Chapter 4

Schizophrenia risk genes are shared with neurodevelopmental disorders

4.1 Introduction

4.1.1 Early evidence for a neurodevelopmental etiology to schizophrenia

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

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

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

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

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

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

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

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

4.1.3 Individual loci increasing risk for schizophrenia and neurodevelopmental disorders

However, the evidence from rare variants for a broader shared genetic etiology between schizophrenia and neurodevelopmental disorders is more mixed. An analysis of *de novo* mutations from schizophrenia probands found a nominal overlap with *de novo* LoF variants from probands with intellectual disability ($P = 0.019$, uncorrected), but this result was based on the observation of a single *de novo* event [98]. A whole-exome sequencing study of 2,536 schizophrenia cases and 2,543 controls tested for a burden of rare LoF and nonsynonymous variants in candidate gene sets for autism and intellectual disability, including genes hit by *de novo* mutations in intellectual disability and autism, but did not observe any overlap [103]. Evidence at individual rare schizophrenia risk loci suggested that a partial, perhaps weaker overlap may exist between psychiatric and neurodevelopmental disorders. First,

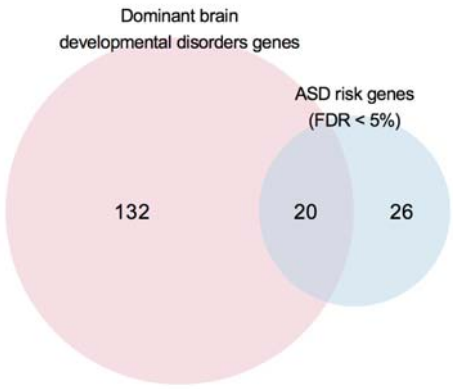

Fig. 4.1 **The overlap between autism risk genes and dominant developmental disorder genes.** This Venn Diagram illustrates the overlap between the autism risk genes implicated by Sanders *et al.* at FDR $< 5\%$ (46 genes) and dominant brain developmental disorder risk genes described in the DDG2P database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

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

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

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

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

4.1.5 Aims and goals

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

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

4.1.6 Publication note and contributions

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

4.2 Methods

4.2.1 Sample collections

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

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

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

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

4.2.2 Rare variant gene set enrichment analyses

Case-control enrichment burden tests

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

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

CNV logistic regression

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

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

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

Weighted permutation-based sampling of *de novo* mutations

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

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

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

4.2.3 Combined joint analysis

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

statistic for each gene set:

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

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

4.2.4 Description of gene sets

Public gene set databases

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

Schizophrenia candidate gene sets

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

Constrained genes

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

Risk genes for autism and neurodevelopmental disorders

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

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

Brain expression gene sets

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

4.2.5 Conditional analyses

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

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

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

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

4.2.6 Rare variants and cognition in schizophrenia

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

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

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

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

4.3 Results

4.3.1 Study design

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

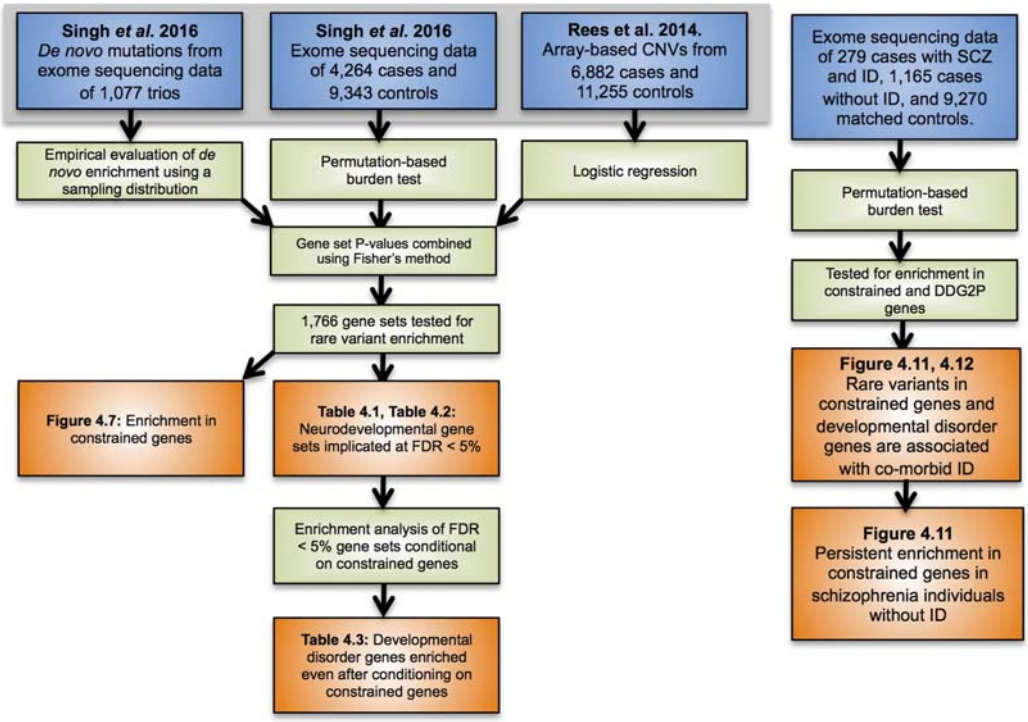

Fig. 4.2 **Analysis workflow.** Data sets are shown in blue, statistical methods and analysis steps are shown in green, and results (figures and tables) from the analysis are shown in orange. The left chart describes analyses testing for enrichment in 1,766 gene sets using the entire data set. The right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

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

4.3.2 Selection of allele frequency thresholds and consequence severity

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

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

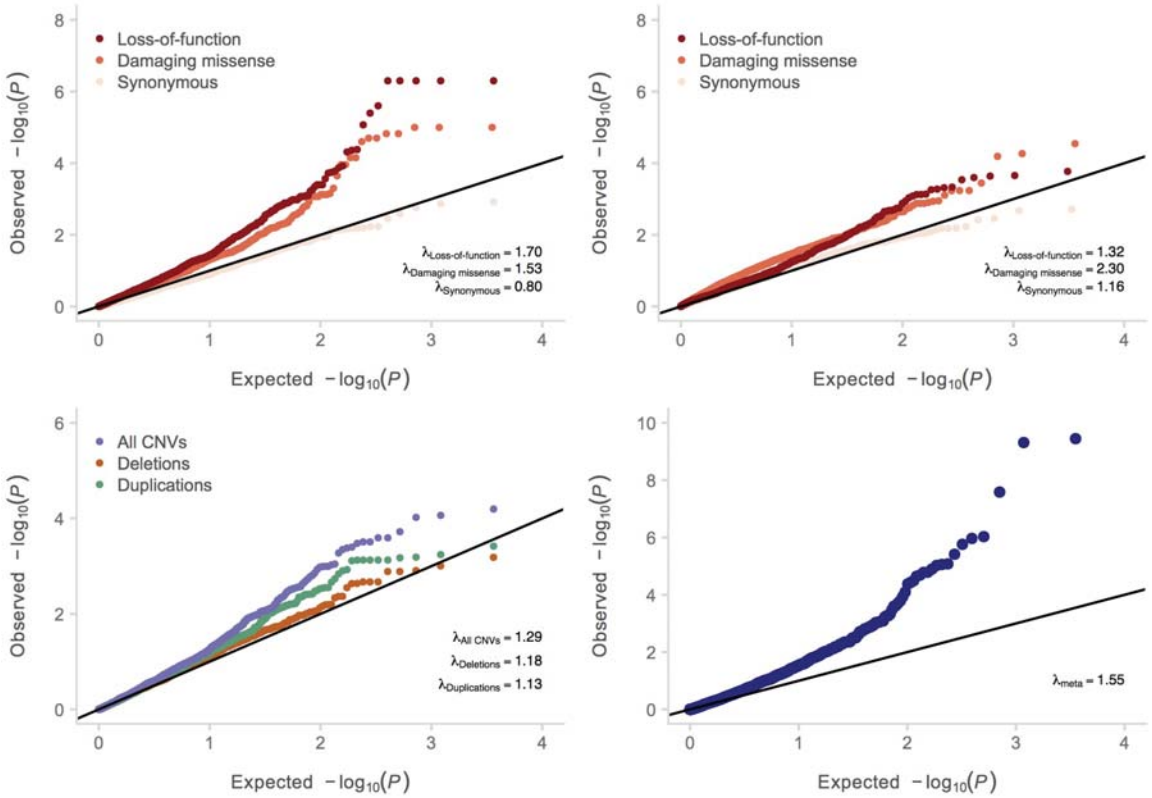

Fig. 4.3 **Q-Q plots of *P*-values from enrichment tests of 1,766 gene sets. Top left:** case-control SNVs from whole-exome sequence data; **Top right:** *de novo* mutations from 1,077 trios; **Bottom left:** case-control CNVs; **Bottom right:** meta-analysed *P*-values from Fisher's method (dark blue).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

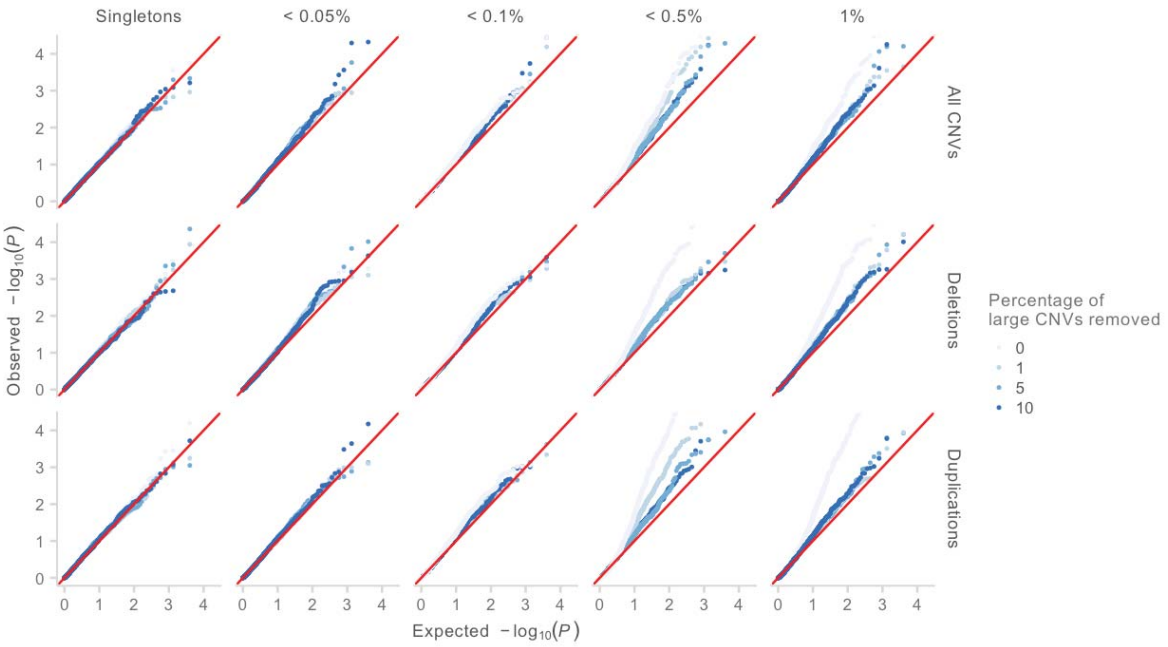

Fig. 4.4 **The use of frequency and size cut-offs in CNV gene sets enrichment tests to reduce genomic inflation.** Q-Q plots were generated based on *P*-values from CNV enrichment tests of random gene sets, using different MAF cut-offs (Singletons, $< 0.05\%$, $< 0.1\%$, 1%) and CNV size cut-offs (removing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

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

4.3.3 Robustness of enrichment analyses

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

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

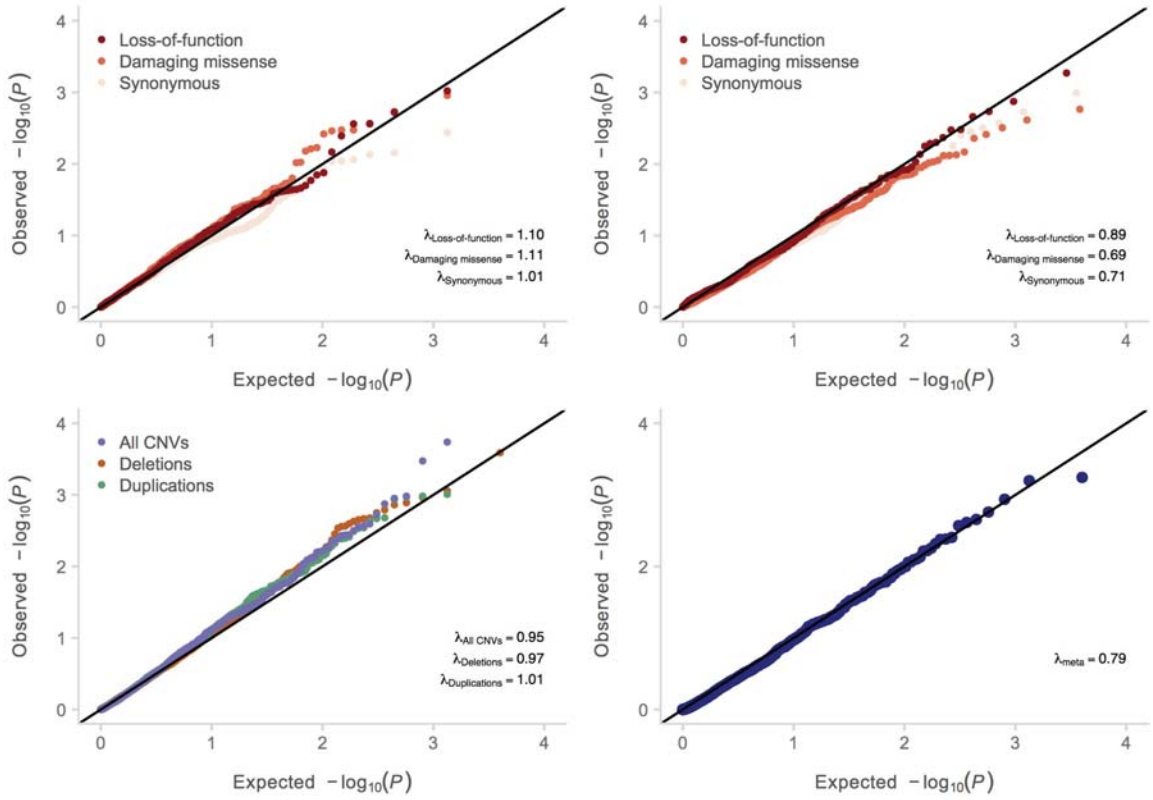

Fig. 4.5 **Q-Q plots of *P*-values from enrichment tests of random gene sets. Top left**: case-control SNVs from whole-exome sequence data; **Top right**: *de novo* mutations from 1,077 trios; **Bottom left**: case-control CNVs. Genes were randomly sampled from the genome to create gene sets with the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

4.3.4 Rare, damaging schizophrenia variants are concentrated in constrained genes

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

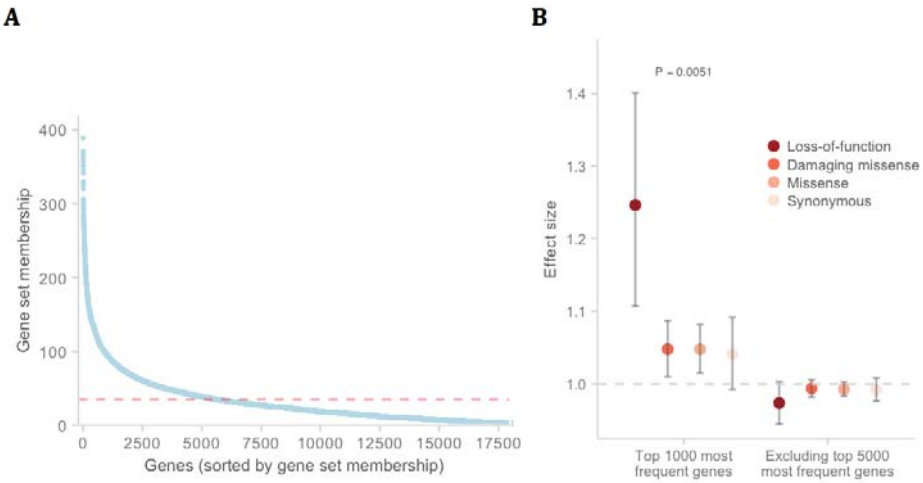

Fig. 4.6 **Non-random sampling of genes in the 1,766 gene sets resulted in non-null enrichment of disruptive variants. A**: Genes were ranked and plotted based on the number of gene sets they belonged in. The top 1000 genes were massively over-represented in gene sets from public databases, and genes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

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

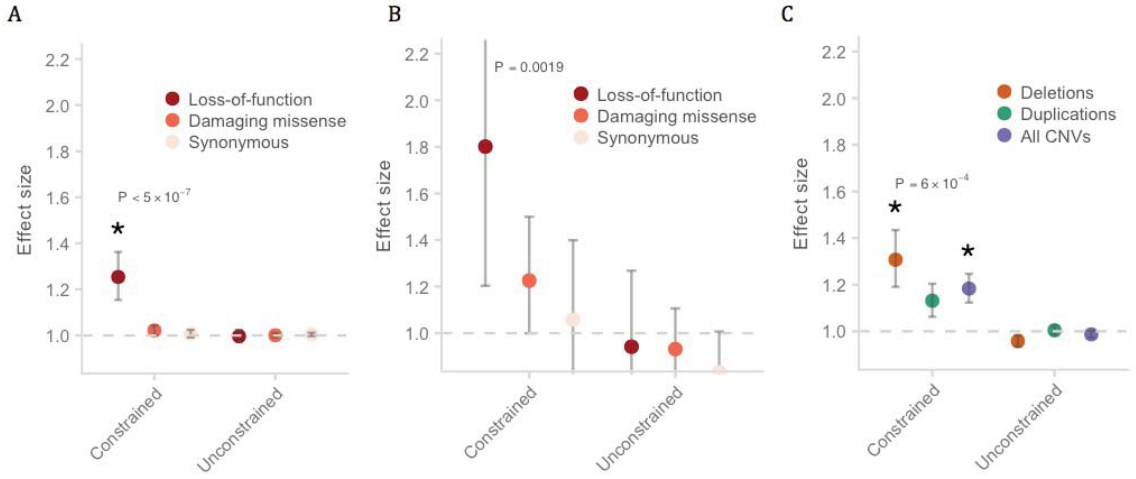

Fig. 4.7 **Enrichment of schizophrenia rare variants in constrained genes. A:** Schizophrenia cases compared to controls for rare SNVs and indels; **B:** Rates of *de novo* mutations in schizophrenia probands compared to control probands; **C:** Case-control CNVs. *P*-values shown were from the test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

4.3.5 Comparing the enrichment in constrained genes across neurodevelopmental disorders

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

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

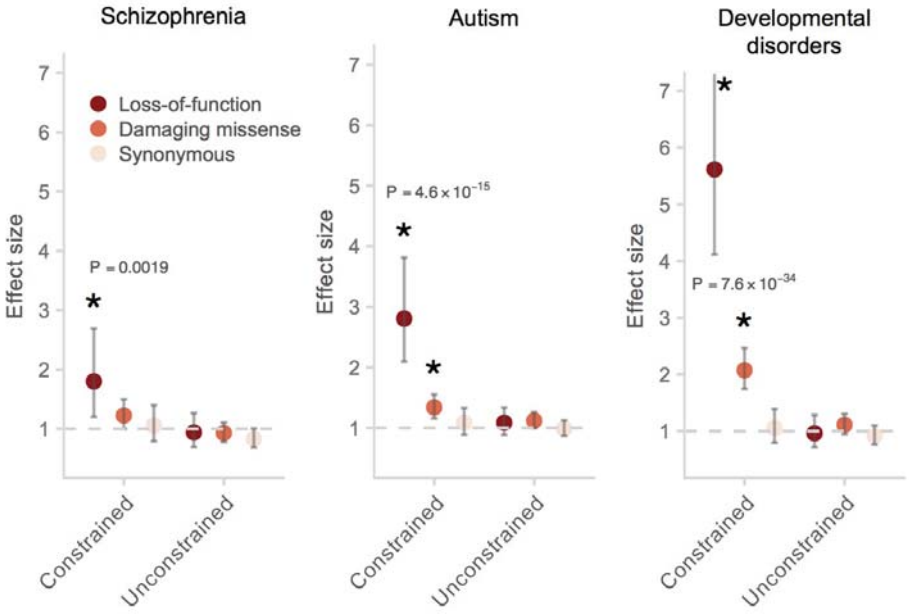

Fig. 4.8 **Enrichment of *de novo* mutations in genes with near-complete depletion of truncating variants across schizophrenia and neurodevelopmental disorders.** In autism, schizophrenia, and severe neurodevelopmental disorders, *de novo* mutations were enriched in a subset of genes under genic constraint,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

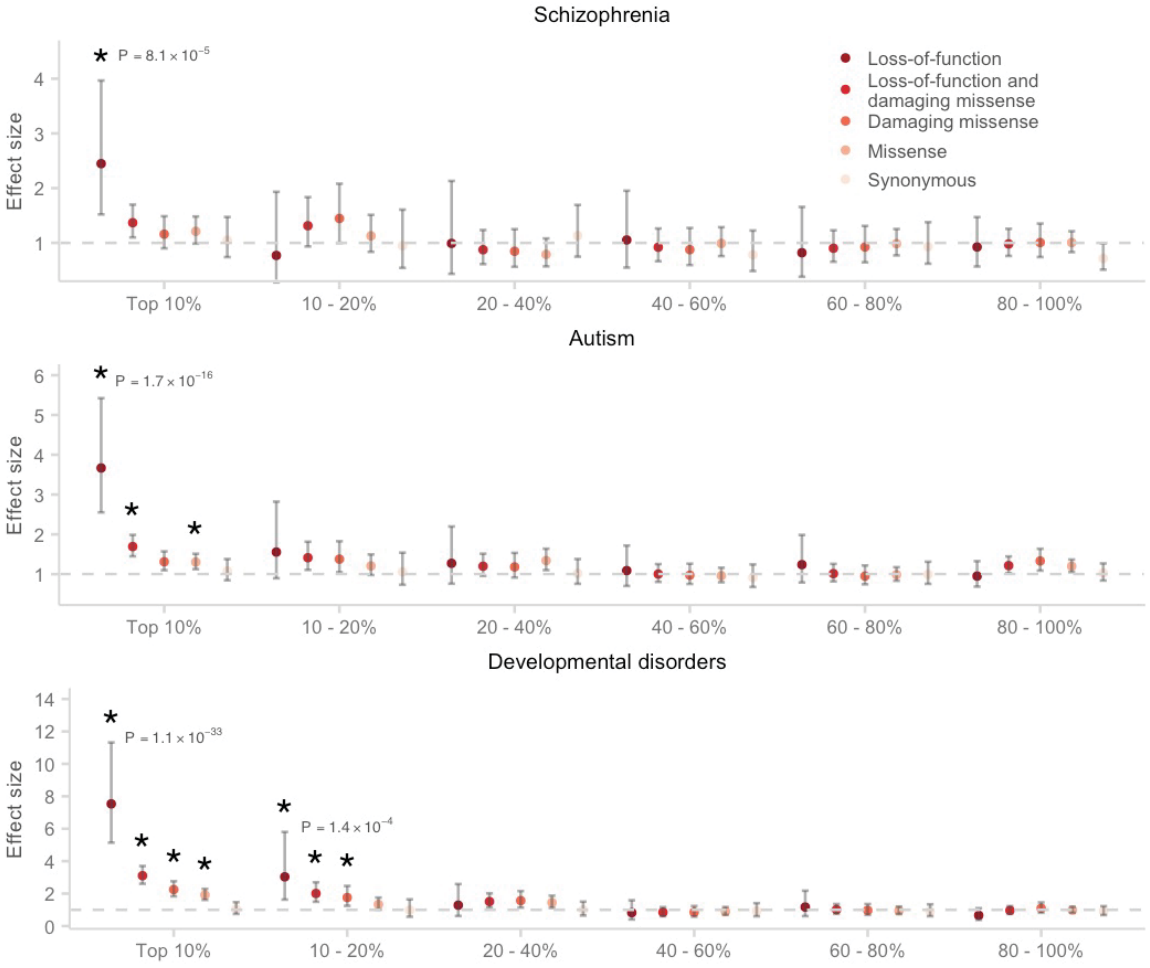

Fig. 4.9 **Enrichment of *de novo* mutations in genes ordered and grouped by genic constraint across schizophrenia and neurodevelopmental disorders.** Genes were ordered by their degree of constraint (pLI score), and grouped into six categories: the 10% most constrained, 10 – 20% most constrained, 20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

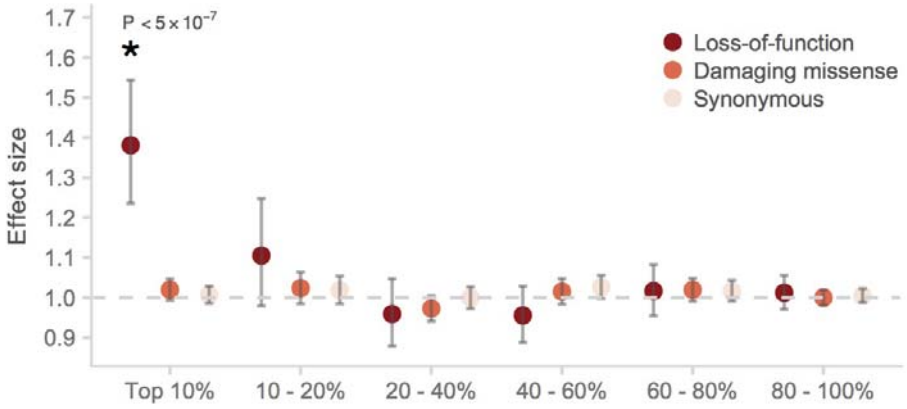

Fig. 4.10 **Enrichment of case-control SNVs in genes ordered and grouped by genic constraint.** Genes were ordered by their degree of constraint (pLI score), and grouped into six categories: the 10% most constrained, $10-20$% most constrained, $20-40$% most constrained, and so on. A significant enrichment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

4.3.6 Schizophrenia risk genes are shared with other neurodevelopmental disorders

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

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

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

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

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

4.3.7 Schizophrenia rare variants are associated with intellectual disability

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

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

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

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

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

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

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

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

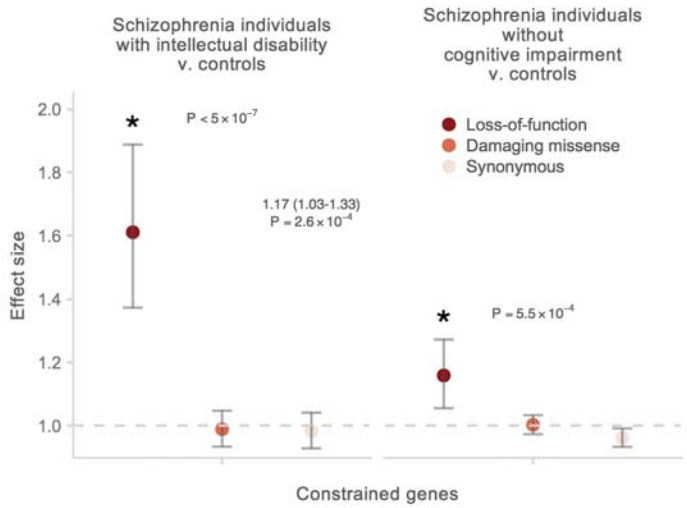

Fig. 4.11 **Enrichment of rare variants in constrained genes between schizophrenia (SCZ) individuals with ID, schizophrenia individuals without ID, and matched controls.** The *P*-values shown were calculated from the burden test of LoF variants between the corresponding cases and matched controls. The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

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

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

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

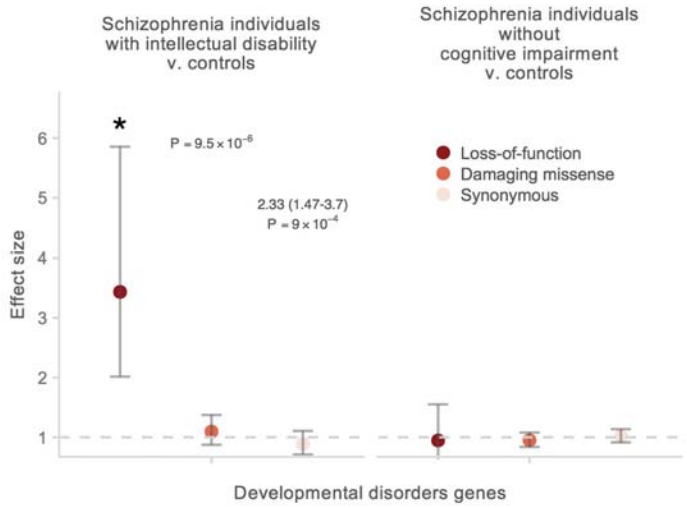

Fig. 4.12 **Enrichment of rare variants in diagnostic developmental disorder genes between schizophrenia (SCZ) individuals with ID, schizophrenia individuals without ID, and matched controls.** The *P*-values shown were calculated from the burden test of LoF variants between the corresponding cases and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

4.4 Discussion

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

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

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

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

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