Neuron

Novel ultra-rare exonic variants identified in a founder population implicate cadherins in schizophrenia

Highlights

- Schizophrenia rare variant discovery is enhanced in the Ashkenazi founder population
- Ultra-rare variant burden is inversely correlated with polygenic risk scores in cases
- Ultra-rare exonic variants in schizophrenia cases are enriched in cadherin genes
- A recurrent case mutation in *PCDHA3* disrupts homophilic interactions in culture

Authors

Todd Lencz, Jin Yu, Raiyan Rashid Khan, ..., Tom Maniatis, Gil Atzmon, Itsik Pe'er

Correspondence

tlencz@northwell.edu (T.L.), itsik@cs.columbia.edu (I.P.)

In brief

Lencz et al. demonstrate that the Ashkenazi Jewish population has an enhanced power for genetic discovery in schizophrenia. Cases had excess missense or loss-of-function ultra-rare variants, enriched in cadherins and neurodevelopmental genes. A recurrent case mutation in *PCDHA3* results in the formation of cytoplasmic aggregates and failure to engage in membrane homophilic interactions.

Lencz et al., 2021, Neuron *109*, 1465–1478
May 5, 2021
https://doi.org/10.1016/j.neuron.2021.03.004

CellPress

Neuron

CellPress

Article

Novel ultra-rare exonic variants identified in a founder population implicate cadherins in schizophrenia

Todd Lencz,[1,2,3,25,27,*] Jin Yu,[2,3,25] Raiyan Rashid Khan,[4] Erin Flaherty,[5,6] Shai Carmi,[7] Max Lam,[2,3] Danny Ben-Avraham,[8,9] Nir Barzilai,[8,9] Susan Bressman,[10] Ariel Darvasi,[11,26] Judy H. Cho,[12,13] Lorraine N. Clark,[14,15] Zeynep H. Gümüş,[13,16] Joseph Vijai,[17] Robert J. Klein,[13,15] Steven Lipkin,[18] Kenneth Offit,[17,19] Harry Ostrer,[20] Laurie J. Ozelius,[21] Inga Peter,[13,16] Anil K. Malhotra,[1,2,3] Tom Maniatis,[5,6,22] Gil Atzmon,[8,9,23] and Itsik Pe'er[4,24,*]

[1]Departments of Psychiatry and Molecular Medicine, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY 11550, USA
[2]Department of Psychiatry, Division of Research, The Zucker Hillside Hospital Division of Northwell Health, Glen Oaks, NY 11004, USA
[3]Institute for Behavioral Science, The Feinstein Institutes for Medical Research, Manhasset, NY 11030, USA
[4]Department of Computer Science, Columbia University, New York, NY 10027, USA
[5]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
[6]Mortimer B. Zuckerman Mind Brain and Behavior Institute, Columbia University, New York, NY 10027, USA
[7]Braun School of Public Health and Community Medicine, Faculty of Medicine, Hebrew University of Jerusalem, Ein Kerem, Jerusalem 9112102, Israel
[8]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[9]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[10]Department of Neurology, Beth Israel Medical Center, New York, NY 10003, USA
[11]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Givat Ram, Jerusalem 91904, Israel
[12]Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[13]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[14]Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY 10032, USA
[15]Taub Institute for Research of Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, New York, NY 10032, USA
[16]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[17]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[18]Departments of Medicine, Genetic Medicine and Surgery, Weill Cornell Medical College, New York, NY 10065, USA
[19]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[20]Departments of Pathology and Pediatrics, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[21]Department of Neurology, Massachusetts General Hospital, Charlestown, MA 02129, USA
[22]New York Genome Center, New York, NY 10013, USA
[23]Department of Human Biology, Haifa University, Haifa, Israel
[24]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA
[25]These authors contributed equally
[26]Deceased
[27]Lead contact
*Correspondence: tlencz@northwell.edu (T.L.), itsik@cs.columbia.edu (I.P.)
https://doi.org/10.1016/j.neuron.2021.03.004

SUMMARY

The identification of rare variants associated with schizophrenia has proven challenging due to genetic heterogeneity, which is reduced in founder populations. In samples from the Ashkenazi Jewish population, we report that schizophrenia cases had a greater frequency of novel missense or loss of function (MisLoF) ultra-rare variants (URVs) compared to controls, and the MisLoF URV burden was inversely correlated with polygenic risk scores in cases. Characterizing 141 "case-only" genes (MisLoF URVs in $\geq$3 cases with none in controls), the cadherin gene set was associated with schizophrenia. We report a recurrent case mutation in *PCDHA3* that results in the formation of cytoplasmic aggregates and failure to engage in homophilic interactions on the plasma membrane in cultured cells. Modeling purifying selection, we demonstrate that deleterious URVs are greatly overrepresented in the Ashkenazi population, yielding enhanced power for association studies. Identification of the cadherin/protocadherin family as risk genes helps specify the synaptic abnormalities central to schizophrenia.

INTRODUCTION

Twin studies and other family-based designs have long demonstrated that schizophrenia (SCZ) is highly heritable ($h^2 \approx 0.6$–0.85) (Hilker et al., 2018; McGue et al., 1983; Sullivan et al., 2003). While large-scale genome-wide association studies (GWASs) have discovered increasing numbers of common (minor allele frequency >1%) variants associated with illness (Lam et al., 2019; Pardiñas et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014, 2020), the cumulative effect of such variants accounts for only approximately one-third of the total heritability of SCZ (Lee et al., 2012; Loh et al., 2015). It is therefore likely that rare genetic variants contribute substantially to the heritability of SCZ (Ganna et al., 2018; Purcell et al., 2014), and such rare variants may have considerably higher effect sizes (odds ratios [ORs]) relative to common variants (Sullivan et al., 2012). For example, several rare (frequency << 1% in the general population) copy-number variants (CNVs) have been reliably associated with SCZ, with odds ratios ranging from 5 to ≥20 (Marshall et al., 2017).

The identification of rare single-nucleotide variants (SNVs) associated with SCZ has proven difficult for several reasons: (1) SCZ is marked by a high degree of locus heterogeneity due to the large "mutational target" (i.e., damage to many different genes can increase risk for the phenotype) (Gratten et al., 2014); (2) at any given gene, a variety of different alleles may have deleterious effects (allelic heterogeneity) (Li and Leal, 2009); (3) deleterious rare variants are generally driven to extremely low frequencies due to purifying selection (Kryukov et al., 2007); and (4) the background rate of benign rare variation across the population is very high (Tennessen et al., 2012). To date, only very large international consortia efforts have identified any SCZ-associated SNVs. The largest such effort, the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium identified only 10 exome-wide significant genes with a sample size of 25,000 cases and nearly 100,000 controls (Singh et al., 2020).

One approach to enhance power in rare variant studies is to examine unusual populations marked by a strong, (relatively) recent founder effect; such populations are enriched for deleterious rare variants due to inefficient purifying selection (Locke et al., 2019; Wang et al., 2014). For example, the Ashkenazi Jewish (AJ) population, currently numbering >10 million individuals worldwide, effectively derives from a mere ~300 founders ~750 years ago (Carmi et al., 2014; Palamara et al., 2012). While the AJ population is well known to be enriched for deleterious variants leading to rare recessive disorders (Baskovich et al., 2016), the AJ population also demonstrates a 10-fold elevated frequency of high-penetrance risk variants for common complex disease, such as the *LRRK2* p.G2019S allele associated with Parkinson's disease (Ozelius et al., 2006) and the *BRCA1* c.66_67AG allele associated with breast cancer (Friedman et al., 1995). Importantly, a recent large-scale (n > 5,000) sequencing study of AJ individuals demonstrated that this enrichment is widespread across the exome, with approximately one-third of all protein-coding alleles demonstrating frequencies in AJ that were an order of magnitude greater than the maximum frequency in any well-characterized outbred population (Rivas et al., 2018).

In the present study, we examined rates of protein-altering ultra-rare variants (URVs) in AJ cases with SCZ compared with AJ controls. Based on prior research (Genovese et al., 2016; Gulsuner et al., 2020; Nguyen et al., 2017; Purcell et al., 2014), we hypothesized that cases would be enriched for rare deleterious variants. We also examined the relationship between URV burden and polygenic risk scores derived from common variant GWAS to test the hypothesis that these would be inversely correlated in SCZ cases as predicted under the additive model. We sought to replicate prior findings that SCZ URVs would be overrepresented in genes expressed at the neuronal synapse and during neurodevelopment, and to extend these results to additional categories of genes that may be detectable due to the greater frequency of rare variants observed in the AJ population. In addition, we attempted to replicate the SCZ risk genes identified by the SCHEMA consortium, and we sought evidence of individual risk variants that could be observed multiple times across our dataset and SCHEMA. For one such variant, we performed *in vitro* experiments to identify potential molecular mechanisms by which risk for illness may be conferred. Finally, we modeled the process of purifying selection in a rapidly expanding, bottlenecked population to quantify the relative power of AJ for rare variant discovery.

RESULTS

Greater rates of novel exonic variants in cases

After quality control (QC) procedures, a total of 786 SCZ cases and 463 controls were available for final analysis. Groups did not significantly differ in the total number of variants called in their whole genome (~3.68 million) or exome (~49,000) (Table 1). Called variants were then filtered for novelty against all of the variants observed in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine TOPMed and gnomAD (version 2.1.1, non-neuro) datasets. This variant filtration procedure was performed equally in both cases and controls in our dataset, and therefore the number of URVs per genome is not a function of asymmetrical sample size; notably, cases and controls did not significantly differ on the total number of novel variants observed genome-wide (~5,000). However, cases had significantly more novel exonic variants, exclusively limited to singletons (17.76 ± 6.24 versus 15.44 ± 6.42, $p = 6.13 \times 10^{-10}$; Table 1).

Against this backdrop, cases and controls were compared on the number of missense or loss of function (MisLoF) variants within the exome in two ways: variant-based tests and gene-based tests. At the variant level, cases manifest a significantly elevated rate of novel MisLoF URVs (Table 1, last row). Cases also demonstrated an elevated rate of novel non-MisLoF URVs (Table 1, second-to-last row); however, even compared to this background elevation of non-MisLoF URVs, there was a significantly elevated proportion of exonic variants classified as MisLoF in cases (Fisher's exact $p = 0.034$). Next, we examined case-control differences in MisLoF versus non-MisLoF URVs at the gene level, as follows: for each class of URV (MisLoF or non-MisLoF), we compared the number of genes hit by one or more such URVs in cases with no such hits in controls ("case-only" genes)

Table 1. Variant counts in cases and controls

					Logistic regression			Linear regression		
	Per case (n = 786)	SD	Per control (n = 463)	SD	OR	95% CI	p	Excess variant per case	95% CI	p
Post-QC										
Variants per genome	3,676,280	17,055	3,676,451	18,899	1	1.000–1.000	0.97	−15.51	−2,060 to 2,029	0.99
Variants per exome	49,437	356	49,454	387	1	1.000–1.000	0.67	−8.62	−51 to 33	0.69
Postfiltering on gnomAD (non-neuro) and TOPMed										
Novel variants per genome	5,028	427	5,034	322	1	1.000– 1.000	0.89	−3.15	−48 to 42	0.89
Novel variants per exome	30.64	6.14	28.66	7.15	1.045	1.027–1.064	1.53E−6	1.99	1.20–2.79	1.04E−6
Novel non-singleton per exome	12.88	2.7	13.23	3.01	0.958	0.920–0.998	3.78E−2	−0.34	−0.67 to −0.02	3.76E−2
Novel singleton (URV) per exome	17.76	6.24	15.44	6.42	1.067	1.045–1.090	1.33E−9	2.34	1.61–3.06	4.38E−10
Non-MisLoF URV	9.86	4.13	8.81	3.97	1.07	1.037–1.103	1.66E−5	1.05	0.58–1.52	1.30E−5
MisLoF URV	7.9	3.54	6.63	3.51	1.116	1.077–1.156	1.74E−9	1.29	0.88–1.69	7.46E−10

CI, confidence interval; MisLoF, missense or loss of function; OR, odds ratio; SD, standard deviation; TOPMed, Trans-Omics for Precision Medicine; URV, ultra-rare variant.

to the number of genes hit by one or more such URVs in controls with no such hits in cases ("control-only" genes), using the formula $\frac{\#CASE\_only - \#CONTROL\_only}{\#CASE\_only}$. As shown in Figure 1, this ratio was much greater (i.e., more case-only genes than control-only genes) for MisLoF URVs relative to non-MisLoF URVs (p = 0.0001 by permutation test). Note that for this analysis, we down-sampled the number of cases to match the number of controls, and iterated across 10,000 permutations.

MisLoF URVs are inversely correlated with common variant polygenic risk score in cases

Next, we tested the liability threshold model of SCZ (Gottesman and Shields, 1967; Kendler, 2015; McGue et al., 1983; Smeland et al., 2020), which suggests that genetic risk factors are largely additive; if true, then it would be expected that cases with a greater URV burden would require a lower burden of common risk variants, as indexed by GWAS-derived polygenic risk score (PRS). By contrast, no relationship between PRS and URV would be expected in controls. For each subject, PRS was calculated based on the large-scale SCZ GWAS reported by the Psychiatric Genomics Consortium (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020), excluding our own Ashkenazi cohort, as described in Common variant polygenic risk score (PRS). The p value threshold (P_T) for calculating the PRS from the GWAS summary statistics was determined by optimizing R^2 for the comparison of AJ cases and controls. At P_T = 0.00725, the Nagelkerke R^2 for PRS of the cases versus controls was optimized at 0.15 on the observed scale (OR = 2.21 for 1 U of standardized PRS, $p < 2 \times 10^{-16}$), consistent with previous estimates (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014, 2020). Using this threshold and controlling for sex and the first 5 principal components derived from the GWAS data, there was a significant inverse relationship between PRS and the total number of MisLoF URVs in cases ($\beta = -0.0232$, SE = 0.0095, p = 0.014), but not in controls ($\beta = -0.0005$, SE = 0.0126, p = 0.967). Results were substantively unchanged when a different PRS threshold was used, based on the optimal level (P_T = 0.05) reported in the original GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020); there remained a significant inverse relationship between PRS and URV for cases ($\beta = -0.0207$, SE = 0.0096, p = 0.031), but not controls ($\beta = -0.0031$, SE = 0.0124, p = 0.805).

Replication of previously identified SCZ risk genes

Given that cases demonstrated significant elevation in genes carrying MisLoF URVs, we next sought to characterize the genes that were hit by MisLoF URVs in cases only, with none observed in controls. As shown in Table S1A, 8 genes had case ≥ 5 and control = 0 MisLoF URVs. Notably, one of these was *SETD1A*, a methyltransferase gene that was the first to reach genome-wide significance in a SCZ rare variant study (Singh et al., 2016). We next tested the set of 9 autosomal exome-wide-significant SCZ genes identified by the SCHEMA consortium (Singh et al., 2020) for overlap with 141 "case-only" genes in which ≥ 3 AJ cases in our dataset had MisLoF URVs, with none found in our AJ controls (full list provided in Table S1A). Three of 9 autosomal exome-wide significant SCHEMA genes (*SETD1A*, *TRIO*, and *XPO7*) were among the 141 case-only genes, a 47-fold overrepresentation relative to chance (hypergeometric test $p = 2.89 \times 10^{-5}$). Results remained significant when permutation tests controlling for gene size (see Method details) were performed (empirical $p = 1.7 \times 10^{-3}$; see also Table S2). Similar results were obtained in examining the overlap of our 141 case-only genes with the set of 29 autosomal genes that met the criteria of FDR < 0.05 in SCHEMA; in addition to the 3 genes above, *STAG1* was also shared between our case-only list and SCHEMA (4/29 genes; hypergeometric p = 5.12 ×

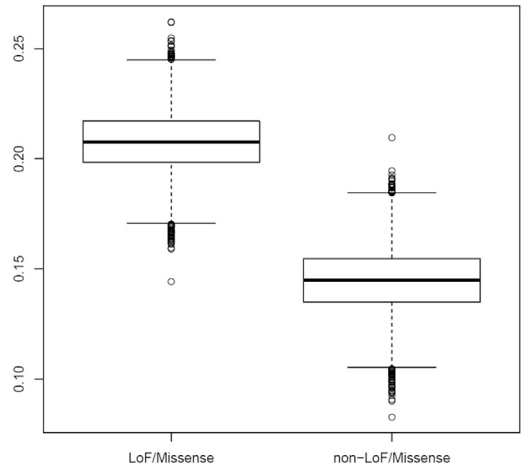

Figure 1. More genes are hit by missense or loss-of-function ultra-rare variants (URVs) in cases relative to controls

The y axis denotes the degree of elevation of "case-only" genes to "control-only" genes, defined by the ratio $\frac{\#CASE\_only - \#CONTROL\_only}{\#CASE\_only}$. The boxplot at left displays this ratio for missense or loss-of-function URVs; there are many more genes in which missense or loss-of-function URVs are observed in cases only, but no controls (ratio ≈0.205). The boxplot at right displays this ratio for all of the other URVs in the exome calling interval (including synonymous exonic, flanking intronic variants, and UTRs); while there are more genes in which such URVs are observed in cases only relative to controls only, the ratio (≈0.145) is much smaller than that observed for missense and loss-of-function URVs (empirical p = 0.0001; note that, for this analysis, we downsampled the number of cases to match the number of controls, and iterated across 10,000 permutations).

10^{-5}; empirical $p = 8.5 \times 10^{-3}$ using permutations controlling for gene size).

In addition to evidence that our dataset provided support for genes identified as significant in SCHEMA, we also used the broader SCHEMA dataset to provide supporting evidence for the genes on our case-only list. We observed that a total of 13 of our case-only genes had nominal ($p < 0.05$) support within the SCHEMA dataset, which represents a significant enrichment relative to chance (hypergeometric p = 0.012; empirical $p < 1 \times 10^{-4}$ using permutations controlling for gene size). These 13 genes include *BSN*, *CACNA1E*, *DNAJC14*, *HMGCR*, *KIAA0586*, *MACF1*, *SPAG5*, *TOP2B*, and *WFS1*, in addition to the 4 genes noted in the paragraph above. In addition, we examined the distribution of p values in SCHEMA for the 141 case-only genes we identified, as compared against all other genes. As shown in Figure S1, there is a distinct enrichment for our case-only genes compared to all other genes, and the λ values are 1.14 and 0.73, respectively.

Within our set of 141 case-only genes, there was no evidence of oligogenic effect; the distribution of cases with >1 MisLoF variant in this set of genes is as expected under the null for multinomial distributions (multinomial goodness of fit p = 0.3825) (Resin, 2020). Moreover, the distribution of multiply "hit" cases did not significantly differ ($\chi^2(4) = 3.15$, p = 0.53) from the distribution of controls with >1 MisLoF variant in a similarly sized set of 148 control-only genes (defined as genes in which ≥ 2 AJ controls had MisLoF URVs with none found in our AJ cases; Table S1B).

We compared the 347 cases that carried at least 1 MisLoF URV in the 141 case-only genes to the 439 cases that were non-carriers of such URVs on 3 clinical variables available in our dataset: (1) age of onset; (2) severity of course (defined as continuous illness without episodes of full or partial remission); and (3) treatment resistance (defined using prescription of clozapine as a proxy for treatment resistance; Ruderfer et al., 2016). There was no significant difference between case carriers and case non-carriers in age of onset (24.3 years versus 24.2 years; t = 0.27, p = 0.79); there was no significant difference in the designation of course as "continuous" (46.8% versus 42.7%, $\chi^2 = 1.10$, p = 0.29); and there was no significant difference in treatment resistance as suggested by the prescription of clozapine (20.6% versus 19.4%, $\chi^2 = 0.18$, p = 0.67). As a test of the sensitivity of our clinical measures, we confirmed a significant relationship between treatment resistance and age of onset; patients who were prescribed clozapine (regardless of URV status) had an earlier age of onset compared to patients not prescribed clozapine (22.5 versus 24.7 years of age; t = −2.94, p = 0.003), as has been consistently reported in other studies (Smart et al., 2021).

Case-only genes are enriched for known neurodevelopmental genes, synaptic genes, and cadherins

We used gene set analyses to characterize our 141 case-only genes and 148 control-only genes. We tested sets of genes selected *a priori* based on previous literature; specifically, previous case-control exome studies in SCZ have identified: (1) overlaps with other developmental brain disorders (DBDs), including autism spectrum disorder (ASD) and intellectual disability (ID); (2) gene sets representing critical synaptic and/or neurodevelopmental functions such as binding partners of FRMP, RBFOX, and CELF4; and (3) constrained genes (i.e., genes with far fewer MisLoF variants than average, presumably due to purifying selection) (Genovese et al., 2016; Gonzalez-Mantilla et al., 2016; Gulsuner et al., 2020; Nguyen et al., 2017; Purcell et al., 2014). As shown in Table 2, each of these gene sets demonstrated significant (by hypergeometric test) overlap with the case-only gene lists; moreover, the difference in enrichment between the case-only and control-only gene sets was statistically significant in all of the cases. Permutation tests accounting for gene size demonstrated similar results, with the exception that the overlaps with ASD/ID gene sets were no longer significant, and the SynaptomeDB gene set was marginal (p = 0.0514) (Table S2). Moreover, the size-matched permutation tests demonstrate that the genes on the control-only list were significantly underrepresented among many of the key gene sets of interest, as indicated by empirical p > 0.95 (i.e., <5% of all gene-size-matched permuted gene lists had less overlap with the target gene set than the control-only list itself).

By contrast, there was no significant overlap of the case-only list with genes identified in the most recent GWAS of the Psychiatric Genomics Consortium (using either broad or narrow criteria for prioritizing genes within a GWAS locus) (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020), nor was there any significant overlap with SCZ risk genes identified in two reports from the PsychENCODE Consortium: (1) 301

Table 2. *A priori* gene set analysis

	Genes in gene set	Case-only overlap	Control-only overlap	p(case)	p(control)	OR	χ^2 p
DBD (Gonzalez-Mantilla et al., 2016)	214	11	1	3.95×10^{-7}	8.01×10^{-1}	12.44	2.40×10^{-3}
ASD (Satterstrom et al., 2020)	102	4	0	6.13×10^{-3}	1.00	inf	3.91×10^{-2}
ASD_ID_DD (Coe et al., 2019; Satterstrom et al., 2020)	274	8	2	7.93×10^{-4}	6.10×10^{-1}	4.39	4.45×10^{-2}
SynaptomeDB (Pirooznia et al., 2012)	1,828	26	15	4.88×10^{-4}	3.93×10^{-1}	2.00	4.31×10^{-2}
CELF4 (Wagnon et al., 2012)	2,504	41	23	1.70×10^{-7}	1.74×10^{-1}	2.23	5.60×10^{-3}
FMRP (Darnell et al., 2011)	1,210	35	19	6.15×10^{-13}	1.74×10^{-3}	2.24	8.98×10^{-3}
RBFOX2 (Weyn-Vanhentenryck et al., 2014)	2,911	48	27	7.10×10^{-9}	1.37×10^{-1}	2.31	2.19×10^{-3}
RBFOX1/3 (Weyn-Vanhentenryck et al., 2014)	3,255	50	27	3.18×10^{-8}	3.10×10^{-1}	2.46	9.35×10^{-4}
Missense constrained (Samocha et al., 2014)	961	23	8	3.13×10^{-7}	4.30×10^{-1}	3.41	2.74×10^{-3}
LoF constrained (Karczewski et al., 2020)	3,264	62	35	1.35×10^{-14}	1.54×10^{-2}	2.53	2.55×10^{-4}
PGC3 broad (Ripke et al., 2020)	426	3	4	5.88×10^{-1}	3.95e−1	0.78	7.51×10^{-1}
PGC3 prioritized (Ripke et al., 2020)	111	1	0	5.49×10^{-1}	1.00	inf	3.05×10^{-1}
PsychENCODE DEG (Gandal et al., 2018)	3,953	35	36	9.31×10^{-2}	1.12×10^{-1}	1.03	9.22×10^{-1}
PsychENCODE TWAS (Wang et al., 2018)	301	4	1	1.68×10^{-1}	8.98×10^{-1}	4.29	1.59×10^{-1}

See also Table S2. ASD, autism spectrum disorder; DBD, developmental brain disorder; DD, developmental delay; DEG, differentially expressed genes; ID, intellectual disability; inf, infinite; LoF, loss of function; OR, odds ratio; TWAS, transcriptome-wide association study.

genes identified by a transcriptome-wide association study (TWAS) derived from GWAS results (Wang et al., 2018); and (2) 3,953 differentially expressed genes (DEGs) identified in post-mortem brain samples of patients with SCZ compared to post-mortem samples from controls (Gandal et al., 2018). Because these results were non-significant, they were not examined further with permutation testing.

Next, we examined the enrichment of our case-only gene list across all Gene Ontology (GO) categories and Panther protein classes (Ashburner et al., 2000; The Gene Ontology Consortium, 2017, 2019; Mi et al., 2019), as well as synaptic components using annotations from SYNGO (Koopmans et al., 2019). As shown in Table 3, the 141 case-only genes were significantly enriched for biological processes related to cell adhesion, and specifically to the cadherin class of proteins. It is noteworthy that the enrichment for cadherins is driven by 3 of the 4 FAT atypical cadherins (*FAT1*, *FAT2*, and *FAT4*), all in different chromosomal regions, appearing on our case-only list (as did their key interacting gene, *DCHS2*); specific mutations observed in these genes are listed in Table S3. By contrast, none of the 18 genes in the cadherin gene set were observed on the list of 148 control-only genes. SYNGO analysis demonstrated that enrichment extended across both presynaptic and postsynaptic genes (Table S4). By contrast, the 148 control-only genes showed no synaptic enrichment (all q > 0.75), no enriched GO categories (all false discovery rate [FDR] > 0.05), and only 1 enriched Panther protein class, in a very small gene set (Hsp90 family chaperone, overlap of 3/9 genes; $p = 6.37 \times 10^{-5}$, FDR = 1.24×10^{-2}).

As shown in Table S2, the lead categories remained significant in the size-matched permutation procedure, and the control-only list was significantly underrepresented in several of these gene sets. In addition, we ran the GO biological processes and PANTHER protein class analyses on 100 permuted gene lists, size matched to the case-only list. For the PANTHER protein class analyses, only 1/100 permutations generated a p value stronger than the one we reported for the cadherin gene set; moreover, 81/100 permutations resulted in no significant (FDR-corrected) protein classes at all. For the GO biological process analyses, permutation testing was somewhat more equivocal; 12/100 permutations generated p values stronger than our top gene set (homophilic cell adhesion via plasma membrane adhesion molecules), although 67/100 permutations resulted in no FDR-significant gene sets.

A damaging URV in *PCDHA3* is observed recurrently in AJ SCZ cases

The foregoing analyses examined singleton URVs only, which have been the primary focus of exome studies in SCZ to date; the SCHEMA dataset of case-only variants contains ~95% singletons, and <1% of all case-only variants in SCHEMA are observed ≥3 times. However, we hypothesized that the AJ population would be more likely than non-founder populations to retain and propagate multiple copies of deleterious variants. Consequently, we merged exome data from our cohort with the AJ SCZ cases (n = 869) and controls (n = 2,415) from SCHEMA to identify individual URVs that were observed ≥3 times in cases (i.e., at least ~1/1,000 allele frequency in the 3,310 AJ case chromosomes available). For further filtering, we used exome data from an additional 1,587 AJ controls from a separate study of longevity. As shown in Table 4, 17 MisLoF URVs were observed in ≥3 cases and 0 controls (nominal p < 0.05 by Fisher's exact test). Notably, the most common variant (observed 5 times, case frequency = 0.15%) is a putatively damaging (Combined Annotation Dependent Depletion [CADD] score = 23.4) missense variant (chr5:140182458:A/G; p.Asn559Ser) in *PCDHA3*, part of the protocadherin gene cluster on chromosome 5. Intriguingly, 2 of the 3 patients in our own dataset who carried this variant had early-onset SCZ (ages at first diagnosis for the 3 patients were 12, 13, and 22, respectively),

Table 3. Hypothesis-free gene set analysis

	No. genes	Overlap	Expected	Fold enrichment	Raw p value	FDR
GO biological process						
Homophilic cell adhesion via plasma membrane adhesion molecules	167	10	1.12	8.92	3.14E−7	4.98E−3
Cell-cell adhesion via plasma-membrane adhesion molecules	256	10	1.72	5.82	1.22E−5	4.84E−2
Cell adhesion	947	19	6.36	2.99	2.32E−5	6.15E−2
Biological adhesion	953	20	6.40	3.13	7.29E−6	3.85E−2
Neurogenesis	1,401	24	9.41	2.55	2.33E−5	5.29E−2
Developmental process	5,765	62	38.71	1.60	2.64E−5	4.66E−2
Nervous system development	2,203	32	14.79	2.16	2.51E−5	4.97E−2
System development	4,317	53	28.99	1.83	3.22E−6	2.55E−2
Multicellular organism development	4,906	56	32.94	1.70	1.47E−5	4.66E−2
Multicellular organismal process	6,985	71	46.90	1.51	3.10E−5	4.92E−2
PANTHER protein class						
Cadherin	18	4	0.12	33.10	1.26E−5	2.46E−3
Cell adhesion molecule	90	5	0.60	8.27	4.40E−4	2.14E−2
Intermediate filament binding protein	15	3	0.10	29.79	2.20E−4	2.15E−2
Intermediate filament	15	3	0.10	29.79	2.20E−4	1.43E−2
Extracellular matrix protein	166	6	1.11	5.38	1.05E−3	4.11E−2

See also Table S2. FDR, false discovery rate; GO, Gene Ontology.

and 1 of these 2 was prescribed clozapine. Of course, these clinical findings must be considered preliminary due to the very small number of patients involved, although the enrichment of very-early-onset cases is statistically significant, given that only 19 other cases in the full cohort had an age of onset $\leq$13 (Fisher's exact p = 0.002).

In vitro characterization of the *PCDHA3* variant

The *PCDHA3* gene is a member of the clustered protocadherin gene family. This family comprises three gene clusters, α, β, and γ, which are stochastically expressed in individual neurons, generating a cell surface "barcode" required for neuronal self-recognition (Canzio and Maniatis, 2019; Mountoufaris et al., 2018; Rubinstein et al., 2017; Wu and Jia, 2021). Individual PCDH isoforms from all three gene clusters form nearly random cis-dimers ($\alpha\beta$, $\alpha\gamma$, β/γ,β/β,γ/γ – but not α/α), which assemble on the cell surface, and engage in highly specific homophilic interactions with cis-dimers on apposing plasma membranes of neurites, which leads to self-recognition and neurite repulsion (Goodman et al., 2016, 2017; Rubinstein et al., 2015; Thu et al., 2014). Mouse studies have shown that mutations in the PCDH gene cluster can result in neural circuit deficits and behavioral phenotypes associated with neuropsychiatric or neurodevelopmental disorders (Chen et al., 2017; Katori et al., 2009, 2017; Mountoufaris et al., 2017).

The p.Asn559Ser URV in *PCDHA3* is part of the DxND motif within the EC5 domain, and lies within a calcium-coordinating residue that is highly conserved across all clustered protocadherin proteins. The EC5 domain is required for cis-dimerization, and PCDHα isoforms must dimerize with either PCDHβ or γ isoforms to localize on the plasma membrane (Goodman et al., 2017; Thu et al., 2014). To investigate the possibility that the p.Asn559Ser URV affects membrane localization, we examined the localization of mCherry-tagged PCDHα3 wild-type and PCDHα3(N559S) mutant isoforms in non-neuronal K562 cells when co-expressed with the PCDHγC3ΔEC1 carrier isoform (Thu et al., 2014). These cells, which do not endogenously express PCDH isoforms, have been used extensively to study PCDH homophilic interactions in prior work (Thu et al., 2014); notably, this previous work has demonstrated the most robust effects for PCDH isoforms lacking the intracellular domain (ICD) of the protein. In the present report, we show primary results for isoforms without the ICD in Figure 2 and comparable results for full-length isoforms in Figure S2; results did not substantively differ according to the expression of the ICD, but they were somewhat more robust for isoforms lacking the ICD.

We found that the wild-type PCDHα3 protein localized primarily to the surface of the K562 cells, although the mCherry signal could also be detected in the cytoplasm (Figure 2A and S2A). By contrast, the PCDHα3(N559S) protein was not detected at the cell surface. Rather, it localized to the cytoplasm, suggesting that the PCDHα3 URV may disrupt dimerization with the PCDHγC3 carrier isoform, which is required for plasma membrane localization, and thus fails to be transported to the cell surface. Alternatively, the amino acid substitution may destabilize the protein, causing it to form aggregates in the cytoplasm.

To determine whether the mislocalization of the PCDHα3(N559S) prevents the protein from engaging in the homophilic interactions required for self-recognition, we performed a cell aggregation assay in K562 cells expressing either the mCherry-tagged wild-type PCDHα3 or the PCDHα3(N559S) mutant isoform along with the PCDHγC3ΔEC1 carrier isoform. As shown in Figure 2B (and Figure S2B), K562 cells expressing the wild-type PCDHα3 protein formed homophilic aggregates

Table 4. Recurrent (≥3 observations) MisLoF URVs in AJ cases

No. cases	Gene	Chr	Position	Ref	Alt	Impact	Coding position	Protein position	AA change
5	*PCDHA3*	5	140182458	A	G	missense	1,676	559	N/S
4	*ACBD6*	1	180382593	T	G	missense	481	161	N/H
4	*IGF1R*	15	99251324	C	T	missense	628	210	R/C
3	*ATAD3C*	1	1389777	CA	C	frameshift	276	92	T/X
3	*ACOX3*	4	8412048	G	A	missense	578	193	A/V
3	*FIGNL1*	7	50513662	G	A	missense	1,324	442	P/S
3	*CEP104*	1	3740011	T	C	missense	2,480	827	H/R
3	*UBR4*	1	19426131	A	G	missense	13,262	4,421	V/A
3	*KLHL30*	2	239049577	T	C	missense	182	61	M/T
3	*TBX18*	6	85457687	G	C	missense	860	297	S/C
3	*CACNA2D1*	7	81599254	A	C	missense	2,287	763	F/V
3	*TENM4*	11	78369465	A	T	missense	7,948	2,650	S/T
3	*TNFRSF1A*	12	6440026	C	A	missense	618	206	E/D
3	*BCAT1*	12	25002856	T	C	missense	574	192	K/E
3	*LCP1*	13	46722531	C	T	missense	934	312	E/K
3	*PCSK2*	20	17240934	A	T	missense	227	76	K/M
3	*PTK6*	20	62164958	A	G	missense	616	206	F/L

AA, amino acid; AJ, Ashkenazi Jewish; Alt, alternate allele; Chr, chromosome; MisLoF, missense or loss of function; Ref, reference allele; URV, ultra-rare variant.

(consistent with previous findings; Thu et al., 2014), while those expressing PCDHα3(N559S) proteins failed to form aggregates, remaining as individual, non-aggregating cells in suspension. These results indicate that the *PCDHα3* URV identified in SCZ cases may prevent the formation of cis-dimers, which are required for membrane localization, homophilic interactions, and proper self-avoidance. Thus, the absence of Pcdhα heterodimerization may interfere with normal cell surface lattice formation, potentially resulting in neural circuit deficits in individuals bearing the p.Asn559Ser variant.

Damaging rare variants escape purifying selection in a founder population

Given that we were able to detect numerous recurrent case-only variants, despite our relatively small sample size compared to SCHEMA, we sought to model the parameters affecting the persistence of deleterious alleles in a founder population. We initiated a series of simulations based on our prior estimates of the size (N = 300) and timing (30 generations ago) of the Ashkenazi bottleneck (Carmi et al., 2014; Palamara et al., 2012) and the population-based estimates (Power et al., 2013) of reduced fecundity in SCZ (fecundity ratio ~0.5 for females and ~0.25 for males). We then generated 10,000 simulations for each of a series of variations on these parameters (Table S5) to model the odds of a deleterious variant, present in a single individual at the time of the bottleneck, escaping extinction to persist in the present AJ population. In addition to the parameters noted above, simulations were performed as a function of penetrance for SCZ. As shown in Figure 3A, between 30% and 50% of such variants escape extinction within the range of penetrance expected, given the genetic architecture of the disorder (Sullivan et al., 2012).

Based on these results and given the rough estimates of the AJ population today (~10 million) and the prevalence of SCZ (~1%), we could then estimate the total number of case and control carriers expected in the contemporary AJ population for each scenario. These calculations allowed us to determine the power of our homogenous discovery cohort (Figure 4) combined with the SCHEMA AJ cohort (total N = 1,637 cases and 2,878 controls) to detect a given variant at exome-wide significance (Figure 3B), which generally ranged between 5% and 20%. However, we also calculated that a slightly larger study of 5,000 cases and 9,000 controls would have power of ~20%–40% to detect any individual variant in the range of realistic penetrance (Figure 3C), and would therefore have >80% power to detect at least 1 variant, assuming there are at least 7 such variants circulating in the population (i.e., even if only 5% of our case-only list were true positives). Such an assumption is likely to be extremely conservative, given the estimated mutational target of ≥1,000 genes (Nguyen et al., 2017; Purcell et al., 2014), the replication of the SCHEMA results in our dataset, the significant findings documented in Table 2, and the long list of variants at greater than doubleton frequency documented in Table 4.

DISCUSSION

The present study demonstrates the enhanced power available to genetic studies performed in populations enriched for rare variants, consistent with recent work in SCZ (Gulsuner et al., 2020) and other phenotypes (Locke et al., 2019; Rivas et al., 2018; Selvan et al., 2020). We further reduced background heterogeneity by using a strict filter against all of the variants reported in non-neuropsychiatric samples across the two largest publicly available sequencing datasets, gnomAD (Karczewski

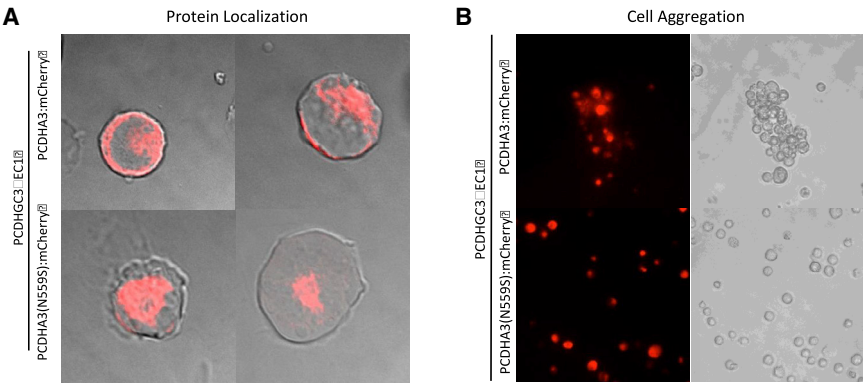

Figure 2. *PCDHA3* p.Asn559Ser variant causes mislocalization of PCDHα3 protein and disrupts homophilic aggregation *in vitro*

(A) Expression of wild-type PCDHα3 or PCDHα3(N559S) mCherry fusion proteins (both excluding the intracellular domain), along with PcdhγC3ΔEC1 carrier protein, in non-neuronal K562 cells to assess cell surface localization. Wild-type mCherry-labeled PCDHα3 protein localized primarily to the cell surface, although some mCherry signal could also be detected in the cytoplasm. By contrast, the PCDHα3(N559S) protein was not detected at the cell surface; rather, it localized to the cytoplasm.

(B) Cell aggregation assay in K562 cells expressing either wild-type PCDHα3 or PCDHα3(N559S) mCherry fusion proteins (both excluding the intracellular domain), along with PcdhγC3ΔEC1 carrier protein to assess PCDH homophilic interaction. K562 cells expressing the wild-type PCDHα3 protein formed homophilic cell-cell aggregates, while those expressing PCDHα3(N559S) proteins failed to aggregate, remaining as individual, non-aggregating cells in suspension.

et al., 2020) and TOPMed (The NHLBI Trans-Omics for Precision Medicine TOPMed Whole Genome Sequencing Program, 2018). Thus, despite relatively modest sample sizes, the present study was able to replicate several previously identified SCZ-associated genes (*SETD1A*, *TRIO*, *XPO7*) (Singh et al., 2016, 2020) and gene sets (synaptic, DBD-related, and constrained genes) (Genovese et al., 2016; Gonzalez-Mantilla et al., 2016; Gulsuner et al., 2020; Nguyen et al., 2017; Purcell et al., 2014), with these analyses serving as a positive control for our approach. Beyond these replications, we were also able to make several additional discoveries, as described below.

We identified several gene sets associated with SCZ that have not been reported in previous studies. The strongest statistical signal was observed for cell adhesion processes, especially the cadherin family genes (Table 3). Cadherins form calcium-dependent adherence junctions at the synapse and are involved in both neuronal migration and mature synaptic activity (Friedman et al., 2015). Surprisingly, cadherins have not received much attention in the SCZ genetics literature, despite the considerable recent focus on both calcium activity and synaptic proteins (Nanou and Catterall, 2018). While there are >100 different proteins in the cadherin superfamily (Friedman et al., 2015), it is noteworthy that 3 of the 4 FAT atypical cadherins, all in different chromosomal regions, appeared on our case-only list (as did their key interacting gene, *DCHS2*). These genes are specifically involved in regulating microtubule polarity, thereby directing cellular migration in the developing nervous system (Avilés and Goodrich, 2017; Fulford and McNeill, 2020). Homozygous mutations in *FAT4* cause Van Maldergem syndrome, a recessive intellectual disability marked by periventricular neuronal heterotopia (Cappello et al., 2013), while mutations in *FAT1* have been observed in ASD (Cukier et al., 2014).

Relatedly, we observed a single missense variant in a protocadherin gene (*PCDHA3*) at a higher rate of recurrence (5 observations) in cases than any other ultra-rare (i.e., not in healthy individuals) variant in the published SCZ literature (although it should be noted that 1 splice acceptor variant in *SETD1A* appears 6 times in the SCHEMA database). *PCDHA3* is a member of the PCDH gene cluster, which, in conjunction with the stochastic expression of PCDHα, β, and γ isoforms, generates a "molecular barcode" on the cell surface required for neuronal self-recognition (Canzio and Maniatis, 2019). Individual PCDH proteins form nearly random cis-dimers, which are transported to the plasma membrane and engage in *trans* with homophilic partners on the apposing cell surface. Ultimately, the assembly of PCDH cis-/trans-tetramers on the plasma membrane creates a lattice-like structure, which may be required for repulsion and self-avoidance (Brasch et al., 2019).

The missense variant in the *PCDHA3* gene is located in the EC5 domain of the protein, which is critical for cis-dimerization and cell surface localization of PCDHα isoforms (Thu et al., 2014). Consistent with the role of the EC5 domain, we found that the PCDHα3 variant protein failed to localize on the surface of K562 cells, instead accumulating as cytoplasmic aggregates. Such an effect in primary neurons may result in the failure of the PCDHα3 variant protein to localize to the plasma membrane, preventing homophilic engagement and disrupting the assembly of the PCDH protein lattice, which would cause deficits in self-avoidance. In patients carrying this variant, these effects would likely have a significant impact on neural circuit formation. For example, previous studies have shown that the PcdhαC2 protein is required for normal serotonergic neuron wiring (Chen et al., 2017; Katori et al., 2009, 2017). Moreover, the disruption of PCDHα isoforms is consistent with other findings in SCZ; for example, altered expression of protocadherins (including *PCDHA3* in SCZ has been implicated by a recent transcriptome-wide association study of both prefrontal cortex and hippocampus (Collado-Torres et al., 2019). This observation is supported by functional interrogation of the SCZ GWAS locus encompassing the PCDHA gene cluster in human induced pluripotent stem cell (hiPSC)-derived neural cells (Rajarajan et al., 2018). In addition, cortical interneurons derived from iPSCs of patients with SCZ showed reduced *PCDHA3* expression compared to similarly derived interneurons from controls (Shao et al., 2019). The latter study further demonstrated that reduced protocadherin expression was associated with deficient synaptic arborization in both rodent and iPSC-derived human interneurons (but not

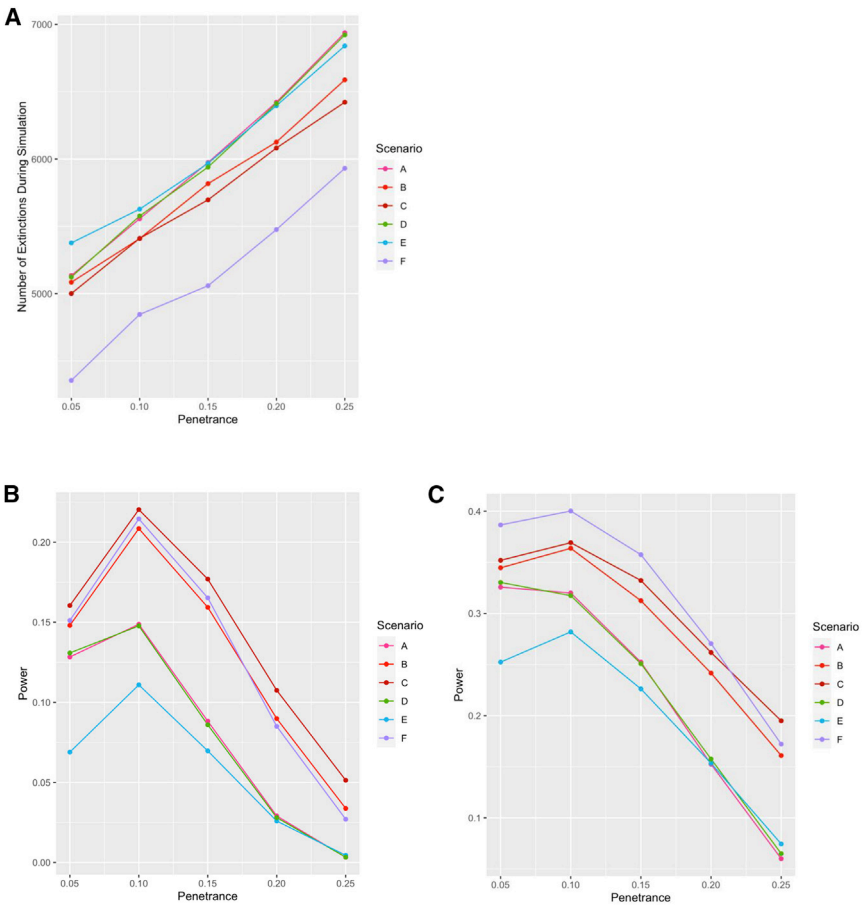

Figure 3. Damaging rare variants escape purifying selection in a founder population

(A) Under a range of scenarios consistent with known population and disease parameters, as many as half of all damaging variants remain in a rapidly expanding founder population. Scenario A represents the best estimates of the size and timing of the AJ population bottleneck based on our prior work (Carmi et al., 2014; Palamara et al., 2012), and effects of schizophrenia on fecundity based on the work of Power et al. (2013); other scenarios (detailed in Table S5) test the sensitivity of the model to variations in fecundity effects (scenarios B and C), size of the bottleneck (scenario D), and number of generations since the bottleneck (scenarios E and F). The y axis denotes the number of extinctions out of 10,000 simulations for each condition.

(B) Given rough estimates of the AJ population today (~10 million) and the prevalence of schizophrenia (~1%), the power of our discovery cohort combined with the SCHEMA AJ cohort (N = 1,637 cases and 2,878 controls) to detect a given variant at exome-wide significance generally ranged between 5% and 20%. Notably, while power tends to decrease at higher levels of penetrance due to increased variant extinction (as shown in A), there also tends to be a decrease in power at the lowest level of penetrance due to an increased frequency of these variants appearing in controls.

(C) For a slightly larger study (N = 5,000 cases and 9,000 controls), power would be ~20%–40% to detect any individual variant in the range of realistic penetrance, resulting in >80% power to detect at least 1 variant, assuming there are at least 7 such variants circulating in the population.

glutamatergic neurons), and that these deficits could be reversed by treatment with an inhibitor of protein kinase C (Shao et al., 2019).

When our samples were combined with AJ patients from the SCHEMA database, the *PCDHA3* missense variant was observed in 0.3% of all AJ cases in the present study. While unusually high for a SCZ-associated URV, this carrier rate is low compared to the ~4% rate observed for the most common *BRCA1* founder variant in AJ breast cancer cases (King et al., 2003), and the ~15% rate of the *LRRK2* G2019S variant among AJ patients with Parkinson's disease (Correia Guedes et al., 2010). These latter disorders have onset in late life, and therefore susceptibility alleles for these diseases are not under the strong purifying selection affecting genes for SCZ (Pardiñas et al., 2018), a disorder that results in markedly reduced fecundity (Power et al., 2013). Nevertheless, our simulations demonstrated the limits of purifying selection in a founder population with a tight bottleneck. One-third to half of all damaging variants escape purifying selection, and these variants tend to become surprisingly frequent in the context of a rapidly expanding population, as described previously for the Finnish population (Wang et al., 2014). Consequently, ascertainment of additional samples from founder populations can be a highly cost-effective way of rapidly enhancing the power of rare variant studies (Locke et al., 2019).

The overlap of our SCZ case-only gene list with gene sets derived from DBDs was notable, insofar as exome studies in these disorders have been more well powered than SCZ studies to date (Myers et al., 2020). Consequently, the overlapping genes indicated in the first three rows of Table 2 (and especially the first row, which remained significant after permutation testing) have a strong prior probability of association, especially given evidence that rare SNVs (e.g., *SETD1A*) (Singh et al., 2016) and CNVs (Kirov, 2015) tend to be shared across SCZ and other neurodevelopmental disorders. In the present study, case-only variants in the 11 genes overlapping the prior report of DBDs (Gonzalez-Mantilla et al., 2016) (*ASXL3*, *BIRC6*, *DIP2A*, *DST*, *LAMA2*, *NSD1*, *PCDH15*, *SETBP1*, *SETD1A*, *TRIO*, *WDFY3*) were overwhelmingly (10:1 ratio) missense rather than LoF; by contrast, the DBD list was generated from prior reports of LoF variation exclusively. Thus, it is possible that our findings represent allelic series at these genes, in which more damaging variants are associated with more severe clinical phenotypes emerging in early childhood (Shohat et al., 2017).

Similarly, we identified 5 cases (and no controls) with novel missense variants in *TSC2*, a gene in which mutations (primarily LoF) are known to cause tuberous sclerosis (TS). TS is an autosomal dominant disorder marked by hamartomas across multiple organs, potentially including the brain (Henske et al., 2016). Case reports of psychotic features in TS patients have

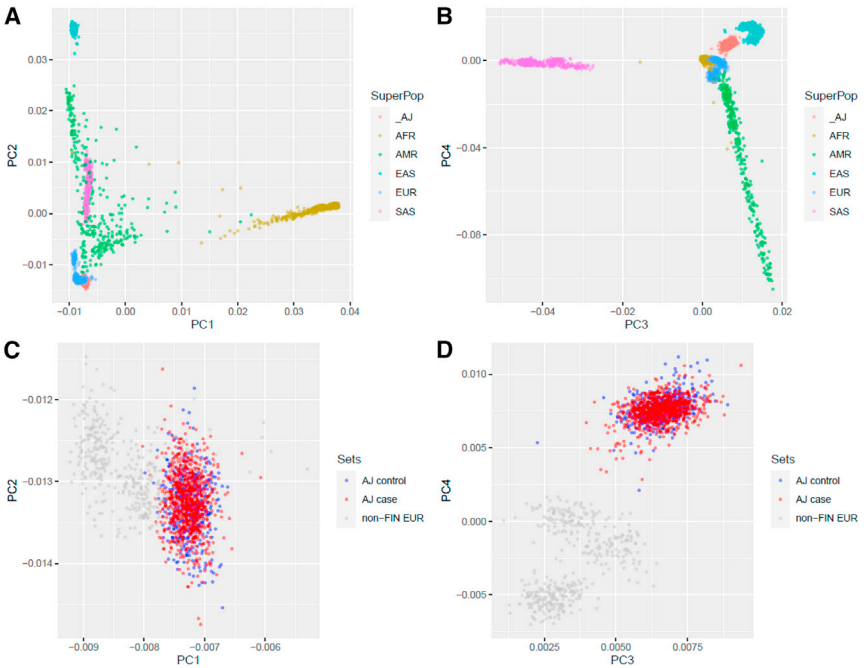

Figure 4. PCA plots of AJ subjects compared to global populations

AJ subjects demonstrate clear separation from other groups at the global scale (A and B) and within populations of European ancestry (C and D; gray dots represent CEU, GBR, IBS, and TSI populations from 1000 Genomes). Importantly, no clear distinction can be drawn between AJ cases and controls (C and D, red and blue dots, respectively).

proliferated for decades (Herkert et al., 1972); a recent survey of a large international cohort of TS patients identified psychosis in 11% of adults (de Vries et al., 2018). Since the affected cases in the present study were not noted in their clinical report to have TS, our results suggest that SCZ can be the primary presenting feature of *TSC2* mutations.

In the last 15 years, genetic research in SCZ has given consistent support to the long-posited liability-threshold model (Gottesman and Shields, 1967; Kendler, 2015; McGue et al., 1983; Smeland et al., 2020), which states that manifestation of illness requires that the additive total of risk factors (including genetic and environmental) crosses an (unknown) threshold. While most SCZ research to date has focused on the total burden of common genetic variants, as captured by the polygenic risk score (Lee et al., 2012; van Rheenen et al., 2019; Purcell et al., 2009), the model suggests that rare variants contribute in the same manner, albeit with much greater individual weighting (penetrance) (Richards et al., 2016). Supporting evidence has come from four very recent studies of SCZ patients carrying known, high-penetrance CNVs (Bergen et al., 2019; Cleynen et al., 2020; Davies et al., 2020; Taniguchi et al., 2020). These studies show that patients with highly penetrant CNVs have lower common-variant PRSs compared to patients not carrying a known CNV, presumably because the CNV has already pushed them closer to the threshold for illness. In the present study, we have demonstrated, for the first time, a similar inverse correlation between common-variant PRS and rare variant burden indexed by missense and LoF single-nucleotide changes and small indels. Interestingly, the cases that carried MisLoF URVs in the 141 case-only genes did not show earlier age at onset or more severe course of illness, suggesting that these variants do not generally manifest in a fundamentally different illness relative to other additive risk factors. However, there was initial evidence that the recurrent *PCDHA3* mutation may convey risk for an early-onset form of SCZ.

This study had several limitations, most notably that the sample size was relatively small for a genetic association study; however, we demonstrated that the AJ population is enriched for rare variants and has substantially greater power than comparably sized studies of outbred populations. Moreover, we were able to use external, well-powered datasets (i.e., SCHEMA and various studies of neurodevelopmental disorders) as validation/replication, and our gene set results for synaptic and constrained genes served as a positive control for our approach. Relatedly, although we used whole-genome sequencing to obtain our data, we restricted our analysis to the exome to make use of the largest possible set of samples for both purposes of these comparisons and filtering of variants. Finally, we did not have neuropsychological testing data available for our cases, so we were unable to differentiate the relative contribution of URVs to cognitive deficits in our samples (Singh et al., 2017). Because of the enhanced power available in the AJ population, future rare variant studies in well-characterized AJ samples can be useful in overcoming a common limitation in large-scale genetic studies of SCZ, namely, understanding the clinical impact of URVs on phenotypic expression.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects
  - Cell lines
- METHOD DETAILS
  - Sequencing and variant calling pipeline
  - Defining ultra-rare variants (URVs) in TAGC samples
  - Plasmid generation

- Cell aggregation assay
- Immunostaining

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Common variant polygenic risk score (PRS)
  - Assigning novel URVs to genes and defining "case-only" and "control-only" genes
  - Replication of SCHEMA genes
  - Gene set analyses
  - Modeling the effects of purifying selection

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2021.03.004.

ACKNOWLEDGMENTS

The authors are extremely grateful to Soren Germer, PhD and his team at the New York Genome Center for performing the Illumina sequencing. We acknowledge financial support from the Human Frontier Science Program (to S.C.); NIH research grants AG042188 (to G.A.), DK62429, DK062422, DK092235 (to J.H.C.), NS050487, NS060113 (to L.N.C.), AG021654, AG027734 (to N.B.), MH089964, MH095458, MH084098 (to T.L.), MH10857, MH114817 (to T.M.), and CA121852 (computational infrastructure, to I. Pe'er); NSF research grants 08929882 and 0845677 (to I. Pe'er); the Rachel and Lewis Rudin Foundation (to H.O.); the Northwell Health Foundation (to T.L.); the Brain & Behavior Foundation (to T.L.); the US-Israel Binational Science Foundation (to T.L. and A.D.); the LUNGevity Foundation (to Z.H.G.); the New York Crohn's Disease Foundation (to I. Peter); Edwin & Caroline Levy and Joseph & Carol Reich (to S.B.); the Parkinson's Disease Foundation (to L.N.C.); the Sharon Levine Corzine Cancer Research Fund (to K.O.); and the Andrew Sabin Family Research Fund (to K.O.).

AUTHOR CONTRIBUTIONS

T.L. and I. Pe'er led the analysis, and T.L. and J.Y. led the writing of the manuscript. J.Y. and R.R.K. conducted the primary analyses of the exome data, with assistance from M.L. and S.C. E.F. and T.M. designed and E.F. conducted the *in vitro* analyses. T.L. led the funding of the study. T.L., A.D., G.A., D.B.-A., N.B., and L.N.C. provided the samples and conducted the lab work. T.L., I. Pe'er, N.B., S.B., A.D., J.H.C., L.N.C., Z.H.G., J.V., R.J.K., S.L., K.O., H.O., L.J.O., I. Peter, A.M.K., and G.A. initiated and designed the study and provided the funding.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 25, 2020
Revised: December 16, 2020
Accepted: March 1, 2021
Published: March 22, 2021

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., and Ostrer, H. (2010). Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. Am. J. Hum. Genet. *86*, 850–859.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Avilés, E.C., and Goodrich, L.V. (2017). Configuring a robust nervous system with Fat cadherins. Semin. Cell Dev. Biol. *69*, 91–101.

Baskovich, B., Hiraki, S., Upadhyay, K., Meyer, P., Carmi, S., Barzilai, N., Darvasi, A., Ozelius, L., Peter, I., Cho, J.H., et al. (2016). Expanded genetic screening panel for the Ashkenazi Jewish population. Genet. Med. *18*, 522–528.

Bergen, S.E., Ploner, A., Howrigan, D., O'Donovan, M.C., Smoller, J.W., Sullivan, P.F., Sebat, J., Neale, B., and Kendler, K.S.; CNV Analysis Group and the Schizophrenia Working Group of the Psychiatric Genomics Consortium (2019). Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia. Am. J. Psychiatry *176*, 29–35.

Brasch, J., Goodman, K.M., Noble, A.J., Rapp, M., Mannepalli, S., Bahna, F., Dandey, V.P., Bepler, T., Berger, B., Maniatis, T., et al. (2019). Visualization of clustered protocadherin neuronal self-recognition complexes. Nature *569*, 280–283.

Canzio, D., and Maniatis, T. (2019). The generation of a protocadherin cell-surface recognition code for neural circuit assembly. Curr. Opin. Neurobiol. *59*, 213–220.

Cappello, S., Gray, M.J., Badouel, C., Lange, S., Einsiedler, M., Srour, M., Chitayat, D., Hamdan, F.F., Jenkins, Z.A., Morgan, T., et al. (2013). Mutations in genes encoding the cadherin receptor-ligand pair DCHS1 and FAT4 disrupt cerebral cortical development. Nat. Genet. *45*, 1300–1308.

Carmi, S., Hui, K.Y., Kochav, E., Liu, X., Xue, J., Grady, F., Guha, S., Upadhyay, K., Ben-Avraham, D., Mukherjee, S., et al. (2014). Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat. Commun. *5*, 4835.

Chen, W.V., Nwakeze, C.L., Denny, C.A., O'Keeffe, S., Rieger, M.A., Mountoufaris, G., Kirner, A., Dougherty, J.D., Hen, R., Wu, Q., and Maniatis, T. (2017). Pcdhαc2 is required for axonal tiling and assembly of serotonergic circuitries in mice. Science *356*, 406–411.

Cleynen, I., Engchuan, W., Hestand, M.S., Heung, T., Holleman, A.M., Johnston, H.R., Monfeuga, T., McDonald-McGinn, D.M., Gur, R.E., Morrow, B.E., et al.; International 22q11.2DS Brain and Behavior Consortium (2020). Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. Mol. Psychiatry. https://doi.org/10.1038/s41380-020-0654-3.

Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A., and Eichler, E.E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat. Genet. *51*, 106–116.

Collado-Torres, L., Burke, E.E., Peterson, A., Shin, J., Straub, R.E., Rajpurohit, A., Semick, S.A., Ulrich, W.S., Price, A.J., Valencia, C., et al.; BrainSeq Consortium (2019). Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. Neuron *103*, 203–216.e8.

Correia Guedes, L., Ferreira, J.J., Rosa, M.M., Coelho, M., Bonifati, V., and Sampaio, C. (2010). Worldwide frequency of G2019S LRRK2 mutation in Parkinson's disease: a systematic review. Parkinsonism Relat. Disord. *16*, 237–242.

Cukier, H.N., Dueker, N.D., Slifer, S.H., Lee, J.M., Whitehead, P.L., Lalanne, E., Leyva, N., Konidari, I., Gentry, R.C., Hulme, W.F., et al. (2014). Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. Mol. Autism *5*, 1.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y.S., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., et al. (2011). FMRP stalls

ribosomal translocation on mRNAs linked to synaptic function and autism. Cell *146*, 247–261.

Davies, R.W., Fiksinski, A.M., Breetvelt, E.J., Williams, N.M., Hooper, S.R., Monfeuga, T., Bassett, A.S., Owen, M.J., Gur, R.E., Morrow, B.E., et al.; International 22q11.2 Brain and Behavior Consortium (2020). Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. Nat. Med. *26*, 1912–1918.

de Vries, P.J., Belousova, E., Benedik, M.P., Carter, T., Cottin, V., Curatolo, P., Dahlin, M., D'Amato, L., d'Augères, G.B., Ferreira, J.C., et al.; TOSCA Consortium and TOSCA Investigators (2018). TSC-associated neuropsychiatric disorders (TAND): findings from the TOSCA natural history study. Orphanet J. Rare Dis. *13*, 157.

Friedman, L.S., Szabo, C.I., Ostermeyer, E.A., Dowd, P., Butler, L., Park, T., Lee, M.K., Goode, E.L., Rowell, S.E., and King, M.C. (1995). Novel inherited mutations and variable expressivity of BRCA1 alleles, including the founder mutation 185delAG in Ashkenazi Jewish families. Am. J. Hum. Genet. *57*, 1284–1297.

Friedman, L.G., Benson, D.L., and Huntley, G.W. (2015). Cadherin-based transsynaptic networks in establishing and modifying neural connectivity. Curr. Top. Dev. Biol. *112*, 415–465.

Fulford, A.D., and McNeill, H. (2020). Fat/Dachsous family cadherins in cell and tissue organisation. Curr. Opin. Cell Biol. *62*, 96–103.

Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al.; PsychENCODE Consortium (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*, eaat8127.

Ganna, A., Satterstrom, F.K., Zekavat, S.M., Das, I., Kurki, M.I., Churchhouse, C., Alfoldi, J., Martin, A.R., Havulinna, A.S., Byrnes, A., et al.; GoT2D/T2D-GENES Consortium; SIGMA Consortium Helmsley IBD Exome Sequencing Project; FinMetSeq Consortium; iPSYCH-Broad Consortium (2018). Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. Am. J. Hum. Genet. *102*, 1204–1211.

Genovese, G., Fromer, M., Stahl, E.A., Ruderfer, D.M., Chambert, K., Landén, M., Moran, J.L., Purcell, S.M., Sklar, P., Sullivan, P.F., et al. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat. Neurosci. *19*, 1433–1441.

Gonzalez-Mantilla, A.J., Moreno-De-Luca, A., Ledbetter, D.H., and Martin, C.L. (2016). A Cross-Disorder Method to Identify Novel Candidate Genes for Developmental Brain Disorders. JAMA Psychiatry *73*, 275–283.

Goodman, K.M., Rubinstein, R., Thu, C.A., Bahna, F., Mannepalli, S., Ahlsén, G., Rittenhouse, C., Maniatis, T., Honig, B., and Shapiro, L. (2016). Structural Basis of Diverse Homophilic Recognition by Clustered α- and β-Protocadherins. Neuron *90*, 709–723.

Goodman, K.M., Rubinstein, R., Dan, H., Bahna, F., Mannepalli, S., Ahlsén, G., Aye Thu, C., Sampogna, R.V., Maniatis, T., Honig, B., and Shapiro, L. (2017). Protocadherin *cis*-dimer architecture and recognition unit diversity. Proc. Natl. Acad. Sci. USA *114*, E9829–E9837.

Gottesman, I.I., and Shields, J. (1967). A polygenic theory of schizophrenia. Proc. Natl. Acad. Sci. USA *58*, 199–205.

Gratten, J., Wray, N.R., Keller, M.C., and Visscher, P.M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat. Neurosci. *17*, 782–790.

Guha, S., Rosenfeld, J.A., Malhotra, A.K., Lee, A.T., Gregersen, P.K., Kane, J.M., Pe'er, I., Darvasi, A., and Lencz, T. (2012). Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. Genome Biol. *13*, R2.

Guha, S., Rees, E., Darvasi, A., Ivanov, D., Ikeda, M., Bergen, S.E., Magnusson, P.K., Cormican, P., Morris, D., Gill, M., et al.; Molecular Genetics of Schizophrenia Consortium; Wellcome Trust Case Control Consortium 2 (2013). Implication of a rare deletion at distal 16p11.2 in schizophrenia. JAMA Psychiatry *70*, 253–260.

Gulsuner, S., Stein, D.J., Susser, E.S., Sibeko, G., Pretorius, A., Walsh, T., Majara, L., Mndini, M.M., Mqulwana, S.G., Ntola, O.A., et al. (2020). Genetics of schizophrenia in the South African Xhosa. Science *367*, 569–573.

Henske, E.P., Jóźwiak, S., Kingswood, J.C., Sampson, J.R., and Thiele, E.A. (2016). Tuberous sclerosis complex. Nat. Rev. Dis. Primers *2*, 16035.

Herkert, E.E., Wald, A., and Romero, O. (1972). Tuberous sclerosis and schizophrenia. Dis. Nerv. Syst. *33*, 439–445.

Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T.M., Nordentoft, M., and Glenthøj, B. (2018). Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. Biol. Psychiatry *83*, 492–498.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Katori, S., Hamada, S., Noguchi, Y., Fukuda, E., Yamamoto, T., Yamamoto, H., Hasegawa, S., and Yagi, T. (2009). Protocadherin-alpha family is required for serotonergic projections to appropriately innervate target brain areas. J. Neurosci. *29*, 9137–9147.

Katori, S., Noguchi-Katori, Y., Okayama, A., Kawamura, Y., Luo, W., Sakimura, K., Hirabayashi, T., Iwasato, T., and Yagi, T. (2017). Protocadherin-αC2 is required for diffuse projections of serotonergic axons. Sci. Rep. *7*, 15908.

Kendler, K.S. (2015). A joint history of the nature of genetic variation and the nature of schizophrenia. Mol. Psychiatry *20*, 77–83.

Kenny, E.E., Pe'er, I., Karban, A., Ozelius, L., Mitchell, A.A., Ng, S.M., Erazo, M., Ostrer, H., Abraham, C., Abreu, M.T., et al. (2012). A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. PLoS Genet. *8*, e1002559.

King, M.-C., Marks, J.H., and Mandell, J.B.; New York Breast Cancer Study Group (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science *302*, 643–646.

Kirov, G. (2015). CNVs in neuropsychiatric disorders. Hum. Mol. Genet. *24* (R1), R45–R49.

Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M.P., Cornelisse, L.N., Farrell, R.J., Goldschmidt, H.L., Howrigan, D.P., et al. (2019). SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. Neuron *103*, 217–234.e4.

Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.

Lam, M., Chen, C.-Y., Li, Z., Martin, A.R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B.C., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Indonesia Schizophrenia Consortium; Genetic REsearch on schizophreniA neTwork-China and the Netherlands (GREAT-CN) (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat. Genet. *51*, 1670–1678.

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R.; Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS) (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.

Lencz, T., Guha, S., Liu, C., Rosenfeld, J., Mukherjee, S., DeRosse, P., John, M., Cheng, L., Zhang, C., Badner, J.A., et al. (2013). Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. Nat. Commun. *4*, 2739.

Lencz, T., Yu, J., Palmer, C., Carmi, S., Ben-Avraham, D., Barzilai, N., Bressman, S., Darvasi, A., Cho, J.H., Clark, L.N., et al. (2018). High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. Hum. Genet. *137*, 343–355.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics *30*, 2843–2851.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, B., and Leal, S.M. (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet. *5*, e1000481.

Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S., et al.; FinnGen Project (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature *572*, 323–328.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al.; Schizophrenia Working Group of Psychiatric Genomics Consortium (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.

Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., et al.; Psychosis Endophenotypes International Consortium; CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat. Genet. *49*, 27–35.

McGue, M., Gottesman, I.I., and Rao, D.C. (1983). The transmission of schizophrenia under a multifactorial threshold model. Am. J. Hum. Genet. *35*, 1161–1178.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. *47* (D1), D419–D426.

Mountoufaris, G., Chen, W.V., Hirabayashi, Y., O'Keeffe, S., Chevee, M., Nwakeze, C.L., Polleux, F., and Maniatis, T. (2017). Multicluster Pcdh diversity is required for mouse olfactory neural circuit assembly. Science *356*, 411–414.

Mountoufaris, G., Canzio, D., Nwakeze, C.L., Chen, W.V., and Maniatis, T. (2018). Writing, Reading, and Translating the Clustered Protocadherin Cell Surface Recognition Code for Neural Circuit Assembly. Annu. Rev. Cell Dev. Biol. *34*, 471–493.

Myers, S.M., Challman, T.D., Bernier, R., Bourgeron, T., Chung, W.K., Constantino, J.N., Eichler, E.E., Jacquemont, S., Miller, D.T., Mitchell, K.J., et al. (2020). Insufficient Evidence for "Autism-Specific" Genes. Am. J. Hum. Genet. *106*, 587–595.

Nanou, E., and Catterall, W.A. (2018). Calcium Channels, Synaptic Plasticity, and Neuropsychiatric Disease. Neuron *98*, 466–481.

Nguyen, H.T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L.M., Munoz-Manchado, A.B., Ruderfer, D.M., Genovese, G., Fromer, M., Xu, X., et al. (2017). Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Med. *9*, 114.

Ozelius, L.J., Senthil, G., Saunders-Pullman, R., Ohmann, E., Deligtisch, A., Tagliati, M., Hunt, A.L., Klein, C., Henick, B., Hailpern, S.M., et al. (2006). LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews. N. Engl. J. Med. *354*, 424–425.

Palamara, P.F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. *91*, 809–822.

Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., et al.; GERAD1 Consortium; CRESTAR Consortium (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat. Genet. *50*, 381–389.

Pirooznia, M., Wang, T., Avramopoulos, D., Valle, D., Thomas, G., Huganir, R.L., Goes, F.S., Potash, J.B., and Zandi, P.P. (2012). SynaptomeDB: an ontology-based knowledgebase for synaptic genes. Bioinformatics *28*, 897–899.

Power, R.A., Kyaga, S., Uher, R., MacCabe, J.H., Långström, N., Landen, M., McGuffin, P., Lewis, C.M., Lichtenstein, P., and Svensson, A.C. (2013). Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. JAMA Psychiatry *70*, 22–30.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190.

Rajarajan, P., Borrman, T., Liao, W., Schrode, N., Flaherty, E., Casiño, C., Powell, S., Yashaswini, C., LaMarca, E.A., Kassim, B., et al. (2018). Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science *362*, eaat4311.

Resin, J. (2020). A Simple Algorithm for Exact Multinomial Tests. ArXiv, 2008.12682v1 http://arxiv.org/abs/2008.12682v1.

Richards, A.L., Leonenko, G., Walters, J.T., Kavanagh, D.H., Rees, E.G., Evans, A., Chambert, K.D., Moran, J.L., Goldstein, J., Neale, B.M., et al. (2016). Exome arrays capture polygenic rare variant contributions to schizophrenia. Hum. Mol. Genet. *25*, 1001–1007.

Ripke, S., Walters, J.T., and O'Donovan, M.C.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. MedRxiv, 2020.09.12.20192922.

Risch, N.J., Bressman, S.B., Senthil, G., and Ozelius, L.J. (2007). Intragenic Cis and Trans modification of genetic susceptibility in DYT1 torsion dystonia. Am. J. Hum. Genet. *80*, 1188–1193.

Rivas, M.A., Avila, B.E., Koskela, J., Huang, H., Stevens, C., Pirinen, M., Haritunians, T., Neale, B.M., Kurki, M., Ganna, A., et al.; International IBD Genetics Consortium; NIDDK IBD Genetics Consortium; T2D-GENES Consortium (2018). Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. PLoS Genet. *14*, e1007329.

Rubinstein, R., Thu, C.A., Goodman, K.M., Wolcott, H.N., Bahna, F., Mannepalli, S., Ahlsen, G., Chevee, M., Halim, A., Clausen, H., et al. (2015). Molecular logic of neuronal self-recognition through protocadherin domain interactions. Cell *163*, 629–642.

Rubinstein, R., Goodman, K.M., Maniatis, T., Shapiro, L., and Honig, B. (2017). Structural origins of clustered protocadherin-mediated neuronal barcoding. Semin. Cell Dev. Biol. *69*, 140–150.

Ruderfer, D.M., Charney, A.W., Readhead, B., Kidd, B.A., Kähler, A.K., Kenny, P.J., Keiser, M.J., Moran, J.L., Hultman, C.M., Scott, S.A., et al. (2016). Polygenic overlap between schizophrenia risk and antipsychotic response: a genomic medicine approach. Lancet Psychiatry *3*, 350–357.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.

Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al.; Autism Sequencing Consortium; iPSYCH-Broad Consortium (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568–584.e23.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J.T., and O'Donovan, M.C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. MedRxiv, 2020.09.12.20192922.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods *9*, 671–675.

Selvan, M.E., Zauderer, M.G., Rudin, C.M., Jones, S., Mukherjee, S., Offit, K., Onel, K., Rennert, G., Velculescu, V.E., Lipkin, S.M., et al. (2020). Inherited rare,

deleterious variants in ATM increase lung adenocarcinoma risk. MedRxiv, 2020.03.19.20034942.

Shao, Z., Noh, H., Bin Kim, W., Ni, P., Nguyen, C., Cote, S.E., Noyes, E., Zhao, J., Parsons, T., Park, J.M., et al. (2019). Dysregulated protocadherin-pathway activity as an intrinsic defect in induced pluripotent stem cell-derived cortical interneurons from subjects with schizophrenia. Nat. Neurosci. *22*, 229–242.

Shohat, S., Ben-David, E., and Shifman, S. (2017). Varying Intolerance of Gene Pathways to Mutational Classes Explain Genetic Convergence across Neuropsychiatric Disorders. Cell Rep. *18*, 2217–2227.

Singh, T., Kurki, M.I., Curtis, D., Purcell, S.M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., et al.; Swedish Schizophrenia Study; INTERVAL Study; DDD Study; UK10 K Consortium (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat. Neurosci. *19*, 571–577.

Singh, T., Walters, J.T.R., Johnstone, M., Curtis, D., Suvisaari, J., Torniainen, M., Rees, E., Iyegbe, C., Blackwood, D., McIntosh, A.M., et al.; INTERVAL Study; UK10K Consortium (2017). The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nat. Genet. *49*, 1167–1173.

Singh, T., Poterba, T., Curtis, D., Akil, H., Eissa, M.A., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., et al. (2020). Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. MedRxiv, 2020.09.18.20192815.

Smart, S.E., Kępińska, A.P., Murray, R.M., and MacCabe, J.H. (2021). Predictors of treatment resistant schizophrenia: a systematic review of prospective observational studies. Psychol. Med. *51*, 44–53.

Smeland, O.B., Frei, O., Dale, A.M., and Andreassen, O.A. (2020). The polygenic architecture of schizophrenia - rethinking pathogenesis and nosology. Nat. Rev. Neurol. *16*, 366–379.

Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch. Gen. Psychiatry *60*, 1187–1192.

Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. *13*, 537–551.

Taniguchi, S., Ninomiya, K., Kushima, I., Saito, T., Shimasaki, A., Sakusabe, T., Momozawa, Y., Kubo, M., Kamatani, Y., Ozaki, N., et al. (2020). Polygenic risk scores in schizophrenia with clinically significant copy number variants. Psychiatry Clin. Neurosci. *74*, 35–39.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. *45* (D1), D331–D338.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. *47* (D1), D330–D338.

The NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Program (2018). BRAVO variant browser (NHLBI). https://bravo.sph.umich.edu/freeze5/hg38/.

Thu, C.A., Chen, W.V., Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between α, β, and γ protocadherins. Cell *158*, 1045–1059.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–11.10.33.

van Rheenen, W., Peyrot, W.J., Schork, A.J., Lee, S.H., and Wray, N.R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. Nat. Rev. Genet. *20*, 567–581.

Wagnon, J.L., Briese, M., Sun, W., Mahaffey, C.L., Curk, T., Rot, G., Ule, J., and Frankel, W.N. (2012). CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. PLoS Genet. *8*, e1003067.

Walter, S., Atzmon, G., Demerath, E.W., Garcia, M.E., Kaplan, R.C., Kumari, M., Lunetta, K.L., Milaneschi, Y., Tanaka, T., Tranah, G.J., et al. (2011). A genome-wide association study of aging. Neurobiol. Aging *32*, 2109.e15–2109.e28.

Wang, S.R., Agarwala, V., Flannick, J., Chiang, C.W.K., Altshuler, D., and Hirschhorn, J.N.; GoT2D Consortium (2014). Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. Am. J. Hum. Genet. *94*, 710–720.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al.; PsychENCODE Consortium (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*, eaat8464.

Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q., et al. (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep. *6*, 1139–1152.

Wu, Q., and Jia, Z. (2021). Wiring the Brain by Clustered Protocadherin Neural Codes. Neurosci. Bull. *37*, 117–131.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
NEB® 5-alpha Competent *E. coli* (High Efficiency)	NEB	C2987H
Deposited data		
	This paper	European Genome-phenome Archive https://www.ebi.ac.uk/ega/ accession code EGAS00001000664
Experimental models: cell lines		
K562 cells (huma leukemia cell line)	ATCC	CCL243
Oligonucleotides		
Primes for cloning see Table S7	This paper	N/A
Recombinant DNA		
pmax_PCDHA3_mCherry	This paper	Addgene plasmid #164616
pmax_PCDHA3(N559S)_mCherry	This paper	Addgene plasmid #164617
Software and algorithms		
ImageJ	Schneider et al., 2012	https://imagej.nih.gov/ij/
Simulation of purifying selection in a founder population	This paper	https://github.com/rrkhan/RareVariantSimulation

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and data should be directed to, and will be fulfilled by, the Lead Contact, Todd Lencz, Ph.D. (tlencz@northwell.edu).

Materials availability

Plasmids generated in this study have been deposited to Addgene (#164616 and #164617).

Data and code availability

DNA sequence data for the AJ cohort are available to qualified investigators through the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession code EGAS00001000664.

Code generated for this study is available at https://github.com/rrkhan/RareVariantSimulation.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Sequenced samples were derived from subjects described previously from multiple case-control cohorts summarized in Table S6. All subjects were self-reported to be Ashkenazi Jewish, and were only selected for sequencing if ancestry was also verified as AJ by principal components analysis of previously collected SNP array data as described in our prior publications (Atzmon et al., 2010; Guha et al., 2012). Moreover, no outliers were observed in principal components analysis of all included subjects in comparison to reference population data (1000 Genomes Phase 3), with cases and controls overlapping within the AJ cluster (Figure 4). Following the variant calling and quality control steps described below (Method details), the final sample available for analysis consisted of 1249 subjects (560 female; 689 male).

Patients with schizophrenia were recruited from hospitalized inpatients at seven medical centers in Israel as described previously (Guha et al., 2013; Lencz et al., 2013). All diagnoses were assigned after direct interview using a structured clinical interview, a questionnaire with inclusion and exclusion criteria, and cross-references to medical records. The inclusion criteria specified that subjects had to be diagnosed with schizophrenia or schizoaffective disorder by the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). The exclusion criteria eliminated subjects diagnosed with at least one of the following disorders: psychotic disorder due to a general medical condition, substance-induced psychotic disorder, or any Cluster A (schizotypal, schizoid or paranoid) personality disorder. Controls were taken from several cohorts, primarily those screened for multiple forms of chronic illness (Lencz et al.,

2013; Walter et al., 2011), but also including a small number of subjects ascertained for non-psychiatric disorders (Inflammatory Bowel Disease or Dystonia)(Kenny et al., 2012; Risch et al., 2007). Informed consent was obtained from all subjects in accordance with institutional policies and the studies were approved by the corresponding institutional review boards.

Cell lines

For plasmid generation, the variable coding sequence of the *PCDHA3* isoform was amplified by PCR from human genomic DNA isolated from HEK293T cells. For isoforms containing the intracellular domain, the constant coding region was amplified from cDNA isolated from SKNSH cells. For the cell aggregation and immunostaining experiments, K562 cells were utilized.

METHOD DETAILS

Sequencing and variant calling pipeline

All samples were sequenced on the Illumina HiSeq X platform, using methods described previously (Lencz et al., 2018). Briefly, genomic DNA was isolated from whole blood and was quantified using PicoGreen, and integrity was assessed using the Fragment Analyzer (Advanced Analytical). Sequencing libraries were prepared using the Illumina TruSeq Nano DNA kit, with 100ng input, and pooled in equimolar amounts (8 samples/pool); a 2.5-3nM pooled library was loaded onto each lane of the patterned flow cell, and clustered on a cBot, generating ~375-400M pass filter 2x150bp reads per flow cell lane. For samples that did not meet 30x mean genome coverage post alignment, additional aliquots of the sequencing libraries were pooled in proportion to the amount of additional reads needed for re-sequencing.

Upon completion of sequencing runs, bcl files were demultiplexed and quality of sequencing data reviewed using SAV software (Illumina) and FastQC for deviations from expected values with respect to total number of reads, percent reads demultiplexed (> 95%), percent clusters pass filter (> 55%), base quality by lane and cycle, percent bases > Q30 for read 1 and read 2 (> 75%), GC content, and percent N-content. FastQ files were aligned to hg19/GRCh37 using the Burrows-Wheeler Aligner (BWA-MEM v0.78)(Li and Durbin, 2009) and processed using the best-practices pipeline that includes marking of duplicate reads by the use of Picard tools (v1.83, http://picard.sourceforge.net), realignment around indels, and base recalibration via Genome Analysis Toolkit (GATK v3.5)(Van der Auwera et al., 2013). A total of 1,310 samples proceeded to the last steps of joint genotyping and VQSR variant filtering after removal of 10 samples (7 cases, 3 controls) with < 80% 20X read depth coverage of the genome, and removal of 26 duplicate samples (1 case and 25 controls) sequenced as part of QC procedures. All remaining samples were jointly genotyped to generate a multi-sample VCF. Variant Quality Score Recalibration (VQSR) was performed on the multi-sample VCF, and variants were annotated using VCFtools (Danecek et al., 2011). After the GATK pipeline, we also filtered the SNP and small INDEL on LCR regions (Li, 2014) and 1000G masked difficult regions (Auton et al., 2015), since these regions are enriched with calling errors that cannot be filtered effectively by the VQSR model, as we demonstrated previously (Lencz et al., 2018). Furthermore, we masked the genotypes with GQ < 20 as "./." and filtered variants where ≤ 80% individuals could not be genotyped confidently. For purposes of downstream case-control analyses, we also removed 1 member of any pair of samples that were related at the first-cousin level or greater (n = 27 controls).

Defining ultra-rare variants (URVs) in TAGC samples

To maximize the power of existing reference population databases, we focused our primary analysis on the exome regions (as defined by gnomAD exome calling intervals (Karczewski et al., 2020)). We focused on URVs that were novel singletons in our TAGC cohort, filtering out all variants called in gnomAD (v2.1.1) non-neuro samples of any ethnicity, regardless of their call quality, and filtering out all variants called in TOPMed (The NHLBI Trans-Omics for Precision Medicine TOPMed Whole Genome Sequencing Program, 2018) freeze 5 release (with coordinates lifted over to hg19). We identified a small set of TAGC samples (n = 34; 21 cases / 13 controls) with an excessive number of exonic URVs (> = 50), shown as outliers in the distribution (Figure S3). Importantly, these outlier samples also demonstrated excess number of intergenic URVs and were not restricted to any sequencing batch. After filtering these outliers, we ended up with 1,249 samples for the downstream analyses (Table S6).

Primary analyses compared potentially functional (missense or loss of function) to putatively silent (synonymous or other) variants; these were defined as MisLoF and non-MisLoF, respectively. Exonic URVs were classified as loss of function, missense, synonymous, or other (generally intronic bases flanking exons as well as UTR regions), based on their most damaging impact annotated for any transcript. Thus, non-MisLoF variants had no missense or loss of function annotation on any known transcript.

Plasmid generation

The variable coding sequence of the *PCDHA3* isoform was amplified by PCR from human genomic DNA isolated from HEK293T cells. For isoforms including the intracellular domain, the constant coding region was amplified from cDNA isolated from SKNSH cells. HiFi DNA assembly was used to generate a mCherry tagged PCDHα3 protein by inserting the PCR fragments into the KpnI and AgeI restriction sites of a plasmid containing the immediate early promoter of cytomegalovirus with intron element (PCMV-IE) and mCherry sequence previously described (Thu et al., 2014). To generate the PCDHα3(N559S) mutant expression plasmid, a single nucleotide

substitution was made using Q5 site directed mutagenesis (NEB) following manufacturer's instructions. Plasmid sequences were confirmed by Sanger sequencing (Genewiz). Primers used to generate the wild-type PCDHα3 expression construct and perform the site directed mutagenesis are found in Table S7.

Cell aggregation assay

K562 cell aggregation assay was performed as previously described (Thu et al., 2014). Briefly, PCDHα proteins require the co-expression of PCDHβ or PCDHγ isoforms to reach the cell surface, wild-type and mutant PCDHA3 expression plasmids were nucleofected together with a PCDHγC3ΔEC1 carrier expression construct into K562 cells (ATCC CCCL243) using the Amaxa 4D- Nucleofector (Lonza) following the manufacturer's recommended protocol. After 24 hours in culture, the cells were allowed to aggregate for 3-4 hours on a rocker kept inside of the incubator. The cells were then fixed in 4% paraformaldehyde (PFA), washed with PBS, and imaged using an Olympus IX71 microscope.

Immunostaining

K562 cells were nucleofected as described above. Following 24 hours post-nucleofection, the cells were fixed with 4% paraformaldehyde and washed in PBS. Cells were then collected onto poly-D-lysine coated coverslips by centrifugation at 800rcf. for 10min. Cells were then mounted onto glass slides and the K562 cells expressing mCherry tagged PCDHα3 isoforms were imaged using an Olympus Fluoview FV1000 confocal microscope.

QUANTIFICATION AND STATISTICAL ANALYSIS

Common variant polygenic risk score (PRS)

A common variant PRS for schizophrenia was calculated for each subject based on summary statistics from the large-scale schizophrenia GWAS reported by the Psychiatric Genomics Consortium (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020), excluding our own Ashkenazi cohort. These summary statistics were filtered to include only high-quality SNPs (imputation INFO score $\geq$ 0.9). Additional filtering was also applied to the genotype data of the AJ samples to maximize quality of the variants used for PRS calculation; variants were only utilized if: 1) $\geq$ 99% of the AJ samples were called with high confident genotypes (GQ $\geq$ 20) at that position; 2) MAF $\geq$ 1% in the AJ samples; 3) HWE $p > 1x10^{-10}$.

The PRS calculation was performed using PRSice-2 (v2.3.3) after applying LD clumping (R^2 threshold = 0.1 within 250kb) on the AJ genotypes. In order to optimize the p value threshold (P_T) for the primary comparison of interest (PRS versus URV burden), we examined the effect of different values of P_T on the amount of PRS variance explained by case/control status. The largest R^2 for the case-control comparison was obtained at P_T = 0.00725, and this threshold was then used for the primary regression analyses examining the relationship between PRS and total number of MisLoF URVs, controlling for sex and the loadings of top five genetic principal components derived from genotypes on common SNP sites, performed separately for cases and controls.

Assigning novel URVs to genes and defining "case-only" and "control-only" genes

Each gene was characterized by the number of cases and the number of controls harboring a novel MisLoF or non-MisLoF variant within it. We focused on genes in which only cases or only controls harbored a MisLoF variant; genes in which both cases and controls were observed to have a given type of URV were excluded from subsequent analyses for that variant type. Of course, it was more likely that a gene would be identified as "case-only" rather than "control-only" for each type of URV due to the unequal numbers of cases relative to controls. Consequently, we controlled for the both effects by utilizing a re-sampling strategy, down-sampling the number of cases to match the number of controls, and iterated 10,000 times. For each iteration for each variant type, the following calculation was performed: $\frac{\#CASE\_only\ genes - \#CONTROL\_only\ genes}{\#CASE\_only\ genes}$

Replication of SCHEMA genes

The SCHEMA consortium lists 10 significant genes using strict exome-wide criteria ($p < 2.2x10^{-6}$) and 32 genes using false discovery rate < 0.05 ($p < 7.9x10^{-5}$) (Singh et al., 2020). We used the hypergeometric test to statistically compare the overlap between case-only MisLoF genes in our AJ sample and these SCHEMA genes (restricting the comparison to the autosome). Note that we did not simply merge datasets due to differences in sequence acquisition, calling, and quality control procedures. To guard against potential confound by gene size, permutations were performed to match the size distribution of our case-only gene set in the following manner: First, we created a ranked list of all autosomal genes in order of size of the coding sequence, and divided this list into ten bins (deciles). In permutation testing, a gene was randomly sampled from the same decile for each gene in the set to be matched. A total of 10,000 iterations were performed, and the empirical p value was defined by the proportion of permuted gene sets with overlap $\geq$ the case-only gene set. If no permutation was observed to meet this criteria, the p value was reported as $< 1x10^{-4}$.

Gene set analyses

We compared the AJ case-only and control-only MisLoF genes to three categories of gene sets based on prior studies: 1. *De novo* mutation genes implicated across multiple developmental brain disorders (DBD) (Gonzalez-Mantilla et al., 2016), a large scale autism

spectrum disorder(ASD) exome study (Satterstrom et al., 2020), and the integration of the ASD set with a large-scale study of ASD, developmental disorder(DD), and intellectual disability(ID) exome sequencing studies (Coe et al., 2019); 2. Genes known to encode proteins of the synapse aggregated in SynaptomeDB (Pirooznia et al., 2012), and genes regulated by known neuronal RNA-binding proteins, including CELF4 (Wagnon et al., 2012), FMRP (Darnell et al., 2011), RBFOX2 and RBFOX1/RBFOX3 (Weyn-Vanhentenryck et al., 2014); 3. Genes constrained by missense (Samocha et al., 2014) and LoF variants (Karczewski et al., 2020). Since X and Y chromosomes of AJ samples are not included, we adjusted the number of genes in each set and total protein-coding genes (n = 19,780; http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database) to autosomal-only to calculate the p values using hypergeometric tests. We further compared the relative enrichment of AJ case-only versus control-only MisLoF genes for these gene sets using the chi-square test based on the 2x2 table of overlap and non-overlap for case-only and control-only. As above, we also performed permutation testing to control for effects of gene size. In addition to the *a priori* gene sets identified above, we performed hypothesis-free testing of the case-only and control-only lists against all GO categories and PANTHER protein classes available from the Gene Ontology Consortium (http://geneontology.org, GO Ontology database https://doi.org/10.5281/zenodo.4081749)(Ashburner et al., 2000; The Gene Ontology Consortium, 2017, 2019; Mi et al., 2019), as well as synaptic components (annotated by SYNGO; Koopmans et al., 2019). To control for gene size, we extended the permutation procedures described above to the significant gene sets identified from the hypothesis-free GO analysis. Moreover, we ran the full set of GO and PANTHER analyses on 100 permuted gene lists, size-matched to the case-only list.

Modeling the effects of purifying selection

We simulated the spread of a single schizophrenia-causing variant in a founder population under a number of empirical conditions. The initial conditions of the simulation assume a single variant carrier in a fixed size population at the time of bottleneck, which we denote as N_B. From there, we modeled the growth of the number of carriers and total population size over a set number of G generations until a maximum population size N_{max} is achieved. The growth rate R is computed as $R = ((N_{max}/N_B)/G)$. We generated the number of offspring per individual within a generation using a Poisson branching process, in which the lambda parameter is a function of growth rate R, sex, and case/control status. A healthy individual will have a lambda equal to R, while an individual with schizophrenia will have reduced relative fecundity, differing by sex (estimated at 0.5 for females and 0.25 for males (Power et al., 2013)). Within our simulated population, we track the number of variant carriers for each of several scenarios, in which we test the sensitivity of the model to variations on our initial estimates of population bottleneck size N_B, number of generations G, and relative fecundity of schizophrenia cases; we report results as a function of the disease penetrance of the variant. We computed 10,000 simulations for each scenario, calculating the proportion of simulations in which the number of variant carriers goes to zero (extinction) within G generations, and the proportion of simulations in which a variant escapes extinction. We then use the number of variant carriers remaining after G generations in the population to calculate the power of Fisher's exact test for detecting the variant effect in a study cohort size of a given size.