Chapter 3

SETD1A is associated with schizophrenia and neurodevelopmental disorders

3.1 Introduction

3.1.1 Motivation behind rare variant analyses in psychiatric disorders

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

Despite the successes of common variant analyses, studies investigating rare coding variation (minor allele frequency $< 0.1\%$) provide an unique opportunity to extend our understanding of the genetic architecture of psychiatric disorders. First, alleles that confer substantial risk for human disease are expected to reside in the rare end of the allele frequency spectrum. These variants are subject to strong negative selection, and thus, are depleted in the general population. In addition, because coding alleles cause changes at the mRNA and protein level, they are easier to fine-map than common intergenic variants, and are more likely to cause obvious cellular changes in human carriers. Both these properties increase the success and interpretability of subsequent functional studies, which are critical for elucidating the biological mechanisms underlying human disorders. Furthermore, while ultra-rare variants explain only a modest fraction of the broad-sense heritability of complex disorders, they contribute substantially to individual liability, and are immediately useful in clinical practice for identifying patients with higher risk for disease [141]. At the moment, genetic counselling and genetic testing are limited to fully penetrant alleles for Mendelian traits (e.g. *HTT* repeat length for Huntington's disease or *HBB* allele for sickle cell anaemia) or rare variants of large effect (e.g. *BRCA1/2* alleles for breast cancer or *APOE* alleles for cardiovascular disease), and many more of these clinically relevant variants remain to be discovered. Fortunately, the decreasing costs of whole-exome sequencing has enabled the identification of very rare, often private, protein-coding variants in sufficiently large populations, and well-designed studies leveraging this technology can advance our limited understanding of the rare variant contribution to complex disorders.

3.1.2 Early studies of rare variants in psychiatric disorders

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

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

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

3.1.3 Emerging results from sequencing studies of neurodevelopmental disorders

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

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

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

Insights into neurodevelopment from the Deciphering Developmental Disorders study

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

3.1.4 Goal and aims

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

3.1.5 Publication note and contributions

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

3.2 Materials and methods

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

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

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

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

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

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

$$X \sim Pois(2N_t\mu)$$

$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} P(X = i)$$

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

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

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

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

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

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

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

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

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

3.2.5 Validation of variants of interest

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

3.2.6 Functional consequence of the exon 16 splice acceptor deletion

The functional assay described here was performed by Sebastian S. Gerety of the DDD study to assess the impact of the exon 16 splice acceptor site variant. A custom minigene construct was first created by cloning the entire 696 bp genomic region encompassing exons 15, 16, 17 and intervening introns of human *SETD1A*, fused in-frame to a C-terminal GFP. We flanked the cassette with a strong upstream promoter and a downstream polyadenylation sequence. We transfected plasmids containing either the reference or deletion-containing forms into HELA cells, and these cells were grown for 2 days under standard conditions. The RNA was extracted (RNEasy, Qiagen) from the transfected cells and cDNA was synthesized (Superscript III, Invitrogen). Minigene-specific primers were designed to avoid amplification of endogenous HELA derived transcripts. The first pair of primers spanned all three exons, thus allowing us to detect overall splicing changes (Pair 1, Forward 2: TCGAAGAGTCATAAACACTGCCATG, Reverse 9: GTGAACAGCTCCTCGCCCTTG). We also designed pairs of exonic, intron-spanning primers to distinguish splicing events upstream (Pair 2, Forward 1: TTTGCAGGATCCCATCGAAGAGTC, exon 16 reverse: CACTGTCCATGATGGCGGAGGTA) and downstream (Pair 3, exon16 forward: CTGCTGAGCGCCATCGGTAC, exon17 reverse: CTGAACTTGTGGCCGTTTACGTC) of exon 16. We performed PCR on the cDNA from two transfection replicates of each sample. Agarose gels were used identify PCR product size differences (DNA ladder: 2-log ladder, New England Biolabs), which were further analysed by capillary sequencing.

3.2.7 Phenotype clustering in DDD probands

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

3.3 Results

3.3.1 Study design

The case-control data set consisted of 357,088 damaging missense and 55,955 LoF variants called in 4,264 cases and 9,343 controls (Figure 3.1). I restricted our analyses to rare variants, stratified by allele frequency (singletons, $< 0.1\%$, and $< 0.5\%$) and function (LoF and damaging missense variants). I first replicated the enrichment of rare LoF variants in the previously implicated set of 2,456 genes [103] in our UK and Finnish schizophrenia data sets ($P = 7 \times 10^{-4}$). Having confirmed that rare disruptive variants spread among many genes are associated with schizophrenia risk, I tested for an excess of disruptive variants within each of 18,271 genes in cases compared to controls using the Fisher's exact test. Despite our large sample size, the per-gene statistics followed a null distribution in all tests, and I was unable to implicate any gene via case-control burden of disruptive variants (Figures 3.2, 3.3).

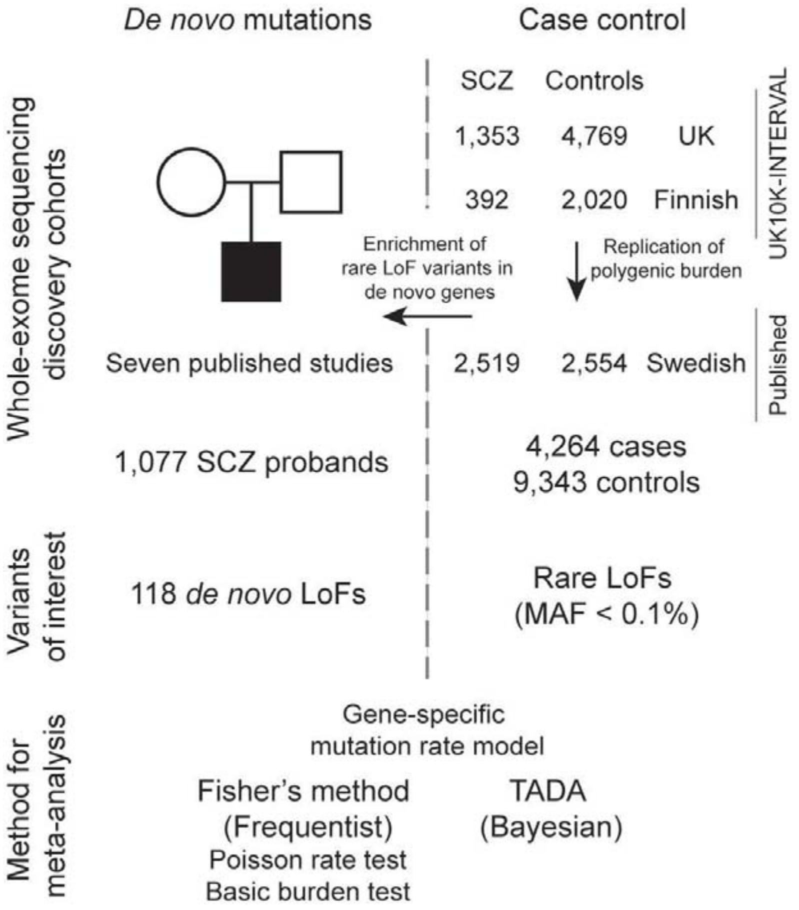


Fig. 3.1 **Study design for the schizophrenia exome meta-analysis.** The source of sequencing data, sample sizes, variant classes, and analytical methods are described. Details on case-control samples are shown on the right, while parent-proband trios are described on the left.

3.3.2 LoF variants in *SETD1A* are associated with schizophrenia

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

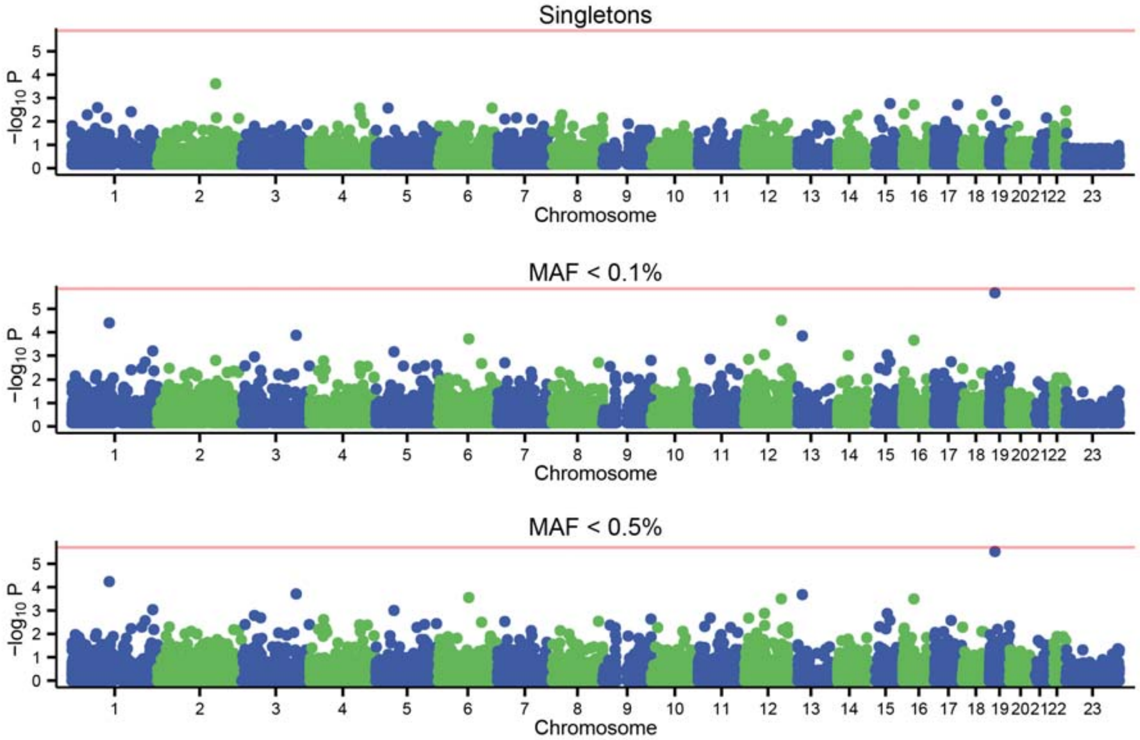


Fig. 3.2 **Manhattan plot of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.** I tested for an excess of LoF variants within 18,271 genes using Fisher's exact test. $-\log_{10}$ P-values were plotted against the chromosomal location (mid-point) of each gene. I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

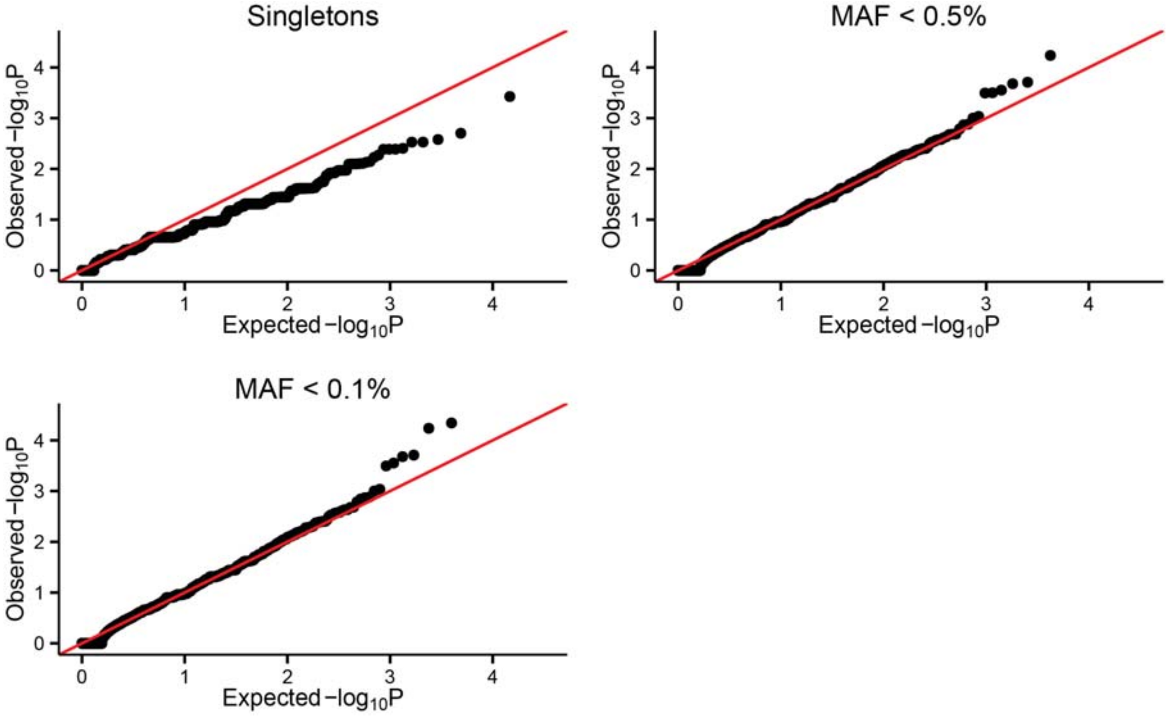


Fig. 3.3 **QQ plots of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.** I tested for an excess of LoF variants within 18,271 genes using Fisher's exact test, and plotted the ordered $-\log_{10}$ P-values against transformed P-values sampled from the uniform distribution.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

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

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

3.3.3 Robustness of the *SETD1A* association

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

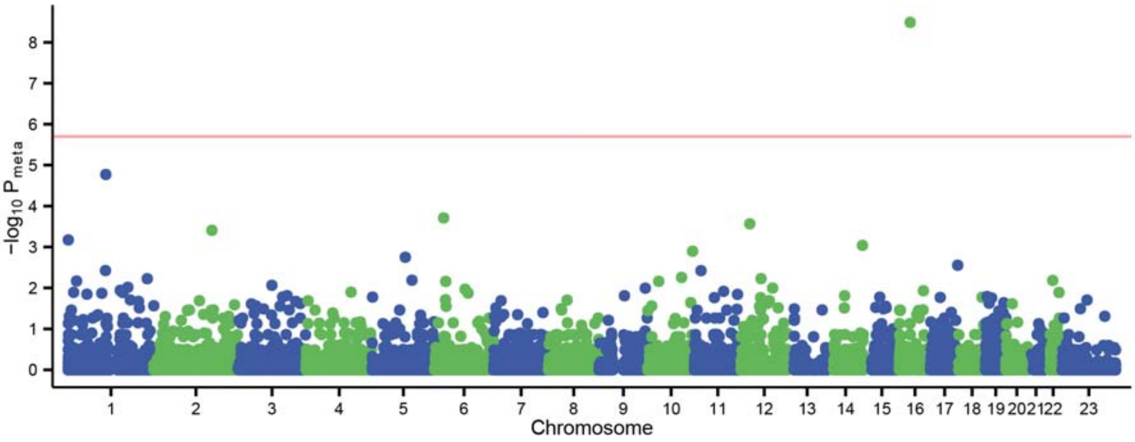
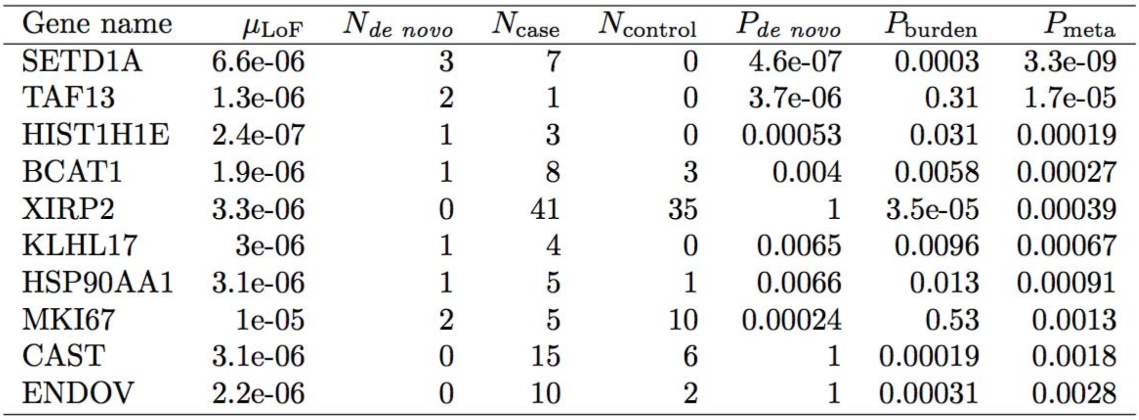


Fig. 3.4 **Manhattan plot of the meta-analysis of *de novo* mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.** *De novo* and case-control burden *P*-values were meta-analysed using Fisher's combined probability method. $-\log_{10}$ *P*-values were plotted against the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



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

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

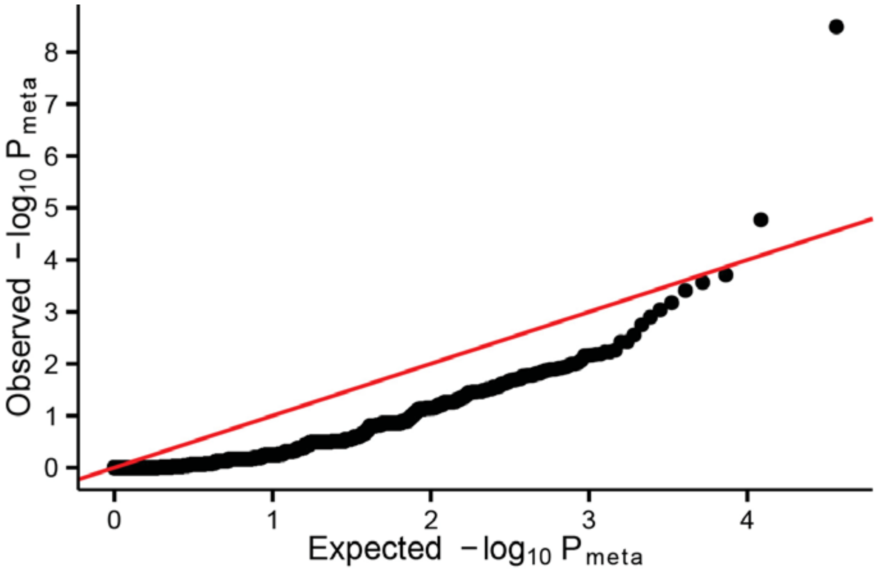


Fig. 3.5 **QQ plot of the meta-analysis of *de novo* mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.** *De novo* and case-control burden *P*-values were meta-analysed using Fisher's combined probability method, and the $\log_{10}$ *P*-values plotted against transformed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

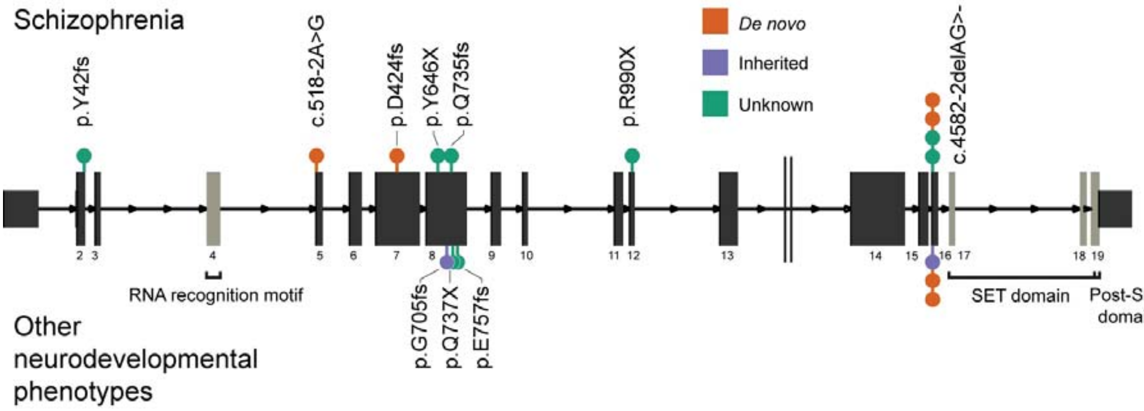
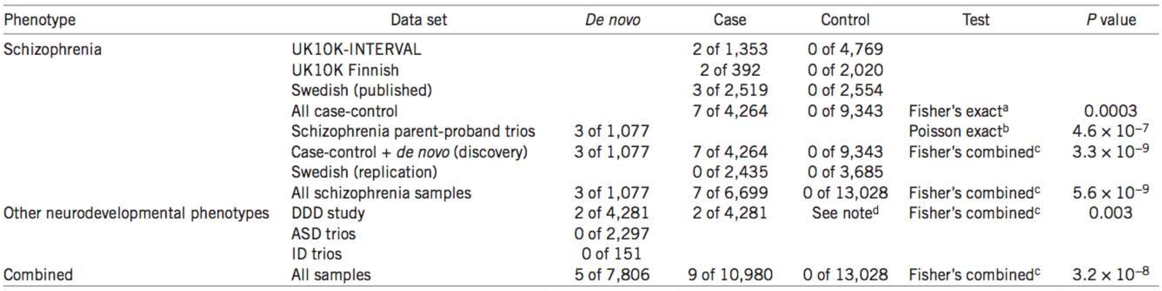


Fig. 3.6 **The genomic position and coding consequences of 16 *SETD1A* LoF variants observed in the schizophrenia exome meta-analysis, the DDD study, and the SiSU project.** Variants discovered in patients with schizophrenia are plotted above the gene, and those discovered in individuals with other neurodevelopmental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



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

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

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

Depletion of *SETD1A* LoF variants in the ExAC database

I performed five analyses to ensure our *SETD1A* association was robust to possible confounders of rare variant association testing. First, to validate our observation of the rarity of disruptive variants in *SETD1A* in unaffected individuals, I examined the Exome Aggregation Consortium (ExAC) v0.3 for the LoF variants in *SETD1A* [112]. All exomes in ExAC were joint-called using the GATK v3.2 pipeline, and included other public exome datasets, such as the 1000 Genomes Project and NHLBI-GO Exome Sequencing Project, with additional quality control compared to their original releases. In 60,706 unrelated exomes, I observed seven LoF variants in *SETD1A*. Since the v0.3 release aggregated studies of psychiatric disorders including the Swedish schizophrenia study, I excluded all samples from these data sets, leaving only four LoF variants in 45,376 exomes without a known neuropsychiatric diagnosis. I next applied the same stringent QC metrics used in our analysis to ExAC data. I found that the 16:30976302-GC/G indel observed in two individuals was located at the same position as a high-quality SNP, and occurred at a homopolymer run of cytosines. At the genotype level, both calls had a genotype quality (GQ) Phred probability of < 40, far lower than used in our study in which I required indels to have a GQ > 90. In addition, the variant has poor allelic balance (AB < 0.15), and the BAM alignment reflected these low-quality metrics [112]. Given this evidence, I excluded the putative indel. Two high-quality *SETD1A* LoF variants in 45,376 unaffected ExAC exomes remained. Following the approach in Samocha *et al.*, I determined the significance of the depletion of *SETD1A* LoF variants in ExAC using a signed Z-score of the χ-squared deviation between observed and expected counts [138]. I scaled the expected LoF counts provided by ExAC (43 in 60,706) to 45,376 exomes (expected 32.5), and calculated the one-tailed *P*-value of the signed Z-score assuming two observed LoF variants. Observing only two LoF variants when expecting 32.5 variants represented a substantial depletion compared to chance expectation ($P = 4.4 \times 10^{-8}$). According to its pLI score, a measure of constraint relative to other coding genes calculated using the ExAC data, *SETD1A* is among the 3% most constrained genes in the human genome [112]; LoF variants in *SETD1A* were almost totally absent in the general population.

Dependence of results on specification of mutation rate

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

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

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

Independence of results on parameterization in TADA

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

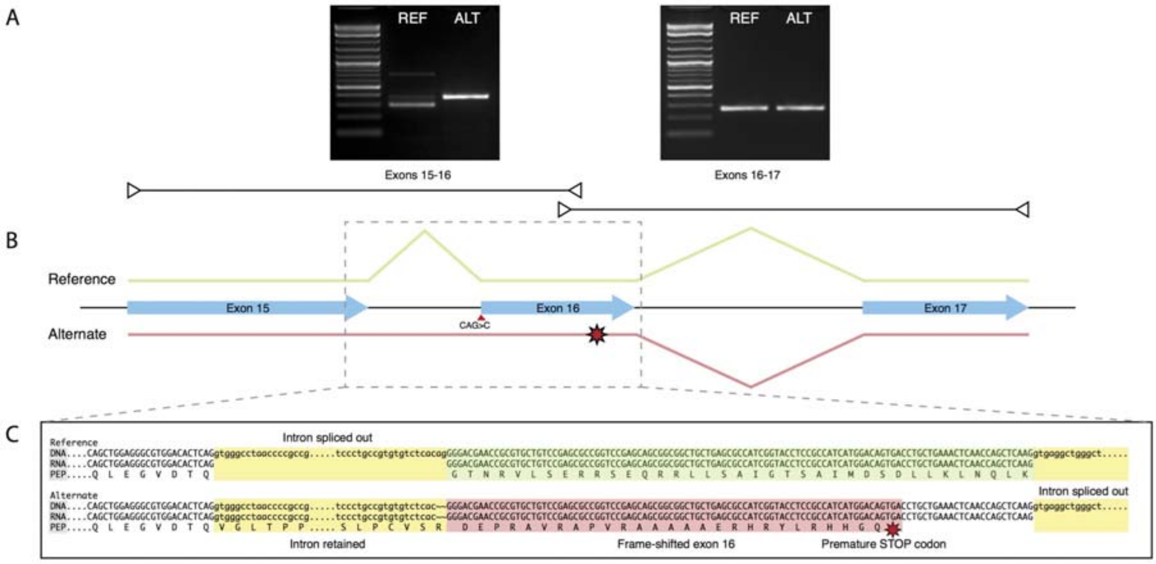

A
REF
ALT
Exons 15-16
REF
ALT
Exons 16-17
B
Reference
Exon 15
Exon 16
Exon 17
CAG>C
Alternate
C
Intron spliced out
Reference
DNA....CAGCTGGAGGGCGTGGACACTCAGgtgggcctaaccccgccg.....tccctgccgtgtgtctcacagGGACGAACCGCGTGCTGTCCGAGCGCCGGTCCGAGCAGCGGCGGCTGCTGAGCGCCATCGGTACCTCCGCCATCATGGACAGTGACCTGCTGAAACTCAACCAGCTCAAGgtgaggctgggct.....
RNA....CAGCTGGAGGGCGTGGACACTCAGGGACGAACCGCGTGCTGTCCGAGCGCCGGTCCGAGCAGCGGCGGCTGCTGAGCGCCATCGGTACCTCCGCCATCATGGACAGTGACCTGCTGAAACTCAACCAGCTCAAG
PEP.....Q L E G V D T Q G T N R V L S E R R S E Q R R L L S A I G T S A I M D S D L L K L N Q L K
Intron spliced out
Alternate
DNA....CAGCTGGAGGGCGTGGACACTCAGgtgggcctaaccccgccg.....tccctgccgtgtgtctcac—GGACGAACCGCGTGCTGTCCGAGCGCCGGTCCGAGCAGCGGCGGCTGCTGAGCGCCATCGGTACCTCCGCCATCATGGACAGTGACCTGCTGAAACTCAACCAGCTCAAGgtgaggctgggct.....
RNA....CAGCTGGAGGGCGTGGACACTCAGgtgggcctaaccccgccg.....tccctgccgtgtgtctcac—GGACGAACCGCGTGCTGTCCGAGCGCCGGTCCGAGCAGCGGCGGCTGCTGAGCGCCATCGGTACCTCCGCCATCATGGACAGTGACCTGCTGAAACTCAACCAGCTCAAG
PEP.....Q L E G V D T Q V G L T P P..... S L P C V S R D E P R A V R A P V R A A A A A E R H R Y L R H H G Q
Intron retained
Frame-shifted exon 16
Premature STOP codon

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

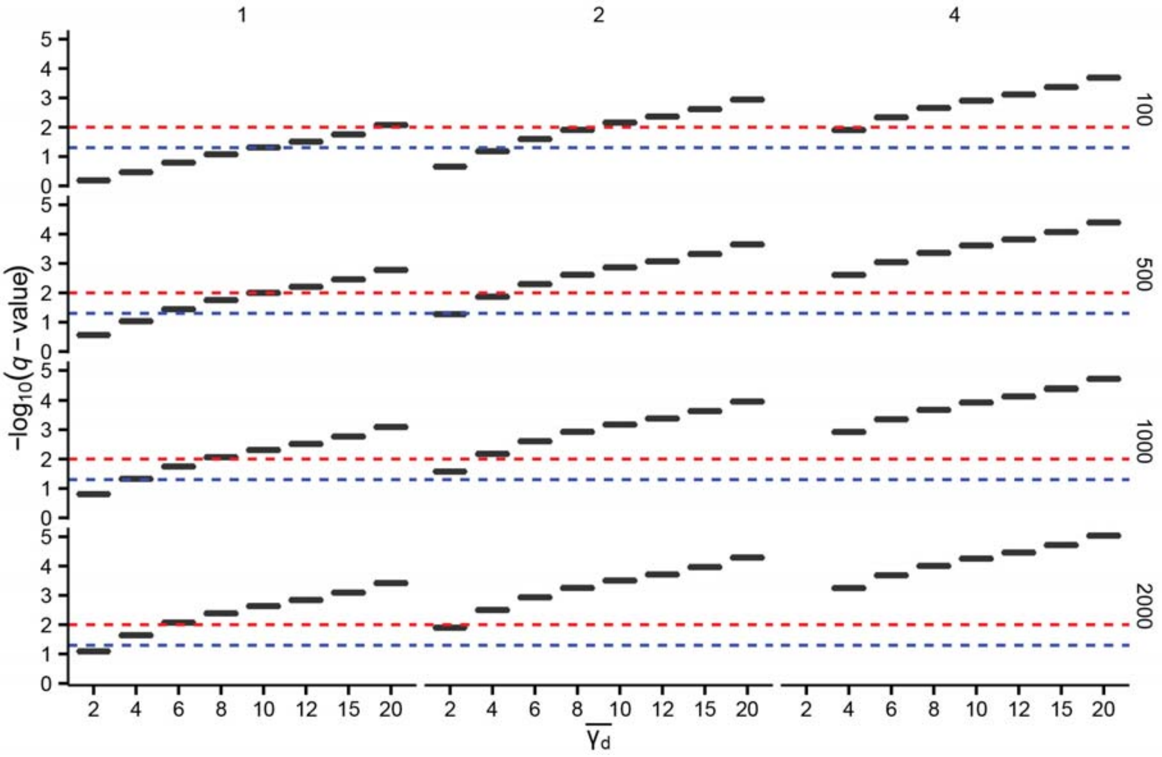
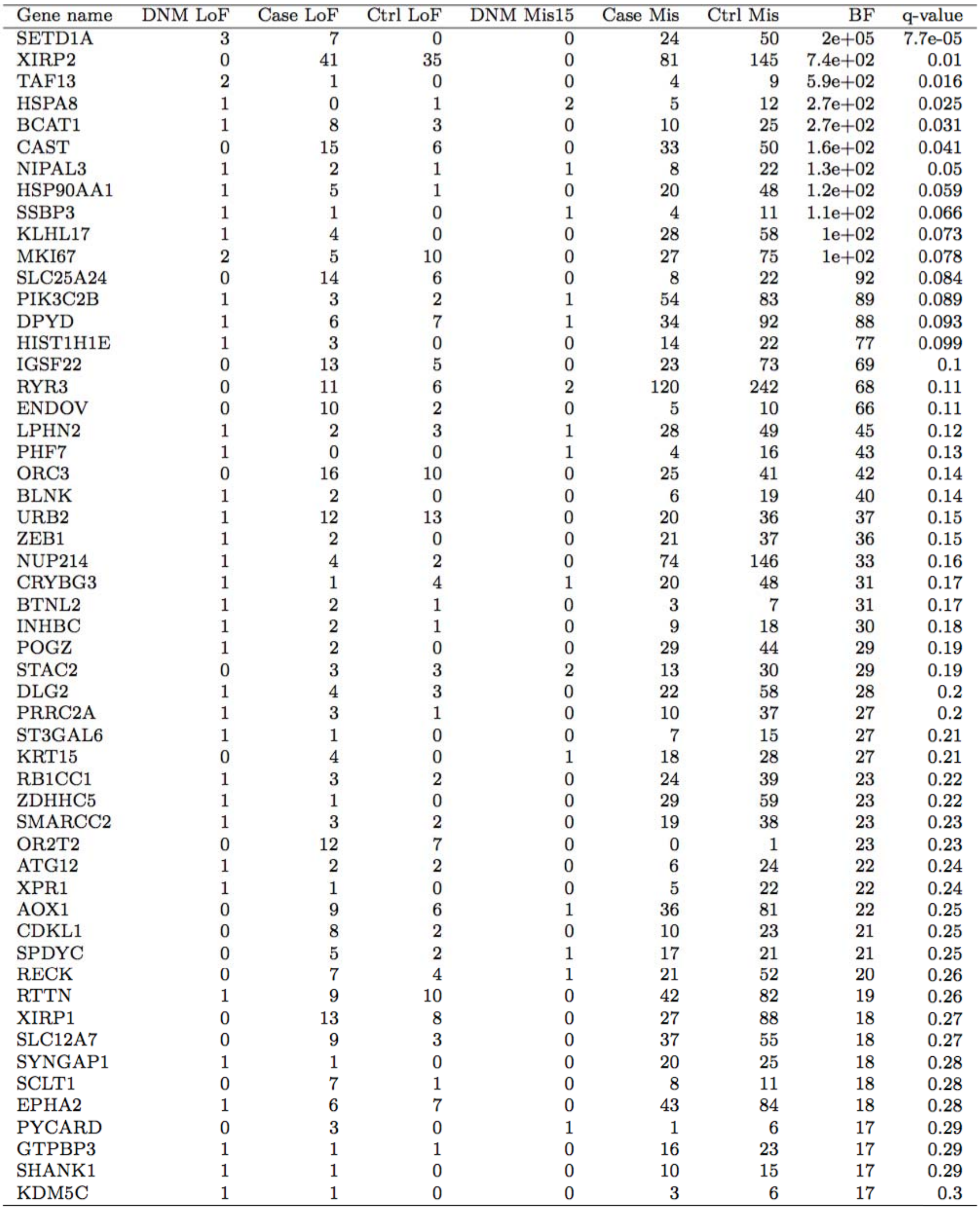
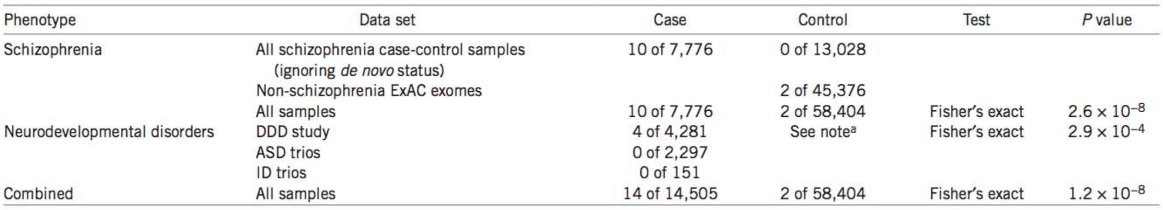


Fig. 3.8 **The robustness of the *SETD1A* result across reasonable parameters in the TADA model.** Because the TADA model depended heavily on the specification of its hyperparameters, I calculated the log q-value of *SETD1A* across different mean relative risk of *de novo* variants ($\bar{\gamma}_d$),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



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

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



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

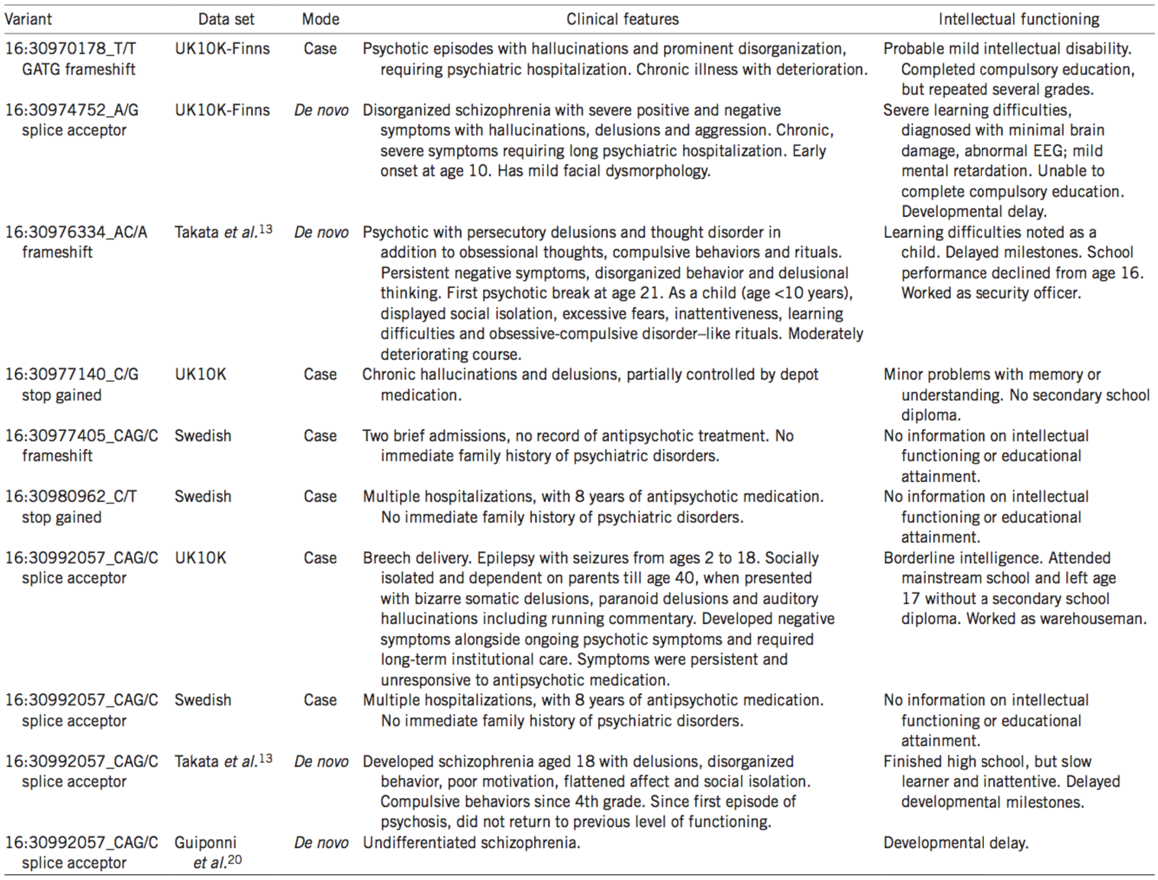
Table 3.4 **Burden tests associating disruptive variants in *SETD1A* to schizophrenia and developmental delay.** *De novo* status of variants was ignored and non-schizophrenia exomes from the ExAC database were incorporated as controls. The number of *SETD1A* LoF variants and the sample size of each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

Burden testing with non-psychiatric ExAC exomes as additional controls

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

3.3.4 *SETD1A* is associated with severe developmental disorders

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

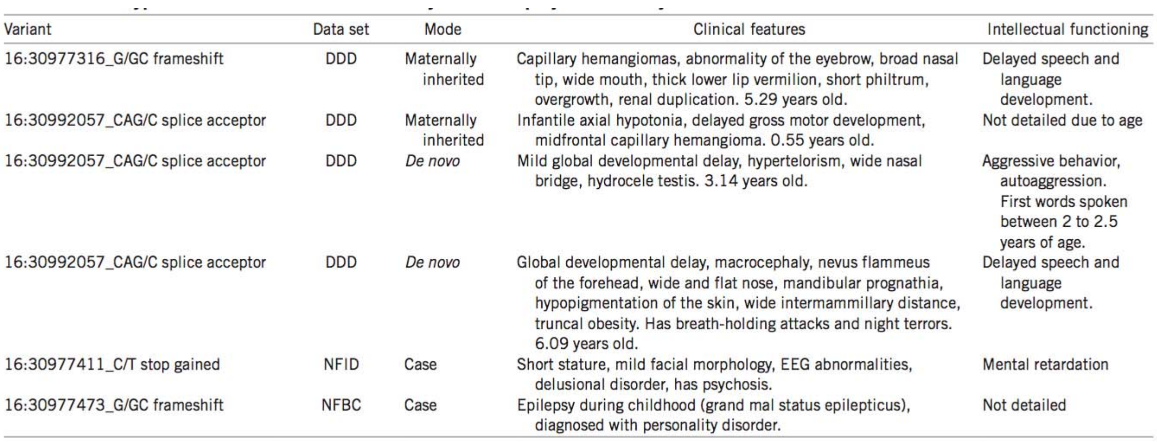


Variant	Data set	Mode	Clinical features	Intellectual functioning
16:30970178_T/T GATG frameshift	UK10K-Finns	Case	Psychotic episodes with hallucinations and prominent disorganization, requiring psychiatric hospitalization. Chronic illness with deterioration.	Probable mild intellectual disability. Completed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

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

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

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



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

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

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

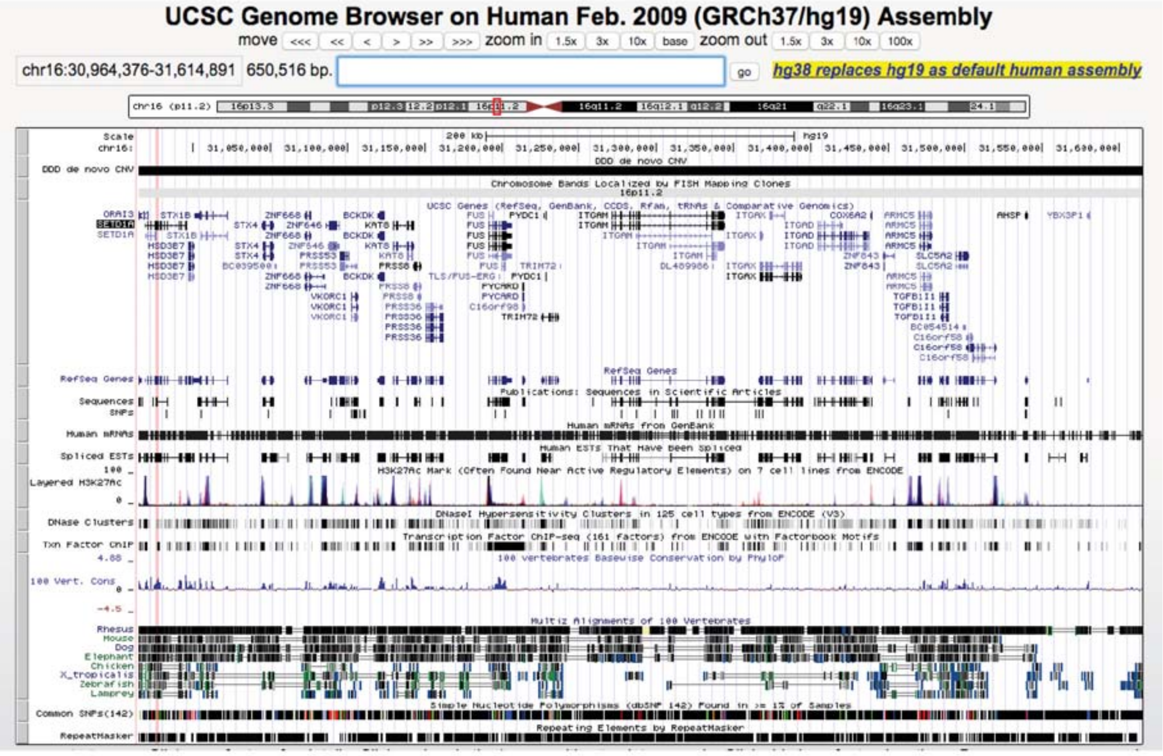

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr16:30,964,376-31,614,891 650,516 bp.
go
hg38 replaces hg19 as default human assembly
chr16 (p11.2)
Scale
chr16:
200 kb
hg19
DDD de novo CNV
Chromosome Bands Localized by FISH Mapping Clones
16p11.2
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ORAI3
SETD1A
STX1B
HSD3B7
STX4
BC039500
ZNF668
ZNF646
PRSS53
VKORC1
BCKDK
KAT8
PRSS8
PRSS36
FUS
TLS/FUS-ERG
PYDC1
PYCARD
C16orf98
TRIM72
ITGAM
ITGAX
ITGAD
DL489986
COX6A2
ARMC5
ZNF843
SLC5A2
TGFB1I1
BC054514
C16orf58
AHSP
YBX3P1
RefSeq Genes
Sequences
SNPs
Publications: Sequences in Scientific Articles
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(142)
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

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

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

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

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

I defined the following:

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

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

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

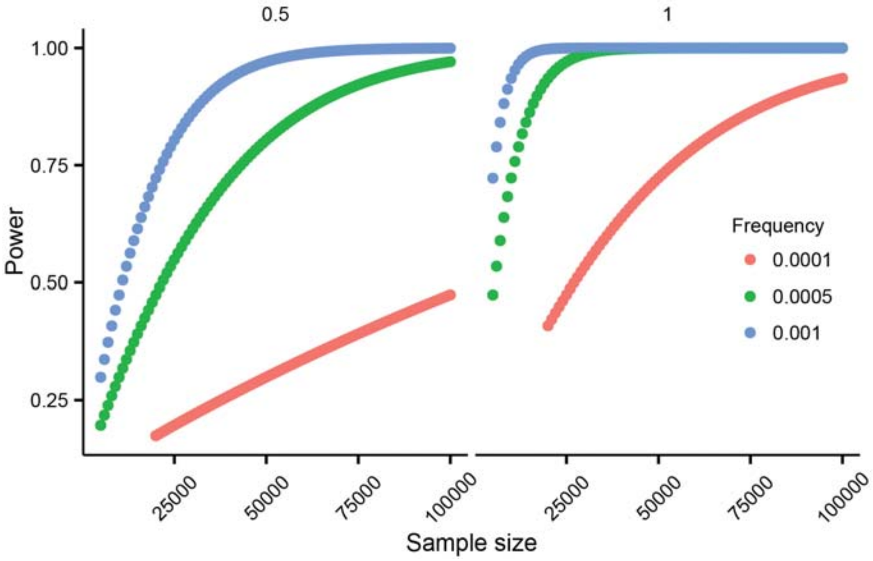


Fig. 3.10 **Sample size curves for detecting an increased risk of pre-morbid cognitive impairment in schizophrenia *SETD1A* LoF carriers.** I performed power calculations using a simple one-sided t-test to identify sample sizes required to show possible cognitive impairment in *SETD1A* schizophrenia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

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

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

3.3.6 *De novo* burden in neurodevelopmental disorders

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

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

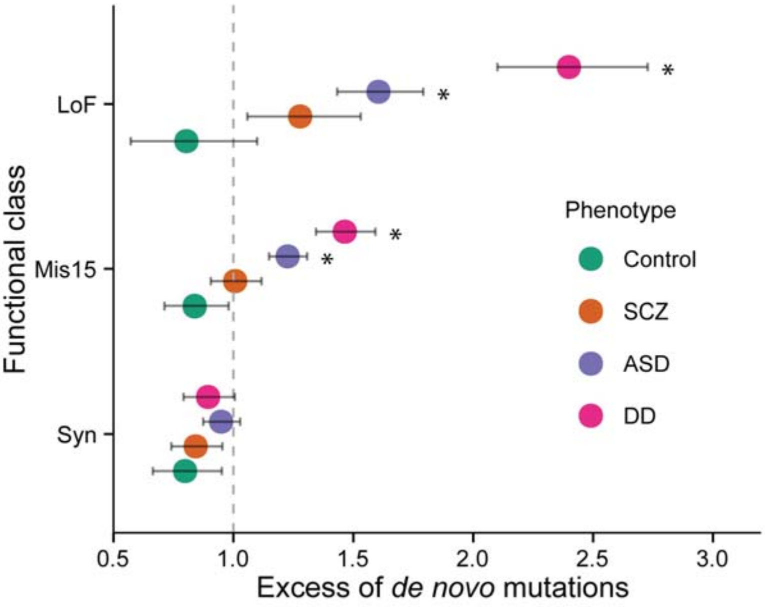


Fig. 3.11 **A comparison of genome-wide *de novo* mutation rates in probands with autism, developmental disorders, schizophrenia, and controls.** Rates were modelled using calibrated genome-wide mutation rates. Significant excess of *de novo* mutations when compared to the baseline model was marked with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

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

3.4 Discussion

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

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

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

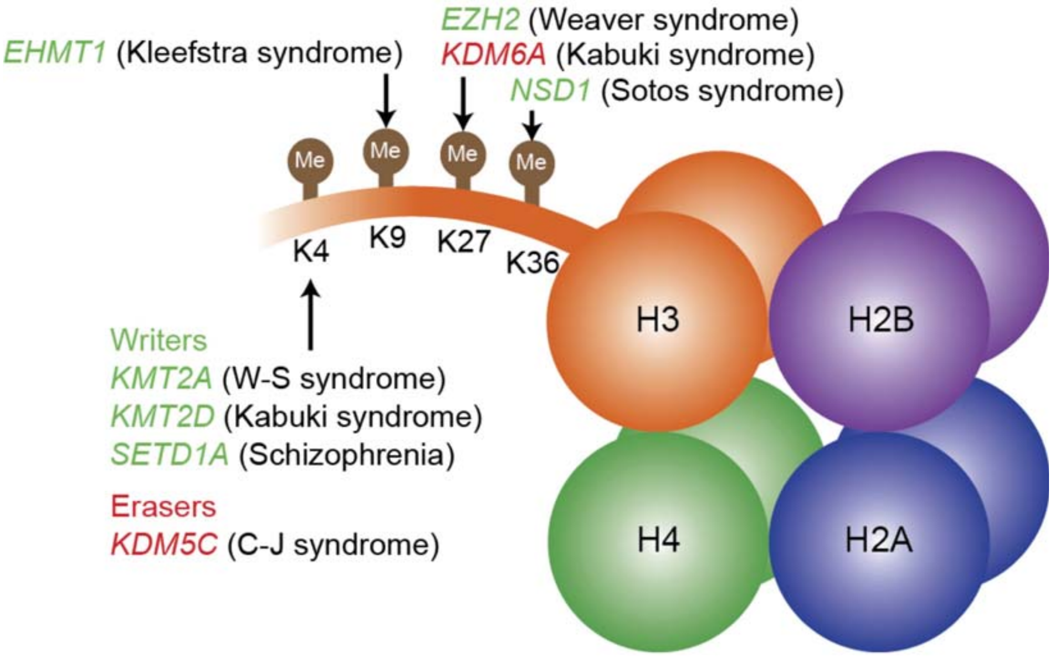


Fig. 3.12 **Mendelian disorders of epigenetic machinery at histone H3.** Writers (in green) add methyl groups at the specified residue of the histone tail, while erasers (in red) perform targeted demethylation. Disrupting variants in writers and erasers described in the figure result in well-known examples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

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

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

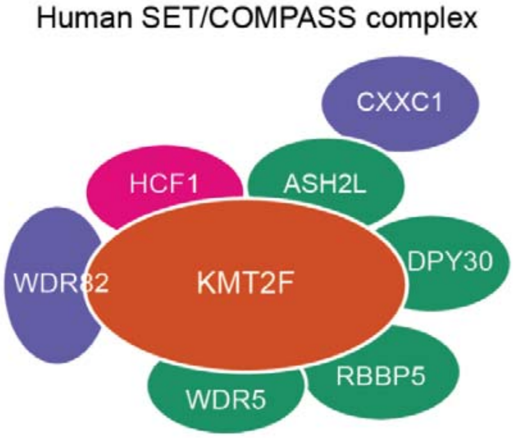


Fig. 3.13 **SET1/COMPASS complex** A highly conserved protein complex that methylates the tail of histone H3. *SETD1A* or *KMT2F* is one of the catalytic cores of this complex.

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

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