Chapter 5

Discussion and future directions

5.1 Summary of findings

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

To better understand if the findings relating to *SETD1A* can be extended and generalized to a larger number of rare schizophrenia risk variants, I performed a series of analyses that explored the potential overlap of genetic risk between schizophrenia and broader developmental disorders. I jointly analysed the trio and case-control exome data set with array-based CNV calls from 6,882 cases and 11,255 controls, and found that individuals with schizophrenia carried a significantly higher burden of rare damaging variants in 3,488 genes with a near-complete depletion of truncating variants across all variant types. This concentration of risk alleles in constrained genes was previously observed in autism, intellectual disability, and severe developmental disorders. I then performed rare variant enrichment analyses in 1,766 gene sets, and found that the rare variant burden was most strongly enriched in known autism risk genes, and genes diagnostic of severe developmental disorders. This result was significant even after controlling for the baseline enrichment in genes depleted of truncating variants. Finally, in a subset of schizophrenia patients with intellectual disability, I showed that this burden is even stronger than in the general schizophrenia population, mirroring previous results comparing autism individuals with and without cognitive impairment. Combined, these results demonstrate that schizophrenia risk loci of large effect across a range of variant types implicate a common set of genes shared with broader neurodevelopmental disorders, suggesting a path forward in identifying additional risk genes in psychiatric disorders and further supporting a neurodevelopmental etiology to the pathogenesis of schizophrenia.

5.2 Limitations of results described in this Thesis

5.2.1 Limitations in the interpretation of protein-coding consequences

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

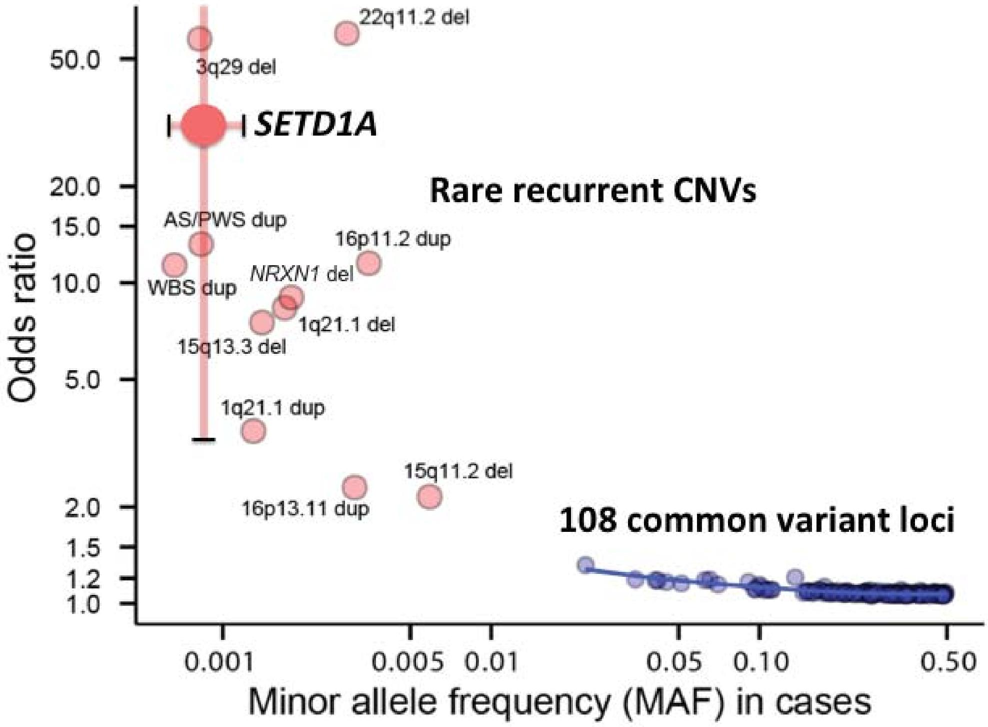

Fig. 5.1 **Risk variants for schizophrenia, with *SETD1A* included.** The effect size of each genome-wide significant risk variant for schizophrenia, as described in Ripke *et al.* and Rees *et al.*, were plotted against its allele frequency in cases [57, 67].

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

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

5.2.2 Insufficient standardisation of clinical data

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

On a similar thread, I identified an enrichment of rare damaging variants in developmental disorder and autism genes, but was unable to acquire the appropriate phenotypic information to determine if carriers of these LoF variants represented a distinct population of patients. Schizophrenia carriers of LoF variants in *CHD8*, an autism risk gene, could potentially have autistic features in addition to psychosis [105]. Furthermore, the disruption of specific developmental disorder genes, such as *KMT2A* and *KMT2D*, causes characteristic facial and physical dysmorphology [157, 118], and our data did not allow us to determine if this were true of the carriers in the schizophrenia data set. Ultimately, the lack of high-quality and standardised clinical data accompanying large-scale genetic data is a severe limitation in our current study, and of association studies of psychiatric traits moving forward. Comprehensive phenotyping would be required to investigate whether carriers of rare LoF variants in constrained genes represented a distinct population of patients when compared to the remaining cases, or if these variants were associated with patterns in age-of-onset, pre-morbid impairment, neurological co-morbidities, relapse, and severity.

5.2.3 Limitations in the definition of the constrained gene list

The enrichment of rare risk variants in constrained genes is among the most striking results in early whole-exome sequencing studies of psychiatric and neurodevelopmental disorders. However, the definition of genic constraint, or the probability of loss-of-function intolerant (pLI), has caveats that need considered when interpreting the significance of our results. First, the pLI score was calculated using an expectation-maximization algorithm that assigned genes to one of three categories: null (in which LoF variants is completely tolerated), recessive (in which homozygous LoF variants is not tolerated), and haploinsufficient (in which a single copy loss is not tolerated). Genes above an arbitrary probability threshold of 0.9 were described as loss-of-function intolerant. From this definition, it is clear that the power to assign a gene to one of these categories is highly dependent on gene length; longer genes would have a greater number of expected loss-of-function variants, enabling more robust estimates of LoF depletion. Despite that notable size of the ExAC study, there may not be sufficient observations of rare LoF variants in smaller genes to detect a deviation from expectation, and these genes may have a pLI score less than 0.9 and defined as unconstrained for this reason. For example, *ARX* is a gene in which LoF variants cause severe mental retardation [157, 118], but its pLI score was estimated to be 0.74 because a 4:0 expected-to-observed LoF variant ratio was insufficient for estimating genic constraint. Therefore, a gene's ranking along the distribution of pLI score is highly dependent on statistical power that is a property of the gene sequence, and the metric itself is not a valid proxy for the strength of selection or degree of constraint. Furthermore, as described earlier in this Section, most genes have a number of transcripts that are sometimes regulated in a tissue or time-specific manner with varying functionality, and this is ignored during the modelling of genic constraint. LoF variants across all transcripts of a gene were aggregated during the calculation of pLI, and it is conceivable that LoF variants in certain transcripts are benign while in others, it is severely pathogenic. Over-simplifying the question of annotation could result in the assignment of a gene to the unconstrained category when it is actually haploinsufficient in one isoform. Finally, the 45,376 exomes used to estimate the depletion of LoF variants were aggregated from studies that include cases diagnosed with complex disorders. While exomes of individuals with psychiatric disorders were explicitly excluded, the ExAC study had an increased incidence of autoimmune disorders, metabolic syndrome, Type II diabetes, and cardiovascular disorders [112], and risk genes for these conditions would have biased pLI scores that would be lower than expected for the general population . Given the many limitations of the ExAC constraint metric, the list of constrained genes is incomplete and imperfect, and should be treated as so. It is a rough but relatively effective tool in identifying a set of genes that are likely to be haploinsufficient in the genome. The significant enrichment of rare variants in constrained genes should be interpreted as simply an indication that there exists a large number of genes that carry rare LoF variants that substantially increase risk for psychiatric and neurodevelopmental disorders in the genome. Given this enrichment, the next step is to identify these genes with a scale-up in sample size of whole-exome sequencing studies.

5.2.4 Interpretation and generalisability of gene set results

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

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

5.3 Future directions

5.3.1 Whole-genome sequencing at the population scale

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

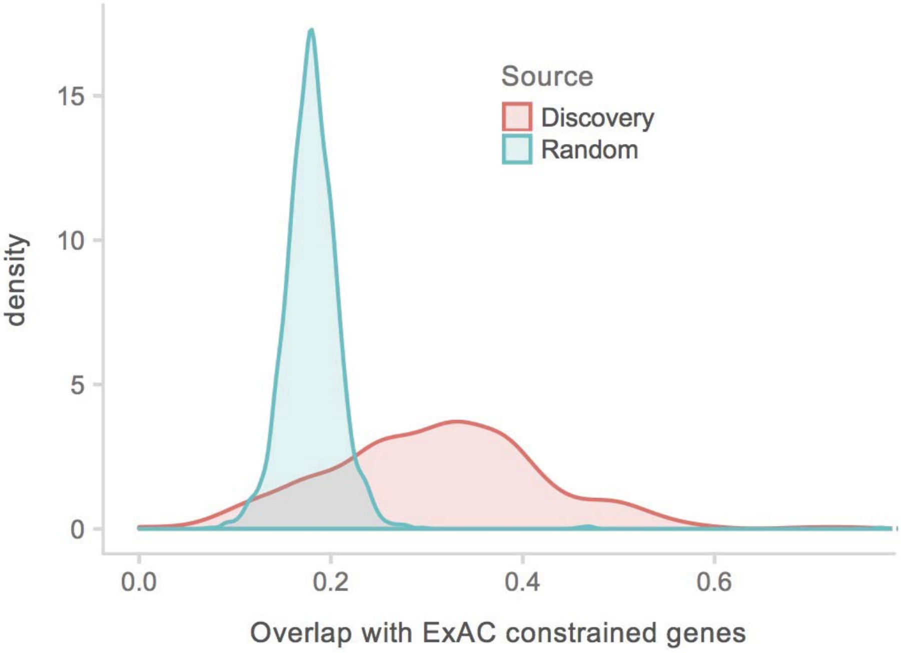

Fig. 5.2 **Distribution of overlap coefficients with the constrained gene set.** The overlap coefficients between each of the 1,766 discovery gene sets described in Chapter 4 and the constrained gene set were calculated. Random gene sets were sampled from the genome with the same size distribution as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

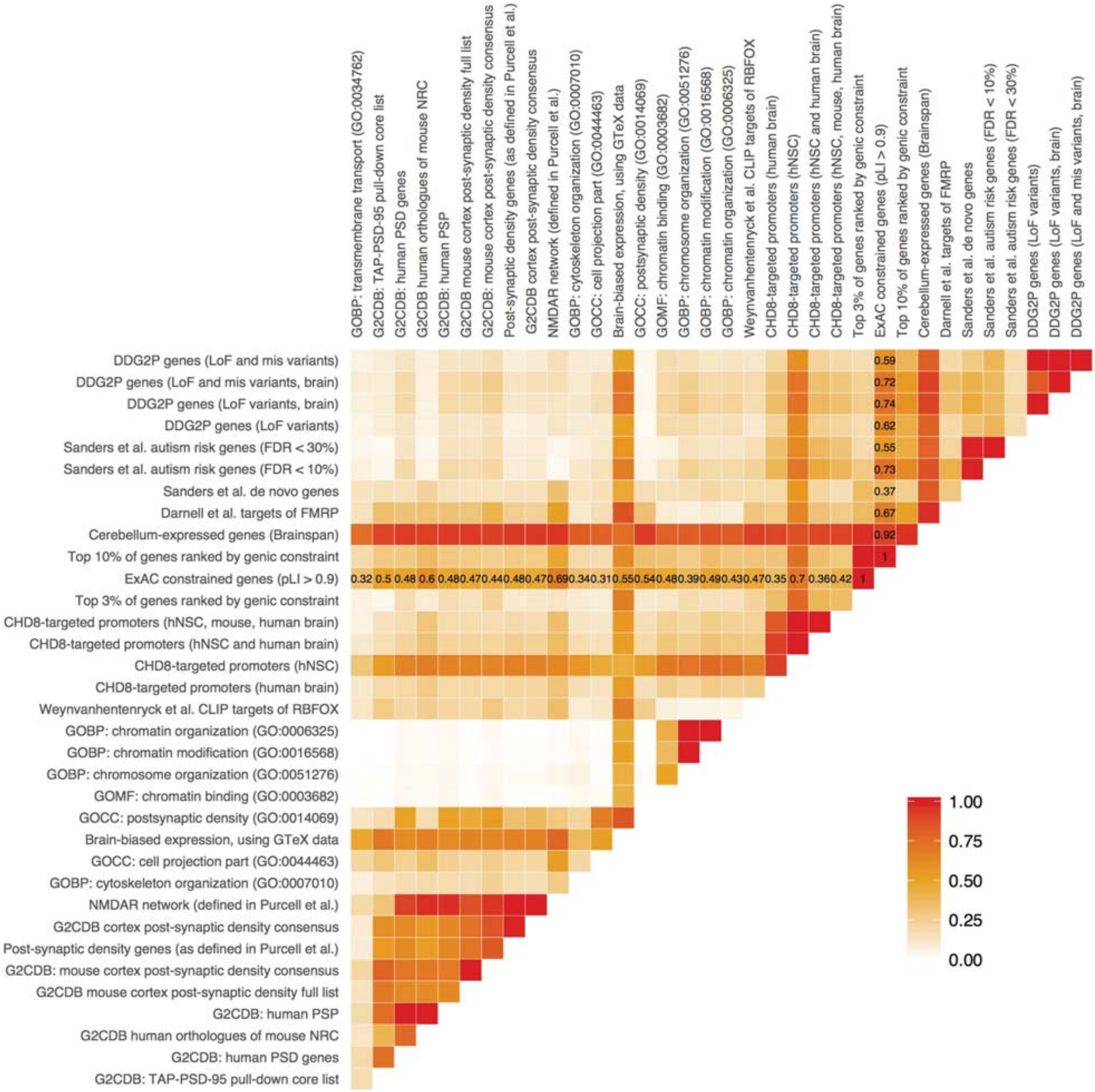

Fig. 5.3 **Heatmap of overlap coefficients calculated between FDR $< 5\%$ gene sets.** The overlap coefficients of gene sets enriched for rare coding variants conferring risk for schizophrenia were computed, clustered, and displayed as a heatmap. The overlap coefficient is a similarity measure defined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

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

5.3.2 Specificity of shared risk alleles for individual psychiatric disorders

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

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

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

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

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

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

We have also engineered two LoF variants in *SETD1A* into human induced pluripotent stem cell (iPSC) lines. We will use a combination of RNA-seq and ChIP-seq to characterize transcriptomic and epigenetic changes in neuronal progenitors. This analysis will provide a complementary set of differentially methylated regions and differentially expressed genes, and results from the mouse model and the iPSC line will be compared and contrasted. We will test these gene sets for enrichment in common schizophrenia risk loci, intellectual disability and autism risk genes. In summary, *in vitro* and *in vivo* models of highly penetrant rare variants have proven useful in studying genes for neurodevelopmental disorders, and will likely be applied to many more risk genes in the future to advance our understanding of the disease mechanisms underlying autism, intellectual disability, and schizophrenia.

5.4 Concluding remarks

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