}
7q11. 23 WBS dup	11.35 (2.58-49.93)	6.9×10^{-5}
15q11.2 del	2.15 (1.71-2.68)	2.5×10^{-10}
15q11-13q AS/PWS dup	13.20 (3.72-46.77)	5.6×10^{-6}
15q13.3 del	7.52 (3.98-14.19)	4.0×10^{-10}
16p11.2 dup	11.52 (6.86-19.34)	2.9×10^{-24}
16p13.11 dup	2.30 (1.57-3.36)	5.7×10^{-5}
22q11.2 del	NA (28.27-∞)	4.4×10^{-40}

Table 1.1. 11 CNV loci that been consistently shown to be highly enriched in schizophrenia cases compared to controls. Odds ratios and p-values from (Rees et al., 2016). Del = deletion, dup = duplication.

1q21.1. This locus contains several blocks of LCRs that can be substrates for NAHR. Most commonly, this results in an ~800kb recurrent deletion/duplication at the distal end of the locus, but in rarer cases can form an ~2mb recurrent deletion that extends across the whole locus (Brunetti-Pierri et al., 2008). Genes within this locus, such as PRKAB2, CHD1L, and GJA8, have been implicated in neuronal development, synaptic plasticity, and neurotransmission (Harvard et al., 2011; Mefford et al., 2008), and these CNVs have been additionally shown to confer risk for intellectual disability and autism spectrum disorders (ASD).

2p16.3 (NRXN1) deletion. *NRXN1* deletions are non-recurrent and can vary in size, the most common encompassing exons in the NRXN1 gene, while others include additional adjacent genes (Kirov et al., 2009). It is the only schizophrenia associated CNV to affect just one gene, and therefore provides a unique opportunity for biological insight. *NRXN1* codes for Neurexin 1, a cell adhesion protein that mediates neurotransmission (Südhof, 2008), indicating that disrupted synaptic activity is important in schizophrenia aetiology.

3q29 deletion. This recurrent deletion typically spans approximately 1.6mb and encompasses around 20 genes (Willsey & State, 2015). It is very rare and has the largest effect size of schizophrenia associated CNVs. Several genes in this region, such as DLG1 and PAK2, play essential roles in synaptic function and neuronal development (Mulle et al., 2010). It has also been associated with a range of neurodevelopmental phenotypes (Glassford et al., 2016).

7q11.23 WBS duplication. The size of the WBS recurrent duplication CNV usually spans 1.5-1.8 megabases (Mb) and includes 26-28 protein-coding genes (Merla et al., 2010). Several genes in this region, such as GTF2I and GTF2IRD1, have been associated with neurodevelopment and cognitive functions (Crespi et al., 2010). It is named for Williams-Beuren Syndrome, a genomic disorder associated with the deletion of this locus (Stromme et al., 2002).

15q11.2 deletion. This recurrent deletion CNV spans approximately 500-700 kilobases (kb) and encompasses four genes: NIPA1, NIPA2, CYFIP1, and TUBGCP5 (Burnside et al., 2011). CYFIP1 has been linked to synaptic function and neuronal development (Oguro-Ando et al., 2015) It is the most common schizophrenia risk CNV, often exceeding 1% frequency in case samples, evidenced by its relatively low effect size.

15q11-13q AS/PWS duplication. This locus contains imprinted genes, which means their expression varies according to their parent of origin (Falls et al., 1999). Deletion of the region in the paternal chromosome causes Prader-Willi syndrome (PWS), while deletion in maternal chromosome causes Angelman syndrome (AS) (Buiting, 2010). Recurrent duplication in the maternal chromosome is associated with schizophrenia. The region spans approximately 4-6mb on chromosome 15 and contains several genes, including UBE3A, SNRPN, and a cluster of small nucleolar RNA (snoRNA) genes, that have been also implicated in neurodevelopmental disorders (Horsthemke & Wagstaff, 2008).

15q13.3 deletion. This recurrent deletion is approximately 1.5mb in size and is likely formed by NAHR due to the presence of LCRs at its breakpoints (Sharp, 2008). Several genes within this region play a role in neuronal function, including CHRNA7 and TRPM1 (Szafranski, 2010). It is also highly associated with epilepsy and occurs in ~1% of epilepsy cases (Helbig, 2009).

16p11.2 duplication. This is one of the most strongly associated recurrent CNVs with schizophrenia. The locus is ~700kb in length and contains 26 genes, at least 5 of which are involved in neuronal development, synaptic plasticity, and neurotransmission (Pucilowska, 2015). Its deletion is strongly associated with autism and developmental delay but is not enriched in schizophrenia cases (Malhotra & Sebat, 2012; Shinawi, 2010).

16p13.11 duplication. This locus is ~600kb in length and contains three intervals subdivided by LCR blocks with near-identical sequence homology (Ullmann et al., 2007). 12 recurrent CNV events have been observed in the region, depending on which LCRs are misaligned during NAHR. Deletions have been strongly associated with developmental delay, autism, and epilepsy (Ramalingam, 2011). An excess of deletions has been reported in schizophrenia cases, but evidence of association is modest (Ingason et al., 2011). While duplications were initially thought to be benign, there is now evidence for their roles in autism, epilepsy, and ADHD, in addition to schizophrenia (Williams et al., 2010). While the functional impact of the duplication is largely unknown, the region contains *NDE1*, which is known to be involved in cortical development (Alkuraya, 2011).

22q11.2 deletion. This deletion is the most strongly associated schizophrenia genetic risk factor. The affected locus is ~3mb in length, encompassing around 40 genes (Shaikh et al., 2000). It contains 4 LCR regions and is thus prone to a high number of different CNV events (Burnside, 2011). Deletion of the entire locus causes DiGeorge Syndrome, a genomic disorder with a broad array of symptoms, including intellectual disabilities, cardiac defects, immune system dysfunctions, facial deformities, and psychiatric issues (McDonald-McGinn, 2015). Around 30% of DiGeorge syndrome cases develop schizophrenia or another psychotic disorder (Murphy, 1999). Deletion of the 1.5mb proximal region can also cause DiGeorge syndrome, but symptoms tend to be less severe (Burnside, 2011). Given the high gene content of the locus, the functional impact of its deletion with regards to schizophrenia is difficult to determine. A likely candidate gene is COMT, which is involved in dopamine metabolism (Egan, 2001). However, *COMT* was not implicated by common variants in the PGC3 schizophrenia GWAS (Trubetskoy et al., 2022). 22q11.2 duplication has been found to be protective against schizophrenia (Rees et al., 2014).

In addition to individual risk variants, an increased burden of rare CNVs genome-wide have also been reported in cases compared to controls (International Schizophrenia Consortium, 2008; Rees & Kirov, 2021; Stefansson et al., 2008), and CNVs that have the largest impacts are enriched for LoFi genes (Marshall et al., 2017) and neurodevelopmental disorder (NDD) risk genes (Kirov et al., 2014). *De novo* SVs have also found to be enriched in schizophrenia cases, which I discuss in the results chapters of projects I undertook to identify such SVs in a WES trios data set (chapters 3 and 4).

2.4 Platforms used for structural variant detection

2.4.1 Genotyping microarrays

Most studies examining CNVs in schizophrenia have utilised genotyping microarrays (Wang & Bucan, 2008), whose primary function is to detect the relative quantities of SNPs in a DNA sample. This method involves the washing of fragmented cDNA from a single genome across millions of probes, each containing oligonucleotides that anneal to fragments exhibiting a specific SNP variant, labelled with a fluorescent marker. The density of hybridised probes at a given locus linearly correlates with the intensity of fluorescence observed, quantified as the normalised log R ratio (LRR) for that SNP. Copy number variation is indicated by deviation from the expected LRR for contiguous probes (Wang & Bucan, 2008). A heterozygous deletion, for example, is evidenced by LRRs that are 50% less than would be expected for diploid copy number, while a heterozygous duplication is associated with a 33% increase in signal intensity (Figure 1.8, A).

If the arrays used can detect different alleles at the same locus, a second metric that can be implemented to detect CNVs is the relative signal intensity of two alleles (A and B), known as B allele frequency (BAF). For diploid copy numbers, BAF at a given locus is expected be 0, 0.5 or 1 for AA, AB, and BB, respectively. For heterozygous deletions, however, BAF will be only 0 or 1, depending on whether the deleted sequence contains a B allele at that locus. Loci in the presence of a heterozygous duplication will have a BAF of ~0.33 if there are two copies of the A allele, and ~0.67 if there are two copies of the B allele (Figure 1.8, B).

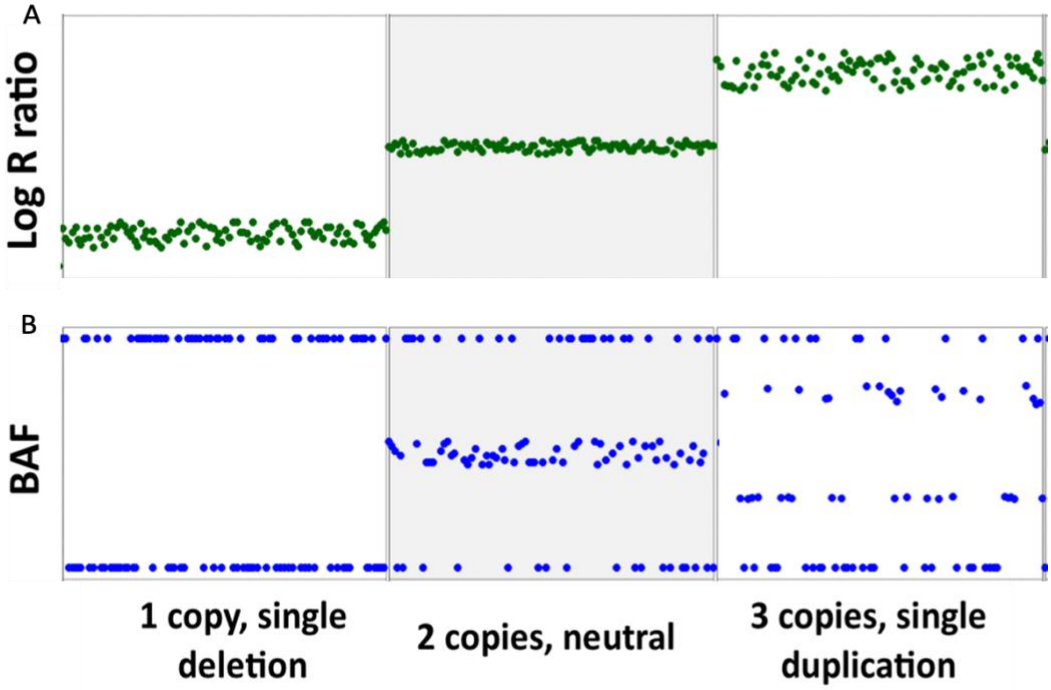


Figure 1.8. Log R Ratios (A) and B allele frequencies (B) for contiguous microarray probes. Heterozygous deletion (left) and heterozygous duplication (right) are contrasted with diploid copy number (centre). Adapted from Nadolo et al. (2008)

2.4.2 Whole exome sequencing

Whole exome sequencing (WES) methods have also been used to successfully detect CNVs, in addition to other structural variant types, in protein-coding regions (Seaby et al., 2016) Similar to microarray methods, genomic cDNA is first isolated and fragmented. Then, exonal fragments are ligated with adaptors and enriched to produce an exome 'insert' library. These inserts are hybridised to baits on a flow cell within the sequencer and fluorescently labelled single nucleotides are introduced along with polymerases. As the polymerases cause labelled nucleotides to successively bind to each fragment, they emit one of four specific fluorescent wavelengths. Analysing the order at which these wavelengths occur, the sequencer thereby determines the sequence of the original fragments (Alekseyev et al., 2018).

Different sequencing platforms are capable of processing inserts of different lengths and are generally divided into long-read (~10-100kb) or short-read (~100-200bp) subtypes (Adewale, 2020). Moreover, short-read platforms can process reads in one direction or in both directions simultaneously. The latter is known as 'paired-end sequencing' and produces two reads per insert that are referred to as read mates (Figure 1.9). While paired-end sequencing typically has a longer run time, it offers improved read assembly and alignment (Seaby et al., 2016). Moreover, as I describe in chapter 2, it can improve detection of balanced SV types that can impact the orientation and distance between read pairs. The average number of times a targeted base in a DNA sample can be sequenced by a platform is known as the platform's coverage depth. Coverage depth can vary greatly between platforms, and generally platforms with a large coverage depth can produce more high-confidence variant calls. I describe how different elements of the data produced by a short-read sequencing platform can be used to detect SVs in chapter 2.

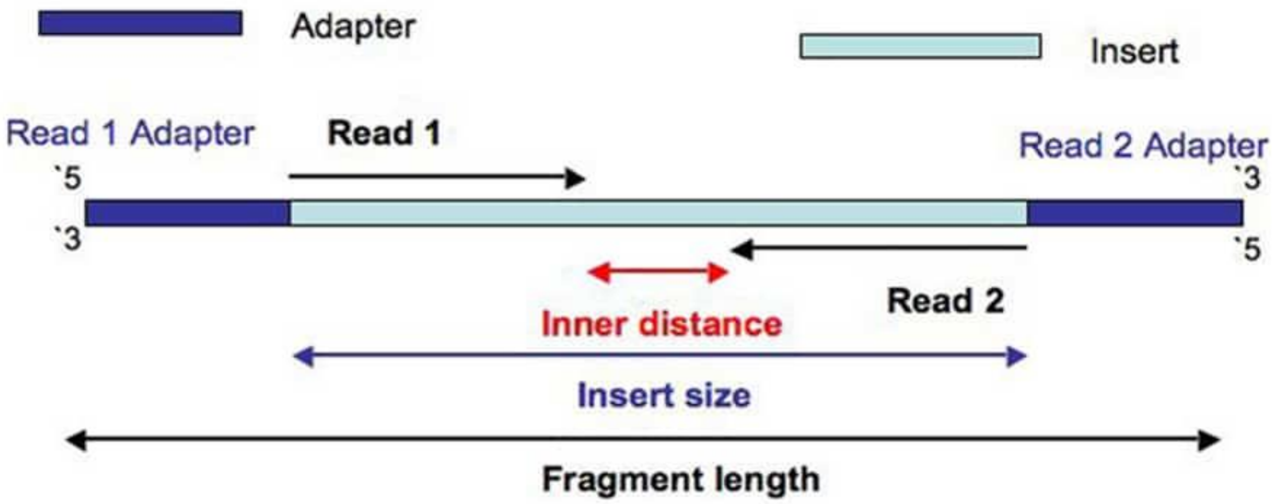


Figure 1.9. Paired-end sequencing fragment. The insert is sequenced in both the forward (Read 1) and backward (Read 2) directions.

2.5 Summary

In this section I have described the different types of SV and their mechanism of formation. I have also given an overview of SVs in the human genome, and types of clinical phenotypes they are associated with. I have described up-to-date knowledge about the role of SVs in schizophrenia, which is largely limited to large, rare recurrent CNVs. Finally, I have described the two platforms that have been most used to detect SVs in previous studies: genotyping microarrays and whole exome sequencing. This chapter will serve as relevant background for understanding the SVs I investigated in my own research. In chapter 2 I describe the processes by which the calling algorithms I used detect SVs in WES data.

3. Aims and objectives

The overarching aim of this PhD was to assess the utility of WES to detect SVs in schizophrenia, and to determine the potential impacts of any SVs identified in disorder aetiology. To achieve this, I called SVs in two data sets using two calling algorithms: CLAMMS (Packer et al., 2016) and InDelible (Gardner et al., 2021). Each caller leverages different aspects of WES data, and therefore produce largely non-overlapping call sets. I describe each algorithm in detail in chapter 2. One of the data sets I analysed consisted of schizophrenia probands and their parents, while the other consisted only of cases who had been extensively tested for cognitive performance. My primary objectives for research were:

1) Assess the overlap and differences between call sets generated by the two calling algorithms, and CNV call sets previously generated from the same data using genotyping microarrays. I particularly focused on the size distributions of each call set, and whether they contained any variants that were known to confer schizophrenia risk. Based on these comparisons, I could conclude whether combining these approaches can produce a more accurate and comprehensive SV call set for analysis in schizophrenia, or whether only one approach is sufficient for future studies.

2) Identify *de novo* SVs in the schizophrenia trios data and use findings from previous rare and common variant studies to determine whether any putative candidate schizophrenia risk genes are impacted by the SVs.

3) Test SVs for association with cognitive deficits in schizophrenia, with particular focus on the role of SVs at the smaller end of the size spectrum. While SVs have been shown to be associated with cognition in schizophrenia (as I describe in the relevant results chapters), no studies have used SVs generated from WES data, and only large (>100kb) variants have been tested. As such there are large knowledge gaps regarding the contribution of smaller SVs (e.g. < 100KB) to cognitive impairments in schizophrenia.

Chapter 2: Structural Variant Callers

In this chapter I give a detailed description of the two structural variant callers I used in the four studies carried out during my PhD: CLAMMS and InDelible. I also discuss the original studies describing these methods as well as benchmarking studies which evaluated the accuracy of these methods to detect known SVs. First, however, I will describe how raw sequencing data is processed and used to call SVs by CLAMMS and InDelible.

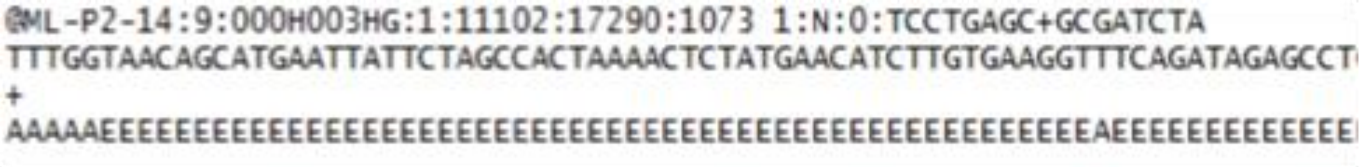
1. Preparation of raw sequencing data

1.1 FASTQ files

In section 2.4.2 of chapter 1 I described the process by which short reads are generated from exome sequencing on Illumina paired-end short-read sequencing platforms. The data that is output by the sequencing machine is stored as a FASTQ file, which contains sequence and quality data for all reads. In a paired-end run, two files are generated, for one for R1 mates and another for R2. Each entry in a FASTQ describes an individual read and typically contains four lines: 1) a unique identifier, including run information and the flow cell lane/tile on which the read was sequenced, and whether the read is a first or second mate; 2) The base pair sequence of the read (a string of A, C, T, G and N); 3) a separator line, consisting of a single "+", which is included for ease of parsing; and 4) A quality score string the same length as the sequence strings, specifying the quality of each base.

The base quality scores in the final line represent the probability that each base has been correctly sequenced, based primarily of the intensity of the fluorescent signals and signal-to-noise ratios. The raw probabilities are transformed to the Phred-scale, defined as -10 log10(P), where P signifies the initial probability. This scale has two advantages: first, it is logarithmic, and so can represent a wider range of values in a smaller amount of data than the raw probability scores. Second, it allows for standardisation of quality scores across a range of sequencing platforms. The maximum possible Phred-score is 40, which corresponds to base call accuracy of 99.99% and above, while a score of 30 corresponds to base call accuracy of 99.9%. For a FASTQ quality score string, Phred-scores for each base are converted to an

ASCII codes between 33 and 73 (https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm). An example of a FASTQ entry, from (https://knowledge.illumina.com/software/general/software-general-reference_material-list/000002211), is shown by Figure 2.1



```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEE
```

Figure 2.1. An example FASTQ file entry

1.2 Read alignment

Before read data can be mined for the presence of SVs, the reads need to be aligned to a reference genome, such as Genome Reference Consortium Human Build 37 (GRCh37). There are several alignment tools that can take FASTQ sequence strings as their input, and which differ according to whether they align DNA or RNA, and the size of sequences they are configured to align. As all the WES data I used in my PhD was aligned using the Burrow-Wheeler Aligner (BWA) (Li & Durbin, 2009), this is the only approach I will provide details for. The BWA is named after the Burrow-Wheeler transform, a method of compressing reference sequences to allow for more computationally efficient mining of FASTQ input sequences in a reference genome. A reference genome input file is provided in FASTA format, which consists of sequences ranging from hundreds to thousands bases in length, paired with unique identifiers specifying their chromosomes and relative positions.

The most recent version of BWA has three algorithms:

1) BWA-backtrack, designed to align short reads generated by Illumina platforms ~100bp in length. It consists of two steps: seed generation and seed extension. In the first step, sequences to align are divided into substrings called seeds, typically 32bp in length. The reference genome is mined for exact matches for each seed. In the second step, seeds that have been exactly matched are then extended to include the rest of the initial sequences,

scoring each alignment based on number of mismatches. It can take account of insertions and deletions (indels) that compromise the unique alignment of a read sequence.

2) BWA-SW is designed to align longer reads than BWA-backtrack. It has an equivalent seed generation step, but then implements the Smith-Waterman algorithm for alignment rather than seed extension, which is more sensitive to mismatches/indels but results in a slower run-time than BWA-backtrack.
3) BWA-mem is a hybrid of the first two algorithms and runs faster than both. It was designed to replace BWA-backtrack and BWA-mem in most scenarios.

In all three algorithms, a Phred-scaled mapping quality (MAPQ) score is calculated for each sequence alignment based on its quality (mismatches and indel content) and uniqueness, and from which the best alignment for a given read sequence is derived. MAPQ scores range from 0-60, with 60 corresponding to a less than 10^{-6} probability that the read is incorrectly mapped. In cases where the alignment process has failed or was not applied, MAPQ is set to 255.

BWA-mem was used to align the WES read data used in the studies I undertook, and was conducted by Elliott Rees. Once all reads (and their mates, if applicable) in a FASTQ file have been aligned to a reference genome, BWA outputs a mapping/alignment file in SAM, BAM, or CRAM format, depending on the compression algorithm used. Mapping/alignment files can be paired with an index file which allows for computationally efficient retrieval of reads at reference loci of interest and are a required input for all SV calling methods. Index files are not output by BWA but can be generated by the Samtools 'Index' function (Li et al., 2009). All data used in my research were in BAM format, whose index files were also generated by Elliott Rees and have the suffix BAI.

1.3 BAM files

Each entry in a BAM file corresponds to an aligned read, and has 11 mandatory fields of information, described in Table 2.1. The number of entries for a given read are based on the number of alignments generated for that read. A range of optional fields can be included based on BWA configuration, but as these are not relevant for the two SV callers, I will not describe them.

Field name	**Description**	**Example**
QNAME	Unique identifier for read	K00267:46:HFYN2BBXX:5:2205:31101:21201
FLAG	An integer between 0 to 65535, calculated based on combinations 12 of read and quality metrics. Described in more detail in 1.3.1	145
RNAME	Chromosome read is aligned to in the reference genome	1
POS	Start position of the read sequence in the reference genome	2237429
MAPQ	Phred-scaled score indicating the likelihood that the read is correctly aligned	60
CIGAR	Acronym for ‘Compact Idiosyncratic Gapped Alignment Report.’ Specifies the number of matches, mismatches and indels in a read. Described in more detail in 1.3.2	75M
MRNM	The reference chromosome to which a read’s mate is aligned. Set as ‘=’ if both reads align to the same chromosome, or ‘*’ no mate information is available	=
MPOS	Start position of the read mate sequence in the reference genome	2237421

ISIZE	Base pair distance between reference position and mate position. Value is negative if read is the second mate (assuming typical read orientation).	-83
SEQ	Base sequence of the read, in term of ‘A’, ‘T’, ‘C’, ‘G’ and ‘N’	CAGTGACCCCGAG […]
QUAL	Base quality scores, specified the FASTQ ASCII string	@BBCD?CDDAFEE […]

Table 2.1. The 11 mandatory fields for each entry in a BAM file. All examples are for a single entry in from a BAM file analysed as part of my research.

1.3.1 BAM FLAG field

The FLAG field is generated by summing the decimal equivalents of hexadecimal values corresponding to 12 read mate and quality properties specified in Table 2.2

Property	**Hexadecimal notation (decimal equivalent)**
Read paired	0x1 (1)
Read mapped in proper pair (i.e. both mares are mapped)	0x2 (2)
Read unmapped	0x4 (4)
Mate unmapped	0x8 (8)
Read reverse strand	0x10 (16)
Mate reverse strand	0x20 (32)
First in pair	0x40 (64)
Second in pair	0x80 (128)
Not primary aligned (i.e not the best quality alignment for the read)	0x100 (256)

Read fails platform/vendor quality control	0x200 (512)
Read is PCR duplicate	0x400 (1024)
Supplementary alignment (sequences within the read align to different reference loci)	0x800 (2048)

Table 2.2. The properties used to calculate the BAM FLAG field, with their hexadecimal notations and decimal equivalents.

In the example entry given in Table 2.2., the alignment has four of these properties: Read paired (0x1), Read is mapped in a proper pair (0x2), Read is on the reverse strand (0x10) and Read is the second read in a pair (0x80). Thus, 0x1 + 0x2 + 0x10 + 0x80 = 1 + 2 + 16 + 128, which gives the FLAG value 145.

1.3.2 CIGAR string

The Compact Idiosyncratic Gapped Alignment Report (CIGAR) specifies the alignment properties of each base in an aligned sequence, including discrepancies with the reference genome that indicate the presence of indels. Each property is signified by a single letter, described in Table 2.3. I have only included those properties that are output from the alignment of DNA reads to a standard reference genome.

Property signifier	Description
M	Matches reference base at same position
I	Insertion, relative to reference sequence
D	Deletion, relative to reference sequence
S	Bases that mismatch with the reference genome but are still included in the read sequence. Known as ‘soft-clipping’
H	Bases that mismatch with the reference genome but are excluded from the read sequence. Aligners will only exclude mismatching bases if they fall below a quality threshold, or are adaptor sequences.

Table 2.3. Signifiers for base alignment properties included in a CIGAR string.

The CIGAR string in the example entry in Table 2.3 is ‘75M’, signifying that the read sequence has 75 bases, all of which match with the reference genome at the aligned position. An example of a more complex CIGAR string is ‘15S5M2D66M’. In this case, the read sequence consists of 87 bases, the first 15 of which are soft-clipped. The next 5 match, followed by an absence of 2 bases relative to reference, and finally 66 bases which also match.

1.3.3 Whole-exome sequencing BAMs

The size a WES BAM file based on an exome varies greatly with sequencing depth but is generally between 5-15GB and contains >10 million entries. Both the SV callers take WES BAM files (and their index files) as inputs, but mine very different aspects of their data, as I describe in the following sections.

2. CLAMMS

2.1 Introduction

The first SV caller I describe is CLAMMS (Copy number estimation using Lattice-Aligned Mixture Models), developed by Regeneron and published in 2016 (Packer et al. 2016, code repository: https://github.com/rgcgithub/clamms). CLAMMS detects CNVs in WES by leveraging coverage depth, based on the observation that coverage is correlated linearly with copy number state. In general, if reference sequence S is affected by a heterozygous deletion in individual X, there will be a ~50% decrease in coverage for all bases within S in X’s WES data, relative to data for individuals with diploid copy number at S. This is because ~50% fewer reads that align to S from X’s DNA will be available for sequencing. Conversely, if S is affected by a heterozygous duplication in individual Y, there will be ~33% increase in coverage for all bases within S in Y’s WES data. Most of CLAMMS’ calling process is dedicated to reducing the impact of factors that confound this correlation between coverage and copy number state, so that copy number states can then be modelled across samples in a WES data set. CLAMMS can call CNVs across the whole allele-frequency spectrum and is designed for use in large-scale analyses.

2.2 CLAMMS algorithm

The algorithm has four main stages: 1) generate windows to be targeted for CNV calling; 2) compute coverage depth for all call windows in each BAM file; 3) model coverage copy number states across all BAM files; and 4) use models to call CNVs.

2.2.1 Generating call windows

The first step of the algorithm generates a file containing the exome sequence windows which will be targeted for CNV calling. Four input files are required: 1) a list of the exome regions captured by the sequencing platform used to generate to the WES data to be analysed; 2) the FASTA file for the reference genome to which reads have been aligned; 3) a list of mappability scores for bases genome-wide; and 4) a list of 1333 special regions, included in the CLAMMS code repository.

File 1) contains three fields, specifying the chromosome, start base and end base of each exome capture region. The base mappability scores contained in file 3) are defined as one divided by the number of places across the genome that a k-mer starting at that base aligns to, with up to two mismatches allowed. Scores range from 0 to 1, with 1 indicating that the k-mer starting at the base does not map to any other position (i.e.. a unique alignment). A reference sequence containing a high proportion of bases with low mappabiltiy is prone to read malignment, such that coverage depth of bases across the sequence is likely to be inaccurate. Mappability scores for 75-mers and 100-mers can be downloaded from the UCSC Institute (https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability). The special regions listed in file 4) are either regions of known extreme sequence polymorphism, such as *HLA* and *KIR* gene clusters, or common duplications. The former are prone to read mismapping, while the latter are prone to copy number states > 3 (homozygous duplication).

CLAMMS also requires a user-defined insert size variable when generating the windows file, for purposes of calculating window GC content. Defined as the number of G-C base pairs in a sequence, GC content has a large impact on base coverage depth. DNA fragments with a high GC content are more difficult to denature during PCR and are therefore likely to be under-represented among sequenced reads (Benjamini & Speed, 2012). The downstream consequence is that bases within such fragments will have a coverage depth that is not representative of their copy number state in the source DNA. (Benjamini & Speed, 2012) found that GC content coverage bias is best estimated by the content of a full insert, not individual reads. The CLAMMS developers therefore recommend an insert size that is ‘a little bit bigger’ than the mean insert size for the sequencing process used, so that most reads will come from inserts of sizes smaller than this value.

The code used to generate the window file first divides exome capture regions that are ≥ 1kb long into equally sized windows 500-1kb long. This ensures that the breakpoints of CNVs that only partially overlap large exons can be more precisely determined. Exome capture regions < 1kb are treated as individual calling windows by default. Windows are then annotated with 5 metrics: 1) the bp interval of the exome capture region from which the window was derived; 2) number of GC-base pairs, extended to fit insert size; 3) GC content as a proportion of the insert size; 4) the mean mappability score for all bases in the window; and 5) ‘-6’ or ‘1’ depending on whether the widow intersects (by at least 1bp) a region of extreme polymorphism or common duplication, respectively.

Windows are filtered if they have a GC content proportion that lies outside a configurable range, which by default is [0.3,0.7]. Investigating the relationship between coverage and GC content in the original CLAMMS study (Packer et al., 2016), the developers found that coverage depth variance for windows that lie outside this range is too large to be adequately modelled (Figure 2.2). Windows are also filtered if their mean mappability score is < 0.75, or if they intersect a region of extreme sequence polymorphism. In total, ~12% of exome capture regions are excluded from the windows file.

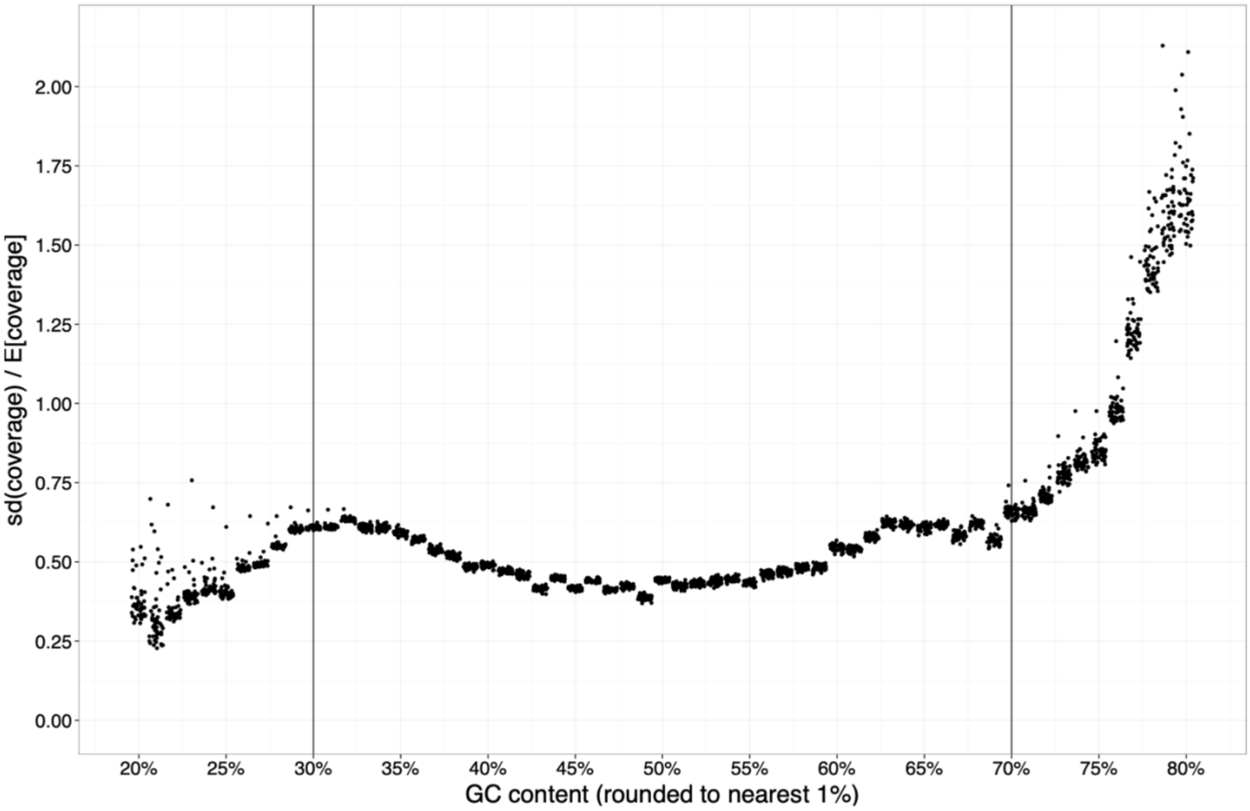


Figure 2.2. Coefficient of variation (standard deviation/mean) of coverage (y axis), conditional on GC content (x axis), for 50 samples from the CLAMMS original study (jittered for visibility). Vertical lines show the upper and lower thresholds for GC content recommended by the developers. Above GC = 70%, coverage variance dramatically increases relative to sample mean. Below GC = 30%, variance of coverage itself is volatile. Adapted from (Packer et al., 2016) supplement.

2.2.2 Computing depths of coverage

In the second stage of the algorithm, depth of coverage is calculated for each call window in each BAM file to be analysed. This is achieved using the ‘bedcov’ command of Samtools, for which 2 input files are required: 1) the calls windows file; and 2) the indexed BAM file of each sample to be analysed. An additional parameter is the minimum MAPQ for a read to be counted, with 30 being the default suggested by the CLAMMS authors. The probability that a read with MAPQ > 30 has been mapped incorrectly is < 1% (Li & Durbin, 2009). As it is implemented in CLAMMS, Samtools determines the number of reads passing the MAPQ filter that are aligned to all bases in the windows file, then calculates the mean base coverage for each window. The output contains 4 fields, the first 3 of which specify the chromosome, start base and end base of call windows, while field 4 gives the mean coverage value.

Depth of coverage for each sample is then internally normalised to control for GC content bias and overall average depth of coverage. Two input files are required for this step: 1) the call windows file; and 2) the coverage file for each BAM, to which the following formula is applied:

$$\mathrm{Cov}_{\mathrm{norm}}(w) = \mathrm{Cov}(w) \,/\, \mathrm{median}(\mathrm{Cov} \mid \mathrm{GC}(w))$$

where w is a call window, Cov is coverage for a given BAM and median(Cov | GC(w)) is the median coverage for that sample conditional on the GC content of w. The conditional median is calculated by binning all call windows for a BAM with similar GC content, computing the median coverage for each bin. A normalisation factor for each window is then calculated using a linear interpolation between the median coverage of the two bins nearest to it. The number of windows per bin is configurable, but the developers give a default size ‘that balances fine-grained binnings with the need to provide each bin with a sufficient sample size for estimation.’

2.2.3 Modelling copy number states

Aside from GC content and mappability, there are other factors whose confounding impact on coverage depth is more stochastic in nature. These include often-subtle differences in sample preparation and input DNA quality and are collectively referred to as ‘batch effects’ in the CLAMMS methods paper. Adequately mitigating the impact of batch effects is the first step of the CLAMMS modelling stage and is achieved by grouping samples according to similarities in quality metrics. The developers recommend using the following seven metrics, that can be generated using Picard (https://broadinstitute.github.io/picard/): GC_DROP_OUT, AT_DROP_OUT, MEAN_INSERT_SIZE, ON_BAIT_VS_SELECTED, PCT_PF_UQ_READS, PCT_TARGET_BASES_10X, and PCT_TARGET_BASES_50X. A description of each is given in Table 2.4, along with the Picard function used to generate it from BAM files.

Picard command	Picard metric	Description
CollectGcBiasSummaryMetrics	GC_DROP_OUT	Produced by calculating (%GC in ref - %GC in reads) for 50bp exonic windows and summing all positive values for GC = [50..100].
	AT_DROP_OUT	Produced by calculating (%GC in ref - %GC in reads) for 50bp exonic windows and summing all positive values for GC=[0..50].
CollectInsertSizeMetrics	MEAN_INSERT_SIZE	The mean insert size, excluding artefactual outliers.
CollectHsMetrics	ON_BAIT_VS_SELECTED	Fraction of bases on or near baits (i.e., targeted sequences) that are covered by baits.
	PCT_PF_UQ_READS	Fraction of aligned unique reads that pass Illumina's internal quality filter (PF) from total number of PF reads
	PCT_TARGET_BASES_10X	The fraction of all target bases achieving 10X or greater coverage.
	PCT_TARGET_BASES_50X	The fraction of all target bases achieving 50X or greater coverage.

Table 2.4. Picard metrics used to group samples for coverage depth models, including commands used to generate each and a brief description.

The CLAMMS authors recommend grouping samples according to the k-nearest neighbours' algorithm. This can be achieved in several ways, but for the studies I conducted, the seven Picard metrics for each sample were imported into an R environment and normalised using min-max scaling. The package 'FNN' (Fast Nearest Neighbour (https://cran.r-project.org/web/packages/FNN/FNN.pdf)) was used to construct a k-d trees from the metrics, from which k-nearest neighbour groupings could be extracted. It cannot be known *a priori* which value of k best controls for batch effects in a given dataset. As a rule, however, larger values of k decrease variance of model parameters, but increase the potential bias of batch effects. The methods by which I optimised k were particular to each study and are therefore described in the methods sections of relevant results chapters. For each sample, a file was created containing its k-nearest neighbours. For the purposes of modelling, these files are referred to as 'reference panels'.

CLAMMS then uses sample reference panels to train exome-wide depth of coverage models for each call window. Two input files are required: 1) a reference panel file annotated with the file path to the normalised coverage files for each sample; and 2) the call window file. Each model has at least 4 sub-components (hence 'mixture'), corresponding to copy number states 0-3, and two free parameters: the mean (μDIP) and standard deviation (σDIP) of the coverage depth corresponding to diploid copy number. For a non-diploid copy number k, the mean is constrained to equal (k/2) x μDIP. Thus, if μDIP = 6, mean coverage for copy number 1 (haploid) is ½ x 6 = 3. Through an examination of male vs female samples on X chromosome, the developers found that haploid samples had approximately half the variance of diploid samples, so set the standard deviation for haploid samples to equal √0.5 x σDIP. Variance for copy numbers > 2 should be greater than for diploid samples, but the developers found that including this in the model increased the rate of false-positive duplications. The standard deviation for all events of copy number > 2 were therefore set to equal σDIP.

While copy number 0 would result in no coverage in the absence of confounders, read mismapping can lead to a small amount of coverage even in the presence of such events. The authors therefore found coverage copy number 0 to fit an exponential distribution, with a mean (1/λ) initially equal to 6.25% of μDIP. In regions of no read mismapping, iterations of the model fitting algorithm will drop the mean to 0, causing numerical instability. To prevent this, if the mean drops below 0.1% of μDIP, the exponential distribution is replaced with a point mass at 0. CLAMMS considers copy numbers 4-6 only for windows intersecting regions where duplication is known to commonly occur (annotated ‘6’ in windows file generation).

Each model is fit to the normalised coverage data for a reference panel by means of the expectation-maximisation (EM) algorithm. μDIP is initialized as the median coverage across all samples in the reference panel, for the call window to be modelled, while σDIP is initialised to the median absolute deviation of coverage (MAD) around the call window median, scaled by a constant factor. Figure 2.3, taken from the CLAMMS methods paper supplement, shows mixture models fit to real normalised coverage distributions for exons of gene *GSTT*1.

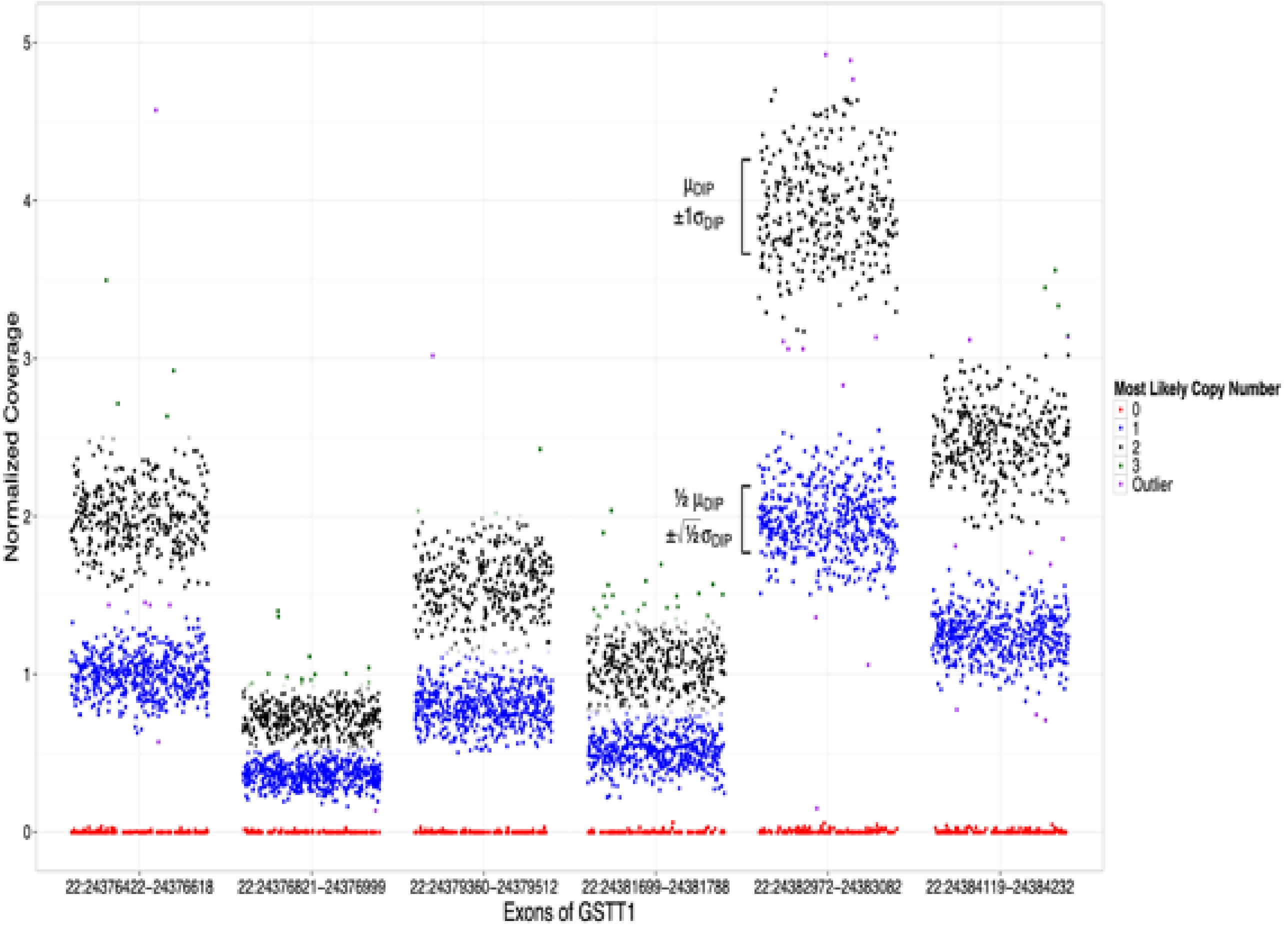


Figure 2.3. Mixture models fit for mean observed normalised coverage distributions of exons of *GSTT1*, annotated with mean and variance parameters for diploid and haploid copy numbers at exon 4. Each point, jittered for visibility, represents an

individual sample from the DiscovEHR Study. Adapted from (Packer et al., 2016) supplement.

A model file is the output containing 17 fields. The first three give the chromosome, start and end base of the calling windows. Field 4 gives the maximum copy number considered (6 if the window intersects a region of common duplication, 3 otherwise). Field 5 and 6 specify GC content proportion and mappability scores, while fields 7-10 are model parameters: 7 gives the flag to introduce the point mass in windows with 0 coverage (set to 1 if not applicable); 8 gives the λ value for the exponential distribution modelling non-zero coverage in windows where copy number is 0; 9 is μDIP and 10 is σDIP. Finally, Fields 11-17 gives the number of samples in the reference panel that have estimated copy numbers 0-6, respectively.

2.2.4 CNV calling

For each sample, the CNV calling stage requires two input files: 1) normalised coverage file; and 2) model file. An optional flag can be added to the call command with the sex of each sample as its argument (M or F), to call CNVs from sex chromosomes. CLAMMS calls CNVs using a Hidden Markov Model (HMM) (Eddy, 2004), whose input is the sample's normalised coverage values at each call window. The states are DEL (deletion), DIP (diploid) and DUP (duplication). Thus, the probability of observing a normalised coverage value x, at a calling window w, given state s, is determined by the mixture model trained at w that correspond to state s. Copy numbers 0-1 correspond to DEL, while copy numbers 3-6 correspond to DUP, and there are two prior assumptions: 1) DEL and DUP are of equal likelihood, and 2) CNV size is exponentially distributed. Using this HMM, CLAMMS identifies CNVs as a series of call windows where the maximum-likelihood sequence of states is non-diploid, predicted by the Viterbi algorithm run in both the 5' to 3' and 3' to 5' directions. The developers found that running the algorithm in only the forward direction introduced a bias to the calling, as the probability threshold required to 'start' a CNV is higher than that to extend it, such that calls would tend to overshoot 3' breakpoints.

Based on the emission probabilities of the HMM, CLAMMS generates 6 quality (Q) metrics: Qsome, Qexact, Qleft_extend, Qright_extend, Qleft_contract, Qright_contract.

Quality metric	Description
Qsome	a Phred-scaled probability the call region contains any CNV
Qexact	a non-Phred-scaled score measuring how closely the coverage profile for the call region matches the exact called CNV state and breakpoints
Qleft_expand	Phred-scaled quality score for the left breakpoint, based on likelihood ratio of called breakpoint compared to breakpoint if call extended by one window
Qright_expand	Phred-scaled quality score for the right breakpoint, based on likelihood ratio of called breakpoint compared to breakpoint if call is extended by one window
Qleft_contract	Phred-scaled quality score for the left breakpoint, based on likelihood ratio of called breakpoint compared to breakpoint if call is contracted by one window
Qright_contract	Phred-scaled quality score for the right breakpoint, based on likelihood ratio of called breakpoint compared to breakpoint if call contracted by one window

Table 2.5. Descriptions of each CNV quality metric generated by CLAMMS CNV calling.

The file output by the CNV calling algorithm contains 18 fields. Fields 1-4 give the chromosome, start window coordinate, end window coordinate and interval for calls. 5 gives sample ID, while 6-7 specify CNV state (DEL or DUP), most likely copy number, and number of windows in the call. Fields 9-18 contain the 6 Q scores in Table 2.5, and the coordinates of the windows if the breakpoints are expanded or contracted by one window. For studies I conducted, CNV files were imported into R and subjected to a series of quality control filters, which I describe in the methods sections of the relevant results chapters.

2.2.5 Manual Inspection

The coverage distribution for individual CNV calls can also be manually inspected using a plotting script included in the CLAMMS code repository. For a given sample, the scripts' inputs are 1) list of CNV calls; 2) normalised coverage scores; and 3) model file. The output is a PNG file illustrating the mean coverage depth per exon impacted by a putative CNV, relative to model μDIP and σDIP. If an exon has been divided into >1 call window, the mean of coverage values for each window are calculated. Figure 2.4 shows an example plot, whose different elements are further explained in the legend. These plots were utilised in my quality control processes, again explained in the methods sections within the relevant results chapters.

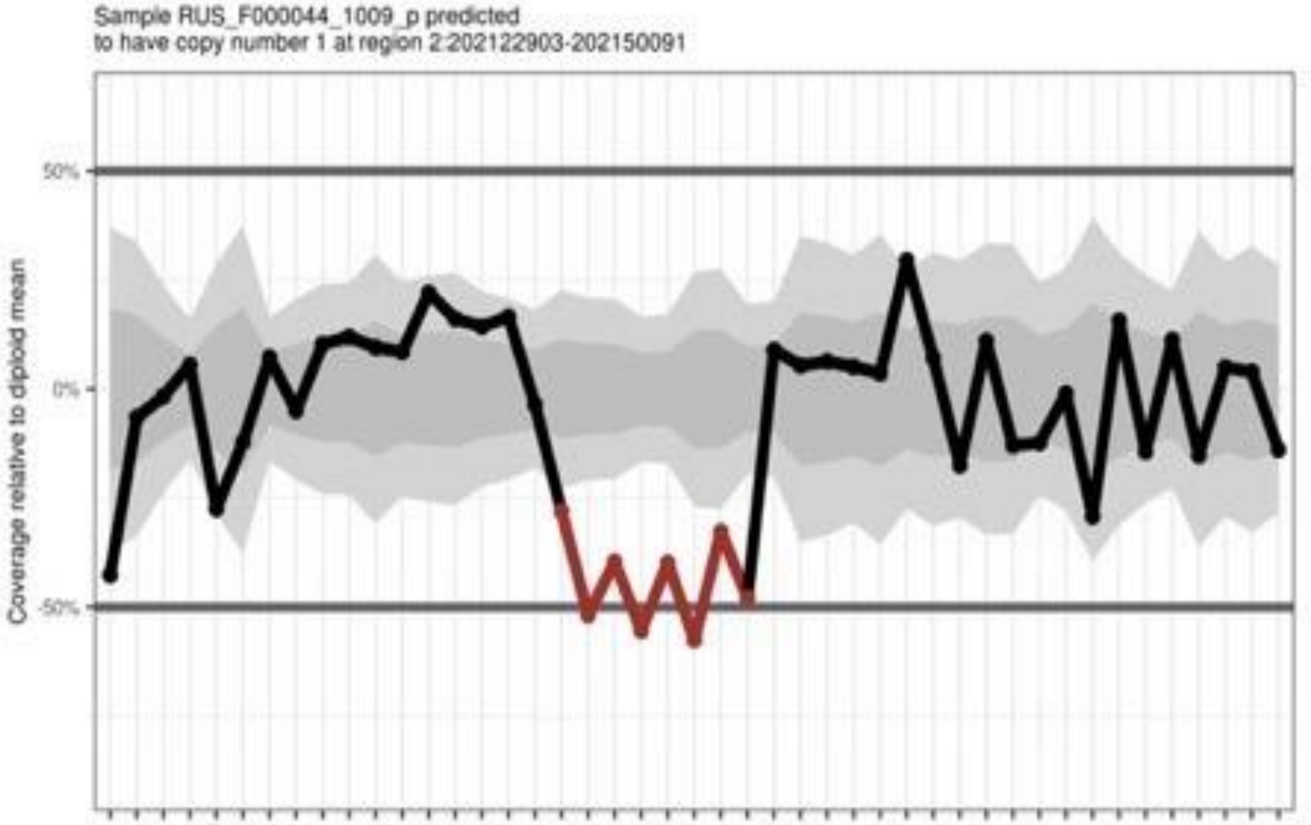


Figure 2.4. A CLAMMS coverage plot, showing a putative heterozygous deletion. The red lines highlight the CNV call region, while the black lines show the coverage profile for this sample +/-100kb beyond the CNV breakpoints. Each node along the line represents the mean coverage depth for an exon. 50% and -50% on the y axis show the expected coverage depth for heterozygous duplications and deletions,

respectively, relative to the mean of the sample's reference panel (µDIP). The dark and light grey regions indicate mean reference panel coverage depth +/- 1 and 2 standard deviations, respectively, for diploid copy number (σDIP).

2.2.6 Summary

In this section I have described the four stages of the CLAMMS algorithm: 1) generating call windows; 2) computing depth of coverage; 3) modelling copy number states; and 4) calling CNVs. Throughout these stages, there are two parameters that vary between different datasets: the insert size variable required to generate the call window, and the size of the reference panel used to model copy number states. Therefore, these are the two aspects of the algorithm that I focus on in the methods sections of the relevant results chapters. Unless specified, all other parameters were kept to their default values. In the next section, I justify my use of CLAMMS with reference to benchmarking analyses carried out by its developers.

2.3 Published CLAMMS studies

The developers tested CLAMMS on two WES datasets:1) Regeneron's WES database, reported in (Packer et al., 2016); and 2) adult participants of Regeneron's DiscovEHR study (Maxwell et al., 2017). The former was used primarily for call validation and method benchmarking, while latter was used primarily to test association of CNVs with clinical phenotypes in participant electronic health records (EHRs), and secondarily for further call validation. I describe the (Packer et al., 2016) study, and only those aspects of the (Maxwell et al., 2017) that pertain to call validation, as the details of the latter's phenotype analyses are not relevant to justify my use of CLAMMs, nor my phenotype of interest. I refer to the quality control procedures of the Maxwell study in the relevant results chapters, however, as they informed my own analyses.

2.3.1 Call validation

CLAMMs was applied to 3,164 samples included in Regeneron's WES database, for which high-quality CNVs had previously been called by PennCNV in their corresponding array data. Each algorithm was run with its 'default parameters'. Coverage depth models were based a on reference panel size of 100, as this value was found to give the best trade-off between the variance of model parameters and batch effect bias. Samples were excluded from the array-based test set if. 1) they had PennCNV calls > 50; 2) standard deviation of log R ratios > 0.23 (95th percentile); or 2) B-allele frequency drift >0.005 (95th percentile). To minimise false positives in the array-based call set, and to ensure that they could be detected by WES-based callers, calls were excluded if: 1) CNV length < 10kb or > 2mb; 2) they did not overlap at least 1 exon and 10 array SNPs; 3) they overlapped a gap in the GRCh37 reference genome; 4) had an AF > 1%. The last criterion is included as PennCNV is not designed to genotype common CNVs, so a high allele frequency call set is likely to be enriched for false positives. After sample and variant exclusions, the array-based call set included 1,715 CNVs (46% deletions, 54% duplications) across 1,240 samples.

Samples were excluded from the WES call set if they had >2x the median number of calls. The median for this study was 14, so 26 (0.8%) samples were excluded for having >28 calls. PennCNV calls for these samples were still included in the test set, however. CLAMMS CNV calls were subject to additional QC that is not specified in the paper but is likely the same as that used in (Maxwell et al., 2017). CLAMMS calls were compared to PennCNV calls according to three metrics: precision, recall and F-score. Precision refers to the percentage of PennCNV calls that were called by CLAMMS, subject to variant-level QC criteria. Recall is the percentage of PennCNV calls that were called by CLAMMS, with no variant-level QC applied. F-score is defined as ‘the harmonic mean of precision and recall.’ Values of these metrics were calculated for three degrees of reciprocal call overlap: any overlap, 33% overlap, and 50% overlap (Table 2.5).

Metric	Any Overlap	33.3% overlap	50% overlap
Precision	78.4	71.9	67.2
Recall	65.4	49.7	41.9
F-score	71.3	58.8	51.6

Table 2.5. Precision, recall and F-score metrics for CLAMMS calls tested for overlap with PennCNV calls. Adapted from (Packer et al., 2016) supplement.

For all three metrics, percentage decrease and degree of overlap are negatively correlated. This is expected, as the assigned breakpoints for each call will differ by method. CLAMMS systemically under-estimates breakpoint distance, as it can only capture the parts of a CNV that occur in exons, and it is unlikely that even one CNV breakpoint will occur within an exon. However, the 20-point difference in F-score between any overlap and 50% overlap suggests that not insignificant proportion of overlaps are missed by the latter threshold. Precision is at least 10 points greater than recall for all overlaps, indicating that quality control is effective at excluding false positive calls in both methods. Sixty-seven to 78% of PennCNV calls overlapped CLAMMS calls, subject to QC, indicating a high validation rate. To benchmark this performance, the authors compared it with that of four similar methods, described in section 2.3.2.

Calls produced from the same WES samples were also validated using TaqMan qPCR. Twenty rare and twenty-three common call loci were randomly selected from the set of all calls that intersected at least one disease-associated gene in the Human Gene Mutation Database ((Stenson et al., 2014), N = 7,430 genes), compared with PCR-based copy number predictions in 56 samples for the rare loci, and 165 samples for the common loci. Nineteen of twenty (95%) of the rare variants were validated, although authors don’t specify precisely what this means (i.e, whether each of the nineteen loci were validated in one, most, or all the 56 samples). Four of the common loci ‘appeared to be correct’, but results were ambiguous and inconclusive to due high variance in the PCR data. Two loci had ambiguous PCR results for two samples but validated the calls in the rest. The remaining seventeen non-ambiguous loci were all validated. The only rare locus that was not validated was also the smallest, at 718bp in length, though there appeared to be no correlation between size and PCR validation among the common loci, the smallest of which was 559bp.

In (Maxwell et al. 2017), the CLAMMS developers identified 475,664 CNVs in 47,349 DiscovEHR participants. The median size of rare (allele frequency < 1%) CNVs was 17.8 kb (Deletions 8.4kb, Duplications 32.8kb), while the ratio of rare duplications to deletions was 1.6:1. There was an average of one very rare (allele frequency <0.1%) CNV per individual. The developers used this call set to extend the PCR validation reported in (Packer et al., 2016), assessing the algorithm's error rates regarding variants of the lowest size range. Testing was carried out on small loci identified as non-transmissions and transmissions in 333 proband-parent duos included in DiscovEHR cohort, to assess the similarity of transmission rates obtained from PCR results with those of the CLAMMS quality-controlled call sets. All loci were 1-3 exons in length, with 21 corresponding to a single call window. 89.4% of predicted non-transmissions were PCR validated (i.e. true positive in parent, true negative in proband), with a 86.9% validation rate for single exon loci. One hundred percent of transmissions were PCR validated (true positive in both parent and child), consistent with the assumption that the transmission status of a call increases confidence in its true positive status in both proband and parent. Collectively, these findings demonstrate that CLAMMS can adequately control for coverage depth confounders even at single exon resolutions. Given that the impact of batch effects can vary between datasets, however, it cannot be concluded that CLAMMS can call small CNVs in all samples with the same error rates.

2.3.2 Benchmarking

In (Packer et al. 2016), Precision, recall and F-score metrics were calculated for calls generated from the same data analysed by 4 additional publicly available callers, all of which leverage coverage depth in a similar way to CLAMMS: 1) XHMM (Fromer et al., 2012), CoNIFER (Krumm et al., 2012), CANOES (Backenroth et al., 2014) and ExomeDepth (Plagnol et al., 2012). All callers were run using default parameters and QC procedures recommended by their respective developers. Using the any overlap criterion, CLAMMS had an F-score 9.3% higher than XHMM, 6.6% higher than ExomeDepth, and 38.2% higher than CoNIFER. CANOES would not run on the server with the 30GB memory limits required by the developers, so no F-score comparison could be made. The CLAMMS developers argue that observed improvements are due to their algorithm's higher precision, which was ~20% higher for all overlap criterion than the next best performing caller (XHMM), reflecting more robust quality control procedure. In summary, CLAMMS is reported by its developers to have a significantly higher rate of validation compared to alternatives, reflecting a greater ability to detect true positive events.

2.3.3 Computational advantages

In addition to improved performance in variant calling, the CLAMMS developers report several advantages over alternative methods in terms of computational efficiency. These are mostly due to the way CLAMMS handles batch effects: a k-d tree for 7 quality metrics requires minimal computational power to produce, even for thousands of samples, and to extract reference panels from. This process is therefore quick and efficient to repeat if more samples are added to a dataset, or a different reference panel size is chosen. In contrast, CoNIFER and XHMM control for batch effects by calculating principal components of a sample-by-exon coverage depth matrix and excluding the contribution of the largest components. This requires construction of a very high dimensional space and is therefore computationally intensive (and likely why CoNIFER failed to run in the benchmarking analysis). Computation time also scales exponentially with sample size, limiting the application of these methods to large samples. ExomeDepth and CaNOES also control for batch effects by direct use of coverage depth data but are similar to CLAMMS in that they use a reference panel approach. Each sample's coverage profile is normalised against the average of a reference panel of samples with whose coverage profile which with it is most highly correlated. Again, however, calculating these coverage correlations across the entire exome is computationally intensive and shares the same scalability issues as the principal component method. CLAMMS is therefore the only the method whose computational power scales linearly with sample size, making it most suited for integration in CNV calling pipelines.

2.4 CLAMMs Summary

In this section I described each stage of the CLAMMS algorithm, showing precisely how it uses coverage depth across exon call window to model copy number states, and thereby call CNVs. I have also presented evidence, reported in the CLAMMS method paper, that CLAMMS both produces higher validation rates than alternative approaches, due its higher precision, and is more computationally efficient. The array call validation rate was important for my purposes, as much of my research involved assessment of the utility of WES and array-based approaches for detecting CNVs that meet different sets of criteria (size, type etc.). Using the approach that can call array-based CNVs with the greatest precision would lead to the most robust comparison of calls. The computational efficiency of CLAMMS was also important, as I would be applying the method to different datasets and therefore needed a method that could be easily integrated into an SV calling pipeline.

3. InDelible

3.1 Introduction

The second SV caller I describe is InDelible, developed by the Hurles lab at the Wellcome Sanger Institute (Gardner et al. 2021, code repository: https://github.com/HurlesGroupSanger/indelible). The impetus for InDelible's development was the observation that SVs < 1kb in size were undetectable by existing array-based and WES-based methods. The potential clinical impact of such variants was therefore unknown, limiting the knowledge and diagnoses of congenital disorders that are driven by structural variation. To detect these small variants, InDelible mines the CIGAR strings of aligned reads in a BAM for soft-clipped bases. If a read has soft-clipped (i.e. misaligned) bases at its 5' or 3' end (or both), it is referred to as 'split'. Split reads (SRs) can indicate the breakpoints of many SV types, allowing for the detection of events at very small bp resolutions, often smaller than the reads themselves.

3.2 Split reads

The number of bases that misalign in a split read is indicated by the integer preceding the 'S' (soft-clipped) signifier in their CIGAR string. Thus, the string '24M12S' corresponds to a split read consisting of 35 bases, whose first 24 bases (in the 5'-3' direction) are matched/aligned to the reference genome, and whose last 12 bases are soft-clipped. Figure 2.5 illustrates a split read that would have this CIGAR string. The junction between the aligned and misaligned bases in a split read is called the split position. The soft-clipped bases may or may not align to another reference sequence, depending on SV type.

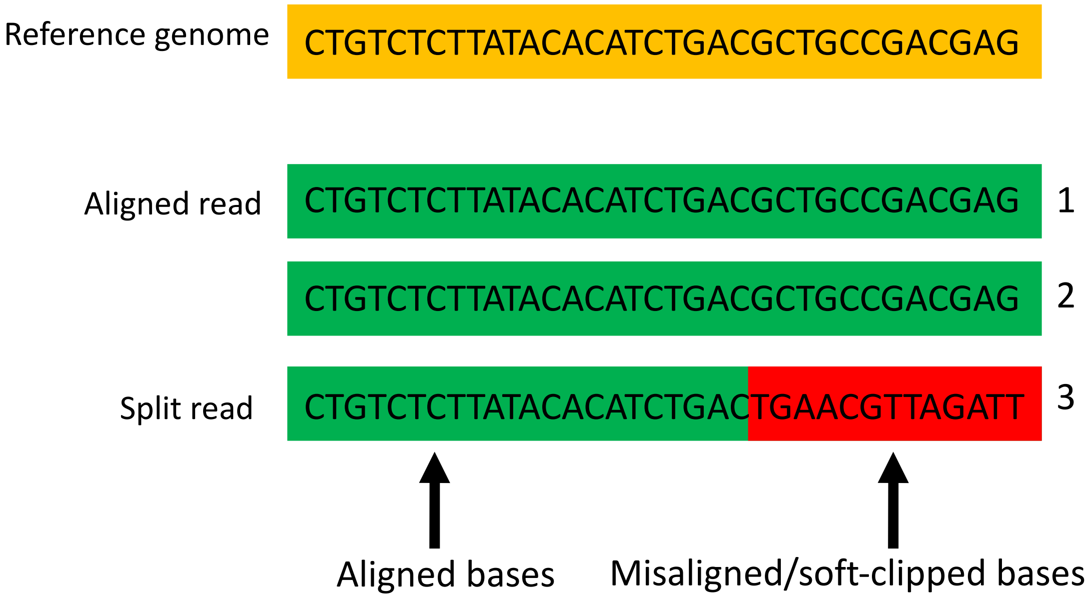

Reference genome
CTGTCTCTTATACACATCTGACGCTGCCGACGAG
Aligned read
CTGTCTCTTATACACATCTGACGCTGCCGACGAG
1
CTGTCTCTTATACACATCTGACGCTGCCGACGAG
2
Split read
CTGTCTCTTATACACATCTGACTGAACGTTAGATT
3
Aligned bases
Misaligned/soft-clipped bases

Figure 2.5, An illustration of a split read with CIGAR strig '24M12S'. The yellow bar shows the reference sequence at this locus. There are 3 reads mapped the reference, the first two of which contain no misaligned bases. The first read is split at its 3' end, indicated by the red region. The junction between the aligned and misaligned bases (where the read turns from green to red) is the split position. The 12 soft-clipped bases may align to another reference sequence, depending on the SV type that caused them.

Insofar as an SR does indicate an SV breakpoint, there are likely to be other reads mapping to the same reference sequence that split at the same position (assuming adequate coverage). This is because, in the absence of read mismapping, they will be based on the same source DNA sequence, and therefore carry the same variant 'signature'. The SRs need not align to precisely the same reference sequence to have the same split positions, however; they just need to overlap at the reference position where the breakpoint occurs. SRs with the same split position are denoted as 'clusters' by the InDelible developers, and the algorithm is designed to detect these clusters and filter those that are unlikely to be caused by real SV events.

3.3 InDelible algorithm.

The algorithm has 6 stages: 1) Fetch: mine WES BAM files for SR; 2) Aggregate: merge SRs with the same split positions into clusters; 3) Score: score each cluster with a probability that it is not an artefact, using a random forest adaptive learning model; 4) Database: build allele frequency database for clusters, and determine the type, size, and other breakpoints of their corresponding SVs; 5) Annotate: annotate clusters with their allele frequency and gene intersects; and 6) denovo: an optional step that mines parental data (if available) for the presence of clusters called in probands. Figure 2.6 is an illustration of the algorithm taken from the InDelible methods paper. In the following six sections I describe each stage and how they are executed.

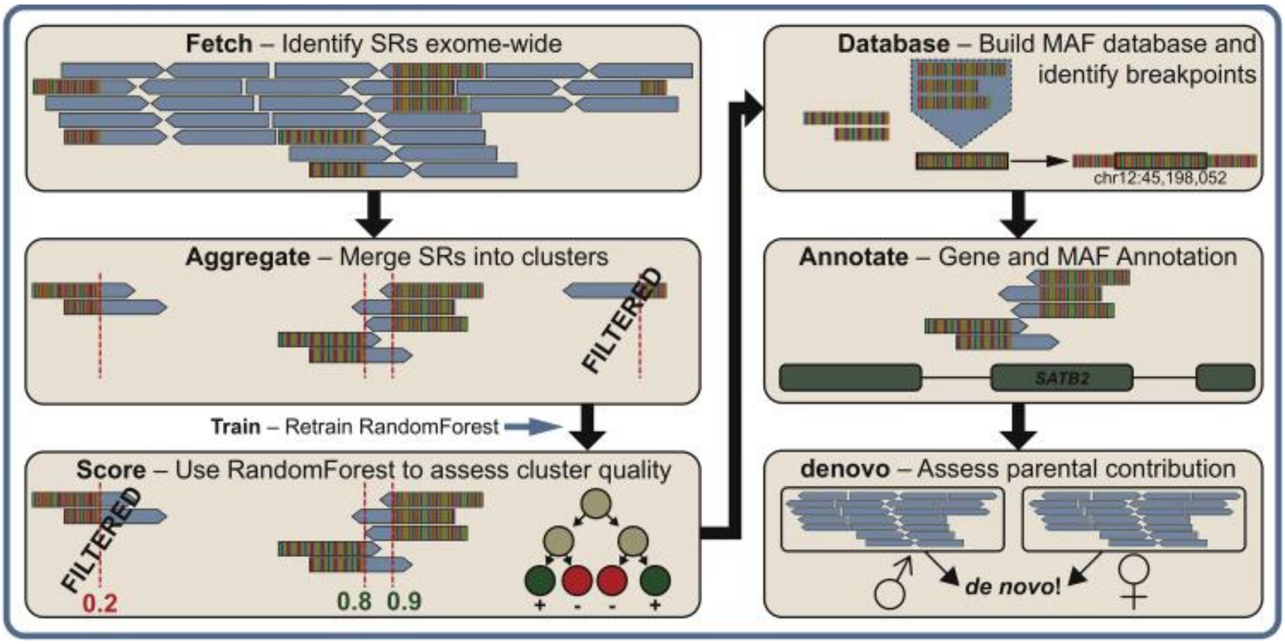

Fetch – Identify SRs exome-wide
Aggregate – Merge SRs into clusters
FILTERED
Train – Retrain RandomForest
Score – Use RandomForest to assess cluster quality
FILTERED
0.2
0.8 0.9
+ - - +
Database – Build MAF database and identify breakpoints
chr12:45,198,052
Annotate – Gene and MAF Annotation
SATB2
denovo – Assess parental contribution
♂
de novo!
♀

Figure 2.6. The 6 steps of the InDelible structural variant calling algorithm, adapted from (Gardner et al. 2021). The horizonal grey lines in each diagram represent reads, and the multi-coloured regions show where reads are split. This is based on how split reads appear in the Integrative Genomics Viewer (IGV), described in section 3.4. SR = split read, MAF = minor allele frequency.

3.3.1 Fetch

In the first stage, soft-clipped reads are ‘fetched’ from all BAM files to be analysed, by interrogating the CIGAR and SEQ fields of all entries. Two input files are required: 1) an indexed BAM file; and 2) a configuration file. The latter is an input for multiple stages and specifies paths to several datasets in the InDelible code repository, along configurable parameters by which the algorithm filters SRs or SR clusters. For Fetch, the parameters are minimum read mapping quality, minimum average base quality, and minimum SR length. Split reads are not considered if they do not meet one of the criteria specified, the defaults for which are 5, 10, and 5, respectively.

The output file generated by Fetch contains 10 fields, with each entry corresponding to a single SR. The first 5 specify its chromosome, split position, split end (5' or 3'), number of soft-clipped bases and sequence of soft-clipped bases. Fields 6-8 specify ASCII base quality of soft-clipped bases, read mapping quality and average base quality; and the final two fields are Boolean variables indicating whether the aligned bases reverse complement the reference, and whether the read is split at both ends, i.e 'double-split'. Double-split reads will have two file entries: one for the soft-clipped bases on their 5' end, and the other for those at their 3' end.

3.3.2 Aggregate

In the second step, SRs identified by Fetch are clustered according to their chromosome and split positions, and features are calculated that will be used as inputs for Score's adaptive learning model. The proportion of reads in each cluster that are double-split is also calculated for downstream QC. Aggregate requires 4 input files: 1) an SR file generated by Fetch; 2) an indexed BAM; 3) a reference genome in FASTA format; and 4) a configuration file. The configuration file can be the same file used in the previous stage. For Aggregate, the configurable exclusion parameter is the minimum number of SRs required to form a cluster, which is set to 3 by default. Another parameter can be set for purposes of cluster annotation, which is the minimum number of soft-clipped bases for an SR in a cluster to be considered 'short'. This is set to 10 by default and informs calculations in both the Score and Database stage. A third parameter specifies the window size around the split position for which coverage depth should be calculated, set to 5 (+/-5bp) by default.

The output file contains 21 headers, with each entry corresponding to a single SR cluster. The first three are chromosome, split position, and coverage depth in set window around the split position. 4 is 'insertion context', which is the number of insertions detected each cluster (derived from its CIGAR strings of its constituent SRs), and 5 is 'deletion context', the same but for deletions. Fields 6-12 give the number of SRs are 'short' and 'long', according to the configuration file parameter, and the number of SRs that are short and long at the 3' end and the 5' of the cluster.

Fields 13-17 gives the entropy (i.e. base variability) of the longest soft-clipped sequence, the sequence +/-20bp around the split position, the sequence from the split position +20bp, and the sequence from the split position -20bp. Fields 18-21 contain a sequence similarity score for SRs in the cluster, the average of the average soft-clipped base quality, average SR mapping quality, the longest soft-clipped base sequence, and finally the proportion of SRs that are double-split.

3.3.3 Score

In the third stage, the 17 metrics calculated in Aggregate (the final 17 fields of its output) are used as inputs for a 500-tree Random Forest adaptive learning model, which is by default trained on a set of 2,000 manually curated SR clusters called from WES data generated from participants of the Deciphering Developmental Disorders (DDD) study. I describe this dataset, and how InDelible was applied to it, in section 3.4. Based on the features exhibited by the training set, the model scores clusters according to their probability of being a true event. The inputs for Score are: 1) a cluster file produced by Aggregate; and 2) a configuration file. Score uses the configuration file only to access the training data included in the InDelible code repository. A subfunction of Score, ‘Train’, can be used to retrain the model based on a truth set supplied by the user. The output of Score is largely identical to that of Aggregate, with the addition of three fields: probability that the cluster is a false positive (prob_n), probability that the cluster is a true positive (prob_y), and Boolean variable that is set to ‘Y’ if prob_y > 0.5 (i.e. likely to be true positive), and otherwise ‘N’.

3.3.4 Database

In the fourth algorithm stage, InDelible calculates the frequency of each cluster, and the type, size, and other putative breakpoint of SVs. Its inputs are: 1) a file containing the file paths all the Score outputs, 2) reference genome, in FASTA format and 3) a configuration file. Frequencies are calculated by counting clusters with same the split positions and dividing by total number of Score outputs. InDelible does not factor in this calculation the possibility that SVs of different types could have the same split position, presumably because it highly unlikely.

Ascertainment of SV type is derived from the longest soft-clipped sequence across all clusters at the same position, defined as that sequence which has at least 60% homology to all other sequences. In the first step of this process, A synthetic FASTQ file is created whose unmapped reads correspond the longest soft-clipped sequences. BWA-mem is used to determine the reference sequence with which the sequences align. As it is implemented by InDelible, BWA-mem will only process sequences that are at least 19bp length. SV type can then be ascertained from the position of the alignment relative to the initial position of the soft-clipped bases. Section 3.5 serves to clarify, using diagrams, the precise relation between SR cluster patterns and SV type that is leveraged by Database.

As some SV types produce split reads that map to several loci across the genome, the longest soft-clipped sequences are separately aligned to a database of repeated sequences (such as mobile elements) using the BLAST aligner. BLAST operates similarly to BWA-backtrack but is designed for alignment of sequences with curated sequence databases, rather than a reference genome. As implemented by InDelible, BLAST will only process sequences that are at least 22bp length. Database uses BLAST ‘hits’ to assign SV type if there is not a unique BWA-mem alignment for a soft-clipped sequence, there are repeat BLAST hits, and these hits correspond to a repeated sequence whose type (e.g. *Alu*, SINE, LINE) is defined in the sequence database.

If the sequence does not have a BWA-mem alignment or a BLAST hit, it’s type cannot be determined. This is expected in the case of simple insertions, which therefore need to be ascertained by other means, described in section 3.5. Otherwise, InDelible can assign 6 SV types: deletions, duplications, translocations, mobile element retrotranspositions, pseudogene retrotranspositions and complex events involving insertions nested within deletions or duplications. Small tandem repeats are detectable from split read data, but high variability in their cluster patterns mean that they cannot be assigned by InDelible. SRs whose soft-clipped sequences are in reverse orientation to the reference genome can be evidence of inversions. However, the InDelible authors claim that this is more likely to be an artefact, and therefore exclude such sequences from type assignment.

Calculation of SV size is also contingent on unique BWA-mem alignment of soft-clipped sequences to the reference genome. In all cases, depending on whether the soft-clipped sequences occur at the 5' or 3' end of their associated SR, the 5' or 3' end of the reference sequence to which they align will correspond to the other/alternate breakpoint of the SV. The precise bp size of the SV is therefore determined by subtracting this alignment position from the cluster split position. If the alignment position is upstream of the split position, size is given as a negative value. Database interrogates allele frequency data to establish if the alternate breakpoint of an SR cluster is equal to the split position of another cluster included in the Sore output, thereby determining if the alternative breakpoint was also called by InDelible.

The output of Database contains 14 fields, which each field corresponding to a cluster that occurs at least once in the Score output. The fields are: chromosome, split position, split position frequency, split position count, total individuals assessed (i.e. number of Score output files), mean coverage around split position (calculated by Aggregate), alignment position, alignment mode (whether BWA-mem or BLAST was used to determine SV type, set to 'FAIL_*' if alignment failed), SV type, SV size, length of alignment, a Boolean variable for whether the alternate breakpoint was included in the Score output, a Boolean variable for whether the split position corresponds to the 5' breakpoint, and a variant coordinate in the format CHR:BP1-BP2.

3.3.5 Annotate

The fifth stage of the algorithm annotates the SR clusters in the output of Score with their corresponding fields contained in the Database output, according to the chromosome and split positions. It also annotates clusters with their gene intersects in the appropriate reference genome. The input files are: 1) A Score output file; and 2) A Database output file; and 3) a configuration file. By default, Annotate uses the configuration file to access gene intersect reference genome coordinates, in ENSEMBL format. In addition, Annotate accesses a file listing exon transcript coordinates. Optional files can be specified in the configuration file for coordinates of genes contained in gene sets of interest. Unless parental data is available for the denovo stage, the output of Annotate is used for downstream QC. It contains 41 fields, with each entry corresponding to an SR Cluster. 24 fields are from the Score output, 9 from the Database output, and three additional fields are added for

ENSEMBL gene intersect, a Boolean variable for whether the cluster intersects an exon transcript, and the IDs of transcript intersects.

3.3.6 denovo

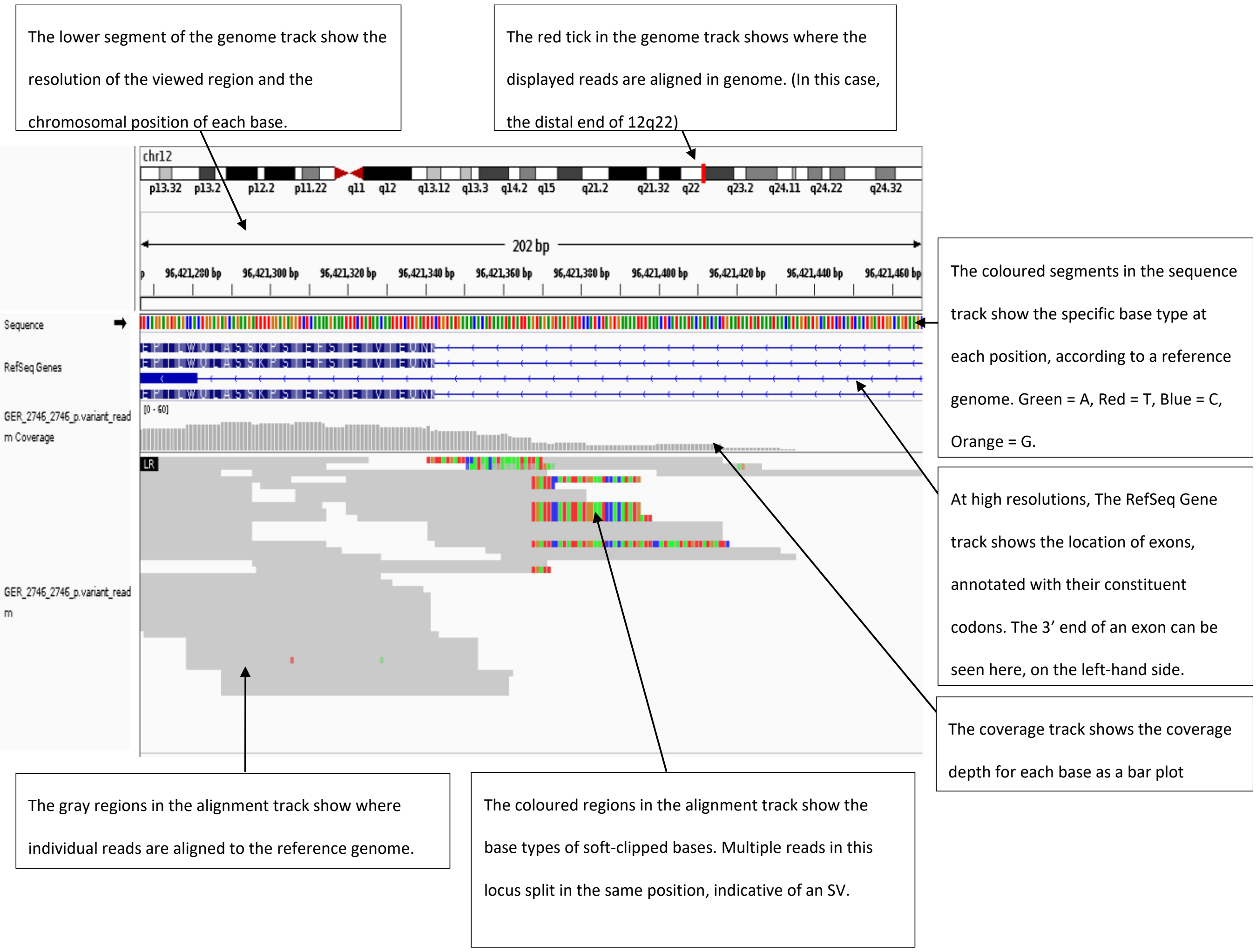
The final stage of the algorithm can only be applied to samples if the indexed BAM files for one or both parents is available. denovo interrogates the split positions of proband clusters in parent data to determine putative transmission status. The required inputs are: 1) an Annotate output file; 2) a maternal or paternal BAM, or both; and 3) a configuration file. Parameters of the configuration file relevant for denovo specify the minimum prob_y score threshold for proband clusters to be considered (default = 0.6), the minimum required coverage around the call position in parent data (specified by window size parameter, default = 9), and maximum number of reads that split at the same position in the parent for the SV to be considered inherited (default = 3). prob_y score threshold 0.6 was found by the developers to optimise false positive and false negative trade-offs in DDD sample calls. The denovo output add 3 fields per parent to the Annotate output: 1) number of SRs at cluster split position in parent; 2) the insertion/deletion context of the reads overlapping the split position in parent; and 3) coverage around the split position in parent.

3.3.7 Summary

In this section I described each of the 6 stages of the InDelible algorithm, along with their required inputs and their output files. The first 5 can be run on any WES BAM dataset, but the 6th can only be run on datasets that incorporate parental data. Any changes I made to default parameters in my own analyses will be explained in the relevant results chapters. In the next section, I explain how InDelible calls can be inspected using alternate methods, and the precise relation between SR patterns and SV types.

3.4 Manual Inspection

InDelible calls can be manually inspected in the Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2013). IGV enables direct viewing of read alignments in a BAM file at a given reference genome locus. Soft-clipped bases can be highlighted, thereby allowing for the visual identification of split reads. An example of the IGV interface (version 2.9.1), with reads loaded from a BAM file analysed in my own research, is shown in Figure 2.7. Explanations of each element of the interface are included alongside the figure.



The lower segment of the genome track show the resolution of the viewed region and the chromosomal position of each base.

The red tick in the genome track shows where the displayed reads are aligned in genome. (In this case, the distal end of 12q22)

The coloured segments in the sequence track show the specific base type at each position, according to a reference genome. Green = A, Red = T, Blue = C, Orange = G.

At high resolutions, The RefSeq Gene track shows the location of exons, annotated with their constituent codons. The 3’ end of an exon can be seen here, on the left-hand side.

The coverage track shows the coverage depth for each base as a bar plot

The gray regions in the alignment track show where individual reads are aligned to the reference genome.

The coloured regions in the alignment track show the base types of soft-clipped bases. Multiple reads in this locus split in the same position, indicative of an SV.

Figure 2.7: An example of the IGV interface loaded with a BAM file, illustrating reads aligned to a chromosome 12 locus. The textboxes alongside include descriptions of each element of the interface.

IGV includes a function which can generate snapshots of the interface from a list of reference loci paired with BAM files, which can be used to inspect calls more efficiently. As I used this function in the same way for all InDelible analyses, I give the relevant details here. First, I created a 200bp (+/-100bp) window around each call position. These windows are large enough to encompass the size of most human exons, thus allowing for the inspection of all reads adjacent to the SR cluster, but small enough to allow inspection of reads at the resolution of individual bases. In addition, the developers report in (Gardner et al. 2021) that InDelible is most sensitive to known variants 11-50bp in size (see section 3.4 for further details), such that most true positive variants I detect should have breakpoints that fit within a 200bp locus. I then created file with 3 fields, with 1 entry per InDelible call: 1) sample ID; 2) the 200bp window in format CHR:BP1-BP2; and 3) a prefix for the snapshot output containing the a unique call ID and the sample ID. I input this file into a script that assigns each sample ID and window as variables and instructs IGV to generate a snapshot of each window in the relevant samples. In addition to sample ID and loci, the reference build must also be specified prior to running the snapshot function. Additional commands allow for the configuration of output image dimensions and how reads in the alignment track are sorted (https://github.com/igvteam/igv/wiki/Batch-commands). I sorted reads by base, which ensures all reads in the same order of their aligned bases in the reference genome.

Insofar as both breakpoints occur with the 200bp window, these snapshots can be used both to confirm the SV type, size, and transmission status of InDelible call, and to assign type and size in cases where the algorithm had failed. In the next section, I describe the pattern of SR clusters associated with each SV called by InDelible, and how to determine SV type and size from IGV snapshots. There are several criteria by which a variant may be excluded by manual inspection. As some are specific to datasets, I describe them in the relevant results chapters.

3.5 Split read patterns and structural variation

Each type of SV is associated with a specific pattern of SRs, according to which their type and size can be determined by InDelible or by inspecting their call positions in IGV. The SR patterns of variants that InDelible can detect are described and illustrated below. Some of the figures depicting reads in IGV were created by the snapshot function, while others are screenshots taken directly within the interface. The elements of the interface are the same in both cases.

3.5.1 Deletions

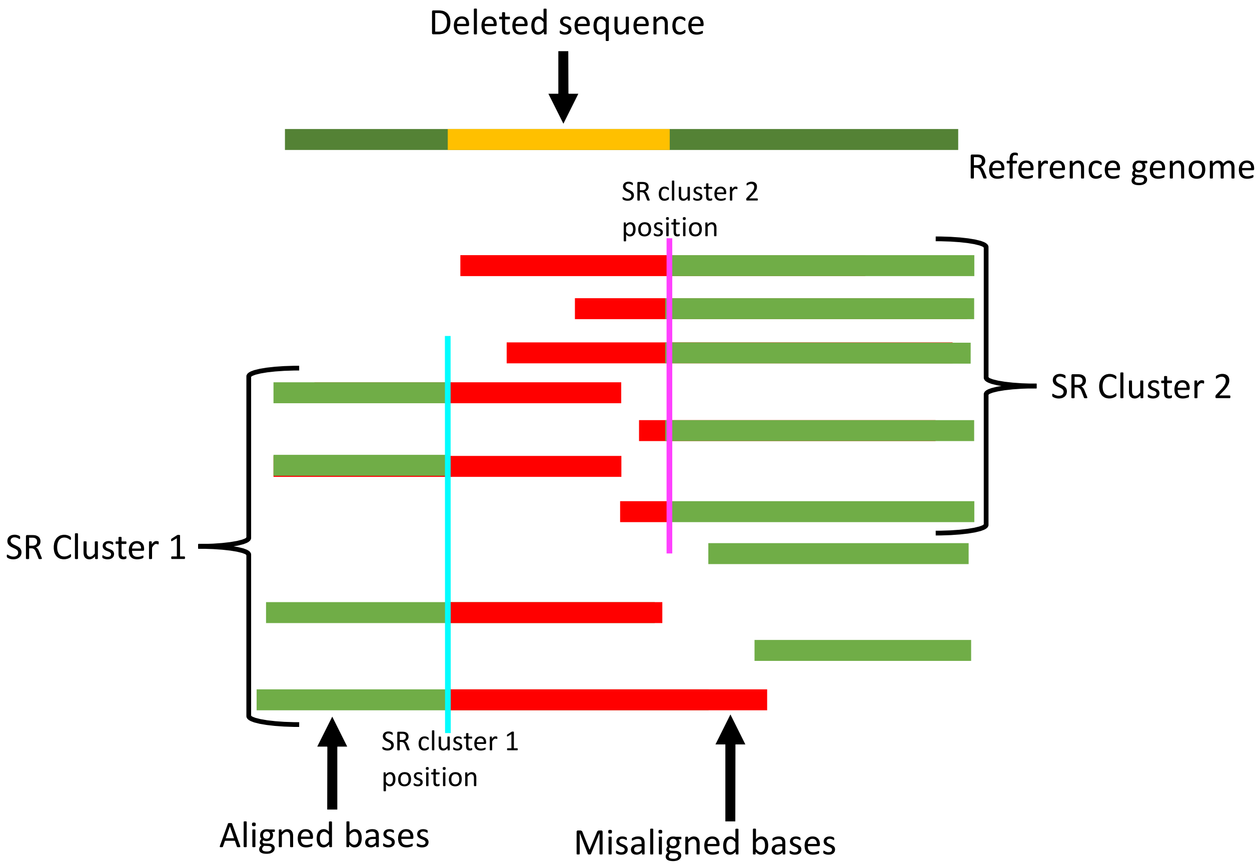


Figure 2.8. The split read pattern associated with a deletion. The reference genome is represented by the dark green line, within which the deleted sequence is coloured in gold. The light green lines represent the aligned bases of reads mapped to this locus, while the red lines represented their misaligned/soft-clipped bases. The vertical cyan and pink lines intersect the split positions of the two split read clusters, SR Cluster 1 and SR Cluster 2. SR = split read.

A deletion is indicated by two SR clusters, corresponding the deletion breakpoints with respect to the reference genome. The 5' cluster (SR Cluster 1) splits at its 3' end, while the 3' cluster (SR cluster 2) splits at its 5' end. The soft-clipped bases of the clusters overlap one another if the event is smaller than the reads themselves (~ <150bp), an example of which is illustrated by Figure 2.8. The position at which SR Cluster 1 splits aligns to the 5' end of the deleted sequence, which is indicated by the cyan line in the figure. The position at which the SR Cluster 2 splits aligns to the 3' end, indicated by the pink line. Thus, the precise size of the deletion corresponds to the distance between the two SR Cluster positions. The soft-clipped bases in SR Cluster 1 align to the reference sequence directly after the 3' end of the deleted sequence, while the soft-clipped bases in SR Cluster 2 align to the bases directly before its 5’ end. Deletions are also often indicated by a drop in coverage across the deleted sequence, as some reads that are impacted by the deletion will have <50% alignment with the reference and are therefore less likely to be successfully mapped than adjacent reads that are unaffected. Figure 2.9 shows an example of this pattern observed in an IGV snapshot.

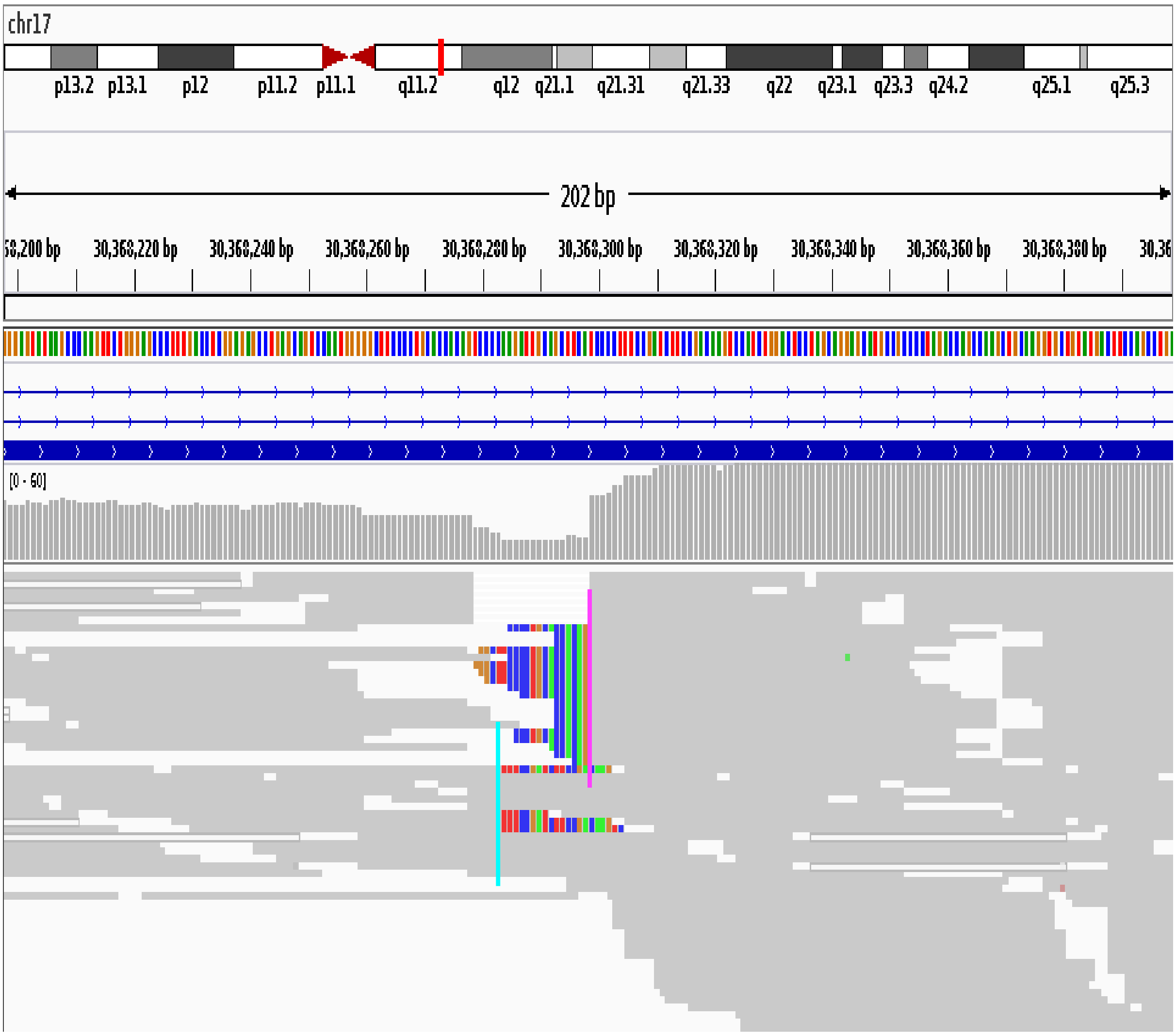


Figure 2.9. IGV snapshot showing evidence of an 18bp deletion on chromosome 17. I have used the same-coloured vertical lines in Figure 2.8 to annotate the split positions of the two SR clusters at this locus. The cyan line indicates the position of SR Cluster 1 and the pink line the position of SR Cluster 2. The size of the SV can therefore be calculated from the snapshot by counting the bases between the split positions. A decrease in coverage across the deleted sequence can be observed in the coverage track.

3.5.2 Duplications

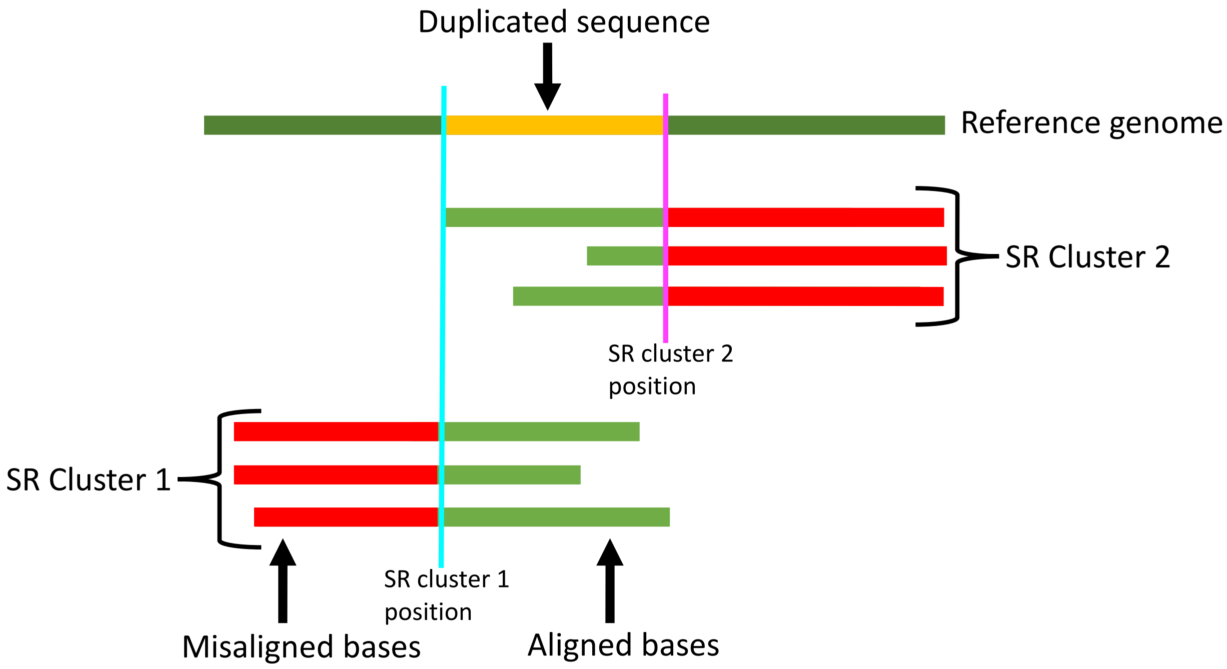


Figure 2.10. The split read pattern associated with a duplication. The reference genome is represented by the dark green bar, within which the duplicated sequence is coloured in gold. The light green bars represent the aligned bases of reads mapped to this locus, while the red bars represented their misaligned/soft-clipped bases. The vertical cyan and pink lines intersect the split positions of the two split read clusters, SR Cluster 1 and SR Cluster 2. SR = split read.

A duplication is also indicated by two SR clusters, corresponding to the duplication breakpoints with respect to the reference genome. The 5’ cluster (SR Cluster 1) splits at its 5’ end, while the 3’ cluster (SR cluster 2) splits at its 3’ end. Reads affected by a duplicated sequence will split at the junction of its first copy and second copy. Depending on which copy aligns to the reference genome, the SRs will be soft-clipped at their 5’ or 3’ ends. In Figure 2.10, the second copy of the duplication in SR Cluster 1 is aligning with the reference, so the reads are split at their 5’ end (cyan line). In SR Cluster 2, the first copy is aligning with the reference, so the reads are split at their 3’ end (pink line). As in the case of deletions, the precise size of the duplicated sequence therefore corresponds to the distance between the two split positions. The soft-clipped bases in SR Cluster 1 will align to bases directly before the 3’ of the duplicated sequence, while those in SR Cluster 2 align to bases directly after the 5’ end. Unlike the SR pattern for deletions, the misaligned bases of the two clusters can never overlap. Duplications are not typically associated with a decrease in coverage, as >50% of the bases in all affected reads can be aligned. If the event is large enough for reads to be entirely nested within the duplicated sequence, there will be an increase in coverage between the breakpoints. Figure 2.11 shows an example of this pattern observed in an IGV snapshot.

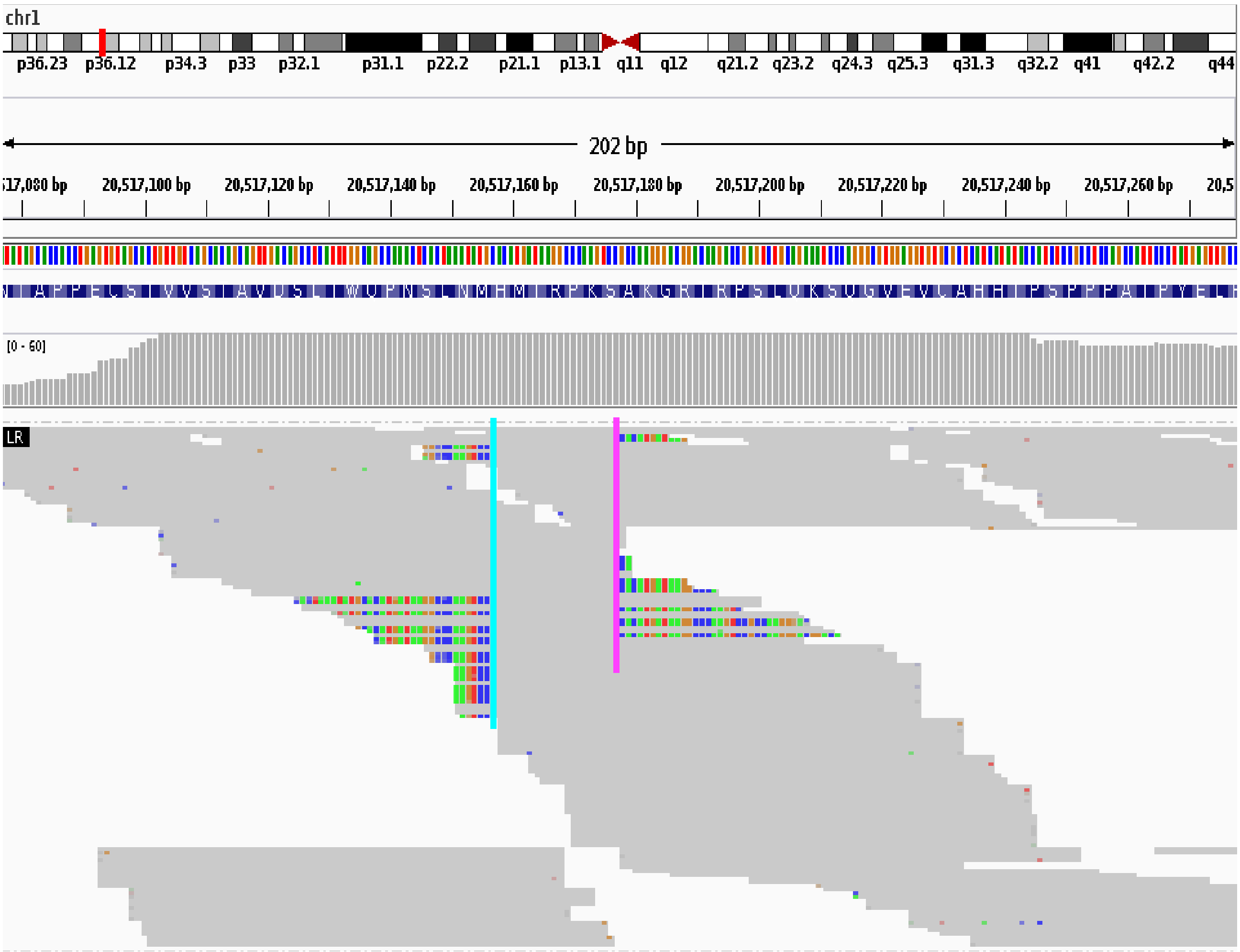


Figure 2.11. IGV snapshot showing evidence of a 20bp duplication on chromosome 1. I have used the same-coloured vertical lines in Figure 2.10 to annotate the split positions of the two SR clusters at this locus. The cyan line indicates the position of SR Cluster 1 and the pink line the position of SR Cluster 2. The size of the SV can therefore be calculated from the snapshot by counting the bases between the split positions.

3.5.3 Simple insertions

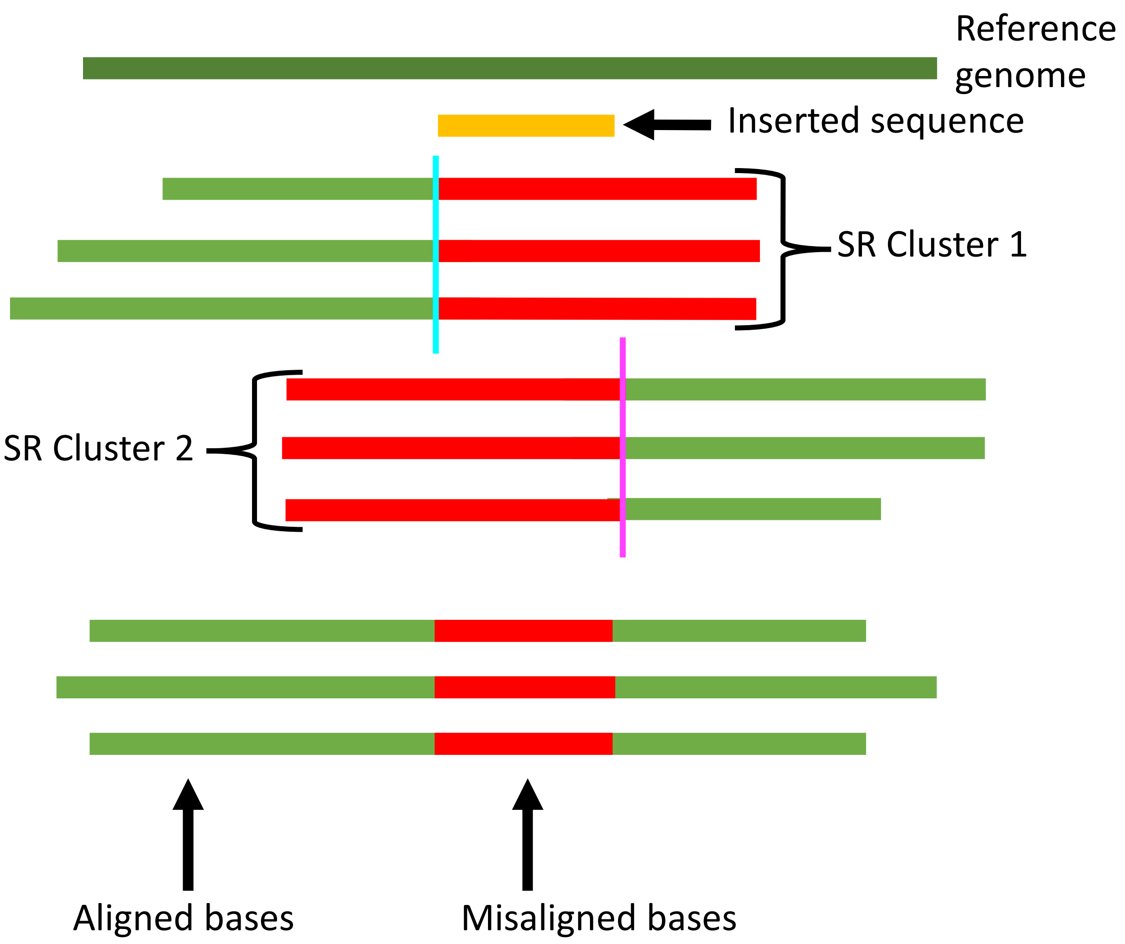


Figure 2.12. The split read pattern associated with a simple insertion. The reference genome is represented by the dark green bar. As the inserted sequence (gold line) is, by definition, not in the reference genome, I have shown its position relative to reference genome just below it. The light green bars represent the aligned bases of reads mapped to this locus, while the red lines represented their misaligned bases. The lower three reads are not split but have misaligned bases nested between their aligned 5' and 3' ends corresponding to the inserted sequence. The misaligned bases in the SR cluster have an 'S' (soft-clipped) signifier in their associated CIGAR string, while those in the lower three reads have and 'I' (insertion) signifier.

Simple insertions are indicated by one or two SR clusters, depending on how reads that mapped to the reference are affect by the inserted sequence. If the inserted sequence is only at the 5' or 3' of all reads, there will be one cluster. It is at the 5' end of some reads, and at the 3' of others, there will be two clusters (Figure 2.12). In the figure, the inserted sequence causes the reads in cluster 1 to split at their 3' ends, shown by the cyan line. However, it also causes the reads in cluster 2 to split at their 5' end, show by the pink line. The first misaligned bases in these reads will correspond to the insertion itself, followed by reference bases that occur after the inserted bases. In the case of simple insertions that are small enough to be nested in reads, this split read pattern occurs alongside reads with nested misaligned bases (the lower three reads) which correspond precisely to the inserted sequence and have an 'I' (insertion) signifier in their associated CIGAR string, rather than an 'S' (soft-clipped) signifier. In IGV, the misaligned bases in these reads are truncated and replaced with a purple marker, which when selected opens a window showing the size and constituent bases of the insertion, based on the read's CIGAR string (Figure 2.14). A decrease in coverage across the inserted sequence is also typical, as impacted reads are less likely be successfully mapped. Figure 2.13 shows an example of this event in an IGV snapshot, in which a single SR Cluster, split at its 3' end, and nested misalignments can be observed.

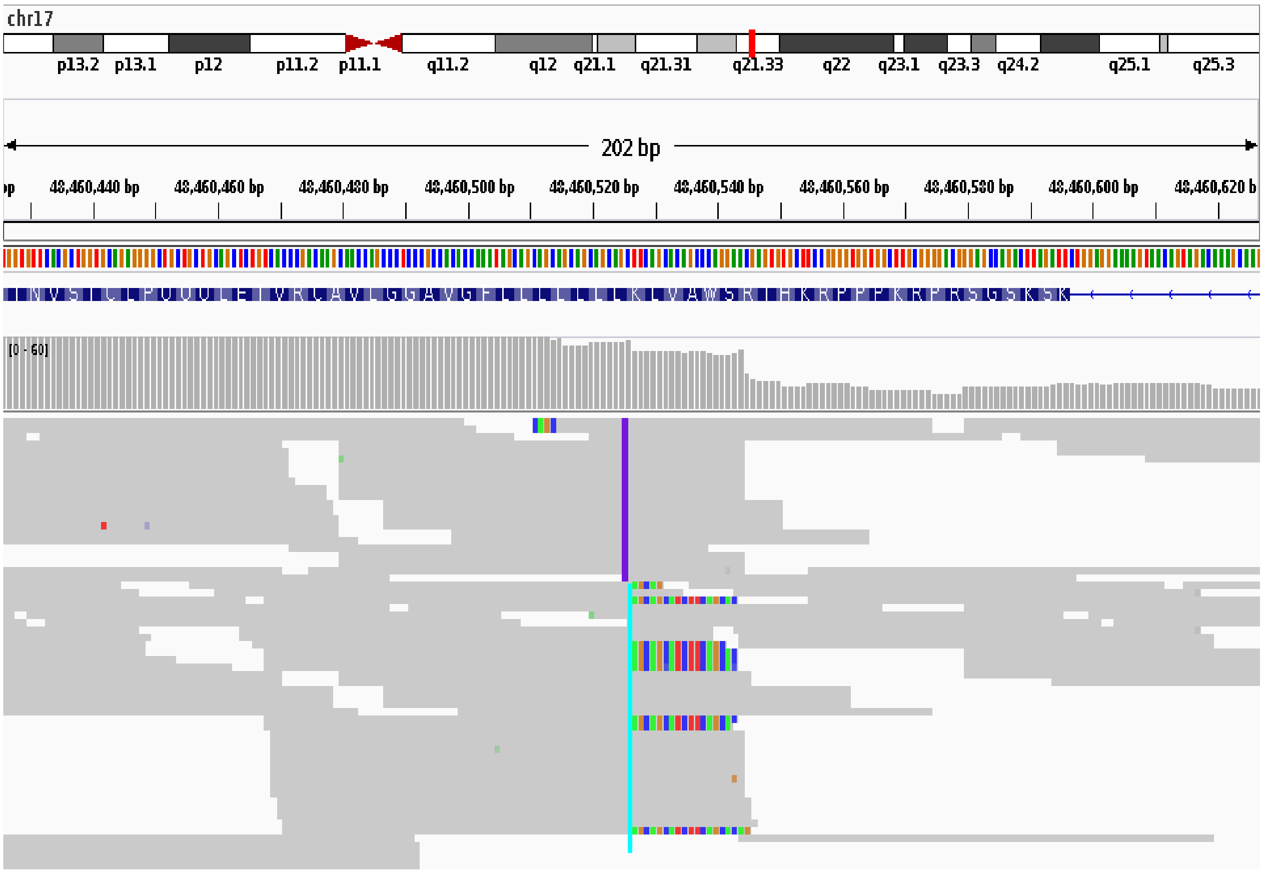
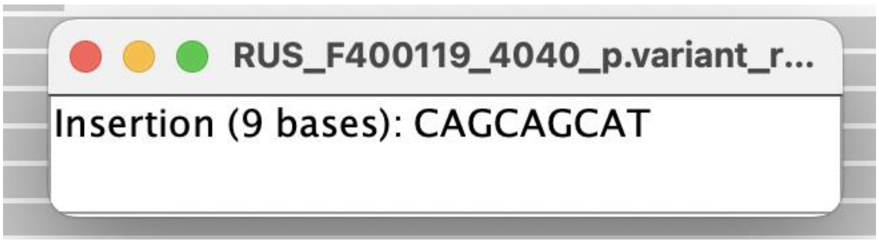


Figure 2.13. IGV snapshot showing evidence of a 9bp simple insertion (purple line) on chromosome 17. I have used the same-coloured vertical line in Figure 2.12 to

annotate the split positions of the single SR cluster at this loci. The cyan line indicates the position of SR Cluster 1. The purple line in the reads above the SR cluster indicate that the inserted sequence is nested within these reads. Clicking on this in the IGV interface will open a window specifying the size and specific base sequence of the insertion (Figure 2.14)



Insertion (9 bases): CAGCAGCAT

Figure 2.14: In the IGV interface, clicking on the purple line shown in Figure 2.13 opens this window, showing the size and constituent bases of the simple insertion based on the read's CIGAR string.

As I explained in section 3.4.4, InDelible can't assign simple insertion as an SV type, as their associated soft-clipped bases do not align to the reference genome (assuming no read mismapping), nor are they among the sequences included in the curated database of repeat sequences interrogated by BLAST. In cases where the inserted sequence is too long to be nested in reads, and where no bases in the adjacent reference sequence are among the soft-clipped bases in the SR cluster (or there are too few of them), it is not possible on the basis of an IGV snapshot to differentiate a simple insertion from a large (>100bp, so the other breakpoint is outside the snapshot window) deletion that failed alignment.

3.5.4 Small tandem repeats

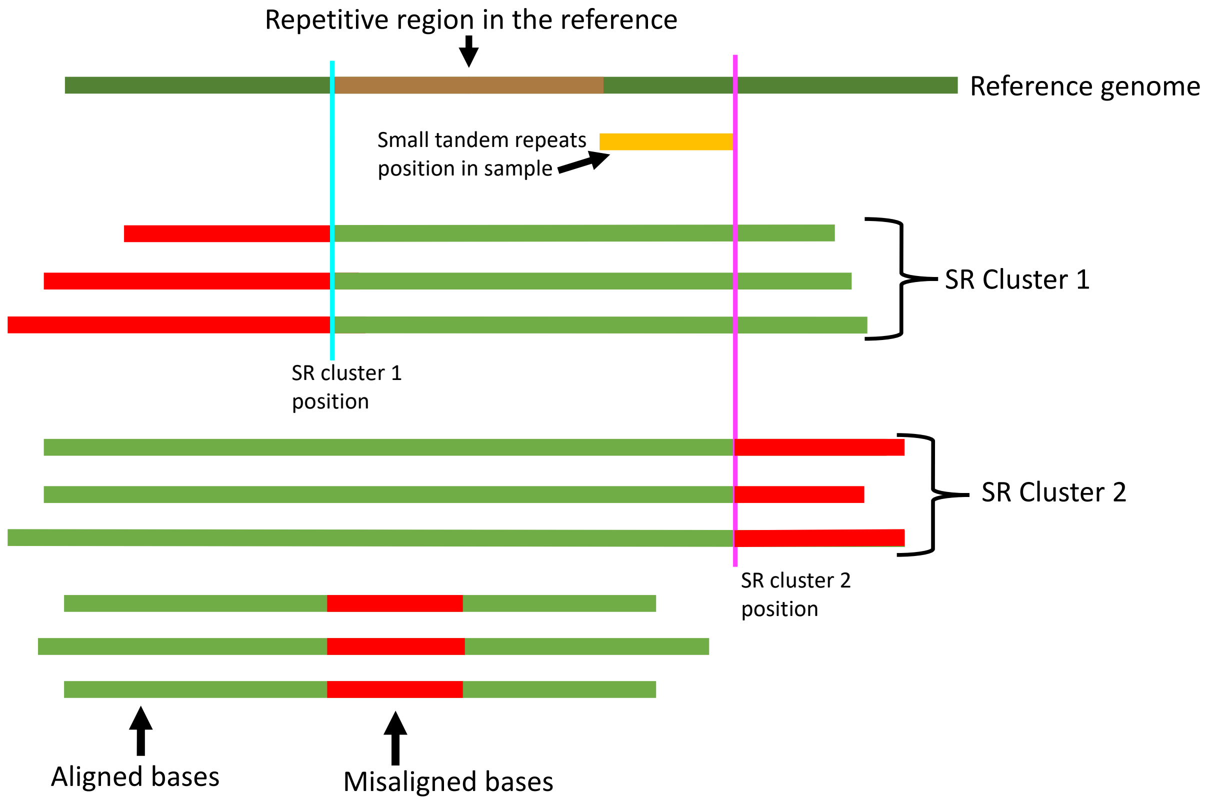
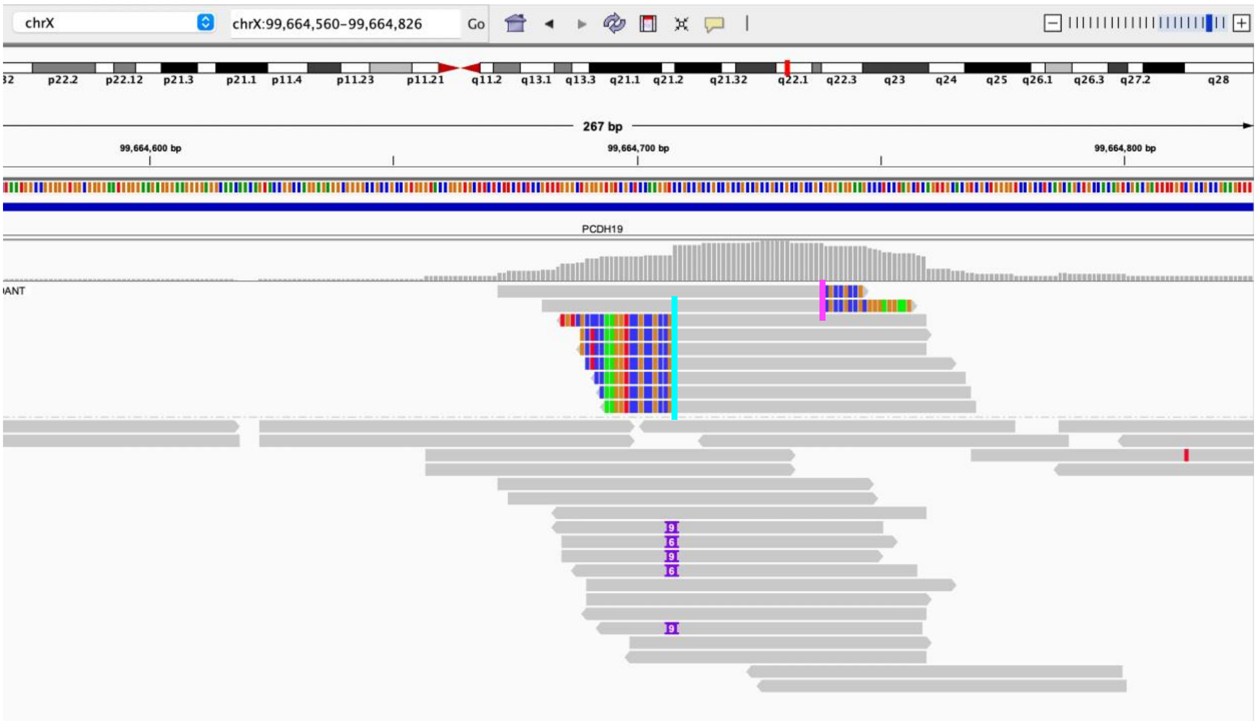


Figure 2.15. The split read pattern associated with a small tandem repeat. The reference genome is represented by the dark green bar. The gold bar shows the position of the small tandem repeat instances in the sample, while the brown bar shows the reference genome position of the repetitive sequence that has been extended in the sample. 33% more copies of the repetitive sequence occur in the sample. As in the case of simple insertions, the lower three reads are not split but have nested misaligned bases, corresponding to the small tandem repeat expansion.

Although small tandem repeats (STRs) are not usually defined as SV, I include them here because they can be detected through SR analysis in the same manner as SVs and are therefore called by InDelible. The SR pattern caused by STRs combines elements of patterns for insertions and duplications. There may be one SR Cluster or two (as in Figure 2.15) depending on how reads are mapped to the repetitive sequence in the reference genome. If the 3’ end of the read is aligned (SR Cluster 1), the sequence directly upstream of the repetitive region will be misaligned by the number of bases corresponding the extended length of the STR in the sample. If the 5’ end of the read is aligned (SR Cluster 2), the sequence directly downstream of the repetitive region will be misaligned by the number of bases corresponding the extended length of the STR in the sample. In both cases, the misaligned bases closest to the split position will correspond to the extension sequence. In some mappings, an STR is recognised as an insertion and the read will not be split, but with contain an insertion signifier in their CIGAR strings. Again, IGV will truncate the STR and replace it with a purple marker, which when selected will show the size of the event and its constituent bases. Figure 2.16 shows an example of this SR pattern in the IGV interface.



chrX

chrX:99,664,560-99,664,826

Figure 2.16: An IGV snapshot showing evidence of a 9bp small tandem repeat on chromosome X. I have used the same-coloured vertical lines in Figure X to annotate the split positions of the two SR clusters at this locus. The repetitive region in the reference genome consists of 9 instances of CCG. 3 further CCG instances occur in the sample genome. In the first SR cluster, the final 9 instances of the repetitive sequence are aligned to the reference, so there are 3 instances just prior to the split position (cyan line). In the second SR Cluster (which was filtered for having only two reads), the first 9 instances align, so there are 3 instances just after the split position (as least in one of the reads, the only contains 2 of the additional 3 CCG instances) (pink line). 5 reads impacted by the sequence are not split as the event was detected as an insertion by the aligner. However, in 2/5 reads the inserted sequence contains only 6 bases, corresponding to two CCG instances. This was likely caused by

replication errors during PCR.

InDelible does not assign small tandem repeats as an SV type, in part because the Database script does not differentiate repetitive soft-clipped bases from non-repetitive soft-clipped bases, but also because the orientation of the cluster split position to its alignment position is the same as that of duplications.

3.5.5 Translocations

Translocations (including segmental duplications) are indicated by the same single cluster SR pattern as insertions, only in this case the soft-clipped bases resolve to a single non-adjacent locus - a different region of the chromosome or a different chromosome. As such these events can only be differentiated from insertions in the InDelible output data if they have a BWA-mem alignment. However, a feature unique to translocations that can be observed in IGV is the discordant mapping of paired reads to non-adjacent loci, which allows for identification by manual inspection even if no BWA-mem alignment was found. Discordant mapping occurs when the first breakpoint of the translocation is in or just flanking the inner sequence between two read mates. If the breakpoint occurs upstream of inner sequence, the first discordant read will be split at its 3’ end. If it occurs downstream, the second discordant read will be split at its 5’ end. If the first breakpoint is upstream of or in the inner sequence and the second breakpoint occurs with the second read, the second read will be split at its 3’ end. However, if the first breakpoint occurs in the inner sequence, and the event is longer than the second read, neither read will be split (as illustrated by Figure 2.17). If the breakpoint occurs before the inner sequence in any of the discordant read pairs at each locus, the translocation event is at the locus to which the first read is mapped. If the breakpoint occurs after the inner sequence, the event is at the locus of the second read. In the highly improbable case in which the breakpoint occurs in the inner sequence of all the discordant reads, the locus of the translocation can still be determined in WES data, as it is highly unlikely that both reads in each pair will map to an exon.

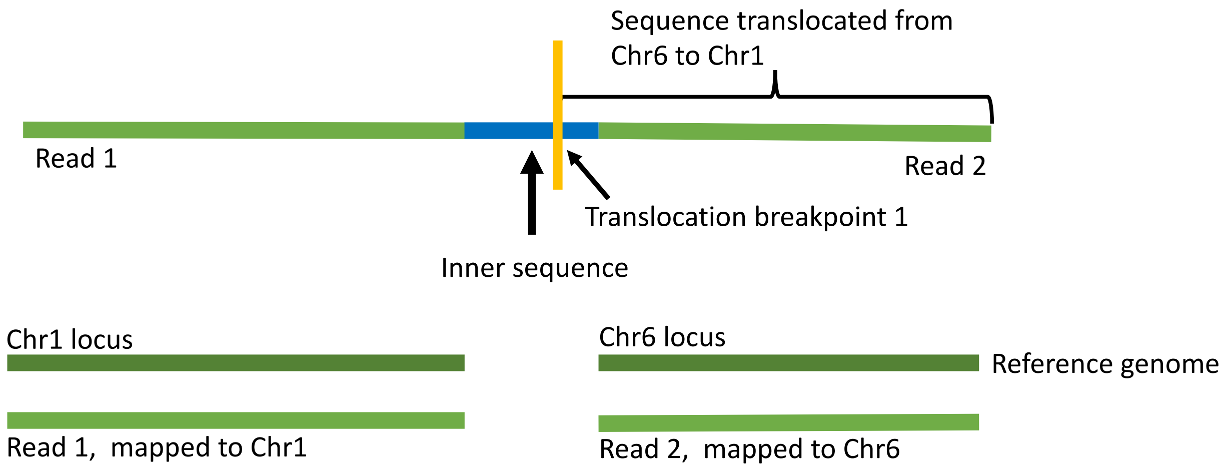


Figure 2.17. The discordant mapping of a read pair when the first breakpoint of a translocation (gold line) occurs in the inner sequence between the reads (blue bar). Both reads are sequenced from chromosome 1 DNA but read 2 consists of a sequence translocated from chromosome 6. As the translocated sequence is larger than the read itself, read 2 maps to chromosome 6 of the reference genome without splitting (bottom right). The same mapping would occur if the reads were sequenced from chromosome 6 and read 1 consisted of a translocated sequence from chromosome 1. However, as it is highly unlikely that both reads will map to an exon, the locus of the event can still be determined in WES data. Chr = chromosome

The precise size of translocated sequences is equal to the distance between the split position of the SR cluster at the first locus and the reference genome position to which the first misaligned base of the SRs in the second cluster aligns. InDelible does not calculate the size of these events. However it can be calculated from IGV snapshots there is a snapshot for each locus and the event is <100bp, allowing the soft-clipped bases at the second locus to be manually aligned to reference genome at the first locus.

In some cases, the translocated sequence is small enough to be nested within the combined length of a discordant read pair and their inner sequence, and neither the first nor second breakpoint occurs in the inner sequence, resulting in read mates that are both split and discordant. The size of the event then corresponds to the combined length of the soft-clipped bases in the first mare, the inner sequence, and aligned bases in the second mate, illustrated in Figure 2.18.

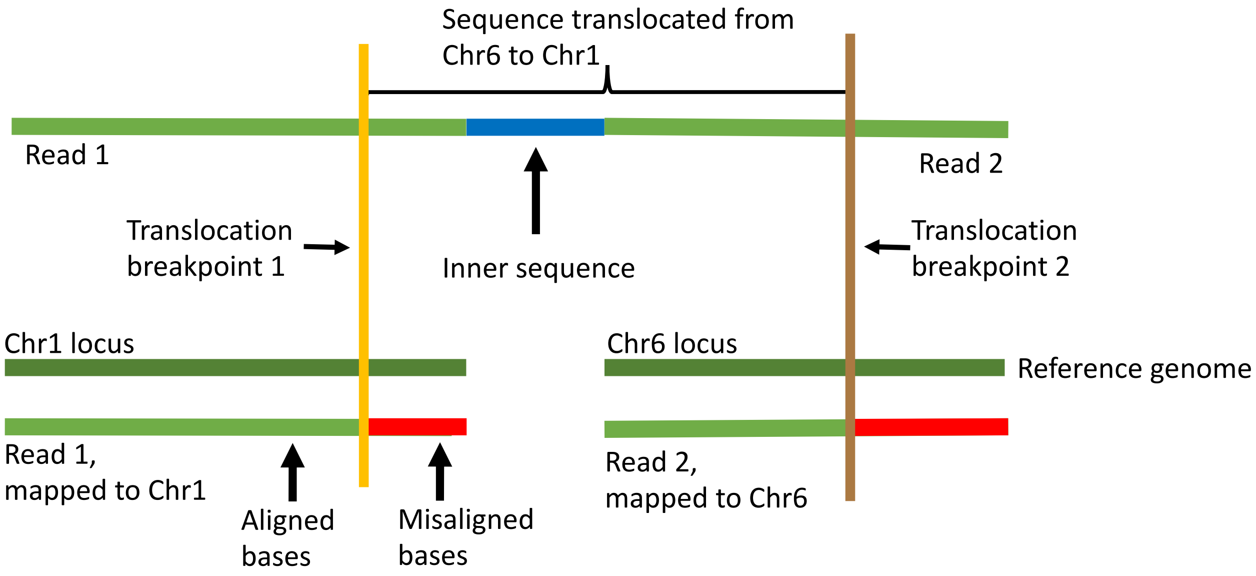


Figure 2.18. The discordant mapping of a read pair when the first breakpoint of a translocation (gold line) occurs in read 1 and the second breakpoint (brown line) occurs in read 2. As in Figure 2.17, both reads are from chromosome 1 DNA, but the translocated sequence originates in chromosome 6. In only this case will discordant reads both be split. The soft-clipped bases of read 1 will align to the reference position of the 5' end of translocated chromosome 6 sequence, while those of read 2 will align to the reference sequence in chromosome 1 that directly follows the second breakpoint. The size of the translocation is thus equal to (n soft-clipped bases in read 1) + (n bases in inner sequence) + (n aligned bases in read 2).

In IGV, discordant reads are colour coded according to the chromosome to which their mate is mapped. The colour for each chromosome is shown in Figure 2.19 below.

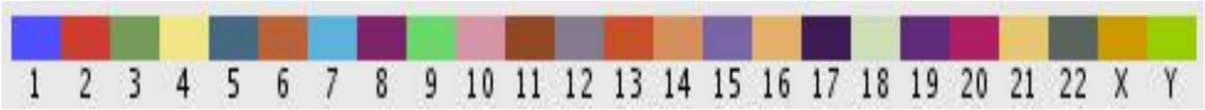

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Figure 2.19. IGV chromosome colour code for discordant reads. A discordant read will be coloured according to the chromosome to which its mate is mapped. E.g. the mate of a read mapped to chromosome 2 is mapped to chromosome 7, the read on chromosome 2 will be coloured light blue. If both reads are mapped to the same chromosome, they will have the same colour.

Figure 2.20 shows an example in IGV of a SR pattern in chromosome 6 that is indicative of a translocation at that locus. Colour coded, non-split and split discordant reads can also be observed. The soft-clipped bases of the SR cluster align to chromosome 11, which is expected given the dark brown colour of the adjacent discordant reads. Figure 2.21 shows the locus corresponding to the misaligned bases of the SRs in Figure 2.20, and the second reads in the discordant pairs. The translocation almost certainly occurs at chromosome 6 because the chromosome 11 locus does not intersect an exon.

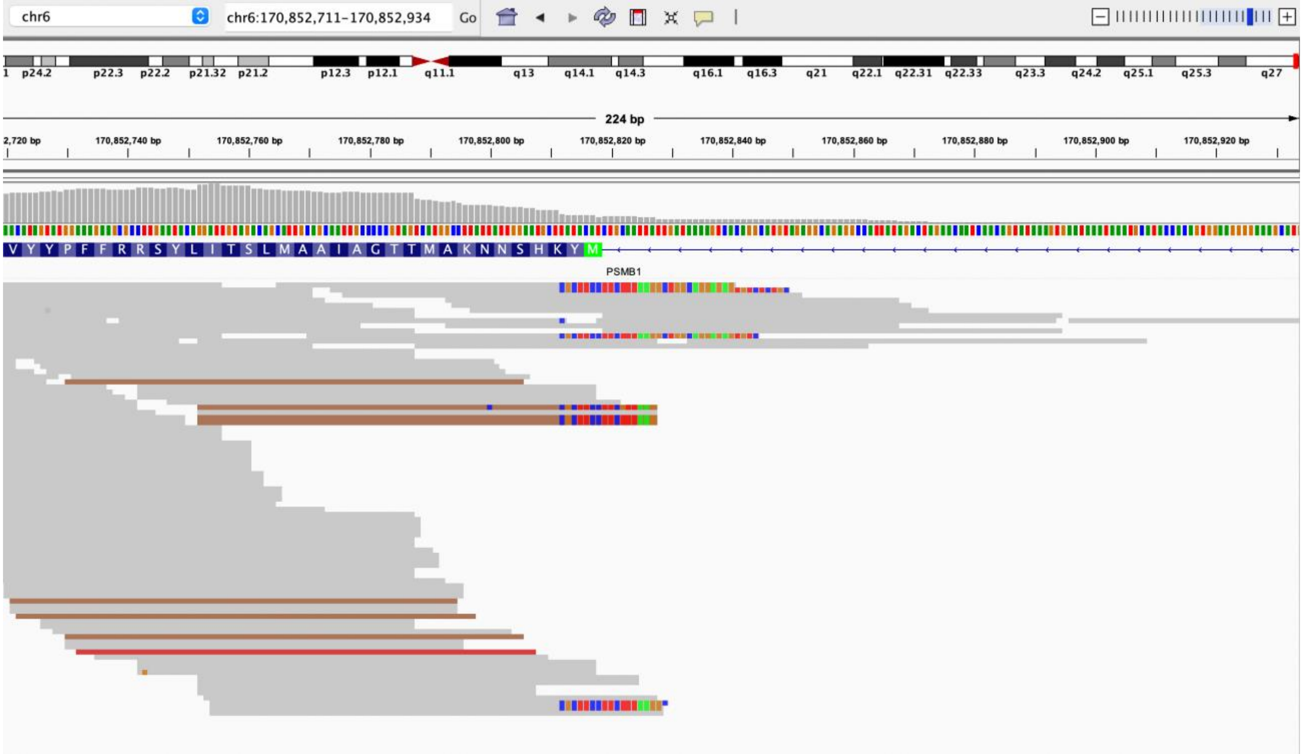
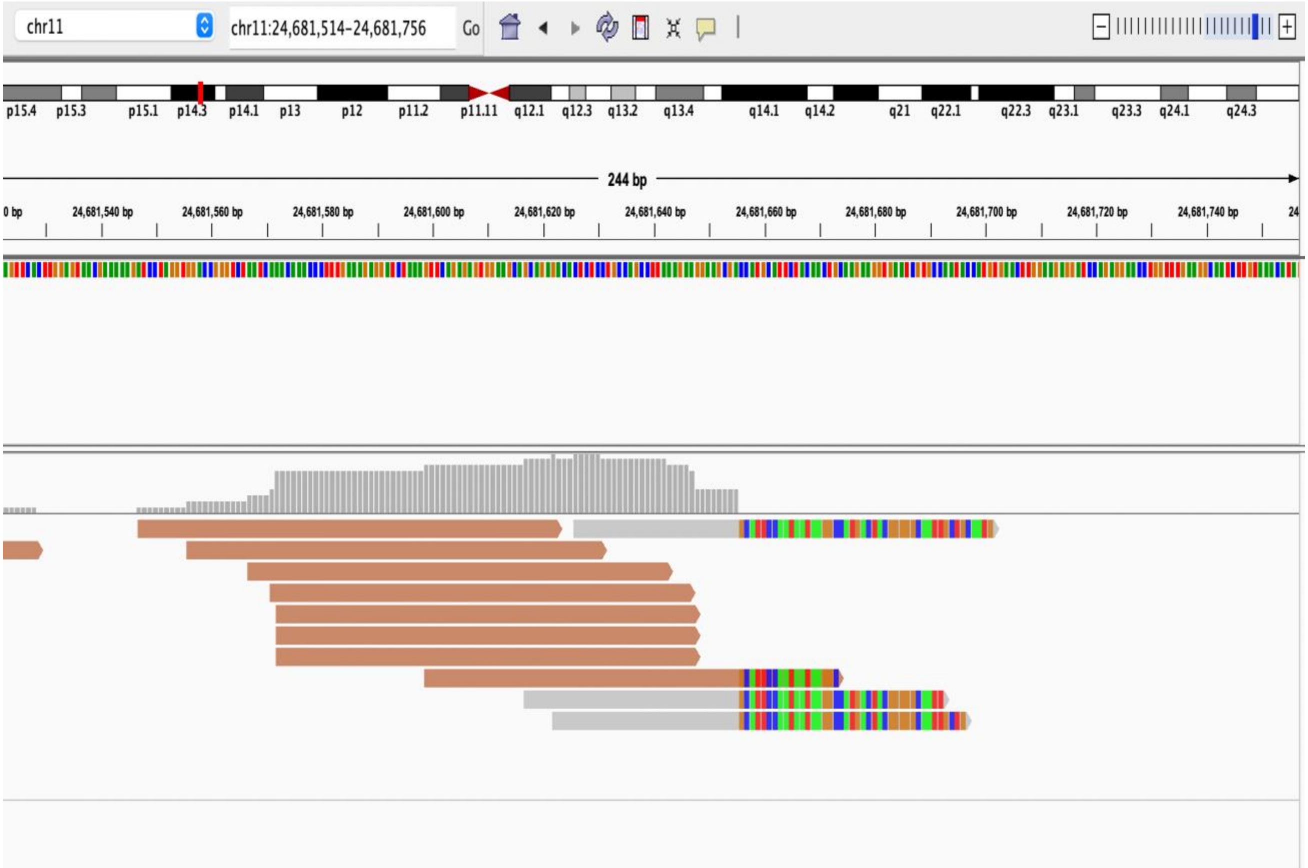

chr6
chr6:170,852,711-170,852,934
Go
1 p24.2 p22.3 p22.2 p21.32 p21.2 p12.3 p12.1 q11.1 q13 q14.1 q14.3 q16.1 q16.3 q21 q22.1 q22.31 q22.33 q23.3 q24.2 q25.1 q25.3 q27
224 bp
2,720 bp
170,852,740 bp
170,852,760 bp
170,852,780 bp
170,852,800 bp
170,852,820 bp
170,852,840 bp
170,852,860 bp
170,852,880 bp
170,852,900 bp
170,852,920 bp
V Y Y P F F R R S Y L I T S L M A A I A G T T M A K N N S H K Y M
PSMB1

Figure 2.20. IGV snapshot showing evidence of a translocation on chromosome 6. The soft-clipped bases of the single SR cluster align to a locus in chromosome 11. The dark brown line represent discordant reads, coloured as such because their mates are mapped to the same chromosome 11 locus. As two of the discordant reads are split at their 3’ ends, the translocation breakpoint 1 is at the position of the split and not at the locus to which both their mates are mapped.


chr11

chr11:24,681,514-24,681,756

Figure 2.21. IGV snapshot of discordant reads mapped to chromosome 11 that are the mates of those shown mapped to chromosome 6 in Figure 2.20. The reads are coloured light brown as that is the colour-code for chromosome 6. One of the discordant reads is split, indicating that its sequence contains the second breakpoint of the translocation. None of the reads are mapped to an exon, increasing confidence that the event is at the chromosome 6 locus. The misaligned bases in the SR Cluster correspond to the sequence directly downstream of the second breakpoint, starting at 6:170,852,752. As the split position of the SR Cluster in Figure X is 6:170,852,764, the size of the event is 12bp (170,852,764 - 170,852,752).

3.3.6 Retrotranspositions

These events can be considered a subtype of translocation that are caused by the insertion of mobile elements (MEIs) by the types of retrotransposon machinery described in section 2.1.5 of chapter 1. Small MEIs are therefore indicated by the same SR pattern as insertions but can be identified in the InDelible output by their assigned type, based on the alignment of their soft-clipped bases to a sequence in the repeat sequence database mined by BLAST. Typically, BLAST alignment of their associated SRs will have >10 genome-wide hits. Figure 2.22 shows the SR pattern associated with a SINE element that has been inserted into chromosome 17. As MEIs are a type of translocation, discordant read pairs are adjacent to the SR Cluster, whose mates map to several other genomic loci where the SINE element has been inserted. In this case the discordant reads cannot be used to determine the source loci of the retrotransposition event, and so its precise size cannot be calculated by manual inspection.

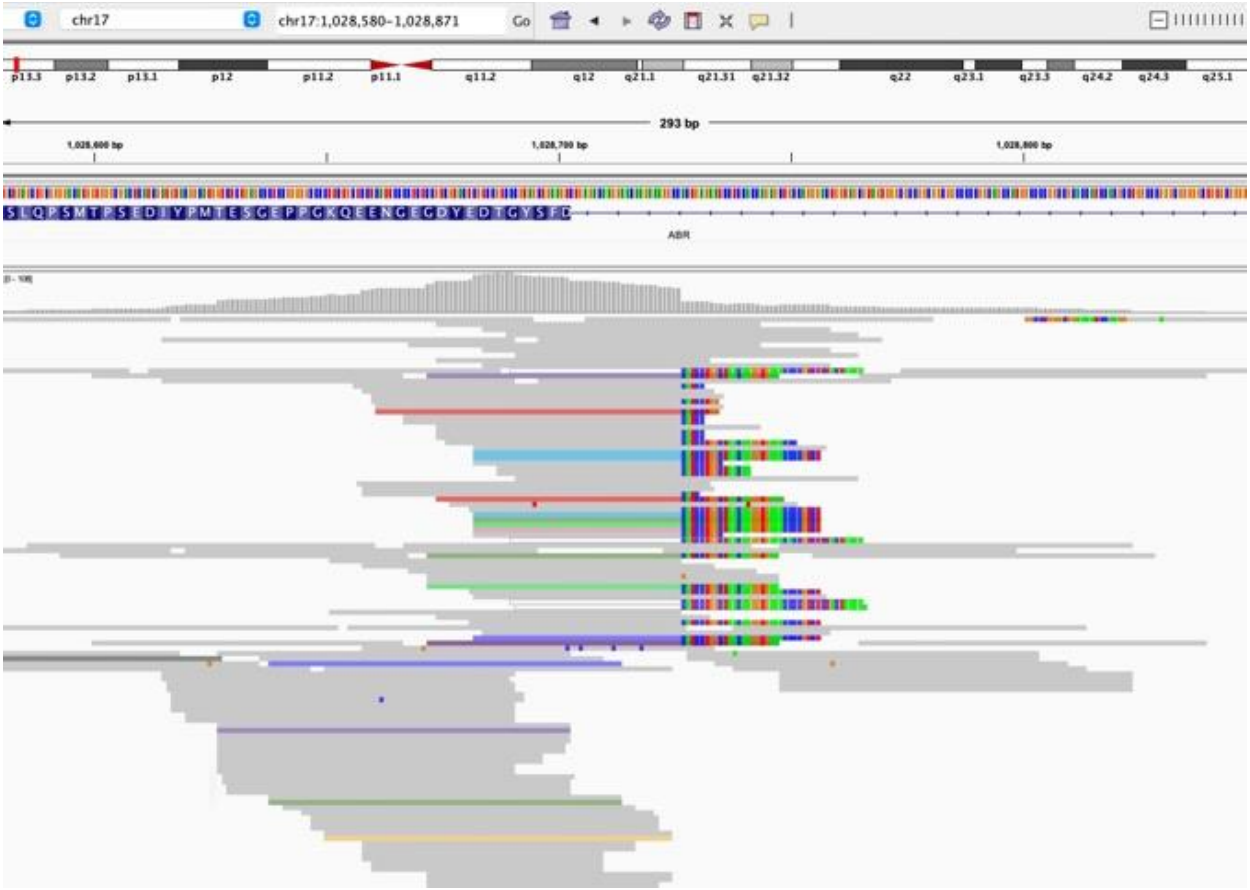


Figure 2.22. IGV snapshot showing evidence of a SINE retrotransposition on chromosome 17. The misaligned bases of the single SR cluster are the 5’ end of the SINE. As this element occurs throughout the genome, discordant reads in this locus have mates which are mapped to 9 different chromosomes: 1, 2, 3, 4, 7, 8, 9, 10 and 15.

A subtype of retrotransposition involves the partial transcription and translocation of successive exons in a gene by the LINE-1 retrotransposon machinery, described in section 2.1.5 of chapter 1. This creates an additional copy of the affected exons called a pseudogene. Reads that are generated from a pseudogene map to the reference at the loci of their source exons. The SR pattern associated with these events is very specific, illustrated in Figure 2.23.

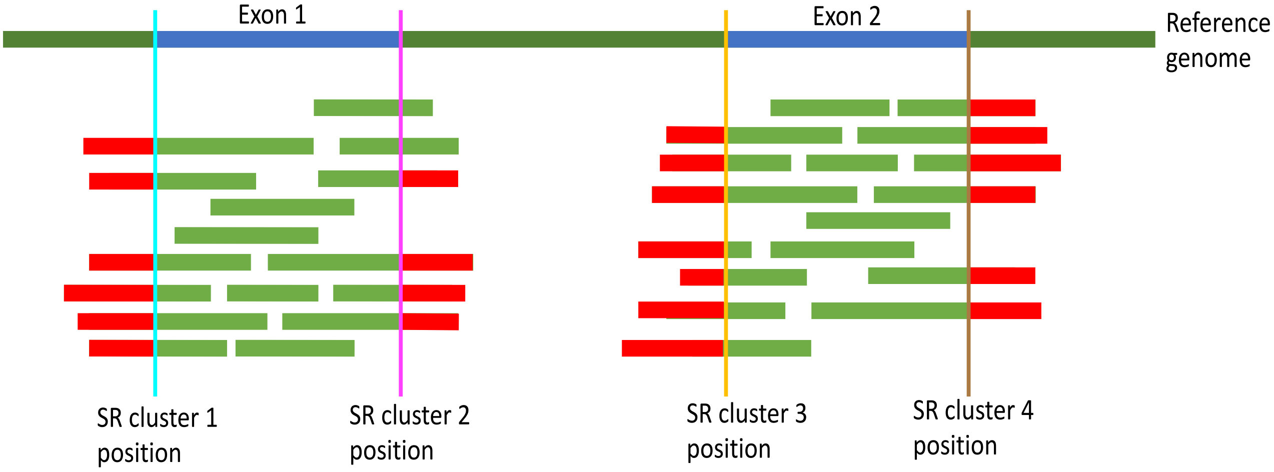


Figure 2.23. The split read pattern associated with a pseudogene retrotransposition spanning 2 exons. The reference genome is represented by the dark green bar, within which the exons are coloured in blue. There are 4 split read clusters, 2 per exon, whose split positions align precisely to the 5' and 3' exon junctions.

As an exon in a pseudogene does not contain any intronic DNA, its associated SRs split at the 5' and 3' junctions of its source exon locus in the reference genome. The soft-clipped bases at the 5' junction will either align to the 3' junction of the upstream exon, or possibly a 5' cap if it is the first exon in the gene in the former case, SRs will also occur at the 3' junction the of the upstream exon. These events can implicate any number of exons in a gene, depending on when the LINE-1 machinery 'hijacks' the transcription process. Figure 2.24 is an IGV snapshot shows SRs associated with this type of retrotransposition at the 5' and 3' junctions of a single exon.

InDelible can only assign this SV type if the retrotransposed pseudogene known to repeat throughout the genome and is therefore included in the curated repeat sequence database. Otherwise, the InDelible is likely to misclassify clusters associated with this event type as deletions that span intronic regions, as the orientation of the split position to alignment position is the same as that of deletions. However, they are straightforward to classify in IGV, given that the split position always align precisely to exon junctions. The size of an event corresponds the combined length of all the exons to which the SR Clusters are mapped. It should be noted that that is no way to determine actual location of an event in their source DNA based on split read data alone. Its gene annotations in the InDelible output are therefore invalid.

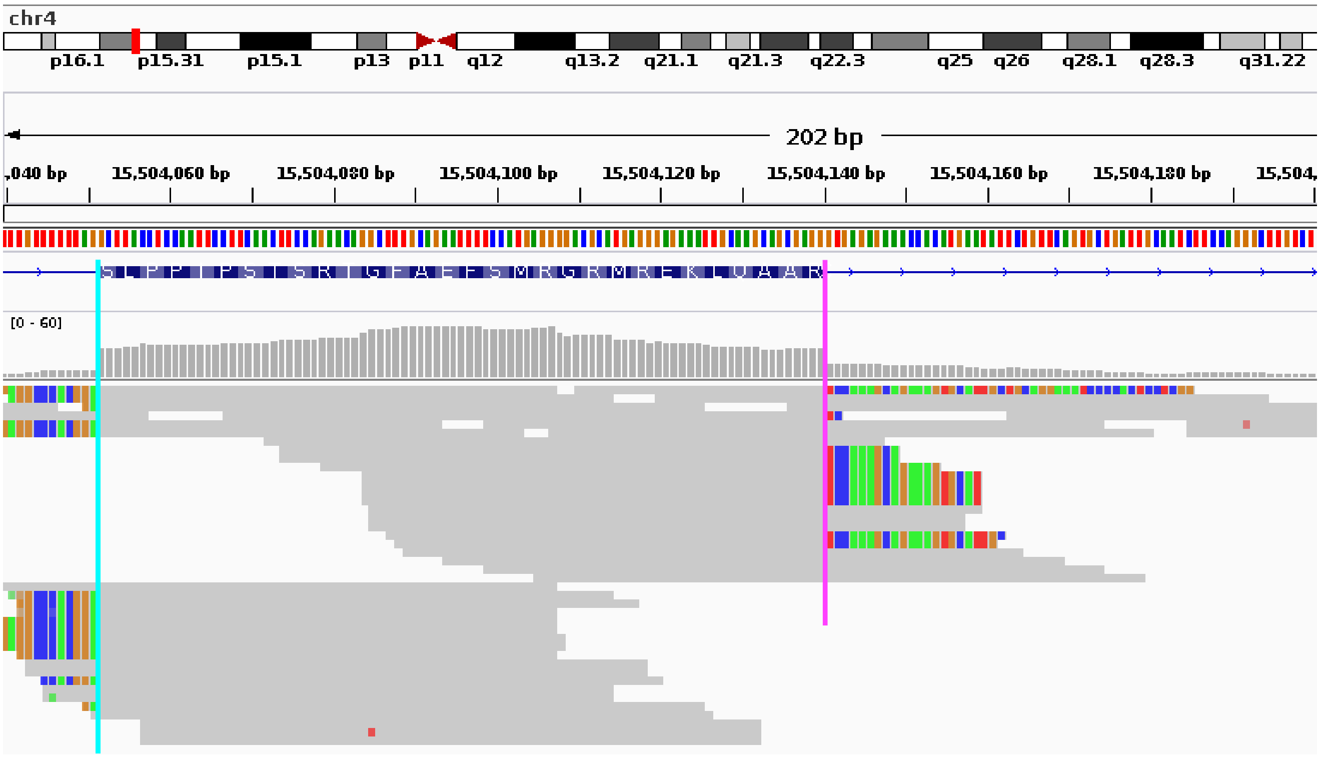


Figure 2.24. IGV snapshot showing evidence of the retrotransposition of a pseudogene based on chromosome 4 gene *CC2D2A*. The split positions of the 2 SR clusters align precisely to the 5’ junction (cyan line) and 3’ junction (pink line) of the exon at this locus. As this is not the first exon implicated in this event, the soft-clipped bases in the 5’ SR cluster align to the 3’ end of the next upstream exon. The soft-clipped bases in the 3’ SR cluster align to the 5’ end of the next downstream exon.

3.3.7 Complex events

Some SV events involve a combination of types and are therefore indicated by SR patterns corresponding to different SV types occurring at the same locus. For example, one copy of a duplicated sequence may contain an insertion, in which case there will be 3 SR clusters associated with this SV: 2 for the duplication breakpoints, and 1 for the 5’ insertion breakpoint (Figure 2.25).

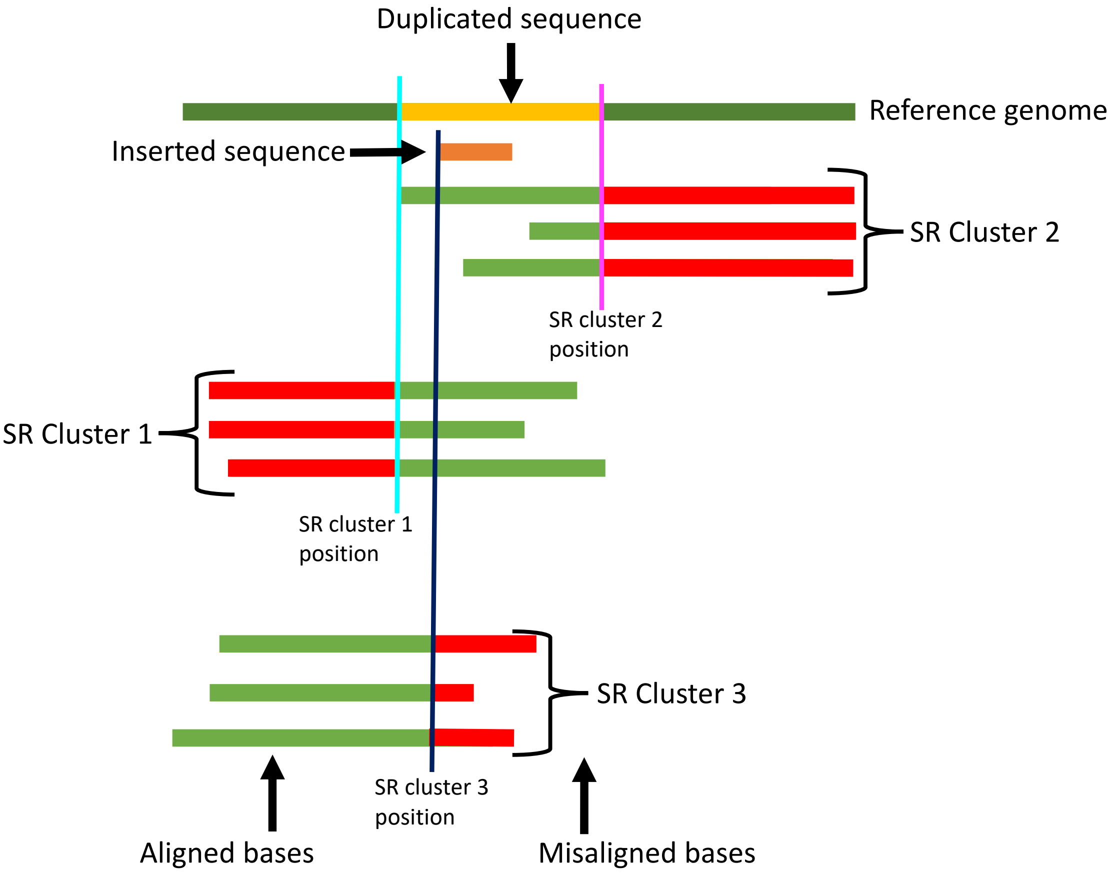


Figure 2.25. The split read pattern associated with a complex-duplication/insertion. The reference genome is represented by the dark green bar, within which the duplicated sequence is coloured in gold. The position of the inserted sequence within the duplicated sequence, relative to the reference genome, is shown by the orange bar. In this example, the insertion only occurs within the first copy of the duplicated sequence, and so would only be visible in the soft-clipped bases of SR Cluster 1. In the aligned bases of SR Cluster 2, the inserted bases would be evident in their CIGAR strings, and a purple maker at the 5' insertion breakpoint in IGV. SR Cluster 3 is produced from reads who's 3' bases partially align to the first copy of the duplicated sequence, but then misalign due to the insertion. As their 3' ends either do not extend beyond the length of the insertion, or do not extend far enough beyond it for the reference bases that occur after the insertion to be aligned, their misaligned bases are soft-clipped.

InDelible assigns this event type by assessing the insertion contexts of soft-clipped bases (produced by BWA-mem, in the same way BWA generates CIGAR strings from non-synthetic FASTQ reads) to which it has already assigned the deletion or duplication type. Size is calculated in the same way as non-complex events, with the added step of adding the length of the insertion context in the case of complex duplications and subtracting it in the case of complex deletions. Size can be determined from IGV by counting the distance between the two SR clusters and adding or subtracting (depending on type) the insertion length revealed by reads in which the insertion is nested (indicated by the purple marker). Figure 2.26 is an IGV snapshot showing the SR patterns that indicate a complex-duplication/insertion.

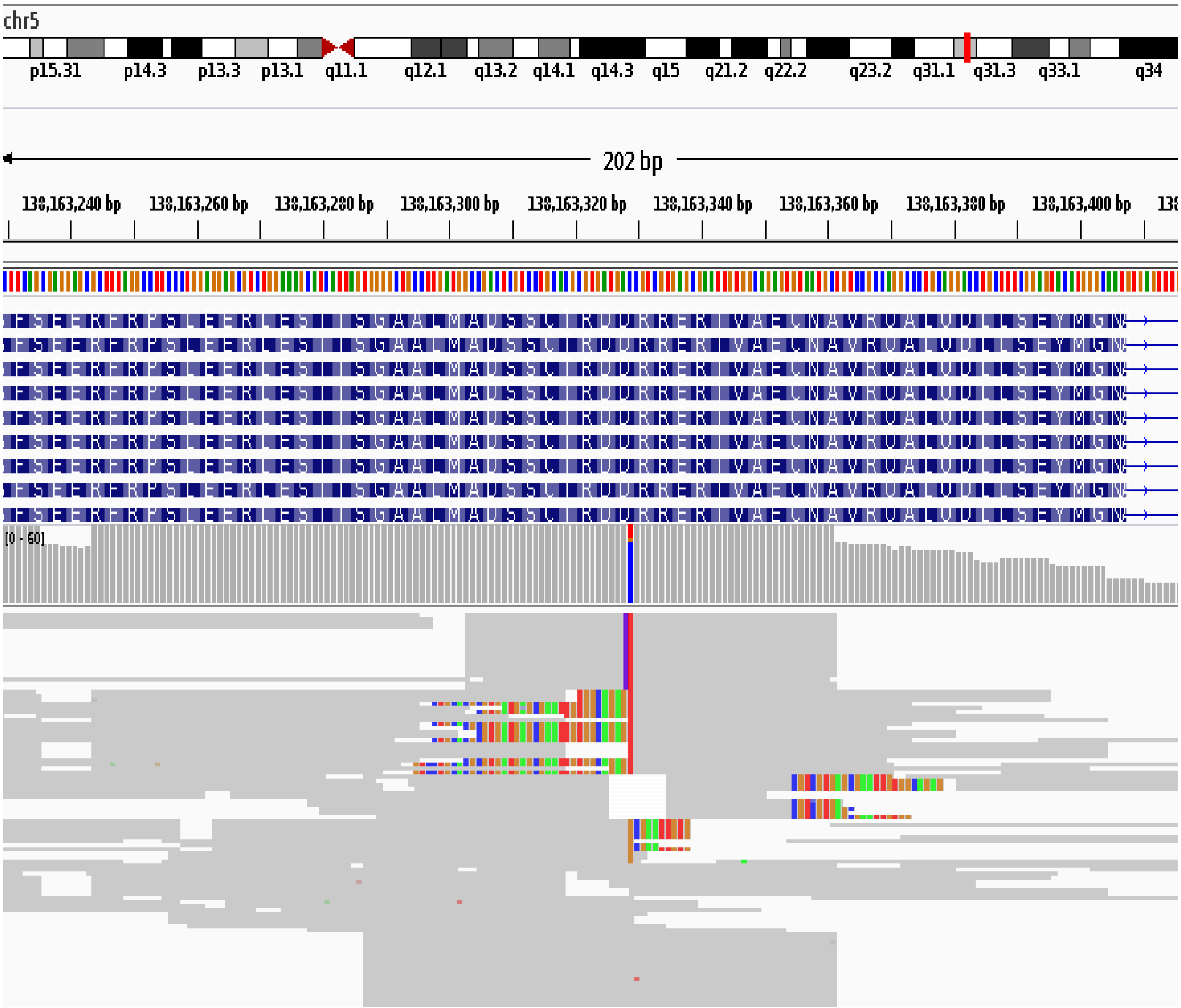


Figure 2.26. IGV snapshot showing evidence of a 32bp complex-duplication/insertion on chromosome 5. The insertion is 3bp and is nested at the 5' end of the first copy of the 29bp duplicated sequence. The lowermost SR cluster is caused by reads that contain the insertion at their 3' ends, but don't extend far enough beyond the length of the insertion for their subsequent reference bases to aligned. The length of the complex event can be calculated in IGV interface by counting the distance between

the outer most SR clusters, then adding the length of the inserted bases, revealed by the purple marker at the 5’ insertion breakpoint.

3.4 Published DDD InDelible study

I now describe the study carried out the developers of InDelible (Garder et al. 2021), which was used to test and benchmark the algorithm, as well as identify SVs contributing to developmental delay. As InDelible was primarily developed to detect SVs of clinical significance that were missed by existing array and WES-based methods, the purpose of the DDD InDelible study was to demonstrate that it can detect putative developmental disorder risk variants in a clinical cohort. However, identified variants were also compared with those previously identified by other methods for validation purposes, and the algorithm was benchmarked against alternative SR-based methods using a well-characterised truth set.

3.4.1 Call validation

13,438 probands with severe developmental disorders were analysed, of which 9,848 has WES data available for both parents. A proband and both parents are collectively referred to as a ‘trio’. To ascertain variants that were likely to be of clinical significance, calls in both trio and non-trio probands were restricted to those with allele frequency < 0.004, and which intersect 399 dominant or X-linked genes that are associated with developmental disorders in the Developmental Disorders Genotype-to-Phenotype database (DDG2P). Calls in trio probands were also excluded if either parent sample had > 2 SRs at the same position, indicating transmission. These criteria gave a preliminary set of 260 candidate DD-risk SVs across all probands. By identifying which of these variants were previously called by other methods, the authors assessed the sensitivity of InDelible to known variants of different size ranges (Figure 2.28). They found that InDelible was most sensitive to known candidate variants in the 21-50bp size range, of which it detected 48.3%. The algorithm was highly insensitive to variants less than 10bp in size (<10% detection rate), and variants larger than 50bp (< 5% detection rate). Among 63/260 novel candidate SVs, 45 (71.4%) were within the 11-20bp and 21-50bp size ranges. 11 (17.5%) were >10bp and 7 (11.1%) were >50bp (Figure 2.28).

InDelible’s sensitivity to such a narrow size range of small variants is likely due to the inverse correlation between allele frequency and the occurrence of an SV breakpoint within an exon. Although the expected frequency of the class of SVs detected by InDelible has not been studied, the frequency of SVs in both clinical and population data has been shown to decrease with size (Sudment et al., 2015). Therefore, a size increases, it is less likely that the breakpoint of an SV affecting a coding region will fall within an exon, and therefore be detectable using split read information.

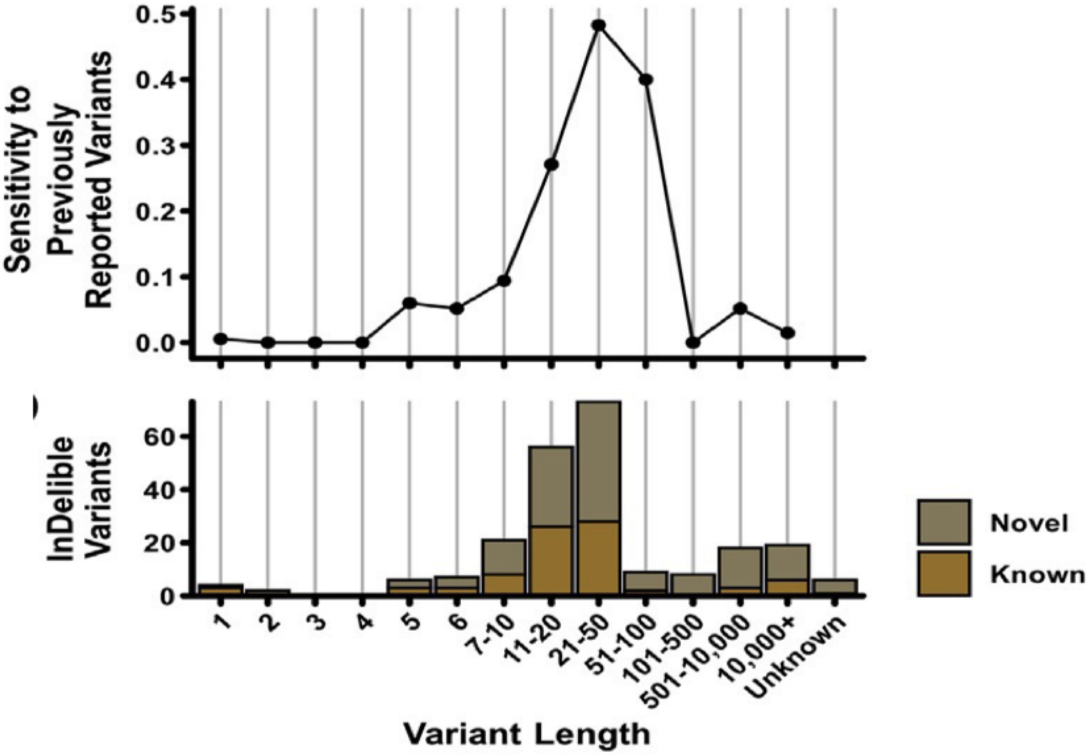


Figure 2.28. The upper plot shows the sensitivity of InDelible to candidate risk SVs previously discovered the Deciphering Developmental Disorders probands, by variant length. The lower plot specifies the number candidate risk SVs discovered by InDelible for each variant length range. These are further subset by novel (brown) and known (orange) variants. Adapted from (Gardner et al. 2021)

Also reported in the paper is the breakdown of the 63 novel candidate DD-risk variants according to their SV type (Figure 2.29). There were 32 (50.8%) deletions, 18 (34%) duplications, 1 (1.6%) simple insertion, 8 (12.7%) complex-insertions/duplications, 3 (4.8%) translocation and 1 (1.6%) *Alu* MEI retrotransposition. I use these type and size reports in the relevant results chapters to assess whether my findings are broadly commensurate with those of the DDD study. For reasons given in the chapters, a robust statistical comparison of results was not appropriate, though I was still able to check that my results contained similar SV types and a had a similar size distribution.

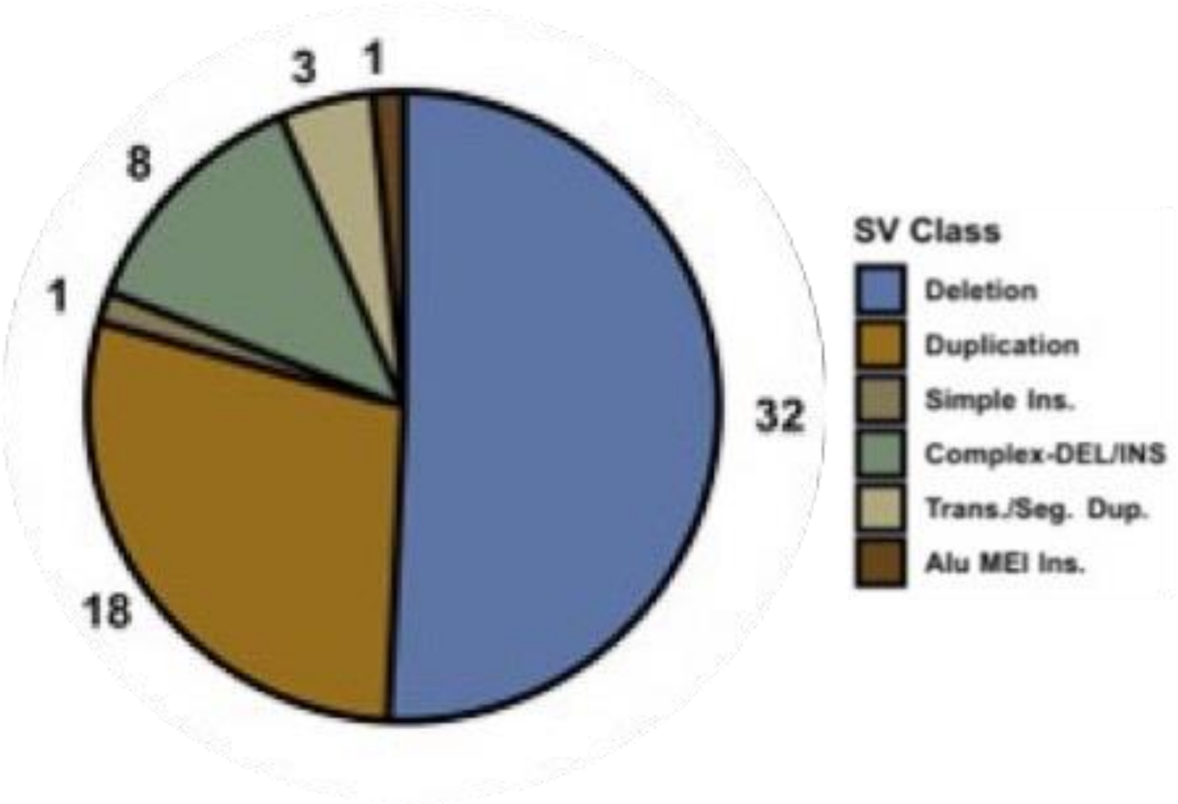


Figure 2.29: Variant types of the 63 novel candidate developmental disorder-risk structural variants detected in the InDelible Deciphering Developmental Disorder study. Figure adapted from (Gardner et al. 2021)

3.4.2 Benchmarking

InDelible was benchmarked against two alternative methods that also use SR information in WES data for variation calling: GATK (Collins et al., 2020) and Manta (Chen et al., 2016). All three callers were applied to a well-characterised control sample produced by the Genome in a Bottle Consortium (GIAB) (Zook et al., 2016). Results were then compared against a gold-standard indel and SV call set produced by GIAB for the same sample, to assess sensitivity and false discovery rates for each algorithm. InDelible was found to equal or exceed the sensitivity of both alternative methods for variants between 21 and 10kb in length (Figure 2.30, A). Assessing sensitivity by SV type, it detected 81.7% more >20bp deletions than GATK, and >15% more than Manta. It also detected 86.9% and 8.2% more duplications and insertions >20 bp in length than each alternative method, respectively. InDelible had lower false discovery rate compared with Manta (Figure 2.30, B), but an increased rate compared with GATK, which the developers attribute to the fact that InDelible is designed (primarily due to the training data applied to its adaptive learning model) to be maximally sensitive to rare, clinically significant variants as opposed to population-level variants. They also note that their analyses likely ‘drastically overstate’ the false positive rates of all callers, as the GIAB gold-standard calls were largely based on long-read sequencing data, which may be more difficult to call using the short-read based methods that query SR information.

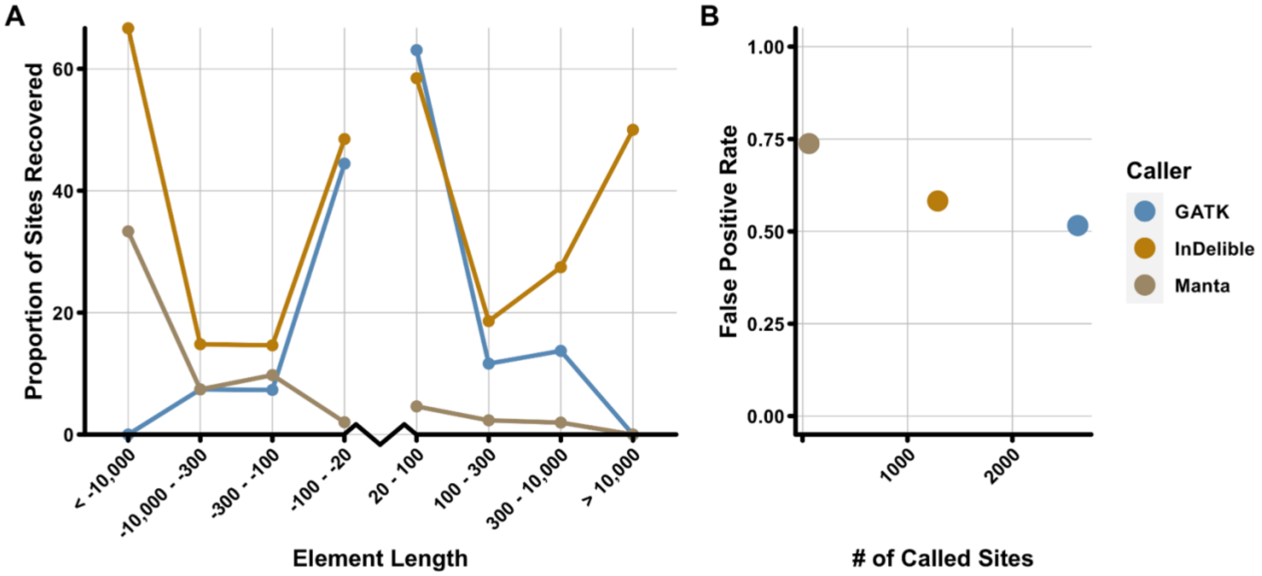


Figure 2.30. Sensitivity and false positive rates of InDelible, GATK and Manta. The methods were benchmarked using a well-characterised control sample produced by the Genome in a Bottle Consortium. A) Sensitivity of the callers to > 20bp deletions of different size ranges is shown on the right side of the plot, while sensitivity to >20bp duplications/insertions is shown on the left. (B) False positive rate for the callers, relative to the total number of sites called by each.

3.5 InDelible summary

In this section, I have described the InDelible algorithm, and the variant types it is designed to detect. I've described also described the validation and benchmarking analyses undertaken by the InDelible authors and reported in the methods paper. Based on comparison with a set of known, putative clinical variants called in the DDD study, the algorithm was found to be most sensitive to variants 11-50bp in length. Peak sensitivity was 0.5, for variants 21-50bp. While this is quite low (compared e.g., to CLAMMS sensitivity to CNVs), it is better than the sensitivity of alternative methods that use the same calling approach and is a size range that is entirely outside the discovery resolution of array-based methods. This is important as Aim 3 of my thesis is to compare the utility of array and WES-based approaches for detecting all types of SV, not just those that can be detected using both platforms. Moreover, InDelible is designed to be maximally sensitive to variants of clinical significance, which is also crucial for my purposes as both Aim 1 and Aim 2 involve assessment of variants for schizophrenia-risk status, and Aim 3 the association of variants with cognition, a clinically significant phenotype.

4. Chapter summary

In this chapter, I have described the algorithms of the of two SV callers I used in my PhD research: CLAMMS and InDelible. These descriptions describe the necessary background to understand which elements of BAM data each caller utilises, and therefore the differences between their call outputs. CLAMMS leverages coverage depth across exon windows to estimate copy number status and is therefore designed to detect both heterozygous and homozygous CNVs at resolutions of single exons (~150bp) and greater. On the other hand, InDelible interrogates aligned reads for split read information, and then uses this information to determine 6 SV types (deletions, duplications, translocations, MEI retrotranspositions, pseudogene retrotranspositions and complex insertions) and breakpoints, but not copy number. It is also most sensitive to variants at resolutions <100bp, and largely insensitive to variants >500bp. Taken together, therefore, the SV callers can detect CNVs across the entire size spectrum. I've also described the validation and benchmarking efforts undertaken by their respective developers, justifying my choice to use these algorithms over similar alternative.

Chapter 3: Using read coverage depth in whole exome sequencing data to detect *de novo* CNVs in schizophrenia

1. Introduction

1.1 *De novo* CNVs in schizophrenia

De novo variants are yet to undergo selection pressure and are therefore more likely to be deleterious than transmitted variants (Acuna-Hidalgo et al., 2016; Rees et al., 2011). Given the strong association between schizophrenia (SCZ) and reduced fecundity (Power et al., 2013), it has been hypothesised that rare (<1% frequency) *de novo* copy number variants (CNVs) may play a role in SCZ aetiology, insofar as disease-causing variants have a reduced likelihood of transmission (Rees et al., 2012). Multiple studies have confirmed this hypothesis by investigating the incidence of rare *de novo* CNVs in SCZ parent-proband trios.

Analysing a sample of 662 Bulgarian proband-parent trios, (Kirov et al., 2012) identified 32 rare *de nov*o CNVs in 32 cases, 8 of which occurred at four known SCZ risk loci: 1 deletion at 3q29, 4 deletions at 15q11.2, 2 deletions at 15q13.3 and 1 duplication at 16p11.2. The median size of *de novo* CNVs was 321kb. The authors found that these SCZ probands had a higher incidence of *de novo* CNVs (5.1%) than two sets of unaffected controls: an Icelandic cohort comprising 2623 individuals (2.2%, p=0.00015) and 872 unaffected siblings from a large family autism study (1.6%, p=0.00008). The difference in *de novo* incidence between control cohorts was found to be non-significant (p = 0.28).

In 177 cases of largely white European ancestry (1 was Hispanic and 1 African-American), (Malhotra et al., 2011) identified 9 rare *de novo* CNVs in 8 cases, with median size 348kb. Though none occur in known SCZ risk loci (1 duplication does occur in 22q11.2 locus, but only deletions at this locus have been found to confer disease risk (Rees et al., 2014)), the authors also compared *de novo* CNV incidence with 426 unaffected controls from the Simons Simplex Collection, again finding that rate of mutation is significantly greater among SCZ cases (4.5% vs 0.9%, p = 0.007). (Xu et al., 2012) analysed 359 proband-parent SCZ trios recruited from the genetically homogeneous Afrikaner population in South Africa. All probands had no history of SCZ among first- or second-degree relatives, so are referred to as 'sporadic' cases. The authors identified 17 rare *de novo* CNVs in 15 cases, 5 of which occur at known SCZ loci: 3 deletions at 22q11.2 and 2 duplications at 16p13.2. Comparing the incidence of *de novo* CNVs with 159 unaffected individuals recruited from the same population, the authors found that these events are ~8 times more frequent in sporadic cases ($p = 7.8 \times 10^{-5}$). Thus, there is strong evidence in existing literature that *de novo* CNVs are implicated in schizophrenia.

1.2 CNV detection methods

Previous studies aiming at detecting *de novo* CNVs in schizophrenia used CNV detection methods based on data generated from genotyping arrays with the number of probes ranging from 900K-2.1M. While such methods have been successful for the detection of large CNVs (>100kb), they are less accurate in detecting smaller events, as it becomes increasingly difficult to separate probe signal from noise at smaller resolutions. It has been estimated that high-density (>1M probes) array platforms are unable to reliably detect CNVs <40kb in size (Carter, 2007). Arrays that have been designed to capture a high density of SNPs in CNV prone regions may improve discovery resolution for some smaller CNVs (Haraksingh et al., 2017), but these have not been used in large *de novo* CNV studies of SCZ. More recently, however, whole exome sequencing (WES) technology has successfully detected CNVs missed by genotyping arrays in protein-coding regions (Zhao et al., 2013). WES can theoretically detect copy number changes that affect single exons (~ 150bp), i.e., much smaller than the resolution provided by SNP arrays (Marchuk et al., 2018). CNV detection methods using WES are thus a promising complement, or even alternative, to array-based methods.

1.3 Study aims

In this chapter I present research completed in the first year of my PhD, in which I explored the utility of CLAMMS for identifying small (<100kb) and rare *de novo* CNVs the WES data for a SCZ trios cohort comprised of 616 probands and their parents. Rare *de novo* CNVs had been called previously in this sample using array-based methods, so I was able to compare the discovery resolution of the sequencing method with that provided by genotyping microarrays. I also compared the size distribution of these calls and assessed the newly identified *de novo* events for possible SCZ risk.

2. Methods

2.1 Trios sample description

The 616 trios that were included in this study were comprised of probands and parents recruited by mental health professionals from inpatient and outpatient psychiatric facilities in 7 European countries: Bulgaria, Italy, Germany, Netherlands, Russia, Spain, and the UK. Probands were diagnosed with schizophrenia or schizoaffective disorder according to DSM-IV or ICD-10 criteria. The number of individuals diagnosed with each disorder were unavailable. Individuals were excluded if they had previously received a diagnosis of a neurodevelopmental disorder or intellectual disability. DNA was obtained from peripheral blood. The breakdown of the cohort by nationality and diagnostic criteria is given in Table 3.1.

Nationality	Diagnostic criteria	N trios
Bulgaria	DSM-IV, SCZ or SAD	69
Germany	ICD-10, SCZ	309
Italy	SCZ, diagnostic system not provided	11
Netherlands	DSM-IV, SCZ	78
Russia	ICD-10, SCZ or SAD	74
Spain	DSM-IV, SCZ	36
UK	DSM-IV, SCZ or SAD	39

Table 3.1. Breakdown of cohort by nationality and diagnostic criteria for inclusion. SCZ = schizophrenia, SAD = schizoaffective disorder

2.2 Genotyping and calling CNVs from array data

Samples were genotyped using OmniExpress-24 and CoreExome-24 Illumina bead arrays. Illumina’s BeadStudio v2.0 was used to call genotypes, to normalize the signal intensity data, and to establish the log R ratio and B allele frequency according to the standard Illumina protocols. PennCNV was used for rare (<1% frequency) *de novo* CNV detection, carried out by Dr Elliott Rees according to protocols described in (Rees et al., 2014). Results of this analysis have not been published.

2.3 Whole Exome Sequencing

Samples were isolated and prepared using the Nextra DNA Exome capture kit, HiSeq 3000/4000 PE Cluster Kit and HiSeq 3000/4000 SBS Kit, and sequenced on Cardiff University on Illumina HiSeq 3000/400 platforms. Raw sequencing reads in FASTQ format were processed according to GATK best practice guidelines (DePristo et al., 2011) and aligned to human reference genome (GRCh37) using BWA-mem version 0.7.1536 to generate BAM files for CNV calling. The mean read depth per sample was 31.7x and mean insert size was 164.2 bp.

2.4 Calling CNVs from WES data

CNVs for all samples were called from WES BAM files using CLAMMS (Packer et al., 2016), full description of which is given section 2 of chapter 2. Two aspects of the CNV calling process vary according to sample quality metrics and thus warrant separate discussion here. CLAMMS requires a user-defined insert size variable when generating the windows files, for purposes of calculating GC content. An insert is the sequence between adaptors and therefore determines the combined size of read pairs. The developers recommend a size that is ‘a little bit bigger’ than the mean insert size for the sequencing process used, such that most reads will come from inserts of sizes smaller than this value. The mean insert size of the sequencing processes used in the present study was estimated by calculating the mean of the mean insert sizes across all samples, which was 164.2 bp. Therefore, the insert variable size was set to 200.

The second sample-specific aspect of CLAMMS involves accounting for ‘batch effects’, where differences in sample preparation and input DNA quality can sometimes introduce stochastic volatility and distort read coverage depth exome-wide. CLAMMS controls for batch effects by clustering samples into reference panels based on 7 quality metrics generated by Picard (http://broadinstitute.github.io/picard.). The appropriate reference panel size (k) cannot be determined *a priori* and varies depending on the level of stochastic volatility in the sample. After QC, the k that minimises the number false positive and false negative calls should be selected. Given that the present study employs a trios design, the best indicator of error rates is the proportion of transmissions to non-transmissions across the sample after excluding low quality and common (>1% frequency) CNV calls, calculated as (n transmissions/(n non-transmissions + n transmissions).

Each parent will transmit only half of the variants they carry to their corresponding proband, and therefore, the expected transmission rate for parental CNVs is 0.5. If the call set contains an excess of false positives, however, there will be an excess of parental calls that are non-transmitted, giving a lower-than-expected transmission rate. If the call set has a high false-negative rate, there is a low likelihood for a transmission to be called in both proband and parent, also giving a lower-than-expected transmission rate. Thus, the reference panel size (k, described in section 2.2.3 of chapter 2) that best minimises both error rates should produce a transmission rate that is closest to 0.5. In the present study, k = 100 was selected. At this reference panel size, the transmission rate among high quality, rare calls was 0.32. When k was set to 50, the transmission rate was 0.25, while k=150 gave a transmission rate of 0.21 (Table 3.2).

K	N Transmission	N Non-transmission	Transmission rate
50	133	396	0.25
100	150	433	0.32
150	143	552	0.21

Table 3.2. Establishing reference panel size associated the best transmission rate. K = reference panel size. T = transmission

2.5 Sample quality control

The criteria used for performing sample- and variant-level quality control (QC) was based on (Maxwell et al., 2017), a DiscovEHR study that also used CLAMMS. First, outlying samples that contained > 2x median calls (n = 22) (Figure 3.1) were filtered out. Several trios were thereby rendered incomplete and were also removed at this stage as it would not be possible to determine the transmission status of their variant calls in later analysis.

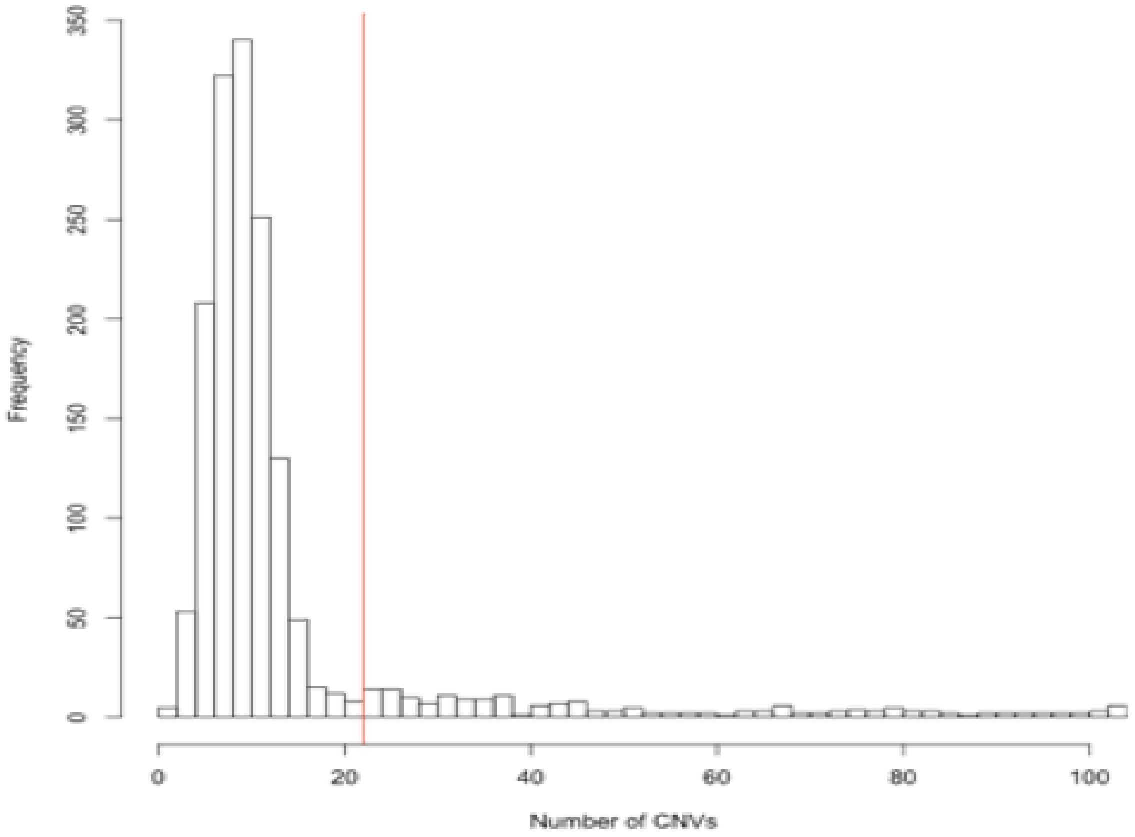


Figure 3.1. Histogram showing distribution of number of CNVs before quality control. The red line intercepts the x axis at n = 22. All samples with N CNVs > 22 were filtered.

2.6 Variant quality control

The first step of variant QC (prior to sample QC in the pipeline) was merging calls that occurred within 10kb of each other, as this may indicate that they are one variant that has been called as separate events by CLAMMS. To ensure that independent *de novo* CNVs that were observed in the same sample were not mistakenly merged together, I manually inspected the coverage profiles for all merged *de novo* CNVs and found no evidence that any *de novo* events were wrongly merged into a single CNV call.

I then filtered out the set of CNV calls in inlying samples that were disproportionally called in samples that were filtered during sample QC, as this suggests that CLAMMS is liable to generate false positive calls in these regions. Isolating the CNV calls in inlier and outlier sample sets independently, I identified inlier calls that overlapped other inlier calls by at least 50% reciprocally. I identified inlier calls that overlapped outlier calls to the same extent. Calls were removed from inliers if 2 x (n overlapping calls in inliers) < (n overlapping calls in outliers).

Calls were also filtered according to two quality metrics generated by CLAMMS: Qsome, a phred-scaled probability that the call region contains any CNV; and Qexact, a non phred-scaled quality score that is a measure of how closely the coverage-profile aligns with the called CNV and breakpoints. Deletions were filtered if Qsome <= 50 AND Qexact <= 0.5, while duplications were filtered if Qany <= 50 AND Qexact <= -1.0. The less stringent criteria for duplications is reflective of the fact that they are more difficult to identify using coverage depth than deletions (Teo et al., 2012).

PLINK 1.9 (Purcell et al., 2007) was used to identify common CNVs in parents, defined as those occurring in greater than 1% of the unrelated parents, which were then filtered out. I did not include probands when estimating variant frequency as this would entail an overrepresentation of transmitted calls among common variants, given that they should appear at least twice in the whole sample by default.

2.6.1 Defining transmitted, non-transmitted and de novo CNV calls

Parental calls were defined as transmitted if they overlapped any call in their respective proband call set by at least 1 base pair. Parental calls were defined as non-transmitted if they did not overlap any call in their respective proband call set, and proband calls were defined as *de novo* if they did not overlap any calls in the call sets for their respective parents. As described in section 2.4, the sample-wide transmission rate for rare variants was then calculated and used to determine the most appropriate reference panel size for CLAMMS.

2.6.2 Manual inspection of putative *de novo* CNVs

CNVs found to be *de novo* were manually checked by inspecting regional coverage plots for the call for each member of the respective trio (described in section 2.2.5 of chapter 2, also see figures below). These plots illustrate the normalised coverage depth of the call region for a given sample relative to the diploid mean (μDIP) of their respective reference panel as well the standard deviation for diploid copy number (σDIP) and expected coverage profiles for heterozygous deletions and duplications (Figure 3.2). Regions flanking the CNV breakpoints of 100kb were also included to assess coverage volatility around the putative events.

De novo and non-transmitted CNVs were accepted as real if most of the exons' relative coverage depths within the putative CNV were increased or decreased to around the + or – 50% level (Figure 3.3). Calls were also filtered out if exons beyond the breakpoints had similar deviations (Figure 3.4), suggesting that call may be the product of general coverage volatility in the wider region. I accepted an event as transmitted, if there was only partial evidence of deviation in the other member of the trio, as these would be unlikely to be real *de novos* or false positives. Therefore, CNVs that showed evidence of being present in the regional coverage plots in either parent, but were not called by CLAMMS, were removed from the *de novo* call set and added to the transmissions call set (Figures 3.5a & 3.5b).

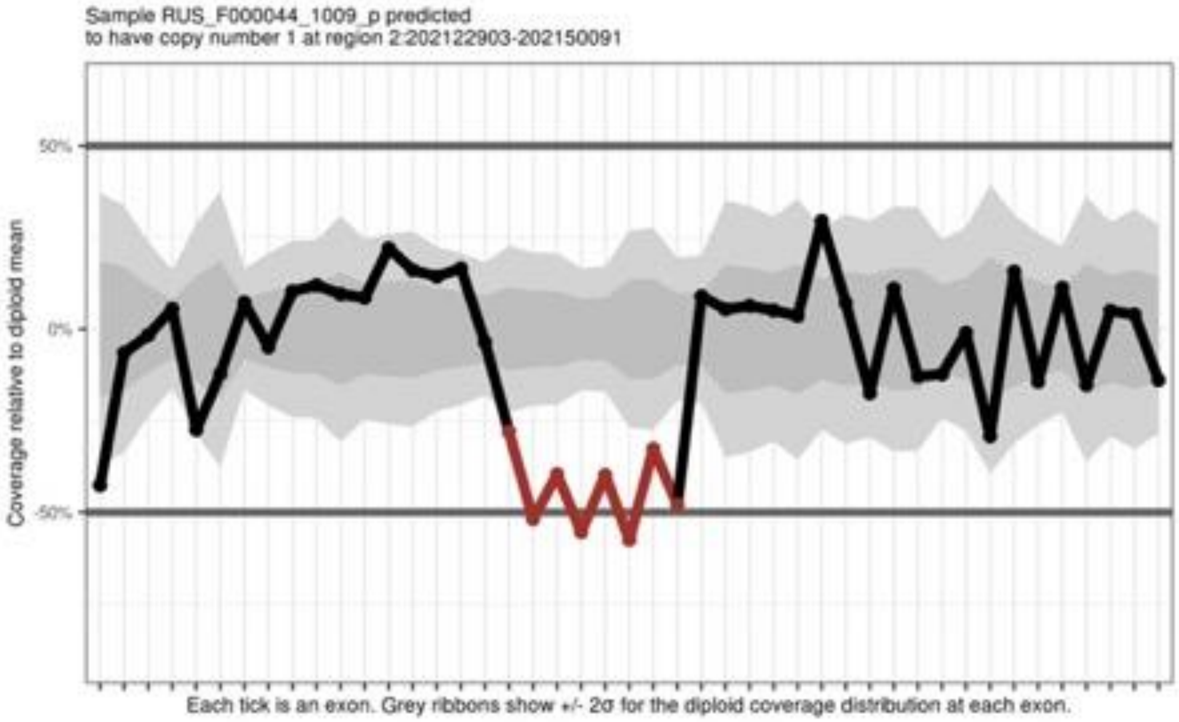


Figure 3.2. Example CLAMMS coverage plot, showing a heterozygous deletion. The red lines highlight the CNV call region, while the black lines show the coverage profile for this sample 100kb beyond the CNV breakpoints. Each node along the line represents the mean coverage depth for an exon. 50% and -50% on the y axis show the expected coverage depth for heterozygous duplications and deletions, respectively, relative to the mean of the sample's reference panel (μDIP). The dark and light grey regions indicate mean reference panel coverage depth +/- 1 and 2 standard deviations, respectively, for diploid copy number (σDIP).

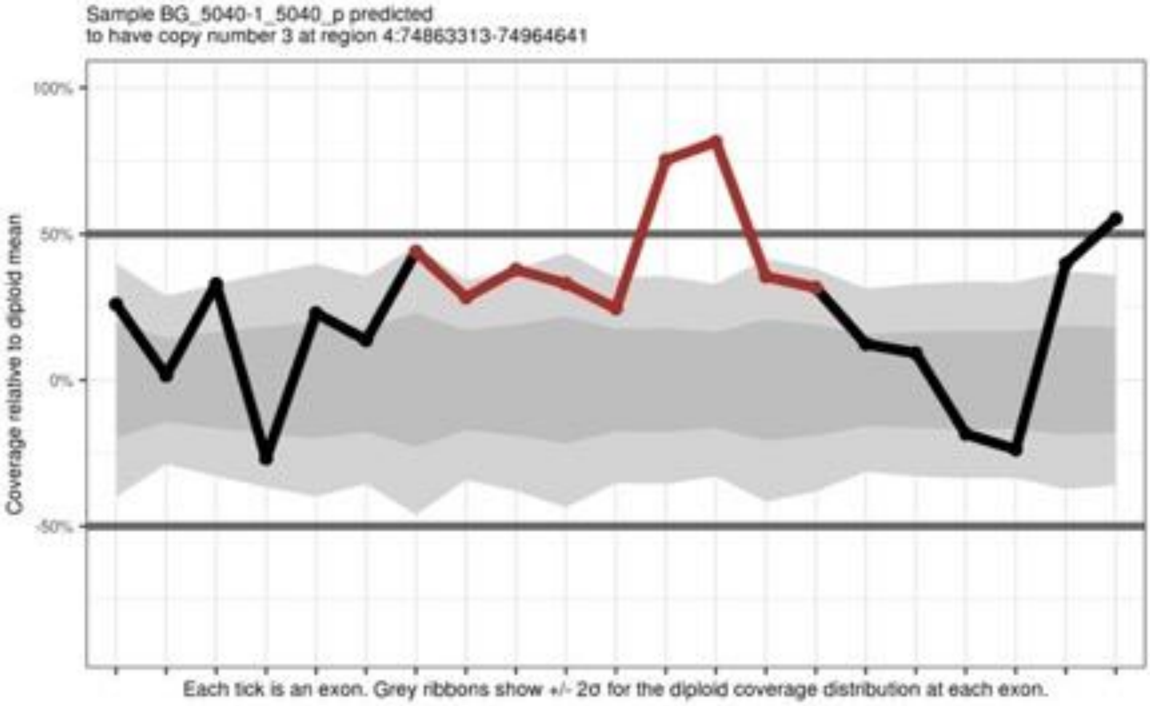


Figure 3.3. CLAMMS coverage plot for a putative heterozygous duplication. This call was excluded however as most of the exons within the call region have a mean coverage depth that lies within the expected variance for diploid copy number.

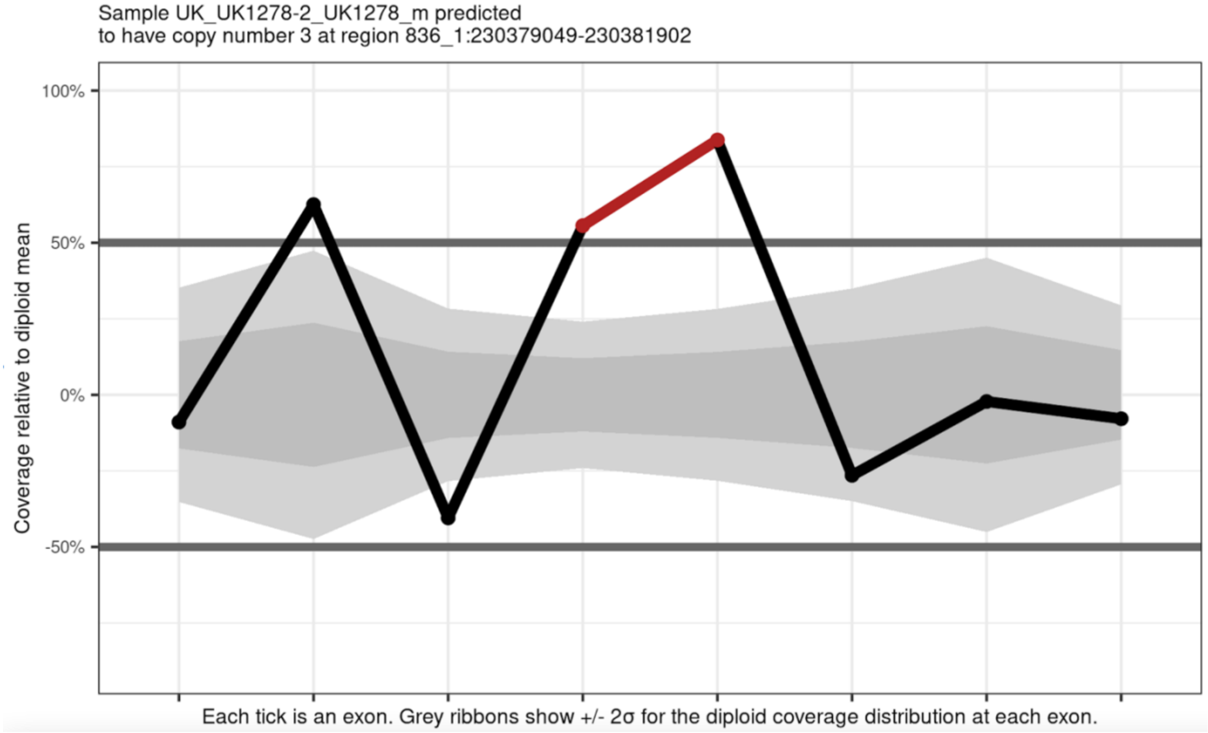


Figure 3.4. CLAMMS coverage plot showing evidence of a duplication. This call was excluded however as many of exons beyond the call region have a mean coverage depth that lies outside the expected variance for diploid copy number.

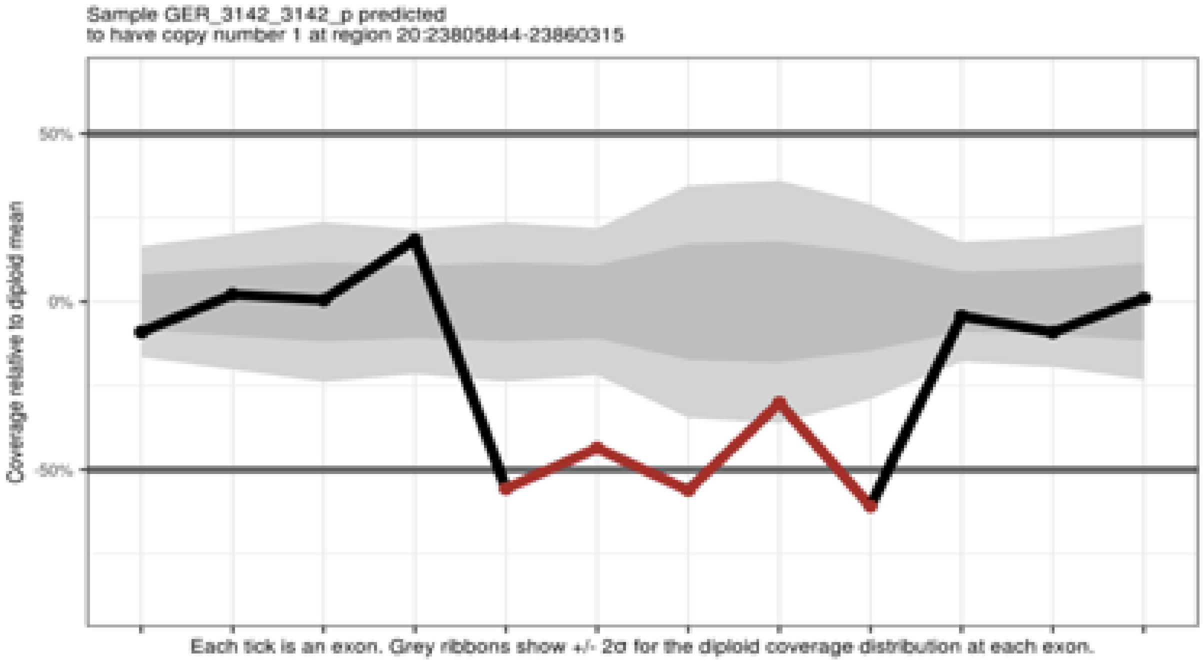
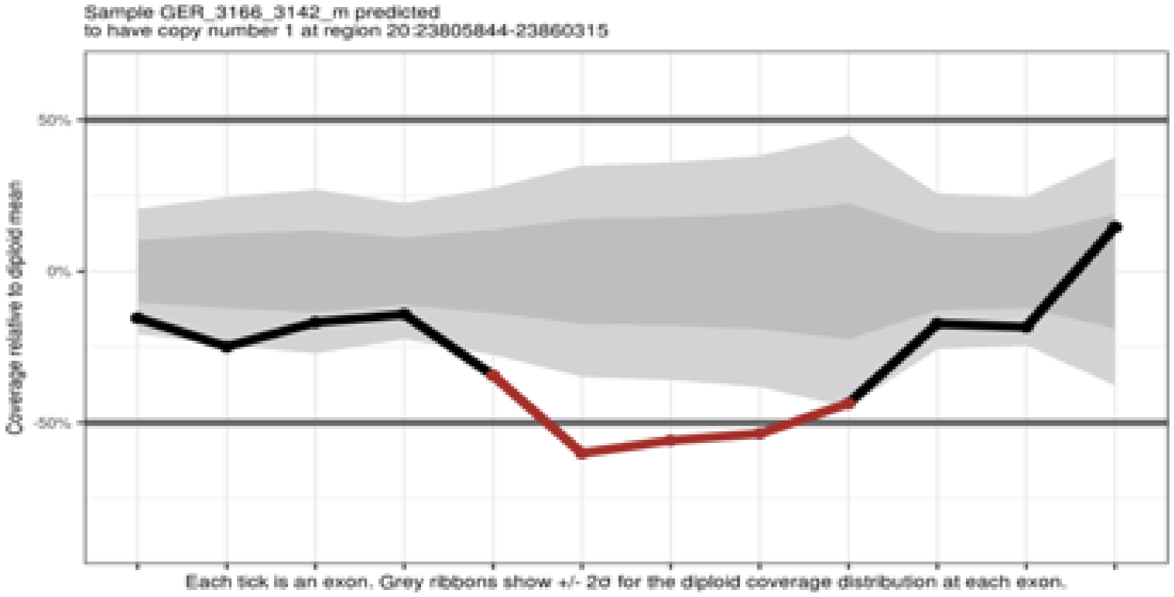


Figure 3.5a



Figures 3.5b. The CLAMMS coverage plot above (a) shows evidence of a deletion in a proband. This was designated as *de novo* in the transmission distortion analysis, as no deletion was called in the same region for either parent. The plot below showing the same genomic region in the mother (b) strongly suggests it is a transmission, however.

2.7 Quality control summary

This section summarises the number of samples or variant calls excluded at each stage of QC. A total of 230,180 CNVs were called by CLAMMS across all samples, reduced to 225,439 after merging adjacent CNV calls in the same sample. Prior to quality control, the median and mean CNVs called per sample were 11 and 122 (standard deviation = 397.8), respectively, with calls per sample ranged from 1 to 3829. From the initial sample of 1863 individuals, 429 were excluded for having an excess of CNV calls (> 2x median). An additional 331 individual were removed for being members of incomplete trios, leaving 369 complete trios for further analysis. After filtering low quality and common variants, 726 rare calls remained. Transmission rate analysis tentatively determined that 205 rare CNVs were transmissions, 433 were non-transmissions, and 88 were de novo, giving a sample-wide transmission rate of 0.32. Manual inspection excluded a further 288 calls, such that 177 transmissions, 252 non-transmissions and 9 de novos remained in the final analysis. Following manual inspection, the transmission rate increased to 0.41, indicating that my criteria for exclusion were effective at filtering false positive calls. Table 3.2 details the N of CNV calls remaining after each QC stage.

Quality control stage	**N CNV calls remaining**
Unfiltered CLAMMS output	230,180
Call merging	225,439
Remove outlying samples (N CNV calls > 22)	12,982
Remove incomplete trios	10,374
Filter calls overrepresented in outlying samples	10,010
Filter calls with low quality scores	6,916
Filter common variants (> 1% frequency)	726

Filter out calls that fail manual inspection	438

Table 3.2 - N CNVs remaining after each quality control step for CLAMMS calls. Entries in the first column state what was filtered out at each stage. Quality control steps are listed in order of application.

2.8 Assessing size distribution of transmitted and non-transmitted CNVs

To assess the size distribution of non-*de novo* CNVs, I separately binned transmitted and non-transmitted CNV into four size ranges: <10kb; 10kb-50kb; 50kb-100kb and >100kb.

2.9 Annotating variants for possible SCZ risk

Rare *de novo* CNVs that passed manual inspection were annotated for 50% reciprocal overlap with any of the 11 SCZ risk loci described in Rees et al. (2016) and shown in Table 1.1 in chapter 1. For *de novo* CNVs that did not occur in known SCZ risk loci, I evaluated whether any affected genes showed evidence for association with SCZ in large case-control exome sequencing or GWAS studies. First, the affected genes were tested via the SCHEMA database (Singh et al., 2020), a consortium of SCZ exome sequencing study results that includes 24,248 cases and 97,322 controls, and *de novo* mutations from 3,444 parent-proband trios. Finally, I tested whether the *de novo* CNVs overlapped any common variant loci implicated by (Trubetskoy et al., 2022), the PGC3 schizophrenia GWAS.

2.10 Assessing overlap with rare *de novo* CNV identified in array data

I then determined which rare *de novo* events that passed manual inspection overlapped those that were identified in the array data by at least one base. For those that did not, I evaluated the BAF and LRR of overlapping probes from the microarray data for evidence of deviations that would be expected for the type of CNV called in case they were real events but were missed by PennCNV. The relation between BAF/LRR and structural variation is described in section 2.4.1 of chapter 1. Rare *de novo* CNVs that were called in the array data but not by CLAMMS were also identified and individually assessed by manual inspection.

3. Results

3.1 Rare de novo CNVs detected by CLAMMS

Following manual inspection, 9 of the 88 rare *de novo* CNVs called by CLAMMS were classed as high-quality: 7 deletions and 2 duplications (Table 3.3). All were heterozygous events, called in 9 (1.5%) separate probands. Figures 3.6 and 3.7 show coverage depth plots for Chr10:18242203-19896831 deletion and 17:44171923-44249519 duplication, and the same regions plotted for the respective parent data, showing that they are not transmitted.

Three deletions occur at known SCZ risk loci:, 3q29 deletion, 16p13.11 deletion, and 22q11.2 deletion. I also found evidence that a further two deletions may confer SCZ risk: Chr18:163305-5478439 DEL was found to disrupt a gene with nominal association ($p < 0.05$) with SCZ in previous exome sequencing studies according to the SCHEMA database (Singh et al., 2020), *DLGAP1* (p = 0.0386, OR = 3.44). Chr10:18242203-19896831 DEL was found to overlap a SCZ-associated locus (Chr10:18538669-18751891, p = 4.80E-13) generated by clumping genome-wide significant signals for common variants (Trubetskoy et al., 2022). This locus encompasses the gene *CACNB2* and was still implicated after fine-mapping for casual variants.

Family ID	Chromosome	Start base	End base	Size (KB)	Type	Known SCZ loci
RUS_1009	2	202122903	202150091	27	DEL	
GER_3911	2	218568674	218604303	35	DUP	
GER_3855	3	195754042	197273323	1519	DEL	3q29 DEL
GER_681	6	10955346	11233735	278	DEL	
SPA_312	10	18242203	19896831	1654	DEL	
RUS_2022	16	15493193	17451938	1958	DEL	16p13.11 DEL
GER_3902	17	44171923	44249519	77	DUP	

GER_3378	18	163305	5478439	5315	DEL	
UK_1238	22	18900636	21411491	2510	DEL	22q11.2 DEL

Table 3.3 - *de novo* CNVs called in the exome sequencing trios data, including size in kilobases, CNV type and intersections with known schizophrenia risk loci. SCZ = schizophrenia, DEL = deletion, DUP = duplication, KB = kilobases.

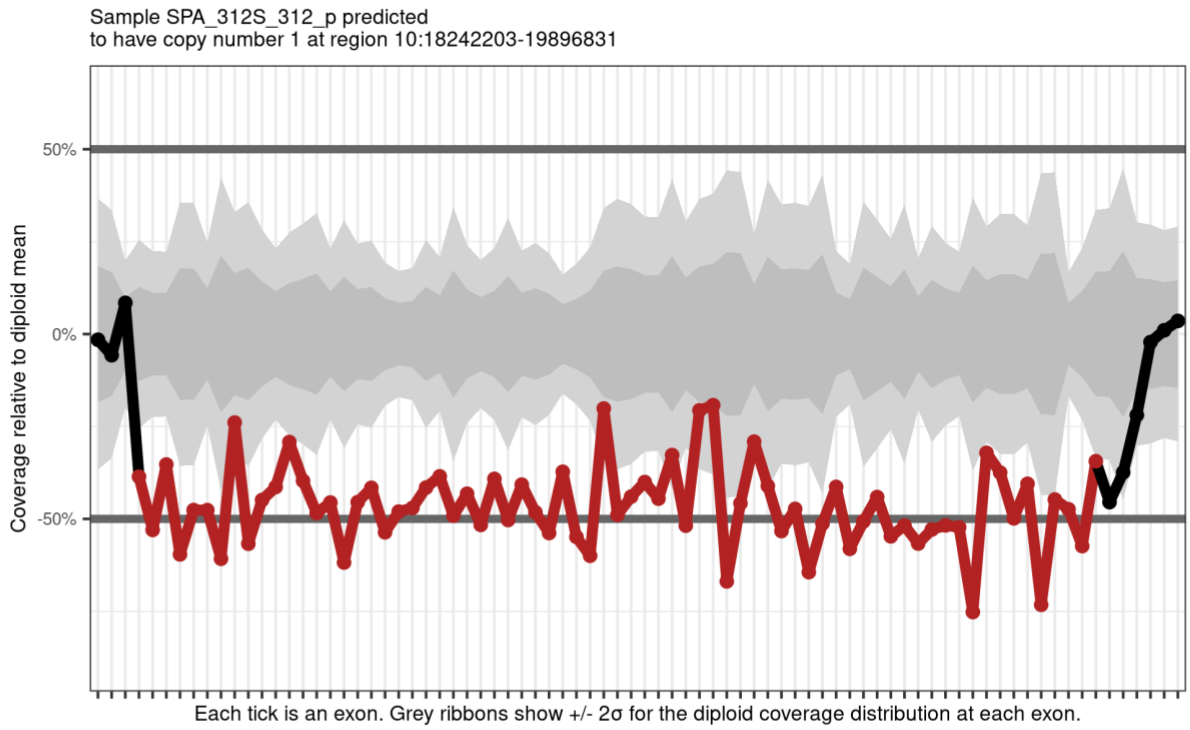
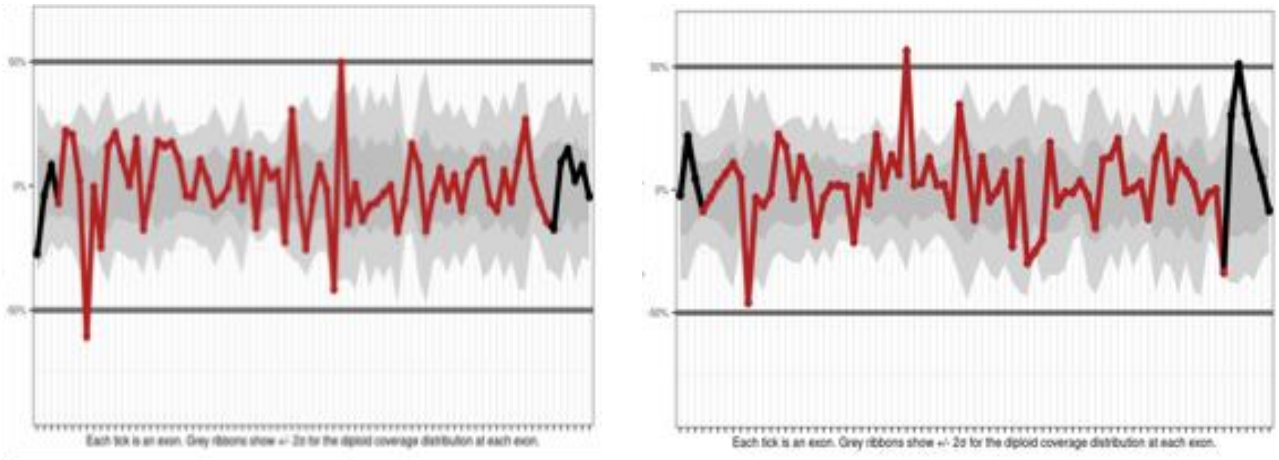


Figure 3.6a - CLAMMS coverage depth plot for *de novo* CNV Chr10:18242203-19896831 DEL.



Figures 3.6b and 3.6c – Proband call region Chr10:18242203-19896831 plotted for the respective parental samples, showing that no CNV is present in either case.

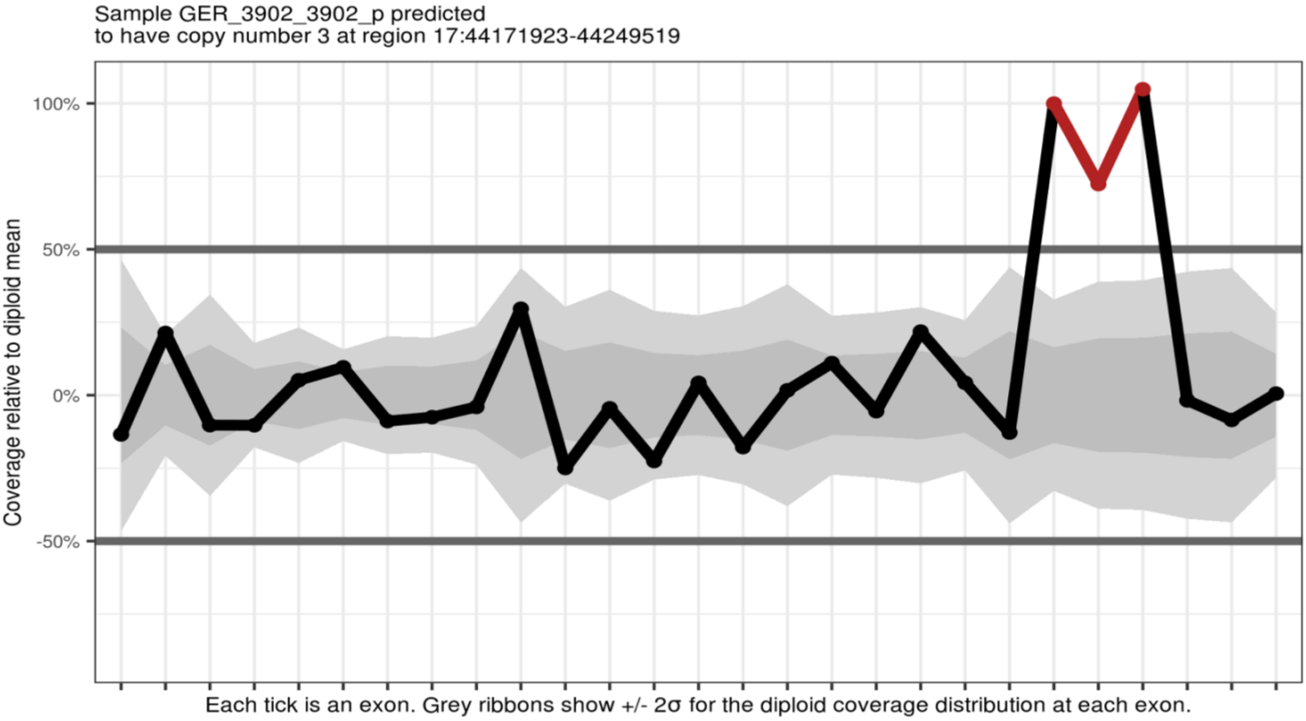


Figure 3.7a – CLAMMSS coverage depth plot for *de novo* CNV Chr17:44171923-44249519 DUP

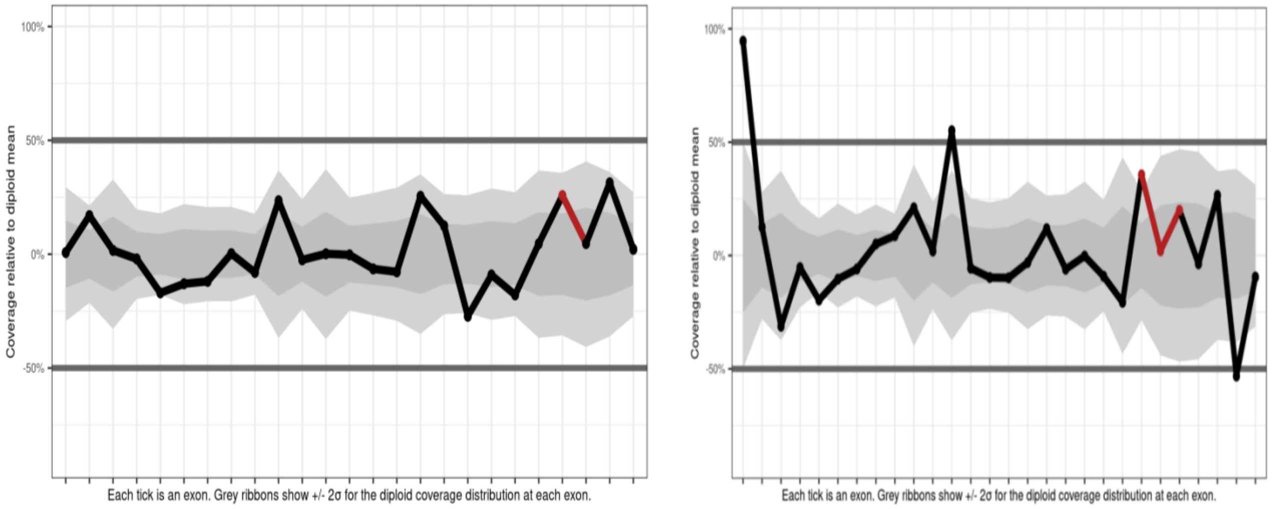


Figure s 7b and 3.7c – Proband call region Chr17:44171923-44249519 plotted for the respective parental samples, showing that no CNV is present in either case

3.2 Presence of rare *de novo* CNVs detected by CLAMMS in array data

5 of 9 rare *de novo* deletions detected by CLAMMS were also identified in the array-based analysis (Table 3.4). Of the remaining 4 CNVs detected by CLAMMS only, evidence was found for the presence of 2 in their corresponding array-based BAF/LRR data. Due to its large size, Chr10:18242203-19896831 DEL was likely called in the array analysis, but its corresponding sample did not pass QC. However, the BAF/LRR data clearly indicates a deletion in this region (Figure 3.8). Chr17:44171923-44249519 DUP and Chr2:218568674-218604303 DUP were missed entirely by the array-based analysis as they are both <100kb in size, and occur in regions with few array probes, shown in Figures 3.9 and 3.10. However, most probes in Chr17:44171923-44249519 DUP call region have an LRR around 0.33, the expected signal intensity for a heterozygous duplication. There are not enough probes in the array data of Chr2:218568674-218604303 DUP to confirm its presence. While array data for the sample carrying Chr22:18900636-21411491 was unavailable, it is a large 22q11.2 deletion event, and so would most likely have been identified in the array-based analysis if its carrier were included.

CNV detected by CLAMMS (CHR:BP1-BP2)	Size (KB)	Type	Identified in array data	Supported by manual inspection of BAF/LRR
2:202122903-202150091	27	DEL	Yes	-
2:218568674-218604303	35	DUP	No	No
3:195754042-197273323	1519	DEL	Yes	-
6:10955346-11233735	278	DEL	Yes	-
10:18242203-19896831	1654	DEL	No	Yes
16:15493193-17451938	1958	DEL	Yes	-
17:44171923-44249519	77	DUP	No	Yes
18:163305-5478439	5315	DEL	Yes	-
22:18900636-21411491	2510	DEL	No	Array data unavailable

Table 3.4 - Status of rare *de novo* CNVs detected by CLAMMS in the array-based analysis. DEL = deletion, DUP = duplication, BAF = beta allele frequency, LRR = log R ratio.

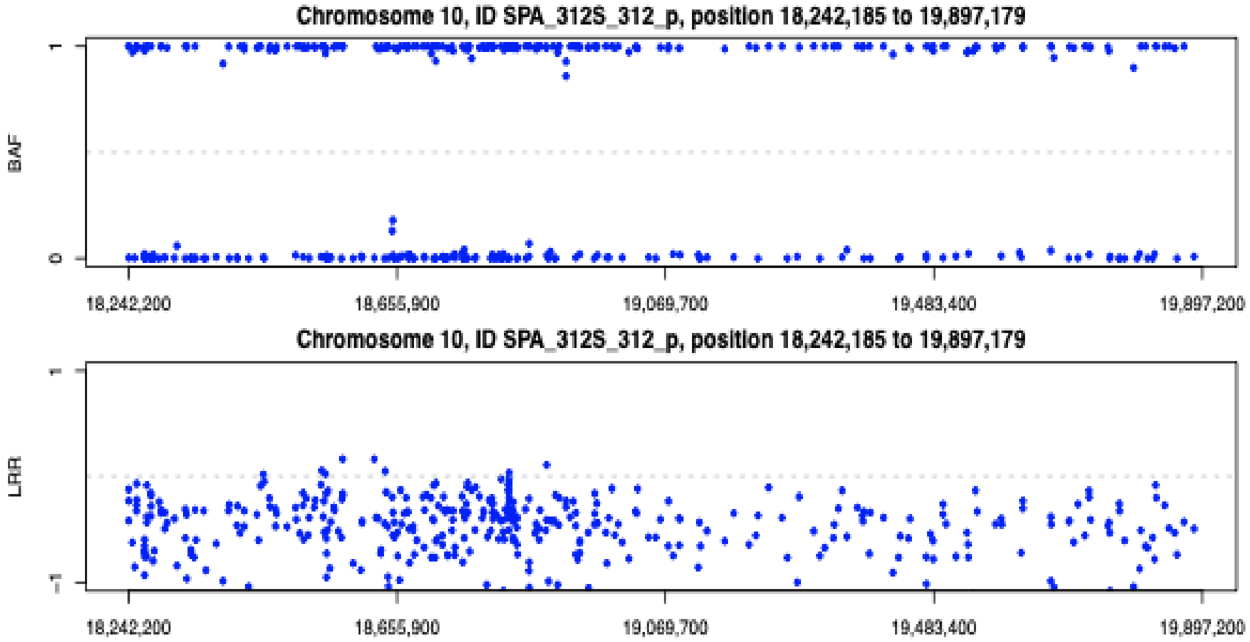


Figure 3.8. B allele frequency (BAF) and Log R Ratio plots (LRR) for *de novo* CNV discovered in the WES analysis, Chr10:18242203-19896831 DEL. Start and end base pairs differ from the sequencing calls as these are the closest recorded SNPs to the sequencing loci. The blue points represent array probe signals. Due to the deletion of an allele at every recorded SNP along the region, the beta allele for each probe is either absent or the only signal present. Thus, the frequencies for beta alleles cluster around 0 and 1 in the BAF plot. LRR is a measure of probe signal intensity. As a deleted sequence entails 50% fewer SNPs present to bind array probes, the LRRs cluster around –0.5 on the normalised scale.

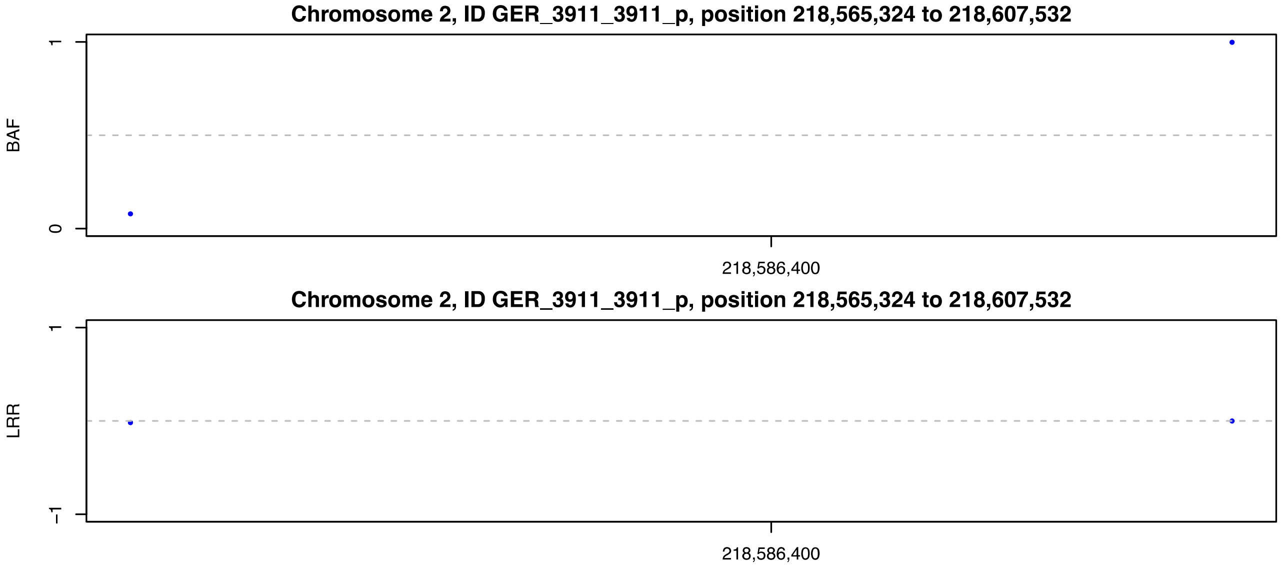


Figure 3.9. B allele frequency (BAF) and Log R Ratio plots (LRR) for *de novo* CNV discovered in the WES analysis, Chr2:218568674-218604303 DUP. This CNV's small size (35kb) is reflected by the fact that there are only two probes in its call region. I was therefore unable to confirm the presence of this CNV in the array data.

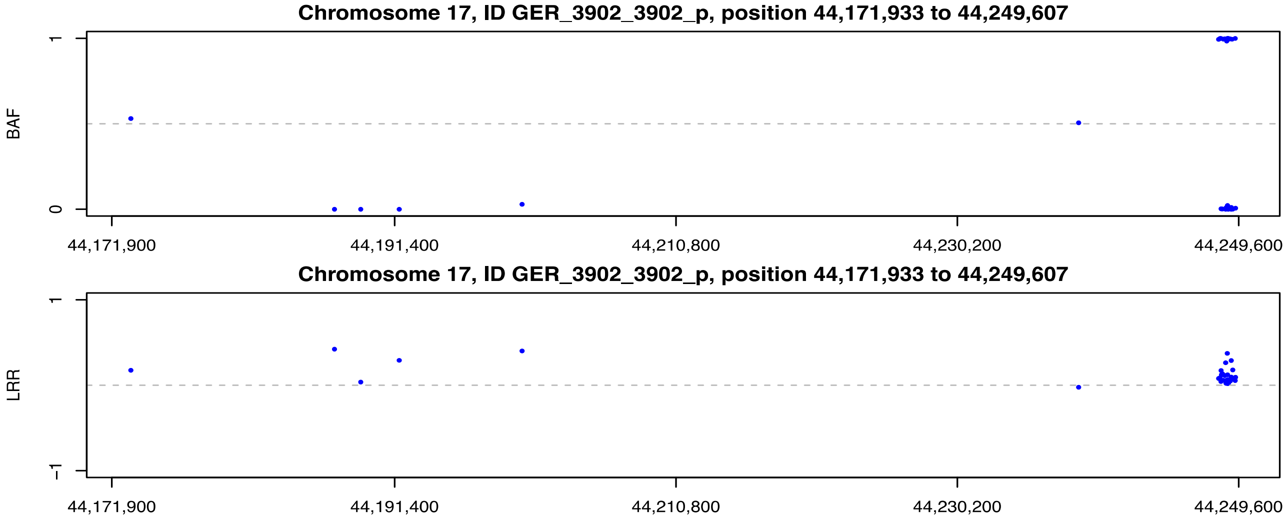


Figure 3.10. B allele frequency (BAF) and Log R Ratio plots (LRR) for *de novo* CNV discovered in the WES analysis, 17:44171923-44249519 DUP. While this CNV's small size (77kb) also means that there are a small number of probes in its call region, most of the probes have an LRR above the 0 baseline, the expected signal intensity for a heterozygous duplication. However, the BA plot is uninformative. I therefore concluded that the array data only partially confirmed the presence of this CNV.

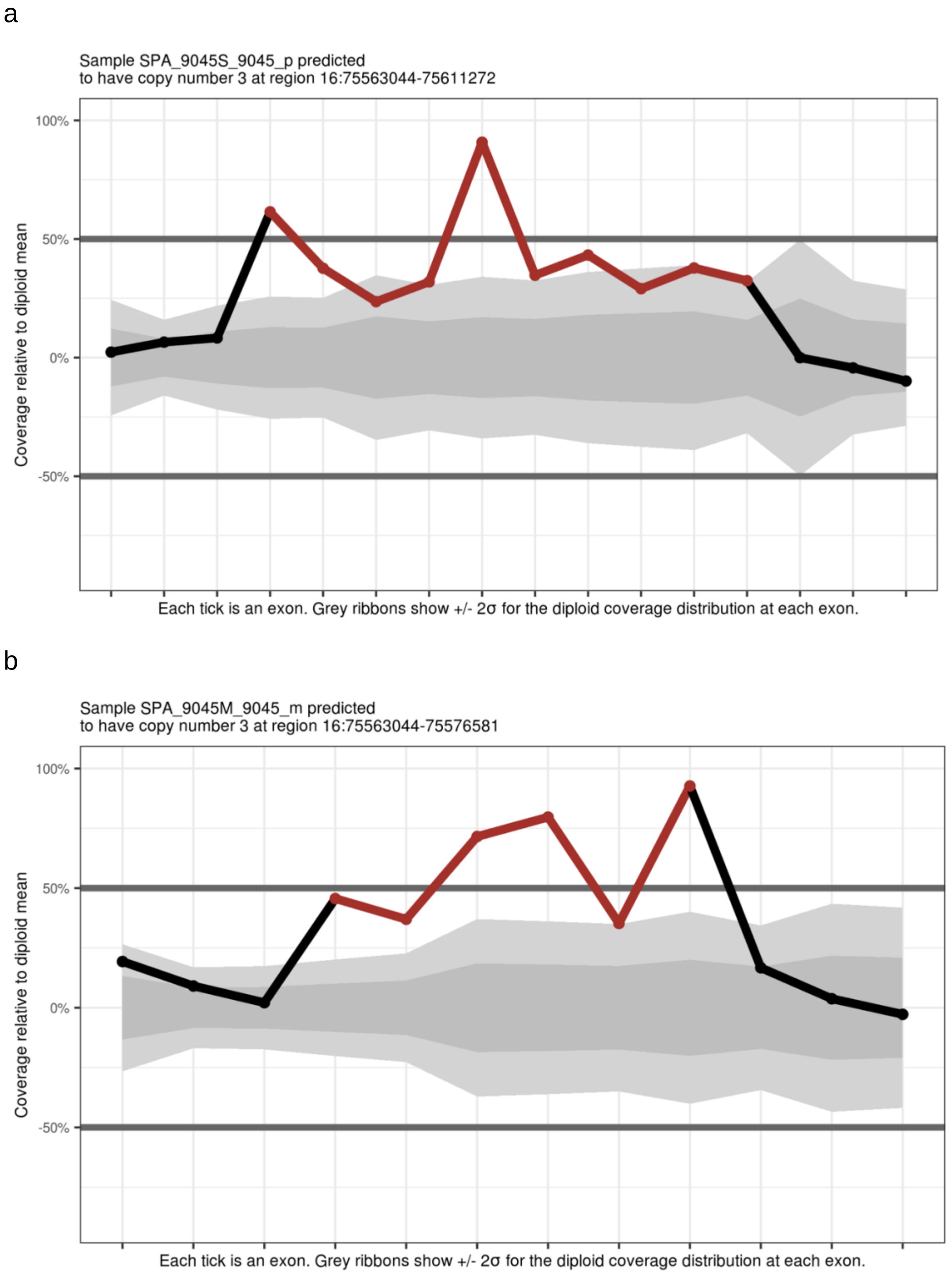
3.3 Rare de novo CNVs previously detected in the microarray data and CLAMMS validation

Ten rare *de novo* CNVs were previously discovered in the microarray data that could potentially be identified in the WES data: 5 deletions and 5 duplications. All 5 deletions and one duplication were detected by CLAMMs (Table 3.5), though the latter was found to be a transmission in the sequencing data (Figure 3.11). Three duplications found in microarray data were called by CLAMMS but were removed during QC as their respective samples had an excess of CNV calls, while the remaining duplication was missed entirely by CLAMMS.

Family ID	CHR	Start base	End base	Size (KB)	Type	Validated by CLAMMS
RUS_1009	2	202102685	202149628	46	DEL	Yes
GER_3855	3	195750742	197346566	1595	DEL	Yes
GER_681	6	10955408	11227987	272	DEL	Yes
GER_2630	7	73184318	74115258	930	DUP	Excluded during QC
GER_2747	9	105765465	105767917	2	DUP	Not called
GER_2114	10	131753010	132043419	290	DUP	Excluded during QC
GER_3241	13	65296926	69633096	4336	DUP	Excluded during QC
RUS_2022	16	15493046	18166320	2673	DEL	Yes
SPA_9045	16	75558483	75577559	19	DUP	Found to be transmission

GER_3378	18	61355	5952544	5891	DEL	Yes

Table 3.5 - *de novo* CNVs discovered in microarray data, along with size in bases and CNV type. Microarray calls were considered validated if they were also called from sequencing data by CLAMMS and passed all QC, including manual inspection. DEL = deletion, DUP = duplication.



Figures 3.11 a & b. CLAMMS coverage depth plots for Chr16:75558483-75577559 DUP in proband (a) and parent (b). This CNV was considered de novo in the microarray analysis but was determined to be transmitted in the exome sequencing data

3.4. Size distribution of rare, transmitted and non-transmitted CNVs

Though rare transmitted and non-transmitted CNVs were called across a broad size spectrum in the WES data, the majority were smaller than 100kb (Fig 3.12). The median size was 34.4Kb. Of 177 transmissions that passed QC: 47 were < 10kb; 59 were 10-50kb, 37 were 50kb-100kb and 34 were >100kb. Of 252 non-transmissions that passed QC: 117 were < 10kb; 102 were 10-50kb, 23 were 50kb-100kb and 10 were >100kb.

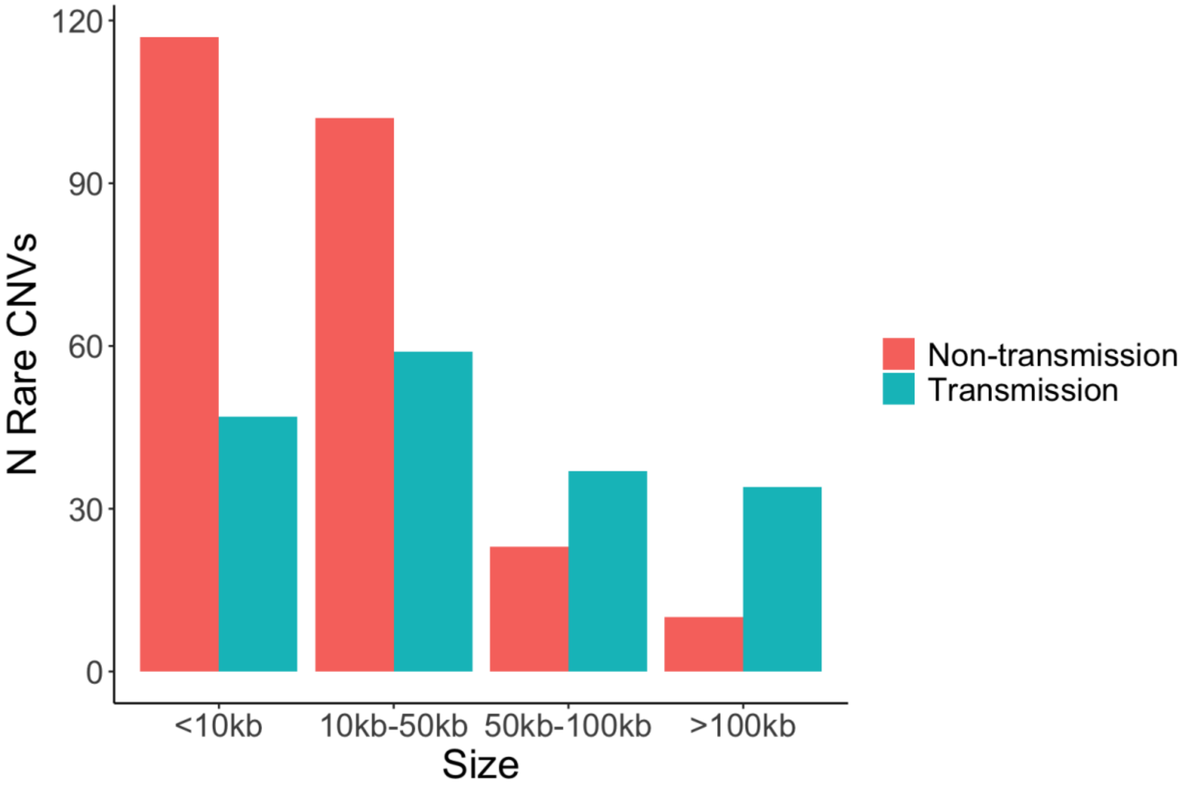


Figure 3.12. Size distribution of rare transmitted and non-transmitted CNVs, subset by 4 size ranges. kb = kilobases.

4. Discussion

My results demonstrate that large, rare de novo CNVs known to be pathogenic for schizophrenia can be reliably called from exome sequencing data and strengthen the case for the role of de novo CNVs in the aetiology of schizophrenia. In addition to the three *de novo* CNVs that overlapped known SCZ loci, I've reported evidence of SCZ association for two novel loci, drawing on results from previous sequencing and common variant studies.

Evidence for the role of *DLGAP1* is increased here, as it was found to be affected by *de novo* CNV carried by a SCZ individual in an earlier study (Kirov et al, 2012), and there is evidence that it harbours more protein-truncating and damaging missense variants in schizophrenia cases and controls in SCHEMA (OR = 3.44). *CACNB2* was also implicated by a *de novo* CNV in the present study and has been shown to harbour putative causal SNPs in the PGC3 schizophrenia GWAS. Both *DLGAP1* and *CACNB2* are expressed at the post-synapse and have roles in synaptic organisation and plasticity (Rasmussen et al., 2017; Dolphin, 2012), and are thus highly plausible risk genes.

I've also identified putative de novo CNVs in the sequencing data that were not detected in the microarray data, suggesting that a combination of both types of data could increase sensitivity for discovering CNVs. Two of these CNVs were <100kb in size, and manual inspection of their array data showed that the most likely reason for their being missed was a low number of probes in the call regions. This demonstrates that CNV calling using WES data can mitigate the resolution issues of an array-based approach.

However, while CLAMMS detected all five de novo deletions observed in the microarray data, four of five array data duplications were either filtered or not called in the array data. In the three cases where the CNVs were filtered, this was due to an excess of CNV calls (indicating poor quality) in their respective samples. The array duplication that was not called by CLAMMS is only 2kb in length, and therefore likely to be a false positive.

Assessing the size distribution of rare transmissions and non-transmissions called by CLAMMS, this method is sensitive to variants across a broad size spectrum. Indeed, 164 of these manually inspected events (117 non-transmissions and 47 transmissions) are smaller than 10kb and so would likely be missed entirely in array-based analysis. The excess of small events among non-transmissions is indicative of an increased false positive rate compared with transmissions, which is expected given that transmissions are more likely to be real by virtue of their being called in both proband and parent. It cannot be ascertained from my results, however, what proportion of small variants were not detected by CLAMMS, and my findings require orthogonal support from a separate approach, such as quantitative PCR.

Chapter 4: Detecting small structural variants in schizophrenia proband-parent trio exome sequencing data

1. Introduction

1.1 Background

One of the aims of my PhD was to identify structural variants (SV) in WES that had been undetectable in array data, and to assess their association with schizophrenia. I partly met this aim in the previous chapter, where I used the CLAMMS algorithm to detect putative de novo CNVs in the WES SCZ trios samples that were smaller than the approximate discovery resolution of PennCNV (~100kb) (Wang et al., 2007). However, 5/9 of the *de novo* CNVs were also detected in the array data, demonstrating that CLAMMS and PennCNV are also sensitive to the same variants. For the current chapter, I progressed this aim by applying the SV calling algorithm InDelible (Gardner et al., 2021) to the same trios sample. As described in section 3.4.1 of chapter 2, InDelible is most sensitive to variants 21-50bp in size, and is largely insensitive to variants >500bp, such that the SVs it is designed to detect from WES data are undetectable in array data. In addition to deletions and duplications, InDelible can detect balanced events (insertions, translocations), which are also undetectable by both PennCNV and CLAMMS. When the research presented in this chapter was conducted, the role of structural variants detectable by InDelible in schizophrenia had not been studied.

1.2 Study Aims

Completed in the second year of my PhD, the primary aim of this research was to analyse WES from a schizophrenia trio sample to detect rare (allele-frequency <1%) *de novo* SVs that cannot be discovered from array data. While large (>100kb) *de novo* SVs have been shown to be important risk factors for schizophrenia (Rees et al., 2012) it is unknown whether smaller SVs, including those within the size range detected by InDelible (<100bp), also contribute to disease risk. The second aim of this chapter was to therefore assess whether any *de novo* variants discovered using the InDelible algorithm are likely to contribute to increased risk for schizophrenia. This was done by evaluating whether any *de novo* SVs intersected genes that have previously been associated with SCZ in small variant studies (Singh et al., 2022).

2. Methods

2.1 Trios recruitment and whole exome sequencing

The samples used in the current study are the same as those described previously in chapter 2, section 2.1. Thus, the recruitment protocols, sample size and sequencing procedures were the same as those specified that study.

2.2 Structural variant calling: InDelible

SVs were called using InDelible (Gardner et al., 2021), described in section 3 of chapter 2. In this chapter, I produced a call set that was designed to identify de novo SVs, which required all 6 steps of the algorithm to be applied to proband WES data. Here, the 'denovo' step interrogates proband SV call regions in both parents, determining the number of reads that split at the same base positions as in the proband. If no split reads are identified in proband call regions in either parent, the proband call is likely to be *de novo*.

InDelible requires a configuration file as input, specifying parameters that are used for variant processing at different stages. I configured the algorithm to exclude reads with mapping quality < 5, base quality < 10, and SR length < 5 at first processing step (Fetch), and to exclude SR clusters containing <3 SRs at the second step (Aggregate). These thresholds were recommended by the InDelible authors (Gardner et al., 2021). I also configured InDelible to exclude SR clusters if the read depth coverage at the same position in either parent was <9. At lower coverage, there may be too few split reads to call SVs in parents even if an SV is present. By default, the 'denovo' step also outputs only those proband calls with a prob_y quality score >0.6 of being real events, according to the random forest model generated in InDelible step 3 (Score). In the initial pre-QC proband-only call set, 268,526 SV calls were identified in 604 probands. For 12 probands, all SV calls were filtered by InDelible's default variant processing.

2.3 Quality control

2.3.1 Sample-level quality control

I excluded samples that were outliers for their total number SV calls. The number of SVs calls per individual was normally distributed (Figure 4.1). The lower and upper thresholds for sample exclusion based on number of SV calls, which were determined by inspecting the distribution of SV calls, were < 190 calls and > 780 calls, respectively.

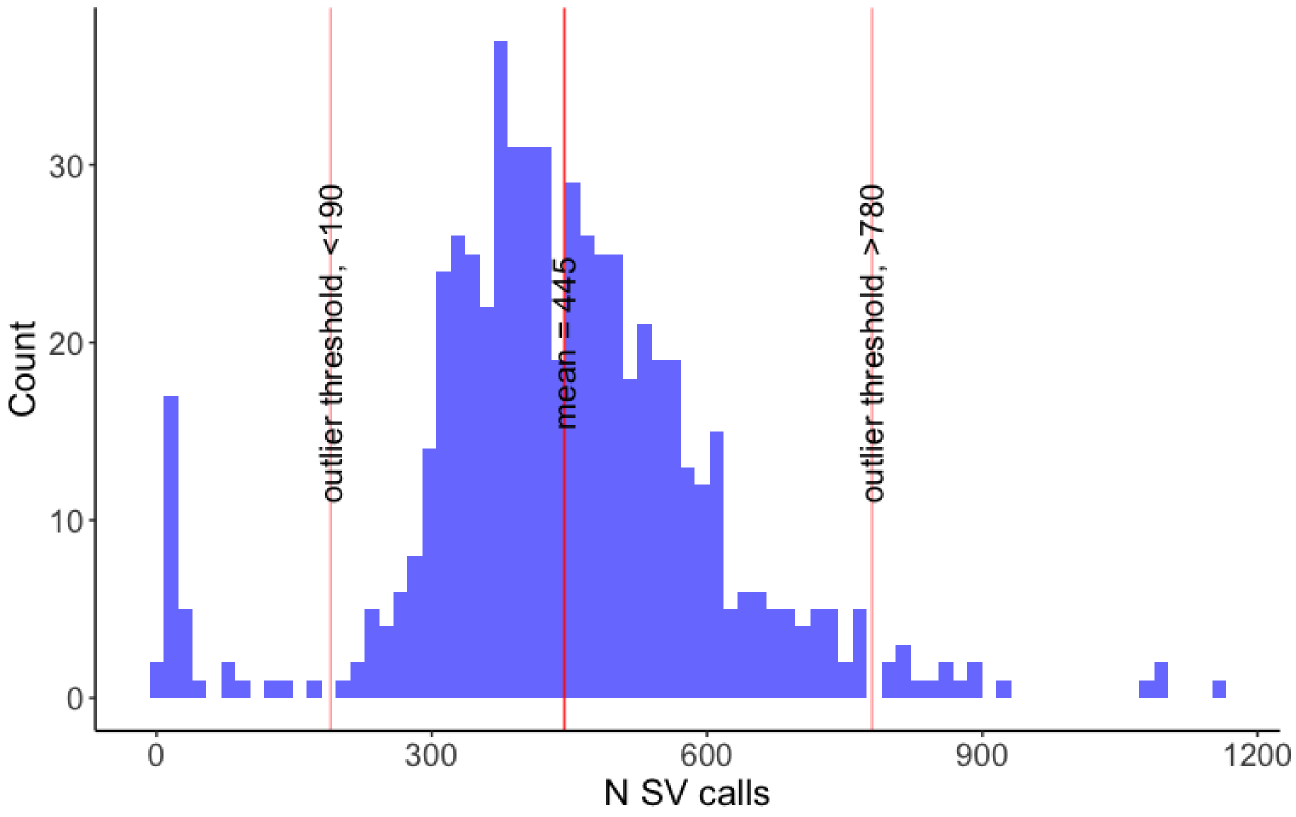


Figure 4.1: number SV calls histogram for the proband-only call set, annotated with the mean number and outlier thresholds.

48 samples were excluded in the proband-only call set. 16,569 SV calls were filtered from the proband-only call set by sample exclusion. 251,957 variants in the proband-only call set were retained for further analysis.

2.3.2 Variant-level quality control

Using the fourth step of the InDelible algorithm (Annotate), calls were annotated with their allele-frequencies (AF) in the Deciphering Developmental Disorders (DDD) AF database (sample size = 13,438). Additionally, I created an AF database from the parent samples of the SCZ trios (sample size = 1238) using the InDelible command

‘Database’. The inputs and output files for this command are described in section 3.3.4 of chapter 2. In my analysis of SVs, calls were retained if they had an AF <= 0.01 in both the DDD data and among the SCZ parents or were absent from the DDD database and/or SCZ parents database. After applying SV AF filters, 227,137 calls were excluded from the proband-only set.

Additional variant level SV QC was applied following the recommended criteria outlined in the original InDelible DDD study (Gardner et al. 2022). First, SV calls were filtered if the average mapping quality (MAPQ) of their constituent SRs was < 20. As soft-clipped bases are not included in the calculation for a read’s MAPQ, this filter ensures that the aligned portions of the SRs are unlikely to be mis-mapped. Second, calls were excluded if >10% of their SRs were split at both the 5’ and 3’ end (double split), but only if they didn’t also have a valid bwa alignment. These alignments are generated in the Database step of the InDelible algorithm and are used to determine likely variant type and breakpoints. SR clusters with a high proportion of double split reads and with no BWA alignment or BLAST hit are most likely the result of errors when adaptors were cleaved from reads during sequencing. Calls were then excluded if they were aligned to non-exonic sequences according to human reference genome GRCh37. Given that all reads are expected to be derived from the exome, this step excludes any calls that are the consequence of read misalignment or non-exonic DNA contamination. Finally, remaining calls were excluded if they had < 5 SRs in their corresponding cluster.

Finally, SV calls in the proband-only call set were filtered if either parent sample had >2 SRs at the same position. According to InDelible author Eugene Gardner, this threshold was chosen to account for expected noise in WES data which can produce SRs even when no structural variant is present, usually because of read misalignment. After applying sample and variant level QC, 59 putative *de novo* calls remained in 25 samples. Twenty-three samples had one call, while one sample had ten de novo calls and the remaining sample had 26. In each of the latter two samples, all calls intersected the same gene. As I describe in section 3.6.6 of chapter 2, this call pattern is indicative of pseudogene retrotranspositions, and so did not in itself provide sufficient grounds for exclusion.

2.3.3 Manual Inspection

IGV snapshots were created for each putative *de novo* call, along with snapshots for the same locus in both parents, according to procedures described in section 3.4 of chapter 2. Based on these snapshots, there were four criteria by which calls were excluded:

1) One of the two (or more in case of complex events and pseudogene retrotranspositions) SR clusters in the proband were also found in a parent sample (i.e., it was a transmission and not a de novo event). In some cases, the other SR cluster associated with an SV at the called loci was filtered for having too few SRs but would have been identified in a parent according to the parent SR > 2 criteria. If the SR cluster that passed QC happens to a have <2 SRs in parents, the call is misidentified as *de novo.* Figure 4.2 shows an example of this in proband and parent, respectively. Seven calls were excluded by this criterion.

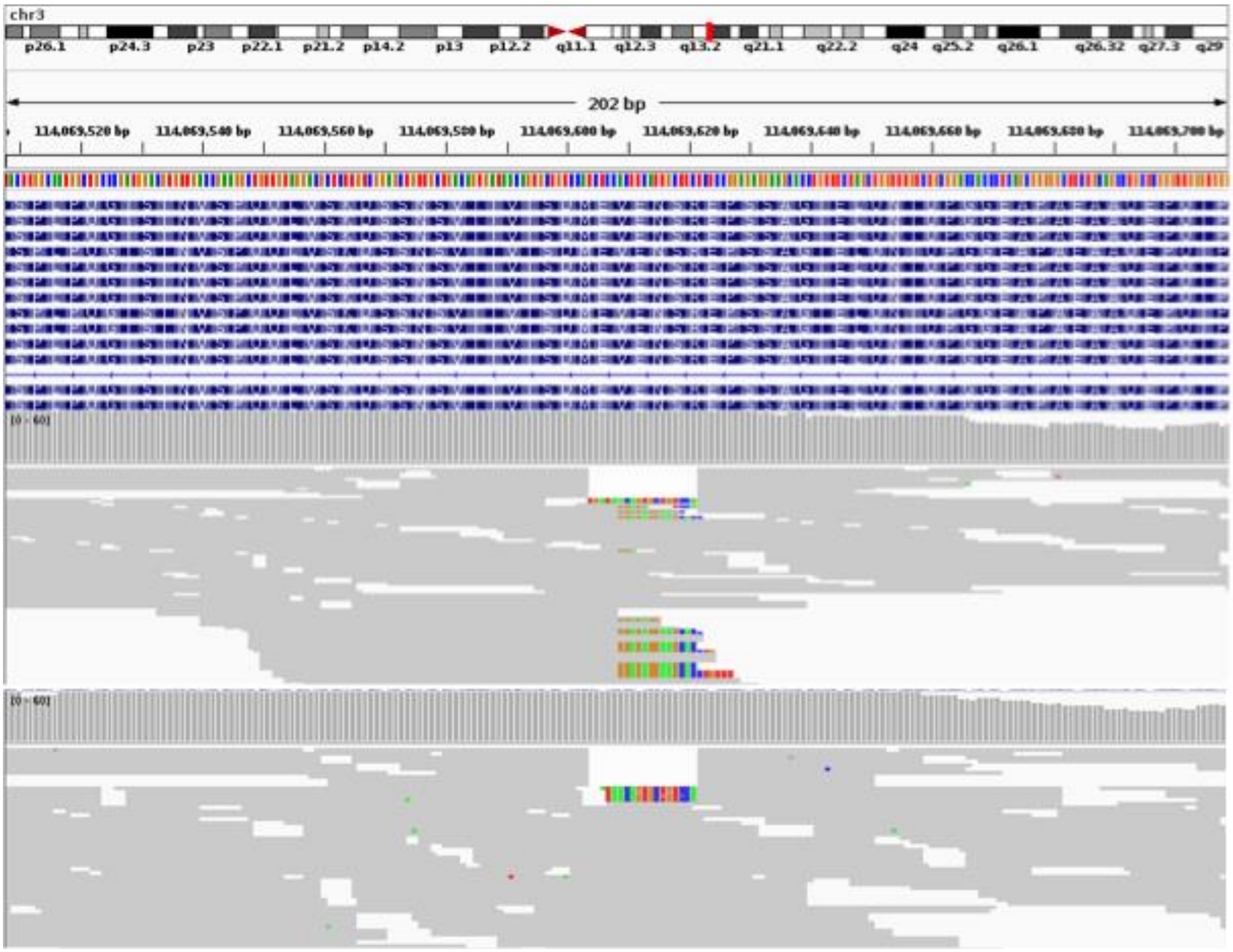


Figure 4.2. The top alignment track in this IGV snapshot shows an 18bp

chromosome 3 deletion called in a proband. The topmost SR is part of a cluster of 3, the other 2 SRs of which were not included in the snapshot but were identified by inspecting all reads at this locus in the IGV interface. This cluster was filtered during the aggregate stage of calling for having < 5 SRs. However, it also occurs in a parent (bottom alignment track), in which it has 4 SRs. It would therefore have been identified as a transmission in the proband according to the parent SRs > 2 criterion. The other SR cluster observed in the proband passed all *de novo* QC as it unusually does not appear at all in the parent data.

2) The longest SR in the cluster had no valid BWA-mem alignment and there was a significant discrepancy between the misaligned bases in each SR. This indicates that multiple errors were introduced in these reads during DNA processing, and so it is probable that any structural change corresponding to the SR cluster is a consequence of these errors and does not occur in the source DNA. An example is shown in Figure 4.3, the only call excluded by this criterion.

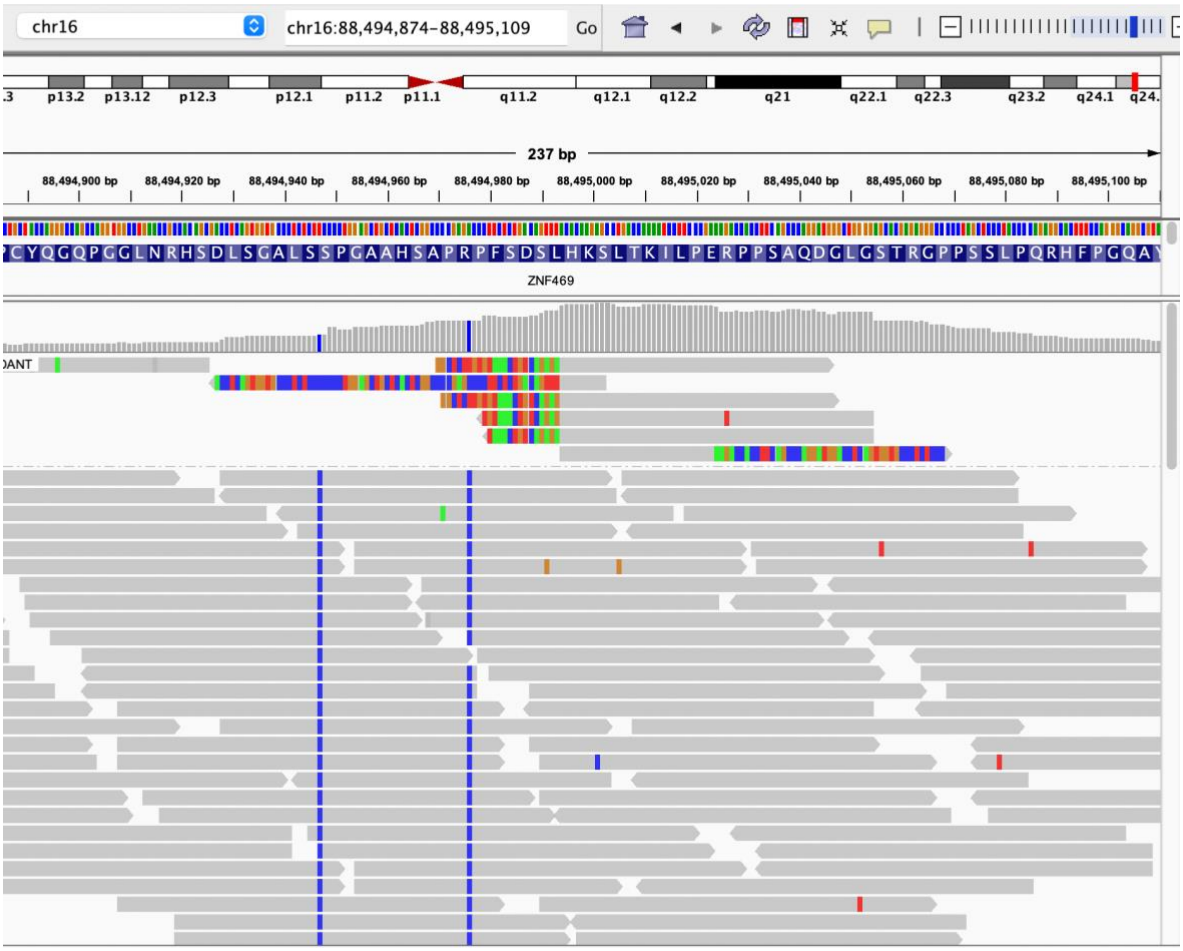


Figure 4.3. In this IGV snapshot, the misaligned bases in the SR cluster either do not themselves align, or in one case are significantly different from those in the other SRs. It is likely that errors were introduced into the reads at the loci during the sequencing process, and so we cannot be confident that the called SR cluster is indicative of a real event.

3) The SR cluster is mapped to a sequence that contains successive instances of a single base, and all the misaligned bases in the SR are also the same base. Such loci are prone to replication errors, so again we cannot be confident that any structural change in the proband DNA is not a consequence of these errors. Figure 4.4 shows the only event to be excluded by this criterion.

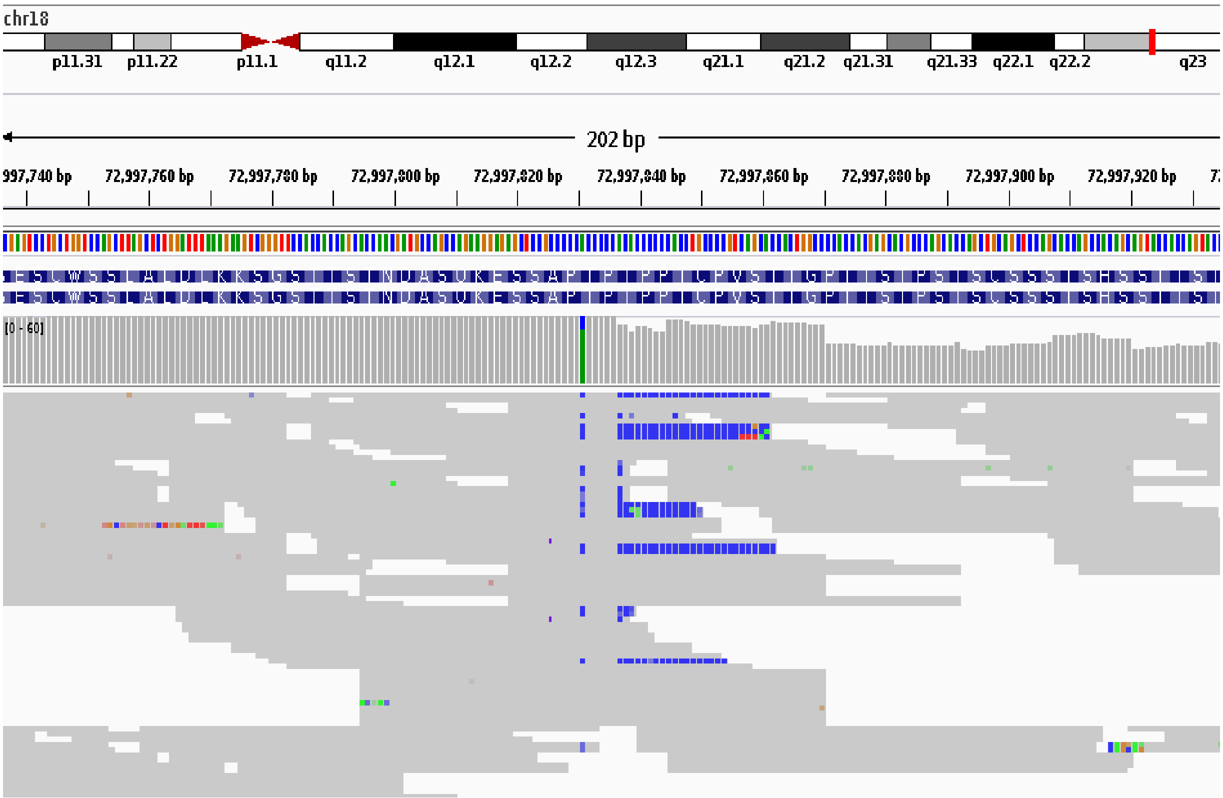


Figure 4.4. The misaligned base of the SR cluster in this IGV snapshot are all cytosines and are mapped to a locus that contains many successive cytosines. Given that sequences containing successive single bases are prone to replication error, it is possible that the SR cluster is a consequence of such errors and does not indicate an event that is carried by the proband.

4) The SR Cluster had fewer than 5 reads in the proband. In one proband, InDelible called an SR Cluster with 6 reads, which therefore passed the N SR >5 QC criterion. However, upon manual inspection I found that there was only one SR at the called position. I have not been able to find the reason for this discrepancy.

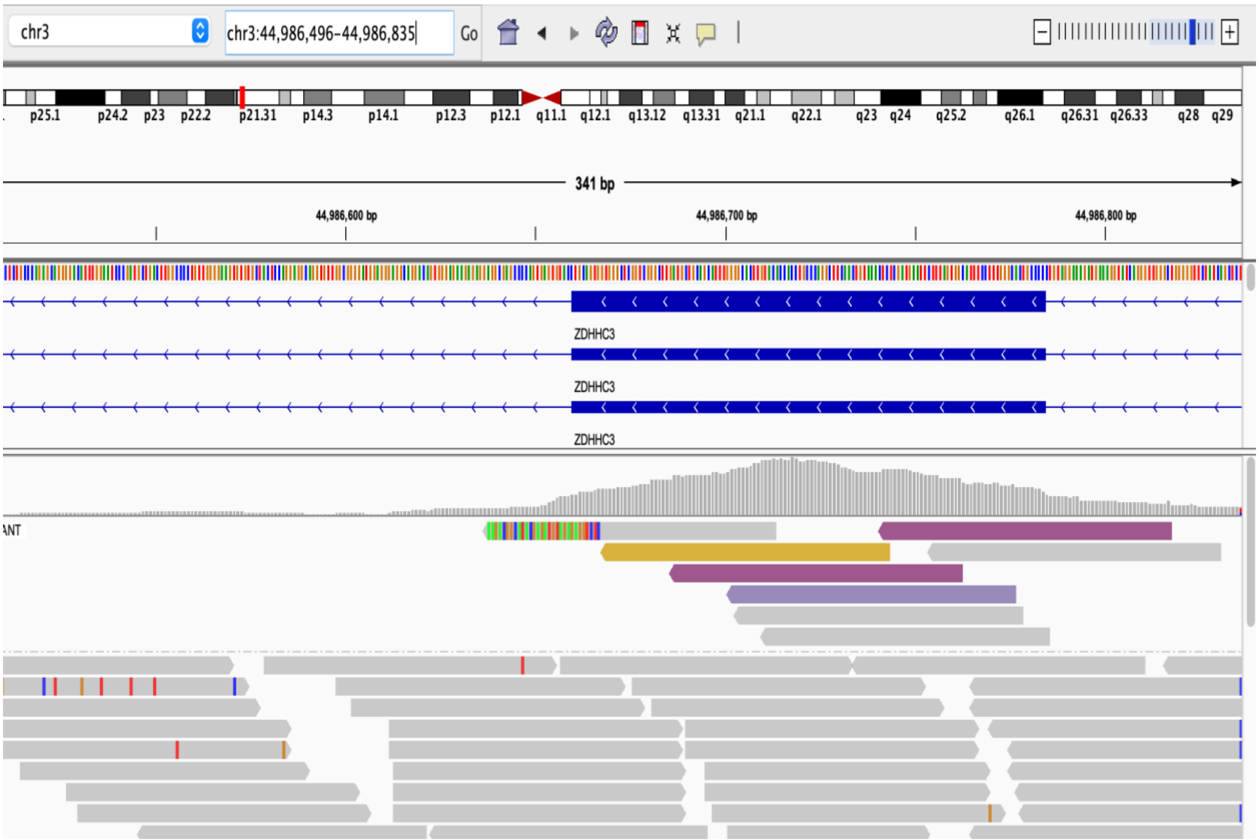


Figure 4.5. IGV snapshot showing a locus at which InDelible called an SR Cluster containing 6 reads. However manual inspection revealed there is only 1 read. The adjacent discordant reads may indicate a translocation, but the misaligned bases in the SR do not have a valid bwa alignment.

2.3.4 Quality control summary

Sample-level and variant level quality control are summarised in Tables 4.1 and 4.2, respectively.

N probands in initial output	604
N probands retained	556

Table 4.1. Summary of sample-level quality control. Samples were excluded for having too few or an excess of calls.

Quality control step	**N calls**
Initial output	268,526
Prob_Y > 0.6	-
Sample-level quality control	251,957
Allele-frequency < 0.01	24,822
MAPQ > 20	24,820
% double split filter	22,552
Exonic	14,115
N SRs > 5	97
N SRs in parents < 2	59
Manual inspection	49

Table 4.2. Summary of variant-level quality control. The call set fields show number of calls remaining after each quality control step was applied. Prob_y quality filter applied by default at the denovo processing stage so would have excluded 0 calls in the proband-only call set. MAPQ = mapping quality, SR = split reads

2.4 Structural variant annotation

In InDelible's Annotate step, SVs are annotated with the Ensembl gene they affect. I converted Ensembl IDs into their corresponding HGNC symbol using R package 'annotables (https://www.rdocumentation.org/packages/annotables/versions/0.1.1). I subsequently annotated the genes affected by SVs with SCZ case-control association statistics taken from the SCHEMA analysis ((Singh et al., 2022); described in section 1.8.2.2 of chapter 1); these included the number of PTVs and missense variants with a MPC ≥ 3 observed in the affected gene in SCZ cases and controls, and the SCHEMA meta p-value. I also annotated genes that were among 120 prioritised genes in the PGC3 SCZ GWAS ((Trubetskoy et al., 2022); criteria for prioritisation described in section 1.8.1.1 of Chapter 1). Finally, I annotated genes as being ‘loss-of-function intolerant’ if they had a probability of loss of function (pLI) score > 0.9 in the gnomAD database (Karczewski et al., 2020).

3. Results

3.1 Identification of putative de novo structural variants

After QC, 49 putative *de novo* split read clusters were identified in 15 probands; however, 36 of these clusters mapped to two pseudogene retrotransposons events. Therefore, a total of 15 putative *de novo* SVs were identified in 15 probands (2.4% of all probands; Table 4.3). These included 2 pseudogene retrotransposons, 7 deletions, 1 duplication, 1 insertion, 1 complex-insertion/deletion, 1 complex-insertion/duplication, and 2 SVs whose type could not be determined. The smallest SV was 19bp and the largest was 12.1kb. The mean and median sizes were 2,571.7bp and 51bp, respectively. Two chromosome 22 deletions were found to be instances of the same SV event in separate probands. Two chromosome 15 SVs, whose type I was unable to determine, were also found to be instances of the same event in separate probands.

Table 4.3 shows the 15 SV events along with their type and size, as determined through manual inspection and by InDelible. For pseudogene retrotranspositions, their genomic position is the 5’ and 3’ position of the pseudogene. I have given their size in exons rather than base pairs. Although InDelible assigned estimated SV sizes to most of the SR clusters for these events, these sizes correspond to the distance between a cluster and the next downstream exon (Figure 2.23), which are precisely those sequences that are not part of the retrotransposed sequence. The sum of these sizes cannot be meaningfully compared with the sizes I have determined through manual inspection, so I recorded the InDelible assigned size for the events as ‘NA’.

Chromosome	Position	SV Type, Manual Inspection	SV Type, InDelible	Size, Manual Inspection	Size, InDelible
2	133,427,484	Duplication	Duplication	20bp	20bp
4	5’: 15,504,140 3’: 5,518,240	Pseudogene retrotransposition	Deletion	5 exons	NA
5	138,163,329	Complex-insertion/duplication	Unknown	32bp	NA
11	33,360,930	Complex-insertion/deletion	Unknown	19bp	NA
11	35,684,964	Deletion	Unknown	21bp	NA
12	123,341,130	Deletion	Unknown	110bp	0
12	7,046,560	Deletion	Deletion	39bp	39bp
14	22,992,575	Deletion	Unknown	26bp	NA
14	10,541,1147	Insertion	Unknown	Unknown	NA
15	81,558,066	Unknown	Unknown	Unknown	NA
15	81,558,066	Unknown	Unknown	Unknown	NA
17	5’: 28,525,550 3’: 28,549,020	Pseudogene retrotransposition	Deletion	13 exons	NA
22	23,223,571	Deletion	Unknown	12493bp	NA
22	23,223,570	Deletion	Unknown	12493bp	NA
X	47,836,307	Deletion	Unknown	421bp	421bp

Table 4.3: Putative *de novo* events identified in the proband-only call set. The two call positions for pseudogene retrotranspositions correspond to the position for the 5’-most and 3’-most calls that were associated with these events. SV type and size determined through manual inspection and by the InDelible algorithm are included for comparison. The InDelible SV type for the pseudogene retrotranspositions is the mode of the type assigned to their constituent calls. Bp = base pairs.

Of the 4 SV calls that InDelible could assign an SV type, 2 of the types were confirmed by manual inspection, a deletion and duplication. InDelible incorrectly assigned the SV type ‘deletion’ to most of the constituent calls of the 2 pseudogene retrotranspositions, as it misidentified intronic regions between SR Cluster positions and their bwa alignments as deletions. InDelible estimated the size of 4 events, of which 2 sizes were confirmed by manual inspection: a 39bp deletion and a 421bp deletion. A duplication that was estimated by InDelible to be 20bp in size was found to be 63bp in manual inspection. InDelible estimated the size of a deletion as 0 as the bwa alignment for the call was upstream of its SR Cluster position. I determined the size of this event to be 110bp.

3.2 Examples of SR clusters for putative de novo structural variants

Figures 4.6-9 show the SR patterns for three *de novo* SV, as they appear in IGV: a deletion, a complex-insertion/duplication and 1 exon of a pseudogene retrotransposition. Reads from both parents at the same locus in are included below the probands sequencing reads, as evidence that the SV was *de novo*. Figure 4.6 shows an example of a SR cluster for which neither myself nor InDelible could assign a SV type.

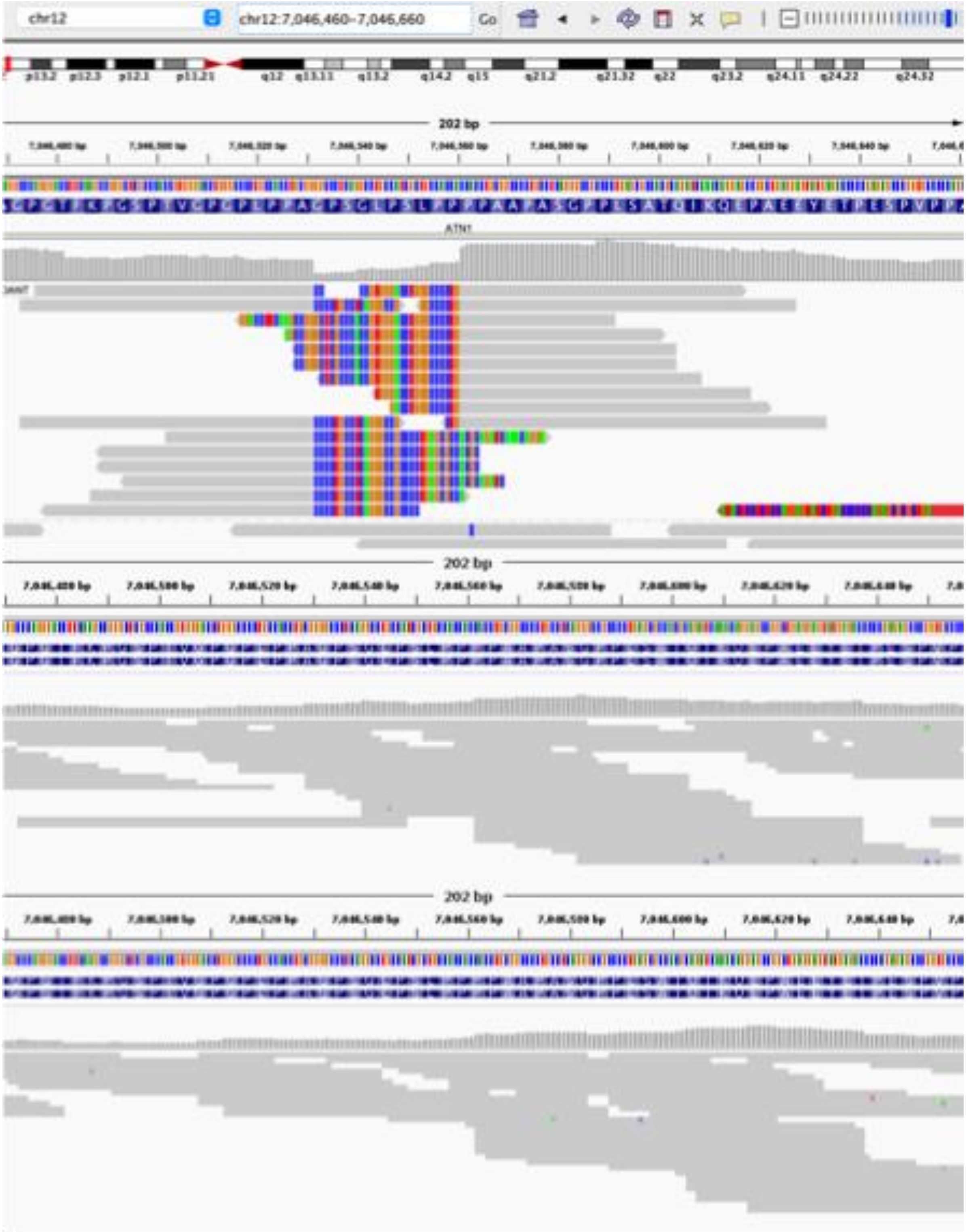


Figure 4.6: The topmost alignment track shows evidence of a 39bp deletion on chromosome 12, called in a proband. The two bottom alignment tracks show the same locus in each parent. Neither parent sample has any SRs, confirming the de novo status of this SV.

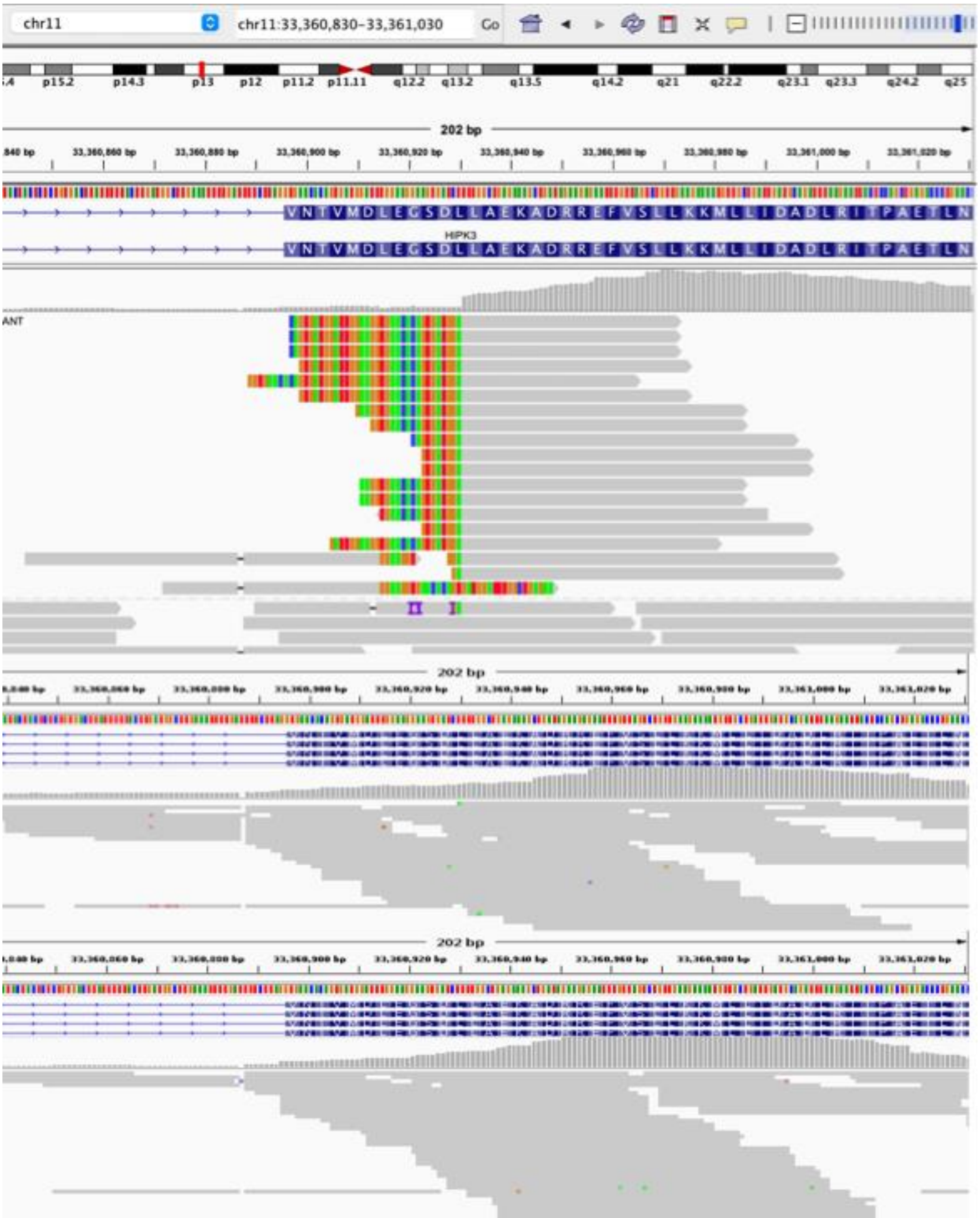


Figure 4.7: The topmost alignment track shows evidence of a 19bp complex-insertion/deletion on chromosome 12, called in a proband. The two bottom alignment tracks show the same locus in each parent. Neither parent has any SRs at this locus, confirming the *de novo* status of this SV.

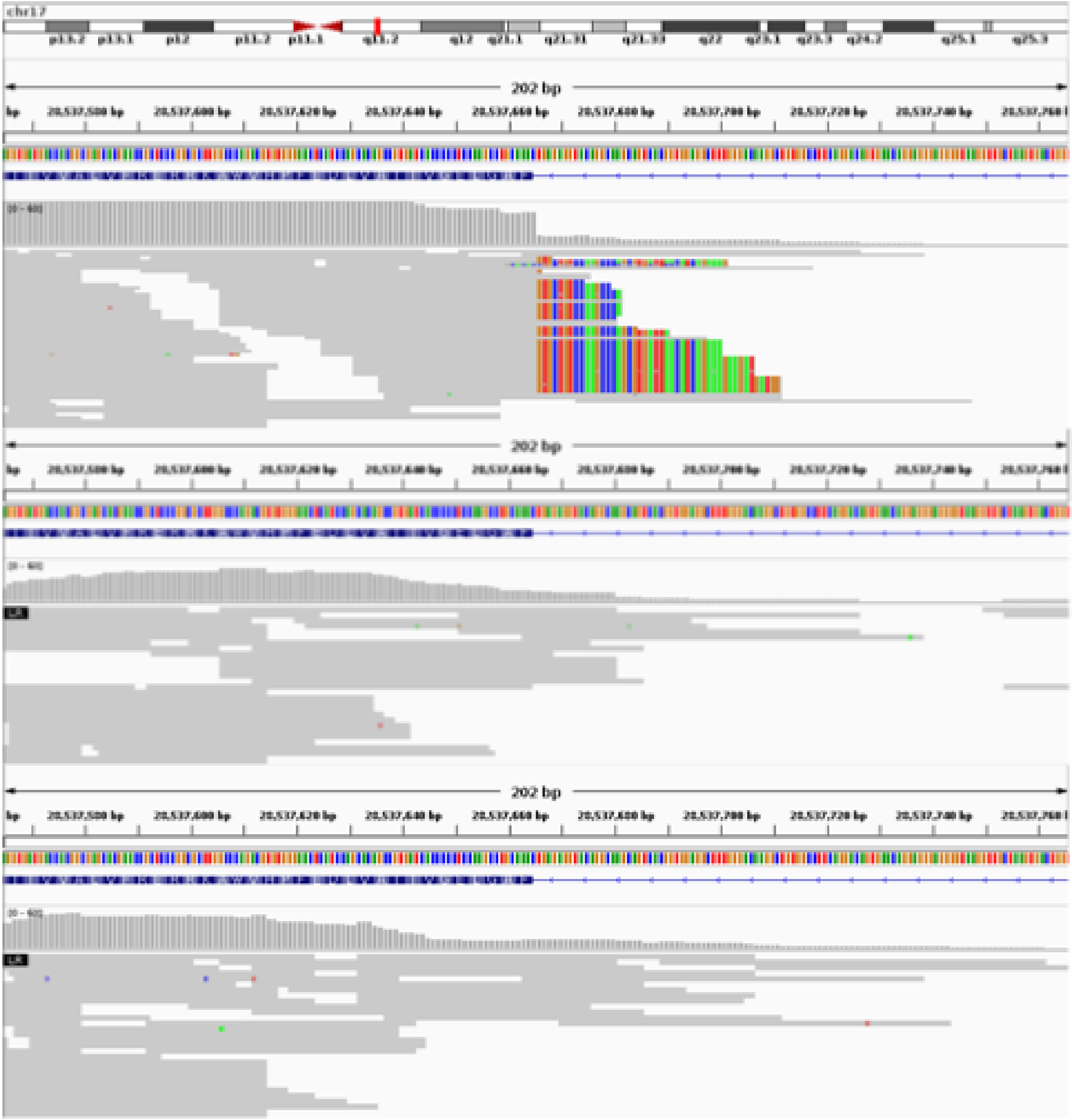


Figure 4.8: The topmost alignment track shows evidence of a pseudogene retrotransposition, called in a proband, and constituted of 13 exons of gene *SLC6A4*. The SR Cluster is one of 26 associated with this event. The two bottom alignment tracks show the same locus in each parent. Neither parent sample has any SRs, confirming the de novo status of this SV.

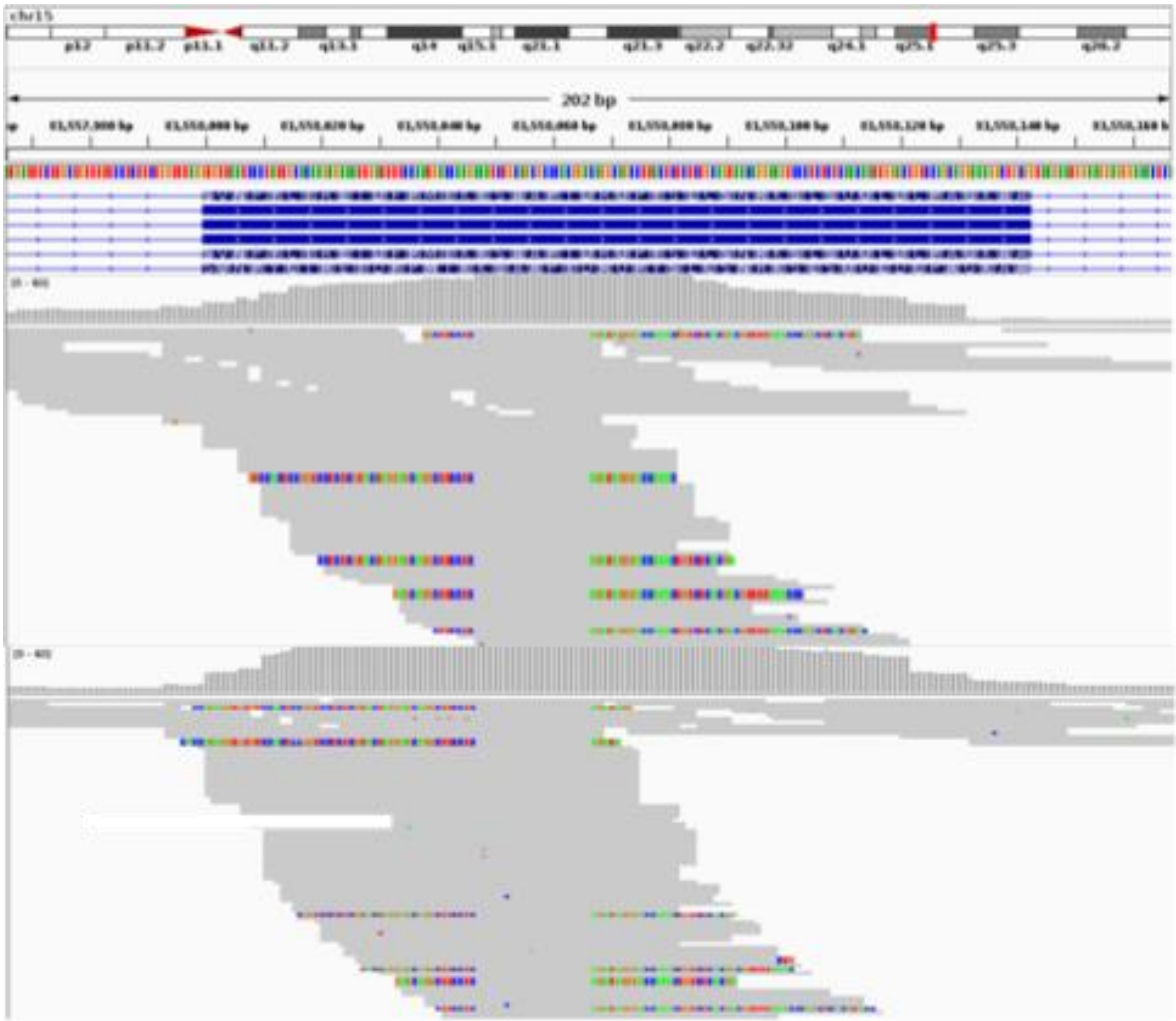


Figure 4.9: IGV snapshots of the same locus in 2 probands, showing evidence of 2 instances an SV whose type I have been unable to determine. It is characterised by a single SR cluster which splits at both ends.

3.3 Genes affected by de novo SVs

Twelve of fifteen putative *de novo* SVs were annotated by InDelible with the ENSEMBL gene they intersect. These genes are: *LYPD1, CC2D2A, CTNNA1, HIPK3, TRIM44, ATN1, HIP1R, AHNAK2, IL16, IL16, SLC6A4* and *ZNF81.* The three SVs that do not have a gene annotation were designated as ‘exonic’ by InDelible and intersect regions in the Ilumina Nextera exome capture kit. Two were the chromosome 22 deletions called at the same position and occur ~7kb upstream of *IGLL5.* The other was the chromosome 14 deletion, which occurs ~32kb upstream of *LINC02332*. The two pseudogene retrotranspositions were annotated with genes *CC2D2A (*chromosome 4) and *SLC6A4* (chromosome 17). Although the retrotransposed elements are composed of exons from these genes, however, they do not actually intersect the genes in the proband DNA. These gene annotations, therefore, cannot be used to acquire biological insights about the possible impact of the SVs themselves.

3.3.1 Previous evidence for association between genes affected by de novo SVs and schizophrenia

None of the genes affected by de novo SVs were previously enriched for PTVs and missense variants with MPC scores ≥ 3 in schizophrenia in the SCHEMA study with a P-value > 0.05 (Table 4.4), nor were they among the 120 genes that were previously prioritised as being likely to underpin schizophrenia GWAS common allele loci. However, two genes affected by de novo SVs are loss-of-function intolerant: *CTNNA1* (pLI = 0.97) and *ATN1* (pLI = 1). The SV intersecting *CTNNA1* is a complex-insertion/duplication, while the SV intersecting *ATN1* is a deletion. gnomAD proability of loss-of-function (pLI) results for all gene annotations, excluding those for pseudogene retrotranspositions, are show in Table 4.4.

Chromosome	**Position**	**SV Type**	**Gene**	**SCHEMA class 1 statistics**		**pLI**
				OR	**P meta**	
2	133427484	Duplication	*LYPD1*	1.46	0.506	0.25
4	15504140	Pseudogene retrotransposition	NA	NA	NA	NA
5	138163329	Complex-insertion/duplication	*CTNNA1*	0.535	0.639	0.97
11	33360930	Complex-insertion/deletion	*HIPK3*	0.463	0.489	1
11	35684964	Deletion	*TRIM44*	2.68	0.17	0
12	7046560	Deletion	*ATN1*	1.15	0.724	NA
12	123341130	Deletion	*HIP1R*	0.912	0.778	NA
14	22992575	Deletion	NA	NA	NA	0.55
14	105411147	Insertion	*AHNAK2*	1.46	0.194	0.05
15	81558066	Unknown	*IL16*	0.683	0.448	NA
15	81558066	Unknown	*IL16*	0.683	0.448	0

17	28525550	Pseudogene retrotransposition	NA	NA	NA	NA
22	23223571	Deletion	NA	NA	NA	0
22	23223570	Deletion	NA	NA	NA	0
X	47836307	Deletion	*ZNF81*	0.256	0.926	0

Table 4.4: SCHEMA association statistics and probability of loss-of-function (pLI) statistics for genes affected by *de novo* structural variants. Gene annotations for pseudogene retrotranspositions have been replaced with ‘NA’, as these events do not occur at their called positions. SV = structural variant, OR = odds ratio, pLI = probability of loss-of-function.

4. Discussion

4.1 Discussion of *de novo* structural variants discovered by InDelible

The primary aim of this study was to detect rare *de novo* SVs in 621 schizophrenia probands that could not be detected using array-based methods. Using InDelible, I identified 15 putative *de novo* events with AF < 1% in 15 (2.4%) probands: 7 deletions, 1 duplication, 1 insertion, 1 complex insertion-deletion, 1 complex insertion-duplication, 2 pseudogene retrotranspositions and 2 SVs whose type I couldn’t determine. The largest of these events is a 12.4kb deletion carried by two separate probands and is about half the size of than the smallest *de novo* CNV detected by CLAMMS (35kb). Moreover, while the size of this event is significantly smaller than the ~100kb resolution required for reliable detection using array-based methods, it is possible that it could have been called from array data for this sample but removed following a CNV size QC filter. As I reported in chapter 2, the smallest putative *de novo* SV detected from the array data was a 19kb duplication (which was found using CLAMMS to be a transmission). However, it is extremely unlikely that a region smaller than ~50kb is targeted by enough array probes to differentiate true CNV signals from noise. The next largest *de novo* SV detected by InDelible was a pseudogene retrotransposition that constituted of 13 partially transcribed exons. Given that the average length of a human exon is 150bp, this SV is ~2kb.

In effect, pseudogenes are an additional copy of each of the exons of the genes from which they were generated. While technically the probes in microarrays that overlap exons associated with pseudogenes might be able to detect the extra copy of this coding region, the size of the exons, and the distance between the exons, would preclude pseudogenes retrotranspositions being accurately called from array data. It is possible that these events could be called by CLAMMS, however, as in WES data they would also be associated with an increase in coverage depth at each of their constituent exons. Given that CLAMMS only assess coverage at exons, its calling algorithm could extrapolate the impacted sequence to include intronic regions too, such that it would call a duplication whose size spans the entire sequence from the first exon in the pseudogene to the last. In the present study, however, neither of the *de novo* pseudogene retrotranspositions were also called by CLAMMS.

Unless insertions are already known (as in the case of MEIs), and therefore can be targeted by an array, this SV type cannot be detected in array data either. A fragmented insertion would either not hybridise any array probes or would off-target hybridise to probes with which it happens to have sequence similarity. Complex events involving insertions would therefore be undetectable in array data too, as there would be no way to differentiate their component events using probe signal alone. In summary, all *de novo* SVs detected by InDelible were not detected in the array data due to their small size or type. However InDelible cannot be used to validate small CNVs called on array data, as its sensitivity to events larger than 500bp is < 0.1

4.2. De novo structural variants with unknown type

InDelible was not able to determine the type of 2 de novo SVs, and I could not determine the type either through manually inspecting the sequence reads in IGV. They were both called on chromosome 15 at position 81,558,066 and have the same SR pattern (Figure 4.9), involving a single SR Cluster that is split at both its 5’ and 3’ ends. The reads are mapped to an exon of gene *IL16* and the split positions occur close the centre of the exon. The misaligned bases of the second cluster align to 161kb contig GL000220.1. It is possible that these events indicate pseudogene retrotranspositions in which the retrotransposed element is a partially or alternatively spliced form of the IL16 exon, which has been inserted into an instance of GL000220.1.

4.3 De novo structural variants with unknown size

I was unable to identify the sizes of three de novo events: 2 SVs of unknown type on chromosome 15, and a simple insertion called on chromosome 14. As I could not determine the event type in former cases, I was unable to infer their size. In the latter case, there are no reads in which the insertion is nested such that its size and constituent bases were identified during alignment, and none of the misaligned bases in the single SR cluster align to any adjacent sequence (Figure 4.10). There is thus no method, based on the alignment data available, to identify the second breakpoint position of the insertion, and hence its size.

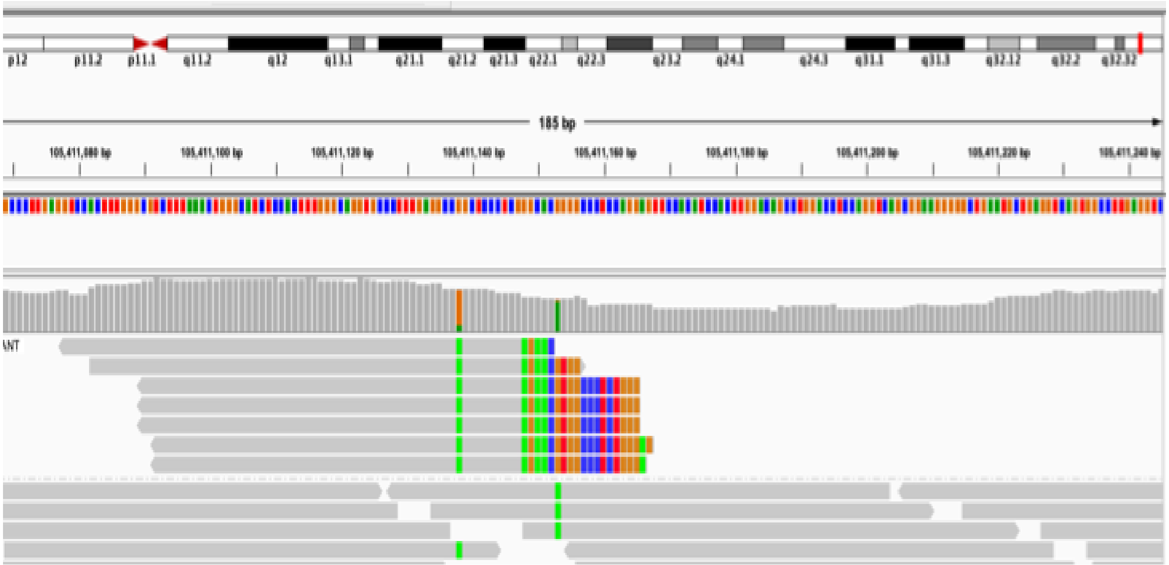


Figure 4.10. IGV snapshot showing evidence of a simple insertion on chromosome 14, whose size I could not determine.

4.4 Comparison of SVs in the current study with those presented in the DDD study

The number of *de novo* SVs I identified in the SCZ probands is too few to enable a statistically meaningful comparison between the size distributions and variant types of these events with those identified in the original DDD study (cite Gardner et al. 2021). Moreover, restriction of calls to those that intersect DD-risk genes in the DDD study may have resulted in an over-representation of variants in a particular size range or type. However, I can still assess whether the size and SV types of the *de novo* SVs is broadly commensurate with the findings of the DDD study.

The size distribution of the 15 de novo SVs is similar to that of the DDD call set, though more skewed toward larger SVs. No variants were <10bp, 6 (40%) were 11-50bp, 6 (40%) were >50bp and 3 (20%) were of unknown size. My results indicate that InDelible is sensitive to *de novo* variants in the 11-50bp size range and insensitive to de novo variants <10bp, which is in line with the findings reported in the original InDelible study. In combination with the results of chapter 3, my results also indicate that InDelible is insensitive to de novo variants >20kb in size, as InDelible detected 0/9 of the >20kb de novo CNVs I identified in the same sample using CLAMMS.

As in the DDD call set, about half of the *de novo* SCZ SVs were deletions (7/15, 46.8%). I detected one duplication, one insertion and two complex insertion events, but no mobile element insertions or translocations. The InDelible authors reported no pseudogene retrotranspositions in their call set, as they only assessed SVs whose gene intersects could be ascertained. In summary, the size distribution, and types of the *de novo* SVs are broadly commensurate with those reported in the DDD study: 40% were 11-50bp in size, the range InDelible is most sensitive to detect, and 4/5 SV types were also reported among the novel DDD candidate SVs.

4.5 Genes impacted by de novo SVs and schizophrenia risk

A secondary aim of this study was to assess whether any identified *de novo* SVs intersect known SCZ-risk genes. My initial source for SCZ-risk genes was SCHEMA, a meta-analysis of PTV and missense mutations identified in schizophrenia cases, controls and trios across multiple WES studies. None of the genes affected by putative de novo SVs in the SCZ probands were nominally associated with SCZ case status in SCHEMA. I also found that of the genes affected by de novo SVs were among the 120 genes prioritised by the PGC3 SCZ GWAS. Finally, as loss-of-function intolerant genes are enriched for SCZ-risk variants, I investigated whether any of the gene intersects had a pLI > 0.9 according to the gnomAD database. I found that *ATN1* and *CTNNA1* had pLI scores of 0.97 and 1, respectively. The SVs that intersect these genes are a 19bp deletion (*ATN1*) and a 32bp complex-insertion/duplication (*CTNNA1*).

ATN1 and *CTNNA1* are both plausible schizophrenia risk genes. The former encodes Atrophin 1, which is especially enriched in sub-cortical brain regions including amygdala, hippocampus, and hypothalamus (Palmer et al., 2019). Higher expression of *ATN1* has been observed in foetal brain tissue, suggesting a role for this gene in neurodevelopment (Palmer et al., 2019). Further evidence for a neurodevelopmental role is the epigenetic regulation of ATN1 expression by lysine-specific demethylase 1 (LSD1), which has been reported to control the differentiation of neural progenitor cells (Zhang et al., 2014). Downregulation of ATN1 causes early differentiation of these cells (Zhang et al., 2014). These findings are significant as multiple schizophrenia risk genes have also been shown to be preferentially expressed in foetal brain tissue (Cameron et al., 2023), and to play a role in neuronal stem cell differentiation (Iannitelli et al., 2017).

Exon 5 of *ATN1* contains a CAG repeat region, which when expanded to ≥48 copies is known to cause Dentatorubral-pallidoluysian atrophy (DRPLA) (Carroll et al., 2018). The extended polyglutamine tract that results in the Atrophin 1 protein is thought to impede protein-protein interactions and thus has a similar functional consequence as haploinsufficiency (Ross, 2002). The core symptoms of DRPLA are ataxia and cognitive impairment (Carroll et al., 2018). In teenaged and early adulthood onset, psychiatric symptoms such as irritability, depression, and psychosis have also been reported (Carroll et al., 2018).

In the context of schizophrenia, these findings are significant as SCZ most commonly develops in early adulthood, and psychosis is among its hallmark symptoms. The complex-insertion/deletion I identified also occurs in exon 5, ~800bp downstream of the CAG repeat. This exon is very large (~1kb), and so the *de novo* SV is unlikely to have the same impact on protein function as the CAG repeat expansion. However, insofar as disruption of this exon can cause SCZ-like symptoms in people with DRPLA, it is possible from my findings that this SV could also play a role in SCZ more generally.

CTNNA1 encodes the cell adhesion protein Catenin alpha-1 (α-cat). α-cat is most highly expressed in epithelial and muscle tissue but is also expressed throughout the brain where it plays a role in synapse formation and maintenance (Arikkath & Reichardt, 2008). Florescent-tagged α-cat localizes to dendritic spines and axons (Chiarella et al., 2018). In knock-models, dendritic spines are misshapen and more motile than in wild-type, with disorganized filopodia along the synaptic cleft, resulting in dysfunctional synapses that are less responsive to local signal changes (Arikkath et al., 2009). Aberrant synapse formation has been widely reported in schizophrenia (Glausier & Lewis, 2013). Many SCZ-risk genes encode proteins that localize to the post-synapse in particular, and are directly involved in the regulation of synaptic activity (Fromer et al., 2014).

In α-cat knock-out mice, cerebellar ataxia has been reported, along with deficits in fear-potentiated startle response (Park et al., 2002). The latter phenotype is observed in schizophrenia cases and has been used to indicate a SCZ-like phenotype in multiple animal models. I have not found studies reporting the phenotypic consequences of *CTNNA1* disruption in human, but given its function and behavioural consequence of knock-out in mice, it is a plausible schizophrenia risk gene.

Experimental validation of the SV calls and independent replication in case-control data is required before they can be considered novel SCZ risk factors. Nevertheless, my results support the utility of using WES data to identify small SVs that may contribute to schizophrenia liability. In summary, there is evidence that *ATN1* and *CTNNA1* could be schizophrenia risk genes. They both play roles in brain that have been implicated by known SCZ-risk genes, and SCZ-like phenotypes have been observed in humans or mouse models in cases of gene disruption.

4.6. Limitations of study design

The main limitation of this study is that I have not experimentally validated the putative de novo variants that were identified using InDelible. It was not possible to use array data from the trios analysed in this study to validate the de novo SVs, given the size and/or types of these SVs are not detectable from array data. I originally intended on validating InDelible de novo SV calls using PCR, which is how putatively pathogenic variants detected in the DDD study were validated. However, due to significant limitations imposed on laboratory research during the COVID-19 pandemic, this was not possible. Nevertheless, SVs discovered using InDelible had high validation rates in the original DDD study, where for 23 variants that the DDD study authors were able to obtain PCR results for, 100% were found to be true positives. As my quality control procedures were very similar to those implemented in the DDD study, these results suggest that I can have a high degree of confidence that the SCZ trios *de novo* call set are real events.

The dearth of studies assessing the rate of rare, small SVs in schizophrenia case/control data also limits the current study's ability to predict pathogenic *de novo* SVs. Given that rare CNVs that are smaller than 500kb, but larger than those detected by InDelible, have been found to have a more limited clinical impact than large events (Hollenbeck et al., 2017), it can be reasonably assumed that the impact of variants <500bp in size will be more modest still. However, I hypothesised that SVs which impact schizophrenia risk genes, based on findings from GWAS or sequencing studies, are more likely to have a role in schizophrenia. While none of the *de novo* SVs discovered in the current study affected genes robustly associated with schizophrenia the in SCHEMA or the PGC3 GWAS, 2 SVs did disrupt loss-of-function intolerant genes, and therefore might have pathogenic effects. However, these findings are preliminary and require replication in larger case-control studies.

4.7 Summary

In this chapter I have presented research that involved using the InDelible algorithm to detect small, rare *de novo* SVs in a WES trios sample consisting of 621 schizophrenia probands and both parents. I identified 15 putative *de novo* SVs in 15 (2.4%) probands, ranging in size from 19bp to 12kb. I then assessed the possible pathogenicity of these SV by ascertaining whether they intersected genes that have been previously associated with schizophrenia. I found that two *de novo* variants, a 19bp deletion and a 31bp complex-insertion/duplication intersected genes with pLI > 0.9: *ATN1* and *CTANN1*, respectively. Both genes have roles in the activity of the post-synapse, and their disruption has been associated with phenotypes relevant to schizophrenia, either in clinical cases or in animal models. Thus, they are both plausible candidate risk genes, suggesting the two *de novo* variants may contribute to the development of schizophrenia in their respective carriers. However, further genome-wide significant support from larger case-control studies of SVs is required before *ATN1* and *CTANN1* can be considered true SCZ risk genes. While the SVs identified in the current study require experimental validation via orthogonal methods, my results suggest that InDelible can be used to detect clinically relevant *de novo* small SVs in a small proportion of people with schizophrenia. By combining InDelible with an approach like CLAMMS, clinically impactful SVs across a very broad size spectrum can be detected using WES data.

Chapter 5: Combining whole exome sequencing and microarray data to identify rare CNVs impacting cognition in schizophrenia

1. Introduction

1.1 Cognition in schizophrenia

In addition to the core positive and negative symptoms, cognitive impairments have been widely reported in schizophrenia. Studies have identified deficits in attention, working memory, processing speed, problem solving, planning, abstract thinking, visual and verbal learning and social cognition (Bowie & Harvey, 2006; Heinrichs & Zakzanis, 1998; Lynham et al., 2018; Mesholam-Gately et al., 2009). A meta-analysis of 204 studies showed that cognitive performance among SCZ patients is at least 1 SD lower on several cognitive tests, especially memory and executive functioning (Heinrichs & Zakzanis, 1998). Cognitive symptoms are also extremely common, affecting up to 98% of patients according to some estimates (Keefe et al., 2005), and are among the earliest signs of disease onset (Häfner et al., 1992; Rund, 1998). The presence and severity of such impairments are highly correlated with poor functional, occupational and social outcomes (Green, 2006; Kraus & Keefe, 2007), and are the symptom dimension that is least responsive to therapeutic interventions (Tripathi et al., 2018).

Cognitive symptoms in SCZ are associated with both common and rare genomic factors (Calafato & Bramon, 2019; Creeth et al., 2022; Hubbard et al., 2021; Smeland et al., 2017), and genomic factors that impact loss-of-function intolerant genes and neurodevelopmental risk genes are particularly enriched. Among rare variants, large copy number variants (CNVs) have been found to impact cognition in SCZ. CNVs that been shown to confer risk for SCZ are associated with lowered cognition in both SCZ cases and in the general population. However, the literature is inconsistent regarding the magnitude of their impact.

1.2 Impact of CNVs on cognition in schizophrenia.

(Hubbard et al., 2021) identified rare (<1% frequency) CNVs >10kb occurring in 11 SCZ-risk loci (described in (Rees et al, 2016), and shown in Table 1.1 of chapter 1) in 15 participants of the Cardiff Cognition in Schizophrenia (COGS) cohort, a UK-based case sample consisting of 875 SCZ individuals tested for general and premorbid cognitive ability. CNVs either overlapped risk loci by ≥50% or affected a critical gene. The authors report an association between SCZ-risk CNV carrier status and decreased general cognitive ability (β = −0.66, p = .047), which they replicated in an independent Irish sample comprised of 679 SCZ cases (β = −0.91, p = .025), of whom 7 carried a SCZ-risk CNV. In addition, premorbid cognition was strongly affected among SCZ-risk CNV carriers (β = −7.16, p = .008). Secondary analyses were carried out exploring the impact of SCZ risk-CNVs across seven cognitive domains: attention, working memory, reasoning/problem solving, speed of processing, visual learning, verbal learning, and social cognition. Patients with SCZ who were carriers of CNVs were more impaired across all domains, compared to SCZ patients with no such CNVs, with the differences reaching about 0.5SD for the different cognitive domains tested.

The authors also found that the burden of CNVs >10kb that impact loss-of-function intolerant (LoFi) genes are associated with general cognitive deficits, even after controlling for the impact of the 12 SCZ-risk CNVs (β = −0.15, p = .048). Deletions affecting LoFi genes had a stronger effect than duplications (deletions: β = −0.21, p = .055; duplications: β = −0.05, p = .513). Deletions affecting genes that code for synaptic proteins were also associated with lowered general cognition (β = −0.22, p = .035) when covarying for SCZ-risk carrier status. Finally, burden of all CNVs >100kb in size was tested for association with general cognition, according to three metrics: n CNVs, total length of CNVs, and N genes impacted by CNVs. No association was found with any burden metric.

(Thygesen et al., 2021) identified 29 rare (<1% frequency) SCZ-risk CNVs >100kb in 29 participants of the Psychosis Endophenotypes International Consortium (PEIC) family study, comprising 749 individuals diagnosed with a psychotic disorder (576 with SCZ (74.9%)), 646 of their unaffected relatives, and 2013 non-relative unaffected controls. The authors defined SCZ-risk CNVs as those affecting one of 27 loci with ‘good evidence’ of an association with SCZ, as described by (Marshall et al., 2017), (Kirov et al., 2014), and (Stefansson et al., 2014a). Only those CNVs that overlapped a SCZ-risk loci by >10% were included. Across the whole sample, SCZ-risk CNV carriers demonstrated deficits in immediate (β = −8.0, p = 0.0036) and delayed (β = −3.3, p = 0.0115) verbal recall, measured by the ability to repeat 15 words that were read to participants at rate of 1 word per second, either immediately or after 30-minute delay. SCZ-risk CNV carrier stratus was also nominally associated with poorer block design score (β = −10.0, p = 0.031), a measure of visuospatial reasoning.

In the same study, when restricting the analysis to individuals with a diagnosis of SCZ, their relatives and controls, the association between SCZ-risk CNV carrier status and immediate verbal recall remained at the same significance level (β = -8.39, p = 0.004), though the association with delayed verbal recall was weaker (β = -3.10, p = 0.025). The association with block design score was also slightly weaker (β = -9.787, p = 0.046). The authors also tested the burden of rare (<1% frequency) CNVs >200kb for association with cognition but reported no significant results.

(Foley et al., 2020) found increased cognitive deficits among SCZ patients who carried rare (<1% frequency) SCZ-risk CNVs >20kb, defined according to 15 loci described in (Rees et al., 2014). The case sample comprised 1215 Irish individuals, of whom 19 carried a SCZ-risk CNV. Specifically, the authors identified three phenotypic variables that were significantly associated with carrier status: ‘history of developmental delay’ (OR = 5.19, p = 0.003), ‘comorbid neurodevelopmental disorder’ (OR = 5.87, p = 0.009) and ‘specific learning disorder’ (OR = 8.12, p = 0.012). Collectively, these results suggest a neurodevelopmental basis for the observed cognitive deficits, and that SCZ-risk CNVs impact cognitive ability prior to disease onset.

(van Scheltinga et al., 2013) tested whether two CNV burden metrics: the total number of CNVs and the total number of genes affected by CNVs, were associated with IQ in 350 SCZ patients and 322 controls. Though the authors do not specify a base pair size threshold, they did not consider events that spanned fewer than 10 consecutive probes. Also, they did not filter CNVs by allele frequency. No result was statistically significant.

In summary, there is clear evidence that SCZ-associated CNVs, and deletions affecting LoFi genes, cause deficits in the cognitive performance of patients affected with SCZ. These deficits appear to occur before disorder onset, and lead to worse cognitive performance compared to SCZ patients do not carry a CNV.

1.3 Impact of SCZ-risk CNVs on cognition in the general population

CNVs associated with increased SCZ risk have also been found to be associated with lowered cognitive ability in the general population, indicating that their impact on cognition might be partly independent of other genetic risk factors for SCZ. (Kendall et al., 2017) investigated the impact of rare SCZ-risk CNVs on cognition in 152,000 participants in the UK Biobank, a large data resource that is partially representative of the UK general population, although is older and healthier in general. Individual calls were filtered if they spanned <10 probes or had a density coverage of <1 probe per 20k base pairs.

CNVs were defined as SCZ-risk if they overlapped any of the 11 loci described by (Rees, 2016b). The authors reported significantly impaired performance in carriers of SCZ-risk CNVs, compared with CNV non-carriers, across 7 tests designed to evaluate the following cognitive domains: episodic memory, processing speed, reasoning, numeric working memory and visual attention. SCZ-risk CNV carriers were less likely to complete higher education (OR = 0.61, $p = 2.4 \times 10^{-18}$), and tended to have occupations that require fewer academic skills (OR = 0.64, $p = 3.7 \times 10^{-11}$). The authors also found that SCZ case status (not all of whom were SCZ-risk CNV carriers) was significantly associated with lowered cognition across all 7 domains, and cognitive impairment was significantly greater for SCZ cases than for SCZ-risk CNV carriers, indicating that other risk factors or disorder progression itself may impair cognition beyond the impact of known CNV risk factors.

(Stefansson et al., 2014b) investigated the impact of schizophrenia risk CNVs on cognitive function in an Icelandic sample with no schizophrenia diagnoses. Several CNVs were found to negatively impact cognition, including 15q11.2 deletion, 16p11.2 deletion, 1q21.1 duplication, and 1q21.1 deletion. However, the effects of some CNVs were limited to specific cognitive domains. For the 15q11.2 deletion carriers, performance on reading and arithmetic test was significantly lower compared to non-carriers (reading: $p = 3.2 \times 10^{-3}$; arithmetic: $p = 1.6 \times 10^{-3}$), while impacts on other domains more modest. There was also higher prevalence of dyslexia and dyscalculia among carriers of this CNV, even after adjusting for IQ, suggesting that it contributes to language and numeracy deficits independently of general cognitive ability. Similarly, carriers of 16p11.2 deletion showed significantly reduced performance in verbal memory ($p = 3.4 \times 10^{-3}$) and processing speed ($p = 5 \times 10^{-4}$).

1.4 CNV detection methods

Previous studies used CNV detection methods based on data generated from genotyping microarrays (hereafter referred to as arrays) with the number of probes ranging from 550K-1.1M (Table 5.1). In section 1.2 of chapter 3, I discussed that a limitation of these approaches is the ability to differentiate true deviations in probe signals caused by CNVs from noise at base pair resolutions < ~100kb, and it has been estimated that even high-density (>1M probes) array platforms are unable to reliably detect CNVs < 40kb in size (Carter, 2007).

Study	Genotyping arrays used	N probes
Hubbard et al., 2021	Illumina HumanOmniExpressExome-8v1	951,117
Thygesen et al., 2020	Affymetrix Human SNP Array 6.0	946,000
Foley et al., 2020	Affymetrix Human SNP Array 6.0	946,000
	Illumina HumanCoreExome	542,586
van Scheltinga et al., 2013	Illumina HumanHap550 beadchip	550,000
Kendall et al., 2016	UK Biobank Axiom Array	820,967
	UK BiLEVE array	807,411
Stefansson et al., 2014	Ilumina Human610-Quad	610,000

Table 5.1. Number of probes on the genotyping arrays used in the 6 studies described in sections 1.2 and 1.3.

However, there is evidence that whole exome sequencing (WES) technology can successfully detect CNVs missed by genotyping arrays in protein-coding regions, particularly for smaller CNVs (Zhao et al., 2013). In chapter 3, I reported that ~81% (143/177) of transmitted CNVs that were called in a trio sample using the WES-based CNV calling algorithm CLAMMS were <100kb in size, and of these 47 were <10kb in size. Therefore, exome-sequencing studies have the potential to advance our understanding of the genetics of cognition in SCZ by analysing CNVs that are typically missed in arrays based CNV studies.

1.5 Study aims

The primary objective of the research presented in this chapter, which was conducted in the second and third years of my PhD, was to use CNV calls generated by CLAMMS to further our understanding of the genetic contribution to cognitive impairment in schizophrenia, progressing the broader thesis aim 3 specified in section 3 of chapter 1. This required completing the following aims:

1) Use CLAMMS to call rare (>1% frequency) CNVs from WES data for 875 individuals with schizophrenia or a related psychotic disorder, recruited in the Cardiff Cognition in Schizophrenia (COGS) study (the same cohort analysed by (Hubbard et al. 2021)). Each participant of Cardiff COGs has been assessed for general cognitive ability (hereafter referred to as ‘current cognition’) and estimated premorbid cognitive ability, which were used as the primary cognitive phenotypes in this chapter.
2) Compare the sensitivity to detect known pathogenic or schizophrenia-risk CNVs between WES and array CNV call sets.
3) Assess whether analysing a consensus CNV call set based on both WES and array-based approaches produces a more accurate CNV call set and increases power to identify CNVs contributing to cognition in SCZ.
4) Explore the impact of small CNVs typically missed in array studies (i.e. CNVs (<100kb) on cognition in SCZ.
5) Determine if CNVs intersecting genes previously implicated in cognition in schizophrenia are associated with cognitive deficits in this sample. Variant sets were further refined to identify those that affect loss-of-function intolerant

(LoFi) genes and neurodevelopmental disorder (NDD) risk genes, given the strong impact of variants affecting these genes on cognition in SCZ.

I hypothesised that large deletions affecting LoFi genes would have the largest impact on both cognitive metrics, and that the effects of small events and duplications would be more modest. Combining platform call sets may not affect any observed association of large events with cognition, given that both approaches are able to detect such events. However, if small CNVs (and particularly small deletions affecting LoFi or NDD-risk genes) do indeed impact cognition in SCZ, their effects are more likely to be captured by WES-based call sets than the array-based call sets.

2. Methods

2.1 Sample description

2.1.1 Recruitment

The Cardiff Cognition in Schizophrenia (COGS) cohort (Lynham et al., 2018) consists of 927 individuals recruited by mental health professionals from in-patient, out-patient, and volunteer mental health services across the UK. After initial screening of medical records, individuals were excluded if they had a previous diagnosis of intellectual disability, a neurological disorder known to impact cognitive ability, or a current substance abuse disorder. Participants were aged between 17 and 82 years at recruitment. Mean age was 43.3 and 60% of participants are male.

2.1.2 SCAN instrument

Participants were assessed using The Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al., 1990). SCAN consists of 22 segments designed to identify and rate symptom dimensions that occur among known neuropsychiatric disorders. It is divided into two parts, the first of which is concerned with general neurotic symptoms, anxiety, eating disorders and substance abuse. After a preliminary screening, interviewers can move to part two if appropriate, which assesses psychotic symptoms and disorders of affect, speech, and behaviour. Trained psychologists and psychiatrists carried out the assessments under the supervision of principal investigator James Walters.

2.1.3 Psychiatric diagnoses

If applicable, participants were then given a best-estimate lifetime diagnosis of a neuropsychiatric disorder based on SCAN outcomes and medical records, in line with DSM-IV criteria (Bell, 1994) (Table 5.2). Interrater reliability for diagnosis was strong; schizophrenia = 0.83, schizoaffective depressive = 0.63, schizoaffective bipolar = 0.72, bipolar disorder = 0.85 (Lynham 2018). In the present study, although CNVs were called in every sample to improve overall call quality, for downstream variant-phenotype association analyses individuals were excluded if they had no accompanying phenotypic information, or had a diagnosis of Mania, Bipolar Disorder, Major Depressive disorder or Other.

DSM-IV Diagnosis	N participants
Schizophrenia	598
Schizoaffective disorder (depressive type)	136
Schizoaffective disorder (bipolar type)	72
Other non-affective psychotic disorder	69
Mania	16
Bipolar disorder	5
Major depressive disorder	17
Other	4
NA	10

Table 5.2. Psychiatric diagnoses received by Cardiff COGS participants, according to DSM-IV criteria.

2.2 Assessing cognitive ability

2.2.1 Current cognitive ability

All participants had been tested for current cognitive ability. This was carried out using the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Consensus Cognitive Battery (CCB) (Marder & Fenton, 2004), which assesses cognitive performance across seven domains: processing speed, attention/vigilance, working memory, verbal learning, visual learning, reasoning/problem solving and social cognition. Description of all tests administered per cognitive domain are given in Table 5.3.

Cognitive domain	**Test**	**Description**
Processing speed	Trail making test, part A; BACS, symbol coding subtest; Category fluency test (animal naming)	Trail making test, part A: subjects draw a line between numbered circles in ascending numerical order, from 1 to 25, as quickly as possible. BACS symbol coding subtest: According to provided key, subjects are given 90 seconds to assign numbers to non-meaningful symbols. Category fluency test (animal naming): In 60 second, subjects produce as many examples as they possible of animal names.
Attention/vigilance	The Continuous Performance Test, Identical Pairs version	Observing changing symbols on a screen, subjects respond as quickly as possible when two identical symbols are presented in a row.
Working memory	Spatial span subtest of WMS, 3rd ed.; Letter-number span test	WMS spatial span subtest: Subjects observe squares in a grid as they change colour in a particular order and are then required to specify the order as quickly as possible.

		Letter-number span test: subjects listen to a series of letters/digits, then repeat series with letters in alphabetical order, or digits in ascending order.
Verbal learning	Hopkins Verbal Learning Test	Subject listens to utterances of 12 nouns, then is required to repeat the words back in any order, both immediately and then after a 25 minute delay.
Visual learning	Brief Visuospatial Memory Test	Subjects observe a visual display of 6 basic figures on a 2x3 grid, and are then required to draw each figure as accurately as possible in the correction locations on a new 2x3 grid, both immediately and then after a 25 minute delay.
Reasoning/problem solving	Neuropsychological Assessment Battery, mazes subtest	Consists of 7 printed mazes of increasingly difficultly that subjects are required to trace through as quickly as possible.
Social cognition	Mayer-Salovey-Caruso Emotional Intelligence Test, managing emotions branch	Questions are designed to test how effectively subject can regulate their own emotion in decision-making, and incorporate the emotions of others into their decision-making.

Table 5.3. Tests comprising the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Consensus Cognitive Battery and their corresponding cognitive domains.

Raw scores of all MATRICS tests were normalised to produce z-scores against the mean and standard deviation of 103 healthy controls, recruited across the UK and matched to case samples on sex and age (50% male, mean age = 41.7 years). For participants with test results across 5 or more domains (926 of 927 participants), composite z-scores were calculated according to MCCB manual procedures by Amy Lynham. These scores are an estimate of general, current cognitive ability.

2.2.2 Premorbid cognitive ability

Premorbid cognitive ability had been estimated using the National Adult Reading Test (NART) (Nelson, 1982). The NART consists of 50 words which participants are instructed to read aloud. As the words have irregular pronunciation, the test is designed to evaluate vocabulary rather than the ability to apply standard rules of pronunciation.

2.3. Sequencing and genotyping

2.3.1 Sample preparation, whole exome sequencing and genotyping.

DNA samples were extracted from whole blood. The exomes of 498 samples were isolated using the Nextera DNA Exome capture kit and sequenced on an Illumina HiSeq X platform at the Broad Institute (hereafter referred to as the Broad subcohort). The exomes of remaining 429 samples were also isolated using the Nextera DNA Exome capture kit but were sequenced at Cardiff University on an Illumina HiSeq 3000/4000 platform (hereafter referred to as the Cardiff subcohort). The GATK best practice pipeline was used to process raw paired-end sequencing reads, which were then aligned to human genome reference build 37 (GRCh37/hg19) with the Burrow-Wheeler Aligner (BWA) v0.7.15 (Li & Durbin, 2009). Genotyping was carried out at the Broad Institute, Massachusetts, on the HumanOmniExpressExome-8v1 combo array, consisting of 951,117 individual SNP probes.

2.3.2 Sequencing Coverage Depth

The coverage depth of a given sequenced base in a sample’s raw alignment data is the number of reads that are aligned to that base. Though coverage depth is primarily determined by sequencing platform, features of the genome itself can lead to large differences in coverage both within and between samples, as discussed in chapter 2. The mean target coverage depth metric for a sample is the mean coverage depth for all bases in the targeted region (exome). In the present study, sample coverage metrics were generated using Picard (http://broadinstitute.github.io/picard/). Across all samples sequenced at Cardiff University, mean target coverage depth is 32.4, while for those sequenced at the Broad Institute it is 84.3. Figure 5.1 shows the mean coverage depth for all samples in Cardiff COGS, coloured by sequencing site.

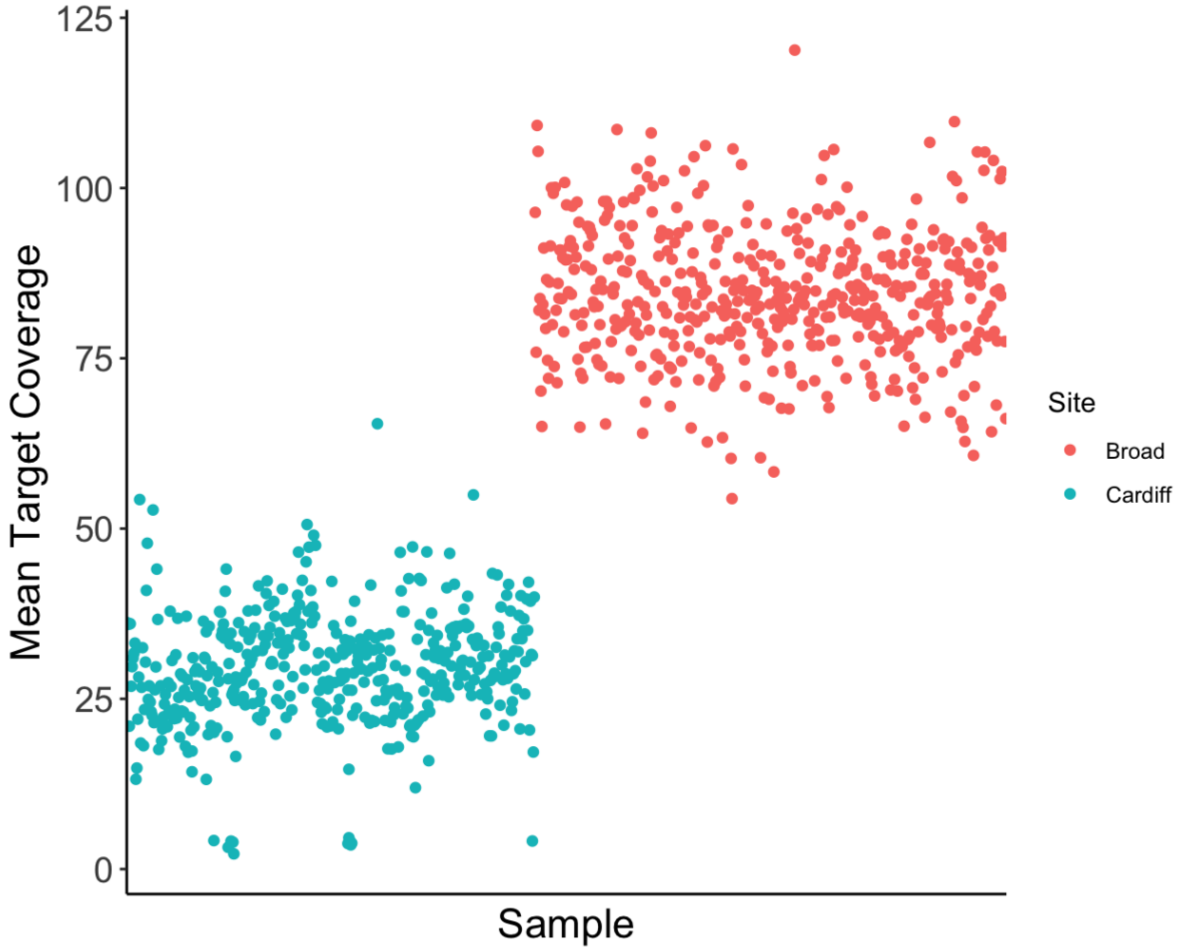


Figure 5.1: Mean target coverage for 927 Cardiff COGS WES samples, coloured by whether they were sequenced at Cardiff University (blue) or at the Broad Institute (red).

2.4 CNV calling

2.4.1 CLAMMS

I called CNVs from the WES data using the CLAMMS algorithm (Packer et al., 2016). CLAMMS estimates the copy number of a given exonic sequence in a sample by comparing its coverage depth to those of the same sequence in a reference panel of samples with similar quality metrics. A full description of the CLAMMS algorithm is given in section 2 of chapter 2. As there is a significant difference in mean target coverage between samples sequenced at each site, I called CNVs for Broad-sequenced and Cardiff-sequenced subcohorts separately.

Two aspects of the CNV calling process vary according to sequencing protocol and sample quality metrics and thus warrant description here. For purposes of calculating GC content, CLAMMS requires a user-defined insert size variable when generating the windows file. The authors recommend a size that is ‘a little bit bigger’ than the mean insert size for the sequencing process used, such that most reads will come from inserts of sizes smaller than this value. The mean insert size of the sequencing processes used in the present study was estimated by calculating the mean of the mean insert sizes across samples from each cohort, generated by the Picard command ‘CollectInsertSizeMetrics’. The mean insert size for the Cardiff and Broad sequencing processes is 164.7 and 376.1, respectively. A separate windows file was therefore generated for each cohort, using window sizes 200 for Cardiff subcohort and 400 for the Broad..

The second sample-specific aspect of CLAMMS is the accounting for batch effects, i.e. differences in sample preparation and input DNA quality that may introduce stochastic volatility and distort read coverage depth exome-wide. In CLAMMS, batch effects are controlled for by clustering samples into k reference panels based on 7 quality metrics generated by Picard (http://broadinstitute.github.io/picard), described in chapter 2. As it cannot be known *a priori* which value of k can adequately control for batch effects, in the current study 40 different values were applied for each sample, from 10-400 in increments of 10. CNVs were called for samples based on each reference panel size, and all QC steps were applied accept allele-frequency filters.

To determine the k that maximised the quality of CNV calls, I evaluated for each value of k the number of samples that failed QC, assuming that higher quality CNV calls will result in lower sample drop-out rates. Additionally, for each value of k, I calculated the number of WES-based CNV calls that overlapped a CNV called in the same sample from the array data, under the assumption that higher quality WES CNV calls are more likely to also be observed in the array CNV call set.

Figures 5.2 and 5.3 show the number of samples failing QC plotted against number of CNVs called in array data. In both Broad and Cardiff subcohorts, a strong correlation was observed between sample drop-out rate and the number of CNVs observed in both WES and array call sets. While CNV quality initially increases as reference panel size increases, this trend reverses for larger reference panel sizes. For the Cardiff subcohort, the optimal reference panel size was 130, while for the Broad subcohort it was 50. Correlation between reference panel size and CNV quality was lower for Cardiff-sequenced samples, suggesting a greater amount of stochastic volatility in coverage depth across samples in this cohort.

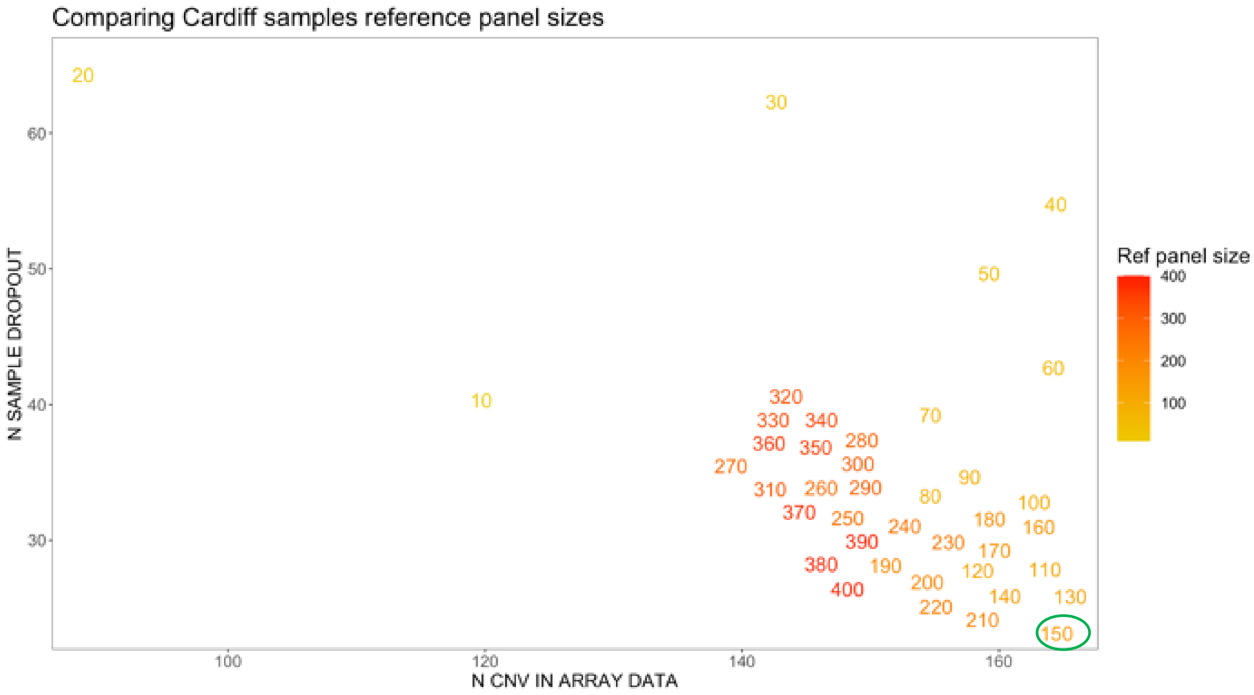


Figure 5.2. Performance of different reference panel sizes for Cardiff subcohort. A reference panel size (k) of 150 was chosen (circled), as it produced the lowest sample drop-out and the 2nd highest number of calls that were also identified in array data.

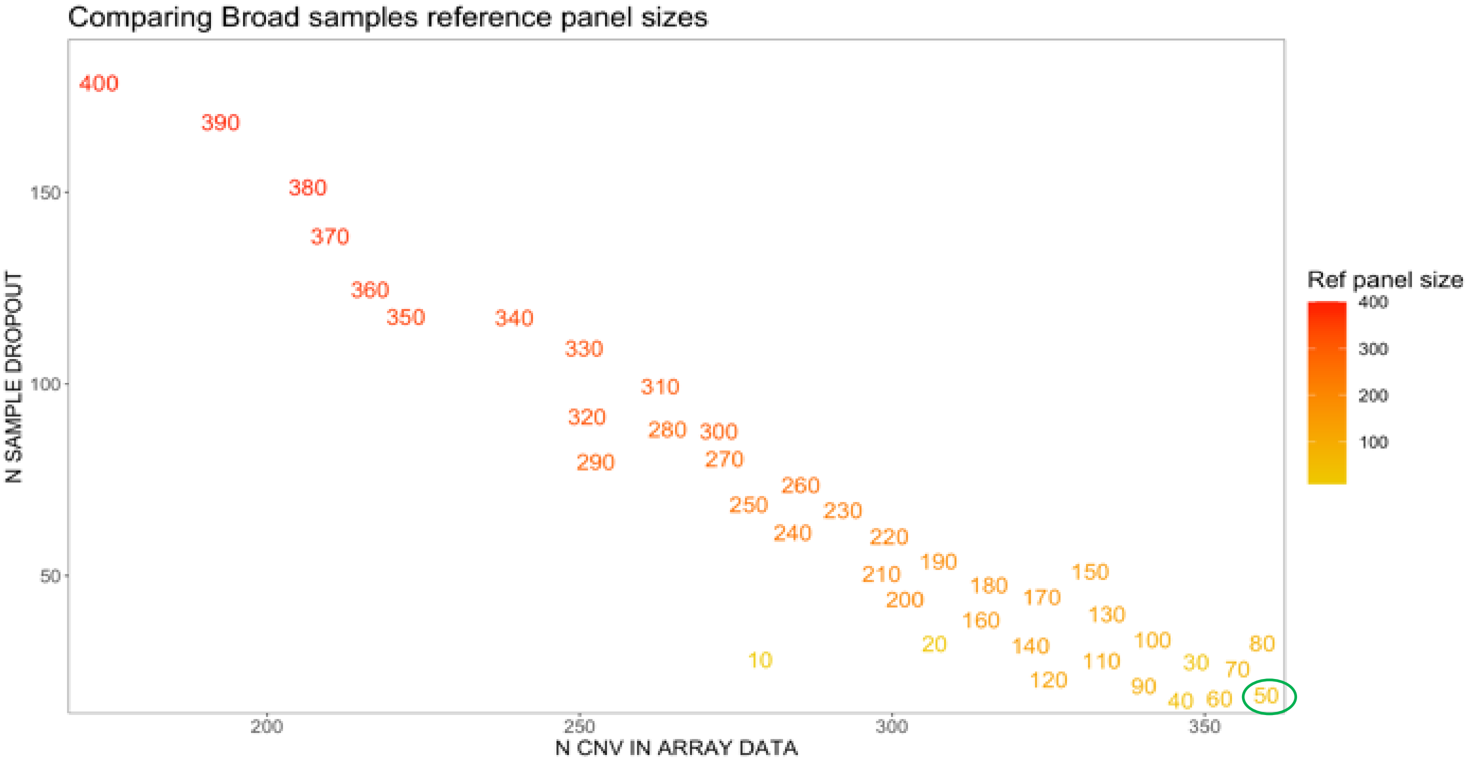


Figure 5.3. Performance of different reference panel sizes for Broad subcohort. A reference panel size (k) of 50 was chosen (circled), at it produced both the lowest sample drop-out and the highest number of of calls that were also identified in array data.

2.4.2 PennCNV

BeadStudio v2.0 was used to call genotypes, to normalize the signal intensity data, and to establish the logRratio (LRR) and B allele frequency (BAF) according to the standard Illumina protocols. Elliott Rees used PennCNV (Aug 2009 version (Wang & Bucan, 2008)) for CNV detection and conducted all QC according to protocols described in (Rees et al., 2014). Both unfiltered and high-quality calls with allele-frequency <1% were analysed.

2.5 Quality control

2.5.1 CNV Merging

For the calls generated from each cohort using the chosen reference panel sizes discussed in the previous section, one sample level and five variant level quality control (QC) criteria were applied, based on (Maxwell et al., 2017). First, separate calls for a sample that were the same CNV type and whose start and end coordinates were within 10kb were merged into a single call using bedtools' merge command, as it is unlikely that they are separate CNV events. In the raw output, 5,684 calls were generated for Cardiff subcohort, while 13,945 calls were generated for the Broad subchort. After merging, there were 5,684 and 13,665 calls, respectively.

2.5.2 Sample level quality control

The data used in the current study had previously undergone sample-level QC as described in (Creeth et al., 2022). This included excluding samples that did not have a diagnosis of schizophrenia, schizoaffective disorder, or other non-affective psychotic disorder (n = 52). Additionally, samples were excluded if their inferred sex did not match their expected sex or were in a second-degree, or closer kinship (Creeth et al., 2022). I applied the following additional sample-level QC that is based on the distribution of SVs called per individual by CLAMMS. The number of CNVs per sample followed a normal distribution (Figure 5.4). In both subcohorts, I decided to use 2 x median number of CNVs as the upper threshold beyond which samples could be considered outliers. A lower limit was not applied as there was no long tail at the left side of the distribution in either case and no sample had 0 CNV calls. For the Cardiff subcohort, the median n CNV calls was 9 and for Broad subcohort the median was 15. Figures 5.4 and 5.5 are histograms for N CNV calls, illustrating sample medians and upper thresholds. Twenty-five samples were filtered from the Cardiff-sequenced subcohort and 22 from the Broad-sequenced subcohort.

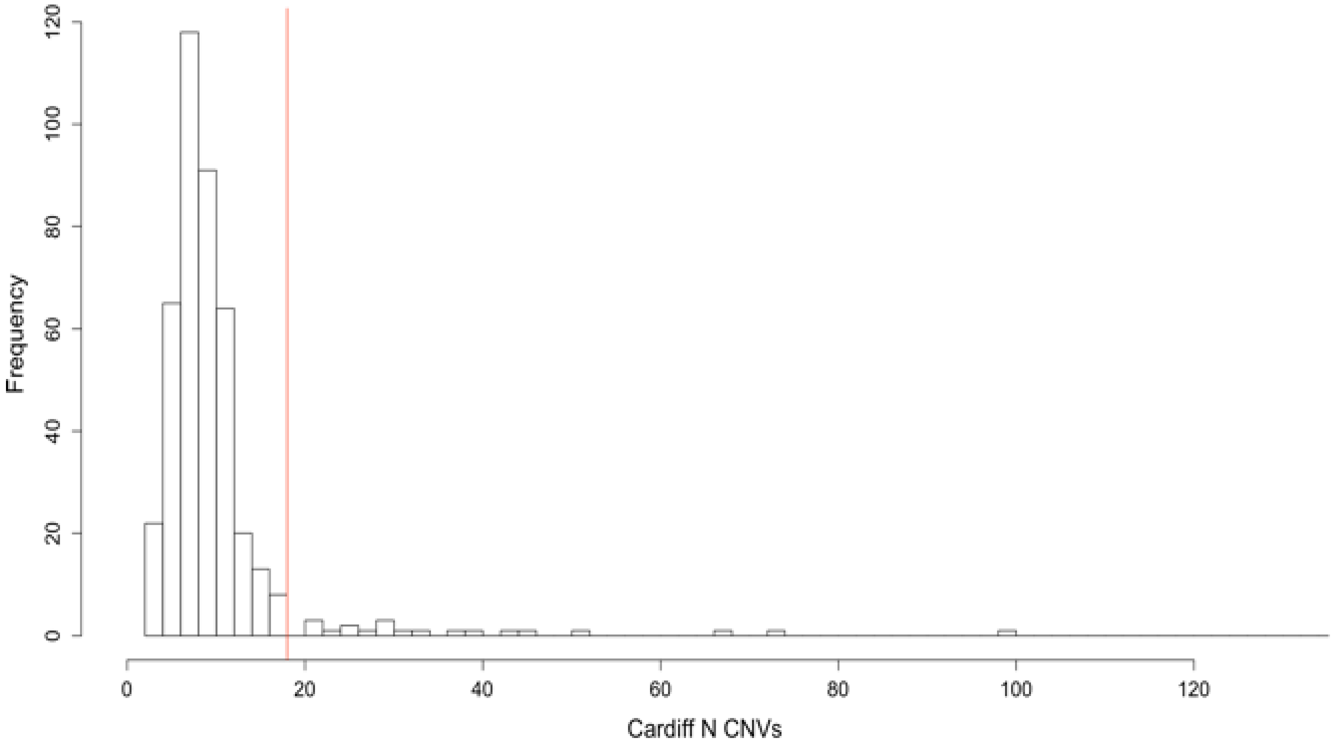


Figure 5.4. Histogram for n CNV calls per Cardiff-sequenced sample after merging. Median number of calls = 9. The red vertical line shows the threshold (2 x median = 18) above which samples were excluded for having an excess of calls.

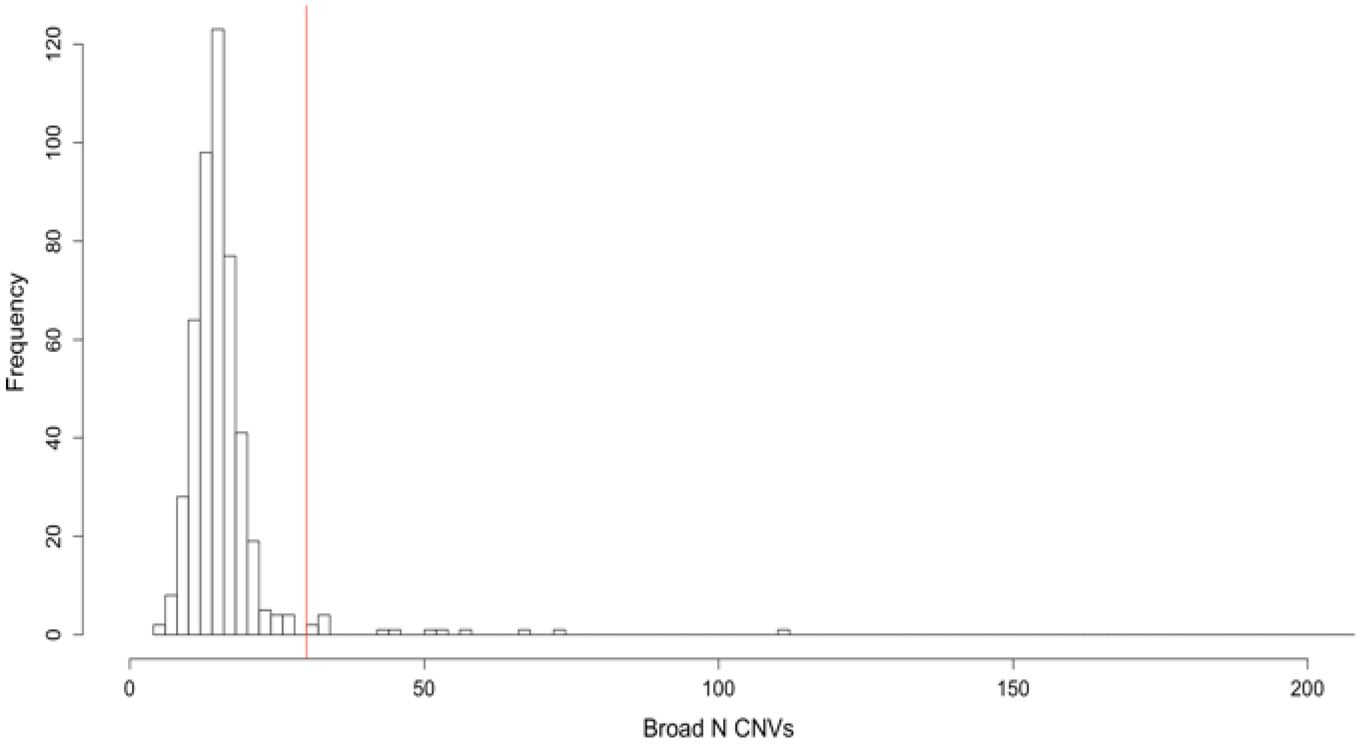


Figure 5.5. Histogram for n CNV calls per Broad-sequenced sample after merging. Median n calls = 15. The red vertical line shows the threshold (2 x median = 30) above which samples were excluded for having an excess of calls.

2.5.3 Variant level quality control

First, specific CNV regions were removed that were disproportionally prevalent among samples excluded in the sample level QC step, indicating loci for which CNV calling is problematic in each subcohort but were not included in the list supplied by the CLAMMS authors. Nine-hundred and ten call loci were identified among samples passing QC in Cardiff subcohort that occurred at least once. Of these, 43 were found to be overrepresented among outlier samples. Among samples passing QC in Broad subcohort, 1328 call loci occurred at least once and 34 were overrepresented among those samples that failed. Filtering these loci left 3,484 calls for Cardiff-sequenced samples and 7,146 for Broad-sequenced samples.

All remaining calls were combined into a single set comprising 10,610 calls. Calls were then filtered according to two quality scores generated by CLAMMS and described in Chapter 2: Qsome and Qexact. Different criteria were applied to deletions and duplications, based on observations of differing performance for CNV type, as described by the developers of CLAMMS, and reflective of the fact that duplications are more difficult to identify using coverage depth than deletions (Mei Teo et al, 2012). Deletions were filtered if Qsome <= 50 AND Qexact <= 0.5, while duplications were filtered if Qany <= 50 AND Qexact <= -1.0, leaving 4,551 CNV calls in total. Finally, Plink 1.9 ((Purcell et al., 2007), https://zzz.bwh.harvard.edu/plink/contact.shtml) was used to exclude remaining calls if they occurred in > 1% of samples. 1,137 rare CNVs remained after frequency filtering.

2.5.3.1 Manual inspection of CNV calls

To further minimise the CNV false positive rate, the sequencing coverage profiles for all CNVs that passed QC were manually inspected, according to the same methods outlined in section 2.6.2 of chapter 3. Figs. 5.6 and 5.7 show plots for calls that were excluded. Following manual inspection, 868 calls remained.

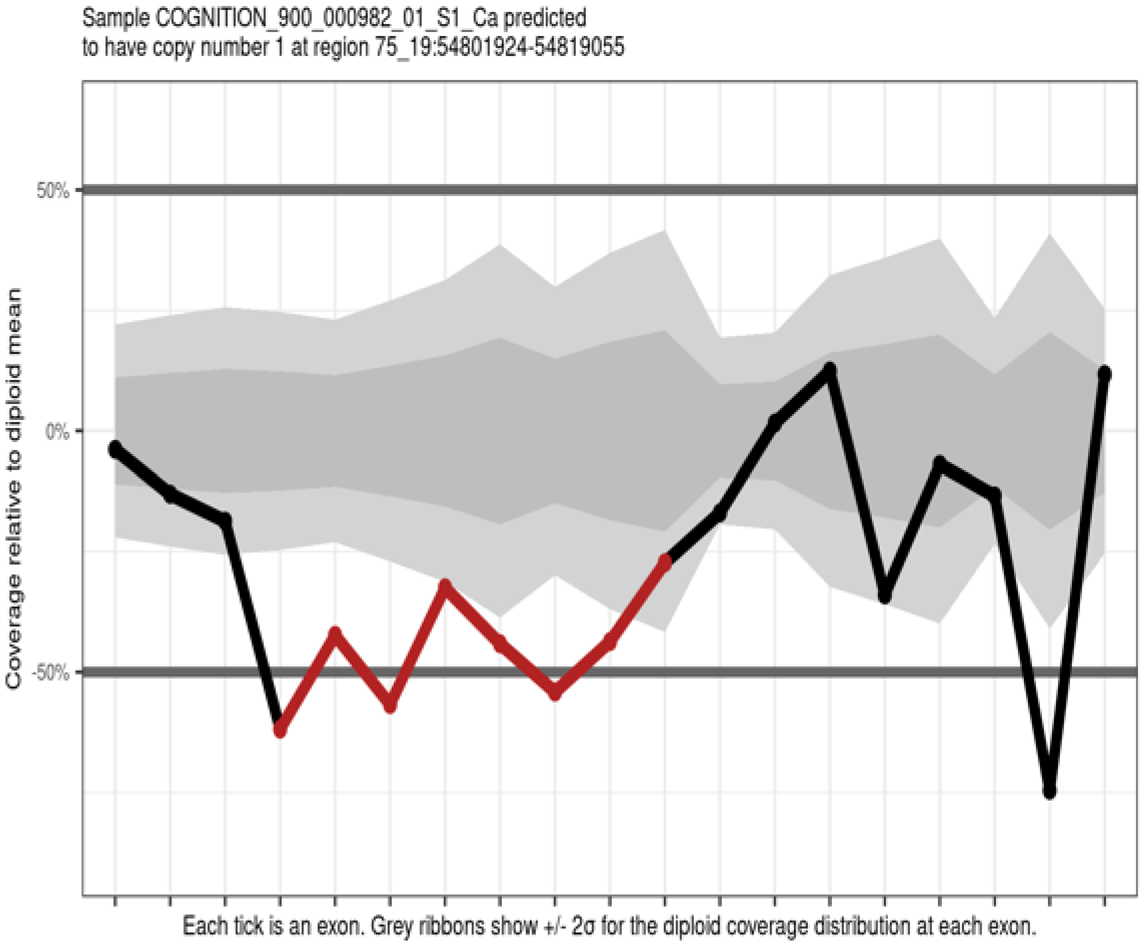


Figure 5.6. Coverage plot for a deletion that failed manual inspection. The red line shows mean coverage per exon between the CNV breakpoints, relative to the model's diploid mean, while the black line shows coverage for exons outside of the CNV region. The dark and light grey regions are the coverage depth within 1 and 2 standard deviations of the diploid mean, respectively.

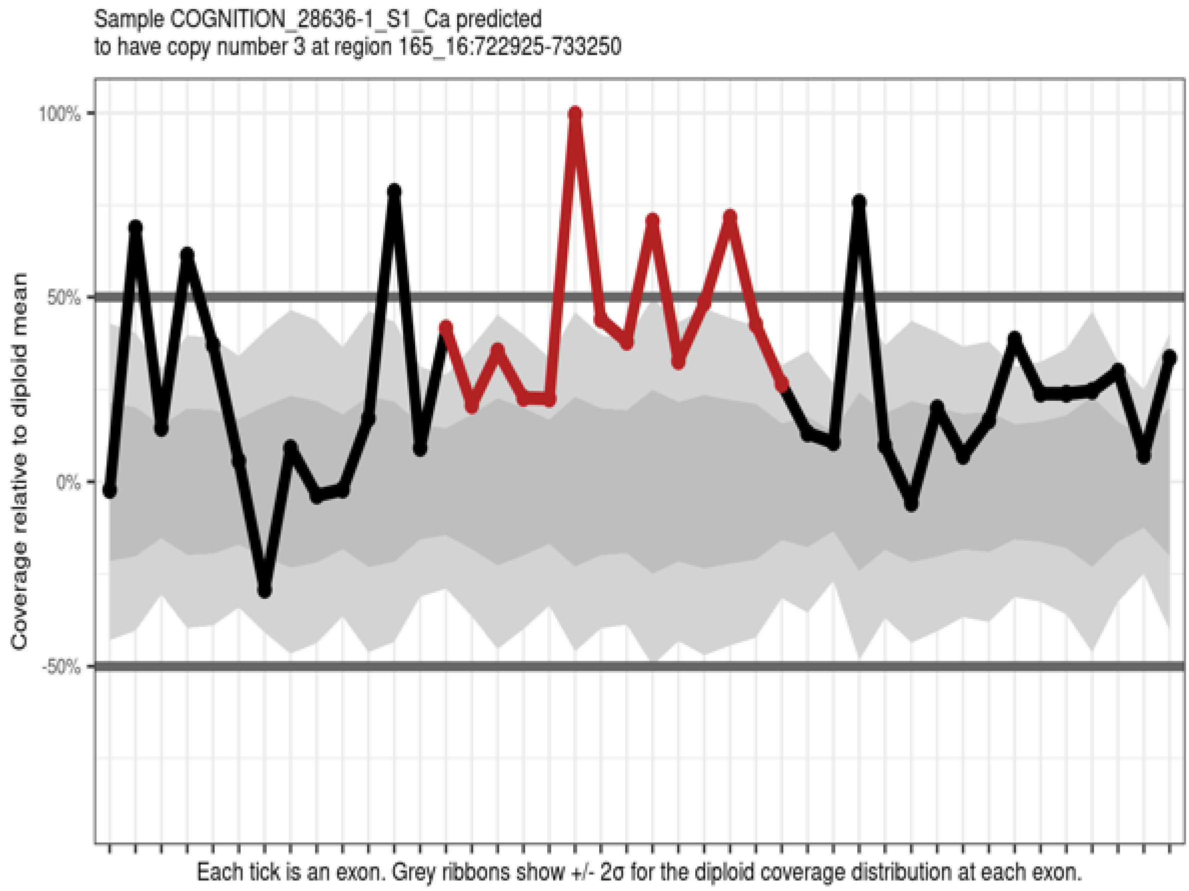


Figure 5.7. Coverage plot for a duplication that failed manual inspection.

2.6 Quality control summary

13,945 CNVs were called by CLAMMS from the Broad subcohort, and 5,684 for the Cardiff. From the initial sample of 927 individuals, 47 were excluded for having an excess of CNV calls (> 2x median), 22 from the Broad subcohort and 25 from the Cardiff subcohort. The number of calls remaining after each stage of variant QC across all samples are summarised in Table 5.4

Quality control stage	N CNV calls remaining
Unfiltered CLAMMS output	19,629
Call merging	19,349
Remove outlying samples (N CNV calls > 22)	10,711
Filter calls overrepresented in outlying samples	10,630
Filter calls with low quality scores	4,551
Filter common variants (> 1% allele-frequency)	1,137
Filter calls that fail manual inspection	979

Table 5.4 - N CNV calls remaining after each quality control step, across all COGS samples. Entries in the first column state what was filtered out at each stage. Quality control steps are listed in order of application.

2.7 CNV calls rate and mean target coverage

Figures 5.8 and 5.9 show the relation between subcohort and the average number of CNV calls per sample, before and after application of all variant-level QC criteria except allele frequency filters and manual inspection. Broad-sequenced samples have more CNV calls per individual in both cases. However, variant-level QC reduces the mean difference in the number of CNVs called (Figures 5.8 and 5.9).

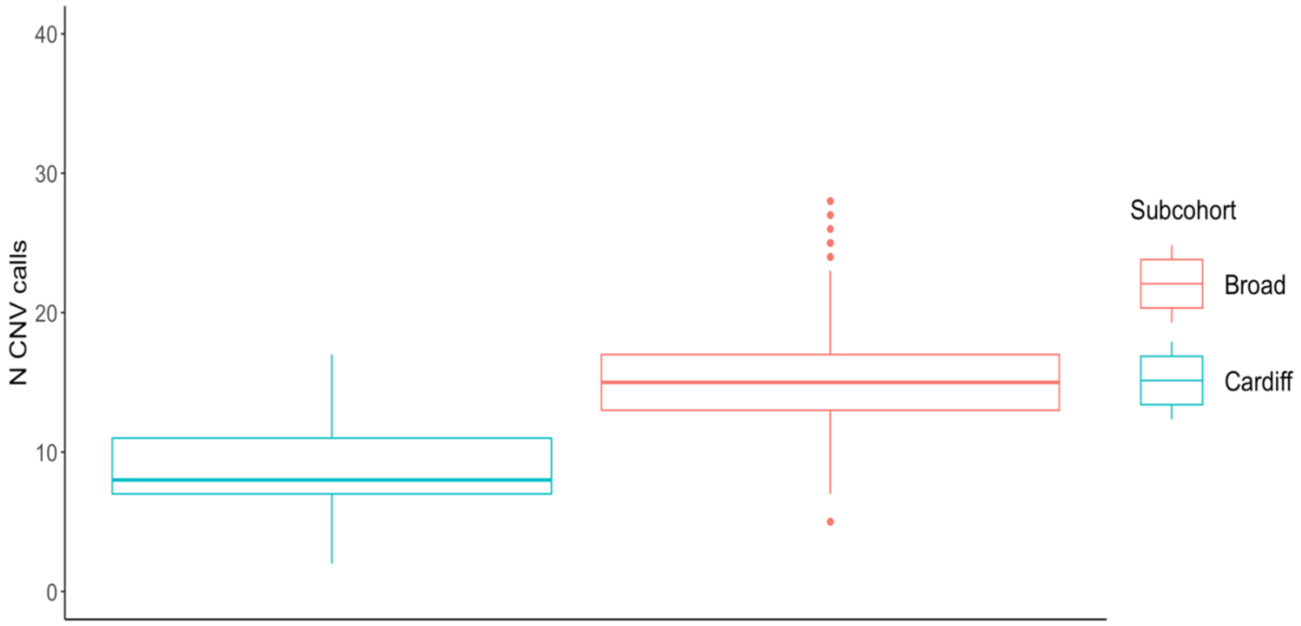


Figure 5.8: Mean number of CNV calls before variant-level QC, stratified by subcohort/sequencing site.

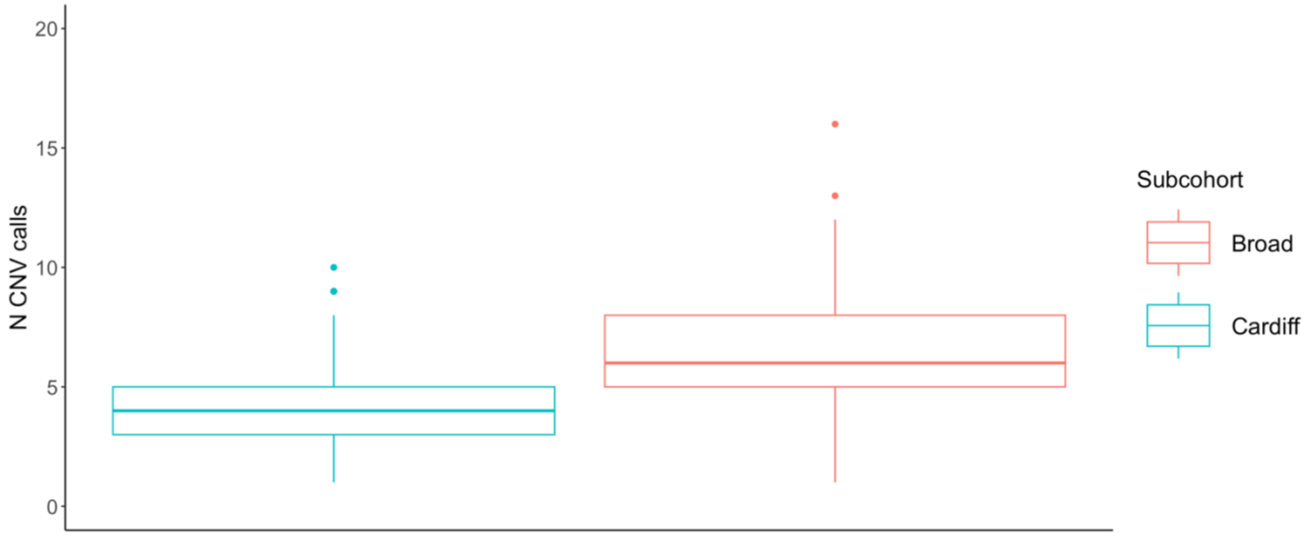


Figure 5.9: Mean number of CNV calls after all variant-level QC except for allele frequency filter and manual inspection, stratified by subcohort/sequencing site.

2.8 Comparing sequencing and array calls

2.8.1 Sample exclusion

For analyses that compared CNV calls between WES and array data, I only included CNVs from samples that passed QC on both platforms. Of 1,326 samples in the entire Cardiff COGS cohort, 1,097 had been either sequenced or genotyped. Of these, 927 had been sequenced and 993 had been genotyped, while 808 had been both sequenced and genotyped. Table 5.5 shows the number of samples excluded from the high-quality (HQ) and unfiltered (ANY) call sets that were not both sequenced and genotyped.

	ANY SEQ	HQ SEQ	ANY ARRAY	HQ ARRAY
Total N samples	927	880	993	983
N samples excluded	127	108	186	184

Table 5.5 N samples in unfiltered and high-quality call sets for both platforms, and n samples excluded from call sets for not having been both sequenced and genotyped. HQ = high-quality, SEQ = sequencing,

Using bedtools, individual CNV calls were then removed from the unfiltered and high-quality array call sets if they did not overlap any of the targeted regions listed in the Nextera exome capture file by at least one base, ensuring that all remaining array calls had the potential to be detected in the WES data. 8,720/12,671 (69%) of CNV calls were thereby excluded from the unfiltered array call set, and 402/1075 (37.3%) from the high-quality array call set.

2.8.2 CNV size comparison

To compare the numbers of CNVs identified by both approaches across a range of CNV sizes, high quality calls from each data set were separated into 6 subsets according to their size: <= 20kb, 20kb-50kb, 50kb-100kb, 100kb-500kb, 500kb-1mb and >= 1mb. For each sample, bedtools was used to count the number of CNVs called in both datasets, where concordant CNVs were defined as those overlapping by one base-pair or more.

2.8.3 Known schizophrenia risk CNVs

Sixteen known schizophrenia risk CNVs were previously called in the Cardiff COGs sample using array data (Table 5.6), based on the 11 CNVs described in (Rees et al, 2016) and presented in Table 1.1 of chapter 1. To ascertain whether any of these CNVs had been called by CLAMMS, high quality CNVs called in the sequencing data that reciprocally overlapped a pathogenic CNV by at least 66% were identified. This constraint ensured that no smaller calls within the breakpoints of a pathogenic CNV would be included.

CNV Locus	**Type**	**N**
1q21.1	DUP	2
2p16.3 (*NRXN1*)	DEL	1
7q11. 23	DUP	1
15q11.2	DEL	4
16p11.2dup	DUP	1
16p13.11dup	DUP	4
22q11.2del	DEL	3

Table 5.6. Known schizophrenia risk CNVs identified in Cardiff COGS array data. CNV = copy number variant, DUP = duplication, DEL = deletion, N = number of CNVs

2.8.4 Overlapping sequencing and array calls

Unfiltered CNV calls were annotated if they overlapped different subsets of calls in the other platform’s dataset. Those that passed all QC in the sequencing data were separately annotated if they overlapped, by at least a single base, calls in the array data that also passed all QC, and separately, unfiltered array calls. Similarly, calls that passed all QC in the array data were separately annotated if they overlapped, by at least a single base, calls in the sequencing data that also passed all QC, and separately, unfiltered sequencing calls. Three additional subsets per platform were thereby created, described in Table 5.7, and illustrated in Figures 5.10a-c.

Platform subset	**Description**
HQ SEQ	High quality, rare sequencing calls
HQ SEQ & ANY ARRAY	Intersect of high quality, rare sequencing calls and unfiltered array calls

HQ_SEQ & HQ_ARRAY	Intersect of high quality, rare sequencing calls and high quality, rare array calls
HQ ARRAY	High quality, rare array calls
HQ ARRAY & ANY SEQ	Intersect of high quality, rare array calls and unfiltered sequencing calls

Table 5.7. All platform subsets created by examining overlaps between sequencing and array calls, in addition to separate high quality call sets. HQ ARRAY & HQ SEQ was not included in later testing as it is identical to HQ SEQ & HQ ARRAY.

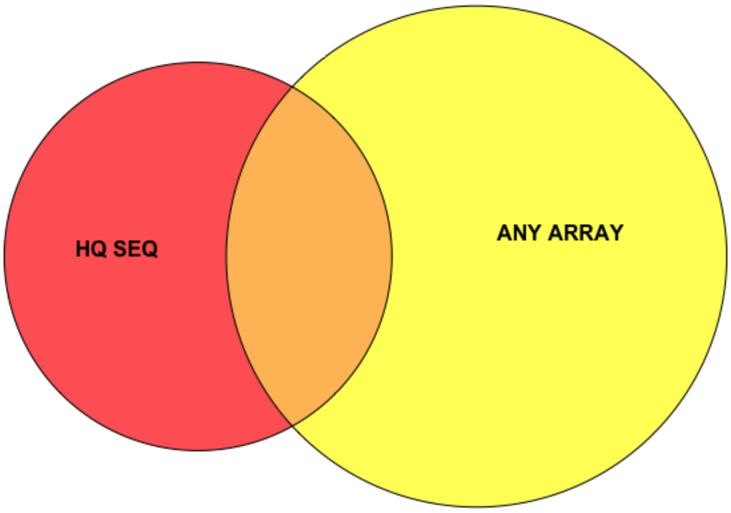


Figure 5.10a: Venn diagram illustrating the intersect (orange) of high-quality rare calls from the WES data (HQ SEQ, red), and unfiltered called from WES array data (ANY ARRAY, yellow). Circle size is not representative of actual data.

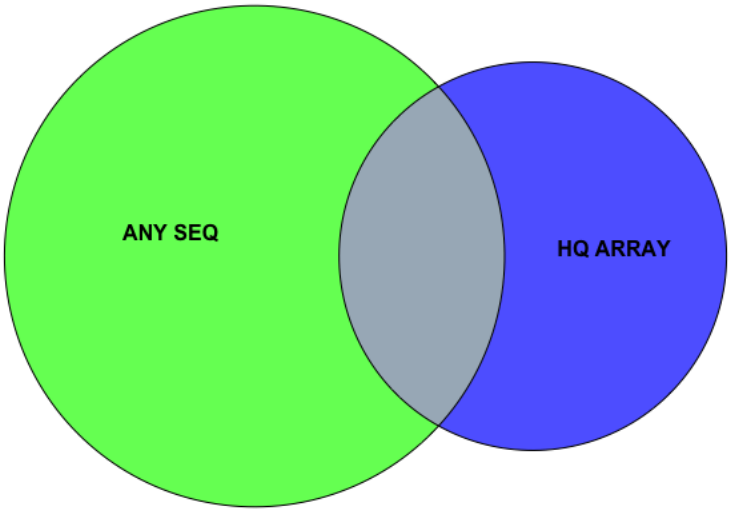


Figure 5.10b: Venn diagram illustrating the intersect (grey) of high-quality rare calls from the array data (HQ ARRAY, blue), and unfiltered called from the WES data (HQ ARRAY, blue). Circle size is not representative of actual data.

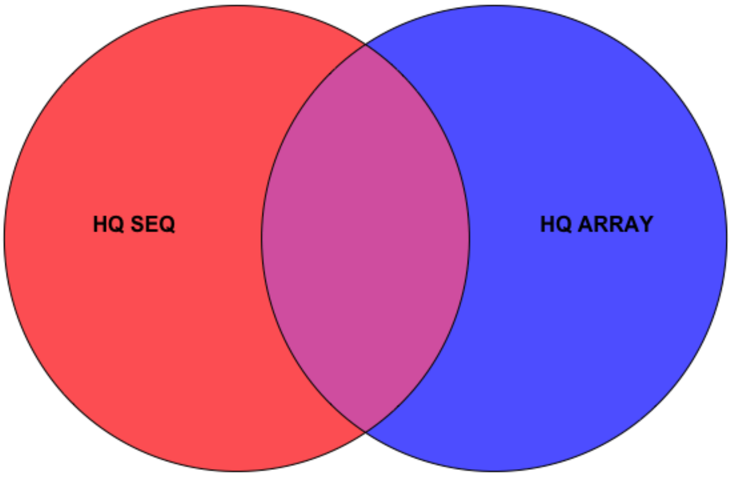


Figure 5.10c: Venn diagram illustrating the intersect (purple) of high-quality rare calls from the WES data (HQ SEQ, red), and high-quality rare calls from the array data (HQ ARRAY, blue). Circle size is not representative of actual data.

2.9 Subsetting variants by type, size, platform, and gene set.

All high quality CNVs that passed QC calls were annotated according to their size: LARGE if >100kb, or SMALL if <100kb. 100kb is the approximate discovery resolution for reliable CNV detection in array data. Genes that occurred within the breakpoints of these CNVs were identified using a script written by Elliott Rees whose inputs were the call sets and complete human gene list indexed to the GRCh37/hg19 build. Calls were further annotated if any genes they encompass are included in the Developmental Disorders Genotype-Phenotype (DDG2P) (Wright et al., 2015) or in the Genome Aggregation v2.1.1. (gnomAD) (Karczewski et al., 2020) databases. The former is a curated set of genes (n = 2579) that have been reported to be implicated in developmental disorders, including whether risk mutations are mono- or bi-allelic, effects on phenotype and degree of confidence in association (4 categories: limited, moderate, strong, definitive). In the present study, the database used was downloaded on 22/04/22 and the set was limited to genes harbouring monoallelic autosomal mutations that have a ‘strong’ or ‘definitive’ degree of confidence (n = 772).

The gnomAD database contains constraint metrics for the majority of known human genes (n = 19,704). One such metric is probability of loss-of-function (pLI), derived from the deviation of the observed n of protein-truncating-variants (PTV) from the expected n, accounting for sequencing content, coverage, and methylation, across multiple studies. Calls were annotated if they overlapped genes whose pLI score > 0.9 (n = 3,063), indicating a high degree of constraint.

One hundred and sixty subsets of variants were produced in total, based on the total number of intersects of five platform subcategories, three CNV type subcategories, three size subcategories and three gene set subcategories. These are shown in Table 5.8.

Category	Subcategory
Platform	HQ SEQ; HQ ARRAY; HQ SEQ-ANY ARRAY; HQ ARRAY-ANY SEQ; HQ SEQ-HQ ARRAY; HQ SEQ-HQ ARRAY
CNV Type	Any type; Deletion; Duplication

Size	Any size; Small (<=100kb); Large (>100kb)
Gene set	All genes; LoFi genes; NDD-risk genes

Table 5.8. All Intersections of these subcategories determined the total n of CNV sets that would be tested for association with cognition (n = 160). HQ = high quality calls, SEQ = sequencing, ANY = unfiltered calls, LoFi = loss-of-function intolerant, NDD = neurodevelopmental disorder

2.10 Statistics

All CNV subsets from both platforms were tested for association with current and premorbid cognition using linear regression. Models were generated using R package ‘speedglm’, which transforms the data so that models can be generated faster than by the using the equivalent base R functions, without compromising their validity (https://cran.r-project.org/web/packages/speedglm/speedglm.pdf). I covaried for age at disorder onset, age^2, sex, sequencing site, and principal components 1-10 that had previously been derived from common SNP variation. Beta coefficients and 95% confidence intervals produced using R package ‘confint’, for current and premorbid cognition were extracted from each model and plotted using R package ggplot2.

2.10.1 Multiple testing correction

For aim 3, I compare CNV calls between WES and array data, which does not require correction for multiple testing as it has already been demonstrated by previous studies that large CNVs (and particularly deletions) called from array data impact cognition. The objective of these analyses is only to assess whether the same variant types called by CLAMMS alter the strength of previous findings. For aim 4, however, I analyse WES data to investigate whether small CNVs typically missed by arrays contribute to cognitive deficits in schizophrenia. As the cognitive impacts of small CNVs have not previously been investigated (at least in the context of schizophrenia), correction for multiple testing is required. This analysis involved testing HQ SEQ deletions and duplications < 100KB in two genes sets: LoFi genes and NDD risk genes, for two cognitive phenotypes. Therefore, I applied FDR correction for 8 independent tests (two CNV types (deletions, duplications), two gene sets and two cognitive phenotypes).

3. Results

3.1 Rare CNVs identified in sequencing data.

A total of 979 high quality, rare CNVs were identified in the sequencing data. Three-hunded and twenty-seven were called in the Cardiff-sequencing samples and 652 in the Broad-sequenced samples. 556 (61%) participants were found to carry at least one rare CNV, while 278 (30%) were found to carry 2 or more. 352 (36%) were deletions and 627 (64%) were duplications. Table 5.9 shows the number of CNVs observed for each predicted copy number. Figures 5.11-14 are regional coverage plots showing evidence of small (<100kb) and large (>100kb) deletions and duplications.

CNV Type	**N**	**Predicted copy number**	**N**
Deletion	352	0	0
		1	352
Duplication	627	3	592
		4	9
		5	10
		6	6

Table 5.9. The total N of rare deletions and duplication identified across all WES data samples and for each integer copy number.

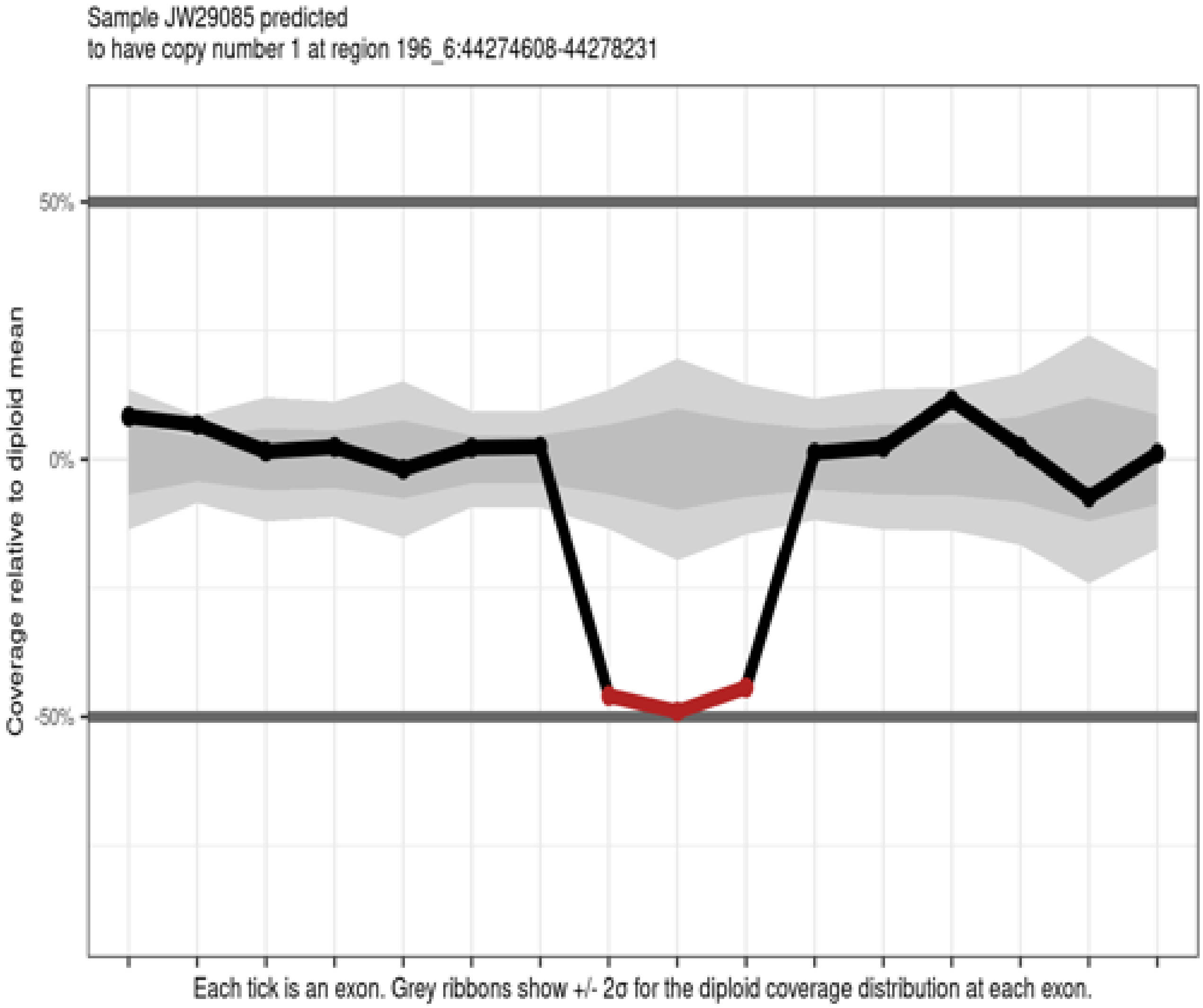


Figure 5.11. Regional coverage plot showing evidence for 3.6kb heterozygous deletion on chromosome 6.

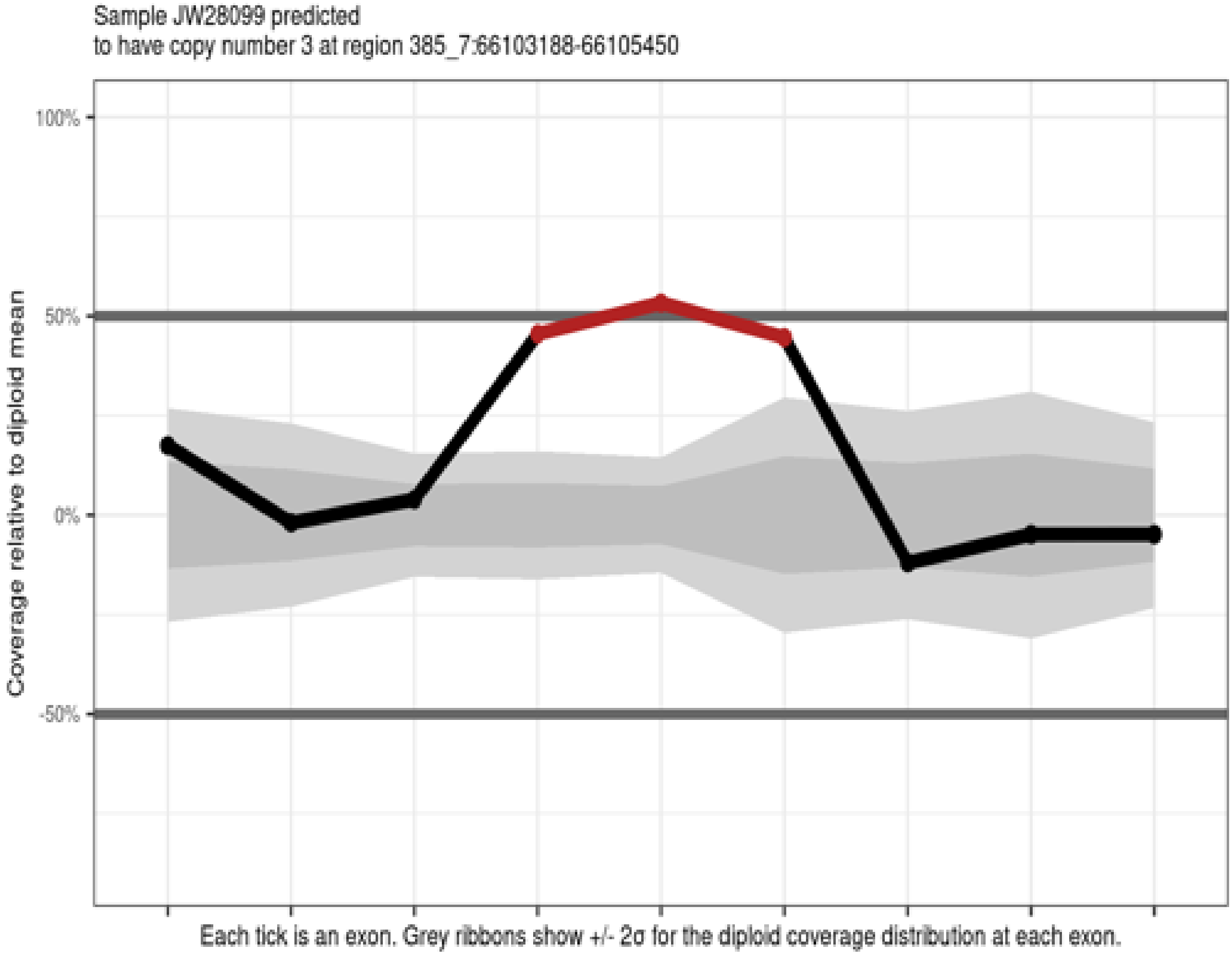


Figure 5.12. Regional coverage plot showing evidence for a 2.2kb heterozygous duplication on chromosome 7.

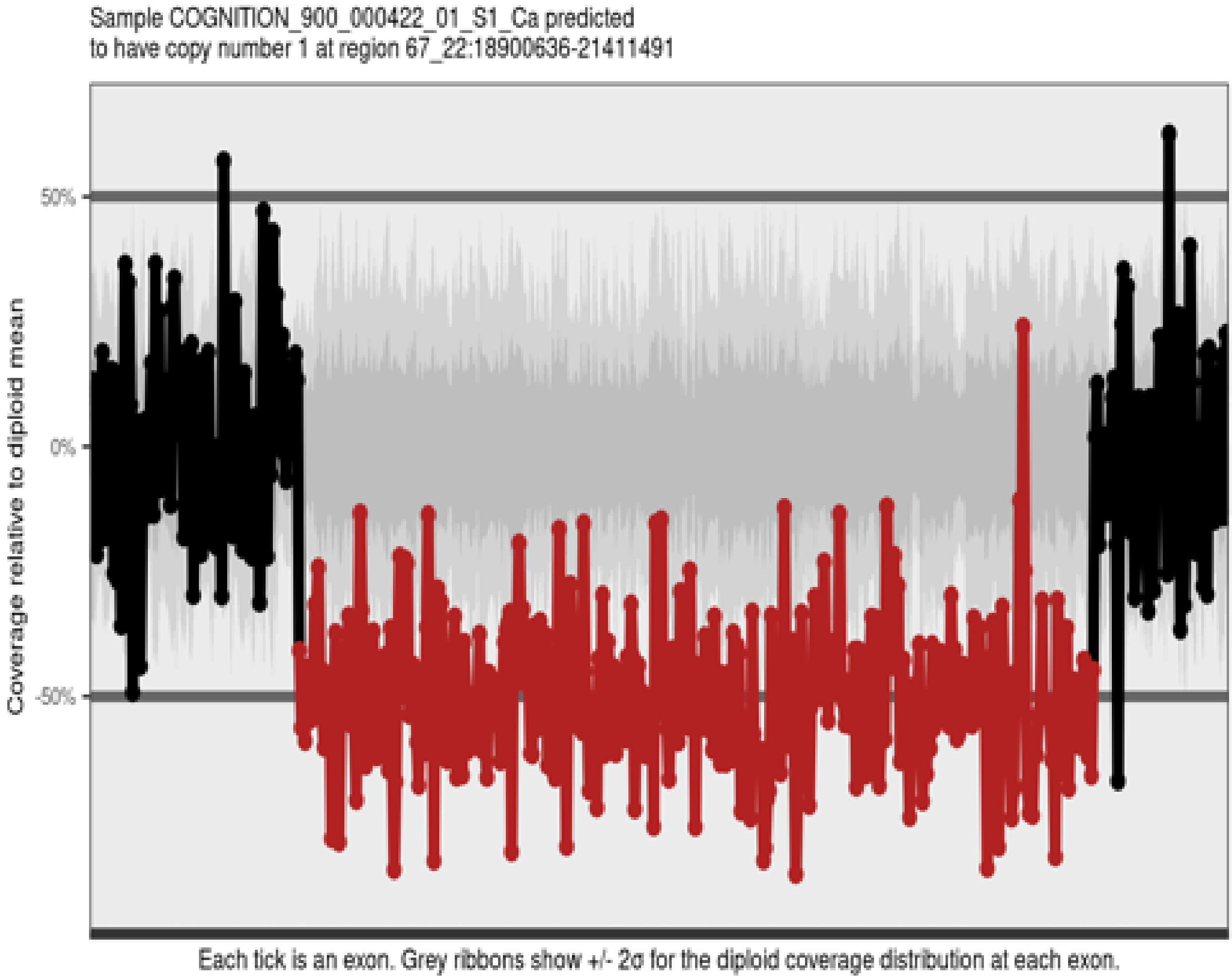


Figure 5.13. Regional coverage plot showing evidence for a 2.5mb heterozygous deletion on chromosome 22, in the DiGeorge Syndrome locus (22q11.2).

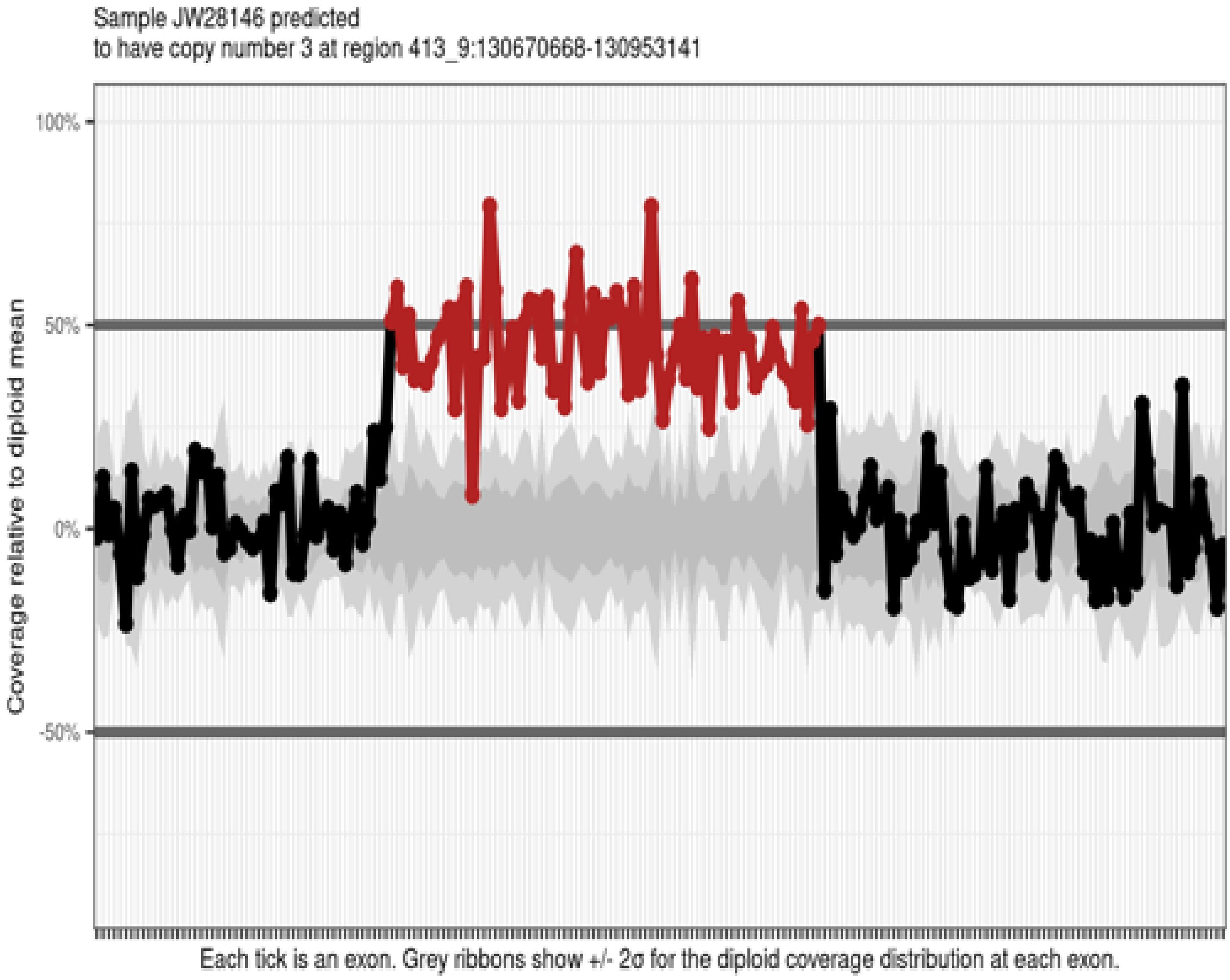
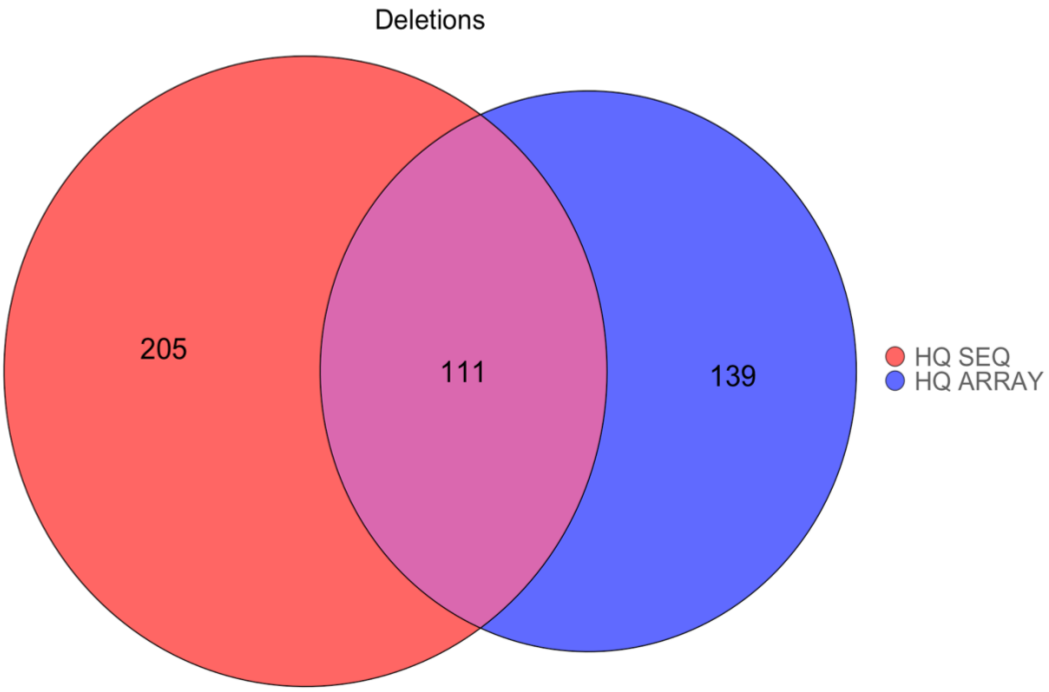
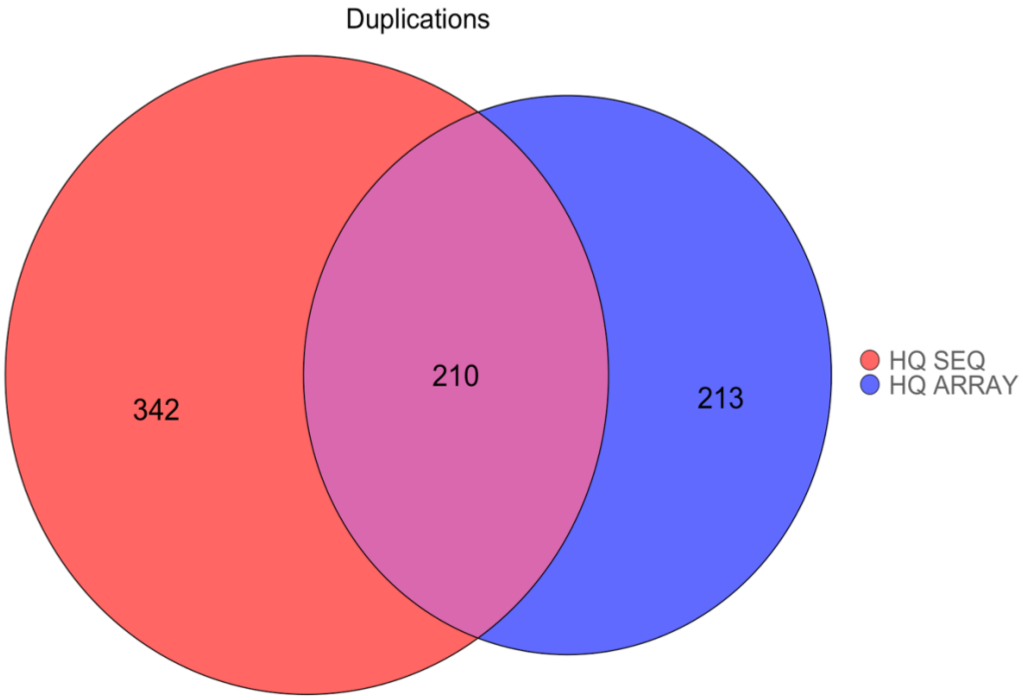


Figure 5.14. Coverage plot showing evidence for a 282kb heterozygous duplication on chromosome 9.

3.2 Comparison of sequencing and array CNV calls

Eight hundred and sixty-eight rare sequencing CNVs were called in samples that were also genotyped. Six hundred and seventy-three rare CNVs in the array data were called in exome capture regions in samples that were also sequenced: 250 (36%) deletions and 423 (64%) duplications. Five CNVs in the sequencing data were called as separate events in the array data, and two CNVs in the array data were called as separate events in the sequencing data. The calls for each CNV were merged, reducing the total number of calls in the sequencing set by two and in the array set by six. After merging, 321 rare CNVs were identified in both call sets, comprising 37% (321/866) of the sequencing calls, and 48% (321/667) of the array calls. One hundred and eleven (35%) of the intersecting calls were deletions and 210 (65%) were duplications. Figures 5.15a & b illustrate the platform overlap for deletions and duplications, respectively. 35/158 (22%) of rare sequencing calls that failed manual inspection were identified in the array data.





Figures 5.15a &b: Venn diagrams showing the platform overlap of calls in the high quality, rare call sets, by deletion (above) and duplication (below). HQ = high-quality, SEQ = sequencing.

Summary statistics for the size ranges of rare CNV called in both platforms' data sets are given in Table 5.10; they specify the largest and smallest CNVs, as well as the mean and median sizes. Figures 5.16-17 show density plots of CNV size for both platforms.

CNV size metric	**Size – WES calls**	**Size - Array calls**
Smallest	160bp	13kb
Largest	6.2mb	10mb
Mean	128.4kb	244kb
Median	34.4kb	107kb

Table 5.10. CNV size metrics for high quality, rare CNVs called in sequencing and array data. CNV = copy number variant, WES = whole exome sequencing, bp = base pairs, kb = kilobases, mb = megabases

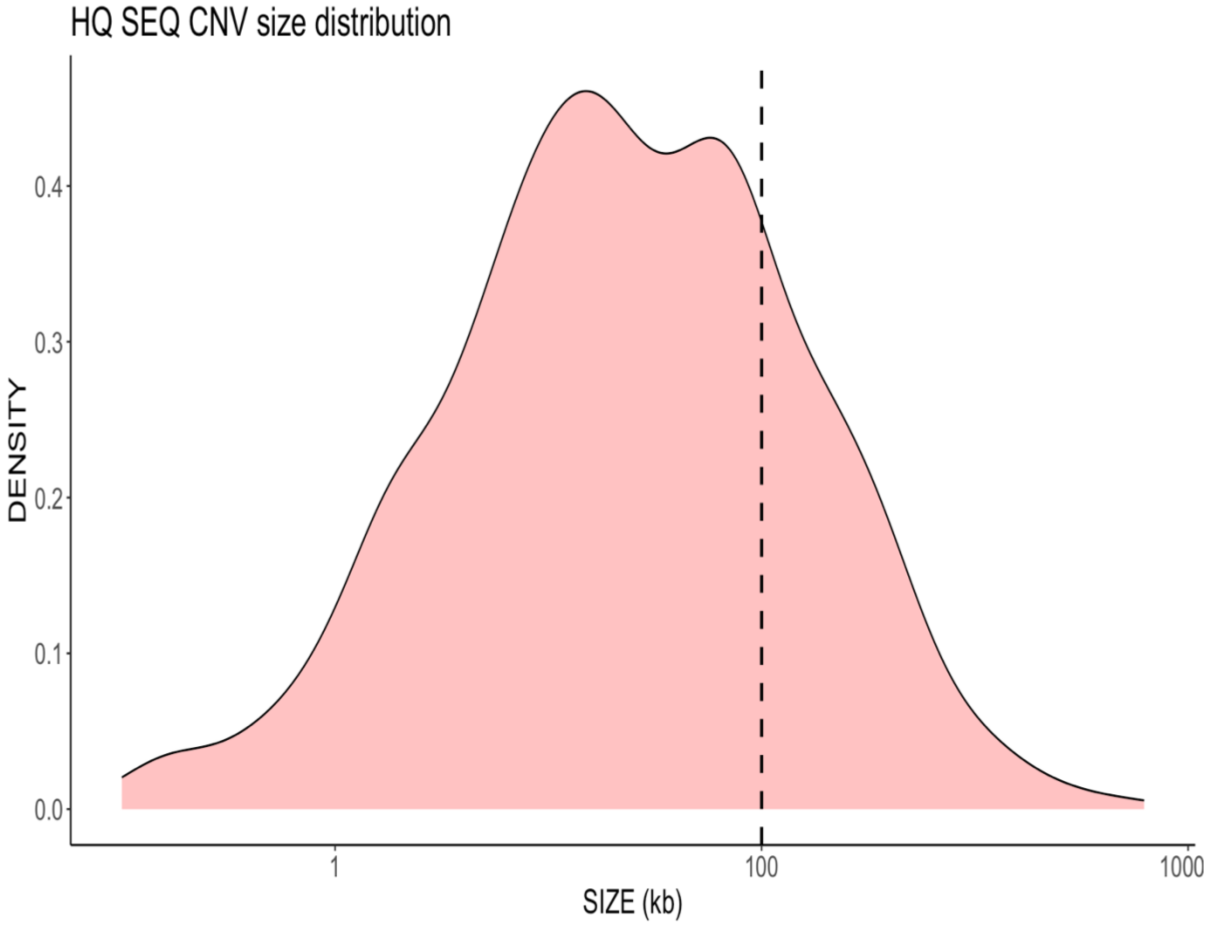


Figure 5.16. Density plot for size of high quality, rare CNVs called in sequencing data. X-axis scale has been log10 transformed. The dashed vertical line intersects at size 100kb, the approximate discovery resolution for reliable CNV detection in array data, and in the present study a chosen threshold separating ‘small’ and ‘large’ events.

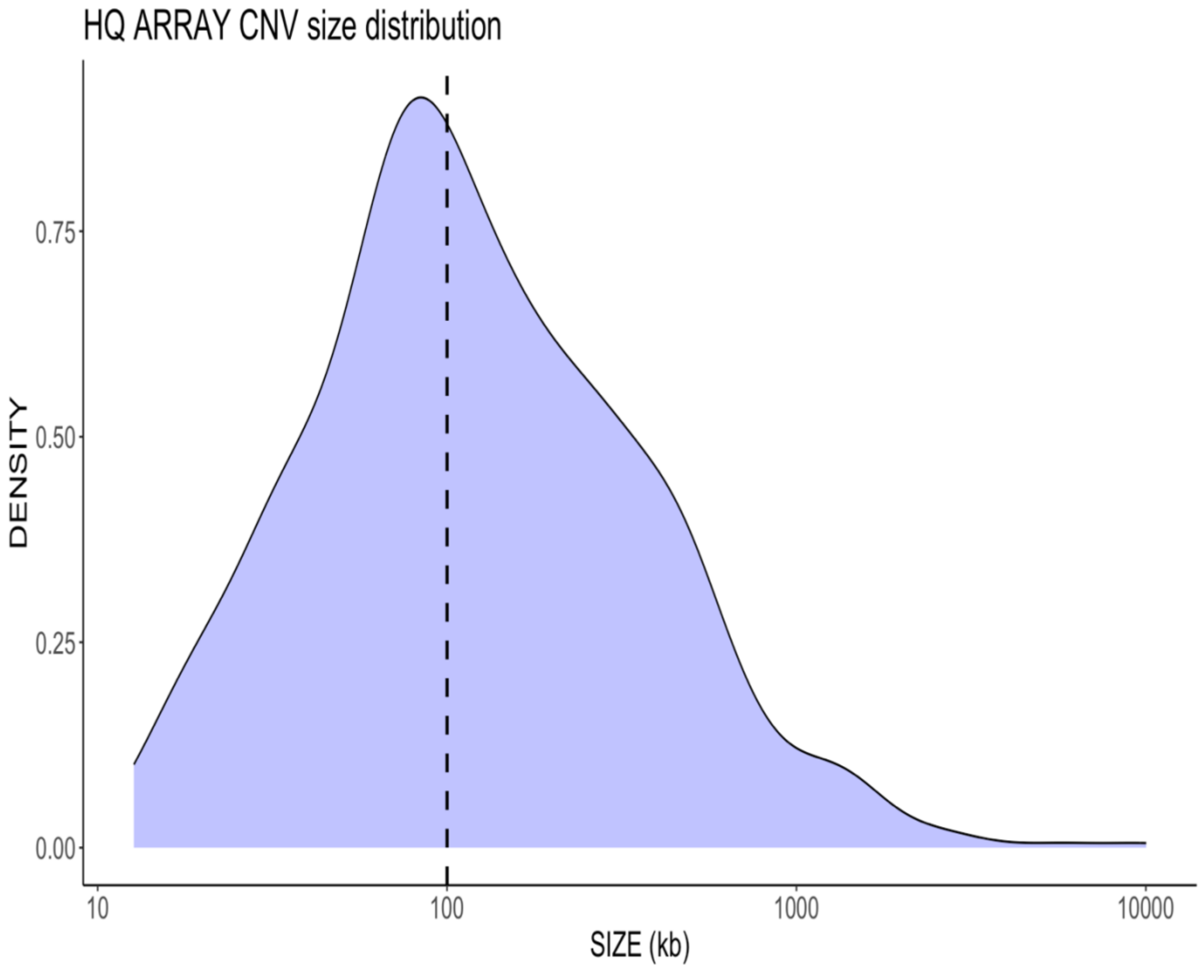


Figure 5.17. Density plot for size of high quality, rare CNVs called in array data. X-axis scale has been log10 transformed. The dashed vertical line intersects at size 100kb, the approximate discovery resolution for reliable CNV detection in array data, and in the present study a chosen threshold separating ‘small’ and ‘large’ events.

Figures 5.18-19 illustrate the platform intersects of rare CNVs within small (<100kb) and large (>100kb) size ranges. As each CNV has different breakpoints called in each data set, the intersect of CNVs for each size range differs according to which set is taken as primary in the comparison. For example, if a CNV is 90kb in the sequencing data but 110kb in the array data, it will only be included in the intersect of small CNVs if the sequencing set is taken as primary. These size discrepancies produced a net difference of 41 CNVs between the large and small intersects for both platforms.

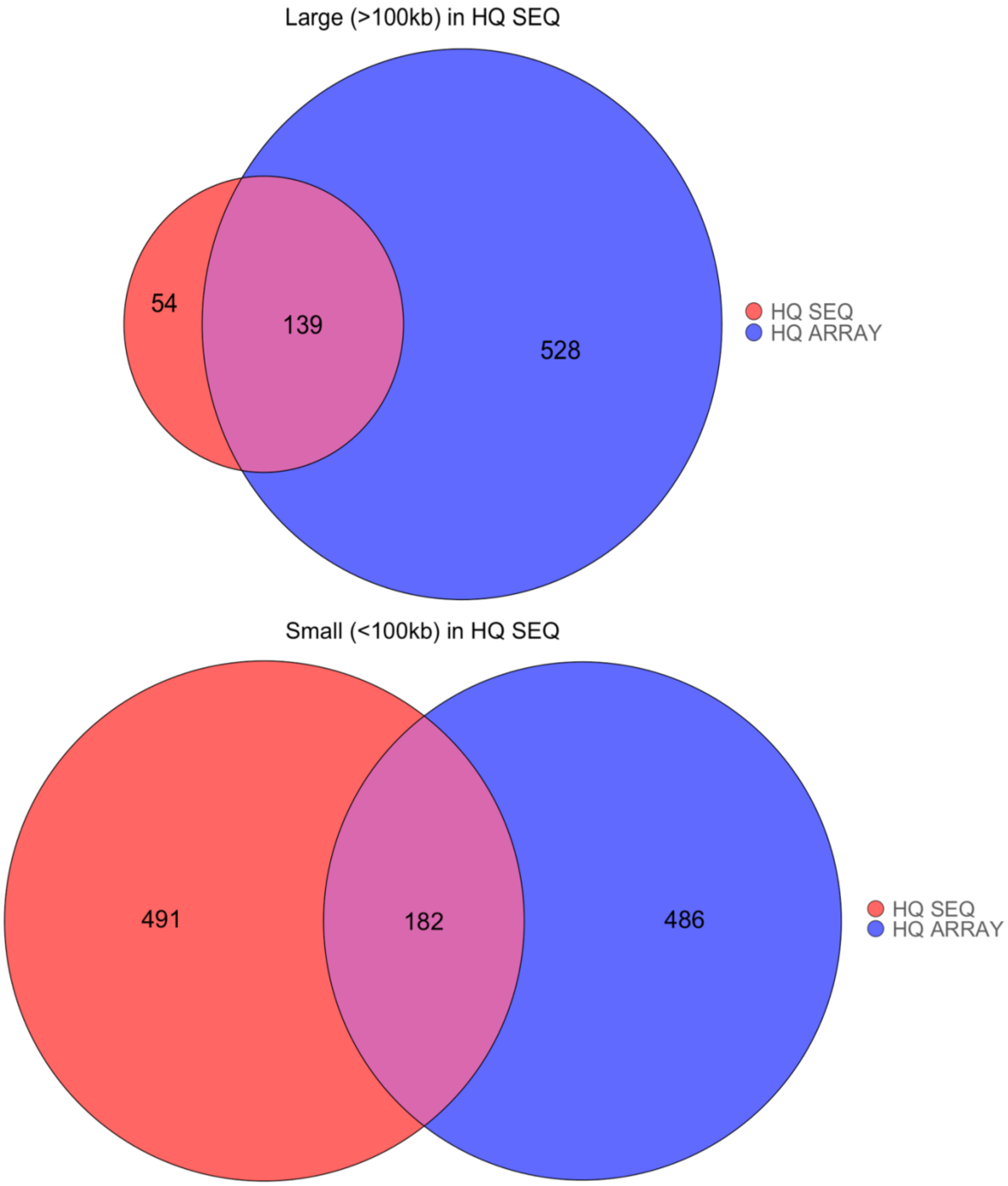
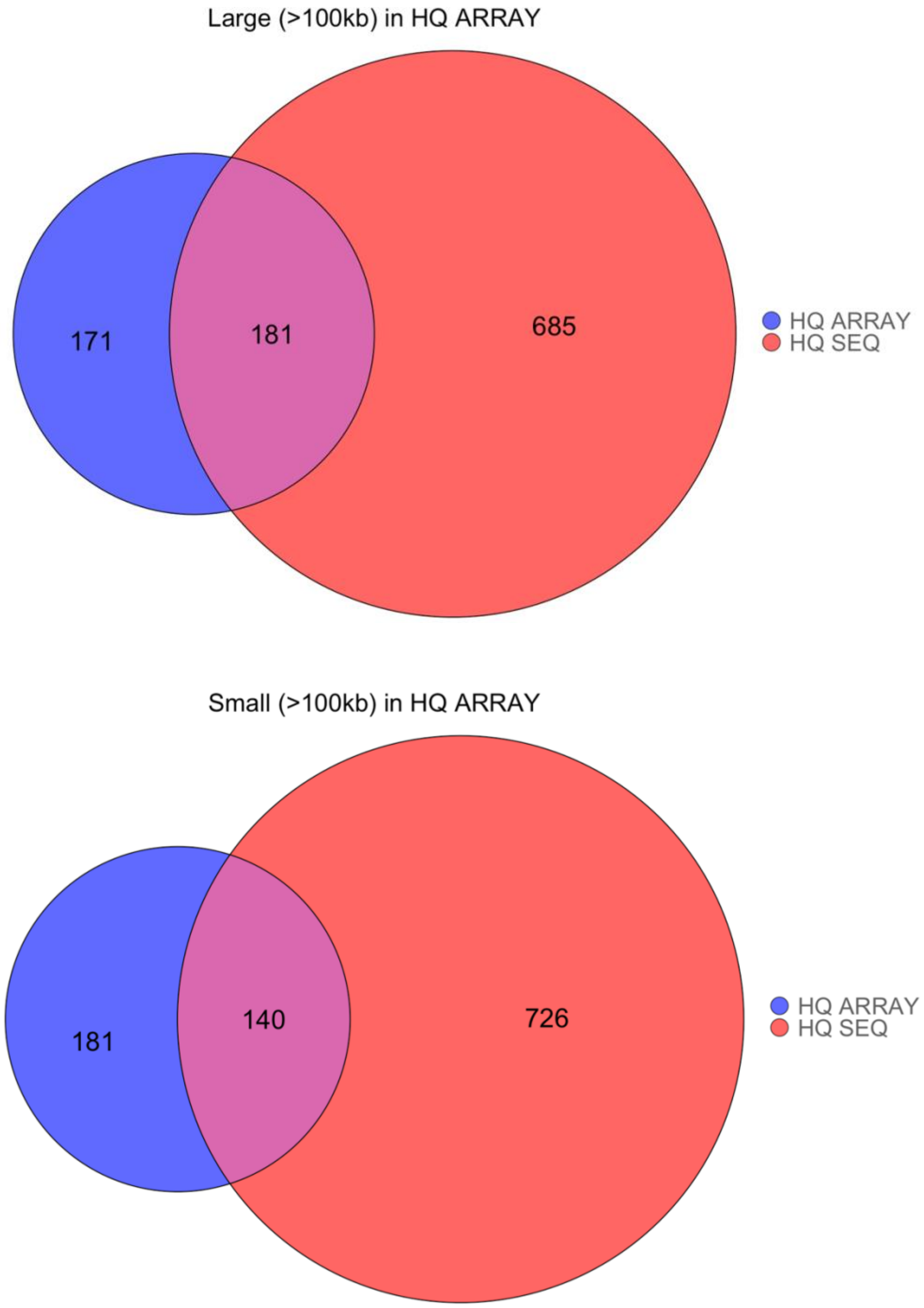


Figure 5.18a & b: Intersect of rare CNVs that are large (>100kb) and small (<100kb) in the sequencing data with all CNVs in the array data. HQ SEQ = high-quality, rare sequencing calls. HQ_ARRAY = high-quality, rare array calls.



Figures 5.19a & b: Intersect of rare CNVs that are large (>100kb) and small (<100kb) in the array data with all CNVs in the sequencing data. HQ SEQ = high-quality, rare sequencing calls. HQ_ARRAY = high-quality, rare array calls.

Table 5.11 shows the percentages of rare CNVs identified in each data set that were also identified in the other set, binned into six size ranges.

CNV size	% HQ SEQ in HQ ARRAY (N)	% HQ ARRAY in HQ SEQ (N)
under 20kb	10.6 (44/415)	42 (13/31)
20kb-50kb	49.6 (67/138)	46.3 (50/108)
50kb-100kb	59.2 (71/120)	42.3 (77/182)
100kb-500kb	70 (112/160)	47.8 (137/287)
500kb-1mb	88.2 (15/17)	68.4 (26/38)
over 1mb	75 (12/16)	85.6 (18/21)

Table 5.11. Percentage of rare CNVs identified in each platform’s data that were also identified in the other platform’s data, binned into 6 size ranges. CNV = copy number variant, N = N CNV calls.

3.3 Schizophrenia-risk CNVs

16 instances of CNVs impacting schizophrenia-risk loci were identified in the array data. All these events were detected by CLAMMS and passed variant QC except for one 16p11.2 duplication, whose corresponding sample was filtered in the sequencing analysis for having an excess of calls (Table 5.12). No additional schizophrenia CNVs were discovered in the WES data.

SCZ-risk locus	CNV Type	N in array	N in sequencing	Details
1q21.1	DUP	2	2	
2p16.3 (*NRXN1*)	DEL	1	1	
7q11. 23	DUP	1	1	
15q11.2	DEL	4	4	
16p11.2	DUP	1	1	Event was called in sequencing, but sample didn’t pass QC.
16p13.11	DUP	4	4	
22q11.2	DEL	2	2	

Table 5.12. CNVs affecting schizophrenia risk loci identified in sequencing and array data. All events called by CLAMMS were also called in the array data. SCZ = schizophrenia, DEL = deletion, DUP = duplication.

3.4 CNVs affecting loss-of-function intolerant and neurodevelopmental disorder risk genes

Table 5.13 shows the number of rare CNVs for each platform that were found to intersect at least one LoFi gene or one NDD-risk gene.

Platform	N LoFi gene intersects (%)	N NDD-risk gene intersects (%)	N LoFi and NDD-risk gene intersects (%)
HQ SEQ	151 (17.4)	54 (6.2)	44 (5.0)
HQ ARRAY	109 (16.3)	40 (6.0)	33 (5.0)

Table 5.13. N rare CNVs, for each platform, that were found to intersect one LoFi gene, NDD-risk gene, or both. % = percent of total number of CNVs.

3.5 Impact of CNVs on cognitive function

3.5.1 Current cognitive function

In the primary analyses, burden of small deletions and duplications in HQ SEQ were not found to be nominally associated with current cognition, even when impacting LoFi and NDD-risk genes (Table 3.14). In the secondary analyses, however, all large deletions in HQ SEQ, HQ SEQ & ANY ARRAY, and HQ SEQ & HQ ARRAY subsets were nominally associated (p<0.05) with current cognitive deficits in SCZ (Table 5.20), though no association was found for large duplications (Table 5.20). Moreover, large deletions affecting LoFi and NDD-risk genes in every platform subset were more strongly associated with cognitive deficits than those that were not restricted by gene set, while restricting by gene set did not produce any significant associations for large duplications (Figures 5.21 and 5.22). Table 5.14 specifies effect size, and nominal and corrected p values for all tests in the primary analysis, while figures 5.20-22 show effect sizes (beta coefficients) and 95% confidence intervals for all current cognition regression models, for both the primary and secondary analyses. Results for HQ ARRAY & HQ SEQ (wherein HQ ARRAY was taken as primary in the comparison) have been omitted, as they do not contain any CNVs that are not accounted for in the large and small HQ SEQ & HQ ARRAY intersects.

Gene set	CNV type	N variants	Beta	p	q
LoFi genes	Deletion	8	-0.0596	0.355	0.61
	Duplication	43	0.0894	0.379	0.61
NDD-risk genes	Deletion	3	-0.694	0.897	0.9
	Duplication	16	0.299	0.661	0.88

Table 5.14. Impact of small (<100kb) SVs affecting loss-of-function intolerant and NDD-risk genes on current cognitive ability. LoFi = loss-of-function intolerant; NDD = neurodevelopmental disorder.

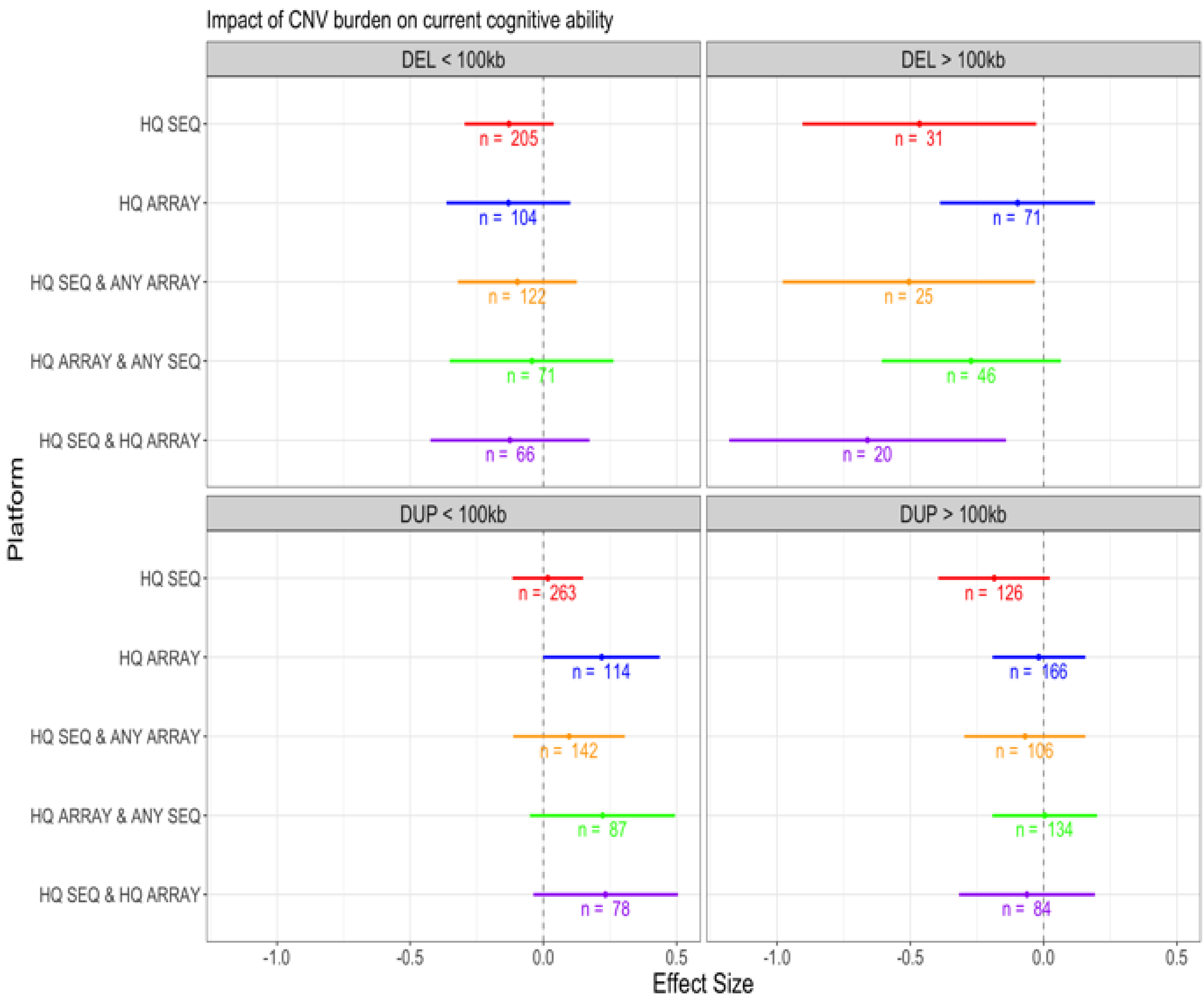


Figure 5.20. Impact of rare CNV burden on current cognition. n = ‘n’ refers to the number of variants tested in each subset. Effect sizes (beta coefficients) are shown as coloured points. The lines extending from each point are 95% confidence intervals. HQ = high-quality call set, ANY = unfiltered call set, DEL = deletion, DUP = duplication.

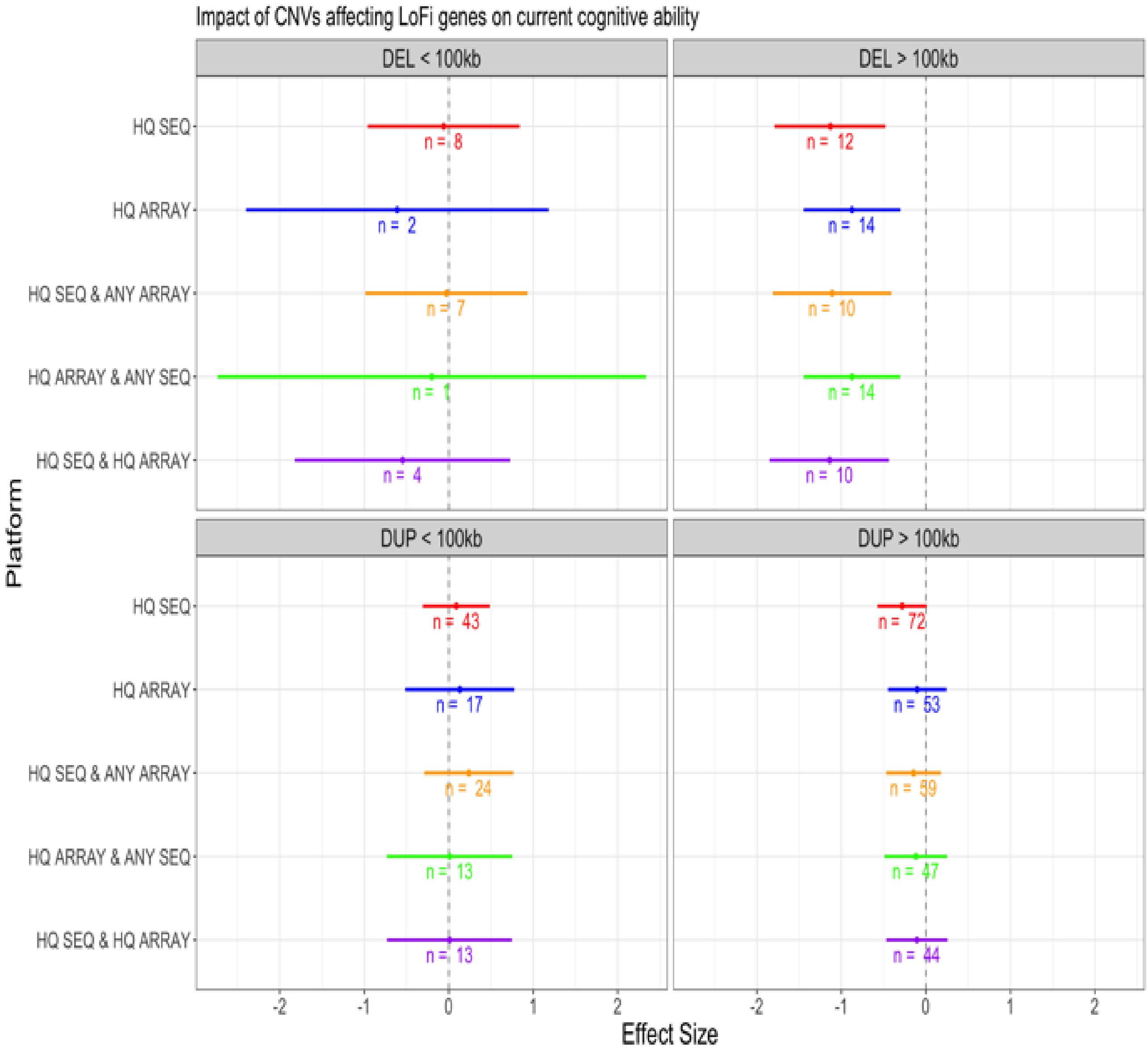


Figure 5.21. Impact of rare CNVs affecting loss-of-function intolerant (LoFi) genes on current cognition. ‘n’ refers to number of variants tested in each subset. Effect sizes (beta coefficients) are shown as coloured points. The lines extending from each point are 95% confidence intervals. HQ = high-quality call set, ANY = unfiltered call set, DEL = deletion, DUP = duplication.

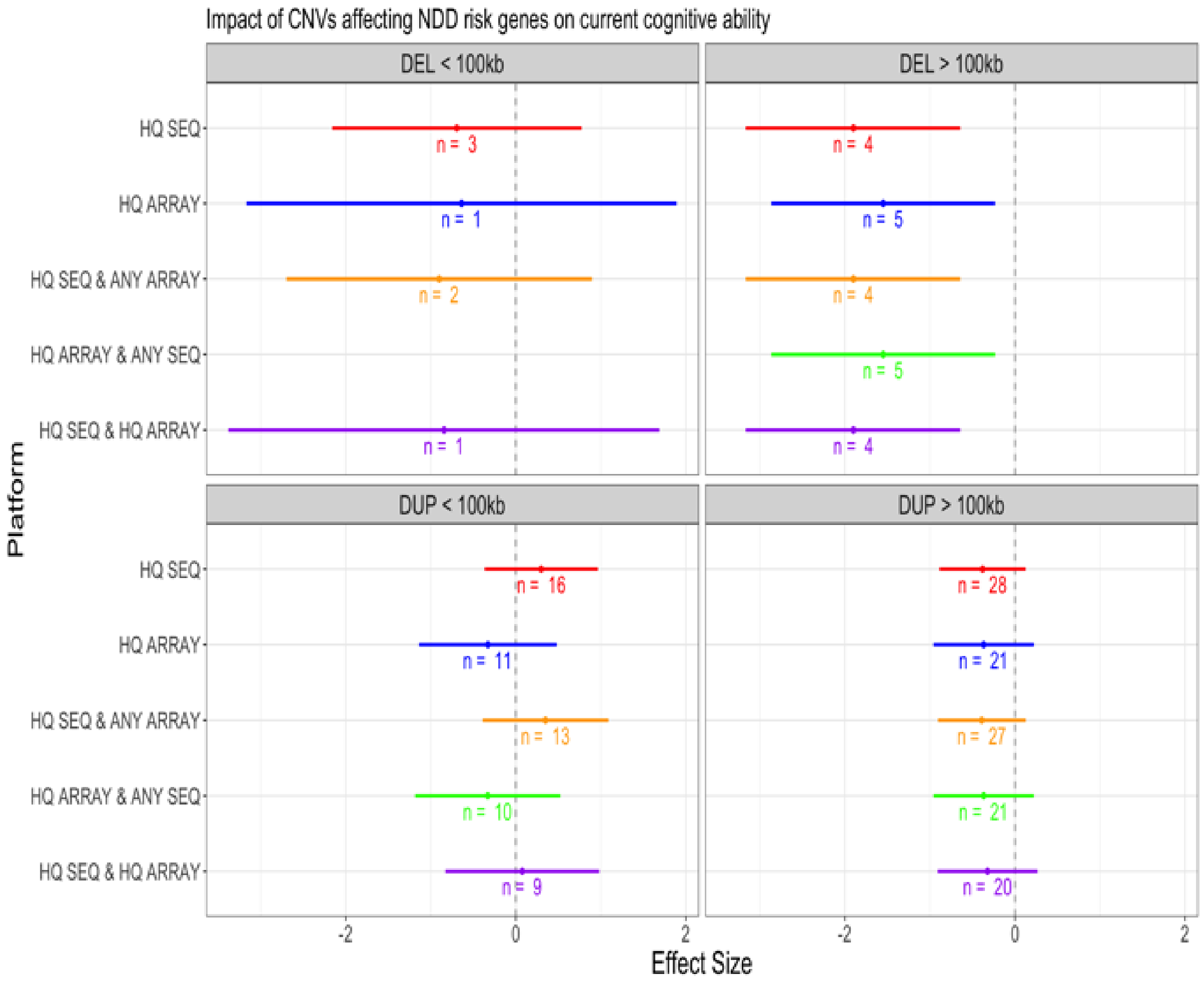


Figure 5.22. Impact of rare CNVs affecting neurodevelopmental disorder risk (NDD risk) genes on current cognition. ‘n’ refers to number of variants tested in each subset. Effect sizes (beta coefficients) are shown as coloured points. The lines extending from each point are 95% confidence intervals. HQ = high-quality call set, NDD = neurodevelopmental disorder, DEL = deletion, DUP = deletion, kb = kilobases. No result is given for small deletions in HQ ARRAY & ANY SEQ as no variants met these criteria.

Of particular interest are the four of large deletions affecting NDD-risk genes in HQ SEQ. They all occur with the 22q.11.2 locus, whose deletion is a cause of DiGeorge/VCF syndrome (DGS) (Cirillo, 2022), and is a risk factor for schizophrenia. Two are typical instances of the DGS-causing deletion, while the others impact only the distal end of the whole 3mb locus and are 673kb in size (Table 5.15).

CHR	Start	End	Type	Size	N genes	NDD-risk genes	LoFi genes
22	18,900,636	21,411,491	DEL	2.5mb	65	TBX1, LZTR1	HIRA, UFD1L, DGCR8, RTN4R, SCARF2, MED15
22	18,900,636	21,411,491	DEL	2.5mb	65	TBX1, LZTR1	HIRA, UFD1L, DGCR8, RTN4R, SCARF2, MED15
22	20,738,965	21,411,491	DEL	673kb	19	LZTR1	SCARF2, MED15
22	20,738,965	21,411,491	DEL	673kb	19	LZTR1	SCARF2, MED15

Table 5.15. Four large deletions that impact NDD-risk genes in the HQ SEQ call set. CHR = chromosome, NDD = neurodevelopmental disorder, LoFi = loss-of-function intolerant, DEL = deletion.

3.5.2 Estimated premorbid cognitive function

In the primary analyses, only small deletions called in HQ SEQ that impacted NDD-risk genes were found to be nominally associated with estimated premorbid cognition, though the effect size for small deletions impacting LoFi genes was trending in the expected direction (Table 5.15). In the secondary analyses, only large deletions in the HQ SEQ subset were nominally associated with estimated premorbid cognitive deficits (Figure 5.23), though restricting to LoFi and NDD-risk genes also produced significant associations for the HQ SEQ & ANY ARRAY and HQ SEQ & HQ ARRAY sets (Figures 5.24 and 5.25). A slightly significant association was also found for small deletions in HQ SEQ (p = 0.041, Figure 5.23). Again, large duplications were not found to be associated with estimated premorbid cognition (Figure 5.23), even when restricting variants by gene sets (Figures 5.24 and 5.25). Table 5.16 specifies the effect sizes, and nominal and corrected p values for all tests in the primary analysis, while figures 16-18 show effect sizes (beta coefficients) and 95% confidence intervals for all current cognition regression models, for both the primary and secondary analyses. Results for HQ ARRAY & HQ SEQ (wherein HQ ARRAY was taken as primary in the comparison) have been omitted, as they do not contain any CNVs that are not accounted for in the large and small HQ SEQ & HQ ARRAY intersects.

Gene set	**CNV type**	**N variants**	**Beta**	**p**	**q**
LoFi genes	Deletion	8	-0.106	0.783	0.9
	Duplication	43	0.262	0.0992	0.4
NDD-risk genes	Deletion	3	-1.24	0.0361	0.29
	Duplication	16	0.359	0.167	0.45

Table 5.16. Impact of small (<100kb) SVs affecting loss-of-function intolerant and NDD-risk genes on estimated premorbid cognitive ability. LoFi = loss-of-function intolerant genes. NDD = neurodevelopmental disorder.

Results of primary analyses for current cognition. * = nominal significance

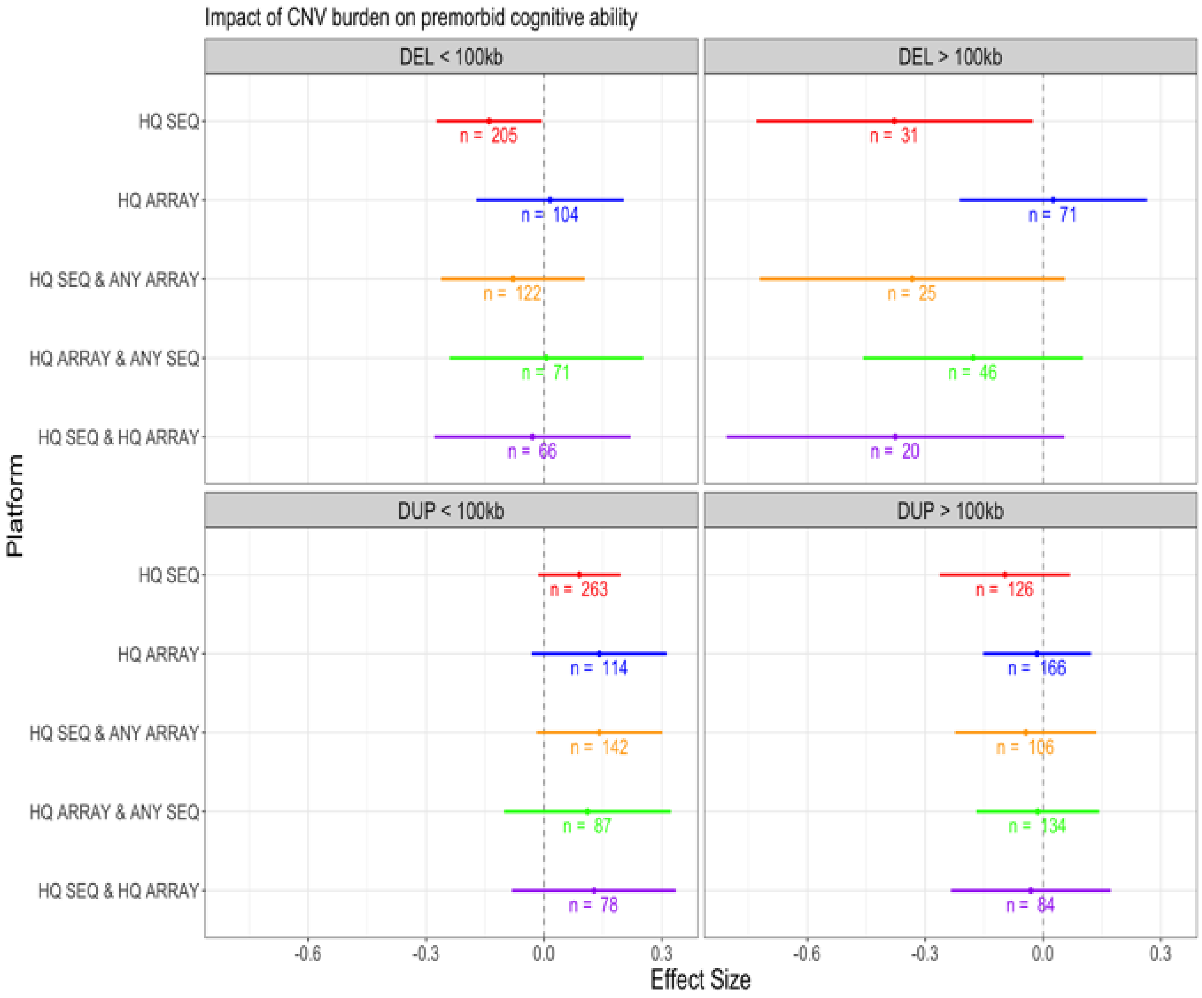


Figure 5.23. Impact of rare CNV burden on premorbid cognition. ‘n’ refers to number of variants tested in each subset. Effect sizes (beta coefficients) are shown as coloured points. The lines extending from each point are 95% confidence intervals. DEL = deletion, DUP = deletion, kb = kilobases.

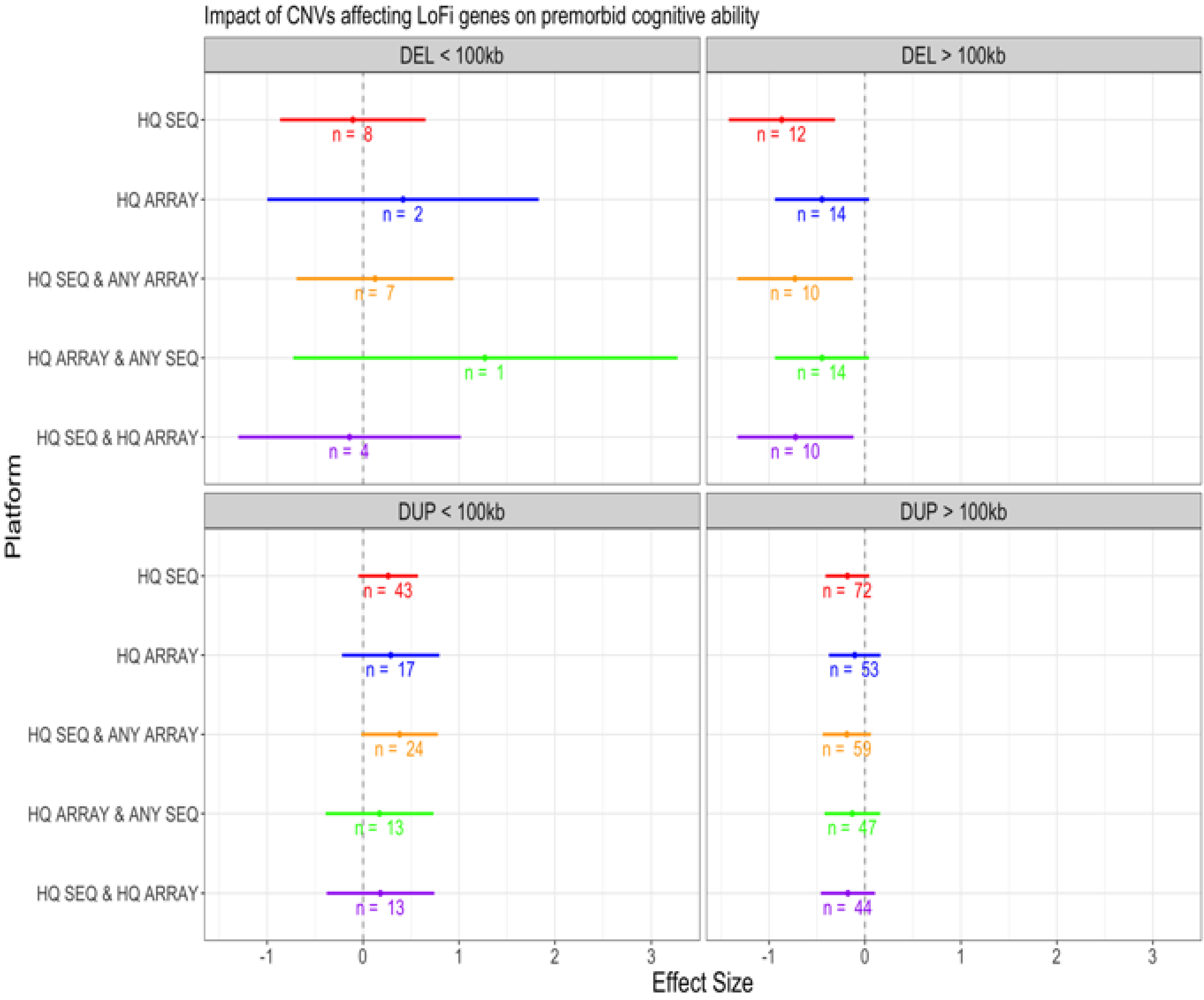


Figure 5.24. Impact of rare CNVs affecting loss-of-function intolerant (LoFi) genes on premorbid cognition. ‘n’ refers to number of variants tested in each subset. Effect sizes (beta coefficients) are shown as coloured points. The lines extending from each point are 95% confidence intervals. DEL = deletion, DUP = deletion, kb = kilobases.

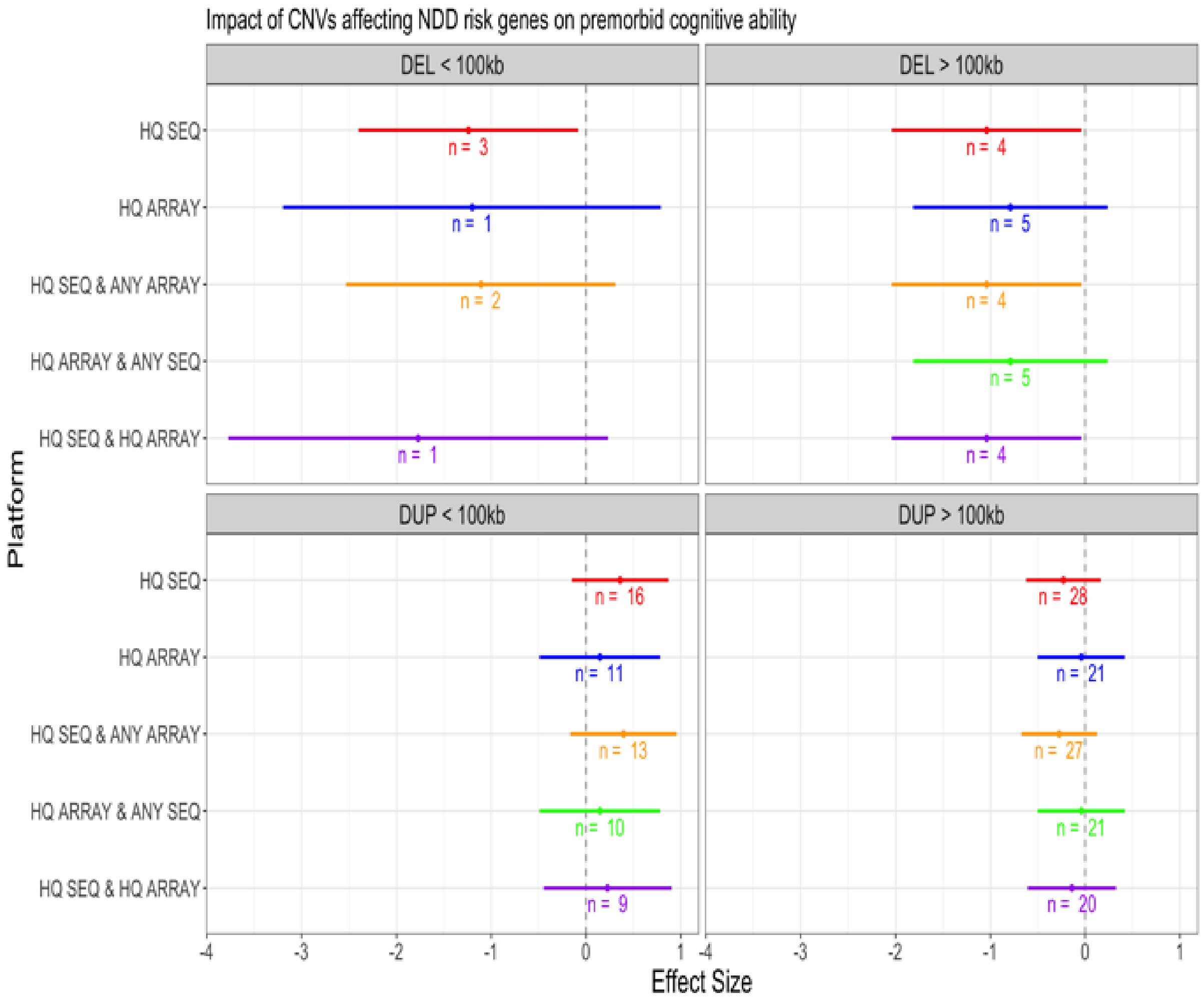


Figure 5.25. Impact of all CNVs affecting DDG2P genes (developmental disorder risk genes) on premorbid cognition. ‘n’ refers to number of variants tested in each subset. Effect sizes (beta coefficients) are shown as coloured points. The lines extending from each point are 95% confidence intervals. NDD = neurodevelopmental disorder, DEL = deletion, DUP = deletion, kb = kilobases.

Small deletions affecting NDD-risk genes in HQ SEQ were nominally associated with premorbid cognition. Only one of these was identified in array data (2:50847158-50883558 *NRXN1*). All three are singletons and impact one gene (Table 5.17), such that it may be simpler to derive their neurobiological effects than it is for larger variants affecting several genes.

CHR	Start	End	Size	Type	NDD-risk gene
11	77034140	77103567	69kb	DEL	PAK1
11	1477564	1482259	5kb	DEL	BRSK2
2	50847158	50883558	36kb	DEL	NRXN1

Table 5.17. Small (<100kb) deletions affecting neurodevelopmental disorder (NDD) risk genes in the HQ SEQ call set. CHR = chromosome, AF = allele frequency, DEL = deletion.

4.Discussion

4.1 Summary of aims

The research presented in this chapter had five aims:

1) Identify rare CNVs in the Cardiff COGS WES data set using CLAMMS.
2) Compare the sensitivity to detect known pathogenic or schizophrenia-risk CNVs between WES and array CNV call sets.
3) Assess whether analysing a consensus CNV call set based on both WES and array-based approaches produces a more accurate CNV call set and increases power to identify CNVs contributing to cognition in SCZ
4) Explore the impact of small CNVs typically missed in array studies (i.e. CNVs (<100kb)) on cognition in SCZ.
5) Determine if CNVs intersecting genes previously implicated in cognition in schizophrenia (LoFi and NDD-risk genes) are associated with cognitive deficits in this sample.

4.2 Schizophrenia-risk CNVs

For aim 1, I identified 868 putative, rare (AF < 1%) CNVs in 556 (61%) COGS participants: 352 deletions and 627 duplications. Sixteen schizophrenia-risk CNVs that were previously detected in the array were called by CLAMMS, and all but one CNV passed QC. This demonstrates that it is not necessary to implement both approaches if the only objective of a study is to detect large schizophrenia-risk CNVs (aim 2), assuming that the WES sample is large enough to sufficiently control for batch effects.

4.3 CNV sizes across platforms

CNVs across a broad size spectrum were detected by both approaches. More large (>100kb) CNVs were detected in the array data (n = 354) than in the WES data (n= 193), though given that 65 CNVs identified as small in the WES data were found to be large in the array data, this difference is likely to be primarily due to the systematic underestimation of breakpoint distances by CLAMMS. This is a limitation of setting a hard threshold (100kb) when comparing large/small CNVs across platforms. Seventy-two percent of large WES CNVs were identified in the array data and 51.4% of large array CNVs were identified in the WES data, which may suggest that the WES large CNV call set contains fewer false positives than that of the array data, as events that were called by both approaches are more likely to be real. It could also mean that CLAMMS has lower sensitivity to large events, however.

Approximately twice as many small (<100kb) variants were identified in the WES data (n = 673) than in the array data (n = 321), indicating that CLAMMS is more sensitive to variants in this size range. This is also reflected by the difference in median CNV size: 34.4kb for the WES call set and 107kb for the array call set. One hundred and eighty-two small CNVs in the WES data were detected in the array data, demonstrating that combining approaches can enable the detection of events that are too small to be reliably called in array data alone. This intersect constituted 27% of the small WES CNVs. This could indicate a high rate of false positive calls, though it may primarily be reflective of the fact that small events are harder to detect in the array data. Orthogonal validation of both intersecting and non-intersecting CNVs would be required to confirm this. 56.3% of small array CNVs were detected in the WES data, suggesting that there can reasonable confidence in the validity of small events called from array data.

4.4 Cognitive impact of CNVs by type, size and gene set

For my primary analyses of cognition, I tested the association of rare, small (<100kb) CNVs called in the WES data that affect LoFi and NDD-risk genes with current and premorbid cognition, as the impact of CNVs in this size range had not been explored by previous studies (aims 4 and 5). While all effect sizes for small deletions affecting these gene sets were trending in the expected direction, there were no significant associations found for current cognition, though small deletions in NDD-risk genes were associated with estimated premorbid cognition. There were only three variants in this subset, however, limiting conclusions that can be drawn about the impact of this gene set as whole.

Each of the three small deletions affecting NDD-risk genes impact only one gene: *PAK1*, *BRSK2* and *NRXN1* (Table 5.17). The proteins encoded by these genes are functionally related to synapse formation and maintenance; for example, PAK1 encodes p21-activated kinase 1, an enzyme that has been demonstrated to play a role in cortical development (Bokoch, 2003). A reduced number of pyramidical neurons, particularly in upper cortical layers, has been observed in PAK1 knock-out mouse models compared to wild type, indicating disrupted neuronal migration (Kelly & Chernoff, 2012). BRSK2 encodes Brain-selective kinase 2 which plays a significant role in neuronal polarisation; compared to wild-type, knock-out mice show a marked decrease in distinct axon and dendrite formation (Hiatt et al., 2019). NRXN1 deletion is known SCZ-risk CNVs, and the gene encodes the neurexin 1 cell adhesion protein. This trains-membrane protein is localised in the pre-synapse, and binds to neuregulins on the post-synaptic membrane, primarily to maintain synaptic structure and organisation (Südhof, 2008). It is therefore plausible that the CNVs impacting these genes are clinically impactful for their respective carriers.

In the secondary analysis, rare, small deletions unrestricted by gene set in HQ SEQ (n = 205) were nominally associated with estimated premorbid cognition, but those in HQ ARRAY were not. While the latter result was expected (as arrays are mostly insensitive to such variants), the former is quite surprising given that small deletions affecting LoFi genes (n = 8) in HQ SEQ were not found to be associated with this cognitive metric. The likely implication is that there is a sufficient number of small deletions in the larger, unrestricted call set to partly overcome the power limitation that besets the LoFi analysis. While non-significant, the effect size for small deletions called in HQ SEQ on current cognition is trending in the expected direction, suggesting that reproducing my analysis with a larger sample would give a significant association for this cognitive measure also. I can thus conclude there is evidence in my results that small deletions called in WES data do negatively impact cognition in SCZ (aim 4).

Corroborating previous studies, I've also showed that rare, large deletions, particularly those affecting LoFi and NDD-risk genes, are associated with deficits in current and estimated premorbid cognition in schizophrenia (aim 5). There is no evidence in my results that large deletions impact one cognitive measure more greatly than the other. When variants are not restricted by gene set, however, only large deletions detected in the WES data were associated with either cognitive measure. As it is already known that large deletions negatively impact cognition, these findings support the hypothesis that the WES large CNV call set contains fewer false positives than the array-based call set. Restricting the array-based large deletions to LoFi and NDD-risk genes produced significant associations with current, but not estimated premorbid cognition, though in the latter case effect sizes were trending in a more negative direction.

Of particular interest were the rare, large deletions affecting NDD-risk genes called in the WES data that are associated with current cognition. All four of these occur with the 22q.11.2 locus whose deletion is reported to cause DiGeorge/VCF syndrome (Cirillo et al., 2022). However, two only affect the distal end of the locus, and are around 700kb in length compared to the 2.5mb deletion of the whole region (Table 5.15). This smaller subtype, known as a nested central deletion, has not been studied in the context of cognition beyond small clinical samples (Burnside, 2015; Karbarz, 2020), and is explored further in a follow up study presented in section 5 of the current chapter.

Duplications of any size were not associated with either cognitive measure, even when restricted to those affecting LoFi and NDD-risk genes. However, effect sizes for all sets of large duplications are trending in the direction of lower cognitive ability for both measures. This may be indicative that such variants are indeed associated with cognitive deficits but have a more limited impact than large deletions. Repeating analyses on larger cohorts would be required to confirm this hypothesis.

4.5 Combining WES and array call sets

Combining WES and array-based call sets consistently produced weaker effect sizes for both cognitive measures than WES call sets tested separately (aim 3). However, effect size of the large deletions in the HQ SEQ & HQ ARRAY intersect on current cognition is trending in a more negative direction than that of HQ SEQ and HQ ARRAY tested separately, suggesting that combining call sets might improve power to detect effects from CNVs on cognition by removing false positive calls. Similarly, the effect size of large deletions in the HQ ARRAY & ANY SEQ intersect on current and estimated premorbid cognition was trending in a more negative direction than large deletions in HQ ARRAY, suggesting that even the unfiltered call set from the other platform can be used to effectively exclude false positives from the putatively high-quality calls.

4.6 Summary

My results demonstrate that schizophrenia-risk CNVs can be effectively identified in both WES and array data, such that only one type of data may be used for a study if this is the primary aim. My results suggest that CLAMMS is more sensitive to small (<100kb) variants than PennCNV, though orthogonal evidence is required to confirm this. In addition, I have corroborated previous findings that large (<100kb) deletions, particularly those affecting LoFi and NDD-risk genes, impact cognitive in schizophrenia. My results also suggest that small (>100kb) deletions impact cognition in schizophrenia, though reproducing my analysis in a larger sample will be required to draw a more robust conclusion. There is limited evidence to suggest that combining WES and array data improves power to detect CNVs impacting cognition, as WES call sets tested separately consistently produced stronger effect sizes. In section 5, I report a follow-up study in which I tested the impact of central deletions in 22q11.2 locus on cognition in a non-clinical cohort, following my findings that they were associated with current cognitive deficits in Cardiff COGS.

5. Follow-up study: 22q11.2 central deletions in the UK Biobank

5.1 Introduction

5.1.1 22q11.2

In section 3.5.1 I described four rare, large (>100kb) deletions impacting NDD-risk genes that were associated with current cognitive deficits in SCZ. All four occur within the proximal q arm of chromosome 22 (22q11.2), a locus consisting of ~3mb that is prone to structural variation due to the occurrence of 4 low copy repeat regions (LCRs), referred to as LCR22A–D (Karbarz, 2020). Homologous non-allelic recombination at these LCR regions can generate at least 5 CNVs of varying size. These are grouped into two categories based on their position within the locus: proximal and central. A-D proximal events are the largest subtype, spanning the entire locus. Proximal A-B spans 50% of the locus, while proximal A-C spans ~66%. The central locus is nested at the distal end of the proximal locus. Central B-D is ~30% the size of A-D, while central C-D is ~50% of B-D (Karbarz, 2020). Figure 5.26 shows the whole locus annotated with known CNV deletions and intersecting genes, coloured by NDD-risk status (purple) and high loss-of-function intolerance (red) according to criteria described in section 2.9.

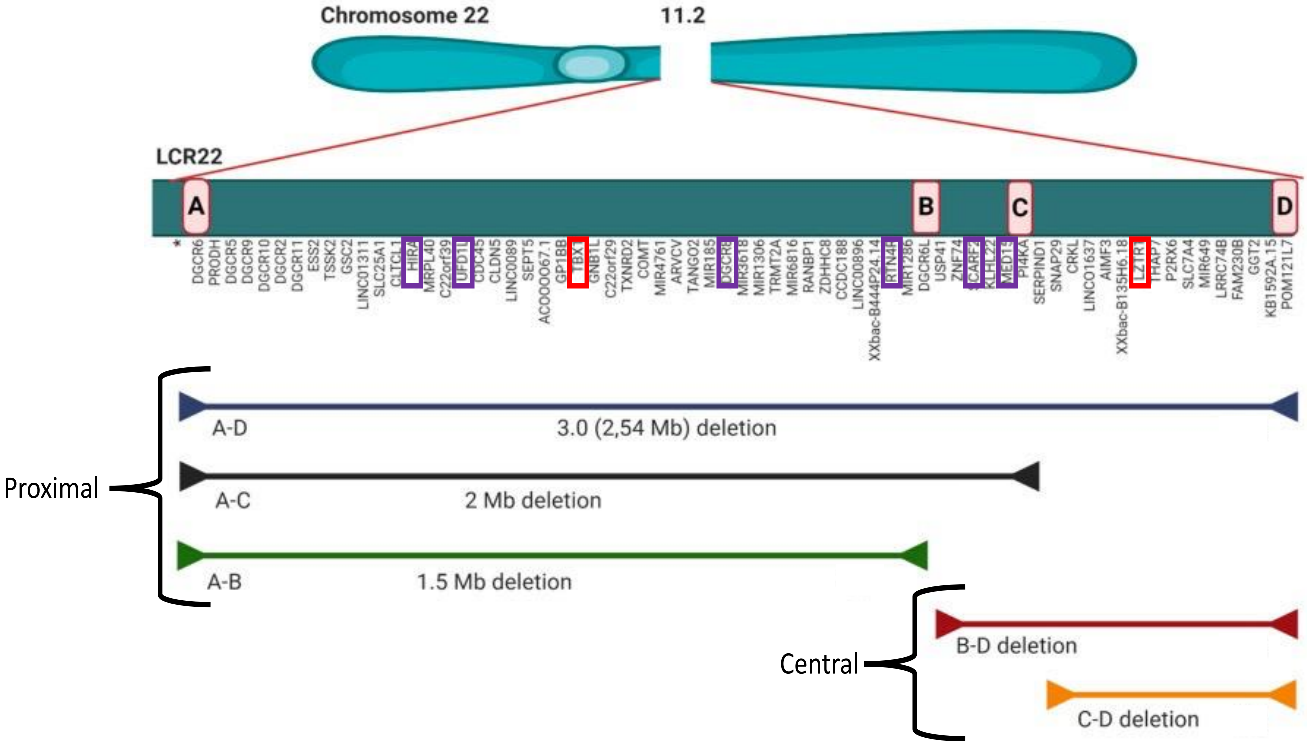


Figure 5.26: The 22q11.2 locus, annotated with gene positions, LCRs and reported deletion events, grouped by proximal and central types. Genes in purple boxes are neurodevelopmental disorder risk genes, while those in red boxes are loss-of-function intolerant genes, defined according to criteria described in section 2.X. This Figure is adapted from (Karbarz, 2020).

Deletion of the proximal A-D locus is the primary cause of DiGeorge Syndrome (DGS), whose symptoms include multiple neurological, morphological, and cardiac abnormalities (Cirillo et al., 2022). Haploinsufficiency of genes *HIRA*, *TBX1*, and *COMT* is thought to be critical in DiGeorge Syndrome (DGS) aetiology (Gothelf et al., 2004; Merscher et al., 2001; Ye et al., 2021). The central locus does not include the critical DGS genes. However, (Burnside, 2015) reviewed symptom reports of 45 carriers of central deletions in previous clinical literature in addition to 23 carriers of this event recruited though her own lab (total n = 68) and reported a less severe DGS-like phenotype. The most common features were growth restriction (16/68 (24%)), developmental delay (16/68 (24%)), intellectual disability (17/68 (25%)), language delay (15/68 (22%)), and dysmorphic features (31/68 (46%)). Of the 35 central deletion carriers that underwent follow-up analysis, 14 (40%) were found to have inherited the central deletion event, a stark contrast with the ~90% *de novo* rate for proximal deletions.

5.1.2 Aims

Two of the deletions identified in Cardiff COGS are 2.5MB instances of the proximal A-D event, while the other two are 673KB instances of the central B-D event. Given the association of the B-D deletions with current cognitive deficits in Cardiff COGs, and Burnside's finding of cognitive deficits as the most common clinical features of carriers, I investigated whether this event also impacts cognition and functional outcomes in the UK Biobank (Sudlow et al., 2015), a much larger population sample (n = 502,485). I also used the UK Biobank dataset to undertake secondary analysis to determine the association of the central deletions with 5 neuropsychiatric phenotypes in which cognitive symptoms are prevalent: schizophrenia ((Owen et al., 2016), bipolar disorder (Bora & Pantelis, 2015), anxiety disorders (Beesdo et al., 2010), major depressive disorder (Rock et al., 2014), and neurodevelopmental disorders (Craig et al., 2016).

5.2 Methods

5.2.1 UK Biobank

The UK biobank (UKB) is large data resource based on the UK population. It includes >500,000 participants, for which multiple environmental, demographic, health exposures have been recorded, many of which are relevant for the study of mental health disorders. Most participants were recruited between 2006 and 2010, though a minority have been recruited since. The version of UKB used in the present study is from February 2021 and includes 502,485 participants aged 40-69 years at recruitment. The mean age at recruitment is 56 years and 54% of participants are female.

5.2.2 Cognition and functional outcomes

UKB participants have been tested for cognitive function using a battery of tests designed to measure separate cognitive domains, carried out at UK Biobank recruitment centres. A subgroup also completed online follow-up tests. For testing association with 22q11.2 carrier status, I only selected those tests that had been completed by at least ~20% of participants, following the methods of (Kendall et al., 2017). Details for each selected test, including the number of participants available in the February 2021 version of the data, are given in Table 5.18.

Test	Cognitive domain	Description	N participants (%)
Pairs Matching	Episodic memory	Participant is shown 6 pairs of cards for 3 seconds, then asked to identify matching pairs among overturned cards. The total number of errors is the outcome measure, transformed by log + 1	420882 (84%)
Reaction Time	Processing speed	Participant is shown two cards simultaneously and are asked to press a button as quickly as possible if both are the same. Outcome measure is the log-transformed mean reaction time of correct responses.	418406 (83%)

Fluid Intelligence	Reasoning/problem solving	Participant asked to complete as many verbal and numerical reasoning questions as possible within 2 minutes. Total number of correct answers is used as the outcome measure.	134610 (25%)
Digit Span	Numeric working memory	Participant shown progressively longer strings of numbers on a screen, then asked to recall them once they had disappeared. Length of longest recalled string is used as the outcome measure.	92445 (18%)
Symbol Digit Substitution	Complex processing speed	Participant is required to match numbers with symbols according to a key shown at the top of the test page. The correct number of substitutions is used as the outcome measure.	102118 (20%)
Trail Making A & B	Visual processing speed	In test A, participant is required to connect 24 randomly positioned numbered circles in ascending order as quickly as possible. In Test B, the circles' labels alternate between letters and numbers. The log-transformed time taken to complete each test are used as outcome measures.	90165 (18%)

Table 5.18 Cognitive metrics analysed in this study. 'N participants' refers to the number of UK Biobank participants for which test data was available in the February 2021 version.

Educational, vocational, and economic outcomes are highly associated with cognitive ability (Deary et al., 2007; Strenze, 2007), and are collectively referred to as 'functional outcomes'. I tested three such measures that had data available for most participants: educational qualifications, household income and the Townsend Deprivation Index. Descriptions of each are given in Table 5.19, including the number of participants for which data was available in the February 2021 version of the data.

Functional Outcome	**Description**	**N participants (%)**
Educational qualifications	Participant asked to specify which of the following qualifications they have attained: 1) College or University degree; 2) A levels/AS levels or equivalent; 3) O levels/GCSEs or equivalent; 4) CSEs or equivalent; 5) NVQ or HND or HNC or equivalent; 6) Other professional qualifications e.g. nursing/teaching. Outcome measure is a binary variable, with participants who had attained 1) or 2) as their highest-level qualification coded as ‘1’.	395545 (79%)
Household income	Participant asked the average total income, before tax, received by their household. Options are 1) Less than £18,000; 2) £18,000 to £30,999, 3) £31,000 to £51,000; 4) £52,000 to £100,000; 5) Greater than £100,000. Outcome measure is a binary variable, in which individuals who responded 4) or 5) are coded as ‘1’.	362527 (72%)
Townsend Deprivation Index	An index of social deprivation, assigned to participants based on their post code at recruitment. For a given post code, it is calculated according to four census metrics: percentage households without a car, percentage over-crowded households,	420777 (84%)

	percentage households not occupied by owner, and percentage unemployed persons. Overcrowding and unemployment are log-transformed, then all metrics are standardised and summed. A higher index score indicates a greater degree of social deprivation.	

Table 5.19. 3 functional outcomes analysed in this study. ‘N participants’ refers to the number of UK Biobank participants for which data was available in the February 2021 version.

5.2.3 Neuropsychiatric disorders

UKB includes ICD-10 participant diagnoses for many neuropsychiatric phenotypes, derived from at least one of: primary care data, hospital admissions data and death registers. As a separate analysis, I tested 5 neuropsychiatric phenotypes for which there is robust evidence of association with cognitive deficits: schizophrenia, bipolar disorder, anxiety disorders, major depressive disorder, and neurodevelopmental disorders. The anxiety disorders metric combines two anxiety disorder subtypes: ‘phobic anxiety disorders’ and ‘other anxiety disorders’, the latter of which includes panic disorder and generalised anxiety disorder. The neurodevelopmental disorders metric also combines two disorder subtypes: specific developmental disorders of speech and language, and specific developmental disorders of scholastic skills. Table 5.20 gives the number of cases of each neuropsychiatric disorder in the February 2021 version of the data.

Neuropsychiatric disorder	**N cases (%)**
Schizophrenia	1,351 (0.3)
Bipolar disorder	2,205 (0.4)
Major depressive disorder	58,414 (11.6)
Anxiety disorders	15,347 (3.1)

Neurodevelopmental disorders	98 (0.02)

Table 5.20. Number of cases of each neuropsychiatric disorder analysed in this study, according to ICD-10 diagnoses derived from at least one of multiple types of health record.

Prior to testing association with central deletion carrier status, I determined how many carriers had been diagnosed with at least one of these five disorders, and disorder prevalence in carriers vs non-carriers across the UKB.

5.2.4 22q11.2 deletions in UK Biobank

22q11.2 deletion events were previously called from array data by Kim Kendall and George Kirov, according to protocols described in (Kendall et al., 2017). In total, they identified 47 carriers of proximal and central deletions. N carriers of each CNV subtype are given in table 5.21.

Deletion subtype	Interval (mode)	Size (mean)	N carriers (%)
Proximal A-B	22:18876630-20311646	1.2mb	7
Proximal A-C	-	-	0
Proximal A-D	22:18876630-21505360	2.6mb	5
Central B-D	22:20457855-21505360	818kb	15
Central C-D	22:21052014-21505360	433kb	22

Table 5.21: N carriers of each 22q11.2 deletion subtype previously called in the UK Biobank.

As (Kendall et al., 2017) were interested in the association of known pathogenic and SCZ risk CNVs with cognitive ability and functional outcomes in UKB, they only considered the proximal A-D subtype. The authors note that one would expect about 37 A-D carriers in a sample of this size, given that the incidence in new-borns is approximately 1:4000, while they only detected 5. They argue that the likely explanation for this is the UKB recruitment strategy. As the median age of death for individuals with DGS is 46.4 (Van et al., 2019), many carriers are likely to have died younger than recruitment age-range 40-69. The low number of carriers meant that were too few who were participants in all 7 of the cognitive tests they investigated to produce valid models.

5.2.5 Statistics

I tested the impact of central deletion carrier status (i.e either subtype, n = 37), central B-D deletion carrier status (n = 15) and central C-D carrier status (n = 22) on the cognitive, functional, and neuropsychiatric outcomes. Using R package speedglm, I generated linear models for all cognitive test scores and Townsend Deprivation Index. Binary models for educational qualifications, household income and all neuropsychiatric outcomes were generated using package logistf, which mitigates bias produced from the low carrier numbers using Firth's penalized likelihood approach. Age at recruitment, age squared, sex, and the first ten principal components were included as covariates in all models. The model for educational qualifications included an additional binary covariate based on age to account for CSE introduction in 1965, produced by binning participants according to whether this qualification was available to them at 15 years of age. I applied an FDR correction of 10 for the 7 cognitive tests and three functional outcomes, and an FDR correction of 5 for the 5 neuropsychiatric outcomes.

5.3 Results

5.3.1 Cognitive ability

All point estimates for the effect sizes of carrier status on cognitive tests scores trended in the expected direction for lower cognition (Figure 5.27). Central deletion carrier status was nominally associated ($p < 0.05$) with lower performance in Pairs Matching (beta = 0.44, nominal p = 0.016, n carriers = 31) and Symbol Digit Substitution tests (beta = -1.51, nominal p = 7.0 x $10^{-4,}$ n carriers = 4). Testing the deletion subtypes separately, only C-D carrier status was nominally associated with lower performance in Pairs Matching (beta = 0.47, nominal p = 0.046, n carriers = 19) and Symbol Digit Substitution tests (beta = -1.73, nominal p = 7.8 x 10^{-4}, n carriers = 3). After FDR correction, central deletion carrier status and C-D carrier status were still associated with lower performances in the Symbol Digit Substitution test (central: q = 7.0 x 10^{-3} ;C-D: q = 7.8 x 10^{-3}). Effect sizes and 95% confidence intervals for all cognitive tests are shown in Figure 5.27.

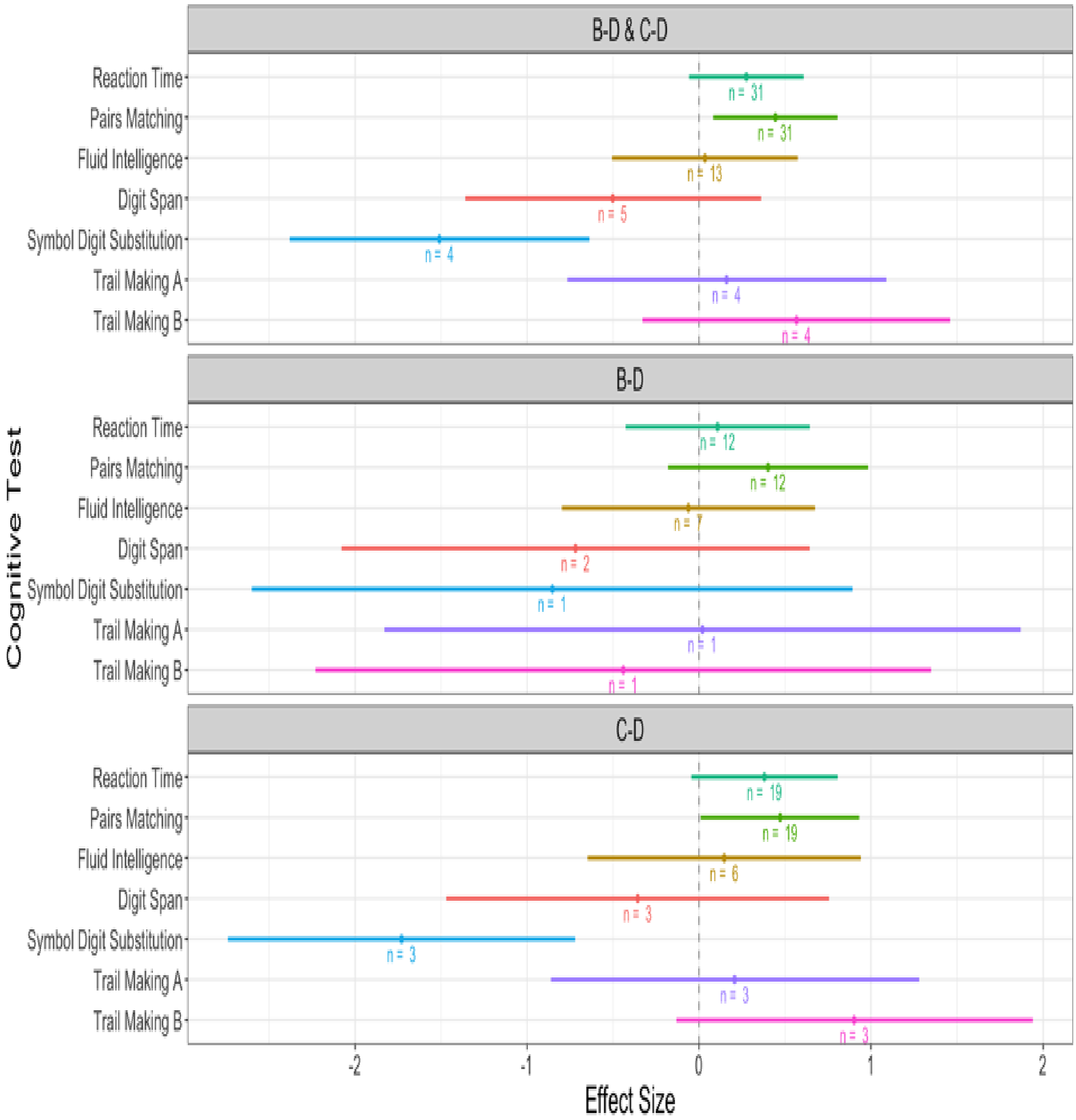


Figure 5.27: Association between central 22q11.2 deletions and 7 tests of cognition in the UK Biobank. B-D = deletion of the central B-D interval, C-D = deletion of the central C-D interval. ‘n’ refers to the number of carriers of each event type that were tested.

5.3.2 Functional outcomes

Central deletion carrier status was nominally associated with lower educational qualifications level (beta = -1.02, p = 0.019, n carriers = 30) lower household Income (beta = -1.62, p = 0.029, n carriers = 28) and higher Townsend Deprivation Index score (beta = 0.54, p = 2.3 x 10-3, n carriers = 31) (Figure 5.28). B-D deletion carrier status was nominally associated with lower household income (beta = -2.25, p = 0.02, n carriers = 11), while C-D deletion carrier status was nominally associated with higher TDI score (beta = 0.55, p = 0.015, n carriers = 19). After FDR correction, central deletion carrier status was associated with Townsend Deprivation Index score (q = 0.028). Effect sizes and 95% confidence intervals for all functional outcome tests are shown in Figure 5.28.

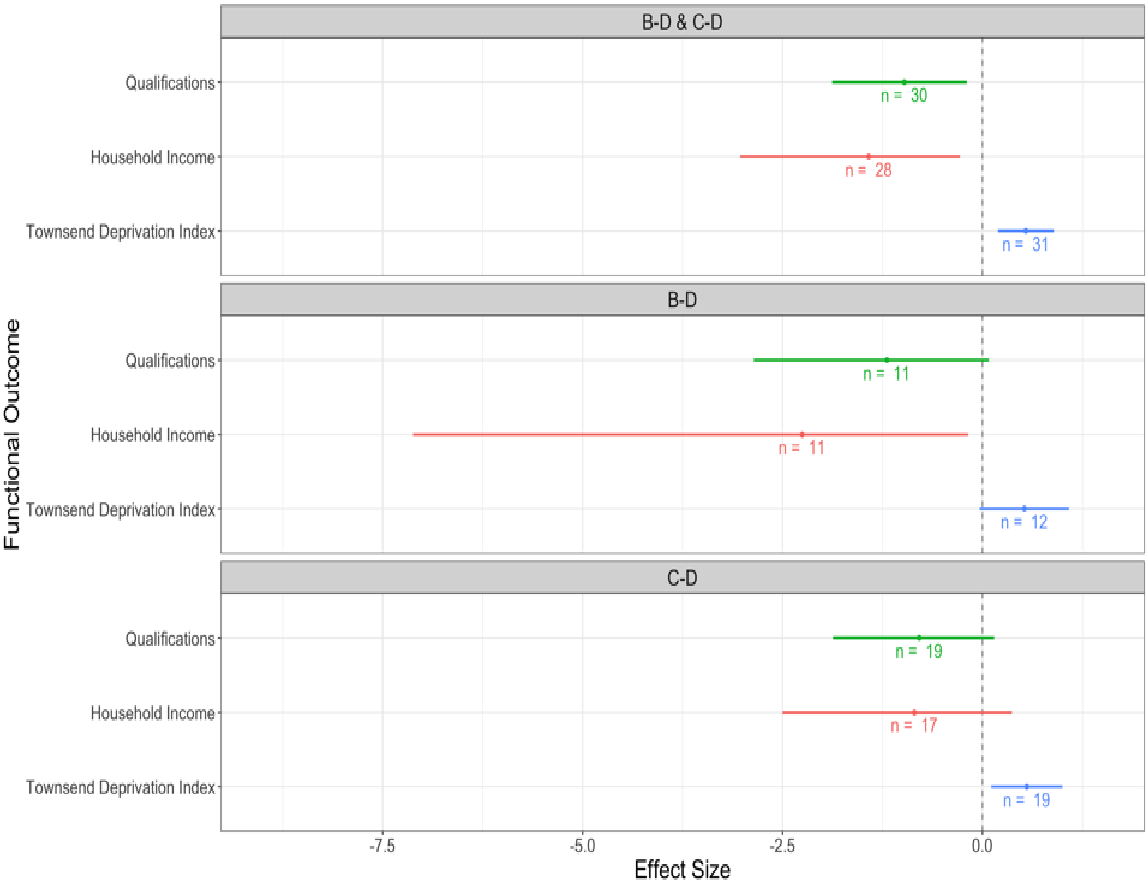


Figure 5.28: Impacts of central 22q11.2 deletions on 3 functional outcomes in the UK Biobank. B-D = deletion of the central B-D interval, C-D = deletion of the central C-D interval. ‘n’ refers to the number of carriers of each event type that were tested.

5.3.3 Neuropsychiatric disorders

Among all central deletion carriers, 7/37 (19%) reported/diagnosed with an AD, 8/37 (22%) reported/diagnosed with MDD, while 3/37 were comorbid for an AD and MDD.

There were 0 cases of SCZ, BPD or NDD among carriers, though these disorders are rare in UKB, and for 25/37 (68%) none of the five disorders were reported or diagnosed. Across the whole cohort, AD and MDD case status was more prevalent among carriers than non-carriers, and being reported/diagnosed with none of the five disorders was more prevalent among non-carriers (Figure 5.29)

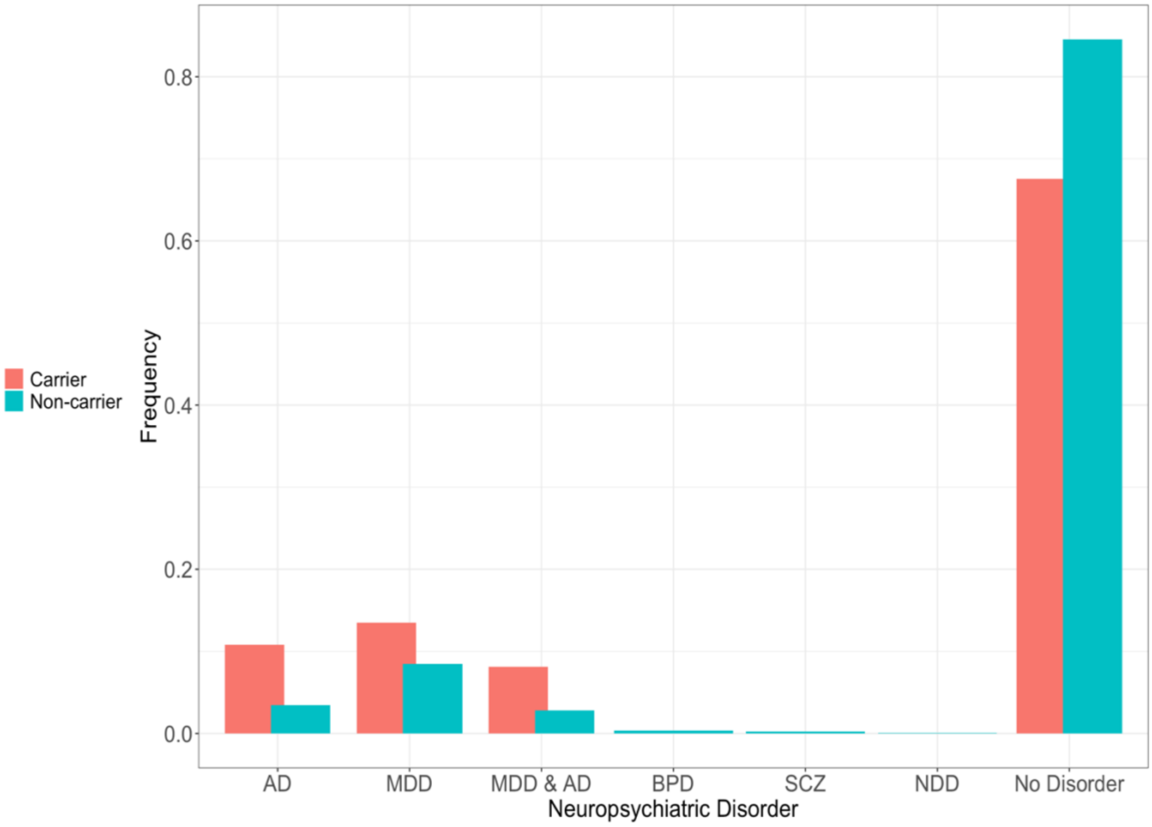


Figure 5.29: The frequencies of 6 neuropsychiatric disorders and their comorbidities among carriers and non-carriers of central 22q11.2 deletions in the UK Biobank. AD = anxiety disorders, MDD = major depressive disorder, BPD = bipolar disorder, SCZ = schizophrenia, NDD = neurodevelopmental disorders.

Central deletion carrier status was nominally associated with AD (beta = 1.32, p = 4.0 x 10-3, n = 31) and B-D deletion individually is nominally associated with AD (beta = 1.66, p = 0.013). After FDR correction, only central deletion carrier status was associated with AD (q = 0.02). Effect sizes and 95% confidence intervals for AD and MDD tests are shown in Figure 5.30.

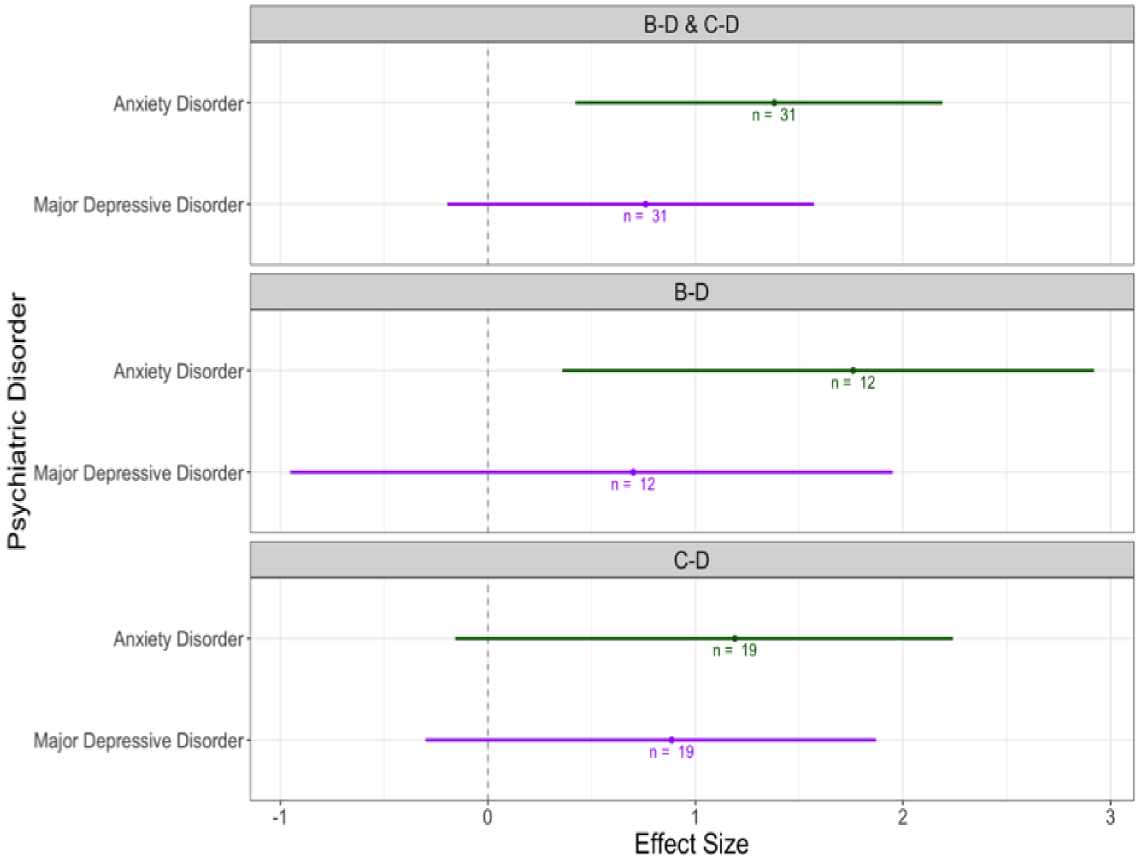


Figure 5.30: Associations of central 22q11.2 deletions with anxiety disorder and major depressive disorder case status in the UK Biobank. B-D = deletion of the central B-D interval, C-D = deletion of the central C-D interval. ‘n’ refers to the number of carriers of each event type that were tested.

5.4 Discussion

I found that 22q11.2 central deletion carrier status is significantly associated with worse performance in the Symbol Digit Substitution test, a measure of complex speed processing. However, confidence in this result is limited by the fact that there were only four carriers who had scores for the test. Moreover, carrier status was correlated with deficits across all cognitive tests, suggesting that the impact of this CNV is not limited to a particular cognitive domain, and that the absence of significant association with other test scores is a function of low power. My results therefore appear to corroborate RD Burnside’s report of cognitive deficits among carriers. I also found that central deletion carrier status was nominally associated with all three functional outcome metrics I tested, though after FDR correction was only associated with Townsend Deprivation Index score. The results provide further evidence that this CNV negatively impacts a broad range of cognitive abilities, as these functional outcomes are themselves highly correlated with general measures of cognition. Twenty-eight of 31 carriers had data for all three metrics, such that power is less of a limitation than in the case of some of the cognitive metrics.

Anxiety disorders, major depressive disorders, and their comorbidity were found to be more prevalent among central deletion carriers than non-carriers. There were too few cases of schizophrenia, bipolar disorder and neurodevelopmental disorders to enable a comparison, An absence of any of these five neuropsychiatric disorders was found to be more prevalent among non-carriers. Carrier status was significantly associated with anxiety disorder diagnosis, and the effect size for major depressive disorder was approaching significance. It is therefore possible that the central deletion is causative of neuropsychiatric symptoms, though given the small numbers of cases for more severe phenotypes I cannot conclude whether its effects are limited to the milder end of the disorder spectrum. It is also not clear whether the impact of the CNV on neuropsychiatric symptoms is mediated by, or independent of, its effects on cognition; or, conversely, whether its effects on cognition are in part mediated by neuropsychiatric symptoms.

The findings presented in this section provide preliminary evidence that the central deletion increases risk for psychiatric disorders. However, the ascertainment biases that are associated with the UK Biobank sample significantly lower my study's power to detect significant associations between genetic factors and psychiatric disorders. Therefore, future research could test the association between the central 22q11.2 deletion and schizophrenia using large schizophrenia case-control samples (Marshall et al., 2017; Rees et al., 2014). As deletion of the larger A-D 22q11.2 region is strongly associated with schizophrenia, it is reasonable to hypothesise that overlapping deletions of the B-D/C-D loci may also be associated with the disorder, albeit with lower penetrance. Moreover, it is also possible that duplication of the central loci is protective for schizophrenia. It is important for future work to refine the penetrance estimates for the different 22q11.2 CNV breakpoints for schizophrenia, and/or cognitive deficits, as this would inform clinical management of people who present with these CNVs.

Chapter 6: Investigating the contribution from small structural variants to cognition in schizophrenia

1.Introduction

1.1 Background

As I described in section 1.1 of chapter 5, several studies have demonstrated the impact of CNVs on cognition in schizophrenia (Foley et al., 2020; Hubbard et al., 2021; Thygesen et al., 2021). However, none have studied the impact of small structural variants. In Chapter 5 I reported evidence that deletions <100kb in size negatively impacted estimated premorbid cognitive ability in Cardiff COGS participants. In the present chapter I further my investigation of the impact of small SVs on cognition schizophrenia, by assessing whether SVs called by InDelible in this samples are also associated with cognitive deficits. Given small SVs have been shown to contribute to developmental disorders by the InDelible developers (Gardner et al., 2021), and other types of small mutation (SNVs/indels) are associated with lower cognition in schizophrenia (Creeth et al., 2022), I hypothesized that small SVs called by InDelible will also contribute to cognitive deficits in schizophrenia.

In a previous study of Cardiff COGS, colleagues at Cardiff University reported an association between burden of ultra-rare point mutations under selective constraint (URCVs) and current cognitive ability (β = −0.18; p = .005) (Creeth et al., 2022). URCVs are defined as singletons that are either protein-truncating variants (PTVs) in LoFi genes with a gnomAD probability of loss-of-function (pLI) scores ≥ 0.9 25) or damaging-missense variants with a constrained pathogenicity classification (MPC) ≥ 2, that do not occur in gnomAD's non-neuro data set. The study also found that URCVs that occur within 348 NDD-risk genes described in (Satterstrom et al., 2020), (Singh et al., 2020) and (Kaplanis et al., 2020), have a larger impact on both current and premorbid cognition than URCVs that do not occur within these genes (Table 6.1), though the difference between the effect sizes was not statistically significant.

Cognitive measure	Constrained variant set	N Variants	Effect size (SE)	Z-Test
Premorbid IQ	In NDD genes	51	−0.26 (0.14)	0.43
	Non-NDD genes	341	−0.1 (0.05)	
Current cognition	In NDD genes	52	−0.36 (0.18)	0.42
	Non-NDD genes	348	−0.16 (0.07)	

Table 6.1. Impact of ultra-rare coding variants in neurodevelopmental disorder risk (NDD) genes on cognition in Cardiff COGS. The differences in effect size between those variants in NDD and not in NDD risk genes was evaluated using a Z-test, the p-values for which are given in the last column. SE = standard error.

1.1 Study aims

Conducted in the second year of my PhD, the primary aim of the present study was to investigate whether rare (allele-frequency < 1%), small (<1kb) structural variants (SVs) that are typically missed using microarray technology contribute to cognitive impairments in schizophrenia (SCZ). SVs of this size are under-reported in the literature and there have been no studies investigating association between small SVs and cognition in schizophrenia. To test whether small SVs impact cognition in schizophrenia, I used InDelible to call SVs from whole exome sequencing (WES) data generated from the Cardiff Cognition in Schizophrenia (COGS) cohort DNA samples (n = 927). Individuals in this schizophrenia sample have been assessed for current cognitive ability and estimated premorbid cognitive ability, allowing for detection of variants that may affect deficits on cognition that predate disorder onset or are a consequence of schizophrenia progression. I therefore tested whether rare SVs discovered by InDelible were associated with measures of current cognitive ability and estimated premorbid cognitive ability.

Given previous findings from this sample that show damaging rare coding variants in constrained genes or known NDD genes are associated with lower cognition, I hypothesized that rare deletions in these genes would have the greatest effects on cognition in schizophrenia. To increase the power of the current study, I performed a secondary analysis where I jointly analysed deletions with the damaging rare coding variants that have been previously called in the current sample and evaluated whether this increases power to detect associations between rare variants and lower cognition in schizophrenia.

2. Methods

2.1 Sample description

The samples used in the current study are the same as those described previously in chapter 5, section 2.1. Thus, the recruitment protocols, sample size and sequencing procedures were the same as those specified that study.

2.2 Cognitive phenotypes

The cognitive phenotypes used in the current study are the same as those described previously in chapter 5. Briefly, I tested two measures of cognition: 1) current cognition and 2) estimated premorbid IQ for association with the burden of rare SVs.

2.3 Calling structural variants

SVs were called from BAM files using InDelible (Gardner et al., 2021), described in section 3 of chapter 2. Given that Cardiff COGS is not a trios sample, I omitted the 'denovo' step and only applied the first 5 steps of the algorithm: Fetch, Aggregate, Score, Database and Annotate. As the algorithm does not require modelling aspects of the data that differ systemically between the subcohorts (e.g coverage depth), variants were called for all samples in the same run. InDelible requires a configuration file as input, specifying parameters that are used for variant processing at different stages. I configured the algorithm to exclude reads with mapping quality < 5, base quality < 10, and SR length < 5 at the Fetch step, and to exclude SR clusters containing <3 SRs at the Aggregate step. These thresholds were recommended by the InDelible developers and were also applied in the SCZ trio analysis reported in chapter 4.

2.4 Quality control

2.4.1 Initial InDelible output

The initial InDelible output contains 119,191,134 calls across all samples, 111,976,650 of which were in the Broad subcohort and 7,214,484 were in the Cardiff subcohort. No sample had 0 calls. The former subcohort is 1.2x larger than the latter but has 15.5x the number of calls. InDelible assigned an SV type to 750,776 (0.67%) of the Broad subcohort calls and 219,612 (3.04%) to the Cardiff subcohort calls and calculated a size for 162,145 (0.14%) of the Broad subcohort calls and 37,673 (0.52%) of the Cardiff subcohort calls.

2.4.2 Random forest quality score

In the Score step, InDelible uses a random forest adaptive learning model score each call according to its probability of being a real event, the details of which are described in section 3.3.3 of chapter 2. This score is output as the metric ‘prob_y’. In chapter 4, the chosen prob_y threshold for the trio analysis was > 0.6. However, when this threshold was applied to the Cardiff COGS WES data, the number of SVs called was too high for them to all be manually inspected (>5000 calls). Given manual inspection of the reads that map to an SV is a critical step in SV quality control, I applied a more stringent prob_y threshold of > 0.8, which excluded 103,931,750 SVs from the initial call set (88.8% of the initial SVs from the Broad subcohort, and 74.1% of the initial SVs from the Cardiff subcohort).

2.4.3 Coverage depth and number of calls

As the InDelible authors found in the DDD sample that sample coverage depth was highly correlated with the number of InDelible calls exome-wide, I assessed the correlation between these metrics in the Cardiff COGs data after application of the prob_y filter. Figure 6.1 shows a strong positive correlation between sample coverage depth and number of calls, which explains much of the discrepancy between the subcohorts. However, this figure also shows that there is significantly greater variance in the Broad subcohort than in the Cardiff that is not explained by coverage depth.

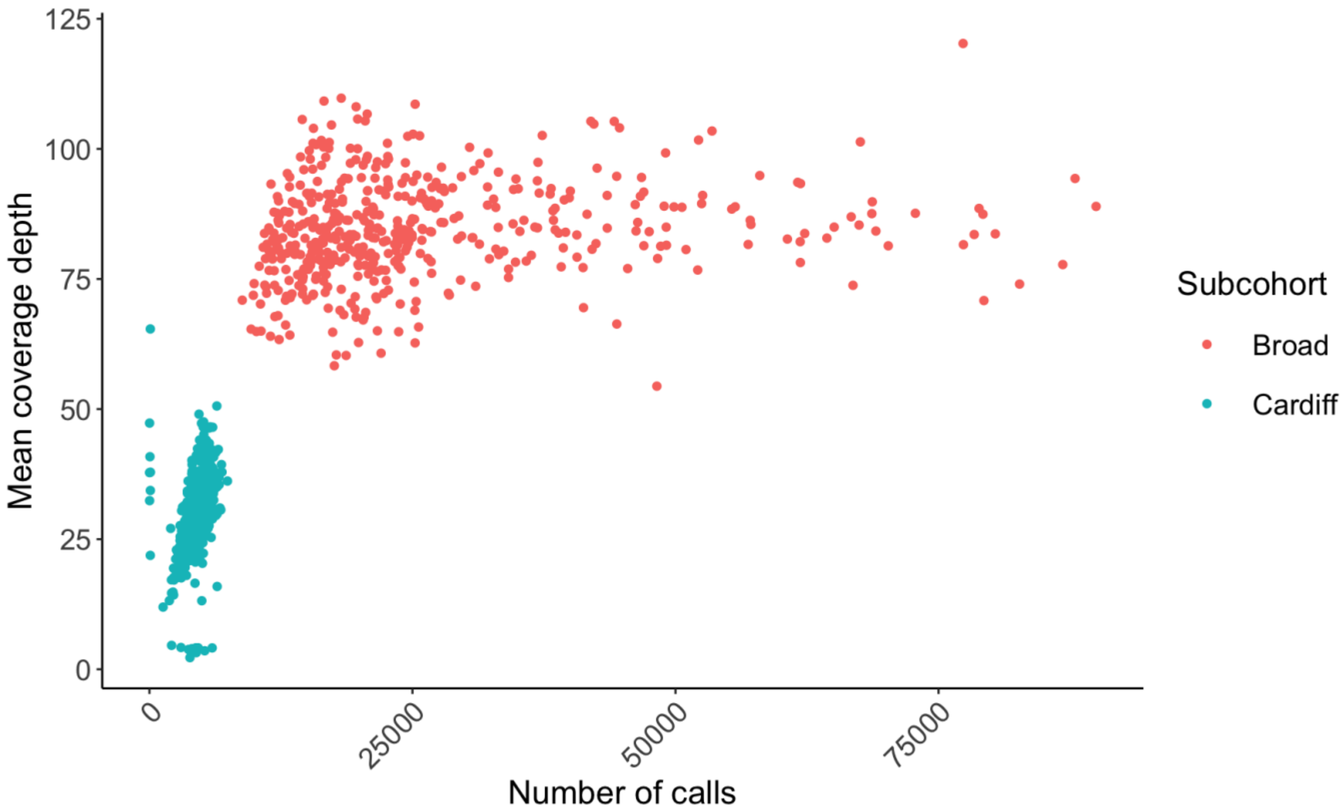


Figure 6.1 Mean coverage depth plotted against number of InDelible calls.

2.4.4 Sample-level quality control

The data used in the current study had previously undergone sample-level QC as described in (Creeth et al., 2022). This included excluding samples that did not have a diagnosis of schizophrenia, schizoaffective disorder, or other non-affective psychotic disorder (n = 52). Additionally, samples were excluded if their inferred sex did not match their expected sex or were in a second-degree, or closer kinship (Creeth et al., 2022). I applied the following additional sample-level QC that is based on the distribution of SVs called per individual by InDelible. This QC was separately applied to the two subcohorts, given the large discrepancy in number of calls between them. Summary statistics for the number of calls for each subcohort's call set, following application of the prob_y quality filter, are shown in Table 6.2. In the Broad subcohort, the number of calls followed a right-skewed distribution with outliers only at the upper end (Figure 6.2). The chosen threshold for sample exclusion was > 72,182 calls, equal to the mean + 3 standard deviations. In the Cardiff subcohort, the number of calls followed a normal distribution with outliers only at the lower end (Figure 6.3), though no samples had 0 calls. The chosen threshold for sample exclusion was < 1,800 calls. Nine samples were excluded from the Broad subcohort and 9 from the Cardiff subcohort.

Subcohort	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**SD**
Broad	8824	15815	20914	25992	29630	89972	15396.8
Cardiff	11	3582	4412	4348	5166	7416	1197.1

Table 6.2. Number of SV summary statistics for the Broad and Cardiff subcohorts, following application of prob_y quality score filter to the initial InDelible output. SD = standard deviation

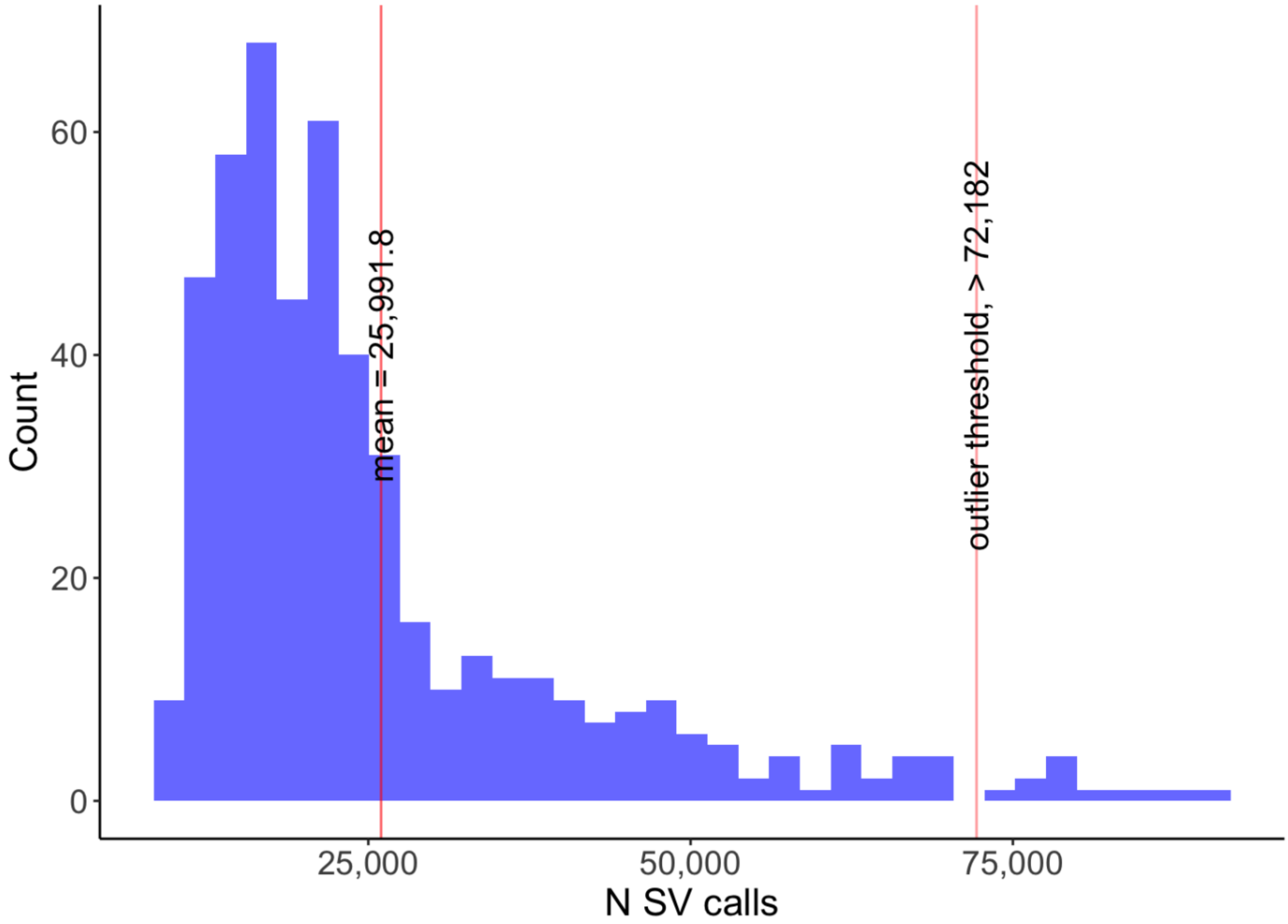


Figure 6.2. Number of SV calls distribution for the Broad subcohort, annotated with mean and outlier threshold.

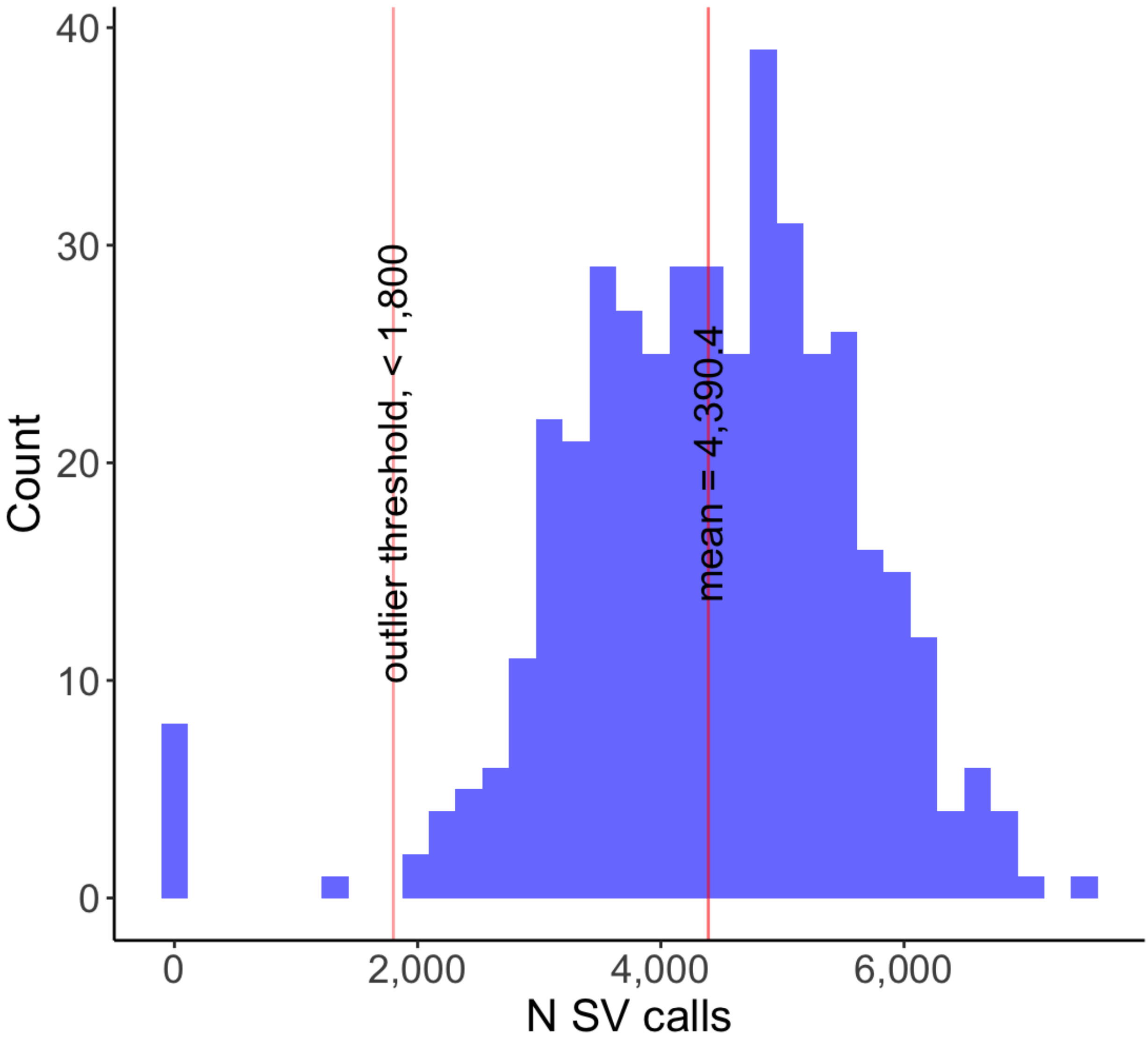


Figure 6.3. Number of SV calls distribution for the Cardiff subcohort, annotated with mean and outlier threshold.

2.3.5 Variant-level quality control

In the Annotate step, SV calls were annotated with their frequencies in three allele-frequency (AF) databases: the database based on Deciphering Developmental Disorders study probands (N = 13,438) included in the InDelible code repository, the database produced during my InDelible analysis of the SCZ trios that is based on parents (N = 1,219) , and a database based on all 927 Cardiff COGS participants. AF databases for the samples I analysed were generated using InDelible's Database function, whose parameters are described in section 3.3.4 of chapter 2. Calls were filtered if their call position had an AF > 0.01 in one or more of the databases. 14,155,039 common calls were thereby excluded.

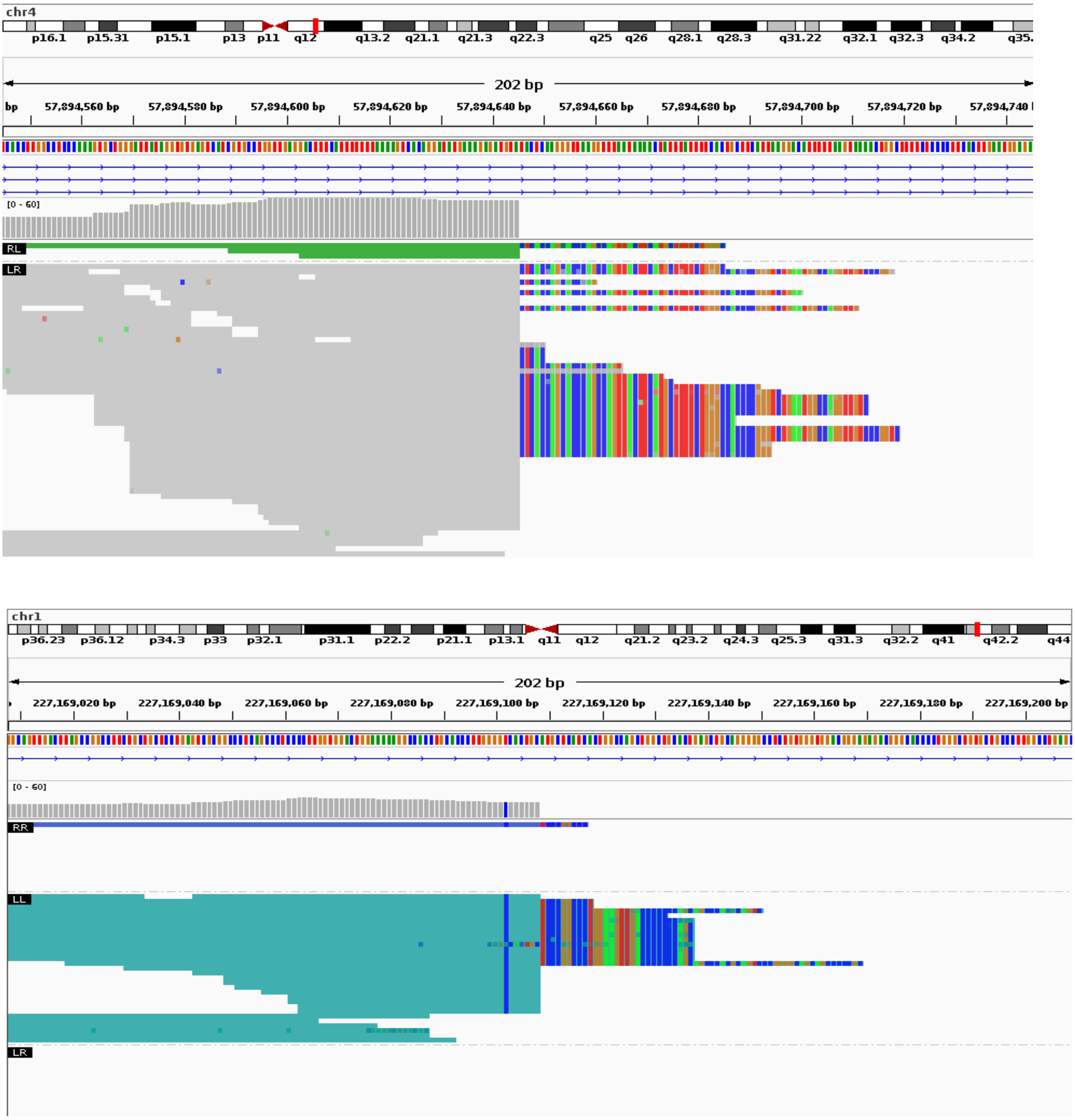
Four additional variant-level QC criteria were applied to all samples, following QC recommendations outlined in (Gardner et al., 2021), and which were previously applied in the SCZ trios study described in chapter 4. First, SV calls also needed to be based on SR clusters whose average mapping quality was >=20, ensuring that the aligned bases if each read were unlikely to be mis-mapped. Second, calls were excluded if > 10% of SRs in their respective cluster were split at both ends, but only if they did not also have a valid BLAST hit. 'Double split' reads are likely to be caused by errors during sequencing but in some cases can result from retrotransposition of the of the types of repetitive sequences (e.g. MEIs) that are detected by InDelible's implementation of BLAST. Third, calls were then excluded if they were aligned to non-exonic sequences, according to human reference genome GRCh37, thereby filtering those that are the consequence of read misalignment or non-exonic DNA contamination. Finally, remaining calls were excluded if they had < 5 SRs in their corresponding cluster.

Collectively, these criteria excluded a further 95,212 calls, leaving 588 calls to be manually inspected. Of these calls, 319 were in the Broad subcohort, and InDelible had assigned an SV type to 104 (32.6%) calls and calculated the size of 95 (29.8%) calls. For the remaining 269 calls in the Cardiff subcohort, InDelible had assigned an SV type to 134 (49.8%) calls and calculated the size of 132 (49.1%) calls.

2.3.5 Manual inspection

Upon manual inspection, there were three criteria by which calls were excluded:

1) Forty-one calls in the Broad subcohort had a single SR cluster that is preceded by a total decrease in coverage at the breakpoint, such that there are no reads mapped to any sequence in the locus downstream of the call position. Inspecting instances in the IGV interface revealed that there were often no mapped reads for several kb, after which coverage will abruptly normalise. The soft-clipped bases in the SR cluster typically align to the locus after the loss of coverage and are often in an atypical orientation to their mate. Ten instances were called deletions by InDelible, 6 as duplications and the rest as unknown. The algorithm also calculated a size for 18, the median of which was 4143.5. IGV snapshots of 2 instances are shown in Figures 6.4 a & b.


chr4
p16.1 p15.31 p15.1 p13 p11 q12 q13.2 q21.1 q21.3 q22.3 q25 q26 q28.1 q28.3 q31.22 q32.1 q32.3 q34.2 q35.
202 bp
57,894,560 bp 57,894,580 bp 57,894,600 bp 57,894,620 bp 57,894,640 bp 57,894,660 bp 57,894,680 bp 57,894,700 bp 57,894,720 bp 57,894,740
[0 - 60]
RL
LR
chr1
p36.23 p36.12 p34.3 p33 p32.1 p31.1 p22.2 p21.1 p13.1 q11 q12 q21.2 q23.2 q24.3 q25.3 q31.3 q32.2 q41 q42.2 q44
202 bp
227,169,020 bp 227,169,040 bp 227,169,060 bp 227,169,080 bp 227,169,100 bp 227,169,120 bp 227,169,140 bp 227,169,160 bp 227,169,180 bp 227,169,200 bp
[0 - 60]
RR
LL
LR

Figures 6.4 a & b. IGV snapshots showing two instances of calls that were excluded according to criterion 1). In both cases a total loss of coverage is observed at the call position. In figure a (top), some reads occur in an RL orientation, which are coloured in green. The majority are in the normal LR orientation. In figure b (bottom) almost every read occurs in an LL orientation, coloured in teal.

2) There were many individual SRs surrounding the call position, often observed with many discrepancies between the mismatched bases in the called SR cluster, indicating that multiple errors were introduced at this locus during the sequencing process. It is therefore probable that any structural change corresponding the SR cluster is a consequence of these errors. Sixty-seven calls were excluded by this criterion, of which 63 were called in the Broad subcohort. An example is show in Figure 6.5

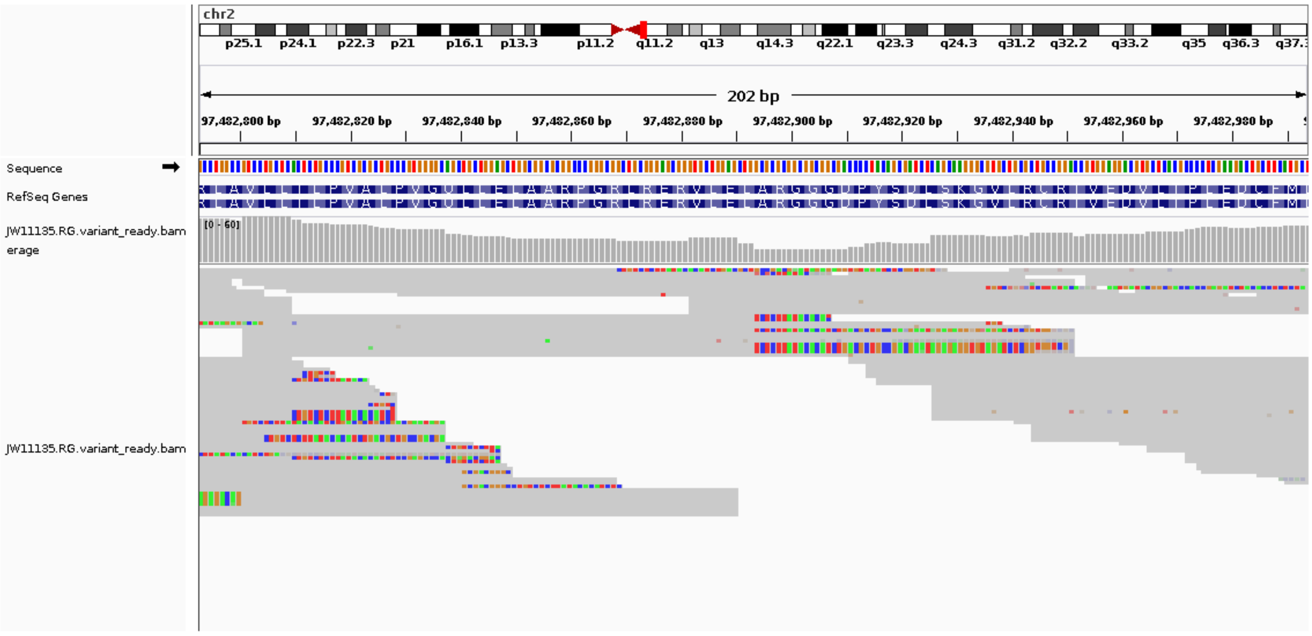


Figure 6.5. IGV snapshot showing a call that was excluded by criterion 2). There are several individual split reads surrounding the call position, indicating that errors were introduced into this locus during sequencing.

3) The SR cluster is mapped to a sequence that contains successive instances of a single base, and all the misaligned bases in the SR are also the same base. Such loci are prone to replication errors during PCR, limiting confidence that any structural change to the proband DNA is not a consequence of these errors. Two calls were excluded by this criterion, both called in the Broad subcohort. An example is shown in Figure 6.6.

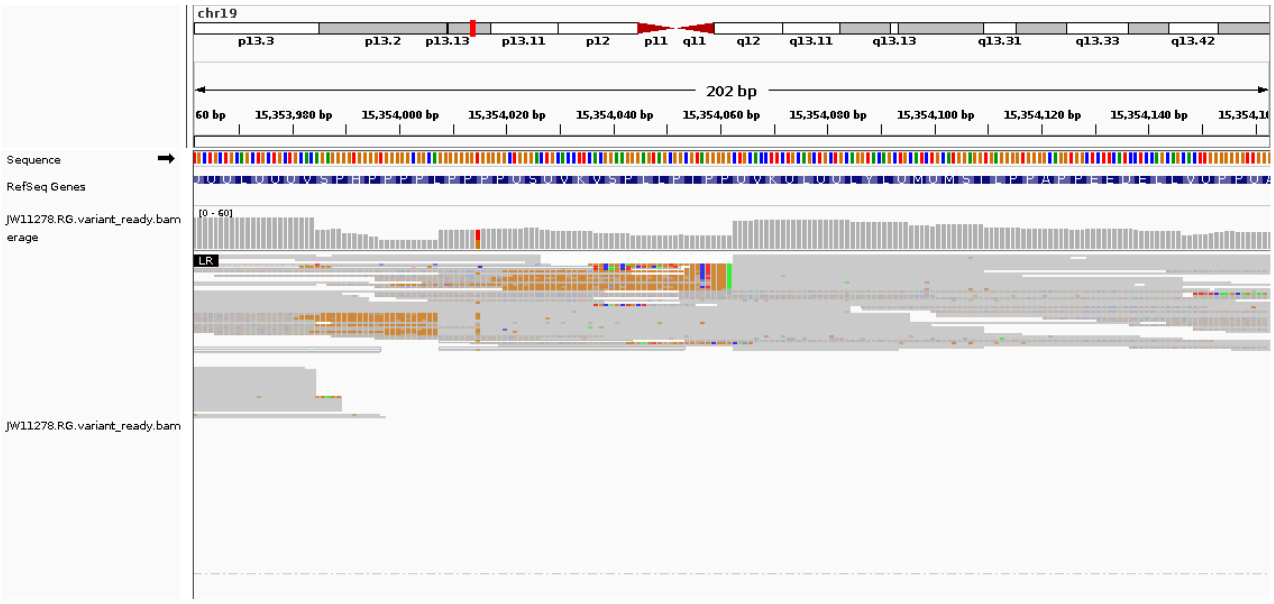


Figure 6.6. IGV snapshot of a call excluded by criterion 3. The misaligned bases contain almost exclusively one base (G), reflecting the sequence slightly upstream on the reference genome. However single base sequences are highly prone to errors during the sequencing process, and so we cannot have confidence that this call is not the result of an artefact.

2.3.5.1 Manual inspection summary

A total of 110 calls were excluded by manual inspection, of which 106 (96.4%) were called in Broad subcohort samples. I was able to determine an SV type for 138 SVs that InDelible failed to classify. Moreover, InDelible misclassified the SV type associated with 91 calls. Sixty-two calls classified as deletions were found to be among the constituent calls of pseudogene retrotranspositions and 14 were artefacts. Four calls classified as duplications were found to be complex-duplication/insertions and 11 were artefacts. One call InDelible classified as a translocation was also likely an artefact: it had BWA-mem alignment on the same chromosome, but a BLAST hit on another chromosome, and the call position was surrounded by single read splits.

I was able to ascertain the size for 134/360 (37.2%) of calls InDelible had also failed to calculate a size for. Forty-three of these were associated with pseudogene retrotranspositions, and so were annotated with the number of exons implicated in their corresponding event. The other 57 were deletions, duplications, simple insertions, small tandem repeats or complex-duplication/insertions < 110 in size (mean size = 26bp). Of the 226 calls neither InDelible nor I ascertained a size for, there were 2 deletions, 28 simple insertions, 16 non-pseudogene retrotranspositions, 1 translocation, 102 calls of unknown type and 77 artefacts.

2.3.6 Quality control summary

Sample-level and variant level quality control are summarised in Tables 6.3 and 6.4, respectively.

	Subcohort	
	Broad	Cardiff
N samples in initial output	498	429
N samples retained	489	420

Table 6.3. Summary of sample-level quality control for each subcohort. Samples were excluded for having too few or an excess of calls.

Quality control step	**N calls retained**
Initial call output	119,191,134
Prob_Y > 0.8	15,259,384
Sample-level quality control	14,250,835
Allele-frequency < 0.01	95,796
MAPQ > 20	95,796
% double split filter	87,635
Exonic	1,666
N SRs > 5	588

Manual inspection	478

Table 6.4. Summary of variant-level quality control across all samples. The second column N calls remaining after each quality control step was applied, which are specified in the first column. MAPQ = mapping quality, SR = split reads.

2.7 Structural variant annotation

The genes impacted by SVs were annotated for their estimated probability of constraint against loss-of-function mutations (pLI), derived from gnomAD (Karczewski et al., 2020). A gene is defined as loss-of-function intolerant (LoFi) if it has a pLI > 0.9. I also annotated variants that affected genes from the DDG2P database (April 2021 version) (Wright et al., 2015) that have either been confirmed to cause a neurodevelopmental disorder, or are thought to be probable candidate, and are associated with a monoallelic mode of inheritance. Genes that fit these criteria are termed NDD-risk genes (N = 726).

2.8 Statistics

Linear regression models were used to test for association between normalised current and estimated premorbid cognition scores. Covariates included were sex (coded as 1 for male, 2 for female), age at interview, age^{2,} , sequencing site (coded as 1 for Cardiff, 2 for Broad) and first 10 principal components for common SNP variation. Using the above linear regression model, the current study tested 4 classes of SV for association with cognition: 1) all SVs; 2) all SVs except those of unknown type; 3) deletions; 4) duplications. Each class of SV was tested under two allele frequency thresholds (< 1% and singletons defined as SVs observed in one sample). I also performed two gene-set analyses for SVs that affected LoFi genes and NDD-risk genes. Combining SVs with the URCVs analysed by (Creeth et al., 2022) produced two further burden metrics: 5) the sum of SVs and URCVs in LoFi genes; and 6) the sum of SVs and URCVs in NDD-risk genes. Thus, a total of 18 variant burden metrics were produced for each individual.

2.8.1 Multiple testing correction

In the primary analysis, 18 different rare variant burden metrics were tested for association with two cognitive phenotypes (current cognition and estimated premorbid cognition). Therefore, I applied FDR correction for 36 independent tests.

3. Results

3.1 Rare structural variants identified by InDelible

Four hundred and seventy-eight rare structural variants were identified, corresponding to 375 individual SV events in 290 (31.2%) individuals. Two hundred and four SVs (53.9%) were called in the Broad subcohort, and 173 (46.1%) in the Cardiff subcohort. There were 119 deletions, 58 duplications, 36 simple insertions, 9 complex-duplication/insertions, 2 small tandem-repeats, 1 translocation, 36 pseudogene retrotranspositions, 16 retrotranspositions of other types (SINE, Alu etc.), and 100 SVs whose type I was unable to determine (Figure 6.7).

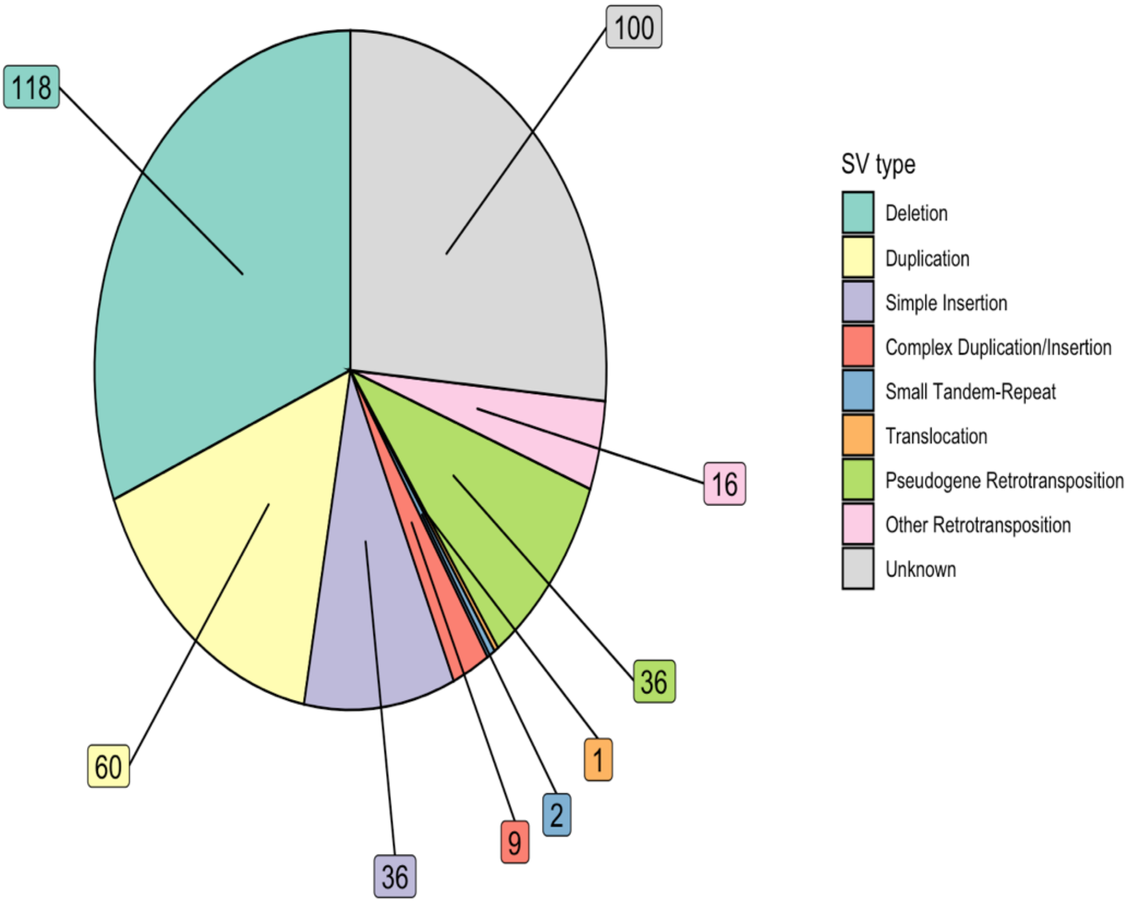


Figure 6.7. Pie chart showing the numbers of each structural variant type identified in all samples. SV = structural variant.

3.2 Structural variant size distribution

Through InDelible and/or manual inspection, the size of 288/375 (76.8%) of SVs could be determined. After excluding pseudogene retrotranspositions, the mean and median SV size was 835.7bp and 32bp, respectively, and ranged from 5bp to 57kb. Table 6.5 gives summary statistics of SV size for all variants, and for each subcohort separately. Figure 6.8 shows the sizes of SVs binned into ranges also used in the InDelible methods paper (Gardner et al. 2021).

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD	NAs
All samples	5	22	32	835.7	78	57118	5032.4	147
Broad subcohort	5	29.25	50.5	1698	281.25	57118	7519.4	99
Cardiff subcohort	6	19	25	171.2	41	7204	737	48

Table 6.5. Summary statistics for structural variant size, for all samples and by subcohort. All sizes are given in base pair length. ‘NAs’ refers to the number of SVs for which both I and InDelible failed to calculate a size. SD = standard deviation

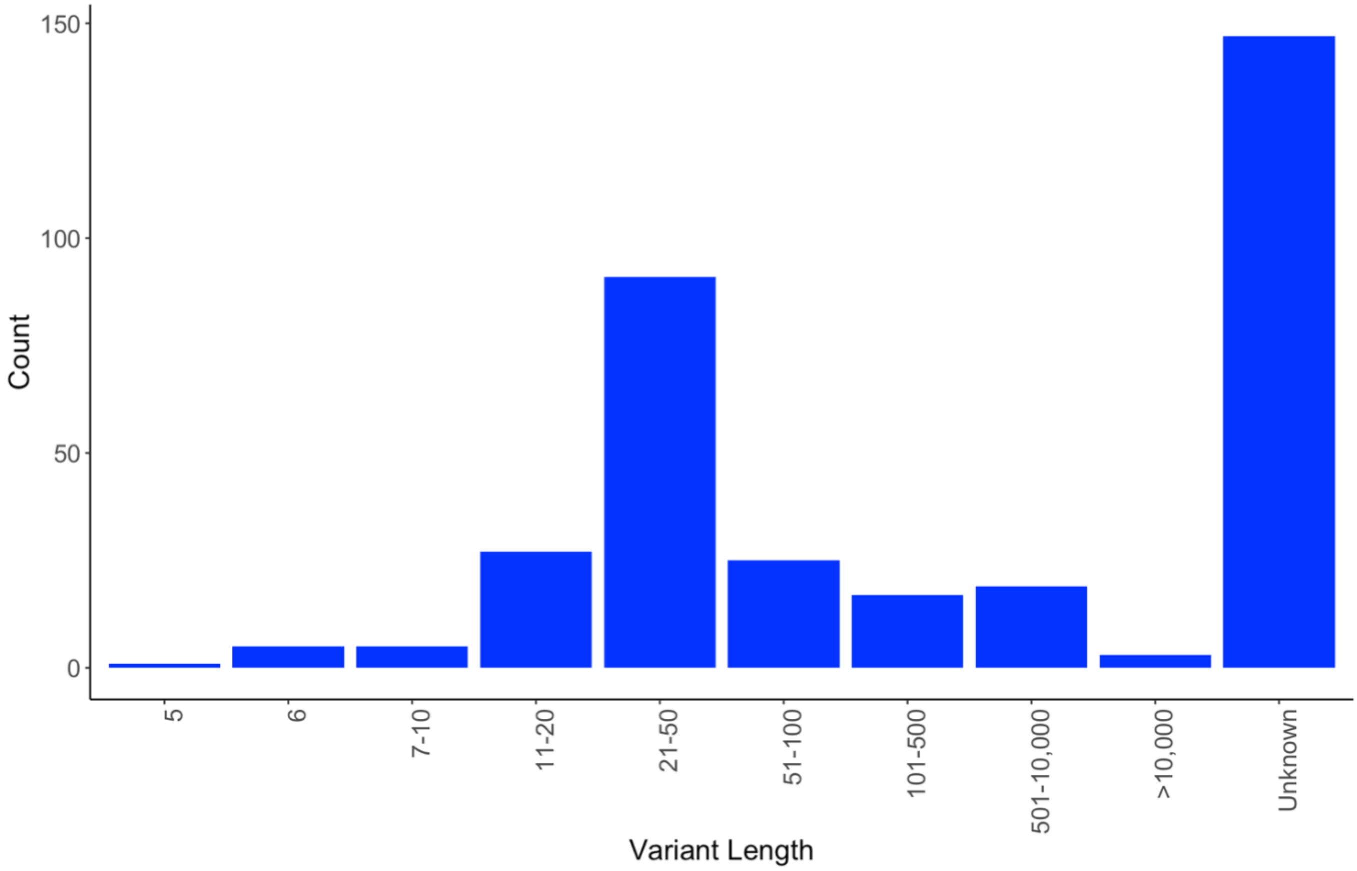


Figure 6.8. Structural variants binned into size ranges also used in the InDelible paper (Gardner et al. 2021). All sizes are given in base pairs.

I calculated the size for all 36 of the pseudogene retrotransposition events, in terms of the number of exons implicated. Eighteen of 36 were found to be 1 exon in length. Eleven events were 2 exons in length, 6 in the Broad subcohort and 5 in the Cardiff. 7 events had 4 or more exons, all of which were called in the Cardiff subcohort. The largest event consisted of 8 exons.

3.3 Examples of SR clusters for structural variants identified in the current sample

Figures 6.9-12 provide examples of the SR clusters for the following SV types: deletion, a duplication, a simple insertion, and a complex-insertion/duplication.

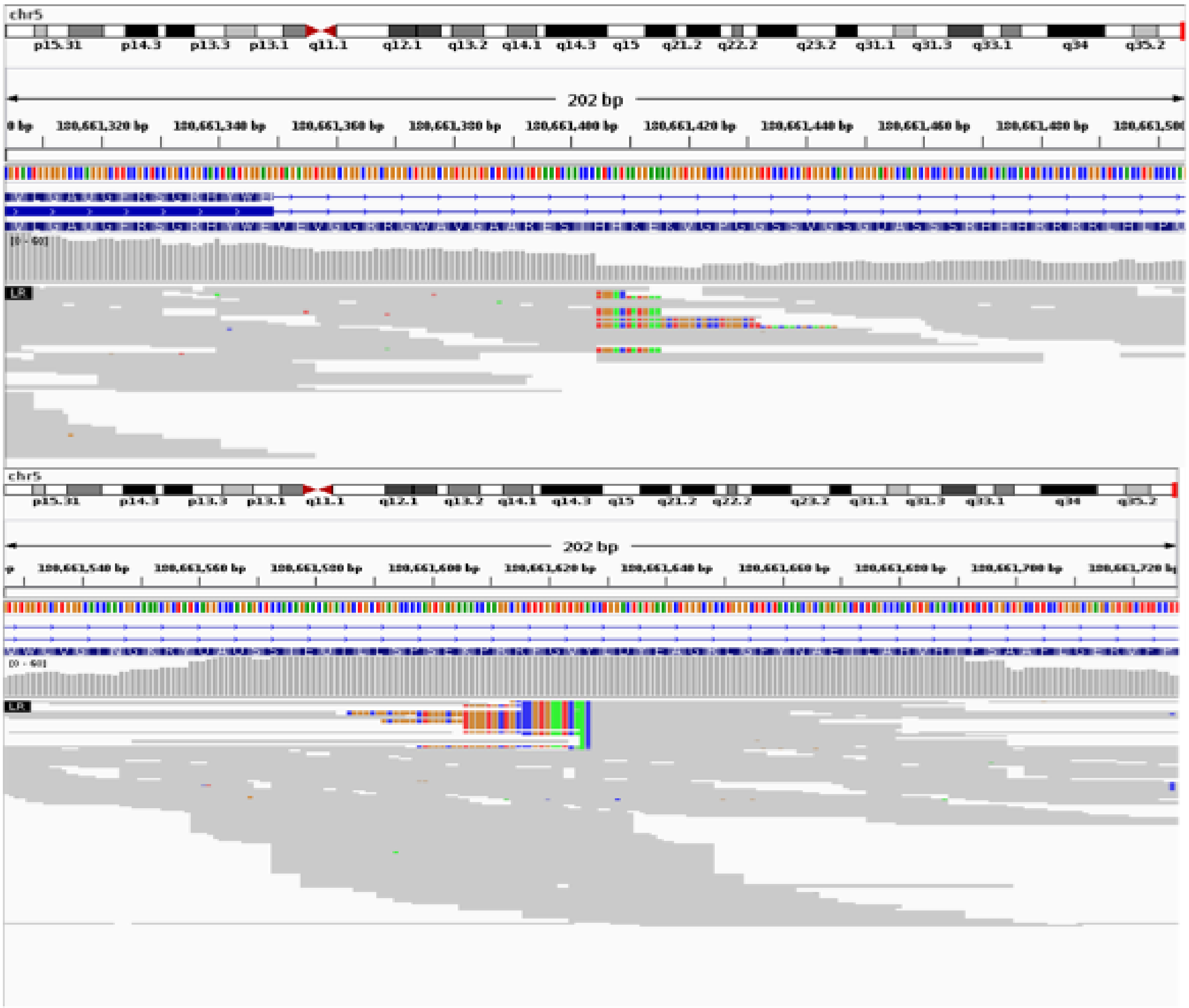


Figure 6.9. IGV snapshots of 2 SR Clusters that indicate the 5’ (top) and 3’ (bottom) breakpoints of a 223bp deletion on chromosome 5.

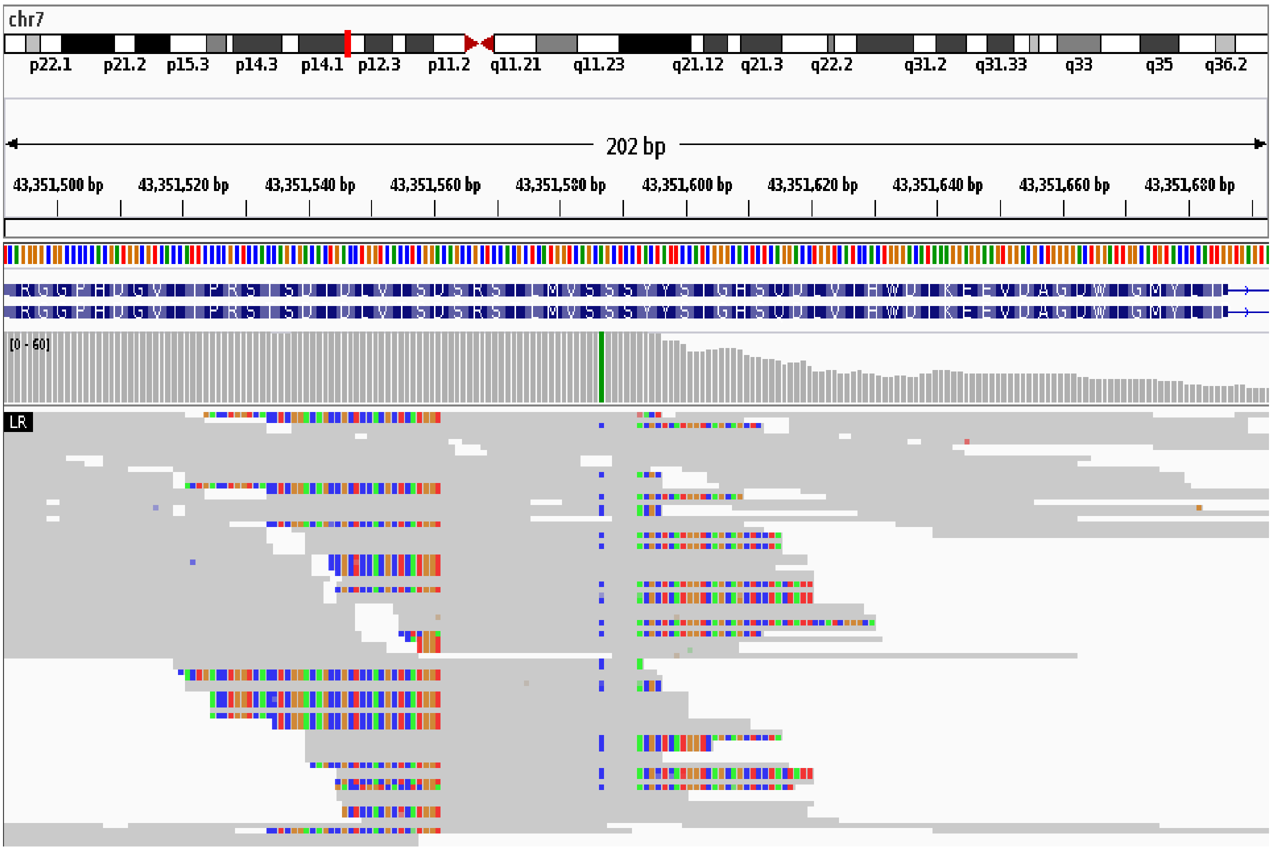


Figure 6.10. IGV snapshot showing 2 SR clusters associated with a 31bp duplication on chromosome 7. An A>C transversion can also be observed in this snapshot, within the duplicated sequence.

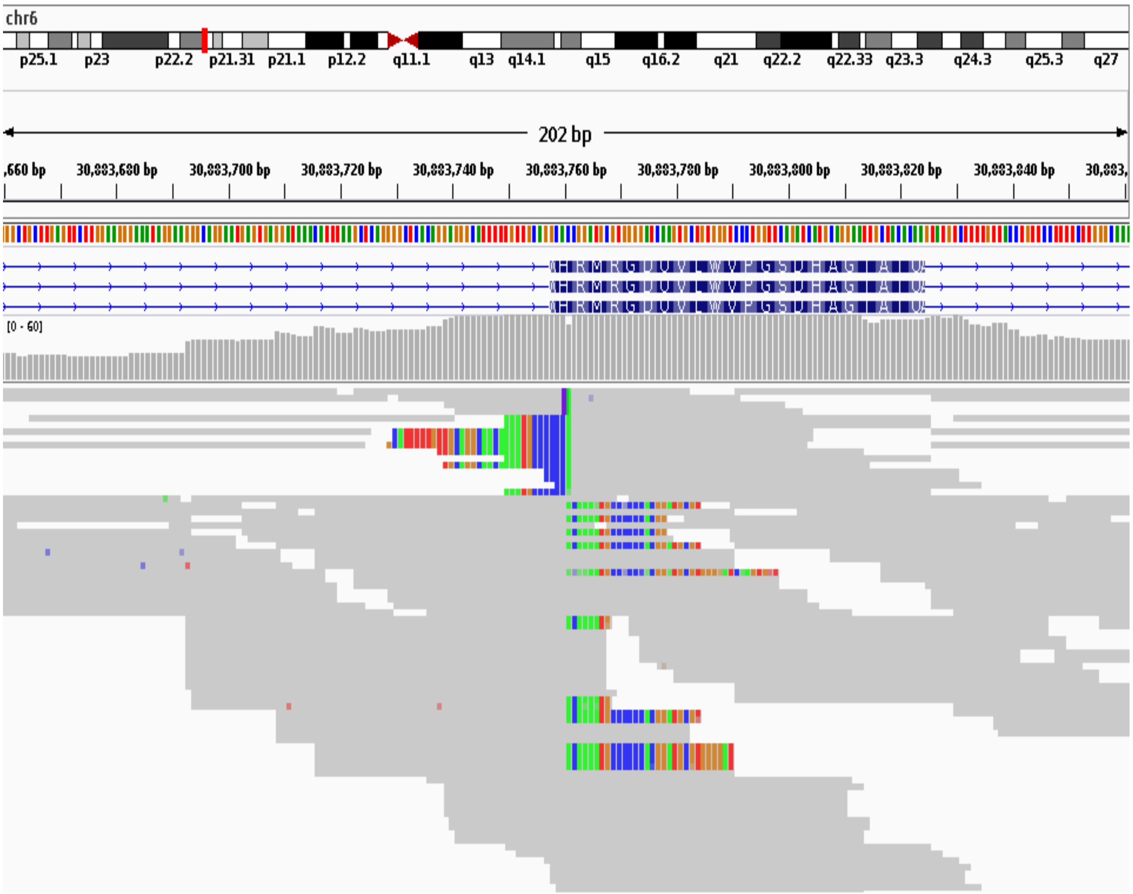


Figure 6.11. IGV snapshot showing 2 SR clusters associated with a 16bp simple insertion on chromosome 6.

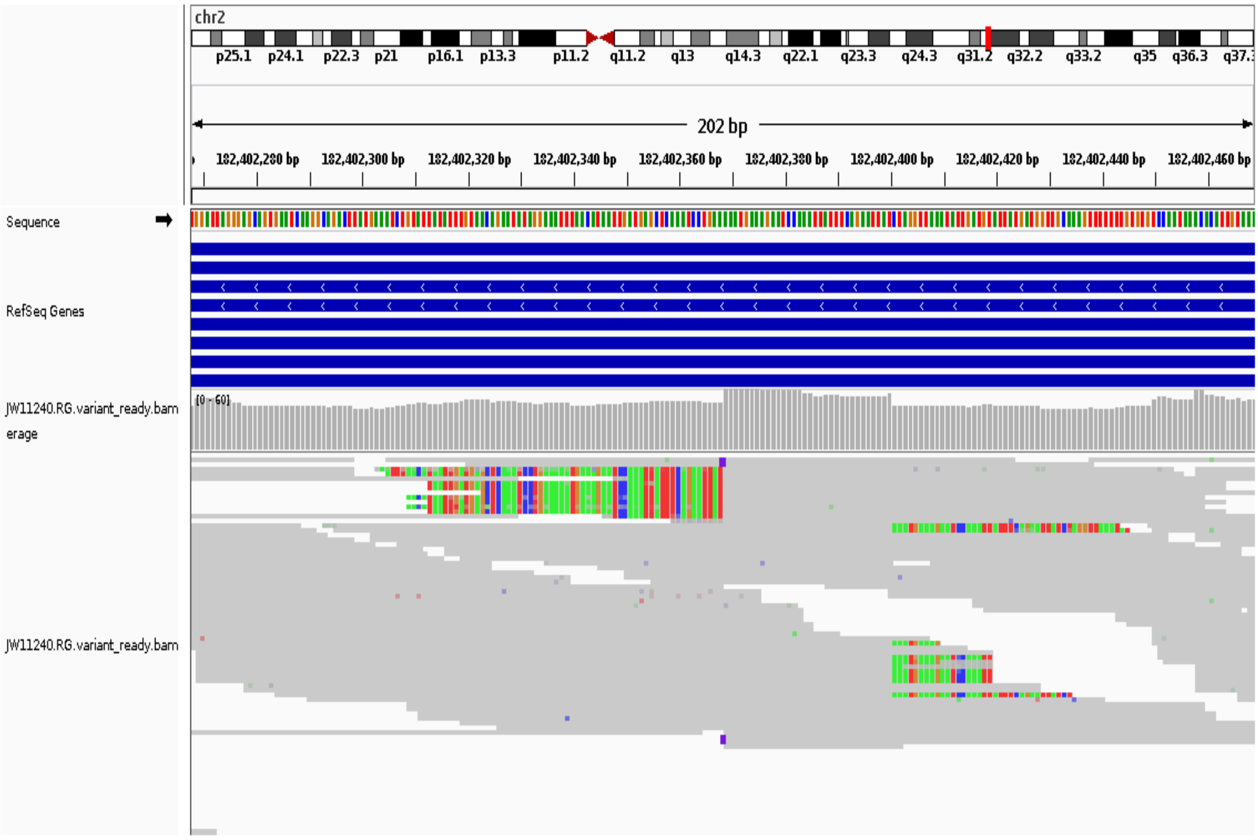

chr2
p25.1 p24.1 p22.3 p21 p16.1 p13.3 p11.2 q11.2 q13 q14.3 q22.1 q23.3 q24.3 q31.2 q32.2 q33.2 q35 q36.3 q37.3
202 bp
182,402,280 bp 182,402,300 bp 182,402,320 bp 182,402,340 bp 182,402,360 bp 182,402,380 bp 182,402,400 bp 182,402,420 bp 182,402,440 bp 182,402,460 bp
Sequence
RefSeq Genes
JW11240.RG.variant_ready.bam
erage
[0 - 60]
JW11240.RG.variant_ready.bam

Figure 6.12. IGV snapshot showing 2 SR clusters associated with a 39bp complex-duplication/insertion on chromosome 2. The duplicated sequence is 32bp in length, and the insertion, indicated by the purple marker in the top-most reads in the alignment track, is 7bp.

3.3 Comparison of InDelible and CLAMMS call sets

Thirty-five InDelible deletions and duplications were larger the 160bp, the size of the smallest CNV identified by CLAMMS. Three of these (8.6%) were detected by CLAMMS: a 7kb deletion, a 1kb deletion and 30kb duplication (Table 6.6). CLAMMS underestimated the size of all three events, indicating that their second breakpoints do not occur within exons (Table 6.6). This validation rate is consistent with the differential sensitives of the two call algorithms: InDelible is largely insensitive to variants >100bp, while CLAMMS is largely insensitive to variants <1kb. Therefore, CLAMMS cannot be used to validate large InDelible CNVs (and vice versa).

Chromosome	Type	InDelible call position	InDelible size	CLAMMS call interval	CLAMMS size
5	DUP	110447233	30,284	110427935-110447042	19,107
9	DEL	139934690	1,373	139934220-139934521	301
11	DEL	93459167	6,933	93466380-93469938	3,558

Table 6.6. CNVs identified by InDelible in current study that were also identified by CLAMMS.

I discussed in section 4.1 of chapter 4 that pseudogene retrotranspositions are unbalanced events and could therefore by mis-called as duplications by CLAMMS at the locus of their constituent exons. Two of 36 (5.6%) of the pseudogene retrotranspositions called by InDelible were called as duplications by CLAMMS: one spanned 6 exons, and the other spanned 1 exon (Table 6.7). CLAMMS overestimated the size of both as included intervening intronic regions. Moreover, the 1 exon event was called by CLAMMS as larger than the 6 exon event (Table 6.7).

Chromosome	InDelible call positions	InDelible size	CLAMMS call interval	CLAMMS size

19	5':30476211; 3':30505791	6 exons	30477186- 30506520	29,334
10	97023620	1 exon	96997679- 97031577	33,898

Table 6.7. Pseudogene retrotranspositions called as duplications by CLAMMS. 5' and 3' call positions for the chromosome 19 event refer to 5'-most and 3'-most InDelible calls.

3.4 Association between rare structural variants and cognitive ability

3.4.1 Current cognition

Figure 6.13 shows the effect sizes for all SV variant burden metrics on current cognition, including the number of variants tested and 95% confidence intervals. Figure 6.14 shows effect sizes for singleton SV variant burden metrics on current cognition. I found no nominally significant association ($p < 0.05$) between the overall burden of rare SVs and current cognitive ability (Figure 6.13; 'No Gene Set' panel). Restricting variants by gene set produced a nominally significant association for SVs impacting NDD-risk genes, but the effect size was trending in the positive direction and did it not survive multiple testing correction (beta = 0.86, p = 0.027, q = 0.98). (Figure 6.13; NDD-risk panel). Restricting the analysis specific types of SV (e.g. deletions or duplications) (Figure 6.13), or singletons (Figure 6.14), did not produce significant results. The strongest effects on lower current cognition was observed for deletions affecting NDD-risk genes (beta = -0.71, p = 0.43). However, this is based on only 2 deletion SVs were found in NDD-risk genes and therefore the confidence interval is very large (95% confidence interval: -2.4800, 1.0600)

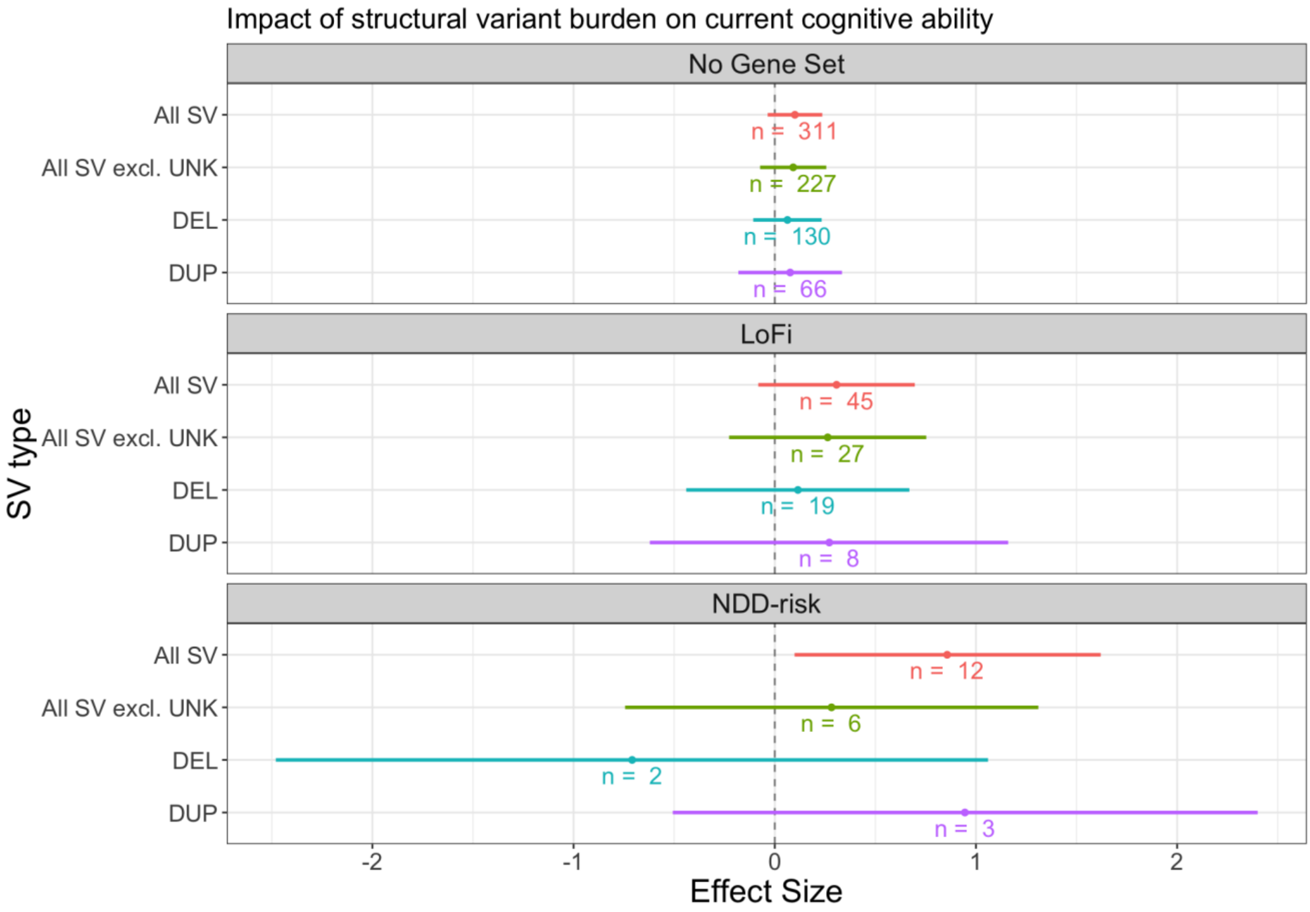


Figure 6.13 Impact of SV variant burdens, including non-singletons, on current cognitive ability. The coloured points represent effect sizes, and the horizontal lines 95% confidence intervals. SV = structural variant, UNK = unknown type, DEL = deletion, DUP = duplication, No Gene Set = variants not restricted by any gene set, LoFi = variants impacting loss-of-function intolerant genes, NDD-risk = variants impacting neurodevelopmental disorder risk genes.

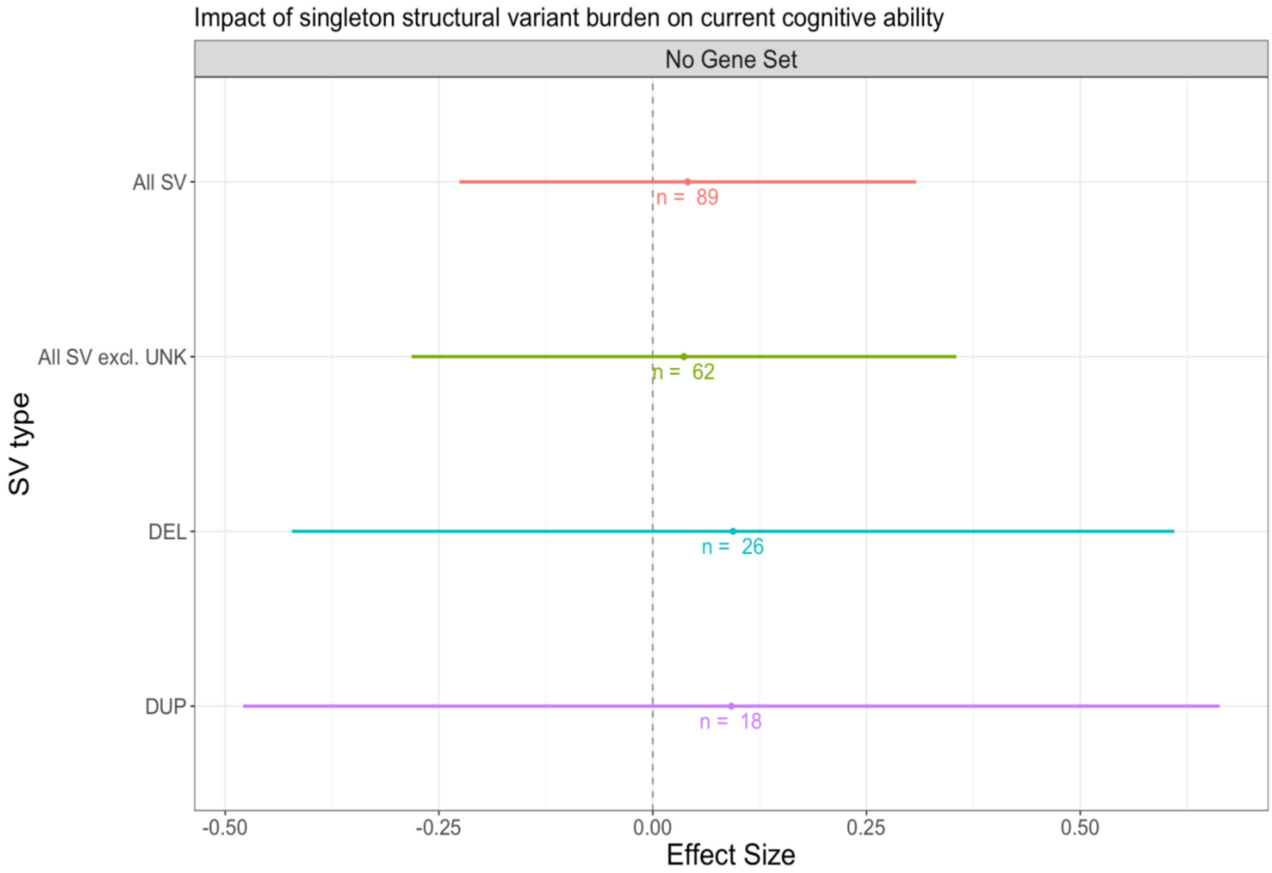


Figure 6.14. Impact of singleton SV variant burdens on current cognitive ability. The coloured points represent effect sizes, and the horizontal lines 95% confidence intervals. SV = structural variant, UNK = unknown type, DEL = deletion, DUP = duplication, No Gene Set = variants not restricted by any gene set.

3.4.2 Estimated premorbid cognition

Figure 6.14 shows the effect sizes for all SV variant burden metrics on premorbid cognition, including number of variants tested and 95% confidence intervals. Figure 6.15 shows effect sizes for singleton SV variant burden metrics on estimated premorbid cognition. Nominally significant associations were found for the overall burden of SVs (p = 0.0173), burden of SVs excluding SVs of unknown type (p 0.0229), and burden of duplications (p = 0.0455), but effect sizes were positive (overall burden: beta = 0.13; excluding unknown types: beta = 0.15; duplications: beta = 0.2) and no result was significant after multiple testing correction (overall burden: q = 0.42; excluding unknown types: q = 0.42; duplications: q = 0.45). Null results were found when restricting the analysis to singleton SVs (Figure 6.16) and SVs affecting LoFi or NDD-risk genes (Figure 6.15).

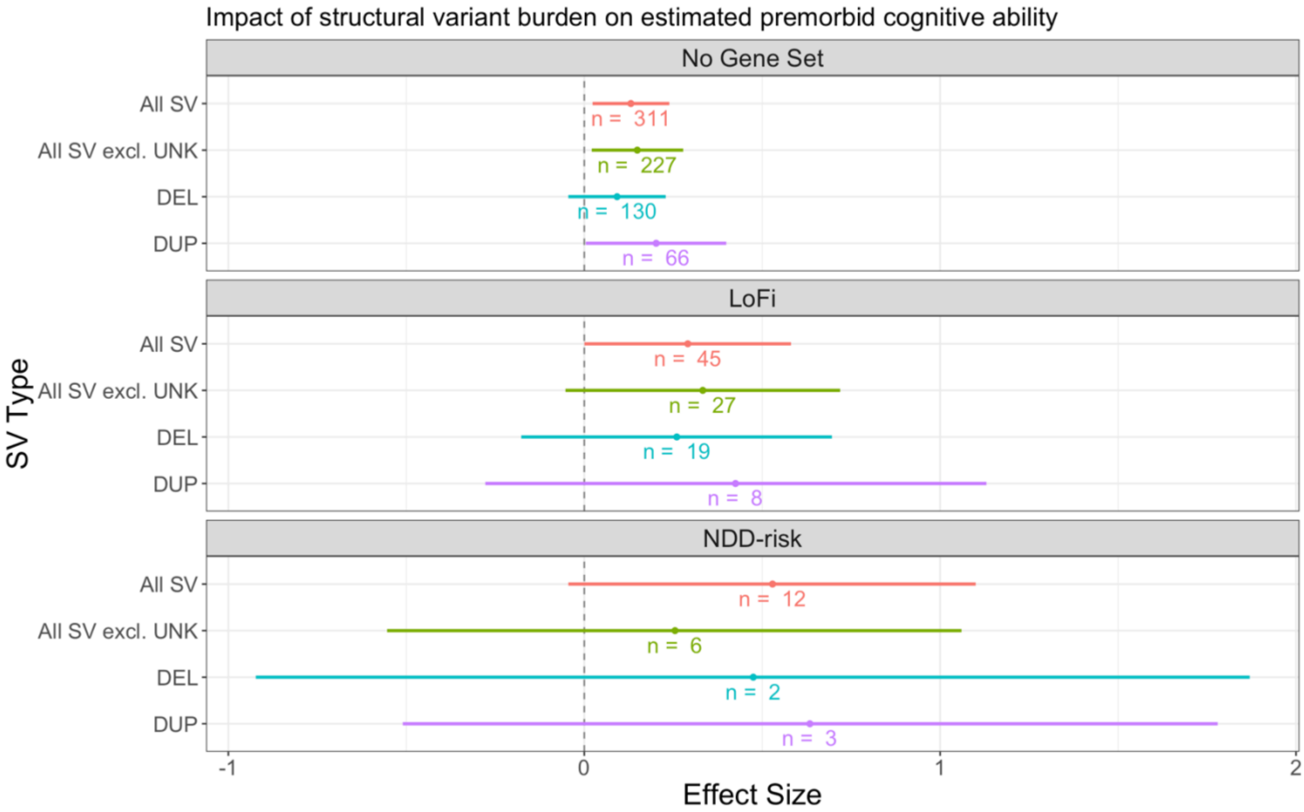


Figure 6.15. Impact of SV variant burdens, including non-singletons, on estimate premorbid cognitive ability. The coloured points represent effect sizes, and the horizontal lines 95% confidence intervals. SV = structural variant, UNK = unknown type, DEL = deletion, DUP = duplication, No Gene Set = variants not restricted by any gene set, LoFi = variants impacting loss-of-function intolerant genes, NDD-risk = variants impacting neurodevelopmental disorder risk genes.

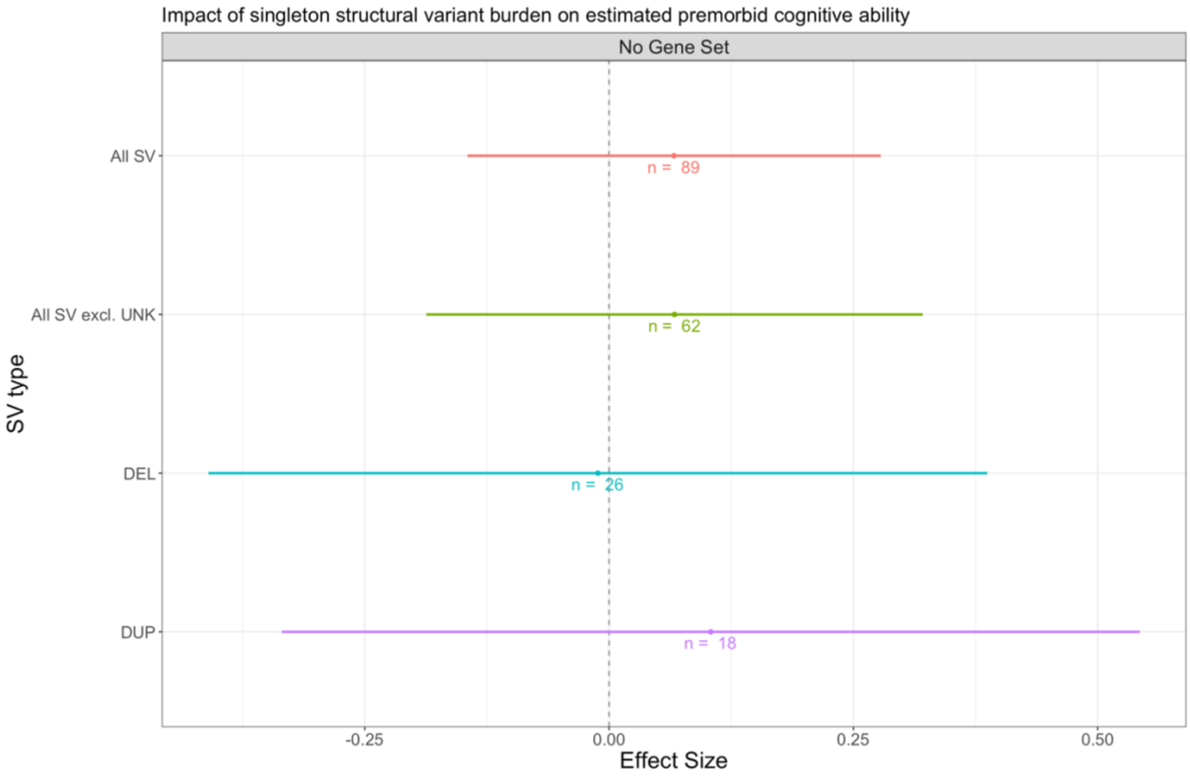


Figure 6.16. Impact of singleton SV variant burdens on current cognitive ability. The coloured points represent effect sizes, and the horizontal lines 95% confidence intervals. SV = structural variant, UNK = unknown type, DEL = deletion, DUP = duplication, No Gene Set = variants not restricted by any gene set.

3.4.3 Combined analysis of SVs and rare coding variants

In previous work, ultra-rare coding variants (URCVs) were previously associated with current cognition and estimated premorbid cognition in the current sample (Creeth et al 2021). For both cognitive measures, I performed a combined analysis of URCVs and deletions in LoFi genes and NDD genes, given deletions had the strongest point estimate effect size for lower current cognition. However, the combined analysis of URCVs and deletions in LoFi genes did not increase the negative effect sizes of the URCVs tested separately on current cognition (Figure 6.17). For current cognition, adding deletions in NDD genes to the equivalent URCV burden test did marginally increase the negative effect size (difference = -0.015), however there were only two deletions impacting NDD-risk genes (Figure 6.17). This trend was not observed for estimated premorbid cognition. Figures 6.17 and 6.18 show the effect sizes for gene set variant burden metrics tested separately and combined, for each cognitive measure, including number of variants tested and 95% confidence intervals.

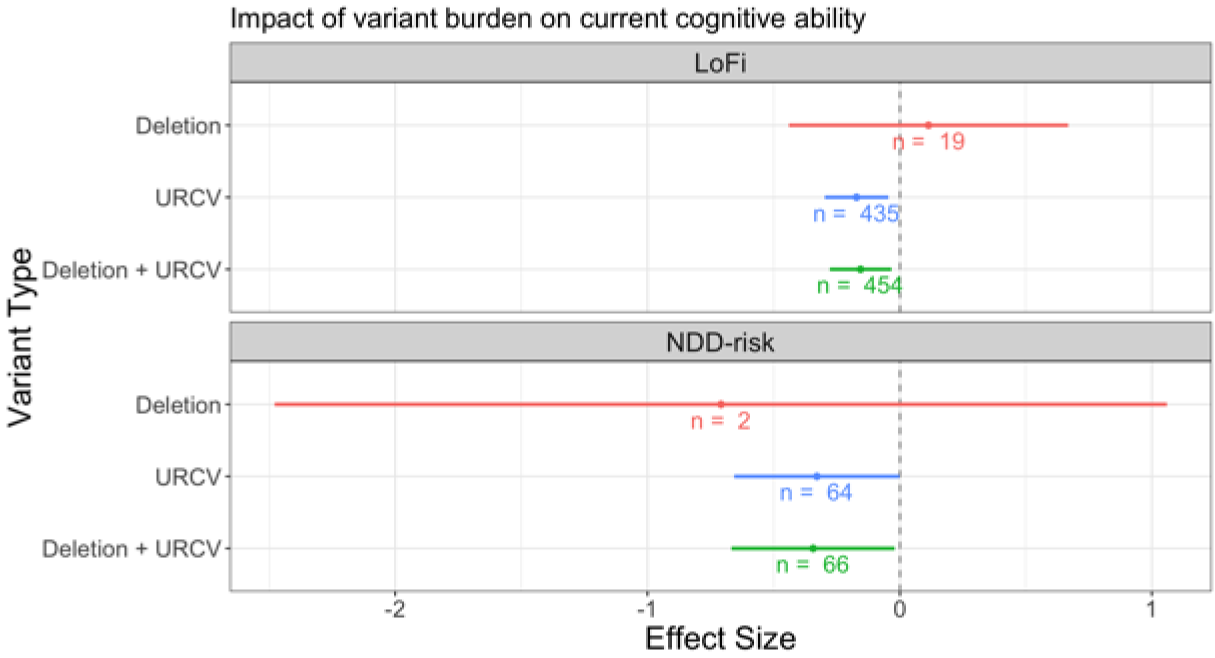


Figure 6.17. Impact of combined structural variant and ultra-rare constrained variant burdens affecting loss-of-function intolerant and neurodevelopmental disorder risk genes on current cognitive ability. The coloured points represent effect sizes and the horizontal lines are 95% confidence intervals. SV = structural variant, URCV = ultra-rare constrained variant, LoFi = variants impacting loss-of-function intolerant genes,

NDD-risk = variants impacting neurodevelopmental disorder risk genes.

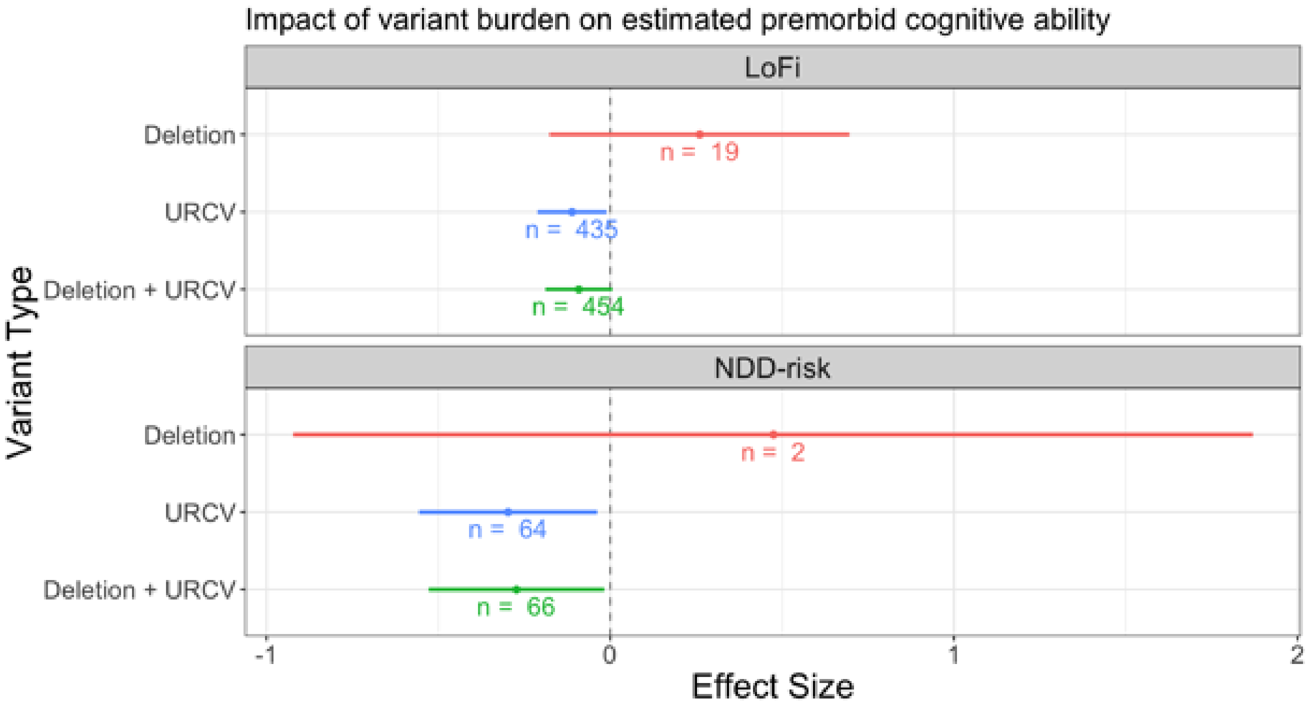


Figure 6.18. Impact of combined structural variant and ultra-rare constrained variant burdens affecting loss-of-function intolerant and neurodevelopmental disorder risk genes on premorbid cognitive ability. The coloured points represent effect sizes, and the horizontal lines are 95% confidence intervals. SV = structural variant, URCV = ultra-rare constrained variant, LoFi = variants impacting loss-of-function intolerant genes, NDD-risk = variants impacting neurodevelopmental disorder risk genes.

4. Discussion

4.1 Discussion of rare structural variants called by InDelible

The research reported in this chapter aimed to test whether the InDelible SV calling algorithm could identify SVs that are typically missed from array-based SV studies that contribute to variation in cognitive ability in schizophrenia. To do this, I identified 375 rare SV events in 290 (31.2%) COGs participants, with a median SV size of 32bp. To evaluate whether the size and frequency characteristics of the SVs identified in the current study are as expected, it would have been useful to compare the SV calls I made with those presented in the the original InDelible DDD study (Gardner et al., 2021), which to my knowledge is the only published study that has used InDelible. However, as described in section 3.4 chapter 2, the DDD study designs differs too much from the current study to enable a statistically meaningful comparison of results. Nevertheless, I can still assess whether my findings are broadly commensurate with the SV types and sizes described in the DDD study.

In both the DDD and current study, deletions are the most prevalent SV type. They constitute 32 (51%) of the novel SVs that were assigned a type in the InDelible paper, and 118 (43%) of SVs that were assigned a type in my results. Duplications are second most common SV type in both studies, making up 18 (28.6%) of novel SVs and 60 (21.8%) of the SVs I identified. The third most common type in the DDD study are complex-deletion/insertions, which make up 8 (12.7%) of the novel SVs. I did not detect any instances of this event type in the current study, but I did identify 9 (3.3%) complex-duplication/insertions. The third most common event type in the current study was both pseudogene retrotranspositions and simple insertions, which constituted 9.6% of the SVs with 36 instances of each. The former are not reported in the InDelible DDD study, as that study only considered SVs that affected known NDD-risk genes, and a limitation of pseudogene retrotranspositions is that it is not possible from exome sequencing data to determine where in the genome these events are inserted, thus we cannot know if they impact NDD-risk genes.

There was only 1 simple insertion in the DDD results. It is possible that that some of the 36 simple insertions I reported have been misclassified, given that they can only be classified through manual inspection, and in some cases would be indistinguishable from an SR cluster associated with a large deletion that happened to fail both BWA and BLAST alignment. Sixteen retrotranspositions of other types were identified in COGS, and one in the DDD results (an *Alu)*. Two translocations were identified in DDD results and one in my results. Thus, the only event type in the DDD results that was not identified in COGS is complex-deletion/insertion.

26.8% (n = 91) of the SVs identified were between 21-50bp in length, confirming that it is variants in this size range that InDelible is most sensitive to detect. Also in keeping with the InDelible DDD results, the size ranges with the fewest number of variants were 1-5 bp (0.2%, n = 1) and >10kb (0.8%, n = 3), demonstrating that InDelible is largely insensitive to point mutations/small indels and large structural variants. As the DDD authors suggest, then, the algorithm does indeed target a class of variants that are under-reported in the literature, and whose clinical impacts are largely unknown.

In summary, the types and size distribution of the rare SVs identified in the current study are broadly commensurate with those reported in the DDD study: Deletions and duplications are the most common SV type in both results, and all but one event type identified in the DDD study were also in my results. The vast majority of SVs I identified were also in the size range InDelible is most sensitive to, validating the findings of its developers.

4.2 Comparison of InDelible and CLAMMS call sets

3 of 35 (8.6%) of InDelible CNVs that were larger than the smallest CLAMMS CNV (160bp) were called by CLAMMS. This validation rate is consistent with the differential sensitives of the two call algorithms: InDelible is largely insensitive to variants >100bp, while CLAMMS is largely insensitive to variants <1kb. Therefore, CLAMMS cannot be used to validate large InDelible CNVs (and vice versa). In addition, 2/36 (5.6%) pseudogene retrotranspositions called by InDelible were called as duplications by CLAMMS. The CLAMMS call intervals were much larger than the size dictated by the number of exons that were found to be included in these events by InDelible. This is expected, as CLAMMS intervals include the intronic regions between the exons. However, the event that was found to implicate only 1 exon was called as larger than the event that implicated 6 exons, suggesting that InDelible may also have missed SR clusters associated with the former. These findings also suggest that CLAMMS is largely insensitive to pseudogene retroreptranspositions (as it missed 94% of those detected by InDelible), though this may more a function of their small size rather than their type.

4.3 Association between rare structural variants and cognitive ability in schizophrenia

For any type or frequency of SV tested in the current study, no significant association was found between SV burden and current cognition. Overall SV burden, SV burden excluding SVs of unknown type, and duplication burden were nominally associated with estimated premorbid cognition. However, effect sizes were all positive and no result remained significant after correcting for multiple testing, indicating that these results were likely false positive.

When I restricted the analysis to the overall burden of deletions, the point-estimate for the effect size of these SVs on both cognitive measures was trending in a more negative direction than that of duplications. This is in line with previous findings (including those reported in chapter 5) that show deletions have stronger effects on cognition in schizophrenia compared with other types of SV. However, if InDelible has a greater accuracy for calling deletions compared with other SV types, this might also contribute to the weak evidence for deletions having an effect on lower cognition.

When the analysis of deletions was further restricted to those occurring in LoFi or NDD-risk genes, weaker effects were found on both current and estimated premorbid cognition in all cases except small deletions affecting NDD-risk genes on estimated premorbid cognition. This pattern of effects contrasts with findings made from recent studies of rare coding variants and cognition in schizophrenia, where rare coding variants in LoFi and NDD genes had stronger effects on cognition. However, the effects from deletions on observed in the current study are based on a very small number of variants, have very large confidence intervals, and are not significant. Therefore, larger, and better powered studies of SVs discovered using InDelible are required to determine whether this type of SV truly contributes to cognition in schizophrenia.

Power could be improved by increasing the number of variants tested through implementing a less conservative QC thresholds, such as lowering the prob_y filter from > 0.8 to > 0.6; however, a critical issue of this would be the time-based restrictions on effective manual inspection that comes with a significantly larger call set, as I have shown that manual inspection is crucial for excluding artefacts in this data, particularly in the Broad subcohort. Without manual inspection of SVs in the larger call set, relaxing the QC would likely increase the number of false positive calls, and thus reduce the power of the current study.

The InDelible SV calling algorithm is relatively new and has not been widely tested. Therefore, it is important to note that unlike PTVs or large CNVs, no methods have been developed to prediction the impact of small SVs discovered by InDelible on protein function. While this is a limitation of the current study, I hypothesised that InDelible SVs affecting LoFi genes were more likely to have pathogenic effects than SVs in non-LoFi genes, given the previous evidence for association between other classes of rare mutation and lower cognition in schizophrenia. Nevertheless, the current study could be improved by first predicting the consequence of each variant on protein conformation and stability, and selecting only those that are likely to cause disruption (Sudmant et al., 2015). Bioinformtic tools such as I-Mutant (Capriotti et al., 2005) can use thermodynamic data to predict the impact of point mutations on protein stability.

4.4 Combined analysis of deletions and ultra-rare constrained variants

Adding deletions impacting LoFi genes to the URCVs did not increase the negative impact of URCVs tested separately for either cognitive measure. Adding deletions impacting NDD-risk genes to URCVs did marginally increase the impact of URCVs test separately, but only in the cases of current cognition. Moreover, there were only 2 deletions in this set, and they were not separately associated with cognitive deficits. Despite affecting fewer bases, the likely reason why the URCVs have a greater impact on cognitive deficits than the SVs is due to how they are defined. While it is more likely that a random SV of any size will be more deleterious than a random point mutation of the same AF, just as a function of the number of bases impacted, the selection criteria for URCV is based on whether they alter protein coding to such an extent they are selected against. As no analogous criteria could be applied to the SVs, given that their impact on protein coding and their relative prevalence in the general population is unknown, many if not most will have a more minimal impact on protein coding of LoFi and NDD genes than any of the URVCs. Deriving and applying such criteria to the SVs would make for a more valid comparison of the two variant types.

4.4 Differences between Cardiff and Broad subcohort

Throughout the QC process, multiple differences between the COGs subcohorts were observed. First, despite consisting of 1.2x more samples, 15.5x more calls were in the initial InDelible output for the Broad subcohort. Figure 6.1 shows there is a strong correlation between sample coverage and number of calls, such that coverage difference between subcohorts is the main factor behind this discrepancy. However, the shapes of the distributions for number of calls was also different: The Broad subcohort had a right-skewed distribution (Figure 6.2), while the Cardiff had a normal distribution (Figure 6.3). As the number of calls in the SCZ trios also had a normal distribution, it seems that the distribution for the Broad subcohort is atypical, possibly indicating that a greater proportion of samples have an excess of artefacts/false positive calls. Moreover, 14.7% more calls were filtered from the Broad subcohort by the prob_y filter, which is intended to exclude calls with a low probability of indicating real SVs and therefore also suggests that these samples contain a higher proportion of false positive calls.

Manual inspection of rare SVs confirmed this hypothesis, as 96% of calls that were excluded as artefacts were called in Broad-sequenced samples. I also observed a kind of artefact that was unique to this subcohort, indicating systematic errors during the sequencing process that did not occur for the Cardiff subcohort. This goes to show that batch effects can produce SR patterns that cannot be detected in all WES data samples, and so may not be detected as artefacts if the SV caller is tested/trained on one dataset only. It also highlights the critical importance of manual inspection of split reads, to detect idiosyncrasies in data to which the algorithm and automated QC steps are not sensitive.

4.5 Summary

In this chapter I have presented research that involved application of the InDelible algorithm to detect small, rare structural variants in WES trios sample consisting of 927 cases. I identified 375 rare SV events in 290 (31.2%) participants, ranging from 5bp to 57kb in size. I then assessed the impact of SV burden on cognition, finding no association of any variant burden metric with either cognitive measure. Combining variants with URCV also did not improve the effect size observed for the latter tested separately. It is likely that my study had limited power, however, and further work needs to be carried out to understand the deleteriousness of small SVs, to design a study that will be able to effectively reveal their impact on cognition.

Chapter 7: General Discussion

1 Summary of thesis aims

In section 3 of chapter 1, I listed 3 overarching thesis aims, which are restated below:

1) Assess the overlap and differences between rare SV call sets generated by the two calling algorithms in WES data, and call sets previously generated from the same data using genotyping microarrays.

2) Identify *de novo* SVs in the schizophrenia trios data and use findings from previous rare and common variant studies to determine any putative candidate schizophrenia risk genes that are impacted by SVs.

3) Test rare SVs for association with cognitive deficits, with particular focus on the role of SVs at the smaller end of the size spectrum.

In the following sections I discuss each of these aims in turn, assessing the extent to which my research has met them. I also discuss the limitations of CLAMMS and InDelible and how future studies of SVs in schizophrenia could build on my findings

2 Discussion of thesis aims

2.1 Discussion of Aim 1

Using CLAMMs and InDelible, I generated four sets of calls, two per WES data set analysed. The first two were *de novo* call sets, called across 616 schizophrenia probands. CLAMMS identified 9 putative *de novo* CNVs in 9 (1.5%) individuals: 7 deletions and 2 duplications (Table 3.3). InDelible, meanwhile, identified 15 putative de novo SVs in 15 (2.4%) individuals: 2 pseudogene retrotransposons, 7 deletions, 1 duplication, 1 insertion, 1 complex-insertion/deletion, 1 complex-insertion/duplication, and 2 SVs whose type could not be determined (Table 4.3). There was no overlap between the call sets due to the differential sensitivities of the callers. CLAMMS is only sensitive to CNVs and its lower discovery resolution is approximately the size of a human exon (~150bp). InDelible, on the other hand, calls all SV types except inversions, is most sensitive to SVs 21-50bp in size, and is largely insensitive to variants >100bp. The smallest *de novo* event identified by CLAMMS was 27kb, and the largest identified by InDelible was 12.5kb.

My findings suggest that the rate of *de novo* SVs that can be called by InDelible is greater than the rate that can be called by CLAMMS, at least in the context of schizophrenia. As about the same number of *de novo* CNVs were identified by each algorithm (8 by InDelible, 9 by CLAMMS), this is a function of the broader number of SV types called by InDelible. However, given that InDelible had a significantly higher false negative rate, according to call validation tests carried out by its developers, compared to CLAMMS (see sections 2.3.1 and 3.4.1 of chapter 2), it is plausible that the actual difference between *de novo* SV rates in the small (<150bp) and large size ranges is greater than is indicated by my results. To my knowledge, however, the expected difference has never been robustly investigated, likely because small SV callers like InDelible are a very recent development.

Seven of nine of the *de novo* SVs identified by CLAMMS were also identified by PennCNV in the array data for the same individuals, though one of the carriers was not genotyped. Moreover, 6/10 de novo CNVs identified by PennCNV were validated by CLAMMS, though one was found to be a transmission. An additional 3 were called by CLAMMS but were filtered during sample-level quality control. A 2kb duplication was not called by CLAMMS, but due its small size is likely to be a false positive in the array data. Thus, combining CLAMMS and PennCNV can increase power to detect *de novo* CNVs if both WES and array data are available, or can used as a method for call validation. Given the high degree of overlap between these call sets, however, it can also be argued that the use of only one method is sufficient.

The second rare SV call sets I produced were called across 927 schizophrenia cases in the Cardiff COGS cohort. CLAMMS identified 977 rare CNVs in 556 (61%) participants: 352 deletions and 627 duplications. InDelible identified 478 rare SVs in 375 individuals: 119 deletions, 58 duplications, 36 simple insertions, 9 complex-duplication/insertions, 2 small tandem repeats, 1 translocation, 36 pseudogene retrotranspositions, 16 retrotranspositions of other types (SINE, Alu etc.), and 100 SVs whose type could not be determined (Figure 6.7). The smallest CNV called by

CLAMMS was 160bp, ~100bp larger than InDelible's target size range. However, 35 (8.6%) of InDelible CNVs were larger than 160bp, 3 of which were also called by CLAMMS: two deletions and a duplication (Table 6.6). In addition, 2 CNVs called as duplications by CLAMMS were identified by InDelible as pseudogene retrotranspositions (Table 6.7). A discussed in chapter 6, this low rate of validation between the call sets consistent with their differential sensitives and demonstrates that the two callers can be used for mutual validation of calls. However, it also shows there can be convergence of coverage depth and split reads as evidence for the same events.

Eight hundred and sixty-six of the rare CNVs identified CLAMMS were carried by individuals for whom array data was available. Three hundred and twenty-one of 866 (37%) of the CNV were identified by PennCNV, a modest degree of overlap. Much of this was driven by large (>100kb) events. One hundred and thirty-nine of 193 (72%) of large events called by CLAMMS were also in the array data, compared with 182/673 (27%) of small events. Orthogonal evidence is required to determine if the low degree of overlap between small events is primarily due to a high positive rate among CLAMMS calls, or the low sensitivity of PennCNV to these events. As CLAMMS has been demonstrated to have a high (>90%) validation rate for variants as small as 1 exon in length (section 2.3.1 of chapter 2), there is reason to suppose the latter. However, of small 327 events called in the array data, 140 (44%) were validated by CLAMMS, indicating that the discrepancy in my results is not simply due to the lower number of small events detected in the array data. Again, there was no overlap between the InDelible and PennCNV call sets.

In summary, assessing the overlaps between call sets generated by CLAMMS and InDelible showed that these methods are highly complementary, in that they mine different aspects of the data and are sensitive to different SV types and size ranges. One cannot be used to validate calls produced by the other, though there can be some overlap between their call sets. By applying both callers to the same data, however, SVs from across the broadest possible size range can be identified, ranging from 10bp to several megabases. Assessing overlaps between CLAMMS and PennCNV call sets showed significant overlap in the case of large (>100kb) CNVs, such that only one approach is sufficient for studies that are primarily interested in such events. There is evidence that CLAMMS is much more sensitive to smaller events, though orthogonal support is required to confirm this.

2.2 Discussion of Aim 2

As mentioned above, 9 *de novo* CNVs were detected in the WES trios data by CLAMMS, and 15 by InDelible. I found that 2 of the *de novo CNVs* called by CLAMMS were instances of the 11 known SCZ-risk CNVs: 3q29 deletion and 22q11.2 deletion. This demonstrates that pathogenic CNVs can be successfully identified by CLAMMS. Another CNV was a deletion of 16p13.11, a locus for which only duplications have been conclusively shown to increase SCZ risk, though an excess of the deletion has been previously reported in cases (Ingason et al., 2011). I also found evidence that a further 2 deletions may confer SCZ risk: Chr18:163305-5478439 deletion was found to disrupt *DLGAP1*, a gene with nominal association ($p < 0.05$) in SCHEMA, and which had been implicated previously in a SCZ *de novo* CNV study (Kirov et al. 2012); and Chr10:18242203-19896831 deletion, which overlaps a locus within the CACNB2 gene that was significantly associated with SCZ in the PGC3 GWAS (Trubetskoy et al. 2022), and was still implicated after fine-mapping. *DLGAP1* and *CACNB2* are expressed at the post-synapse and play roles in synaptic organisation and plasticity (Rasmussen et al., 2017; Dolphin, 2012). It is therefore plausible that both CNVs confer SCZ risk in their respective carriers.

Variants affecting genes with a high probability of loss-of-function (pLI) have consistently been found to be enriched in SCZ cases (Singh et al., 2022). Two of the small SVs identified by InDelible occurred within high pLI genes: a 19bp deletion in *ATN1* (pLI = 0.97) and a 32bp complex-insertion/duplication in *CTNNA1* (pLI = 1). While neither of these genes have been previously implicated in SCZ case/control association studies, they are plausible candidate risk genes. *ATN1* has been implicated in neurodevelopment, specifically the differentiation of neural progenitor cells, psychotic symptoms have been reported in dentatorubral-pallidoluysian atrophy cases, a neurological disorder caused by a CAG trinucleotide repeat expansion in this gene. *CTNNA1* has a clear role in synapse development and maintenance (Arikkath & Reichardt, 2008), and knock-out mice show deficits in fear-potentiated startle response *(Park et al., 2002),* a phenotype that is also evident in schizophrenia cases.

Thus, I have shown that both CLAMMS and InDelible can detect *de novo* variants in schizophrenia cases that affect established and plausible schizophrenia-risk genes. Deriving the precise functional consequences of the CLAMMS variants is complicated, as they impact more genes than just those I have discussed, and likely also affect promotor regions for genes that are outside the deleted locus. It is much simpler in the case of the InDelible variants, however, as their functional impact is likely to be limited to their predictable effects on protein conformation and stability. While it was beyond the scope of my thesis to explore these effects, I think it will prove to be an important aspect of detecting small SVs that are likely to be pathogenic in future work.

2.3 Discussion of Aim 3

I tested both the CLAMMS and InDelible rare SV call sets generated from Cardiff COGS participants for association with current and estimated premorbid cognitive ability, focusing primarily on small (<100kb) variants that are outside the discovery resolution for reliable detection by PennCNV. I found that small CLAMMS CNVs affecting loss-of-function intolerant (LoFi) genes were not associated with current or premorbid cognition, though the effect sizes for deletions were trending in the expected negative direction (Tables 5.14 & 5.16). Small deletion affecting neurodevelopmental disorder-risk genes were associated with estimated premorbid cognition, though there were only three variants in this test set. The effect size for small deletions affecting LoFi genes was trending in the expected negative direction, and there is no sign that small duplications negatively impact cognition. Testing small deletions without gene set restrictions produced a nominal association with estimated premorbid cognitive ability, however, suggesting that the absence of signal for the smaller, theoretically more deleterious variant sets is due to low power. Taken together, these results do suggest that small deletions impact cognition in schizophrenia and provide sufficient grounds for repeating the analysis in larger samples.

My exploratory analysis of larger variants showed that large deletions, but not duplications, are associated with lower current and estimated premorbid cognitive ability. Consistently with previous studies, negative effect sizes increased when variants were restricted to those affecting LoFi and NDD-risk genes, providing further evidence that disruption of these gene sets produces cognitive deficits that are partly independent of schizophrenia progression. There is no evidence in my results that large deletions impact one cognitive measure more greatly than the other. Restricting variants to those that were also identified in the array data did not, generally, improve power to detect variants that impact cognition. Given that large CNVs are known to negatively impact cognition, this may indicate that the large CLAMMS CNV call set contains fewer false positives than the PennCNV call set.

In a follow-up study, I tested a subtype of large, rare deletion in the 22q11.2 locus – central deletion - that was associated with current cognitive deficits in Cardiff COGS for cognitive and neuropsychiatric effects in the UK Biobank (UKB). This event occurs at the distal end of the DiGeorge Syndrome locus and has been associated with a less severe DGS-like phenotype in clinical studies (Burnside, 2011). I found that it was associated with speed-of-processing deficits in the UKB and observed non-significant deficits for the other 5 domains tested. In addition, central deletions negatively impacted functional outcomes that are highly correlated with cognitive ability and were associated with anxiety disorder case status. In line with previous studies, these findings show that CNVs associated with lowered cognition in schizophrenia also impact cognition in the general population. It is unknown whether 22q11.2 central deletions confer risk for schizophrenia, however, and there were not enough SCZ cases in the UKB to test this. Future work on this variant will involve testing its incidence in a large schizophrenia case-control data set.

Testing rare SVs produced by InDelible for association with cognition gave highly inconclusive results. Without gene set restrictions, no SV burden was associated with cognitive deficits, though deletions had a marginally more negative effect size than duplications. Overall SV burden and duplication burden were nominally associated with estimated cognitive enhancements, though these results were not significant after multiple testing corrections were applied. Restricting deletions by gene set only produced a more negative effect size for deletions in NDD-risk genes on current cognitive ability, though only 2 variants were included in this test and the effect was non-significant. Combining deletion burdens with URCVs called by (Creeth et al. 2021) had minimal or no impact of the negative effects of URCVs tested separately. Overall, these results are not surprising when compared to those of the small CLAMMS deletions. If there is limited power in this data to detect the impact of variants <100kb but >100bp, power should be even more limited to detect the impact of variants < 100bp in size. As discussed in chapter 6, however, the reason why significant associations were detected for URCV is because they were restricted to those variants that are known to have a deleterious impact on protein structure – PTVs and damaging missense variants. To increase power to detect any impacts of InDelible SVs in future studies, similar criteria should be implemented before testing.

To summarise, my findings corroborate previous studies that showed associations between large, rare deletions, particularly those affecting LoFi and NDD-risk genes, and cognitive deficits in schizophrenia. They also contain suggestive evidence for association of rare deletions <100kb with cognitive deficits, a variant class that has been hitherto untested in the context of schizophrenia. Future studies of these variants should be conducted in larger samples, and for rare SVs <100bp more work should be conducted to determine their precise functional impacts in order to the isolate variants that are most likely to be clinically significant.

3 Limitations of SV callers

3.1 Limitations of CLAMMS

The main limitations of the CLAMMS algorithm are associated with choosing a reference panel size (k) that optimally controls for batch effects. While the minimal computational resources required to generate the reference panels themselves is the main reason why CLAMMS is more computationally efficient than alternative algorithms, the optimal k cannot be known *a priori,* requiring implementation of the whole calling and qc pipeline for number of values. This could be prohibitively time consuming for larger samples, in addition to producing a large amount of output data. Appropriate metrics by which call error rates can be tested must also be available (transmission rate for SCZ trios; array overlap and sample drop out for Cardiff COGS). However, this is not strictly necessary for producing calls, as the CLAMMS authors recommend a default k of 100 based on their own analyses. Also, it is likely that only a few k's need to be tested to gauge the most appropriate size, as

I showed in the CLAMMS analyses of the trios. I decided to test 40 values of k in the COGS analyses as I wanted to assess the differences in batch effects between the Cardiff and Broad subcohorts, in addition to selecting the optimal k.

A limitation of the CLAMMS CNV output is the imprecise estimation of CNV size, a consequence of basing calls on exon coverage only. CNVs whose breakpoints occur with intronic regions will always be called smaller than they are by CLAMMS. This is not a large limitation if researchers are primarily interested in the genes that are affected by CNVs, but it does mean that functional analyses of CNV impacts cannot be precise. For example, there may be a promotor region in an intronic region that is affected by a CNV but is not included in the CLAMMS call interval for that CNV as it occurs prior to the first exon which that CNV affects. Array-based approaches face a similar limitation if probe density around CNV breakpoint is low, but typically allow for more precise estimation of breakpoint positions. Application of methods similar to CLAMMS in whole genome sequencing (WGS) data would not have this limitation however, as coverage for all intronic regions can be modelled, allowing for much more precise breakpoint estimation than WES and array-based methods.

Another limitation that is very specific, and observed during my analysis, is the tendency of CLAMMS to misclassify pseudogene retrotranspositions as duplications. This does not appear to be a significant issue, as only 2/35 such events were identified by CLAMMS in the Cardiff COGs data, and the others were either not called or filtered by quality score criteria. However, it does highlight the fact that non-CNV events can produce deviations in coverage that can be captured by a coverage-based WES caller, and there is no way to determine such errors based on the call data alone. Researchers ought to be cognisant of this, and check for misclassifications using alternate methods (such as InDelible) if possible. However, this would not be a limitation for WGS analysis, which would likely call such events as duplications at successive exons, given that the corresponding lack of coverage deviation between exons would be recognised.

3.2 Limitations of InDelible

Although InDelible is most sensitive to variants 21-50bp in size, it still only detected ~50% of known variants in this size range that had been previously detected in the DDD study (Gardner et al 2021). Thus, InDelible cannot be used, at least on its own, to assess the general population rates of its targeted variants even in sufficiently large samples, or to ascertain variant prevalence in case/control samples. However, this does not undermine the utility of InDelible to detect potentially pathogenic variants in clinical samples, as the authors demonstrated in the DDD study, and is indicated by own analyses.

InDelible assigned an SV type to 1.2% and 3% of calls in the initial output of the SCZ trios and COGS Cardiff subcohort analyses, respectively. In the DDD study, the authors reported that InDelible assigned a type to 10.2% of the initial output. However, the median coverage depth across the DDD cohort is 90X, compared with 30x for the SCZ trios and the COGS Cardiff subcohort. Thus, the greater proportion of variants whose type could be determined in the DDD study is most likely due to differences in coverage depth. At higher coverage depth, not only will a given SR cluster will have more SRs, but the average length of misaligned bases in the SRs will be longer, thus increasing the likelihood of a valid alignment when the longest misaligned sequence is run through BWA-MEM or BLAST, and in turn the likelihood of InDelible assigning an SV type to a call.

The Broad subcohort does have approximately the same coverage depth as the DDD cohort, but an even lower percentage of SV type assignments than both the Cardiff subcohort and the SCZ trios (0.7%). My analyses strongly suggest, however, that there were large, systematic errors introduced during the sequencing process for the Broad subcohort, producing many split reads artefacts that were detected by InDelible, and are likely to have confounded the correlation between coverage and SV type assignment rate.

Moreover, to account for sequence homology and SNPs the recommended minimum sequence length is 22 for both BLAST and 19 for BWA-MEM, which are both implemented in InDelible. 76.6% of calls in the initial SCZ trios' output have a longest misaligned sequence that is < 19 bases, and thus could not be assigned an SV type by default. Of those calls whose longest misalignment is > 22 bases, there was no hit for 61%. This does not in itself present an issue, as most unfiltered calls will be based on artefacts in the data and should therefore be expected to not have a valid SV type. However, there will be a significant number of calls that are based on real events that have a longest misalignment < 19 due to low coverage. Altering implementation of the alignment tools in InDelible such that the minimum sequence length can be lowered to better fit the data is unlikely to significantly improve the rate of SV type assignment in the SCZ trios, as homology among small sequences would result in multiple off-target hits. While SV type can also be independently ascertained from manual inspection, in the analysis of larger samples or more common variant sets this could be prohibitively time consuming, limiting the scalability of InDelible for data with lower coverage.

I found that InDelible misclassifies calls indicative of pseudogene retrotranspositions as deletions affecting inter-exonic regions. This is not unexpected, given that the SR pattern of clusters at the 5’ and 3’ ends of successive retrotransposed exons corresponds to that of a deletion, and in its current form InDelible does not include code to correctly identify these events if they do not appear in the repeated sequence database mined by BLAST. However, an additional criterion could be added to the SV type identifier script, which assigns this SV type when successive clusters occur at the junctions of exons. It is extremely unlikely that any other SV types will have breakpoints at precisely these positions so the possibility of misclassification would be minimal.

Calculation of SV size by InDelible is also dependent on there being a unique BWA-MEM alignment for soft-clipped bases, as the algorithm uses this alignment to estimate the other breakpoint position for a given call. Thus, size can only be calculated for unbalanced events, as insertions will have no unique alignment and/or repeat BLAST hits (in the case of MEIs), while translocations will align in most cases to another chromosome. This is not a limitation of the algorithm in the case of large insertions and MEIs, as it is not possible to determine the other breakpoint if no unique alignment position has been found. However, in the case of small insertions that can be nested within reads (<10bp), code could be implemented to mine CIGAR alignment strings in reads adjacent to SR clusters to discern the length of any nested insertions that occur at the same positions. InDelible already checks this information in SR Clusters to identify complex events, for example where an insertion is nested within a copy of duplicated sequences. I described in section 3.5.5 of chapter 2 how in some cases the precise size of translocations can be calculated from their associated SR Cluster positions on different chromosomes. It is possible to implement this calculation in a script, though unlike in the case of deletions and duplications, both SR Clusters would need to be paired at a prior analysis stage as there is no way to determine both breakpoints from one cluster alone.

4. Implications for future studies

My findings have several implications for how future studies of SVs in schizophrenia should be conducted. First, if the purpose of a study is to detect rare, large CNVs, large de novo CNVs, or CNVs impacting schizophrenia-risk loci, impacting protein-coding regions, it is not necessary to implement both WES and array-based methods, as I found CLAMMS and PennCNV have approximately the same sensitivity for detecting such events. Call sets do not overlap completely though, and if both approaches can be used it is likely to yield a small number of events that would not be detected if only one of them was applied. Given that only large deletions called by CLAMMS were negatively associated with current cognition in COGS, however, there is some evidence that the CLAMMS call set may contain fewer false positive calls. And based on the raw number of small (<100kb) deletions called by each algorithm, as well the nominal association of small deletions called by CLAMMS with estimated premorbid cognition in COGS, there is evidence that CLAMMS is more sensitive to small events than PennCNV and should thus be the preference in studies investigating these events. My analysis of the impact of small events on cognition was limited by power, however, so future studies should also include larger samples.

I have shown that mining split read information in WES data can yield calls for events of types and sizes that are entirely undetectable in array data. Thus, WES data should be generated over array data if researchers are to investigate the broadest possible range of SVs. While my *de novo* analysis of InDelible calls produced events that impact plausible schizophrenia risk genes, my analysis of the impact of these events on cognition was highly inconclusive. Limited power was likely the main issue, but also not including methods and criteria than can filter variants based on their likely functional impacts (in addition to their occurrence within LoFi/NDD-risk genes). Future studies, even in larger samples, may still be limited by power issues if they do not include methods to this effect. Moreover, the InDelible algorithm could be improved by including code that is able to assign a broader range of SV types more accurately to calls, such as pseudogene retrotranspositions and simple insertions. However, as split read patterns associated with some events are complex, and batch effects can produce artefacts that are specific to data sets, manual inspection will always be needed to assign types to some events.

Wide-spread implementation of short-read WGS will overcome many of the issues that beset WES SV callers. For example, the primary reason why InDelible is insensitive to large events is because there are significantly fewer of them than small events such that their breakpoints are unlikely to occur in exons. For the same reason, a caller like CLAMMS cannot use split read information to support its coverage-based calls. However, if all non-coding regions are also available for analysis, there would be no such limitations. SV callers could then implement both split read information and coverage depth, among other metrics, into their algorithms to call large events, producing a more valid call set than algorithms that only use one of these aspects of the data. In addition, breakpoints could be determined with high precision, so that functional analysis of calls can be much more precise than is currently possible with both WES and array-based SV calls.

There are already several short-read WGS SV callers that implement such approaches. For example, SoftSV (Bartenhagen & Dugas, 2016) and Wham (Kronenberg et al., 2015) use both discordant read pairs and split reads to discern SV breakpoints, while SoloDel (Kim et al., 2015) uses discordant reads and coverage. Manta (Chen et al., 2016) and GRIDDSS (Cameron et al., 2017) leverage three aspects of WGS data: discordant read pairs, split reads, and assembly of sample genome relative to the reference. The latter is a relatively novel technique which uses graph-based assemblies of genomic loci, in which nodes correspond to loci while edges represent loci overlaps. Edges in a sample genome assembly that deviate from the reference can be evidence of structural variation. A benchmarking study found that methods that use multiple aspects of the data, and especially those that implement some kind of genome assembly approach, tend to outperform others in both precision and recall (D. L. Camerson et al., 2019). In addition, methods that use split reads were found to be highly effective at resolving both breakpoints of SVs at single nucleotide-resolutions. These methods will be instrumental in driving forward the understanding of structural variation generally, as well as in clinical contexts like schizophrenia.

Bibliography

Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., Layer, R. M., Neale, B. M., Salerno, W. J., Reeves, C., Buyske, S., Matise, T. C., Muzny, D. M., Zody, M. C., Lander, E. S., Dutcher, S. K., Stitziel, N. O., & Hall, I. M. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, *583*(7814), 83–89. https://doi.org/10.1038/s41586-020-2371-0

Abi-Dargham, A., Mawlawi, O., Lombardo, I., Gil, R., Martinez, D., Huang, Y., & Laruelle, M. (2002). Prefrontal dopamine D1 receptors and working memory in schizophrenia. *Journal of Neuroscience*, *22*(9), 3708–3719. https://doi.org/10.1523/JNEUROSCI.22-09-03708.2002

Abyzov, A., Li, S., & Gerstein, M. B. (2016). Understanding genome structural variations. *Oncotarget*, *7*(7), 7370–7371. https://doi.org/10.18632/oncotarget.6485

Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, *17*(1), 241. https://doi.org/10.1186/s13059-016-1110-1

Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *African Journal of Laboratory Medicine*, *9*(1). https://doi.org/10.4102/ajlm.v9i1.1340

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212.

Alekseyev, Y. O., Fazeli, R., Yang, S., Basran, R., Maher, T., Miller, N. S., & Remick, D. (2018). A Next-Generation Sequencing Primer—How Does It Work and What Can It Do? *Academic Pathology*, *5*, 2374289518766521. https://doi.org/10.1177/2374289518766521

Aleman, A., Kahn, R. S., & Selten, J.-P. (2003). Sex differences in the risk of schizophrenia: Evidence from meta-analysis. *Archives of General Psychiatry*, *60*(6), 565–571.

Aleman, A., & Larøi, F. (2008). *Hallucinations: The science of idiosyncratic perception*. American Psychological Association.

Alkuraya, F. S. (2011). Human mutations in NDE1 cause extreme microcephaly with lissencephaly. *American Journal of Human Genetics*, *88*(5), 536–547.

Alvarez-Jimenez, M., Priede, A., Hetrick, S. E., Bendall, S., Killackey, E., Parker, A. G., & Gleeson, J. F. (2012). Risk factors for relapse following treatment for first episode psychosis: A systematic review and meta-analysis of longitudinal studies. *Schizophrenia Research*, *139*(1–3), 116–128.

Anderson, K. K., Cheng, J., Susser, E., McKenzie, K. J., & Kurdyak, P. (2015). Incidence of psychotic disorders among first-generation immigrants and refugees in Ontario. *Canadian Medical Association Journal*, *187*(9), E279–E286. https://doi.org/10.1503/cmaj.141420

Andreasen, N. C. (1979). Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. *Archives of General Psychiatry*, *36*(12), 1315–1321.

Angrist, B., Rotrosen, J., & Gershon, S. (1980). Differential effects of amphetamine and neuroleptics on negative vs. positive symptoms in schizophrenia. *Psychopharmacology*, *72*(1), 17–19. https://doi.org/10.1007/BF00433822

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., Lee, P. H., Turley, P., Grenier-Boley, B., Chouraki, V., Kamatani, Y., … Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395). https://doi.org/10.1126/science.aap8757

Arikkath, J., Peng, I. F., Ng, Y. G., Israely, I., Liu, X., Ullian, E. M., & Reichardt, L. F. (2009). Delta-catenin regulates spine and synapse morphogenesis and function in hippocampal neurons during development. *Journal of Neuroscience*, *29*(17), 5435–5442.

Arikkath, J., & Reichardt, L. F. (2008). Cadherins and catenins at synapses: roles in synaptogenesis and synaptic plasticity. *Trends in Neurosciences*, *31*(9), 487–494.

Asim, A., Kumar, A., Muthuswamy, S., Jain, S., & Agarwal, S. (2015). "Down syndrome: an insight of the disease." *Journal of Biomedical Science*, *22*(1), 41. https://doi.org/10.1186/s12929-015-0138-y

Association, A. P. (2013). DSM 5. In *American Journal of Psychiatry*. https://doi.org/10.1176/appi.books.9780890425596.744053

Auwerx, C., Lepamets, M., Sadler, M. C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Porcu, E., Reymond, A., Kutalik, Z., Esko, T., Metspalu, A., Milani, L., Mägi, R., & Nelis, M. (2022). The individual and global impact of copy-number variants on complex human traits. *The American Journal of Human Genetics*, *109*(4), 647–668. https://doi.org/10.1016/j.ajhg.2022.02.010

Awad, A. G., & Voruganti, L. N. (2008). The burden of schizophrenia on caregivers. *PharmacoEconomics*, *26*(2), 149–162.

Babulas, V., Factor-Litvak, P., Goetz, R., Schaefer, C. A., & Brown, A. S. (2006). Prenatal Exposure to Maternal Genital and Reproductive Infections and Adult Schizophrenia. *American Journal of Psychiatry*, *163*(5), 927–929. https://doi.org/10.1176/ajp.2006.163.5.927

Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., Lifton, R., Goldmuntz, E., Chung, W. K., & Shen, Y. (2014). CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Research*, *42*(12), e97–e97. https://doi.org/10.1093/nar/gku345

Barlow, J. H., Faryabi, R. B., Callén, E., Wong, N., Malhowski, A., Chen, H. T., Gutierrez-Cruz, G., Sun, H.-W., McKinnon, P., Wright, G., Casellas, R., Robbiani, D. F., Staudt, L., Fernandez-Capetillo, O., & Nussenzweig, A. (2013). Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. *Cell*, *152*(3), 620–632. https://doi.org/10.1016/j.cell.2013.01.006

Bartenhagen, C., & Dugas, M. (2016). Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. *Briefings in Bioinformatics*, *17*(1), 51–62. https://doi.org/10.1093/bib/bbv028

Beesdo, K., Pine, D. S., Lieb, R., & Wittchen, H. U. (2010). Incidence and risk patterns of anxiety and depressive disorders and categorization of generalized anxiety disorder. *Archives of General Psychiatry*, *67*(1), 47–57.

Bendall, S., Jackson, H. J., Hulbert, C. A., & McGorry, P. D. (2008). Childhood trauma and psychotic disorders: a systematic, critical review of the evidence. *Schizophrenia Bulletin*, *34*(3), 568–579.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), e72–e72. https://doi.org/10.1093/nar/gks001

Blanchard, J. J., Bradshaw, K. R., & Garcia, C. P. (2011). Core psychopathology in schizophrenia: A review of the evidence for the “Negative Symptom Syndrome.” In *The Clinical Neuropsychiatry of Schizophrenia* (2nd ed., pp. 55–82). Taylor & Francis.

Bleuler, E. (1911). *Dementia praecox or the group of schizophrenias*. International Universities Press.

Bloomfield, M. A., Morgan, C. J., Egerton, A., Kapur, S., Curran, H. V, & Howes, O. D. (2014). Dopaminergic function in cannabis users and its relationship to cannabis-induced psychotic symptoms. *Biological Psychiatry*, *75*(6), 470–478.

Bokoch, G. M. (2003). Biology of the p21-Activated Kinases. *Annual Review of Biochemistry*, *72*(1), 743–781. https://doi.org/10.1146/annurev.biochem.72.121801.161742

Bora, E., & Pantelis, C. (2015). Meta-analysis of cognitive impairment in first-episode bipolar disorder: comparison with first-episode schizophrenia and healthy controls. *Schizophrenia Bulletin*, *41*(5), 1095–1104.

Bossong, M. G., Berckel, B. N., Boellaard, R., Zuurman, L., Schuit, R. C., Windhorst, A. D., & Ramsey, N. F. (2009). Δ9-Tetrahydrocannabinol induces dopamine release in the human striatum. *Neuropsychopharmacology*, *34*(3), 759–766.

Bourque, F., van der Ven, E., & Malla, A. (2011). A meta-analysis of the risk for psychotic disorders among first- and second-generation immigrants. *Psychological Medicine*, *41*(5), 897–910. https://doi.org/10.1017/S0033291710001406

Bowie, C. R., & Harvey, P. D. (2006). Cognitive deficits and functional outcome in schizophrenia. In *Neuropsychiatric Disease and Treatment*. https://doi.org/10.2147/nedt.2006.2.4.531

Brady, P. D., & Vermeesch, J. R. (2012). Genomic microarrays: a technology overview. *Prenatal Diagnosis*, *32*(4), 336–343. https://doi.org/10.1002/pd.2933

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, *100*(9), 5280–5285. https://doi.org/10.1073/pnas.0831042100

Brown, A. S., Begg, M. D., Gravenstein, S., Schaefer, C. A., Wyatt, R. J., Bresnahan, M., & Susser, E. S. (2004). Serologic evidence of prenatal influenza in the etiology of schizophrenia. *Archives of General Psychiatry*, *61*(8), 774–780.

Brunetti-Pierri, N., Berg, J. S., Scaglia, F., Belmont, J., Bacino, C. A., Sahoo, T., & Shinawi, M. (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature Genetics*, *40*(12), 1466–1471.

Buiting, K. (2010). Prader-Willi syndrome and Angelman syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *154C(3*, 365–376.

Burns, K. H., & Boeke, J. D. (2012). Human transposon tectonics. *Cell*, *149*(4), 740–752.

Burnside, R. D. (2011). Microdeletion/microduplication of proximal 22q11.2: a susceptibility region for neurological and psychiatric disorders. *Journal of Medical Genetics*, *48*(9), 600–608.

Burnside, R. D. (2015). 22q11.21 Deletion Syndromes: A Review of Proximal, Central, and Distal Deletions and Their Associated Features. *Cytogenetic and Genome Research*, *146*(2), 89–99. https://doi.org/10.1159/000438708

Burnside, R. D., Pasion, R., Mikhail, F. M., Carroll, A. J., Robin, N. H., Youngs, E. L., Gadi, I. K., Keitges, E., Jaswaney, V. L., Papenhausen, P. R., Potluri, V. R., Risheg, H., Rush, B., Smith, J. L., Schwartz, S., Tepperberg, J. H., & Butler, M. G. (2011). Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. *Human Genetics*, *130*(4), 517–528. https://doi.org/10.1007/s00439-011-0970-4

Burssed, B., Zamariolli, M., Bellucco, F. T., & Melaragno, M. I. (2022). Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*, *15*(1), 23. https://doi.org/10.1186/s13039-022-00600-6

Calafato, M. S., & Bramon, E. (2019). The interplay between genetics, cognition and schizophrenia. In *Brain*. https://doi.org/10.1093/brain/awy345

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 3240. https://doi.org/10.1038/s41467-019-11146-4

Cameron, D., Mi, D., Vinh, N.-N., Webber, C., Li, M., Marín, O., O'Donovan, M. C., & Bray, N. J. (2023). Single-Nuclei RNA Sequencing of 5 Regions of the Human Prenatal Brain Implicates Developing Neuron Populations in Genetic Risk for Schizophrenia. *Biological Psychiatry*, *93*(2), 157–166. https://doi.org/10.1016/j.biopsych.2022.06.033

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

Carlsson, A., & Carlsson, M. L. (2006). A dopaminergic deficit hypothesis of schizophrenia: the path to discovery. *Dialogues in Clinical Neuroscience*, *8*(1), 137–142.

Carroll, L. S., Massey, T. H., Wardle, M., & Peall, K. J. (2018). Dentatorubral-pallidoluysian Atrophy: An Update. *Tremor and Other Hyperkinetic Movements (New York, N.Y.)*, *8*, 577. https://doi.org/10.7916/D81N9HST

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, *39*(S7), S16–S21. https://doi.org/10.1038/ng2028

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., & Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, *18*(8), 495–506. https://doi.org/10.1038/nrm.2017.48

Charlson, F. J., Ferrari, A. J., Santomauro, D. F., Diminic, S., Stockings, E., Scott, J. G., McGrath, J. J., & Whiteford, H. A. (2018). Global epidemiology and burden of schizophrenia: Findings from the global burden of disease study 2016. *Schizophrenia Bulletin*, *44*(6), 1195–1203.

Chen, L., Zhou, W., Zhang, L., & Zhang, F. (2014). Genome Architecture and Its Roles in Human Copy Number Variation. *Genomics & Informatics*, *12*(4), 136. https://doi.org/10.5808/GI.2014.12.4.136

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Chiarella, S. E., Li, W., Christie, K. J., & Frick, L. R. (2018). A new view of α-catenin in neurons. *Molecular and Cellular Neuroscience*, *89*, 59–67.

Chong, H. Y., Teixeira-Pinto, A., Abdin, E., Vaingankar, J. A., Subramaniam, M., Heng, D., & Chua, H. C. (2016). Global economic burden of schizophrenia: A systematic review. *Neuropsychiatric Disease and Treatment*, *12*, 357–373.

Cirillo, A., Lioncino, M., Maratea, A., Passariello, A., Fusco, A., Fratta, F., Monda, E., Caiazza, M., Signore, G., Esposito, A., Baban, A., Versacci, P., Putotto, C., Marino, B., Pignata, C., Cirillo, E., Giardino, G., Sarubbi, B., Limongelli, G., & Russo, M. G. (2022). Clinical Manifestations of 22q11.2 Deletion Syndrome. In *Heart Failure Clinics*. https://doi.org/10.1016/j.hfc.2021.07.009

Coid, J. W., Kirkbride, J. B., Barker, D., Cowden, F., Stamps, R., Yang, M., & Jones, P. B. (2008). Raised Incidence Rates of All Psychoses Among Migrant Groups. *Archives of General Psychiatry*, *65*(11), 1250. https://doi.org/10.1001/archpsyc.65.11.1250

Collins, F. S., Green, E. D., Guttmacher, A. E., & Guyer, M. S. (2003). A vision for the future of genomics research. *Nature*, *422*(6934), 835–847. https://doi.org/10.1038/nature01626

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., & Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712.

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703.

Craig, F., Margari, F., Legrottaglie, A. R., Palumbi, R., Giambattista, C., & Margari, L. (2016). A review of executive function deficits in autism spectrum disorder and attention-deficit/hyperactivity disorder. *Neuropsychiatric Disease and Treatment*, *12*, 1191–1202.

Creeth, H. D. J., Rees, E., Legge, S. E., Dennison, C. A., Holmans, P., Walters, J. T. R., O'Donovan, M. C., & Owen, M. J. (2022). Ultrarare Coding Variants and Cognitive Function in Schizophrenia. *JAMA Psychiatry*, *79*(10), 963. https://doi.org/10.1001/jamapsychiatry.2022.2289

Crespi, B. J., Hurd, P. L., & Dinsdale, N. (2010). Cognitive-behavioral phenotypes of Williams syndrome are associated with genetic variation in the GTF2I gene, in a healthy population. *BMC Neuroscience*, *11*(1), 139.

Curley, A. A., Arion, D., Volk, D. W., Asafu-Adjei, J. K., Sampson, A. R., Fish, K. N., & Lewis, D. A. (2011). Cortical deficits of glutamic acid decarboxylase 67 expression in schizophrenia: Clinical, protein, and cell type-specific features. *The American Journal of Psychiatry*, *168*(9), 921–929.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, *11*(1), 126. https://doi.org/10.1186/1741-7015-11-126

de Leon, J., & Diaz, F. J. (2005). A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. *Schizophrenia Research*, *76*(2–3), 135–157. https://doi.org/10.1016/j.schres.2005.02.010

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21.

Deininger, P. (2011). Alu elements: know the SINEs. *Genome Biology*, *12*(12), 236.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Di Forti, M., Marconi, A., Carra, E., Fraietta, S., Trotta, A., Bonomo, M., Bianconi, F., Gardner-Sood, P., O'Connor, J., Russo, M., Stilo, S. A., Marques, T. R., Mondelli, V., Dazzan, P., Pariante, C., David, A. S., Gaughran, F., Atakan, Z., Iyegbe, C., … Murray, R. M. (2015). Proportion of patients in south London with first-episode psychosis attributable to use of high potency cannabis: a case-control study. *The Lancet Psychiatry*, *2*(3), 233–238. https://doi.org/10.1016/S2215-0366(14)00117-5

Eaton, W. W., Byrne, M., Ewald, H., Mors, O., Chen, C. Y., Agerbo, E., & Mortensen, P. B. (2006). Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *American Journal of Psychiatry*, *163*(3), 521–528.

Egan, M. F. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences*, *98*(12), 6917–6922.

Ellison-Wright, I., & Bullmore, E. (2009). Meta-analysis of diffusion tensor imaging studies in schizophrenia. *Schizophrenia research*, *108*(1–3), 3–10. https://doi.org/10.1016/j.schres.2008.11.021

Erp, T. G., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., & Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 585–594. https://doi.org/10.1038/mp.2015.63

Esnault, C., Maestre, J., & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, *24*(4), 363–367.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, *29*(1), 51–76.

Falls, J. G., Pulford, D. J., Wylie, A. A., & Jirtle, R. L. (1999). Genomic imprinting: implications for human disease. *The American Journal of Pathology*, *154*(3), 635–647.

Fazel, S., & Grann, M. (2006). The population impact of severe mental illness on violent crime. *The American Journal of Psychiatry*, *163*(8), 1397–1403.

Fazel, S., Gulati, G., Linsell, L., Geddes, J. R., & Grann, M. (2009). Schizophrenia and violence: Systematic review and meta-analysis. *PLoS Medicine*, *6*(8), 1000120.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*(2), 85–97. https://doi.org/10.1038/nrg1767

Fink, M., & Taylor, M. A. (2006). *Catatonia: A clinician's guide to diagnosis and treatment*. Cambridge University Press.

Foley, C., Heron, E. A., Harold, D., Walters, J., Owen, M., O'Donovan, M., Sebat, J., Kelleher, E., Mooney, C., Durand, A., Pinto, C., Cormican, P., Morris, D., Donohoe, G., Gill, M., Gallagher, L., & Corvin, A. (2020). Identifying schizophrenia patients who carry pathogenic genetic copy number variants using standard clinical assessment: Retrospective cohort study. *British Journal of Psychiatry*. https://doi.org/10.1192/bjp.2019.262

Foussias, G., & Remington, G. (2010). Negative symptoms in schizophrenia: Avolition and Occam's razor. *Schizophrenia Bulletin*, *36*(2), 359–369.

Friston, K. (2011). Functional and effective connectivity: a review. *Brain Connectivity*, *1*(1), 13–36. https://doi.org/10.1089/brain.2011.0008

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., & Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2012.08.005

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., & Owen, M. J. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature*, *506*(7487), 179–184.

Frydecka, D., Krzystek-Korpacka, M., Lubeiro, A., Stramecki, F., Stańczykiewicz, B., Beszłej, J. A., Piotrowski, P., Kotowicz, K., Szewczuk-Bogusławska, M., Pawlak-Adamska, E., & Misiak, B. (2018). Profiling inflammatory signatures of schizophrenia: A cross-sectional and meta-analysis study. *Brain, Behavior, and Immunity*, *71*, 28–36. https://doi.org/10.1016/j.bbi.2018.05.002

Garden, G. A., & Campbell, B. M. (2016). Glial biomarkers in human central nervous system disease. *Glia*, *64*(10), 1755–1771. https://doi.org/10.1002/glia.22998

Gardner, E. J., Sifrim, A., Lindsay, S. J., Prigmore, E., Rajan, D., Danecek, P., Gallone, G., Eberhardt, R. Y., Martin, H. C., Wright, C. F., FitzPatrick, D. R., Firth, H. V., & Hurles, M. E. (2021). Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *The American Journal of Human Genetics*, *108*(11), 2186–2194. https://doi.org/10.1016/j.ajhg.2021.09.010

Garety, P. A., & Freeman, D. (1999). Cognitive approaches to delusions: a critical review of theories and evidence. *British Journal of Clinical Psychology*, *38*(2), 113–154.

Glassford, M. R., Rosenfeld, J. A., Freedman, A. A., Zwick, M. E., & Mulle, J. G. (2016). Novel features of 3q29 deletion syndrome: Results from the 3q29 registry. *American Journal of Medical Genetics Part A*, *170*(4), 999–1006. https://doi.org/10.1002/ajmg.a.37537

Glausier, J. R., & Lewis, D. A. (2011). Selective pyramidal cell reduction of GABA(A) receptor α1 subunit messenger RNA expression in schizophrenia. *Neuropsychopharmacology*, *36*(10), 2103–2110.

Glausier, J. R., & Lewis, D. A. (2013). Dendritic spine pathology in schizophrenia. *Neuroscience*, *251*, 90–107.

Goff, D. C., & Wine, L. (2008). Glutamate in schizophrenia: clinical and research implications. *Schizophr Res*, *100*(1–3), 2–12.

Goldman-Rakic, P. S., Castner, S. A., Svensson, T. H., Siever, L. J., & Williams, G. V. (2000). Targeting the dopamine D1 receptor in schizophrenia: insights for cognitive dysfunction. *Psychopharmacology*, *147*(3), 227–233. https://doi.org/10.1007/s002130051018

Goldsmith, D. R., Rapaport, M. H., & Miller, B. J. (2016). A meta-analysis of blood cytokine network alterations in psychiatric patients: comparisons between schizophrenia, bipolar disorder and depression. *Molecular Psychiatry*, *21*(12), 1696–1709.

Gothelf, D., Eliez, S., Thompson, T., Hinard, C., Penniman, L., Feinstein, C., & Reiss, A. L. (2004). COMT genotype predicts longitudinal cognitive decline and psychosis in 22q11.2 deletion syndrome. *Nature Neuroscience*, *8*(11), 1500–1502.

Gottesman, I. I., & Shields, J. (1972). *Schizophrenia and genetics: A twin study vantage point*. Academic Press.

Grace, A. A. (2016). Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression. *Nature Reviews Neuroscience*, *17*(8), 524–532. https://doi.org/10.1038/nrn.2016.57

Green, M. F. (2006). Cognitive impairment and functional outcome in schizophrenia and bipolar disorder. In *Journal of Clinical Psychiatry*.

Gur, R. E., Cowell, P. E., Latshaw, A., Turetsky, B. I., Grossman, R. I., Arnold, S. E., & Gur, R. C. (2000). Reduced dorsal and orbital prefrontal gray matter volumes in schizophrenia. *Archives of General Psychiatry*, *57*(8), 761–768. https://doi.org/10.1001/archpsyc.57.8.761

Häfner, H., Riecher-Rössler, A., Hambrecht, M., Maurer, K., Meissner, S., Schmidtke, A., Fätkenheuer, B., Löffler, W., & van der Heiden, W. (1992). IRAOS: an instrument for the assessment of onset and early course of schizophrenia. *Schizophrenia Research*. https://doi.org/10.1016/0920-9964(92)90004-O

Haijma, S. V, Haren, N., Cahn, W., Koolschijn, P. C., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain volumes in schizophrenia: a meta-analysis in over 18,000 subjects. *Schizophrenia Bulletin*, *39*(5), 1129–1138. https://doi.org/10.1093/schbul/sbs118

Hakulinen, C., McGrath, J. J., Timmerman, A., Skipper, N., Mortensen, P. B., Pedersen, C. B., & Agerbo, E. (2019). The association between early-onset schizophrenia with employment, income, education, and cohabitation status: nationwide study with 35 years of follow-up. *Social Psychiatry and Psychiatric Epidemiology*, *54*(11), 1343–1351. https://doi.org/10.1007/s00127-019-01756-0

Haraksingh, R. R., Abyzov, A., & Urban, A. E. (2017). Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics*, *18*(1), 321. https://doi.org/10.1186/s12864-017-3658-x

Haren, N. E., Hulshoff Pol, H. E., Schnack, H. G., Cahn, W., Brans, R., Carati, I., & Kahn, R. S. (2011). Progressive brain volume loss in schizophrenia over the course of the illness: evidence of maturational abnormalities in early adulthood. *Biological Psychiatry*, *69*(1), 106–113. https://doi.org/10.1016/j.biopsych.2010.06.006

Harrison, P. J. (1999). The neuropathology of schizophrenia: a critical review of the data and their interpretation. *Brain*, *122*(4), 593–624. https://doi.org/10.1093/brain/122.4.593

Hartz, S. M., Pato, C. N., Medeiros, H., Cavazos-Rehg, P., Sobell, J. L., Knowles, J. A., & Pato, M. T. (2014). Comorbidity of severe psychotic disorders with measures of substance use. *JAMA Psychiatry*, *71*(3), 248–254.

Harvard, C., Malenfant, P., Koochek, M., Creighton, S., Mickelson, E. C., Holden, J. J., & Lewis, M. E. (2011). A variant in the 3' untranslated region of the GRIA3 transcript is associated with gene expression changes in subsets of autism brain. *Journal of Medical Genetics*, *48*(5), 317–321.

Hashimoto, T., Bazmi, H. H., Mirnics, K., Wu, Q., Sampson, A. R., & Lewis, D. A. (2008). Conserved regional patterns of GABA-related transcript expression in the neocortex of subjects with schizophrenia. *The American Journal of Psychiatry*, *165*(4), 479–489.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, *10*(8), 551–564. https://doi.org/10.1038/nrg2593

Heinrichs, R. W., & Zakzanis, K. K. (1998). Neurocognitive deficit in schizophrenia: A quantitative review of the evidence. *Neuropsychology*, *12*(3), 426–445.

Helbig, I. (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature Genetics*, *41*(2), 160–162.

Hiatt, S. M., Thompson, M. L., Prokop, J. W., Lawlor, J. M. J., Gray, D. E., Bebin, E. M., Rinne, T., Kempers, M., Pfundt, R., van Bon, B. W., Mignot, C., Nava, C., Depienne, C., Kalsner, L., Rauch, A., Joset, P., Bachmann-Gagescu, R., Wentzensen, I. M., McWalter, K., & Cooper, G. M. (2019). Deleterious Variation in BRSK2 Associates with a Neurodevelopmental Disorder. *The American Journal of Human Genetics*, *104*(4), 701–708. https://doi.org/10.1016/j.ajhg.2019.02.002

Hjorthøj, C., Stürup, A. E., McGrath, J. J., & Nordentoft, M. (2017). Years of potential life lost and life expectancy in schizophrenia: A systematic review and meta-analysis. *The Lancet Psychiatry*, *4*(4), 295–301.

Ho, B. C., Andreasen, N. C., Nopoulos, P., Arndt, S., Magnotta, V., & Flaum, M. (2003). Progressive structural brain abnormalities and their relationship to clinical outcome: a longitudinal magnetic resonance imaging study early in schizophrenia. *Archives of General Psychiatry*, *60*(6), 585–594. https://doi.org/10.1001/archpsyc.60.6.585

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

Hollenbeck, D., Williams, C. L., Drazba, K., Descartes, M., Korf, B. R., Rutledge, S. L., Lose, E. J., Robin, N. H., Carroll, A. J., & Mikhail, F. M. (2017). Clinical relevance of small copy-number variants in chromosomal microarray clinical testing. *Genetics in Medicine*, *19*(4), 377–385. https://doi.org/10.1038/gim.2016.132

Hor, K., & Taylor, M. (2010). Suicide and schizophrenia: A systematic review of rates and risk factors. *Journal of Psychopharmacology*, *24*(4 Suppl), 81–90.

Horsthemke, B., & Wagstaff, J. (2008). Mechanisms of imprinting of the Prader-Willi/Angelman region. *American Journal of Medical Genetics Part A*, *146A(16*, 2041–2052.

Howe, L. J., Lee, M. K., Sharp, G. C., Davey Smith, G., St Pourcain, B., Shaffer, J. R., Ludwig, K. U., Mangold, E., Marazita, M. L., Feingold, E., Zhurov, A., Stergiakouli, E., Sandy, J., Richmond, S., Weinberg, S. M., Hemani, G., & Lewis, S. J. (2018). Investigating the shared genetics of non-syndromic cleft lip/palate and facial morphology. *PLOS Genetics*, *14*(8), e1007501. https://doi.org/10.1371/journal.pgen.1007501

Howes, O. D., Kambeitz, J., Kim, E., Stahl, D., Slifstein, M., Abi-Dargham, A., & Kapur, S. (2012). The nature of dopamine dysfunction in schizophrenia and what this means for treatment: Meta-analysis of imaging studies. *Archives of General Psychiatry*, *69*(8), 776–786. https://doi.org/10.1001/archgenpsychiatry.2012.169

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: version III—the final common pathway. *Schizophrenia Bulletin*, *35*(3), 549–562.

Howrigan, D. P., Rose, S. A., Samocha, K. E., Fromer, M., Cerrato, F., Chen, W. J., Churchhouse, C., Chambert, K., Chandler, S. D., Daly, M. J., Dumont, A., Genovese, G., Hwu, H.-G., Laird, N., Kosmicki, J. A., Moran, J. L., Roe, C., Singh, T., Wang, S.-H., … Neale, B. M. (2020). Exome sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by protein-coding de novo mutations. *Nature Neuroscience*, *23*(2), 185–193. https://doi.org/10.1038/s41593-019-0564-3

Hubbard, L., Rees, E., Morris, D. W., Lynham, A. J., Richards, A. L., Pardiñas, A. F., Legge, S. E., Harold, D., Zammit, S., Corvin, A. C., Gill, M. G., Hall, J., Holmans, P., O'Donovan, M. C., Owen, M. J., Donohoe, G., Kirov, G., Pocklington, A., & Walters, J. T. R. (2021). Rare Copy Number Variants Are Associated With Poorer Cognition in Schizophrenia. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2020.11.025

Iannitelli, A., Quartini, A., Tirassa, P., & Bersani, G. (2017). Schizophrenia and neurogenesis: A stem cell approach. *Neuroscience & Biobehavioral Reviews*, *80*, 414–442. https://doi.org/10.1016/j.neubiorev.2017.06.010

Ingason, A., Rujescu, D., Cichon, S., Sigurdsson, E., Sigmundsson, T., Pietiläinen, O. P. H., Buizer-Voskamp, J. E., Strengman, E., Francks, C., Muglia, P., Gylfason, A., Gustafsson, O., Olason, P. I., Steinberg, S., Hansen, T., Jakobsen, K. D., Rasmussen, H. B., Giegling, I., Möller, H.-J., … Clair, D. M. S. (2011). Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Molecular Psychiatry*, *16*(1), 17–25. https://doi.org/10.1038/mp.2009.101

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

Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., & Eichler, E. E. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *The American Journal of Human Genetics*, *84*(2), 148–161.

Jääskeläinen, E., Juola, P., Hirvonen, N., McGrath, J. J., Saha, S., Isohanni, M., & Miettunen, J. (2013). A systematic review and meta-analysis of recovery in schizophrenia. *Schizophrenia Bulletin*, *39*(6), 1296–1306.

Jablensky, A. (2000). *Epidemiology of schizophrenia: The global burden of disease and disability*.

Javitt, D. C. (2010). Glutamatergic theories of schizophrenia. *Israel Journal of Psychiatry and Related Sciences*, *47*(1), 4–16.

Jeevakumar, V., & Kroener, S. (2016). Ketamine Administration During the Second Postnatal Week Alters Synaptic Properties of Fast-Spiking Interneurons in the Medial Prefrontal Cortex of Adult Mice. *Cerebral Cortex*, *26*(3), 1117–1129. https://doi.org/10.1093/cercor/bhu293

Jeremy Willsey, A., & State, M. W. (2015). Autism spectrum disorders: from genes to neurobiology. *Current Opinion in Neurobiology*, *30*, 92–99. https://doi.org/10.1016/j.conb.2014.10.015

Jones, D. K., Knösche, T. R., & Turner, R. (2013). White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *Neuroimage*, *73*, 239–254. https://doi.org/10.1016/j.neuroimage.2012.06.081

Jones, R. M., Lichtenstein, P., Grann, M., Långström, N., & Fazel, S. (2011). Alcohol use disorders in schizophrenia: a national cohort study of 12,653 patients. *The Journal of Clinical Psychiatry*, *72*(6), 775–779.

Kaplan, H. I., & Sadock, B. J. (2015). *Kaplan and Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry*. Wolters Kluwer.

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., Short, P. J., Torene, R. I., de Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., … Retterer, K. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, *586*(7831), 757–762. https://doi.org/10.1038/s41586-020-2832-5

Karbarz, M. (2020). Consequences of 22q11.2 Microdeletion on the Genome, Individual and Population Levels. *Genes*, *11*(9), 977. https://doi.org/10.3390/genes11090977

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Daly, M. J. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. https://doi.org/10.1038/s41586-020-2308-7

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Karlsgodt, K. H., Erp, T. G., Poldrack, R. A., Bearden, C. E., Nuechterlein, K. H., & Cannon, T. D. (2008). Diffusion tensor imaging of the superior longitudinal fasciculus and working memory in recent-onset schizophrenia. *Biological Psychiatry*, *63*(5), 512–518. https://doi.org/10.1016/j.biopsych.2007.06.017

Katsel, P., Byne, W., Roussos, P., Tan, W., Siever, L., & Haroutunian, V. (2011). Astrocyte and glutamate markers in the superficial, deep, and white matter layers of the anterior cingulate gyrus in schizophrenia. *Neuropsychopharmacology*, *36*(6), 1171–1177.

Kazazian Jr., H. H. (2004). Mobile elements: drivers of genome evolution. *Science*, *303*(5664), 1626–1632.

Keefe, R. S. E., Eesley, C. E., & Poe, M. P. (2005). Defining a cognitive function decrement in schizophrenia. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2005.01.003

Keefe, R. S. E., & Harvey, P. D. (2012). *Cognitive Impairment in Schizophrenia* (pp. 11–37). https://doi.org/10.1007/978-3-642-25758-2_2

Kelly, B. D., O'Callaghan, E., Waddington, J. L., Feeney, L., Browne, S., Scully, P. J., Clarke, M., Quinn, J. F., McTigue, O., Morgan, M. G., Kinsella, A., & Larkin, C. (2010). Schizophrenia and the city: A review of literature and prospective study of psychosis and urbanicity in Ireland. *Schizophrenia Research*, *116*(1), 75–89. https://doi.org/10.1016/j.schres.2009.10.015

Kelly, M. L., & Chernoff, J. (2012). Mouse models of PAK function. *Cellular Logistics*, *2*(2), 84–88. https://doi.org/10.4161/cl.21381

Kendall, K. M., Rees, E., Escott-Price, V., Einon, M., Thomas, R., Hewitt, J., O'Donovan, M. C., Owen, M. J., Walters, J. T. R., & Kirov, G. (2017). Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biological Psychiatry*, *82*(2), 103–110. https://doi.org/10.1016/j.biopsych.2016.08.014

Keshavan, M. S., Morris, D. W., Sweeney, J. A., Pearlson, G., Thaker, G., Seidman, L. J., Eack, S. M., & Tamminga, C. (2011). A dimensional approach to the psychosis spectrum between bipolar disorder and schizophrenia: The Schizo-Bipolar Scale. *Schizophrenia Research*, *133*(1–3), 250–254.

Khandaker, G. M., Cousins, L., Deakin, J., Lennox, B. R., Yolken, R., & Jones, P. B. (2015). Inflammation and immunity in schizophrenia: implications for pathophysiology and treatment. *The Lancet Psychiatry*, *2*(3), 258–270.

Kim, J. S., Kornhuber, H. H., Schmid-Burgk, W., & Holzmüller, B. (1980). Low cerebrospinal fluid glutamate in schizophrenic patients and a new hypothesis on schizophrenia. *Neurosci Lett*, *20*(4), 379–382.

Kim, J., Kim, S., Nam, H., Kim, S., & Lee, D. (2015). SoloDel: a probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data. *Bioinformatics*, *31*(19), 3105–3113. https://doi.org/10.1093/bioinformatics/btv358

Kimberley M. Kendall, George Kirov, & Michael J. Owen. (2017). Genetics of Schizophrenia. In *Kaplan and Saddock's Comprehensive Textbook of Psychiatry*. Wolters Kluwer.

Kinon, B. J., Chen, L., Ascher-Svanum, H., Stauffer, V. L., Kollack-Walker, S., Sniadecki, J. L., & Kane, J. M. (2010). Predicting response to atypical antipsychotics based on early response in the treatment of schizophrenia. *Schizophrenia Research*, *120*(1–3), 23–30.

Kirkbride, J. B., Hameed, Y., Ioannidis, K., Ankireddypalli, G., Crane, C. M., Nasir, M., Kabacs, N., Metastasio, A., Jenkins, O., Espandian, A., Spyridi, S., Ralevic, D., Siddabattuni, S., Walden, B., Adeoye, A., Perez, J., & Jones, P. B. (2017). Ethnic Minority Status, Age-at-Immigration and Psychosis Risk in Rural Environments: Evidence From the SEPEA Study. *Schizophrenia Bulletin*, *43*(6), 1251–1261. https://doi.org/10.1093/schbul/sbx010

Kirkpatrick, B., & Galderisi, S. (2008). Deficit schizophrenia: An upyear. *World Psychiatry*, *7*(3), 143–147.

Kirov, G., Gumus, D., Chen, W., Norton, N., Georgieva, L., Sari, M., O'Donovan, M. C., Erdogan, F., Owen, M. J., Ropers, H.-H., & Ullmann, R. (2007). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Human Molecular Genetics*, *17*(3), 458–465. https://doi.org/10.1093/hmg/ddm323

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L., Grozeva, D., Fjodorova, M., Wollerton, R., Rees, E., Nikolov, I., Van De Lagemaat, L. N., Bayés, A., Fernandez, E., Olason, P. I., … Owen, M. J. (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry*. https://doi.org/10.1038/mp.2011.154

Kirov, G., Rees, E., Walters, J. T. R., Escott-Price, V., Georgieva, L., Richards, A. L., Chambert, K. D., Davies, G., Legge, S. E., Moran, J. L., McCarroll, S. A., O'Donovan, M. C., & Owen, M. J. (2014). The Penetrance of Copy Number Variations for Schizophrenia and Developmental Delay. *Biological Psychiatry*, *75*(5), 378–385. https://doi.org/10.1016/j.biopsych.2013.07.022

Kirov, G., Rujescu, D., Ingason, A., Collier, D. A., O'Donovan, M. C., & Owen, M. J. (2009). Neurexin 1 (NRXN1) Deletions in Schizophrenia. *Schizophrenia Bulletin*, *35*(5), 851–854. https://doi.org/10.1093/schbul/sbp079

Kishino, T., Lalande, M., & Wagstaff, J. (1997). UBE3A/E6-AP mutations cause Angelman syndrome. *Nature Genetics*, *15*(1), 70–73.

Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Möller, H. J., & Riecher-Rössler, A. (2016). Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain*, *139*(7), 2059–2073.

Kraepelin, E. (1896). *Psychiatrie. Ein Lehrbuch für Studierende und Ärzte*.

Kraepelin, E. (1919). *Dementia praecox and paraphrenia*. Chicago Medical Book.

Kraus, M. S., & Keefe, R. S. E. (2007). Cognition as an outcome measure in schizophrenia. In *British Journal of Psychiatry*. https://doi.org/10.1192/bjp.191.50.s46

Kroken, R. A., Sommer, I. E., Steen, V. M., Dieset, I., Johnsen, E., & Jørgensen, H. A. (2018). Constructing the immune signature of schizophrenia for clinical use and research; an integrative review translating descriptives into diagnostics. *Frontiers in Psychiatry*, *9*, 753.

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., NHLBI Exome Sequencing Project, Quinlan, A. R., Nickerson, D. A., & Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, *22*(8), 1525–1532. https://doi.org/10.1101/gr.138115.112

Kronenberg, Z. N., Osborne, E. J., Cone, K. R., Kennedy, B. J., Domyan, E. T., Shapiro, M. D., Elde, N. C., & Yandell, M. (2015). Wham: Identifying Structural Variants of Biological Consequence. *PLOS Computational Biology*, *11*(12), e1004572. https://doi.org/10.1371/journal.pcbi.1004572

Krystal, J. H., Karper, L. P., Seibyl, J. P., Freeman, G. K., Delaney, R., Bremner, J. D., Heninger, G. R., MB Jr, B., & Charney, D. S. (1994). Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. Psychotomimetic, perceptual, cognitive, and neuroendocrine responses. *Arch Gen Psychiatry*, *51*(3), 199–214.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., & Funke, R. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. *Cell*, *177*(1), 70–84. https://doi.org/10.1016/j.cell.2019.02.032

Larøi, F. (2012). How do auditory verbal hallucinations in patients differ from those in non-patients? *Frontiers in Human Neuroscience*, *6*, 25.

Larsson, S., Andreassen, O. A., Aas, M., Røssberg, J. I., Mork, E., Steen, N. E., & Agartz, I. (2013). High prevalence of childhood trauma in patients with schizophrenia spectrum and affective disorder. *Comprehensive Psychiatry*, *54*(2), 123–127.

Lau, C. G., & Zukin, R. S. (2007). NMDA receptor trafficking in synaptic plasticity and neuropsychiatric disorders. *Nature Reviews Neuroscience*, *8*(6), 413–426.

Lavoie, S., Murray, M. M., Deppen, P., Knyazeva, M. G., Berk, M., Boulat, O., & Conus, P. (2007). Glutathione precursor, N-acetyl-cysteine, improves mismatch negativity in schizophrenia patients. *Neuropsychopharmacology*, *32*(9), 2154–2163.

Lee, E. E., Martin, A. S., Tu, X., Palmer, B. W., & Jeste, D. V. (2018). Childhood Adversity and Schizophrenia. *The Journal of Clinical Psychiatry*, *79*(3). https://doi.org/10.4088/JCP.17m11776

Leucht, S., Leucht, C., Huhn, M., Chaimani, A., Mavridis, D., Helfer, B., & Davis, J. M. (2017). Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: Systematic review, Bayesian meta-analysis, and meta-regression of efficacy predictors. *The American Journal of Psychiatry*, *174*(10), 927–942.

Levinson, D. F., Duan, J., Oh, S., Wang, K., Sanders, A. R., Shi, J., Zhang, N., Mowry, B. J., Olincy, A., Amin, F., Cloninger, C. R., Silverman, J. M., Buccola, N. G., Byerley, W. F., Black, D. W., Kendler, K. S., Freedman, R., Dudbridge, F., Pe'er, I., … Gejman, P. V. (2011). Copy Number Variants in Schizophrenia: Confirmation of Five Previous Findings and New Evidence for 3q29 Microdeletions and VIPR2 Duplications. *American Journal of Psychiatry*, *168*(3), 302–316. https://doi.org/10.1176/appi.ajp.2010.10060876

Li, B.-J., Liu, P., Chu, Z., Shang, Y., Huan, M.-X., Dang, Y.-H., & Gao, C.-G. (2017). Social isolation induces schizophrenia-like behavior potentially associated with HINT1, NMDA receptor 1, and dopamine receptor 2. *Neuroreport*, *28*(8), 462–469. https://doi.org/10.1097/WNR.0000000000000775

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Ma, X., Wang, G., Yang, J., & Wang, C. (2016). Why sex differences in schizophrenia? *Journal of Translational Neuroscience*, *1*(1), 37–42.

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *The Lancet*, *373*(9659), 234–239. https://doi.org/10.1016/S0140-6736(09)60072-6

Lieberman, J. A., Kane, J. M., & Alvir, J. (1990). Provocative tests with psychostimulant drugs in schizophrenia. *Psychopharmacology*, *101*(3), 331–333. https://doi.org/10.1007/BF02244129

Lieberman, J. A., Tollefson, G. D., Charles, C., Zipursky, R., Sharma, T., Kahn, R. S., & Vanover, K. (2005). Antipsychotic drug effects on brain morphology in first-episode psychosis. *Archives of General Psychiatry*, *62*(4), 361–370. https://doi.org/10.1001/archpsyc.62.4.361

Liloia, D., Ginevrino, M., & Galderisi, S. (2021). Brain changes in ultra-high risk subjects: a systematic review of MRI studies. In *European archives of psychiatry and clinical neuroscience* (pp. 1–18). https://doi.org/10.1007/s00406-021-01240-w

Lisman, J. E., Coyle, J. T., Green, R. W., Javitt, D. C., Benes, F. M., Heckers, S., & Grace, A. A. (2008). Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. *Trends in Neurosciences*, *31*(5), 234–242.

Liu, D., Meyer, D., Fennessy, B., Feng, C., Cheng, E., Johnson, J. S., Park, Y. J., Rieder, M.-K., Ascolillo, S., de Pins, A., Dobbyn, A., Lebovitch, D., Moya, E., Nguyen, T.-H., Wilkins, L., Hassan, A., Aghanwa, H. S., Ansari, M., Asif, A., … Charney, A. W. (2023). Schizophrenia risk conferred by rare protein-truncating variants is conserved across diverse human populations. *Nature Genetics*, *55*(3), 369–376. https://doi.org/10.1038/s41588-023-01305-1

Lupski, J. R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, *14*(10), 417–422.

Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nature Genetics*, *39*(7S), 43– 47.

Lynham, A. J., Hubbard, L., Tansey, K. E., Hamshere, M. L., Legge, S. E., Owen, M. J., Jones, I. R., & Walters, J. T. R. (2018). Examining cognition across the bipolar/schizophrenia diagnostic spectrum. *Journal of Psychiatry and Neuroscience*. https://doi.org/10.1503/jpn.170076

Malaspina, D., Owen, M. J., Heckers, S., Tandon, R., Bustillo, J., Schultz, S., Barch, D. M., Gaebel, W., Gur, R. E., & Tsuang, M. (2013). Schizoaffective disorder in the DSM-5. *Schizophrenia Research*, *150*(1), 21–25.

Malhotra, D., McCarthy, S., Michaelson, J. J., Vacic, V., Burdick, K. E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, E. S., Gill, M., Karayiorgou, M., Kelsoe, J. R., Krastoshevsky, O., Krause, V., Leibenluft, E., Levy, D. L., Makarov, V., Bhandari, A., … Sebat, J. (2011). High Frequencies of De Novo CNVs in Bipolar Disorder and Schizophrenia. *Neuron*, *72*(6), 951–963. https://doi.org/10.1016/j.neuron.2011.11.007

Malhotra, D., & Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, *148*(6), 1223–1241.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

March, D., Hatch, S. L., Morgan, C., Kirkbride, J. B., Bresnahan, M., Fearon, P., & Susser, E. (2008). Psychosis and Place. *Epidemiologic Reviews*, *30*(1), 84–100. https://doi.org/10.1093/epirev/mxn006

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511.

Marchuk, D. S., Crooks, K., Strande, N., Kaiser-Rogers, K., Milko, L. V., Brandt, A., Arreola, A., Tilley, C. R., Bizon, C., Vora, N. L., Wilhelmsen, K. C., Evans, J. P., & Berg, J. S. (2018). Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLOS ONE*, *13*(12), e0209185. https://doi.org/10.1371/journal.pone.0209185

Marconi, A., Forti, M., Lewis, C. M., Murray, R. M., & Vassos, E. (2016). Meta-analysis of the association between the level of cannabis use and risk of psychosis. *Schizophrenia Bulletin*, *42*(5), 1262–1269.

Marder, S. R., & Fenton, W. (2004). Measurement and Treatment Research to Improve Cognition in Schizophrenia: NIMH MATRICS initiative to support the development of agents for improving cognition in schizophrenia. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2004.09.010

Mardis, E. R. (2008a). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402.

Mardis, E. R. (2008b). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402.

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajarado, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., … Sebat, J. (2017a). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, *49*(1), 27–35. https://doi.org/10.1038/ng.3725

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fuentes Fajarado, K. V., Maile, M. S., Ripke, S., Agartz, I., Albus, M., … Sebat, J. (2017b). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*. https://doi.org/10.1038/ng.3725

Maxwell, E., Packer, J., O'Dushlaine, C., McCarthy, S., Hare-Harris, A., Staples, J., Gonzaga-Jauregui, C., Fetterolf, S., Faucett, W. A., Leader, J., Moreno-De-Luca, A., Gatta, G. Della, Scollan, M., Persaud, T., Penn, J., Hawes, A., Bai, X., Wolf, S., Lopez, A., … Reid, J. (2017). Profiling copy number variation and disease associations from 50,726 DiscovEHR Study exomes. *BioRxiv*. https://doi.org/10.1101/119461

McDonald-McGinn, D. M. (2015a). 22q11.2 deletion syndrome. *Nature Reviews Disease Primers*, *1*, 15071.

McDonald-McGinn, D. M. (2015b). 22q11.2 deletion syndrome. *Nature Reviews Disease Primers*, *1*, 15071.

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

Mefford, H. C., Clauin, S., Sharp, A. J., Moller, R. S., Ullmann, R., Kapur, R., & Bellanne-Chantelot, C. (2008). Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *The American Journal of Human Genetics*, *83*(5), 572–581.

Merla, G., Brunetti-Pierri, N., Micale, L., & Fusco, C. (2010). Copy number variants at Williams–Beuren syndrome 7q11.23 region. *Human Genetics*, *128*(1), 3–26. https://doi.org/10.1007/s00439-010-0827-2

Merscher, S., Funke, B., Epstein, J. A., Heyer, J., Puech, A., Lu, M. M., & Korenberg, J. R. (2001). TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell*, *104*(4), 619–629.

Mesholam-Gately, R. I., Giuliano, A. J., Goff, K. P., Faraone, S. V, & Seidman, L. J. (2009). Neurocognition in first-episode schizophrenia: A meta-analytic review. *Neuropsychology*, *23*(3), 315–336.

Messias, E. L., Chen, C.-Y., & Eaton, W. W. (2007). Epidemiology of Schizophrenia: Review of Findings and Myths. *Psychiatric Clinics of North America*, *30*(3), 323–338. https://doi.org/10.1016/j.psc.2007.04.007

Miller, B. J., Buckley, P., Seabolt, W., Mellor, A., & Kirkpatrick, B. (2011). *Meta-analysis of cytokine alterations in schizophrenia: clinical status and antips*.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., … Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*(7332), 59–65. https://doi.org/10.1038/nature09708

Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature*, *447*(7147), 932–940.

Moghaddam, B., Adams, B., Verma, A., & Daly, D. (1997). Activation of glutamatergic neurotransmission by ketamine: a novel step in the pathway from NMDA receptor blockade to dopaminergic and cognitive disruptions associated with the prefrontal cortex. *J Neurosci*, *17*(8), 2921–2927.

Moghaddam, B., & Javitt, D. (2012). From revolution to evolution: the glutamate hypothesis of schizophrenia and its implication for treatment. *Neuropsychopharmacology*, *37*(1), 4–15.

Möhler, H. (2012). The GABA system in anxiety and depression and its therapeutic potential. *Neuropharmacology*, *62*(1), 42–53.

Möller, H. J. (2007). Clinical evaluation of negative symptoms in schizophrenia. *European Psychiatry*, *22*(6), 380–386.

Morales, M., & Margolis, E. B. (2017). Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nature Reviews Neuroscience*, *18*(2), 73–85.

Morgan, C., & Fisher, H. (2007). Environment and schizophrenia: environmental factors in schizophrenia: childhood trauma—a critical review. *Schizophrenia Bulletin*, *33*(1), 3–10.

Mulle, J. G., Dodd, A. F., McGrath, J. A., Wolyniec, P. S., Mitchell, A. A., Shetty, A. C., Sobreira, N. L., Valle, D., Rudd, M. K., Satten, G., Cutler, D. J., Pulver, A. E., & Warren, S. T. (2010). Microdeletions of 3q29 Confer High Risk for Schizophrenia. *The American Journal of Human Genetics*, *87*(2), 229–236. https://doi.org/10.1016/j.ajhg.2010.07.013

Müller, N., Weidinger, E., Leitner, B., & Schwarz, M. J. (2015). The role of inflammation in schizophrenia. *Frontiers in Neuroscience*, *9*. https://doi.org/10.3389/fnins.2015.00372

Murphy, K. C. (1999). High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Archives of General Psychiatry*, *56*(10), 940–945.

Mustonen, A., Niemelä, S., Nordström, T., Murray, G. K., Mäki, P., Jääskeläinen, E., & Miettunen, J. (2018). Adolescent cannabis use, baseline prodromal symptoms and the risk of psychosis. *The British Journal of Psychiatry*, *212*(4), 227–233. https://doi.org/10.1192/bjp.2017.52

Nakamura, M., Salisbury, D. F., Hirayasu, Y., Bouix, S., Pohl, K. M., Yoshida, T., & McCarley, R. W. (2007). Neocortical gray matter volume in first-episode schizophrenia and first-episode affective psychosis: a cross-sectional and longitudinal MRI study. *Biological Psychiatry*, *62*(7), 773–783. https://doi.org/10.1016/j.biopsych.2007.03.030

Nambiar, M., & Raghavan, S. C. (2011). How does DNA break during chromosomal translocations? *Nucleic Acids Research*, *39*(14), 5813–5825.

Navari, S., & Dazzan, P. (2009). Do antipsychotic drugs affect brain structure? A systematic and critical review of MRI findings. *Psychological Medicine*, *39*(11), 1763–1777. https://doi.org/10.1017/S0033291709005315

Nawka, A., Kalisova, L., Raboch, J., Giacco, D., Cihal, L., Onchev, G., Karastergiou, A., Solomon, Z., Fiorillo, A., Del Vecchio, V., Dembinskas, A., Kiejna, A., Nawka, P., Torres-Gonzales, F., Priebe, S., Kjellin, L., & Kallert, T. W. (2013). Gender differences in coerced patients with schizophrenia. *BMC Psychiatry*, *13*(1), 257. https://doi.org/10.1186/1471-244X-13-257

Nelson, H. E. (1982). The National Adult Reading Test (NART): Test Manual. *Windsor, UK: NFER-Nelson*.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., & Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, *42*(1), 30–35. https://doi.org/10.1038/ng.499

Oguro-Ando, A., Rosensweig, C., Herman, E., Nishimura, Y., Werling, D., Bill, B. R., Berg, J. M., Gao, F., Coppola, G., Abrahams, B. S., & Geschwind, D. H. (2015). Increased CYFIP1 dosage alters cellular and dendritic morphology and dysregulates mTOR. *Molecular Psychiatry*, *20*(9), 1069–1078. https://doi.org/10.1038/mp.2014.124

Ongür, D., Prescot, A. P., Jensen, J. E., Rouse, E. D., Cohen, B. M., Renshaw, P. F., & Olson, D. P. (2010). T2 relaxation time abnormalities in bipolar disorder and schizophrenia. *Magnetic Resonance in Medicine*, *63*(1), 1–8.

Organization, W. H. (2019). *Global Health Estimates 2019: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2019*.

Owen, M. J. (2012). Implications of Genetic Findings for Understanding Schizophrenia. *Schizophrenia Bulletin*, *38*(5), 904–907. https://doi.org/10.1093/schbul/sbs103

Owen, M. J., Sawa, A., & Mortensen, P. B. (2016). Schizophrenia. *Lancet*, *388*(10039), 86–97.

Owen, M. J., & Williams, N. M. (2017). Genomic complexity of schizophrenia: from 108 loci to a single nuclear membrane protein. *Genome Medicine*, *9*(1), 1–3.

Packer, J. S., Maxwell, E. K., O'Dushlaine, C., Lopez, A. E., Dewey, F. E., Chernomorsky, R., Baras, A., Overton, J. D., Habegger, L., & Reid, J. G. (2016). CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv547

Palmer, E. E., Hong, S., Al Zahrani, F., Hashem, M. O., Aleisa, F. A., Ahmed, H. M. J., Kandula, T., Macintosh, R., Minoche, A. E., Puttick, C., Gayevskiy, V., Drew, A. P., Cowley, M. J., Dinger, M., Rosenfeld, J. A., Xiao, R., Cho, M. T., Yakubu, S. F., Henderson, L. B., … Arold, S. T. (2019). De Novo Variants Disrupting the HX Repeat Motif of ATN1 Cause a Recognizable Non-Progressive Neurocognitive Syndrome. *The American Journal of Human Genetics*, *104*(3), 542–552. https://doi.org/10.1016/j.ajhg.2019.01.013

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., & Carrera, N. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet*, *50*(3), 381–389.

Park, C., Falls, W., Finger, J. H., Longo-Guess, C. M., & Ackerman, S. L. (2002). Deletion in Catna2, encoding αN-catenin, causes cerebellar and hippocampal lamination defects and impaired startle modulation. *Nature Genetics*, *31*(3), 279–284. https://doi.org/10.1038/ng908

Peters, S. K., Dunlop, K., & Downar, J. (2016). Cortico-striatal-thalamic loop circuits of the salience network: A central pathway in psychiatric disease and treatment.

Frontiers in Systems Neuroscience, *10*, 104. https://doi.org/10.3389/fnsys.2016.00104

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R., & Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, *28*(21), 2747–2754. https://doi.org/10.1093/bioinformatics/bts526

Power, R. A., Kyaga, S., Uher, R., MacCabe, J. H., Långström, N., Landen, M., McGuffin, P., Lewis, C. M., Lichtenstein, P., & Svensson, A. C. (2013). Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *Archives of General Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2013.268

Pucilowska, J. (2015). 16p11.2 deletion syndrome: a role for a network-based neurodevelopmental synaptopathy. *Molecular Psychiatry*, *20*(10), 1161–1167.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. https://doi.org/10.1086/519795

Ramalingam, A. (2011). 16p13.11 duplication is a risk factor for a wide spectrum of neuropsychiatric disorders. *Journal of Human Genetics*, *56*(7), 541–544.

Rapoport, J. L., Addington, A. M., Frangou, S., & Psych, M. R. C. (2005). The neurodevelopmental model of schizophrenia: update 2005. *Molecular Psychiatry*, *10*(5), 434–449. https://doi.org/10.1038/sj.mp.4001642

Rapoport, J. L., Giedd, J. N., & Gogtay, N. (2012). Neurodevelopmental model of schizophrenia: update 2012. *Molecular Psychiatry*, *17*(12), 1228–1238. https://doi.org/10.1038/mp.2012.23

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

Rees, E. (2016a). Analysis of copy number variations at 15 schizophrenia-associated loci. *British Journal of Psychiatry*, *208*(6), 545–552.

Rees, E. (2016b). Analysis of copy number variations at 15 schizophrenia-associated loci. *British Journal of Psychiatry*, *208*(6), 545–552.

Rees, E., Creeth, H. D. J., Hwu, H.-G., Chen, W. J., Tsuang, M., Glatt, S. J., Rey, R., Kirov, G., Walters, J. T. R., Holmans, P., Owen, M. J., & O'Donovan, M. C. (2021). Schizophrenia, autism spectrum disorders and developmental disorders share specific disruptive coding mutations. *Nature Communications*, *12*(1), 5353. https://doi.org/10.1038/s41467-021-25532-4

Rees, E., Han, J., Morgan, J., Carrera, N., Escott-Price, V., Pocklington, A. J., Duffield, M., Hall, L. S., Legge, S. E., Pardiñas, A. F., Richards, A. L., Roth, J., Lezheiko, T., Kondratyev, N., Kaleda, V., Golimbet, V., Parellada, M., González-Peñas, J., Arango, C., … Owen, M. J. (2020). De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia. *Nature Neuroscience*, *23*(2), 179–184. https://doi.org/10.1038/s41593-019-0565-2

Rees, E., & Kirov, G. (2021). Copy number variation and neuropsychiatric illness. *Current Opinion in Genetics & Development*, *68*, 57–63. https://doi.org/10.1016/j.gde.2021.02.014

Rees, E., Kirov, G., O'Donovan, M. C., & Owen, M. J. (2012). De Novo Mutation in Schizophrenia. *Schizophrenia Bulletin*, *38*(3), 377–381. https://doi.org/10.1093/schbul/sbs047

Rees, E., Kirov, G., Sanders, A., Walters, J. T. R., Chambert, K. D., Shi, J., Szatkiewicz, J., O'Dushlaine, C., Richards, A. L., Green, E. K., Jones, I., Davies, G., Legge, S. E., Moran, J. L., Pato, C., Pato, M., Genovese, G., Levinson, D., Duan, J., … Owen, M. J. (2014). Evidence that duplications of 22q11.2 protect against schizophrenia. *Molecular Psychiatry*, *19*(1), 37–40. https://doi.org/10.1038/mp.2013.156

Rees, E., Moskvina, V., Owen, M. J., O'Donovan, M. C., & Kirov, G. (2011). De Novo Rates and Selection of Schizophrenia-Associated Copy Number Variants. *Biological Psychiatry*, *70*(12), 1109–1114. https://doi.org/10.1016/j.biopsych.2011.07.011

Rees, E., Walters, J. T. R., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L., Mahoney-Davies, G., Legge, S. E., Moran, J. L., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., & Kirov, G. (2014). Analysis of copy number variations at 15 schizophrenia-associated loci. *British Journal of Psychiatry*. https://doi.org/10.1192/bjp.bp.113.131052

Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *JAMA*, *264*(19), 2511–2518.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., & Akterin, S. (2014). & Schizophrenia Working Group of the Psychiatric Genomics Consortium. *Nature Genetics*, *45*(10), 1150–1159. https://doi.org/10.1038/ng.2742

Rock, P. L., Roiser, J. P., Riedel, W. J., & Blackwell, A. D. (2014). Cognitive impairment in depression: a systematic review and meta-analysis. *Psychological Medicine*, *44*(10), 2029–2040.

Ross, C. A. (2002). Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders. *Neuron*, *35*(5), 819–822.

Rund, B. R. (1998). A review of longitudinal studies of cognitive functions in schizophrenia patients. *Schizophrenia Bulletin*. https://doi.org/10.1093/oxfordjournals.schbul.a033337

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

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Walters, R. K. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Seaby, E. G., Pengelly, R. J., & Ennis, S. (2016). Exome sequencing explained: a practical guide to its clinical application. *Briefings in Functional Genomics*, *15*(5), 374–384. https://doi.org/10.1093/bfgp/elv054

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183. https://doi.org/10.1038/nature16549

Sharp, A. J. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, *40*(3), 322–328.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., & Eichler, E. E. (2006). Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*, *79*(1), 78–88.

Shinawi, M. (2010). Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *Journal of Medical Genetics*, *47*(5), 332–341.

Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., Pietiläinen, O., Gerety, S. S., Ayub, M., Blyth, M., Cole, T., Collier, D., Coomber, E. L., Craddock, N., Daly, M. J., … Barrett, J. C. (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nature Neuroscience*, *19*(4), 571–577. https://doi.org/10.1038/nn.4267

Singh, T., Neale, B. M., & Daly, M. J. (2020). Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia on behalf of the Schizophrenia Exome Meta-Analysis (SCHEMA) Consortium*. *MedRxiv*.

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., Buckley, P. F., Bunney, W. E., Bybjerg-Grauholm, J., Byerley, W. F., Chapman, S. B., Chen, W. J., Churchhouse, C., Craddock, N., Cusick, C. M., … Daly, M. J. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, *604*(7906), 509–516. https://doi.org/10.1038/s41586-022-04556-w

Smeland, O. B., Frei, O., Kauppi, K., Hill, W. D., Li, W., Wang, Y., Krull, F., Bettella, F., Eriksen, J. A., Witoelar, A., Davies, G., Fan, C. C., Thompson, W. K., Lam, M., Lencz, T., Chen, C. H., Ueland, T., Jönsson, E. G., Djurovic, S., … Andreassen, O. A. (2017). Identification of genetic loci jointly influencing schizophrenia risk and the cognitive traits of verbal-numerical reasoning, reaction time, and general cognitive function. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2017.1986

Snyder, M. A., & Gao, W. J. (2013). NMDA hypofunction as a convergence point for progression and symptoms of schizophrenia. *Frontiers in Cellular Neuroscience*, *7*, 31.

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*, *27*(1), 281–295. https://doi.org/10.1038/s41380-021-01161-7

Sommer, I. E., van Westrhenen, R., Begemann, M. J. H., de Witte, L. D., Leucht, S., & Kahn, R. S. (2014). Efficacy of Anti-inflammatory Agents to Improve Symptoms in Patients With Schizophrenia: An Update. *Schizophrenia Bulletin*, *40*(1), 181–191. https://doi.org/10.1093/schbul/sbt139

St Clair, D. (2005). Rates of Adult Schizophrenia Following Prenatal Exposure to the Chinese Famine of 1959-1961. *JAMA*, *294*(5), 557. https://doi.org/10.1001/jama.294.5.557

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*, 437–455.

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G. B., Jonsdottir, G. A., Doyle, O. M., Tost, H., Grimm, O., Kristjansdottir, S., Snorrason, H., Davidsdottir, S. R., Gudmundsson, L. J., Jonsson, G. F., Stefansdottir, B., Helgadottir, I., … Stefansson, K. (2014a). CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, *505*(7483), 361–366. https://doi.org/10.1038/nature12818

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G. B., Jonsdottir, G. A., Doyle, O. M., Tost, H., Grimm, O., Kristjansdottir, S., Snorrason, H., Davidsdottir, S. R., Gudmundsson, L. J., Jonsson, G. F., Stefansdottir, B., Helgadottir, I., … Stefansson, K. (2014b). CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, *505*(7483), 361–366. https://doi.org/10.1038/nature12818

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J. E., Hansen, T., Jakobsen, K. D., Muglia, P., Francks, C., Matthews, P. M., Gylfason, A., Halldorsson, B. V., Gudbjartsson, D., Thorgeirsson, T. E., … Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*(7210), 232–236. https://doi.org/10.1038/nature07229

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D., & Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, *133*(1), 1–9. https://doi.org/10.1007/s00439-013-1358-4

Stone, J. M. (2011). Imaging the glutamate system in humans: relevance to drug discovery for schizophrenia. *Current Pharmaceutical Design*, *17*(2), 64–73.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426.

Stromme, P., Bjornstad, P. G., & Ramstad, K. (2002). Prevalence estimation of Williams syndrome. *Journal of Child Neurology*, *17*(4), 269–271.

Südhof, T. C. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. *Nature*, *455*(7215), 903–911. https://doi.org/10.1038/nature07456

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, *13*(8), 537–551.

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

Suzuki, T., Remington, G., Mulsant, B. H., Uchida, H., Rajji, T. K., Graff-Guerrero, A., & Pollock, B. G. (2015). Treatment resistant schizophrenia and response to antipsychotics: A review. *Schizophrenia Research*, *133*(1–3), 54–62.

Szafranski, P. (2010). *Structures and molecular mechanisms for common 15q13.3 microduplications involving CHRNA7*.

Szöke, A., Trandafir, A., Dupont, M.-E., Méary, A., Schürhoff, F., & Leboyer, M. (2008). Longitudinal studies of cognition in schizophrenia: Meta-analysis. *British Journal of Psychiatry*, *192*(4), 248–257. https://doi.org/10.1192/bjp.bp.106.029009

Takahashi, T., Suzuki, M., Zhou, S. Y., Tanino, R., Nakamura, K., Kawasaki, Y., & Seto, H. (2006). A follow-up MRI study of the superior temporal subregions in schizotypal disorder and first-episode schizophrenia. *Schizophrenia Research*, *83*(2–3), 131–141. https://doi.org/10.1016/j.schres.2006.01.009

Tandon, R., Nasrallah, H. A., & Keshavan, M. S. (2009a). Schizophrenia, "just the facts" 4. Clinical features and conceptualization. *Schizophrenia Research*, *110*(1–3), 1–23.

Tandon, R., Nasrallah, H. A., & Keshavan, M. S. (2009b). Schizophrenia, "just the facts" 4. Clinical features and conceptualization. *Schizophrenia Research*, *110*(1–3), 1–23.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, *28*(21), 2711–2718. https://doi.org/10.1093/bioinformatics/bts535

Thara, R., & Kamath, S. (2015). Women and schizophrenia. *Indian Journal of Psychiatry*, *57*(6), 246. https://doi.org/10.4103/0019-5545.161487

Thornicroft, G., Brohan, E., Rose, D., Sartorius, N., & Leese, M. (2009). Global pattern of experienced and anticipated discrimination against people with schizophrenia: A cross-sectional survey. *The Lancet*, *373*(9661), 408–415.

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Thygesen, J. H., Presman, A., Harju-Seppänen, J., Irizar, H., Jones, R., Kuchenbaecker, K., Lin, K., Alizadeh, B. Z., Austin-Zimmerman, I., Bartels-Velthuis, A., Bhat, A., Bruggeman, R., Cahn, W., Calafato, S., Crespo-Facorro, B., de Haan, L., de Zwarte, S. M. C., Di Forti, M., Díez-Revuelta, Á., … Bramon, E. (2021). Genetic copy number variants, cognition and psychosis: a meta-analysis and a family study. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-020-0820-7

Tienari, P., Wynne, L. C., Sorri, A., Lahti, I., Läksy, K., Moring, J., Naarala, M., Nieminen, P., & Wahlberg, K.-E. (2004). Genotype–environment interaction in schizophrenia-spectrum disorder. *British Journal of Psychiatry*, *184*(3), 216–222. https://doi.org/10.1192/bjp.184.3.216

Tripathi, A., Kar, S. K., & Shukla, R. (2018). Cognitive Deficits in Schizophrenia: Understanding the Biological Correlates and Remediation Strategies. *Clinical Psychopharmacology and Neuroscience*, *16*(1), 7–17. https://doi.org/10.9758/cpn.2018.16.1.7

Trotta, A., Murray, R. M., & Fisher, H. L. (2015). The impact of childhood adversity on the persistence of psychotic symptoms: a systematic review and meta-analysis. *Psychological Medicine*, *45*(12), 2481–2498.

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., … van Os, J. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508. https://doi.org/10.1038/s41586-022-04434-5

Ullmann, R., Turner, G., Kirchhoff, M., Chen, W., Tonge, B., Rosenberg, C., Field, M., Vianna-Morgante, A. M., Christie, L., Krepischi-Santos, A. C., Banna, L., Brereton, A. V., Hill, A., Bisgaard, A.-M., Müller, I., Hultschig, C., Erdogan, F., Wieczorek, G., & Ropers, H. H. (2007). Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Human Mutation*, *28*(7), 674–682. https://doi.org/10.1002/humu.20546

Usman, D. M., Olubunmi, O. A., Taiwo, O., Taiwo, A., Rahman, L., & Oladipo, A. (2011). Comparison of catatonia presentation in patients with schizophrenia and mood disorders in lagos, Nigeria. *Iranian Journal of Psychiatry*, *6*(1), 7–11.

Van, L., Heung, T., Graffi, J., Ng, E., Malecki, S., Van Mil, S., Boot, E., Corral, M., Chow, E. W. C., Hodgkinson, K. A., Silversides, C., & Bassett, A. S. (2019). All-cause mortality and survival in adults with 22q11.2 deletion syndrome. *Genetics in Medicine*, *21*(10), 2328–2335. https://doi.org/10.1038/s41436-019-0509-y

van Scheltinga, A. F. T., Bakker, S. C., van Haren, N. E. M., Derks, E. M., Buizer-Voskamp, J. E., Cahn, W., Ripke, S., Ophoff, R. A., & Kahn, R. S. (2013). Schizophrenia genetic variants are not associated with intelligence. *Psychological Medicine*, *43*(12), 2563–2570. https://doi.org/10.1017/S0033291713000196

Varese, F., Smeets, F., Drukker, M., Lieverse, R., Lataster, T., Viechtbauer, W., & Bentall, R. P. (2012). Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective-and cross-sectional cohort studies. *Schizophrenia Bulletin*, *38*(4), 661–671.

Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-Analysis of the Association of Urbanicity With Schizophrenia. *Schizophrenia Bulletin*, *38*(6), 1118–1123. https://doi.org/10.1093/schbul/sbs096

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, *13*(8), 565–575. https://doi.org/10.1038/nrg3241

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, *101*(1), 5–22.

Vita, A., Peri, L., & Silenzi, C. (2012). Brain morphology in first-episode schizophrenia: a meta-analysis of quantitative magnetic resonance imaging studies. *Schizophrenia Research*, *138*(2–3), 99–104. https://doi.org/10.1016/j.schres.2012.03.006

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S. M., Rippey, C. F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R. L., Sikich, L., Stromberg, T., Merriman, B., … Sebat, J. (2008). Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science*, *320*(5875), 539–543. https://doi.org/10.1126/science.1155174

Wang, K., & Bucan, M. (2008). Copy Number Variation Detection via High-Density SNP Genotyping. *Cold Spring Harbor Protocols*, *2008*(6), pdb.top46. https://doi.org/10.1101/pdb.top46

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., & Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*. https://doi.org/10.1101/gr.6861907

Wang, Q. S., & Huang, H. (2022). Methods for statistical fine-mapping and their applications to auto-immune diseases. *Seminars in Immunopathology*, *44*(1), 101–113. https://doi.org/10.1007/s00281-021-00902-8

Weckselblatt, B., & Rudd, M. K. (2015). Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics*, *31*(10), 587–599. https://doi.org/10.1016/j.tig.2015.05.010

Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., Boeke, J. D., & Moran, J. V. (2001). Human L1 retrotransposition: cis-preference versus trans-complementation. *Molecular and Cellular Biology*, *21*(4), 1429–1439.

Williams, N. M., Zaharieva, I., Martin, A., Langley, K., Mantripragada, K., Fossdal, R., Stefansson, H., Stefansson, K., Magnusson, P., Gudmundsson, O. O., Gustafsson, O., Holmans, P., Owen, M. J., O’Donovan, M., & Thapar, A. (2010). Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *The Lancet*, *376*(9750), 1401–1408. https://doi.org/10.1016/S0140-6736(10)61109-9

Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D., & Sartorius, N. (1990). Schedules for clinical assessment in neuropsychiatry. In *Archives of General Psychiatry*. https://doi.org/10.1136/bmj.c7160

Woodberry, K. A., Giuliano, A. J., & Seidman, L. J. (2008). Premorbid IQ in Schizophrenia: A Meta-Analytic Review. *American Journal of Psychiatry*, *165*(5), 579–587. https://doi.org/10.1176/appi.ajp.2008.07081242

World Health Organization. (2019). *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)*. World Health Organisation.

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., Van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzetinova, T., Bevan, A. P., Bragin, E., Chatzimichali, E. A., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Miller, R., Morley, K. I., … Firth, H. V. (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. *The Lancet*. https://doi.org/10.1016/S0140-6736(14)61705-0

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J. A., & Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*, *44*(12), 1365–1369. https://doi.org/10.1038/ng.2446

Ye, T., Huang, L., Li, Q., Huang, S., Zhu, S., Chen, S., & Peng, W. (2021). HIRA deficiency leads to DiGeorge syndrome-like phenotypes in zebrafish. *Human Molecular Genetics*, *30*(3), 214–228.

Zhang, F., Xu, D., Yuan, L., Sun, Y., & Xu, Z. (2014). Epigenetic regulation of Atrophin1 by lysine-specific demethylase 1 is required for cortical progenitor maintenance. *Nature Communications*, *5*(1), 5815. https://doi.org/10.1038/ncomms6815

Zhang, W., Deng, W., Yao, L., Xiao, Y., Li, F., Liu, J., Sweeney, J. A., Lui, S., & Gong, Q. (2015). Brain Structural Abnormalities in a Group of Never-Medicated Patients With Long-Term Schizophrenia. *American Journal of Psychiatry*, *172*(10), 995–1003. https://doi.org/10.1176/appi.ajp.2015.14091108

Zhang, Y., Bertolino, A., Fazio, L., Blasi, G., Rampino, A., Romano, R., & Sadee, W. (2007). Polymorphisms in human dopamine D2 receptor gene affect gene expression, splicing, and neuronal activity during working memory. *Proceedings of the National Academy of Sciences*, *104*(51), 20552–20557. https://doi.org/10.1073/pnas.0707106104

Zhang, Z., Carriero, N., & Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in Genetics*, *20*(2), 62–67.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, *14*(S11), S1. https://doi.org/10.1186/1471-2105-14-S11-S1

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487.

Zoghbi, A. W., Dhindsa, R. S., Goldberg, T. E., Mehralizade, A., Motelow, J. E., Wang, X., Alkelai, A., Harms, M. B., Lieberman, J. A., Markx, S., & Goldstein, D. B. (2021). High-impact rare genetic variants in severe schizophrenia. *Proceedings of the National Academy of Sciences*, *118*(51). https://doi.org/10.1073/pnas.2112560118

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*(1), 160025. https://doi.org/10.1038/sdata.2016.25