Article

Rare coding variants in ten genes confer substantial risk for schizophrenia

https://doi.org/10.1038/s41586-022-04556-w

Received: 13 August 2020

Accepted: 16 February 2022

Published online: 8 April 2022

Check for updates

Rare coding variation has historically provided the most direct connections between gene function and disease pathogenesis. By meta-analysing the whole exomes of 24,248 schizophrenia cases and 97,322 controls, we implicate ultra-rare coding variants (URVs) in 10 genes as conferring substantial risk for schizophrenia (odds ratios of 3–50, $P < 2.14 \times 10^{-6}$) and 32 genes at a false discovery rate of <5%. These genes have the greatest expression in central nervous system neurons and have diverse molecular functions that include the formation, structure and function of the synapse. The associations of the NMDA (*N*-methyl-D-aspartate) receptor subunit *GRIN2A* and AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor subunit *GRIA3* provide support for dysfunction of the glutamatergic system as a mechanistic hypothesis in the pathogenesis of schizophrenia. We observe an overlap of rare variant risk among schizophrenia, autism spectrum disorders[1], epilepsy and severe neurodevelopmental disorders[2], although different mutation types are implicated in some shared genes. Most genes described here, however, are not implicated in neurodevelopment. We demonstrate that genes prioritized from common variant analyses of schizophrenia are enriched in rare variant risk[3], suggesting that common and rare genetic risk factors converge at least partially on the same underlying pathogenic biological processes. Even after excluding significantly associated genes, schizophrenia cases still carry a substantial excess of URVs, which indicates that more risk genes await discovery using this approach.

Schizophrenia is a severe psychiatric disorder with signs and symptoms that include hallucinations, delusions, disorganized speech and behaviour, diminished emotional expression, social withdrawal and cognitive impairment. This disorder has a lifetime risk of approximately 0.7%, is often disabling and reduces life expectancy by nearly 15 years[4,5]. Existing therapies largely address primarily positive symptoms (hallucinations and delusions), and the response to existing antipsychotic medications is highly variable, with approximately 30% of patients classified as treatment resistant[6]. The lack of progress in therapeutic development is in part a consequence of the limited understanding of the molecular aetiology of psychiatric disorders[6,7].

Schizophrenia is well established as having a substantial genetic component with contributions from across the allele frequency spectrum[8–11]. It was initially theorized that the high heritability, consistency of prevalence across populations and increased risk observed for individuals in more densely affected families suggested that polygenic predisposition should have a dominant role in defining schizophrenia risk in the population[4,12]. This has been borne out by genome-wide association studies (GWAS), which have now, in a companion paper, identified 287 common risk loci (minor allele frequency (MAF) > 1%) of individually small effect (median odds ratio (OR) < 1.05)[3]. As a class of variation, common variants explain approximately 24% of the variance in disease liability[13]. Several rare (MAF < 0.1%), recurrent copy number variants (CNVs) have also been robustly associated with schizophrenia, as exemplified by substantially higher rates of schizophrenia in carriers of 22q11.2 deletions[10,14]. This suggests a role for rare gene-disrupting mutations with much larger effects on individual risk (OR = 2–60). Although the variants implicated thus far have large effects on risk in the individual, they make only a small contribution to overall heritability in the population owing to their rarity. Despite these successes in locus discovery, moving from individual associations to specific genes and disease mechanisms remains challenging. Because causal variants in schizophrenia GWAS are predominantly non-coding, challenges related to fine-mapping and interpretation of intergenic and intronic elements limit the ability to confidently identify underlying genes, infer the mechanism by which they influence disease risk and determine the direction of effect. CNVs of large effect, however, often disrupt hundreds of kilobases of the genome and multiple genes simultaneously, limiting the ability to derive clear functional insights[10].

Analysis of rare coding variants is a powerful complementary approach to identify genes in complex traits. Theory predicts that the forces of natural selection will tend to keep large-effect risk variants at much lower frequencies in the population, particularly in disorders such as schizophrenia that are associated with reduced fecundity[15]. However, most rare variants will have little or no functional consequence or impact on risk, which represents a substantial challenge in identifying those that are truly causal and complicates required analyses in which rare variants are tested as a group rather than individually. The most natural grouping for rare variants is within a gene, on the basis of predicted functional consequence or evidence for deleteriousness[15,16]. Protein-truncating variants (PTVs) are among the most interpretable associations because they suggest that the effect on disease most commonly tracks with decreasing expression of the gene[17]. Earlier schizophrenia sequencing studies have established that ultra-rare and

A list of authors and their affiliations appears online. ✉e-mail: tsingh@broadinstitute.org

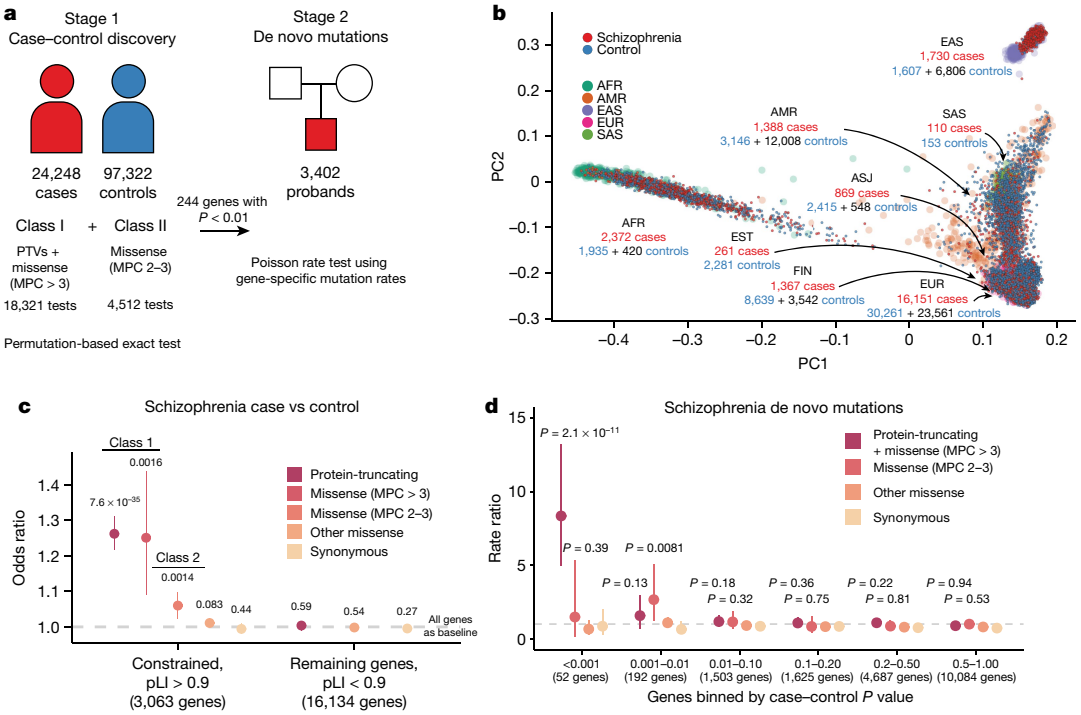


Fig. 1 | Study design and analytic approach. a, Study design. Case–control and parent–proband trio sample sizes, variant classes and analytical methods are described. The case–control stage is shown on the left, and the de novo mutation stage is shown on the right. **b**, Principal-components analysis of SCHEMA samples. 1000 Genomes samples with reported ancestry are plotted in the background, and SCHEMA samples are displayed in the foreground. For each global ancestry group, the numbers of cases and controls in the discovery dataset are in red and blue, respectively, and the number of external controls is in black. AFR, African; ASJ, Ashkenazi Jewish; AMR, Latin American; EAS, East Asian; EST, Estonian; FIN, Finnish; EUR, non-Finnish European; SAS, South Asian. **c**, Case–control enrichment of ultra-rare protein-coding variants in genes intolerant of PTVs (n = 22,444 cases and n = 39,837 controls). The two-sided P values from logistic regression displayed were obtained by comparing the burden of variants of the labelled consequence in cases and controls. By definition, MPC enrichment is shown only for genes with pLI > 0.9. The dot represents the OR and the bar represents the 95% CI of the point estimates. pLI, probability of loss-of-function intolerant in the gnomAD database. **d**, Enrichment of schizophrenia de novo mutations in P-value bins derived from the stage 1 (case–control) gene burden analysis (n = 3,402 schizophrenia trios). The one-sided enrichment P values displayed were calculated as a Poisson probability having equal or greater value than the observed number of mutations given the baseline mutation rate. The relative rate is given by the ratio of the observed to expected rate of de novo mutations. The dot represents the relative rate and the bar represents the 95% CI of the point estimates.

de novo mutations contribute to risk as a category and have prioritized disease-relevant tissues and processes, specifically observing an enrichment in neuronal genes and synaptic processes[9,11,18–22]. Furthermore, these risk alleles are concentrated in genes with a near-complete depletion of PTVs in population studies, a result shared with other neurodevelopmental disorders[9,11] and suggesting strong direct selection against such mutations. However, analysis of ultra-rare coding variants (URVs) has had limited success in delivering individual gene discovery in schizophrenia because of power limitations, with only a single gene, *SETD1A*, identified as robustly associated[15,20].

The Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) consortium was formed as a global collaborative effort to analyse sequence data from many studies to advance gene discovery. Here, we generate, aggregate, harmonize the variant identification of and meta-analyse the exome sequences of 24,248 individuals with schizophrenia and 97,322 controls from seven continental populations. This analysis is, to our knowledge, one of the largest sequencing studies of a complex trait. As predicted by the apparent rare variant burden in schizophrenia, increasing the sample size has led to the identification of ten genes with URVs that confer substantial risk at exome-wide significance. Combining these findings with the results of other large-scale sequencing studies, we find shared and distinct genetic signals for schizophrenia and other neurodevelopmental disorders. In tandem with a companion paper from the Psychiatric Genomics Consortium[3], we provide evidence that common coding variants and URVs identify an overlapping set of genes. Finally, we demonstrate that increased scale following this approach will uncover additional risk genes and help complete the genetic architecture of schizophrenia.

Data description and quality control

We aggregated exome sequence data for 24,248 individuals diagnosed with schizophrenia and 50,437 individuals without a known psychiatric diagnosis recruited in 11 global collections that had previously contributed to common variant association efforts (Supplementary Methods, Fig. 1a, Supplementary Table 1). The sequence data for 7,979 cases have previously been presented in earlier publications[9,11,18–21]; the remaining 16,269 cases are presented here. To ensure calibrated analyses, these samples were included in joint reprocessing and variant calling using a standardized BWA–Picard–GATK pipeline as part of the larger Genome Aggregation Database (gnomAD) effort (Supplementary Methods); SCHEMA case–control samples with appropriate permissions are also included in the gnomAD v2 release[23]. After extracting SCHEMA samples from this callset, we performed quality-control steps to ensure high quality of sequence data, exclude contaminated samples, identify parent–proband trios and other related individuals, and infer global ancestries (Supplementary Methods, Fig. 1b, Supplementary Figs. 1–7, Supplementary Table 2). We subsequently applied site- and genotype-level filters to generate a robust set of coding single-nucleotide polymorphisms (SNPs) and indels for well-matched

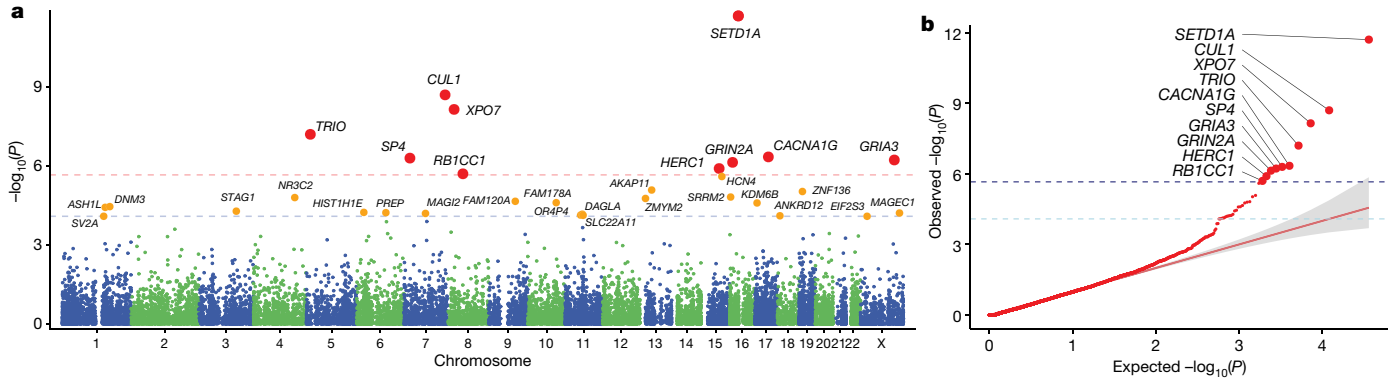


Fig. 2 | Results from meta-analysis of URVs in 3,402 trios, 24,248 cases and 97,322 controls. a, Manhattan plot, with $-\log_{10}$-transformed *P* values plotted against the chromosomal location of each gene. The per-gene *P* values were calculated by meta-analysing the two-sided burden test *P* values from rare coding variants in 24,248 cases and 97,322 controls; the one-sided Poisson rate test *P* values were obtained from de novo mutations in 3,402 trios. (See Supplementary Methods for more information.) Genes reaching exome-wide significance ($P < 2.14 \times 10^{-6}$, corresponding to 0.05 out of 23,321 tests) are shown in red, and those significant at FDR < 5% are shown in orange. Red dashed line, $P = 2.14 \times 10^{-6}$; blue dashed line, FDR < 5%, or $P = 8.23 \times 10^{-5}$. **b**, Quantile–quantile plot. The observed $-\log_{10} P$ values were plotted against the expectation given a uniform distribution. The per-gene *P* values were calculated by meta-analysing the two-sided burden test *P* values from rare coding variants in 24,248 cases and 97,322 controls; the one-sided Poisson rate test *P* values were obtained from de novo mutations in 3,402 trios. (See Supplementary Methods for more information.) Genes reaching exome-wide significance are plotted with a larger size. The direction of effect is indicated by the colour of each point. The grey shaded area indicates the 95% CI under the null. Dark-blue dashed line, $P = 2.14 \times 10^{-6}$; light-blue dashed line, FDR < 5%.

case–control analysis (Supplementary Methods). Previous studies have shown that, in comparison with controls, PTVs were concentrated in 3,063 genes under strong constraint in schizophrenia cases[11,24]. We replicated this result with consistent signals across our major cohorts ($P_{meta} = 7.6 \times 10^{-35}$; OR = 1.26, 95% confidence interval (CI) = 1.22–1.31) (Fig. 1c, Extended Data Fig. 1).

Analysis approach

To increase the power for gene discovery, we incorporated variant counts from additional samples from non-psychiatric and non-neurological collections that were aggregated as part of the gnomAD consortium effort (Supplementary Methods)[23]. We attempted to control technical and methodological batch effects that might arise from this approach in both variant calling and the permutation testing described below. All samples in the gnomAD and SCHEMA consortia were reprocessed and jointly called using the same pipeline, and the same variant filters were applied to achieve high-quality calls. Notably, we restricted our analysis to coding exons with high-quality data across all major exome capture technologies to reduce any artefacts that might arise from coverage differences (Supplementary Methods, Supplementary Figs. 1, 2). After incorporating variant counts from 46,885 additional gnomAD controls, our combined discovery dataset was composed of 24,248 cases and 97,322 population controls (Fig. 1a, b, Supplementary Table 3).

Because only summary-level variant counts were available for the 46,885 external controls, we tested for an excess of disruptive variants per gene using Fisher's exact test, in which the statistical significance was determined by case–control permutations within each stratum (Supplementary Methods, Supplementary Table 3). As in other sequencing studies, we enriched the search for pathogenic variants by restricting our analysis to ultra-rare variants with a minor allele count (MAC) of ≤5 that were either PTVs, defined as stop-gained, frameshift, or essential splice donor or acceptor variants, or damaging missense variants, defined by the MPC pathogenicity score[1,25] (Supplementary Methods). We found that missense variants with an MPC score of >3 had a global signal similar to those of PTVs in schizophrenia, autism spectrum disorders (ASD) and severe neurodevelopmental disorders; however, variants with an MPC score of 2–3 had a significant but weaker signal than PTVs and were therefore analysed separately (Fig. 1c, Extended Data Figs. 2, 3, Supplementary Fig. 8, Supplementary Table 4, Supplementary Methods). Motivated by these observations, we performed a burden test of PTVs and variants with an MPC score of >3 (class I) to generate a *P* value for 18,321 protein-coding genes (Supplementary Methods). For the 4,512 genes with variants with an MPC score of 2–3 (class II), we performed an additional test aggregating these variants and meta-analysed these gene statistics with class I *P* values using a weighted *Z*-score method (Supplementary Methods). To ensure the robustness of the results generated by this approach, we observed the expected null distribution of *P* values in gene-based tests of synonymous variants in each stratum and in the meta-analysis (Supplementary Figs. 9, 10). In addition, we observed no inflation of synonymous *P* values using the Mantel–Haenszel test, even after limiting our analysis to genes with larger total numbers of alleles (gene-wide MAC > 10, 50 or 100), where we had greater power to detect potential artefacts (Supplementary Figs. 11, 12).

Previous studies integrated case–control and trio-based de novo mutations for gene discovery[1,20]. To this end, we aggregated and re-annotated de novo mutations from 3,402 published parent–proband trios (Supplementary Methods). Despite the sizable number of trios, few de novo mutations were available for analysis, with only 325 genes having at least one de novo PTV and only 449 genes having at least one class I or class II mutation. Using Poisson rate tests based on the expected mutation rate[26], we found that these de novo mutations were enriched for the 244 genes with $P < 0.01$ in our case–control analysis (Supplementary Fig. 13, Supplementary Table 5); limited or no signal was found in the remaining genes in the genome (Fig. 1d). The most notable enrichment was observed for the 52 genes with $P < 0.001$ in the case–control analysis (class I mutations: $P = 2.1 \times 10^{-11}$; OR = 8.3, 95% CI = 4.9–13), which provides additional reassurance of the robustness of our case–control gene results. Motivated by these observations, we calculated de novo class I and class II *P* values in the 244 genes with $P_{case-control} < 0.01$ using the Poisson rate test and meta-analysed them with our case–control test statistics using a weighted *Z*-score method to increase the power (Supplementary Methods, Supplementary Figs. 13–15).

Individual genes implicated by URVs

Combined, our meta-analysis of 24,248 cases, 97,322 controls and de novo mutations from 3,402 trios implicates ten genes in which URVs

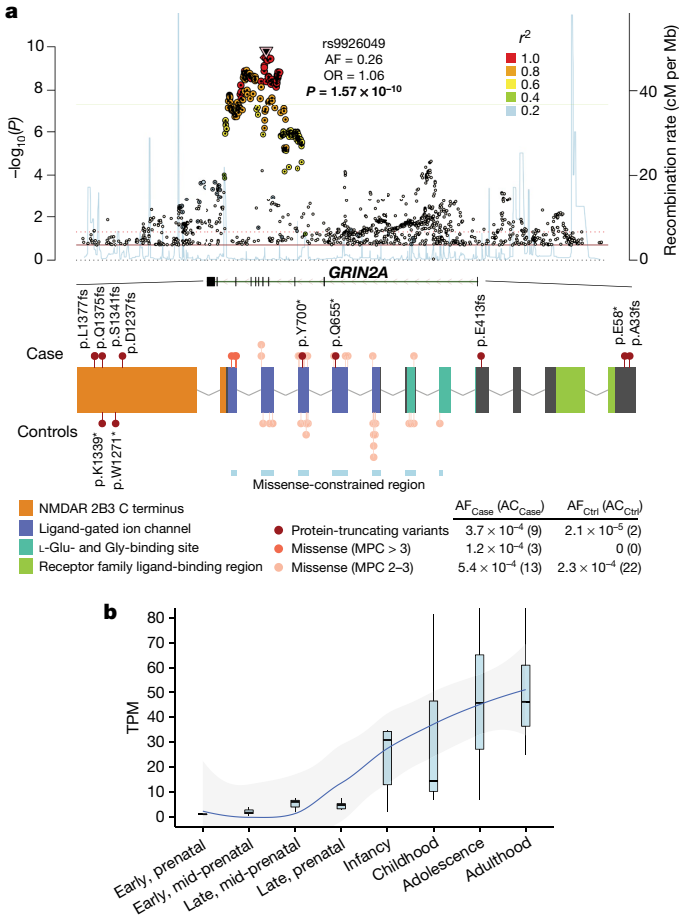


Fig. 3 | Biological insights from exome sequence data. a, Common and rare allelic series at NMDA receptor (NMDAR) subunit *GRIN2A*. Top, the LocusZoom plot displays the common variant (GWAS) association of the gene[3]. The two-sided *P* values of each SNP from the GWAS meta-analysis are shown along the *y* axis. The colour of each dot corresponds to the linkage disequilibrium with the index SNP, and the properties of the index SNP are displayed. Bottom, the gene plot displays the protein-coding variants that contribute to the exome signal in *GRIN2A*. The variants discovered in cases are plotted above the gene, and those from controls are plotted below the gene. Each variant is coloured according to the inferred consequence; the protein domains and missense-constrained regions of the gene are also labelled[25,41]. The frequencies and counts in cases and controls are displayed for each variant class. AF, allele frequency; AC, allele count. **b**, Temporal expression of *GRIN2A* in the human brain ($n = 42$ samples). *GRIN2A* expression is shown in four prenatal and four postnatal periods derived from whole-brain tissue in BrainSpan[28]. The expression values plotted are in transcripts per million (TPM). In the box plot, the lower hinge is the 25% quantile, the middle line is the median, the upper hinge is the 75% quantile, the lower whisker extends to the smallest observation greater than or equal to the lower hinge − 1.5 × interquartile range (IQR) and the upper whisker extends to the largest observation less than or equal to the upper hinge + 1.5 × IQR.

were significantly associated with schizophrenia ($P < 2.14 \times 10^{-6}$, corresponding to 0.05 out of 23,321 tests) (Fig. 2a, b). These top associations as a group are supported by complementary types of variation that include case–control PTVs, damaging missense variants and de novo mutations (Extended Data Table 1 and Supplementary Table 5). Although the CIs were wide, the URVs in these genes appear to confer substantial risk, with ORs for PTVs and class I variants ranging from 3 to 50. As expected, all ten genes were among the most constrained genes in the genome, with a substantial depletion of PTVs compared with the chance expectation[23]. The annotated functions of these genes were diverse and included ion transport (*CACNA1G*, *GRIN2A* and *GRIA3*), neuronal migration and growth (*TRIO*), transcriptional regulation (*SP4*, *RB1CC1* and *SETD1A*), nuclear transport (*XPO7*) and ubiquitin ligation (*CUL1* and *HERC1*). We include a brief discussion of the known biological functions of these genes in the Supplementary Note. Beyond these 10 genes, we identified 22 additional genes at a false discovery rate (FDR) < 5% (Fig. 2a and Supplementary Table 5). We observed notable deviation at the tail of the distribution beyond the associated genes, which suggests that more genes remain to be discovered (Fig. 2b). We report all high-quality variants, relevant annotations and gene-level results on a public browser at https://schema.broadinstitute.org.

The identification of individual genes provides support for more specific mechanistic hypotheses underlying schizophrenia pathogenesis. Developed from neuropharmacological and neuropathological observations, the glutamatergic hypothesis postulates that hypofunction of glutamatergic signalling through NMDA receptors is a possible mechanism of disease[27] (Supplementary Note). Here, we found that PTVs and damaging missense variants in NMDA receptor subunit *GRIN2A* confer substantial risk for schizophrenia ($P = 7.37 \times 10^{-7}$; class I (PTVs and MPC > 3): OR = 24.1, 95% CI = 5.36–221; class II (MPC 2–3): OR = 2.37, 95% CI = 1.1–4.92). Schizophrenia GWAS have also identified a common variant at *GRIN2A* ($P = 1.57 \times 10^{-10}$; OR = 1.057), providing an allelic series in which different perturbations of gene function result in severity of disease risk (Fig. 3a)[8]. The NMDA receptor changes in composition during prenatal to postnatal neurodevelopment, with *GRIN2A* predominantly expressed during late childhood and adolescence, which recapitulates the expected epidemiological observations on schizophrenia age of onset (Fig. 3b, Supplementary Methods)[28]. We also found that risk URVs in AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor subunit *GRIA3* confer substantial risk ($P = 5.98 \times 10^{-7}$; class I (PTVs and MPC > 3): OR = 20.1, 95% CI = 4.28–188) (Extended Data Table 1). Combined, our results from exome sequencing support dysregulation of the glutamatergic system as a mechanistic hypothesis for the development of schizophrenia. Moreover, the specific identification of genes by coding variation may provide an understanding of disease pathogenesis.

Shared genes with GWAS loci

Pathway analyses of common variants have prioritized disease-relevant tissues and cell types and, in some cases, have independently recapitulated known biology[8,29,30]. To derive insights from global patterns of rare coding variants, we tested for an excess burden of URVs in schizophrenia cases compared with controls in 1,732 broadly defined gene sets from databases of biological pathways (that is, Gene Ontology, Reactome and KEGG) and experimental data (Supplementary Methods)[11]. We observed significant enrichment of URVs in 33 gene sets ($P < 2.9 \times 10^{-5}$), which recapitulated consistent and overlapping cellular compartments and biological processes, including definitions of the postsynaptic density (human cortex biopsy postsynaptic density, $P = 1.2 \times 10^{-12}$), chromatin modification (GO:0016568, $P = 1.8 \times 10^{-12}$), regulation of ion transmembrane transport (GO:0034765, $P = 6.7 \times 10^{-7}$), axon guidance ($P = 5.4 \times 10^{-6}$), voltage-gated cation channel activity (GO:0022843, $P = 8.1 \times 10^{-6}$) and synaptic transmission (GO:0007268, $P = 1.79 \times 10^{-5}$) (Supplementary Table 6, Supplementary Fig. 16). Because of the clear synaptic signal, we performed further investigations in the refined synaptic ontology defined by the SynGO consortium[31] and found consistent enrichment for postsynaptic components and processes (GO:0098794, $P = 3.9 \times 10^{-6}$) (Supplementary Table 7). These global observations are consistent with the known functions of the individual risk genes now implicated by rare variation (Supplementary Note). Consistent with earlier reports studying heritability enrichment in Genotype-Tissue Expression (GTEx) tissues[8,30], we found that genes with the highest specific expression in brain regions showed the strongest enrichment of risk URVs, most significantly in the human frontal cortex ($P = 1.63 \times 10^{-8}$) and with limited signals in the other tissue

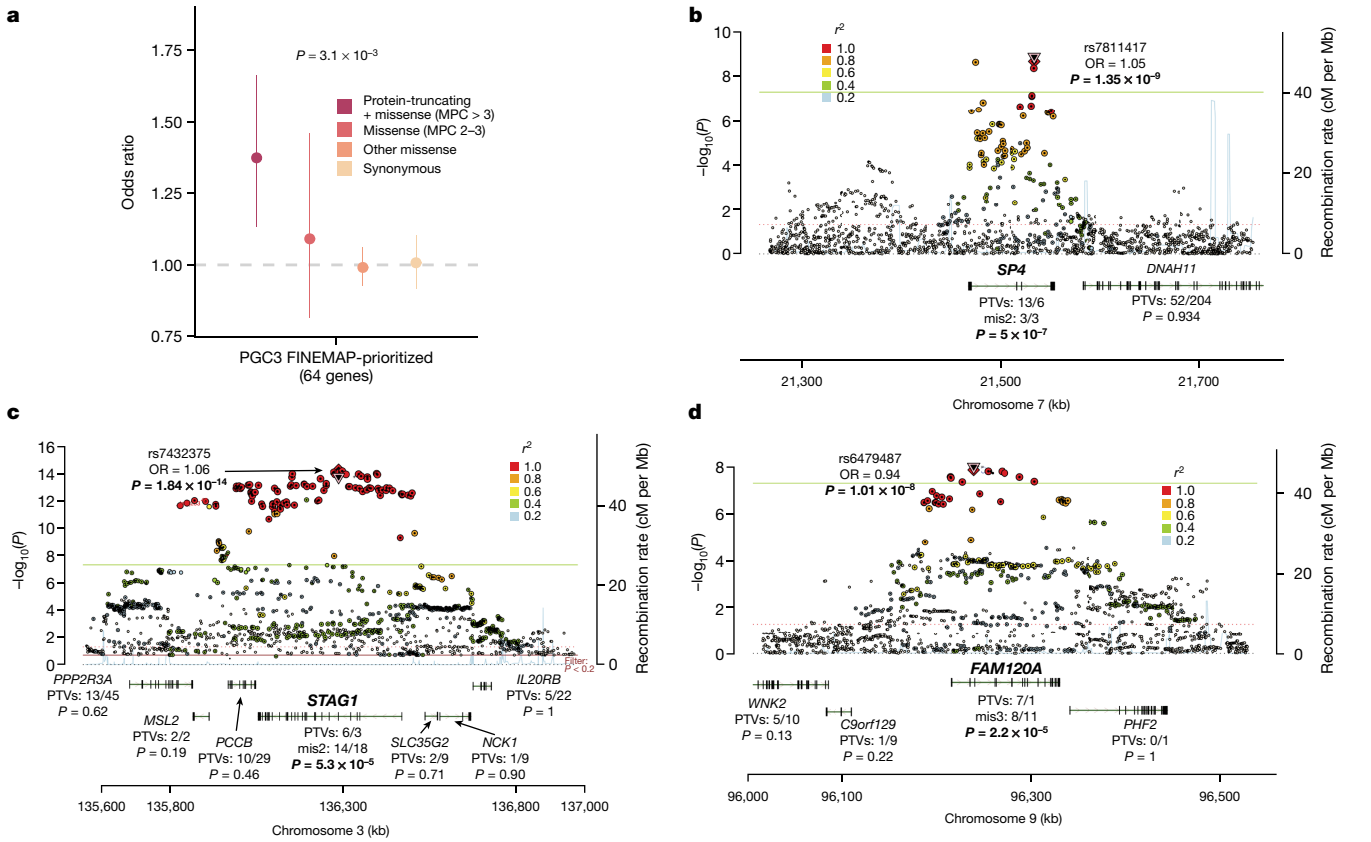


Fig. 4 | Shared genetic signal with schizophrenia GWAS. a, Case–control enrichment of URVs in genes prioritized from fine-mapping of the PGC schizophrenia GWAS (n = 22,444 cases and n = 39,837 controls)[3]. The reported P value was obtained by applying Fisher's combined probability method on the two-sided P values of class I and class II variants. The dot represents the OR and the bar represents the 95% CI of the point estimates. **b**–**d**, Prioritization of GWAS loci using exome data. The LocusZoom plots for three GWAS loci are displayed. The two-sided P values of each SNP from the GWAS meta-analysis are shown along the y axis. Below, for each gene in or adjacent to the region, we show the case–control counts of PTVs in the exome data, along with the two-sided burden test meta-analysis P values. *SP4*, *STAG1* and *FAM120A* are highlighted as the only genes with notable signals in the exome data within each locus.

types (Extended Data Fig. 4, Supplementary Tables 8, 9). To further deconvolute this signal, we investigated which single cell types in the mouse nervous system showed the highest specific expression for the 32 (FDR < 5%) schizophrenia risk genes (Supplementary Methods)[32,33]. Here, we found widespread enrichments across central nervous system neurons, with limited to no signal in glial cells and peripheral nervous system neurons (Supplementary Table 10, Supplementary Fig. 17). Thus, at a high level, global analysis of ultra-rare protein-coding variation independently recapitulates known biology related to schizophrenia pathogenesis, including processes, cellular components and tissues previously implicated by common variant analyses.

To evaluate the overlap of schizophrenia associations from common variant and URV analyses, we jointly analysed our results with the largest GWAS of schizophrenia thus far, which identified common variant associations at 287 distinct loci from the analysis of up to 76,755 cases and 243,649 controls[3]. Statistical fine-mapping prioritized the likely underlying protein-coding gene at 64 of these associations (Supplementary Table 11, Supplementary Fig. 18), and we found case–control enrichment of URVs in these genes ($P_{meta} = 3.1 \times 10^{-3}$; class I: OR = 1.37, 95% CI = 1.13–1.67) (Fig. 4a, Supplementary Table 12). Beyond the statistical enrichment, *GRIN2A* and *SP4*, two of the ten significant rare variant genes, had clear associations in the schizophrenia GWAS (Figs. 3a and 4b). Furthermore, *FAM120A* and *STAG1* resided in more complex GWAS-associated regions containing multiple genes but were prioritized among their neighbours as having FDR < 5% in our sequencing study (Fig. 4c, d). Combined, these results suggest at least partial convergence in the genes and biological processes implicated by common and ultra-rare genetic variation and that URVs can be leveraged to prioritize genes within GWAS loci.

Shared genes with neurodevelopmental disorders

Exome sequencing studies of ASD and severe neurodevelopmental disorders, including developmental delay and intellectual disabilities (DD/ID), have leveraged URVs to identify risk genes. Such studies have established that the genetic signals were concentrated in constrained genes and shared between these disorders[34,35]. More recent analysis of de novo mutations from 31,058 DD/ID trios has implicated 299 genes and that of 11,986 ASD cases has identified 102 genes at FDR < 10% (Supplementary Table 11)[1,2]. We found a significant excess of URVs in schizophrenia cases compared with controls in 299 DD/ID-associated genes ($P_{meta} = 1.5 \times 10^{-14}$; class I: OR = 1.44, 95% CI = 1.3–1.6) and 102 ASD-associated genes ($P_{meta} = 3.7 \times 10^{-7}$; class I: OR = 1.45, 95% CI = 1.23–1.72) (Fig. 5a, Supplementary Table 12). Thus, some schizophrenia rare variant risk appears to be shared with other neurodevelopmental disorders.

With 31,058 trios, the scale of gene discovery in severe DD/ID provided sufficient power to evaluate the individual schizophrenia risk genes associated in our study for a role in broader neurodevelopmental disorders. Nine of the ten schizophrenia-associated genes showed limited de novo PTV signal in DD/ID, with a total of eight de novo PTVs observed in these genes ($X_{exp} = 4.98$; $P_{Poisson} = 0.13$) (Fig. 5b,

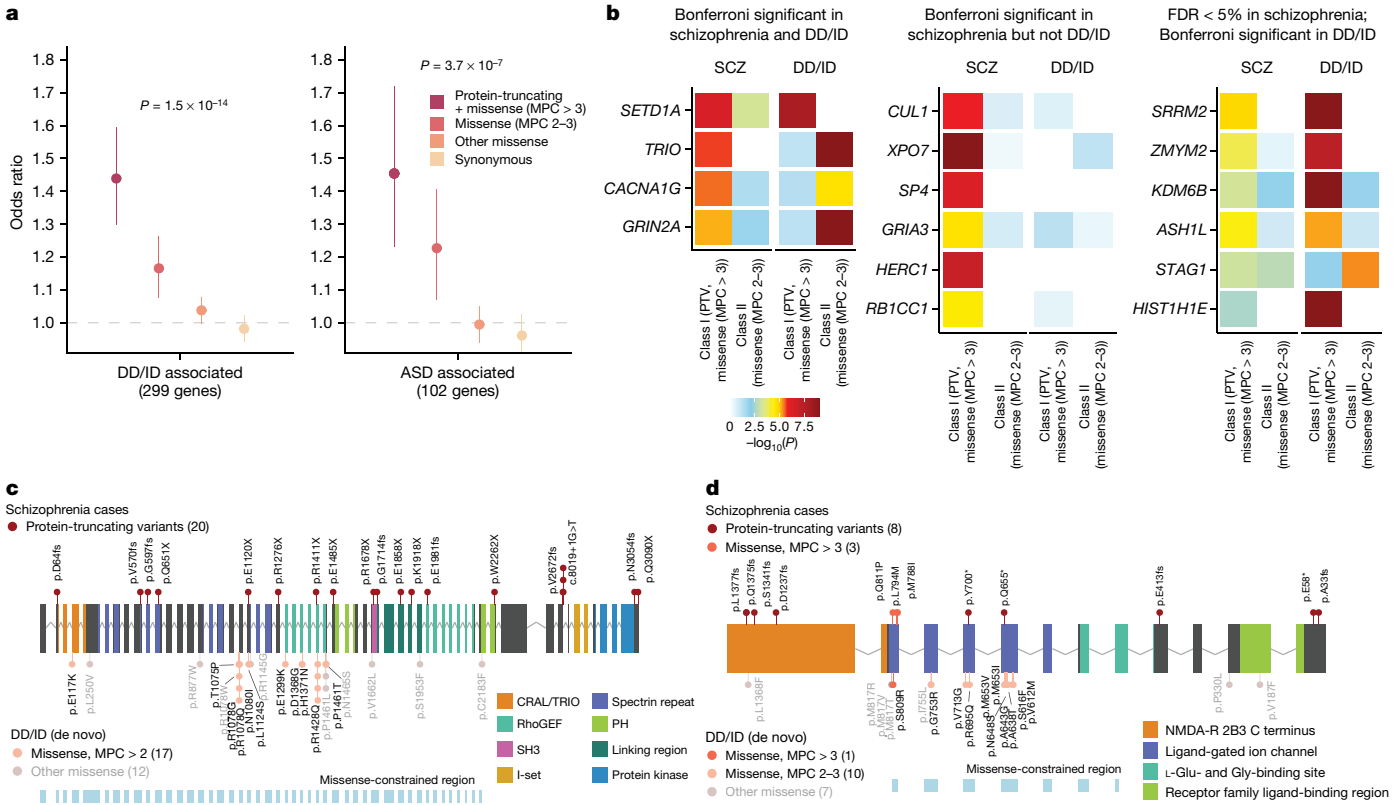


Fig. 5 | Shared genetic signal between schizophrenia and other neurodevelopmental disorders. a, Case–control enrichment of URVs in DD/ID- and ASD-associated genes (n = 22,444 schizophrenia cases and n = 39,837 controls). We tested for the burden of schizophrenia URVs in genes identified in the most recent exome sequencing studies of ASD and DD/ID[1,2]. The reported P value was obtained by applying Fisher's combined probability method on the two-sided P values of class I and class II variants. The dot represents the OR and the bar represents the 95% CIs of the point estimates. **b**, Heatmap displaying the strength of association for schizophrenia-associated genes in our discovery dataset and in genes implicated by de novo mutations in trios diagnosed with DD/ID. We display three groups of genes: genes Bonferroni significant in schizophrenia and DD/ID, genes Bonferroni significant only in schizophrenia, and genes with FDR < 5% in schizophrenia and Bonferroni significant in DD/ID. The degree of association from each sequencing study is displayed as the colour corresponding to $-\log_{10} P$ values in that study. The two-sided case–control burden test P value is reported for schizophrenia; the one-sided P value from de novo enrichment using the Poisson rate test is reported for DD/ID. The results were further stratified to tests of class I (PTVs and missense variants with MPC > 3) and class II (missense variants with MPC of 2–3) variants. **c**, Allelic series in *TRIO* between schizophrenia and DD/ID risk variants. The gene plot displays the protein-coding variants that contribute to the exome signal in *TRIO*. Variants discovered in schizophrenia cases are plotted above the gene, and missense de novo mutations from DD/ID probands are plotted below the gene. Each variant is coloured according to the inferred consequence; the protein domains of the gene are also labelled. The variant counts are displayed for each variant class. **d**, Allelic series in *GRIN2A*, displayed as in **c**.

Supplementary Table 13). *SETD1A* had a significant de novo PTV signal in DD/ID (X_{obs} = 8, X_{exp} = 0.41; $P = 1.3 \times 10^{-8}$), supporting an earlier report that described *SETD1A* as a gene associated with both schizophrenia and broader neurodevelopmental disorders[20]. We also observed a missense signal in *SETD1A* in our study (Extended Data Table 1, Supplementary Fig. 19). Extending this analysis to the additional 22 genes with FDR < 5%, we found that 6 genes (*STAG1*, *ASH1L*, *ZMYM2*, *KDM6B*, *SRRM2* and *HIST1H1E*) were significantly associated with DD/ID in addition to schizophrenia (Fig. 5b, Supplementary Table 13). Among these genes with FDR < 5%, *ASH1L*, *KDM6B* and *NR3C2* were associated with ASD[1] (Supplementary Table 13). Broadly speaking, although PTV mutations in certain genes are joint risk factors for schizophrenia and DD/ID, the majority of schizophrenia associations reported here appeared to have little or no role in DD/ID, despite the enormous power of the DD/ID studies published thus far.

Notably, three of the ten risk genes for schizophrenia (*TRIO*, *GRIN2A* and *CACNA1G*) were associated with risk of severe DD/ID exclusively through de novo missense mutations that cluster within each gene (Fig. 5b, Supplementary Table 13), whereas the schizophrenia signal was largely driven by PTVs. De novo missense mutations in *TRIO* significantly disrupted the exons preceding or encoding the RhoGEF domain (Fig. 5c)[2,36], and those in *GRIN2A* clustered at the base of the ion channel, with the most mutations in the exon encoding the pore of the complex (Fig. 5d). *STAG1*, which had a common and rare variant signal in schizophrenia (Fig. 4d), was associated with DD/ID primarily through de novo missense mutations (Fig. 5b, Supplementary Table 13). These observations suggest that, although schizophrenia and childhood-onset neurodevelopmental disorders share some genes and biological processes, the severity or the nature of the functional impairment differs between disorders, at least in some cases.

We explored what properties may differ between risk genes associated with schizophrenia and DD/ID, proposing that DD/ID genes are under stronger evolutionary constraint with a bias towards prenatal expression when compared with schizophrenia genes. Although schizophrenia genes (FDR < 5%) were under substantial genic constraint compared with the expectation (Mann–Whitney U test, $P = 2.9 \times 10^{-7}$) (Supplementary Fig. 20, Supplementary Methods), they were significantly less constrained than DD/ID-associated genes (Mann–Whitney U test, $P = 3.5 \times 10^{-5}$). Furthermore, schizophrenia genes as a group did not show prenatal or postnatal bias in brain expression (P = 0.21) (Supplementary Fig. 21), whereas DD/ID-associated genes were overwhelmingly prenatal in expression ($P = 7.5 \times 10^{-20}$). Indeed, individual

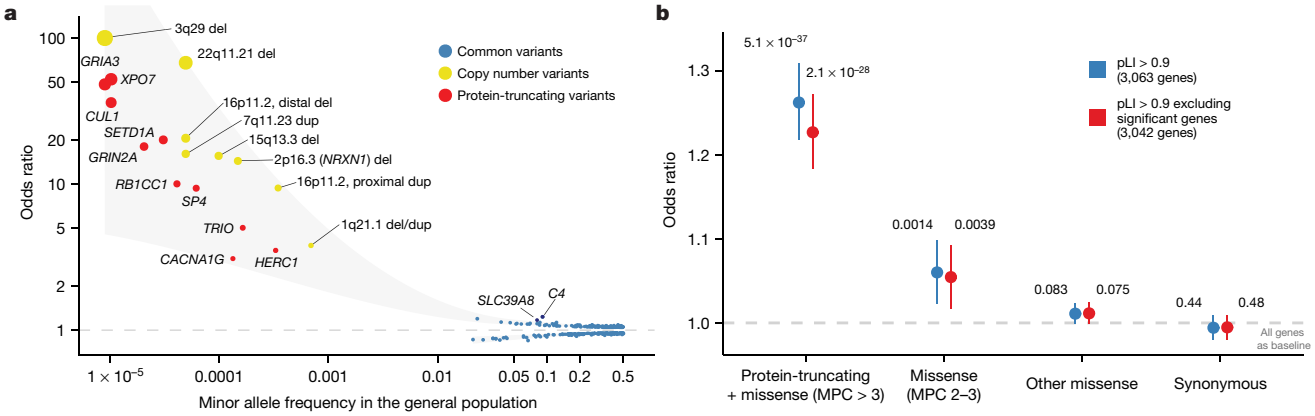


Fig. 6 | Contributions of ultra-rare PTVs to schizophrenia risk. a, Genetic architecture of schizophrenia. Significant genetic associations for schizophrenia from the most recent GWAS, CNV and sequencing studies are displayed. The in-sample OR is plotted against the MAF in the general population. To arrive at bounded odds ratios for visualization, 0.5 was added to all cells of the relevant contingency table that was used to calculate odds ratios if any cell is equal to zero. This modifies the *GRIA3* data point in Extended Data Table 1 to OR = 48.2, 2.7–862.5 95% CI. The colour (size) of each dot corresponds to the source of the association (OR). The shaded area represents the locally weighted scatterplot-smoothed lines of the upper and lower bounds of the point estimates. **b**, Case–control enrichment of URVs in genes intolerant of PTVs after excluding schizophrenia-associated genes (n = 22,444 cases and n = 39,837 controls). We performed the test with all constrained genes (pLI > 0.9) and after excluding all schizophrenia-associated genes with FDR < 5%. The two-sided P values from logistic regression were obtained by comparing the burden of variants of the labelled consequence in the cases compared with controls. The dot represents the OR and the bar represents the 95% CI of the point estimates.

genes such as *SETD1A*, *TRIO* and *SP4* exhibited prenatal expression, whereas *GRIN2A* and *GRIA3* showed postnatal expression (Supplementary Fig. 22). These observations offer the possibility that certain properties may differentiate genes associated with adult psychiatric disorders and more severe DD/ID.

Contribution of ultra-rare PTVs to risk

Efforts from over the past decade are beginning to generate a more comprehensive view of the genetic architecture for schizophrenia, which includes common variants of small effect, large CNVs with elevated frequencies driven by genomic instability and URVs of large effect implicating individual genes (Fig. 6a)[8,10]. Because schizophrenia as a trait is under strong selection[37–39], we expect URVs of large effect to be frequently de novo or of very recent origin and that they contribute to risk in only a fraction of diagnosed patients. We quantified the contribution of PTVs to risk first in our full schizophrenia dataset and then partitioned the de novo and inherited contributions in 2,304 parent–proband trios. We restricted these analyses to the 3,063 PTV-intolerant (pLI > 0.9) genes in which the schizophrenia risk URVs were concentrated. We observed 0.057 (95% CI = 0.049–0.065) extra singleton PTVs per individual in cases compared with controls, suggesting that apparently 5.7% of cases carried a PTV relevant to disease risk. In the 2,304 trios, 0.0394 (95% CI = 0.014–0.065, or 74%) extra singleton PTVs were inherited per proband and 0.0121 (95% CI = 0.0022–0.02, or 26%) extra de novo PTVs in constrained genes were identified in cases compared with controls. By contrast, DD/ID probands had 0.111 (95% CI = 0.103–0.119) extra de novo PTVs in constrained genes, whereas individuals with ASD had 0.0478 (95% CI = 0.0387–0.0568) extra de novo PTVs (Supplementary Fig. 23, Supplementary Methods). In the ten schizophrenia-associated genes, 7 de novo mutations and 13 transmitted variants were observed in 2,304 trios, suggesting that 0.86% of patients are carriers and approximately 35% of variants are de novo. Finally, the genome-wide signal of PTVs in constrained genes (pLI > 0.9; OR = 1.26, $P = 7.6 \times 10^{-35}$) remained significant even after excluding the 32 genes with FDR < 5% (OR = 1.23, $P = 4.3 \times 10^{-27}$) (Fig. 6b, Supplementary Table 4). These findings reaffirm the genetic heterogeneity underlying schizophrenia risk and suggest that the majority of schizophrenia risk genes in which rare variants confer risk remain to be discovered.

Discussion

In one of the largest exome sequencing studies thus far, we identify genes in which disruptive coding variants confer substantial risk for schizophrenia at exome-wide significance. This effort required the reprocessing of a decade of sequence data, harmonization of variant calling and quality control, inclusion of external controls and integration of PTV, damaging missense and de novo variants. Global collaborative efforts such as these provide a template for tackling the genetic contributions in other complex diseases.

The genome-wide analyses recapitulated known biological processes and reaffirmed that schizophrenia risk genes are involved in the postsynaptic density and broader synaptic function and are enriched in expression in neuronal tissues. Furthermore, the identification of specific genes supports more specific mechanistic hypotheses. The association of PTVs in the NMDA receptor subunit *GRIN2A* with schizophrenia risk provides genetic support for the dysregulation of glutamatergic signalling as a possible mechanism of disease. A natural dose–response curve occurs at this gene, in which common regulatory variants modestly influence disease risk while PTVs and predicted damaging missense variants increase risk more substantially. Interestingly, the NMDA receptor is composed of two GRIN2 units (GRIN2A and GRIN2B) and two constitutive GRIN1 units. *GRIN2A* expression increases markedly later in childhood and adolescence, corresponding to the age of onset of disease for schizophrenia. De novo mutations in *GRIN2B* are conversely associated with more severe disorders of neurodevelopment that manifest in childhood, including intellectual disability and autism[40]. Such findings provide a unique opportunity to identify experiments of nature, which help to build and support mechanistic hypotheses that may lead to a better understanding of disease biology.

Joint analysis with genetic data from DD/ID and ASD consortia has provided evidence for shared genes between neuropsychiatric and broader neurodevelopmental disorders. Indeed, 7 of the 32 genes with FDR < 5% in schizophrenia are also associated with DD/ID, providing additional confidence in these associations. The shared genes suggest at least some contribution from early brain developmental processes that predispose to schizophrenia. Despite this overlap, PTVs in nine of the ten genes most confidently associated with schizophrenia are not associated with DD/ID, which may facilitate the identification

of disease-specific processes. Further, we observed allelic series in *GRIN2A*, *TRIO* and *CACNA1G*, in which PTVs increased the schizophrenia risk and de novo missense mutations conferred strong DD/ID risk. De novo missense mutations in these genes clustered in specific domains and are associated with more severe neurodevelopmental, syndromic disorders with cognitive impairment, suggesting an alternate or gain-of-function effect. Analyses estimating relative penetrance for different phenotypes will increase in power as consortium efforts studying specific diseases and biobank efforts continue, the results of which would be useful in distinguishing what is shared or distinct across disorders.

We show here that common regulatory variants from GWAS and URVs disrupt an overlapping set of genes, including an allelic series in four genes in which common variants and rare coding variants increase the risk to varying degrees. Combined, these results suggest that exome sequencing identifies common, shared underlying biology that is dysregulated across the allele frequency spectrum, rather than syndromic forms of disease with unrelated biology regulated by common variation. Furthermore, because of this sharing, coding variants can help refine and fine-map common variant associations such as those at the *STAG1* and *FAM120A* loci. As common and rare variant association studies continue to grow in size, we can better determine the actual degree of overlap of genes that are regulated by both types of variation. Ultimately, the emerging evidence of an overlap between common and ultra-rare variation gives confidence that the integration of results from sequencing consortia with GWAS efforts will have substantial value for identifying specific genes beyond any single strategy.

A decade of genotyping and sequencing studies now establishes specific genetic contributions from common variants, CNVs and URVs as conferring risk for schizophrenia. Despite this progress, it is clear that we are still in the early stages of gene discovery[3]. The vast majority of risk alleles together with their direction and magnitude of effect, mode of action and responsible genes are yet to be discovered. These emerging genetic findings will serve in part to direct and motivate mechanistic studies that begin to unravel disease biology. The success of common variant association studies, and now exome sequencing, suggests concrete progress towards understanding the causes of human complex traits and diseases and provides a clear roadmap for understanding the genetic architecture of schizophrenia.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04556-w.

1. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584 (2020).
2. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
3. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* https://doi.org/10.1038/s41586-022-04434-5 (2022).
4. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30**, 67–76 (2008).
5. Hjorthøj, C., Stürup, A. E., McGrath, J. J. & Nordentoft, M. Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. *Lancet Psychiat.* **4**, 295–301 (2017).
6. Lehman, A. F. et al. Practice guideline for the treatment of patients with schizophrenia, second edition. *Am. J. Psychiat.* **161**, 1–56 (2004).
7. Hyman, S. E. Revolution stalled. *Sci. Transl. Med.* **4**, 155cm11 (2012).
8. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
9. Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
10. Marshall, C. R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
11. Singh, T. et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat. Genet.* **49**, 1167–1173 (2017).
12. Gottesman, I. I. & Shields, J. A polygenic theory of schizophrenia. *Proc. Natl Acad. Sci. USA* **58**, 199–205 (1967).
13. 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).
14. Karayiorgou, M. et al. Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc. Natl Acad. Sci. USA* **92**, 7612–7616 (1995).
15. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
16. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
17. Rivas, M. A. et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
18. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
19. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
20. Singh, T. et al. Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
21. Howrigan, D. P. et al. Exome sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by protein-coding de novo mutations. *Nat. Neurosci.* **23**, 185–193 (2020).
22. Gulsuner, S. et al. Genetics of schizophrenia in the South African Xhosa. *Science* **367**, 569–573 (2020).
23. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
24. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
25. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at bioRxiv https://doi.org/10.1101/148353 (2017).
26. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
27. Hu, W., MacDonald, M. L., Elswick, D. E. & Sweet, R. A. The glutamate hypothesis of schizophrenia: evidence from human brain tissue studies. *Ann. NY Acad. Sci.* **1338**, 38–57 (2015).
28. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
29. Network and Pathway Analysis Subgroup of the Psychiatric Genetics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* **18**, 199–209 (2015).
30. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
31. Koopmans, F. et al. SynGO: an evidence-based, expert-curated knowledge base for the synapse. *Neuron* **103**, 217–234 (2019).
32. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
33. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
34. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
35. The Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
36. Barbosa, S. et al. Opposite modulation of RAC1 by mutations in *TRIO* is associated with distinct, domain-specific neurodevelopmental disorders. *Am. J. Hum. Genet.* **106**, 338–355 (2020).
37. Haukka, J., Suvisaari, J. & Lönnqvist, J. Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland. *Am. J. Psychiat.* **160**, 460–463 (2003).
38. Laursen, T. M. & Munk-Olsen, T. Reproductive patterns in psychotic patients. *Schizophr. Res.* **121**, 234–240 (2010).
39. Power, R. A. et al. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *Arch. Gen. Psychiat.* **70**, 22–30 (2013).
40. Endele, S. et al. Mutations in *GRIN2A* and *GRIN2B* encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–1026 (2010).
41. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

Tarjinder Singh[1,2,25] ✉, Timothy Poterba[1,2], David Curtis[3,4], Huda Akil[5], Mariam Al Eissa[6], Jack D. Barchas[7], Nicholas Bass[6], Tim B. Bigdeli[8], Gerome Breen[9], Evelyn J. Bromet[10], Peter F. Buckley[11], William E. Bunney[12], Jonas Bybjerg-Grauholm[13,14], William F. Byerley[15], Sinéad B. Chapman[2], Wei J. Chen[16], Claire Churchhouse[1,2], Nicholas Craddock[17], Caroline M. Cusick[2], Lynn DeLisi[18], Sheila Dodge[19], Michael A. Escamilla[20], Saana Eskelinen[21,22], Ayman H. Fanous[23], Stephen V. Faraone[24], Alessia Fiorentino[6], Laurent Francioli[1,25], Stacey B. Gabriel[19], Diane Gage[2], Sarah A. Gagliano Taliun[26,27], Andrea Ganna[1,28], Giulio Genovese[2], David C. Glahn[29], Jakob Grove[13,30,31,32], Mei-Hua Hall[33], Eija Hämäläinen[28], Henrike O. Heyne[1,2,28], Matti Holi[34], David M. Hougaard[13,14], Daniel P. Howrigan[1,2], Hailiang Huang[1,2], Hai-Gwo Hwu[35], René S. Kahn[36,37], Hyun Min Kang[38], Konrad J. Karczewski[1,2], George Kirov[39], James A. Knowles[40], Francis S. Lee[7], Douglas S. Lehrer[41], Francesco Lescai[13,42], Dolores Malaspina[36], Stephen R. Marder[36], Steven A. McCarroll[2,43], Andrew M. McIntosh[44], Helena Medeiros[23], Lili Milani[45], Christopher P. Morley[24,46], Derek W. Morris[47], Preben Bo Mortensen[48], Richard M. Myers[49], Merete Nordentoft[13,50,51], Niamh L. O'Brien[6], Ana Maria Olivares[2], Dost Ongur[33], Willem H. Ouwehand[52], Duncan S. Palmer[1,2], Tiina Paunio[53], Digby Quested[54], Mark H. Rapaport[55], Elliott Rees[39], Brandi Rollins[12], F. Kyle Satterstrom[1,2], Alan Schatzberg[56], Edward Scolnick[2], Laura J. Scott[38], Sally I. Sharp[6], Pamela Sklar[36], Jordan W. Smoller[57,58], Janet L. Sobell[59], Matthew Solomonson[25], Eli A. Stahl[36], Christine R. Stevens[2,25], Jaana Suvisaari[60], Grace Tiao[25], Stanley J. Watson[5], Nicholas A. Watts[25], Douglas H. Blackwood[61], Anders D. Børglum[13,30,31], Bruce M. Cohen[33], Aiden P. Corvin[62], Tõnu Esko[45], Nelson B. Freimer[64], Stephen J. Glatt[24], Christina M. Hultman[65], Andrew McQuillin[6], Aarno Palotie[1,2,25,28,57], Carlos N. Pato[8], Michele T. Pato[8], Ann E. Pulver[65], David St. Clair[66], Ming T. Tsuang[67], Marquis P. Vawter[12], James T. Walters[39], Thomas M. Werge[13,51,68,69], Roel A. Ophoff[63,70], Patrick F. Sullivan[71,72], Michael J. Owen[39], Michael Boehnke[38], Michael C. O'Donovan[39], Benjamin M. Neale[1,2,25,73] ✉ & Mark J. Daly[1,2,25,28,73] ✉

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]UCL Genetics Institute, University College London, London, UK. [4]Centre for Psychiatry, Queen Mary University London, London, UK. [5]Department of Psychiatry, Michigan Neuroscience Institute, Medical School, University of Michigan, Ann Arbor, MI, USA. [6]Division of Psychiatry, University College London, London, UK. [7]Weill Cornell Medical College, New York, NY, USA. [8]Department of Psychiatry and Behavioral Sciences, SUNY Downstate College of Medicine, Brooklyn, NY, USA. [9]Social Genetic and Developmental Psychiatry, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [10]Department of Psychiatry and Behavioral Health, Health Sciences Center, Stony Brook University, Stony Brook, NY, USA. [11]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA. [12]Department of Psychiatry and Human Behavior, University of California, Irvine, Irvine, CA, USA. [13]Lundbeck Foundation Initiative for Integrative Psychiatric Research, Copenhagen, Denmark. [14]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark. [15]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA. [16]College of Public Health, National Taiwan University, Taipei, Taiwan. [17]National Centre for Mental Health, Cardiff University, Cardiff, UK. [18]Department of Psychiatry, Cambridge Health Alliance, Cambridge Hospital, Cambridge, MA, USA. [19]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [20]Texas Tech University Health Sciences Center, El Paso, TX, USA. [21]University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [22]Department of Public Health Solutions, Mental Health Unit, National Institute for Health and Welfare, Helsinki, Finland. [23]Department of Psychiatry and Behavioral Sciences, SUNY Downstate Medical Center, Brooklyn, NY, USA. [24]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. [25]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [26]Faculté de Médecine, Université de Montréal, Montreal, Quebec, Canada. [27]Montréal Heart Institute, Montreal, Quebec, Canada. [28]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [29]Department of Psychiatry, Boston Children's Hospital, Boston, MA, USA. [30]Department of Biomedicine and Center for Integrative Sequencing, Aarhus University, Aarhus, Denmark. [31]Center for Genomics and Personalized Medicine, Aarhus, Denmark. [32]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [33]McLean Hospital, Harvard Medical School, Belmont, MA, USA. [34]Department of Psychiatry, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [35]Department of Psychiatry, National Taiwan University, Taipei, Taiwan. [36]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [37]MIRECC, JP Peters VA Hospital, Bronx, NY, USA. [38]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [39]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK. [40]Department of Cell Biology, SUNY Downstate Medical Center, Brooklyn, NY, USA. [41]Department of Psychiatry, Wright State University, Dayton, OH, USA. [42]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [43]Department of Genetics, Harvard Medical School, Boston, MA, USA. [44]University of Edinburgh, Edinburgh, UK. [45]Institute of Genomics, University of Tartu, Tartu, Estonia. [46]Department of Public Health and Preventive Medicine and Department of Family Medicine, SUNY Upstate Medical University, Syracuse, NY, USA. [47]National University of Ireland, Galway, Ireland. [48]Aarhus University, Aarhus, Denmark. [49]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [50]Copenhagen Research Center for Mental Health, Mental Health Services, Copenhagen University Hospital, Copenhagen, Denmark. [51]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [52]University of Cambridge, Cambridge, UK. [53]Department of Psychiatry, University of Helsinki, Helsinki, Finland. [54]Oxford Health NHS Foundation Trust, Oxford, UK. [55]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA, USA. [56]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA. [57]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [58]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [59]Department of Psychiatry and the Behavioral Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [60]Finnish Institute for Health and Welfare, Helsinki, Finland. [61]Department of Psychiatry, University of Edinburgh, Edinburgh, UK. [62]Trinity College Dublin, Dublin, Ireland. [63]Center for Neurobehavioral Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [64]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [65]School of Medicine, Johns Hopkins University, Baltimore, MD, USA. [66]University of Aberdeen, Aberdeen, UK. [67]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA. [68]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Copenhagen, Denmark. [69]Center for GeoGenetics, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [70]Department of Psychiatry, Erasmus Medical Center, Erasmus University, Rotterdam, the Netherlands. [71]Karolinska Institute, Solna, Sweden. [72]University of North Carolina, Chapel Hill, NC, USA. [73]These authors contributed equally: Benjamin M. Neale, Mark J. Daly. ✉e-mail: tsingh@broadinstitute.org; bneale@broadinstitute.org; mjdaly@atgu.mgh.harvard.edu

Methods

Ethics approval and consent to participate

Written institutional review board (IRB) approvals and study consent forms from each of the organizations contributing samples were sent to the Broad Institute of Harvard and MIT before the samples were sequenced and analysed. All relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained. All ethical approvals are on file at the IRB office at Massachusetts General Brigham (MGB), formerly Partners, amended to protocol no. 2014P001342, 'Molecular Profiling of Psychiatric Disease'. These approvals undergo annual continuing review by the MGB Human Research Committee IRB.

Consent for publication

All necessary patient and participant consent has been obtained, and the appropriate institutional forms have been archived.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

We describe all datasets in the manuscript or its Supplementary Information. We provide summary-level data at the variant and gene level in an online browser for viewing and download (https://schema.broadinstitute.org). There are no restrictions on the aggregated data released on the browser. For contributing datasets that are permitted to be distributed at the individual level, we have deposited or are currently depositing the data in a public repository (the database of Genotypes and Phenotypes (dbGaP) and/or the European Genome–Phenome Archive (EGA)), and we provide the accessions in Supplementary Table 1. Whole-exome sequence data generated under this study are currently hosted on and shared with the collaborating study groups via the controlled-access Terra platform (https://app.terra.bio/). The Terra environment, created by the Broad Institute, contains a rich system of workspace functionalities centred on data sharing and analysis. Requests for access to the controlled datasets are managed by data custodians of the SCHEMA consortium and the Broad Institute and are sent to sample contributing investigators for approval.

Code availability

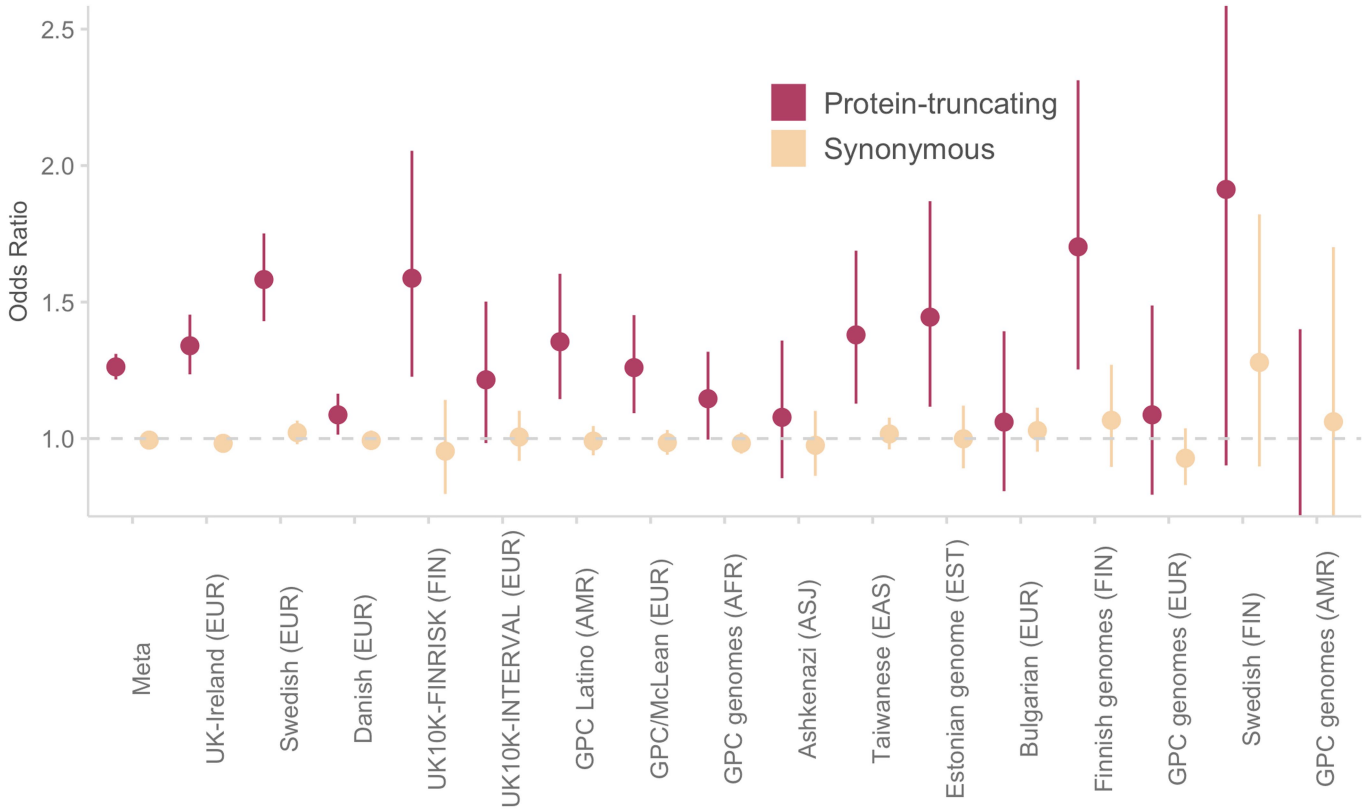
The software and code used are described throughout the Supplementary Methods. In brief, for sequence data generation, we used GATK versions 3.4 and 3.6, Picard version 1.1431 and VerifyBamID version 1.0.0. Sample and variant quality control and analyses were performed using Hail 0.1 and 0.2 (https://hail.is/), with functions and arguments referred to in the Supplementary Methods. Wrappers and methods using Hail code can be found at https://github.com/TarjinderSingh/hailutils. Additional (basic) processing and visualization were performed using base R (version 3.6) with tidyverse libraries (https://www.tidyverse.org/packages/).

Acknowledgements We would like to thank the patients and families who participated in our studies during the past two decades, without whom our research and findings would not be possible. The research reported in this publication was supported by the National Institute of Mental Health (NIMH) and the National Human Genome Research Institute of the National Institutes of Health under award numbers U01 MH105641, U01 MH105578, U01 MH105666, U01 MH109539, R01 MH085548, R01 MH085521, R01 MH124851 and U54 HG003067. We would also like to acknowledge support from K. Dauten and E. Dauten, the Stanley Family Foundation and the Dalio Foundation, which has enabled us to rapidly expand our data generation collections with the goal of moving towards better treatments for schizophrenia and other psychiatric disorders. Further, we wish to acknowledge all of the research participants in the BRIDGES cohort, which wassupported by NIMH under award numbers R01 MH094145 (M.B. and R.M.M., PIs) and U01 MH105653 (M.B., PI). The collection and storage of cases and controls from the Centre for Addiction and Mental Health (CAMH) in Toronto and from the Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, in London was supported by funding from GlaxoSmithKline. CAMH was supported by the Canadian Institutes of Health Research (MOP-172013, J. B. Vincent, PI, CAMH). IoPPN was supported by funding from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and the Maudsley NHS Foundation Trust and by King's College London. The views expressed are those of the author(s) and not necessarily those of the UK NHS, the NIHR or the UK Department of Health. Case and control collection was supported by the Heinz C. Prechter Bipolar Research Fund at the University of Michigan Depression Center to M.G. McInnis. Data and biomaterials were collected for the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD), a multi-centre, longitudinal project selected from responses to RFP NIMH-98-DS-0001, 'Treatment for Bipolar Disorder', which was led by G. Sachs and coordinated by Massachusetts General Hospital in Boston, with support from 2N01 MH080001-001. The Genomic Psychiatric Cohort (GPC) was supported by NIMH (U01 MH105641 (C.N.P., PI), R01 MH085548 (C.N.P. and M.T.P., PIs) and R01 MH104964 (C.N.P. and M.T.P., PIs). The MCTFR study was supported through grants from the National Institutes of Health under numbers DA037904, DA024417, DA036216, DA05147, AA09367, DA024417, HG007022 and HL117626. The work at Cardiff University was supported by Medical Research Council Centre grant no. MR/L010305/1 and programme grant mo. G0800509. We would like to acknowledge the Pritzker Neuropsychiatric Disorders Research Consortium for funding sample collection efforts. The iPSYCH team was supported by grants from the Lundbeck Foundation (R102-A9118, R155-2014-1724, and R248-2017-2003) and the Universities and University Hospitals of Aarhus and Copenhagen. The Danish National Biobank resource was supported by the Novo Nordisk Foundation. A.P. was supported by Academy of Finland Centre of Excellence in Complex Disease Genetics (grant no. 312074, 336824).S.V.F. is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 667302 and 965381; NIMH grants U01MH109536-01, U01AR076092-01A1, R0MH116037 and 5R01AG06495502; Oregon Health and Science University, Otsuka Pharmaceuticals and Supernus Pharmaceutical Company. T.S. was supported by a NARSAD Young Investigator Award from the Brain and Behavior Research Foundation.

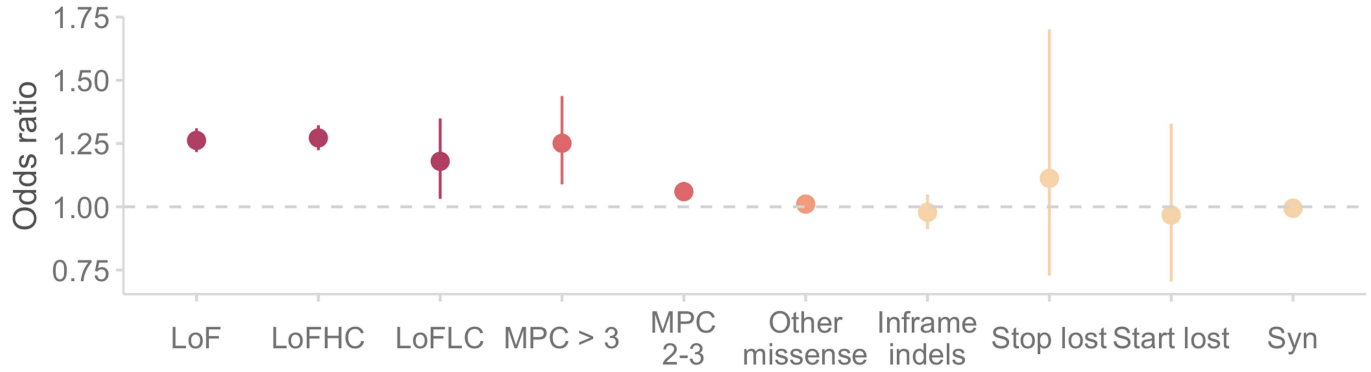
Author contributions , .T.S., E.S., L.J.S., P.S., A.P.C., M.B., M.C.O., B.M.N. and M.J.D. conceived and designed the experiments. T.S., T. Poterba, S.A.G.T., A.G., G.G., H.O.H., D.P.H., H.H., H.M.K., B.R., F.K.S., G.T., J.T.W., R.A.O., M.J.O., M.B., M.C.O., B.M.N. and M.J.D. designed and executed the analysis. H.A., J.D.B., W.E.B., S.B.C., C.C., C.M.C., S.D., S.B.G., D.G., F.S.L., P.B.M., R.M.M., A.M.O., A.S., E.A.S., C.R.S., N.A.W., J.T.W., M.J.O., M.C.O. and B.M.N contributed to project management/sequencing. D.C., M. A. Eissa, N.B., G.B., W.F.B., W.J.C., N.C., L.D., M. A. Escamilla, S.E., A.H.F., S.V.F., A.F., D.C.G., M.H., M.H., H.H., R.S.K., G.K., J.A.K., D.S.L., F.L., S.R.M., S.A.M., A.M.M., H.M., C.P.M., P.B.M., M.N., N.L.O., D.O., W.H.O., T. Paunio, D.Q., M.H.R., E.R., S.I.S., J.W.S., J.L.S., J.S., S.J.W., D.H.B., A.D.B., B.M.C., A.P.C., T.E., N.B.F., S.J.G., C.M.H., A.M., A.P., C.N.P., M.T.P., A.E.P., D.S.C., M.T.T., M.P.V., J.T.W., T.M.W., R.A.O., P.F.S., M.J.O. and M.C.O. recruited, assessed and/or contributed patient samples. T.S., T.B.B., E.J.B., P.F.B., J.B., L.F., J.G., E.H., D.M.H., K.J.K., D.M., A.M.M., L.M., C.P.M., D.S.P., J.W.S., M.S., A.P. and T.M.W. contributed reagents/materials/analysis tools. T.S., D.C., A.M.M., L.J.S., M.P.V., J.T.W., M.J.O., M.C.O., B.M.N. and M.J.D. wrote and/or edited the paper.

Competing interests M.J.D. is a founder of Maze Therapeutics and Neumora Therapeutics. B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora Therapeutics, a member of the scientific advisory committee at Milken and a consultant for Camp4 Therapeutics, Merck and Biogen. A.P. is a member of the genomics advisory board at AstraZeneca. M.C.O., M.J.O. and J.T.W. are supported by a collaborative research grant from Takeda Pharmaceuticals. E.A.S. is currently an employee of the Regeneron Genetics Center. D.S.P. was an employee of Genomics plc; all analyses reported in this paper were performed as part of his employment at Massachusetts General Hospital and the Broad Institute. The remaining authors declare no competing interests. In the past year, S.V.F. received income, potential income, travel expenses continuing education support and/or research support from Aardvark, Akili, Genomind, Ironshore, KemPharm/Corium, Noven, Ondosis, Otsuka, Rhodes, Supernus, Takeda, Tris and Vallon. In previous years, S.V.F. received support from: Alcobra, Arbor, Aveksham, CogCubed, Eli Lilly, Enzymotec, Impact, Janssen, Lundbeck/Takeda, McNeil, NeuroLifeSciences, Neurovance, Novartis, Pfizer, Shire, and Sunovion. With his institution, S.V.F. has US patent US20130217707 A1 for the use of sodium-hydrogen exchange inhibitors in the treatment of ADHD. S.V.F. receives royalties from books published by Guilford Press: Straight Talk about Your Child's Mental Health; Oxford University Press: Schizophrenia: The Facts; and Elsevier: ADHD: Non-Pharmacologic Interventions, and is Program Director of www.adhdinadults.com.

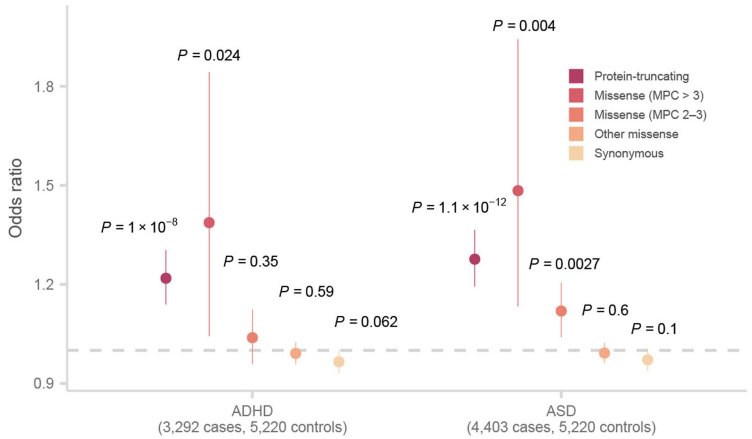
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04556-w.
Correspondence and requests for materials should be addressed to Tarjinder Singh, Benjamin M. Neale or Mark J. Daly.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



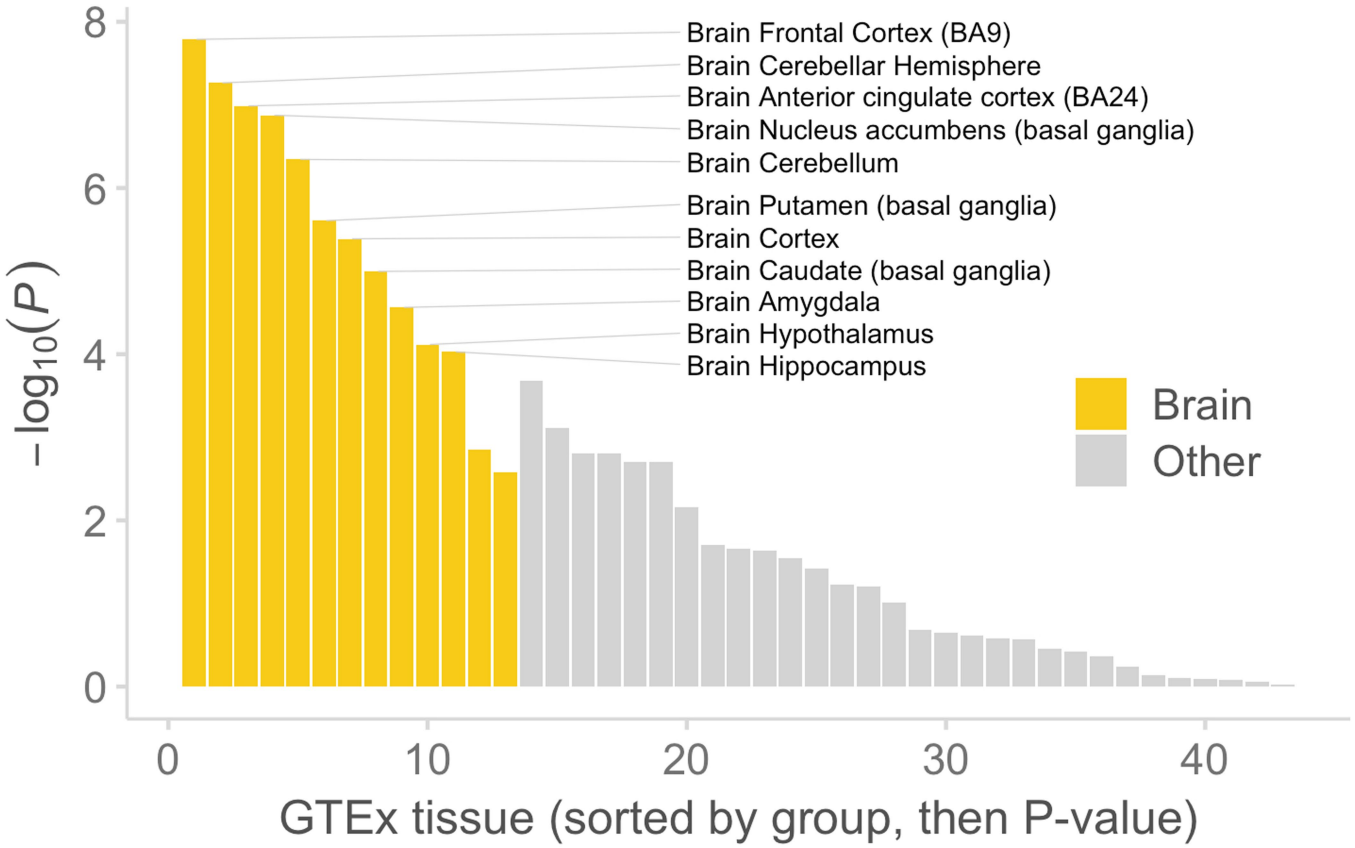
Extended Data Fig. 1 | Schizophrenia case-control enrichment in constrained genes (pLI > 0.9) in different SCHEMA cohorts (n = 22,444 cases and n = 39,837 controls). The odds ratio and standard error of PTVs and synonymous variants are provided for each cohort. The meta-analyzed odds ratio and standard error is calculated using inverse-variance. PTVs show consistent signals across the different cohorts, and synonymous variants do not deviate from expectation. Bars represent the 95% CIs of the point estimates.



Extended Data Fig. 2 | Schizophrenia case-control enrichment in constrained genes (pLI > 0.9) stratified by different variant annotations and inferred consequences (n = 22,444 cases and n = 39,837 controls). LoF: all loss-of-function or PTVs; LoFHC: high-confidence LOFTEE PTVs; LoFLC: low-confidence based on LOFTEE; MPC > 3: missense variants with MPC > 3; MPC 2 - 3: missense variants with MPC 2 - 3; Other missense: missense variants with MPC < 2; Syn: synonymous variants. The dot represents the odds ratio, and the bars represent the 95% CIs of the point estimates.



Extended Data Fig. 3 | Enrichment of URVs in n = 4,403 ASD and n = 3,292 ADHD cases compared to n = 5,220 controls stratified by variant annotation and consequences in constrained genes (pLI > 0.9). Two-sided *P* values from logistic regression displayed are from comparing the burden of variants of the labeled consequence in cases compared to controls. The dot represents the odds ratio, and the bars represent the 95% CIs of the point estimates.



Extended Data Fig. 4 | Schizophrenia case-control gene set enrichment in brain and non-brain GTEx tissues. We test for the burden of rare PTVs in genes with the strongest specific expression in that tissue type relative to other tissues as defined in[30]. Gene set burden statistics are calculated using a logistic regression model of rare variants from n = 22,444 cases and n = 39,837 controls. We report two-sided *P* values. Each bar is a different tissue in GTEx, grouped by whether it is part of the central nervous system and sorted by *P* value (Supplementary Table 8).

Extended Data Table 1 | Case-control and *de novo* counts of the ten Bonferroni significant genes in the main analysis

Gene Symbol	Case PTV	Ctrl PTV	Case mis3	Ctrl mis3	Case mis2	Ctrl mis2	De novo PTV	De novo mis3	De novo mis2	P value	Q value	OR (PTV)	OR (Class I)	OR (Class II)
SETD1A	15	3	3	4	11	10	3			2.00E-12	3.62E-08	20.1 (5.68-108)	10.3 (4.12-29.3)	4.42 (1.7-11.6)
CUL1	8	1	2	0	7	16	3			2.01E-09	1.82E-05	36.1 (5.01-1570)	44.2 (6.42-1880)	1.76 (0.611-4.51)
XPO7	12	1	1	1	10	32	1			7.18E-09	4.34E-05	52.2 (7.84-2190)	28.1 (6.46-253)	1.25 (0.55-2.62)
TRIO	18	16	0	0	24	102	2			6.35E-08	2.88E-04	5.02 (2.47-10.4)	5.02 (2.47-10.4)	0.944 (0.579-1.48)
CACNA1G	10	13	8	4	55	134			1	4.57E-07	1.54E-03	3.09 (1.21-7.63)	4.25 (2.07-8.78)	1.68 (1.21-2.31)
SP4	13	6	3	3	0	2	1			5.08E-07	1.54E-03	9.37 (3.38-29.7)	7.59 (3.2-19.3)	0 (0-21.4)
GRIA3	5	0	3	2	10	24	1	1		5.98E-07	1.55E-03	Inf (4.73-Inf)	20.1 (4.28-188)	1.67 (0.714-3.63)
GRIN2A	9	2	3	0	13	22				7.37E-07	1.67E-03	18.1 (3.74-172)	24.1 (5.36-221)	2.37 (1.1-4.92)
HERC1	28	32	0	0	2	8				1.26E-06	2.54E-03	3.51 (2.04-6.03)	3.51 (2.04-6.03)	1 (0.104-5.03)
RB1CC1	9	4	0	0	0	0	2			2.00E-06	3.63E-03	10 (2.89-43.9)	10 (2.89-43.9)	0 (0-Inf)

Case-control and *de novo* counts of the ten Bonferroni significant genes in the main analysis

Case-control counts displayed are the total counts for variants with minor allele count <= 5. PTV: protein-truncating variant, mis3: missense variants with MPC > 3, mis2: missense variants with MPC 2 - 3; Q value: adjusted *P* value after FDR adjustment; Class I: PTV and missense variants (MPC > 3); Class II: missense variants (MPC 2 - 3). Two-sided gene *P* values for Class I and Class II variants are calculated using the permuted Fisher's exact test. Gene *P* values for *de novo* mutations are calculated using a one-sided Poisson rate test. The meta-analysis gene *P* value is calculated from the weighted Z-score method.

nature research

Corresponding author(s): Mark J. Daly

Last updated by author(s): Mar 21, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software and code was used in data collection.
Data analysis	Software and code used are described in throughout the Supplementary Methods of the manuscript. For sequence data generation, we used GATK v3.4 and v3.6 (GATK nightly-2015-07-31-g3c929b0, 3.4-89-ge494930, and 3.6-0-g89b7209), Picard version 1.1431, and VerifyBamID version 1.0.0. Sample, variant QC, and analyses was performed using functions in Hail 0.1 and 0.2 (website: https://www.hail.is; documentation: https://hail.is/docs/0.1/ and https://hail.is/docs/0.2/; GitHub repository: https://github.com/hail-is/hail), with specific functions and arguments referred to in the Supplementary Text. Variant annotation was performed using the Ensembl Variant Effect Predictor (VEP) v85 tool as implemented in Hail 0.1 with the LOFTEE annotation provided as default (https://github.com/konradjk/loftee/tree/27b0040f524348baa7f3257f1ce58993529e09ef). To further prioritize missense variants, we annotated all variants using Hail (https://hail.is/docs/0.1/annotationdb.html) with classifiers included in the dbNSFP database (such as CADD and PolyPhen). Wrappers and methods using Hail code can be found at https://github.com/TarjinderSingh/hailutils. Additional (basic) processing and visualization was performed using base R (v3.6.1) oriented around built-in functions in the tidyverse libraries (https://www.tidyverse.org/packages/): tidyverse (v1.3.0), tidylog (v1.0.2), extrafont (v0.17), RColorBrewer (v1.1-2), here (v1.0.1), fs (v1.5.0), bookdown (v0.18), knitr (v1.31,), forcats (v0.5.0), stringr (v1.4.0), dplyr (v1.0.2), purrr (v0.3.4), readr (v1.3.1), tidyr (v1.1.2), tibble (v3.0.3), ggplot2 (v3.3.2), devtools (v2.3.0), usethis (v1.6.0), readxl (v1.3.1), metafor (v2.4-0), meta (4.11).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We describe all datasets in the manuscript or Supplementary Information. We provide summary-level data at the variant and gene level in an online browser for viewing and download (https://schema.broadinstitute.org). There are no restrictions on the aggregated data released on the browser. For contributing data sets that are permitted to be distributed at the individual level, we have deposited, or are currently depositing, the data in a public repository (the database of Genotypes and Phenotypes [dbGAP] and/or the European Genome-phenome Archive [EGA]) and provide the accessions in Table S1. Whole Exome Sequence data generated under this study are currently hosted on and shared with the collaborating study groups via the controlled access Terra platform (https://app.terra.bio/). The Terra environment, created by the Broad Institute, contains a rich system of workspace functionalities centered on data sharing and analysis. Requests for access to the controlled datasets are managed by data custodians of the SCHEMA consortium and the Broad Institute and sent to sample contributing investigators for approval.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was not predetermined in this study: we aggregated all available exome and genome data from collections that included individuals with a diagnosis of schizophrenia or schizoaffective disorders. This is the largest sequencing study of psychiatric disorders to date.
Data exclusions	We describe sample ascertainment in detail in the Supplementary Materials. There, we additionally described the criteria for which samples and variants were excluded in our study (see sections on Sample and Variant QC). Samples and variants were excluded if they failed quality control metrics. We included only cases with a clear diagnosis of schizophrenia or schizoaffective disorders, and controls without a known diagnosis of a psychiatric disorder.
Replication	Our main analysis integrated case-control and de novo mutations for gene discovery. De novo mutations were characterized in previously publications, and we observed concordance in genes implicated in both types of data. We did not attempt to reproduce any findings in a separate dataset, as no other data set of comparable size exists.
Randomization	Randomization was not applicable/performed because there was no allocation of samples to experimental groups in our observational genetic study. Case and control status of samples were assigned by clinicians and investigators of contributing collections. We controlled for confounding factors (sequencing technology and population ancestry) by stratifying rare variant counts accordingly in our main analysis.
Blinding	Blinding was not relevant to our study, as the genotype and phenotype data is determined/defined externally and could not be influenced by the analyst or during our aggregation steps.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Table S1 described contributing collections along with the number of samples sequenced, the number of samples retained in the final analysis, and the technological assays used to generate the sequencing data. For each collection, we referred to earlier publications that described the phenotypic ascertainment and analyzed these data for insights related to schizophrenia. Sample recruitment and collection are described extensively in those corresponding studies and collections. To ensure compatibility with Psychiatric Genomics Consortium (PGC) definitions, we included samples with a diagnosis of schizophrenia and schizoaffective disorders in our analysis. In the final analysis, 22,781 individuals were diagnosed with schizophrenia, while 1,467 were diagnosed with schizoaffective disorders. Each collection provided the Consortium with individual-level information on case status according to study-specific criteria. Age was not reported as part of the PGC recruitment progress. 41,584 individuals (15,254 cases) had "XY" as their imputed sex, and 33,101 individuals (8,994 cases) had "XX" as their imputed sex. The number of individuals in each ancestry group include: 4,307 AFR: African, 3,284 ASJ: Ashkenazi Jewish, 4,534 AMR: Latin American, 3,337 EAS: East Asian, 2,542 EST: Estonian, 10,006 FIN: Finnish, 46,412 EUR: non-Finnish European, and 263 SAS: South Asian. The counts in each stratum (by case and control status) is fully reported in Table S3. Each collection defined sample controls as individuals specifically ascertained to not have psychiatric illness, or individuals that were randomly selected from population registers.
Recruitment	The Psychiatric Genomics Consortium (PGC) consortium predefined phenotype criteria for cases and controls, but the specific recruitment was carried out by each contributing study. Patients were recruited originally as a part of numerous cohort studies, described in Table S1. Most participants in our sequencing study are also included in common variant (GWAS) analyses of schizophrenia. The ascertainment strategies in SCHEMA are presented in detail in the PGC GWAS manuscripts (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020), from which the majority of these collections were first recruited. To ensure compatibility with PGC definitions, we included samples with a diagnosis of schizophrenia and schizoaffective disorders in our analysis. Each collection defined sample controls as individuals specifically ascertained to not have psychiatric illness, or individuals that were randomly selected from population registers. In the final analysis, 22,781 individuals were diagnosed with schizophrenia, while 1,467 were diagnosed with schizoaffective disorders. In Table S1, we highlight publications that specifically describe the ascertainment strategies of each collection. In Extended Data Figure 1, we observe variation in the strength of PTV enrichment between collections. While it is difficult to pinpoint the precise reason, differences in ascertainment strategies by cohort in cases and/or controls may have affected the strength of signal by cohort. Most participants in our sequencing study are also included in common variant (GWAS) analyses of schizophrenia. In those analyses, the polygenic risk score explained more variance in liability in collections of European ancestry (likely a result of the ancestry composition of the GWAS), and also in samples by which ascertainment likely include the most severe cases (hospitalized participants or those treated with clozapine). Another study has also reported people with chronic schizophrenia have a notably higher loading of polygenic risk. Therefore, different recruitment strategies in cases and controls may have an effect on enrichment of genetic risk in global schizophrenia collections. While it is difficult to clearly parse the different effects at the moment, we expect that increased phenotypic information available in future collections and more comprehensive sequencing of diverse populations will help unravel these effects. We include a more comprehensive discussion of variability between collections in the Supplementary Methods.
Ethics oversight	Written IRB approvals and study consent forms from each of the sample contributing organizations were sent to the Broad Institute of Harvard and M.I.T. before samples were sequenced and analyzed. All relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained. All ethical approvals are on file at the Massachusetts General Brigham (MGB), formerly Partners, IRB office amended to protocol #2014P001342, title: "Molecular Profiling of Psychiatric Disease" and undergoes annual continuing review by the Mass General Brigham Human Research Committee (MGBHRC) Institutional Review Board (IRB) of Mass General Brigham (Mass General Brigham IRB, Mass General Brigham, 399 Revolution Drive, Suite 710, Somerville, MA 02145). All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

Note that full information on the approval of the study protocol must also be provided in the manuscript.