nature genetics

The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability

Tarjinder Singh[1], James T R Walters[2], Mandy Johnstone[3], David Curtis[4,5], Jaana Suvisaari[6], Minna Torniainen[6], Elliott Rees[2], Conrad Iyegbe[7], Douglas Blackwood[3], Andrew M McIntosh[8], Georg Kirov[2], Daniel Geschwind[9], Robin M Murray[7], Marta Di Forti[7], Elvira Bramon[10], Michael Gandal[9], Christina M Hultman[11], Pamela Sklar[12], INTERVAL Study[13], UK10K Consortium[14], Aarno Palotie[15,16], Patrick F Sullivan[17,18], Michael C O'Donovan[2], Michael J Owen[2] & Jeffrey C Barrett[1]

By performing a meta-analysis of rare coding variants in whole-exome sequences from 4,133 schizophrenia cases and 9,274 controls, *de novo* mutations in 1,077 family trios, and copy number variants from 6,882 cases and 11,255 controls, we show that individuals with schizophrenia carry a significant burden of rare, damaging variants in 3,488 genes previously identified as having a near-complete depletion of loss-of-function variants. In patients with schizophrenia who also have intellectual disability, this burden is concentrated in risk genes associated with neurodevelopmental disorders. After excluding known risk genes for neurodevelopmental disorders, a significant rare variant burden persists in other genes intolerant of loss-of-function variants; although this effect is notably stronger in patients with both schizophrenia and intellectual disability, it is also seen in patients with schizophrenia who do not have intellectual disability. Together, our results show that rare, damaging variants contribute to the risk of schizophrenia both with and without intellectual disability and support an overlap of genetic risk between schizophrenia and other neurodevelopmental disorders.

Schizophrenia is a common and debilitating psychiatric illness characterized by positive symptoms (hallucinations, delusions, and disorganized speech and behavior), negative symptoms (social withdrawal and diminished emotional expression), and cognitive impairment that result in social and occupational dysfunction[1,2]. Operational diagnostic criteria for the disorder as described in the DSM-V require the presence of at least two of the core symptoms over a period of 6 months with at least 1 month of active symptoms[3]. It is increasingly recognized that current categorical psychiatric classifications have a number of shortcomings, in particular that they overlook the increasing evidence for etiological and mechanistic overlap between psychiatric disorders[4].

A diverse range of pathophysiological processes may contribute to the clinical features of schizophrenia[5]. Indeed, previous studies have suggested a number of hypotheses about schizophrenia pathogenesis, including abnormal presynaptic dopaminergic activity[6], postsynaptic mechanisms involved in synaptic plasticity[7], dysregulation of synaptic pruning[8], and disruption to early brain development[9,10]. This complexity is underpinned by the varied nature of genetic contributions to risk of schizophrenia. Genome-wide association studies have identified over 100 independent loci defined by common (minor allele frequency (MAF) > 1%) single-nucleotide variants (SNVs)[11], and a recent analysis determined that more than 71% of all 1-Mb regions in the genome contain at least one common risk allele[12]. The modest effects of these variants (median odds ratio (OR) = 1.08) combine to produce a polygenic contribution that explains only a fraction ($h_g^2 = 0.274$) of the overall liability[12]. In addition, a number of rare variants have been identified that have far larger effects on individual risk. These are best exemplified by 11 large, rare recurrent copy number variants (CNVs) and loss-of-function variants in *SETD1A*, but evidence from whole-exome sequencing studies implies that many other rare coding SNVs and *de novo* mutations also confer substantial individual risk[13–18]. There is growing evidence that some of the same genes and pathways are affected by both common and rare variants[7,18]. Pathway analyses of common variants and hypothesis-driven gene set analyses of rare variants have begun to enumerate some of these specific biological processes, including histone methylation, transmission at glutamatergic synapses, calcium channel signaling, synaptic plasticity, and translational regulation by the fragile X mental retardation protein (FMRP)[11,13,14,19,20].

In addition to exploring the biological mechanisms underlying schizophrenia, genetic analyses can also be used to understand the relationship of schizophrenia to other neuropsychiatric and neurodevelopmental disorders. For instance, schizophrenia, bipolar disorder, and autism spectrum disorder (ASD) show substantial overlap of common risk variants[21,22]. Sequencing studies of neurodevelopmental disorders suggest that this shared genetic risk may extend to rare variants of large effect. In the largest sequencing study of ASD thus far, 20 of the 46 genes and all six CNVs implicated (false discovery rate (FDR) < 5%) had previously been described as dominant causes of developmental disorders[23]. Furthermore, an analysis of 60,706 whole exomes led by the Exome Aggregation Consortium (ExAC) identified 3,230 genes with near-complete depletion of protein-truncating variants, and *de novo*

A full list of affiliations appears at the end of the paper.

Received 29 June 2016; accepted 1 June 2017; published online 26 June 2017; doi:10.1038/ng.3903

loss-of-function mutations identified in individuals with ASD or developmental disorders were concentrated in this set of 'loss-of-function-intolerant' genes[23–25]. Similarly, evidence from rare variants for a broader shared genetic etiology between schizophrenia and neurodevelopmental disorders has begun to emerge. Analyses of whole-exome data provided support for an enrichment of rare variants for schizophrenia in genes associated with intellectual disability, and schizophrenia cases were also found to have a higher concentration of ultra-rare disruptive SNVs in the ExAC loss-of-function-intolerant genes as compared to controls[13,17,26].

However, the contribution of these rare variants to risk in the wider population of individuals diagnosed with schizophrenia, including those without intellectual disability, remains unclear. Intriguingly, the 11 rare CNVs found to be highly penetrant for schizophrenia also increased risk for intellectual disability and other congenital defects[16,27], and, more recently, a meta-analysis of whole-exome sequence data showed that loss-of-function variants in *SETD1A* conferred substantial risk for both schizophrenia and neurodevelopmental disorders[18]. Concurrent analyses of ASD whole-exome data found that *de novo* loss-of-function mutations identified in ASD probands, particularly those that disrupt genes associated with neurodevelopmental disorders, were disproportionately found in individuals with intellectual disability[23,28]. These emerging results raise the possibility that rare risk variants for schizophrenia may be concentrated in a subset of patients with schizophrenia who have comorbid intellectual disability. Here we present one of the largest accumulations thus far of rare variant data for schizophrenia, which we jointly analyze with phenotype data on cognitive function. Using this data set, we attempt to identify groups of genes disrupted by rare risk variants in schizophrenia and to determine whether a subset of patients disproportionately carry these damaging alleles.

RESULTS

Study design

To maximize our power to detect enrichment of damaging variants in schizophrenia cases in groups of genes, we performed a meta-analysis of three different types of rare coding variant data: (i) high-quality SNV calls from the whole-exome sequences of 4,133 schizophrenia cases and 9,274 matched controls, (ii) *de novo* mutations identified in 1,077 schizophrenia parent–proband trios, and (iii) CNV calls from genotyping array data of 6,882 cases and 11,255 controls (**Fig. 1**). The ascertainment of these samples, data production, and quality control were described previously[18,29]. All *de novo* mutations included in our analysis had been validated through Sanger sequencing, and stringent quality control steps were performed on the case–control data to ensure that sample ancestry and batch were closely matched between cases and controls (Online Methods).

For each data type, we used appropriate methods to test for an excess of rare variants (**Fig. 1** and Online Methods). In analyses of case–control SNV data, we applied an extension of the variant threshold burden test that corrected for exome-wide differences between cases and controls[30]. We tested all allele frequency thresholds below 0.1% observed in our data and assessed statistical significance by permutation testing. In analyses of *de novo* SNV data, we compared the observed number of *de novo* mutations to random samples from an expected distribution based on a gene-specific mutation rate model to calculate an empirical *P* value. For both types of whole-exome sequencing data, we restricted our analyses to loss-of-function variants. Finally, in analyses of case–control CNV data, we used a logistic regression framework that compares the rate of CNVs overlapping a specific gene set while correcting for differences in CNV size and number of genes disrupted[7,19,31]. To ensure that our model was well calibrated, we restricted our analyses to small deletions and duplications overlapping fewer than seven genes with MAF <0.1% (Online Methods and **Supplementary Fig. 1**).

We tested for an excess of rare, damaging variants in patients with schizophrenia in 1,766 gene sets (Online Methods, **Supplementary Table 1**, **Supplementary Note**, and detailed results below). Gene set *P* values were computed using the three methods and variant definitions described above, and meta-analysis was then performed using Fisher's method to provide a single *P* value for each gene set. Because we gave each data type equal weight, gene sets achieving significance typically showed at least some signal in all three types of data. We observed a marked inflation in the quantile–quantile plot of gene set *P* values (**Supplementary Fig. 2**), so we conducted two analyses to ensure that our results were robust and not biased as a result of methodological or technical artifacts. First, we observed no inflation of *P* values when testing for enrichment of synonymous variants in our case–control and *de novo* analyses (**Supplementary Fig. 2**). Second, we created random gene sets by sampling uniformly across the genome and observed null distributions in quantile–quantile plots, regardless of variant class and analytical method (**Supplementary Fig. 3**). These findings suggest that our methods sufficiently corrected for known genome-wide differences in loss-of-function variant and CNV burden between cases and controls and for other technical confounders like batch and ancestry.

Rare, damaging schizophrenia variants are concentrated in loss-of-function-intolerant genes

We first tested whether rare schizophrenia risk variants were consistently concentrated in genes defined as loss of function intolerant across study designs and variant types. Because some of our schizophrenia exome data were included in the ExAC database, we focused on the subset of 45,376 ExAC exomes without a known psychiatric diagnosis and that were not present in our study. From this subset, 3,488 genes were found to have near-complete depletion of such variants, which we defined as the loss-of-function-intolerant gene set. We found that rare, damaging variants in schizophrenia cases were enriched in loss-of-function-intolerant genes ($P < 3.6 \times 10^{-10}$; **Fig. 2** and **Table 1**), with support in case–control SNVs ($P < 5 \times 10^{-7}$; OR = 1.24, 95% confidence interval (CI) = 1.16–1.31), case–control CNVs ($P = 2.6 \times 10^{-4}$; OR = 1.21, 95% CI = 1.15–1.28), and *de novo* mutations ($P = 6.7 \times 10^{-3}$; OR = 1.36, 95% CI = 1.1–1.68). Although this result is consistent with observations in ASD and severe developmental disorders[24,32], the absolute effect size is smaller (for example, for *de novo* mutations; **Supplementary Figs. 4** and **5**). We observed no excess burden of rare, damaging variants in the remaining 14,753 genes (**Fig. 2** and **Supplementary Fig. 5**). Furthermore, this signal was spread among many different loss-of-function-intolerant genes: when we ranked genes by decreasing significance, the enrichment disappeared in the case–control SNV analysis ($P > 0.05$) only after exclusion of the top 50 genes. This suggests that the contribution of rare, damaging variants in schizophrenia is not concentrated in just a handful of genes but instead is spread across many genes.

Schizophrenia risk genes are shared with other neurodevelopmental disorders

Given the significant enrichment of rare, damaging variants in loss-of-function-intolerant genes in schizophrenia, ASD, and severe developmental disorders, we next asked whether these variants affected the same genes. We found that ASD risk genes identified from exome sequencing meta-analyses[23] and genes in which loss-of-function variants

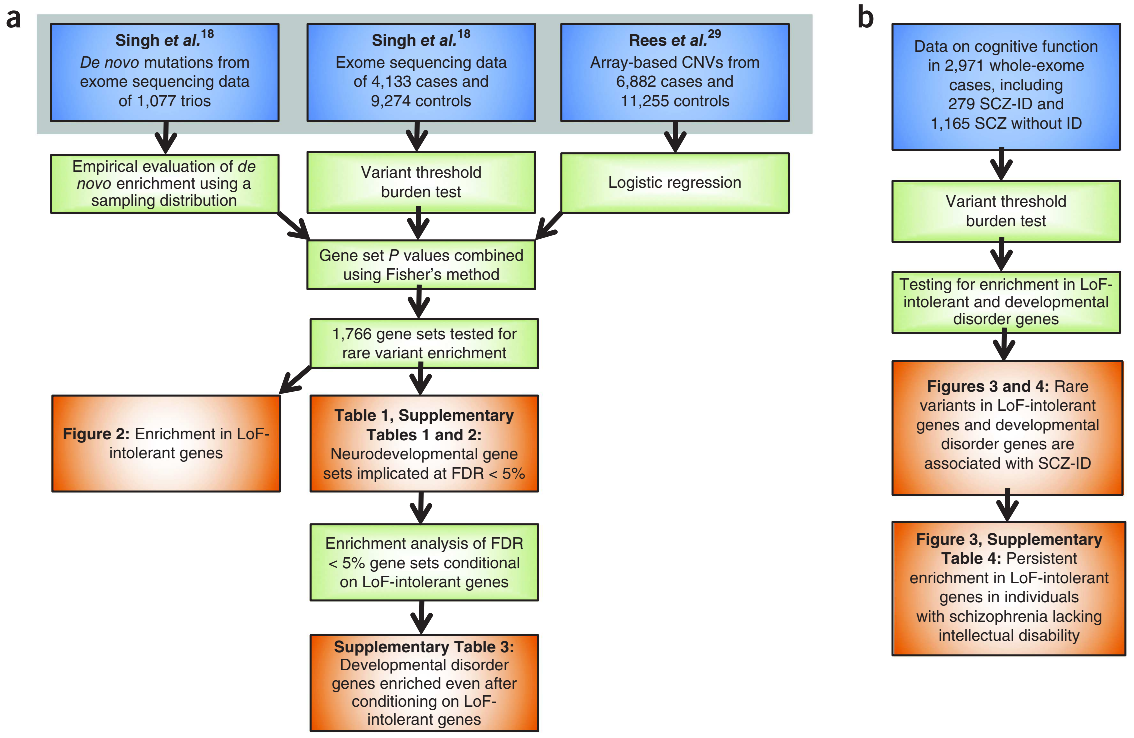
a
Singh et al.18
De novo mutations from exome sequencing data of 1,077 trios
Singh et al.18
Exome sequencing data of 4,133 cases and 9,274 controls
Rees et al.29
Array-based CNVs from 6,882 cases and 11,255 controls
Empirical evaluation of de novo enrichment using a sampling distribution
Variant threshold burden test
Logistic regression
Gene set P values combined using Fisher's method
1,766 gene sets tested for rare variant enrichment
Figure 2: Enrichment in LoF-intolerant genes
Table 1, Supplementary Tables 1 and 2: Neurodevelopmental gene sets implicated at FDR < 5%
Enrichment analysis of FDR < 5% gene sets conditional on LoF-intolerant genes
Supplementary Table 3: Developmental disorder genes enriched even after conditioning on LoF-intolerant genes
b
Data on cognitive function in 2,971 whole-exome cases, including 279 SCZ-ID and 1,165 SCZ without ID
Variant threshold burden test
Testing for enrichment in LoF-intolerant and developmental disorder genes
Figures 3 and 4: Rare variants in LoF-intolerant genes and developmental disorder genes are associated with SCZ-ID
Figure 3, Supplementary Table 4: Persistent enrichment in LoF-intolerant genes in individuals with schizophrenia lacking intellectual disability

Figure 1 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. (**a**) Enrichment analyses in 1,766 gene sets using the entire rare variant data set. (**b**) Enrichment analyses in loss-of-function-intolerant and developmental disorder–associated genes in the subset of cases with information on cognitive function. ID, intellectual disability; SCZ, schizophrenia; SCZ-ID, schizophrenia with intellectual disability; LoF, loss of function.

are known causes of severe developmental disorders as defined by the Deciphering Developmental Disorders (DDD) study[33,34] were significantly enriched for rare variants in individuals with schizophrenia ($P_{\mathrm{ASD}} = 9.5 \times 10^{-6}$; $P_{\mathrm{DD}} = 2.3 \times 10^{-6}$; **Table 1**, Online Methods, and **Supplementary Note**). Previous analyses have shown an enrichment of rare, damaging variants in genes whose mRNAs are bound by FMRP in both schizophrenia and ASD[13,32,35], so we sought to identify further shared biology by testing targets of neural regulatory genes previously implicated in ASD[32,36,37]. We observed enrichment of both such sets: promoter targets of CHD8 ($P = 1.1 \times 10^{-6}$) and splicing targets of RBFOX ($P = 1.3 \times 10^{-5}$) (**Table 1**). We noted that some published gene lists attributed to the same biological process differed owing to choice of assay, cell type, method of sample extraction, and threshold for statistical significance, leading to distinct results in our gene set analyses. For example, we observed significant enrichment in the published FMRP- binding gene set based on mouse brain data[38] but found no signal in one based on data from a human kidney cell line[39].

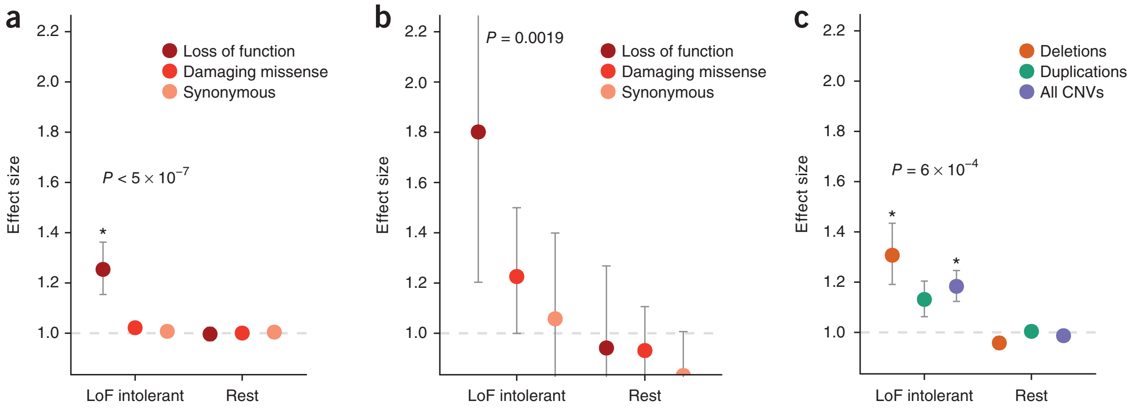

Figure 2 Enrichment of schizophrenia-associated rare variants in genes intolerant of loss-of-function variants. (**a**) Schizophrenia cases compared to controls for rare SNVs and indels. (**b**) Rates of *de novo* mutation in schizophrenia probands as compared to control probands. (**c**) Schizophrenia cases compared to controls for CNVs. *P* values shown are from a test of enrichment for loss-of-function variants in **a** and **b**, and a test for enrichment of all CNVs in **c**. Enrichments are displayed without conditioning on genome-wide differences. Bars represent the 95% CIs of the point estimates. Loss of function intolerant, 3,488 genes with near-complete depletion of truncating variants in the ExAC database; Rest, the remaining genes in the genome with pLI <0.9; damaging missense, missense variants with CADD Phred >15; LoF, loss of function. $^*P < 1 \times 10^{-3}$.

Table 1 **Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR < 1%**

Gene set	N_{genes}	Est_{SNV}	CI_{SNV}, 95%	P_{SNV}	Est_{DNM}	CI_{DNM}, 95%	P_{DNM}	Est_{CNV}	CI_{CNV}, 95%	P_{CNV}	P_{meta}	Q_{meta}
ExAC: LoF-intolerant genes (pLI > 0.9)	3,488	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	$< 3.6 \times 10^{-10}$	4.30×10^{-7}
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	2.3×10^{-6}	0.00067
Sanders *et al.*[23]: ASD 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	9.5×10^{-6}	0.0017
Darnell *et al.*[38]: 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	9.3×10^{-7}	0.00038
Cotney *et al.*[36] : CHD8-targeted promoters (human NSCs and human brain tissue)	2,920	1.09	1.02–1.16	0.0008	1.77	1.36–2.31	0.00025	1.11	1.05–1.18	0.027	1.1×10^{-6}	0.00038
G2CDB: mouse cortex postsynaptic density consensus	1,527	1.20	1.11–1.3	2.5×10^{-6}	1.57	1.06–2.33	0.028	1.04	0.96–1.11	0.32	3.9×10^{-6}	0.00097
Weyn-Vanhentenryck *et al.*[37]: CLIP targets of RBFOX	967	1.21	1.11–1.33	4.8×10^{-5}	1.84	1.21–2.80	0.0085	1.07	0.98–1.17	0.20	1.3×10^{-5}	0.0020
NMDAR network (defined in Purcell *et al.*[35])	61	1.66	1.09–2.54	0.0061	5.60	2.06–16.09	0.017	2.46	1.78–3.4	0.0028	3.7×10^{-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	4.2×10^{-5}	0.0046

The effect sizes and corresponding *P* values from enrichment tests of each variant type (case–control SNVs, *de novo* mutations, 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}). We only show the most significant gene set if there were multiple ones from the same data set or biological process (see **Supplementary Table 1** for all 1,766 gene sets). N_{genes}, number of genes in the gene set; Est, effect size estimate and its lower; CI, upper and lower bounds of the effect size estimate; DNM, *de novo* mutation.

We also tested an additional 1,759 gene sets with at least 100 genes from databases of biological pathways, as we lacked power to detect weak enrichments in smaller sets (Online Methods and **Supplementary Note**). We observed enrichment of rare, damaging variants in schizophrenia cases at FDR $q < 0.05$ in 35 of these gene sets (**Supplementary Tables 1** and **2**). These included previously implicated gene sets, like the NMDA receptor and ARC complexes[13,14,35,38], as well as novel gene sets, such as genes involved in cytoskeleton (GO:0007010), chromatin modification (GO:0016568), and chromatin organization (GO:0006325). Furthermore, the gene sets most significantly enriched (FDR $q < 0.01$) for rare variants in schizophrenia (**Table 1**) had all previously been linked to ASD, intellectual disability, and severe developmental disorders[23,32,33]. Our enrichment results matched some of the findings from a pathway analysis of common risk variants in psychiatric disorders, which also implicated neuronal and chromatin gene sets[20]. However, unlike that study, we found no enrichment of rare variants in immune-related gene sets.

We noticed that the gene sets we tested were collectively enriched with loss-of-function-intolerant genes when compared to a random sampling of genes from the genome (**Supplementary Figs. 6** and **7**). For some of the gene sets associated with schizophrenia, this overrepresentation was quite substantial: 67% of the gene targets of FMRP and 74% of the genes associated with severe neurodevelopmental disorders are loss of function intolerant. To better understand the consequences of this overlap for our results, we extended the gene set enrichment methods (Online Methods) to condition on loss-of-function intolerance and brain expression for the 35 gene sets with FDR $q < 0.05$ in the previous analysis (**Supplementary Table 2**). We first observed that 22 of the 35 gene sets remained significant even after conditioning on brain expression (Online Methods and **Supplementary Table 3**), suggesting that they represent more specific biological processes involved in schizophrenia. However, only known ASD risk genes ($P = 4.4 \times 10^{-4}$) and neurodevelopmental disorder–associated genes ($P = 3 \times 10^{-5}$) had an excess of rare coding variants above the enrichment already observed in loss-of-function-intolerant genes (**Supplementary Table 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.

Patients with both schizophrenia and intellectual disability have a greater burden of rare, damaging variants

In ASD, the observed excess of rare, damaging variants has been shown to be greater in individuals with intellectual disability than in those with normal levels of cognitive function[28]. We observed a similar phenomenon in schizophrenia cases carrying *SETD1A* loss-of-function variants[18], so we next sought to explore whether this pattern is consistent in gene sets implicated in schizophrenia. We acquired relevant cognitive phenotype data for 2,971 of the 4,131 patients with schizophrenia for whom whole-exome sequencing data were available (**Supplementary Fig. 8**). Of these individuals, 279 were clinically diagnosed with intellectual disability in addition to fulfilling the full diagnostic criteria for schizophrenia (SCZ-ID subgroup; Online Methods). We also identified 1,165 individuals for whom we could rule out intellectual disability (by excluding individuals with premorbid IQ <85, fewer than 12 years of schooling, or present in the lowest decile of composite cognitive measures, depending on available data; Online Methods). Finally, we identified 1,527 individuals who were not diagnosed with intellectual disability but in whom some cognitive impairment could not be excluded.

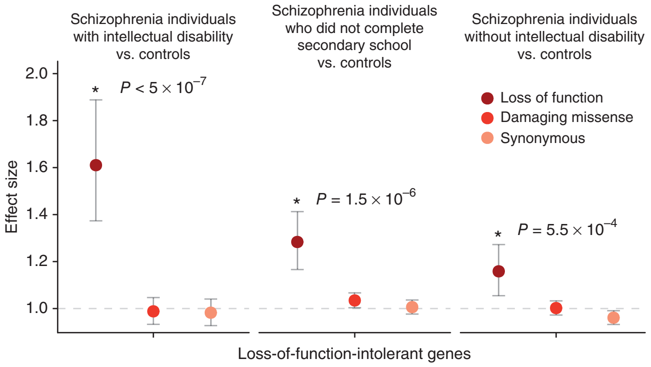

Figure 3 Enrichment of rare loss-of-function variants in loss-of-function-intolerant genes in schizophrenia cases stratified by information on cognitive function as compared to controls. The *P* values shown were calculated using the variant threshold method comparing the burden of loss-of-function variants between the corresponding cases and controls. Bars represent the 95% CIs of the point estimates. Damaging missense, missense variants with CADD Phred >15.

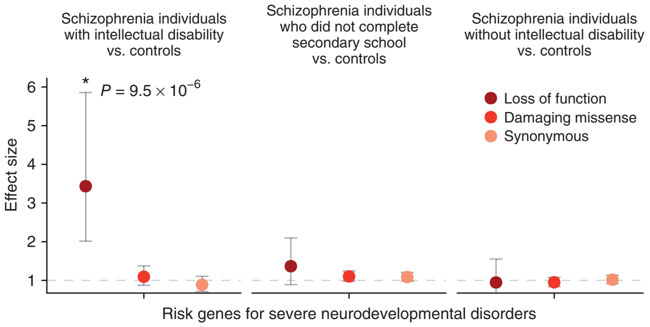

Figure 4 Enrichment of rare loss-of-function variants in genes known to be associated with severe developmental disorders in schizophrenia cases stratified by information on cognitive function as compared to controls. The *P* values shown were calculated using the variant threshold method comparing burden for loss-of-function variants between the corresponding cases and controls. Bars represent the 95% CIs of the point estimates. Damaging missense, missense variants with CADD Phred >15.

When stratifying into these three groups (intellectual disability, no diagnosis of intellectual disability but cognitive impairment not excluded, and no intellectual disability), we observed that the burden of rare, damaging variants in loss-of-function-intolerant genes was significantly greater in the SCZ-ID subgroup than in the remaining schizophrenia cases ($P = 2.6 \times 10^{-4}$; OR = 1.3, 95% CI = 1.12–1.51) or controls ($P < 5 \times 10^{-7}$; OR = 1.61, 95% CI = 1.37–1.89; **Fig. 3**). In the loss-of-function-intolerant gene set, 0.27 (95% CI = 0.20–0.35) extra singleton (defined as having an allele count of one in our data set) loss-of-function variants were observed per exome in SCZ-ID cases as compared to controls, while 0.10 (95% CI = 0.065–0.13) extra singleton loss-of-function variants per exome were observed in the remaining schizophrenia cases as compared to controls (Online Methods). Furthermore, SCZ-ID individuals had significant enrichment of rare loss-of-function variants in developmental disorder–associated genes as compared to the other cases ($P = 9 \times 10^{-4}$; OR = 2.36, 95% CI = 1.41–3.92) or to controls ($P = 9.5 \times 10^{-6}$; OR = 3.43, 95% CI = 2.01–5.86; **Fig. 4**). In comparison to controls, the SCZ-ID individuals carried 0.045 (95% CI = 0.03–0.06) extra singleton loss-of-function variants in developmental disorder–associated genes per exome, suggesting that around 4% of these cases had a loss-of-function variant that is relevant to their clinical presentation. No enrichment in neurodevelopmental disorder–associated genes was observed in patients with schizophrenia who did not have intellectual disability, suggesting that these genes were relevant only for that subset of patients with schizophrenia (**Fig. 4** and **Supplementary Table 4**). Notably, even after excluding known developmental disorder–associated genes from the set of loss-of-function-intolerant genes, we still observed an enrichment of rare variants in SCZ-ID individuals as compared to the remaining cases ($P = 1 \times 10^{-3}$; OR = 1.26, 95% CI = 1.08–1.47) or to controls ($P < 5 \times 10^{-7}$; OR = 1.54, 95% CI = 1.31–1.81; **Supplementary Fig. 9**). Rare variation in these genes contributes more to disease risk in the subset of patients with both schizophrenia and intellectual disability.

Rare variants confer risk for schizophrenia in individuals without intellectual disability

Although rare, damaging variants in loss-of-function-intolerant genes were most enriched in the subset of patients with schizophrenia who also had intellectual disability, we still observed a weaker but significant enrichment in individuals with schizophrenia for whom we could confirm an absence of intellectual disability ($P = 5.5 \times 10^{-4}$; OR = 1.16, 95% CI = 1.05–1.27; **Fig. 3**). Therefore, rare risk variants for schizophrenia follow the pattern previously described in ASD: they are concentrated in individuals with intellectual disability but are not exclusive to that group. To produce a more accurate estimate of the effect of rare, damaging variants on schizophrenia conditional on their effects on overall cognition, we recalculated the enrichment of rare variants in loss-of-function-intolerant genes in a subset of 2,161 schizophrenia cases and 2,398 controls for whom data on years of education were available and for whom intellectual disability could be excluded (**Supplementary Fig. 8**). After controlling for differences in educational attainment (Online Methods), individuals with schizophrenia had a 1.26-fold excess of rare variants in loss-of-function-intolerant genes ($P = 2 \times 10^{-6}$; 95% CI = 1.14–1.38). This increase in our observed odds ratio is consistent with previous accounts that rare, damaging variants also affect educational attainment in controls[40], thus biasing our unconditional estimate.

DISCUSSION

Our integrated analysis of thousands of whole-exome sequences demonstrates that rare, damaging variants increase risk of schizophrenia both with and without comorbid intellectual disability. While the identification of individual genes remains difficult at current sample sizes, we show that the burden of damaging *de novo* mutations and rare SNVs and CNVs in schizophrenia is not scattered across the genome but is primarily concentrated in 3,488 genes intolerant of loss-of-function variants. This observation is shared with ASD, intellectual disability, and severe neurodevelopmental disorders[32,41]. We recapitulate enrichment in previously published gene sets, including transmission at glutamatergic synapses and translational regulation by FMRP, and implicate other gene sets previously linked to ASD, intellectual disability, and severe developmental disorders. However, we find that all of these gene sets share a large number of underlying genes and are especially enriched with the 3,488 genes intolerant of loss-of-function variants. These overlaps among gene sets originating from very different analyses, as well as the subtleties of how they are defined, suggest caution in interpreting biological explanations from observed enrichments.

We jointly analyzed the case–control SNV data with information on cognitive function for 2,971 patients and observed that loss-of-function variants disrupting genes associated with severe developmental disorders are disproportionately found in individuals with schizophrenia with comorbid intellectual disability, with 4% of these cases having a single loss-of-function variant that is relevant to their clinical presentation. Even after excluding variants in known developmental disorder–associated genes, rare variants contribute a greater degree to schizophrenia risk in the SCZ-ID subgroup of patients than in the remaining schizophrenia population. These results show that some of these genetic perturbations have clear manifestations in childhood and that rare risk variants in schizophrenia are particularly associated with comorbid intellectual disability. Our observations are consistent with results in ASD in which rare risk variants are associated with intellectual disability[22,23,28]. Notably, a weaker but still significant rare variant burden was observed in patients with schizophrenia lacking cognitive impairment, and this signal persists even after controlling for educational attainment. Together, these results demonstrate that rare variants have different contributions to schizophrenia risk depending on the degree of cognitive impairment. Notably, these variants do not simply confer risk for a small subset of patients but contribute to disease pathogenesis more broadly.

Our study supports the observation that genetic risk factors for psychiatric and neurodevelopmental disorders do not follow clear diagnostic boundaries. Coding variants disrupting the same genes and, quite possibly, the same biological processes increase risk for a range of phenotypic manifestations. This clinically variable presentation is reminiscent of loss-of-function variants in *SETD1A* and 11 large CNVs from syndromes, previously shown to confer risk for schizophrenia in addition to other prominent developmental defects[16,18]. It is possible that these genes contain an allelic series of variants conferring gradations of risk. A recent schizophrenia GWAS meta-analysis demonstrated that the common variant association signal was similarly enriched in loss-of-function-intolerant genes[42], suggesting that schizophrenia risk genes may be perturbed by common variants of subtle effect and disrupted by rare variants of high penetrance in the population. This possibility is also supported by the overlap in at least some of the pathways affected by both rare and common variation, such as chromatin remodeling. However, the most common deletion in the 22q11.2 locus and a recurrent 2-base deletion in *SETD1A* are associated with both schizophrenia and more severe neurodevelopmental disorders, suggesting that the same variants can also confer risk for a range of clinical features[18,43,44]. Ultimately, it may prove difficult to clearly partition patients genetically into subtypes 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 loss-of-function variants in genes under genic constraint 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 the degree of cognitive impairment, age of onset, and severity[4].

Despite the complex nature of genetic contributions to risk of schizophrenia, it is notable that, across study design (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 ASD, have successfully integrated information across *de novo* SNVs and CNVs to identify novel risk loci[23]. As sample sizes increase, meta-analyses leveraging the shared genetic risk across study designs and variant types, including those we did not consider here, such as classical recessive inheritance, will be similarly well powered to identify additional risk genes in schizophrenia.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully thank all participants in these studies. We thank T. Touloupoulou, M. Picchioni, C. Nosarti, F. Gaughran, and O. Howes for contributing clinical data used in this study. The UK10K project was funded by Wellcome Trust grant WT091310. The INTERVAL sequencing studies are funded by Wellcome Trust grant WT098051. T.S. is supported by the Williams College Dr. Herchel Smith Fellowship. A.P. is supported by Academy of Finland grants 251704 and 286500, NIMH grant U01MH105666, and the Sigrid Juselius Foundation. The work at Cardiff University was funded by Medical Research Council (MRC) Centre (G0801418) and Program (G0800509) grants. P.F.S. gratefully acknowledges support from the Swedish Research Council (Vetenskapsrådet, award D0886501). Creation of the Sweden schizophrenia study data was supported by NIMH grant R01 MH077139 and the Stanley Center of the Broad Institute. Participants in INTERVAL were recruited with the active collaboration of NHS Blood and Transplant England, which has supported fieldwork and other elements of the trial. DNA extraction and genotyping were funded by the National Institute of Health Research (NIHR), the NIHR BioResource, and the NIHR Cambridge Biomedical Research Centre. The academic coordinating center for INTERVAL was supported by core funding from the following: the NIHR Blood and Transplant Research Unit in Donor Health and Genomics, the UK MRC (G0800270), and the British Heart Foundation (SP/09/002). For the CNV analysis, we would like to acknowledge the contribution of data from outside sources: (i) Genetic Architecture of Smoking and Smoking Cessation accessed through dbGaP (study accession phs000404.v1.p1). Funding support for genotyping, which was performed at the Center for Inherited Disease Research (CIDR), was provided by 1 X01 HG005274-01. CIDR is fully funded through a federal contract from the National Institutes of Health to the Johns Hopkins University, contract number HHSN268200782096C. Assistance with genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies (GENEVA) Coordinating Center (U01 HG004446). Funding support for collection of data sets and samples was provided by the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392) and the University of Wisconsin Transdisciplinary Tobacco Use Research Center (P50 DA019706 and P50 CA084724). (ii) High-Density SNP Association Analysis of Melanoma: Case–Control and Outcomes Investigation (dbGaP study accession phs000187.v1.p1). Research support to collect data and develop an application to support this project was provided by 3P50CA093459, 5P50CA097007, 5R01ES011740, and 5R01CA133996. (iii) Genetic Epidemiology of Refractive Error in the KORA Study (dbGaP study accession phs000303.v1.p1). Principal investigators: D. Stambolian (University of Pennsylvania) and H.E. Wichmann (Institut für Humangenetik, Helmholtz Zentrum München; National Eye Institute, National Institutes of Health). Funding was provided by R01 EY020483 from the National Institutes of Health. (iv) WTCCC2 study. Samples were downloaded from EGA (http://www.ebi.ac.uk/ega/) and include samples from the National Blood Donors Cohort (EGAD00000000024) and samples from the 1958 British Birth Cohort (EGAD00000000022). Funding for these projects was provided by the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z), the Wellcome Trust (072894/Z/03/Z, 090532/Z/09/Z, and 075491/Z/04/B), and NIMH grants (MH 41953 and MH083094).

AUTHOR CONTRIBUTIONS

T.S. and J.C.B. conceived and designed the experiments. T.S. performed the statistical analysis. T.S., J.T.R.W., M.J., D.C., J.S., M.T., E.R., and P.F.S. analyzed the data. T.S., J.T.R.W., M.J., J.S., M.T., E.R., C.I., D.B., A.M.M., G.K., D.G., R.M.M., M.D.F., E.B., M.G., C.M.H., P.S., A.P., M.C.O'D., M.J.O., and J.C.B. contributed reagents, materials, or analysis tools. T.S., D.C., M.J.O., and J.C.B. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. van Os, J. & Kapur, S. Schizophrenia. *Lancet* **374**, 635–645 (2009).

2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (American Psychiatric Publishing, 2013).
3. Tandon, R. *et al.* Definition and description of schizophrenia in the DSM-5. *Schizophr. Res.* **150**, 3–10 (2013).
4. Owen, M.J. New approaches to psychiatric diagnostic classification. *Neuron* **84**, 564–571 (2014).
5. Owen, M.J., Sawa, A. & Mortensen, P.B. Schizophrenia. *Lancet* **388**, 86–97 (2016).
6. Howes, O.D. & Kapur, S. The dopamine hypothesis of schizophrenia: version III—the final common pathway. *Schizophr. Bull.* **35**, 549–562 (2009).
7. Pocklington, A.J. *et al.* Novel findings from CNVs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron* **86**, 1203–1214 (2015).
8. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
9. Owen, M.J., O'Donovan, M.C., Thapar, A. & Craddock, N. Neurodevelopmental hypothesis of schizophrenia. *Br. J. Psychiatry* **198**, 173–175 (2011).
10. Rapoport, J.L., Giedd, J.N. & Gogtay, N. Neurodevelopmental model of schizophrenia: update 2012. *Mol. Psychiatry* **17**, 1228–1238 (2012).
11. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
12. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
13. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
14. Kirov, G. *et al. De novo* CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142–153 (2012).
15. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
16. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204**, 108–114 (2014).
17. Zhu, X., Need, A.C., Petrovski, S. & Goldstein, D.B. One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. *Nat. Neurosci.* **17**, 773–781 (2014).
18. Singh, T. *et al.* Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
19. Szatkiewicz, J.P. *et al.* Copy number variation in schizophrenia in Sweden. *Mol. Psychiatry* **19**, 762–773 (2014).
20. Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* **18**, 199–209 (2015).
21. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
22. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552–555 (2016).
23. Sanders, S.J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
24. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
25. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
26. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
27. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biol. Psychiatry* **75**, 378–385 (2014).
28. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
29. Rees, E. *et al.* CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and *SLC1A1* and duplications at 1p36.33 and *CGNL1*. *Hum. Mol. Genet.* **23**, 1669–1676 (2014).
30. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
31. Raychaudhuri, S. *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.* **6**, e1001097 (2010).
32. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
33. Firth, H.V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
34. Deciphering Developmental Disorders Study. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
35. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
36. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* **6**, 6404 (2015).
37. Weyn-Vanhentenryck, S.M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6**, 1139–1152 (2014).
38. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
39. Ascano, M. Jr. *et al.* FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**, 382–386 (2012).
40. Ganna, A. *et al.* Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19**, 1563–1565 (2016).
41. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
42. Pardiñas, A.F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. Preprint at *bioRxiv* http://dx.doi.org/10.1101/068593 (2016).
43. Ben-Shachar, S. *et al.* 22q11.2 distal deletion: a recurrent genomic disorder distinct from DiGeorge syndrome and velocardiofacial syndrome. *Am. J. Hum. Genet.* **82**, 214–221 (2008).
44. Michaelovsky, E. *et al.* Genotype–phenotype correlation in 22q11.2 deletion syndrome. *BMC Med. Genet.* **13**, 122 (2012).

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [2]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [3]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK. [4]University College London Genetics Institute, University College London, London, UK. [5]Centre for Psychiatry, Barts and the London School of Medicine and Dentistry, London, UK. [6]National Institute for Health and Welfare, Helsinki, Finland. [7]Institute of Psychiatry, King's College London, London, UK. [8]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK. [9]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [10]Division of Psychiatry, University College London, London, UK. [11]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [12]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [13]A list of contributors is available from http://www.intervalstudy.org.uk/about-the-study/whos-involved/interval-contributors/. [14]A list of members appears in the **Supplementary Note**. [15]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [16]Program in Medical and Population Genetics and Genetic Analysis Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [17]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [18]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina, USA. Correspondence should be addressed to T.S. (ts14@sanger.ac.uk) or J.C.B. (barrett@sanger.ac.uk).

ONLINE METHODS

Sample collections. Ascertainment, data production, and quality control of the schizophrenia case–control whole-exome sequencing data set have been described in detail in an earlier publication[18]. Briefly, the data set was composed of schizophrenia cases recruited as part of eight collections in the UK10K sequencing project and matched population controls from non-psychiatric arms of the UK10K project, healthy blood donors from the INTERVAL project, and five Finnish population studies. The UK10K data set was combined and analyzed with published data from a Swedish schizophrenia case–control study[35]. Data production, quality control, and analysis of the case–control CNV data set were described in an earlier publication[29]. The schizophrenia cases were recruited as part of the CLOZUK and CardiffCOGS studies and consisted of both individuals with schizophrenia 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, and data were processed and variants called under the same protocol. Sanger-validated *de novo* mutations identified through whole-exome sequencing in seven published studies of schizophrenia parent–proband trios were aggregated and reannotated for enrichment analyses[13,45–50]. A full description of each trio study, including sequencing and capture technology and sample recruitment, was previously provided[18].

Sample and variant quality control. We jointly called each case data set with its nationality-matched controls and excluded samples on the basis of contamination, low coverage, non-European ancestry, and excess relatedness[18]. A number of empirically derived filters were applied at the variant and genotype levels, including filters on GATK VQSR, genotype quality, read depth, allele balance, missingness, and Hardy–Weinberg disequilibrium[18]. After variant filtering, the per-sample transition-to-transversion ratio was ~3.2 across the entire data set, as expected for populations of European ancestry[51]. For the case–control CNV analysis, we similarly excluded samples on the basis of excess relatedness, and only CNVs supported by more than ten probes and greater than 10 kb in size were retained to ensure high-quality calls. All *de novo* mutations in our study had been validated using Sanger sequencing.

We used Ensembl Variant Effect Predictor (VEP) version 75 to annotate all variants (SNVs and CNVs) according to GENCODE v.19 coding transcripts. We defined frameshift, stop-gain, splice acceptor, and splice donor variants as loss of function and missense and initiator codon variants with the recommended CADD Phred score cutoff of greater than 15 as damaging missense variants[52]. A gene was annotated as disrupted by a deletion if part of its coding sequence overlapped the copy number event. We more conservatively defined genes as duplicated only if the entire canonical transcript of the gene overlapped with the duplication event.

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. No genome-wide inflation was observed in burden tests of individual genes[18]. In the curated set of *de novo* mutations, we observed the expected exome-wide number of synonymous mutations given gene mutation rates from previously validated models[24], suggesting that variant calling was generally unbiased across GENCODE v.19 coding genes. Lastly, the case–control CNV data set had previously been analyzed for burden of CNVs affecting individual genes and enrichment analyses in targeted gene sets[7,29].

Case–control enrichment burden tests. For the case–control SNV data set, we performed permutation-based gene set enrichment tests using an extension of the variant threshold method[30]. This method assumed that variants with a 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.*[30], $t(T)$ 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 $<T$. We 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 $<T$ ($X_{\text{in},T}$) while correcting for the total number of damaging alleles across the genome with MAF $<T$ ($X_{\text{all},T}$). $X_{\text{all},T}$ controlled for exome-wide differences between schizophrenia cases and controls, ensuring that any significant gene set result was significant beyond baseline differences. $t(T)$ was defined as the *t* statistic testing whether the regression coefficient of $X_{\text{in},T}$ deviated from 0. We then calculated $t(T)$ for all observed thresholds below a MAF of 0.1% and selected the maximum value for $t_{\max}$ on the basis of the observed data. To calculate a null distribution for $t_{\max}$, we performed 2 million case–control permutations within each population (UK, Finnish, and Swedish) to control for batch and ancestry, and we 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 2 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 across the genome (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% confidence intervals were calculated using only variants observed once in our data set (allele count of 1) to ensure that they were comparable between tested gene sets.

CNV logistic regression. We adapted a logistic regression framework described in Raychaudhuri *et al.* and implemented in PLINK to estimate the case–control differences in the rate of CNVs overlapping a specific gene set while correcting for differences in CNV size and total number of genes disrupted[7,19,31]. We first restricted our analyses to coding deletions and duplications and tested for enrichment using the following model

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

where for individual *i*, p_i is the probability that this individual has 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 control for the genome-wide differences in the rate and size of CNVs between cases and controls, while β_3 captures the true gene set enrichment above this background rate[7,19,31]. For each gene set, we report 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, we first determined the total number of *de novo* mutations observed in the 1,077 schizophrenia trios. We then generated 2 million random samples with the same number of *de novo* mutations, weighting the probability of observing a mutation in a gene by its estimated mutation rate. Baseline gene-specific mutation rates were obtained using the method described in Samocha *et al.*[24] and adapted to produce loss-of-function and damaging missense variant rates for each GENCODE v.19 gene. These mutation rates were adjusted for both sequence context and gene length and were successfully applied in the primary analyses of large-scale exome sequencing of ASD and severe developmental disorders with replicable results[23,32,41]. For each gene set, one-sided enrichment *P* values were calculated as the fraction of the 2 million random samples that had a greater or equal number of *de novo* mutations in the gene set of interest as was observed in the 1,077 trios. 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 2 million random samples. We adapted a method in Fromer *et al.* to calculate 95% credible intervals for the enrichment statistic[13]. We first generated a list of 1,000 evenly spaced values between 0 and 10 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 number observed in the 1,077 trios. We normalized 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.

Combined joint analysis. Gene set *P* values calculated using the case–control SNV, case–control CNV, and *de novo* data were meta-analyzed using Fisher's combined probability method with 6 degrees of freedom to provide a single test statistic for each gene set. We corrected for the number of gene sets tested in the discovery analysis ($n = 1{,}766$) by controlling the FDR using the Benjamini–Hochberg approach and report only results with a *q* value of less than 5%.

Description of gene sets. The full list of tested gene sets is found in **Supplementary Table 1**, and a detailed description is provided in the **Supplementary Note**. Briefly, we tested all gene sets with more than 100 genes from five public pathway databases. We also tested additional gene sets selected on the basis of biological hypotheses about schizophrenia risk and genome-wide screens investigating rare variants in intellectual disability, ASD, and other neurodevelopmental disorders. All gene identifiers were mapped to the GENCODE v.19 release, and all noncoding genes were excluded. A total of 1,766 gene sets were included in our analysis.

Selection of allele frequency thresholds and consequence severity. For the case–control whole-exome data, we applied an extension of the variant threshold model (described above). With this method, we tested damaging variants at a number of frequency thresholds without specifying an a priori MAF cutoff. All thresholds below a MAF of 0.1% observed in our data were tested, and we assessed statistical significance by permutation testing. For all whole-exome data (case–control and trio), we restricted our analyses to loss-of-function variants. These variants have a clear and severe predicted functional consequence in that they putatively cause single-copy loss of a gene. Furthermore, this class of variants has been demonstrated to have the strongest genome-wide enrichment between cases and controls across neurodevelopmental and psychiatric disorders[18,32,41]. When selecting MAF cutoffs for case–control CNVs, we found that, although the bulk of the test statistics were not inflated, the tail of gene set *P* values were dramatically inflated, even when testing for enrichment in the random gene sets (**Supplementary Fig. 1**). This inflation in the tail of the quantile–quantile plot was driven in part by very large (overlapping more than ten genes), more common (MAF between 0.1 and 1%) CNVs observed mainly in cases or controls. Some of these CNVs, such as the known syndromic CNVs, likely harbored true risk genes. However, because these CNVs were highly recurrent in cases and 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 that our model was well calibrated and that its *P* values followed a null distribution for random gene sets, we explored different frequency and size thresholds and conservatively restricted our analysis to copy number events overlapping fewer than seven genes (excluding the largest 10% of CNVs) with MAF <0.1% (**Supplementary Fig. 1**). Our main conclusions remained unchanged even if we selected a more stringent (excluding the largest 15% of CNVs) or less stringent (excluding the largest 5% of CNVs) size threshold.

Robustness of enrichment analyses. We 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, we 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 cutoffs across all variant classes. A quantile–quantile plot was generated using *P* values from enrichment tests of each data set and variant type. Reassuringly, we observed null distributions in all such quantile–quantile plots (**Supplementary Fig. 3**).

Comparison of *de novo* mutation enrichment with that of broader neurodevelopmental disorders. We aggregated and reannotated *de novo* mutations from four studies: 1,113 severe developmental disorder probands[41], 4,038 ASD probands[23,32], and 2,134 control probands[28,32]. We used the Poisson exact test to calculate differences in *de novo* mutation rates in constrained genes between schizophrenia, ASD, and severe developmental disorder cases and controls. Counts in each functional class (synonymous, damaging missense, and loss of function) were tested separately, and the one-sided *P* value, rate ratio, and 95% confidence interval of each comparison are reported and plotted in **Figure 2** and **Supplementary Figures 4** and **5**.

Conditional analyses. In each of the three methods we used for gene set enrichment, we restricted all variants analyzed to those that reside in the background gene list and tested for an excess of rare variants in genes shared by the gene set of interest (*K*) and the background list (*B*). Brain-enriched genes from GTEx and the ExAC loss-of-function-intolerant genes (pLI > 0.9) were used as background (see above). For the case–control SNV data, we 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

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

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, we determined the total number of *de novo* mutations within the background gene list observed in the 1,077 schizophrenia trios and 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 the 2 million random samples that had a greater or equal number of *de novo* mutations in genes in $K \cap B$ as observed in the 1,077 trios. Gene set *P* values were combined using Fisher's method. We restricted our conditional enrichment analysis to gene sets with $q < 5\%$ in the discovery analysis and adjusted for multiple testing using Bonferroni correction ($P = 0.00071$, or 0.05/67 tests; **Supplementary Table 3**).

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 ICD-9). The recruitment guidelines of the MUIR collection were described in detail in a previous publication[53]. In brief, evidence of remedial education was a prerequisite to inclusion, and individuals with premorbid IQs below 50 or above 70, severe learning disabilities, or who were unable to give consent were excluded. The Schizophrenia and Affective Disorders Schedule–Lifetime version (SADS-L) in people with mild learning disability, Positive and Negative Syndrome Scale (PANSS), Research Diagnostic Criteria (RDC), and DSM-III-R, and the St. Louis Criterion were applied to all individuals to ensure that any diagnosis of schizophrenia was robust. Using the clinical information provided alongside the Swedish and Finnish case–control data sets, we identified an additional 182 individuals with schizophrenia who were similarly diagnosed with intellectual disability, for a total of 279 individuals.

Cognitive testing and educational attainment data available for a subset of samples were used to identify individuals with schizophrenia who did not have cognitive impairment. For 502 individuals from the Cardiff collection in the UK10K study, we acquired premorbid IQ as extrapolated from the National Adult Reading Test (NART) and identified 412 individuals for analysis after excluding all individuals with predicted premorbid IQ of less than 85 (or less than 1 s.d. from the population distribution for IQ). We additionally acquired information on educational attainment in 54 individuals with schizophrenia in the UK10K London collection and retained 27 individuals without intellectual disability and who completed at least 12 years of schooling. Lastly, the California Verbal Learning Test (CVLT) was conducted on 124 Finnish individuals with schizophrenia 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, we additionally excluded any individuals who ranked in the lowest decile for CVLT composite score and retained 92 individuals for analysis. According to these criteria, we identified 531 of 697 individuals with schizophrenia from the UK and Finnish data sets with cognitive data as not having intellectual disability. We additionally acquired data on educational attainment for the Swedish schizophrenia cases and controls from the Swedish National Registry. After excluding individuals with intellectual disability, we identified 1,527 individuals with schizophrenia who did not complete secondary school (less than 12 years of schooling) and 634 individuals with schizophrenia who completed at least compulsory and upper secondary schooling (at least 12 years of schooling). The last group with the greatest educational attainment was defined as cases without intellectual disability. In the Swedish sample, 49.4% of control samples had lower educational attainment than the 634 individuals with schizophrenia defined as having no intellectual disability, suggesting that our definition was sufficiently strict. In total, combining the UK, Finnish, and Swedish data, we identified 1,165 individuals with schizophrenia who did not have intellectual disability.

Using the variant threshold method, we tested for differences in rare loss-of-function variant burden between the three case groups (intellectual disability, did not complete secondary school, no intellectual disability) against controls. We restricted these analyses to three gene sets (loss-of-function-intolerant genes, genes in which loss-of-function variants are diagnostic for severe developmental disorders, and loss-of-function-intolerant genes after excluding genes associated with severe developmental disorders) and adjusted for multiple testing using Bonferroni correction ($P = 0.0038$, or 0.05/13 tests). **Supplementary Table 4** enumerates all the statistical tests performed. To estimate the per-exome excess of rare singleton (defined as having an allele count of one in our data set) loss-of-function variants in cases as compared to controls, we regressed X_{in} (the number of loss-of-function variants in the gene set of interest) on case status (0 or 1) while controlling for X_{all} (the total number of loss-of-function variants across the genome) and population (UK, Finnish, and Swedish). The effect size and 95% confidence interval of the regression coefficient of the case status predictor are reported.

Data availability. Sequence data and processed VCFs for the UK10K project were deposited into the European Genome-phenome Archive (EGA) under study accession EGAO00000000079. The processed VCFs from the Swedish case–control study were deposited in dbGaP under accession phs000473. Rare variant counts and gene-level association results from combining the whole-exome sequencing data sets were described in a previous publication[18] and were made available on the PGC results and download page (https://www.med.unc.edu/pgc/results-and-downloads).

45. Guipponi, M. *et al.* Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. *PLoS One* **9**, e112745 (2014).
46. Girard, S.L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat. Genet.* **43**, 860–863 (2011).
47. McCarthy, S.E. *et al. De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol. Psychiatry* **19**, 652–658 (2014).
48. Takata, A. *et al.* Loss-of-function variants in schizophrenia risk and *SETD1A* as a candidate susceptibility gene. *Neuron* **82**, 773–780 (2014).
49. Xu, B. *et al.* Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nat. Genet.* **43**, 864–868 (2011).
50. Xu, B. *et al. De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
51. Do, R. *et al.* Exome sequencing identifies rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015).
52. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
53. Doody, G.A., Johnstone, E.C., Sanderson, T.L., Owens, D.G. & Muir, W.J. 'Pfropfschizophrenie' revisited. Schizophrenia in people with mild learning disability. *Br. J. Psychiatry* **173**, 145–153 (1998).

doi:10.1038/ng.3903