Novel Ultra-Rare Exonic Variants Identified in a Founder Population Implicate Cadherins in Schizophrenia

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

1 Departments of Psychiatry and Molecular Medicine, Hofstra Northwell School of Medicine, Hempstead, New York 11550

2 Department of Psychiatry, Division of Research, The Zucker Hillside Hospital Division of Northwell Health, Glen Oaks, NY, 11004

3 Institute for Behavioral Science, The Feinstein Institutes for Medical Research, Manhasset, NY, 11030

4 Department of Computer Science, Columbia University, 500 W 120th St, New York, NY, 10027

5 Braun School of Public Health, Faculty of Medicine, Hebrew University of Jerusalem, Ein Kerem, Jerusalem, 9112102, Israel

6 Department of Genetics, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461

7 Department of Medicine, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461

8 Department of Neurology, Beth Israel Medical Center, New York, New York, USA. 10003

9 Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Givat Ram, Jerusalem, Israel, 91904

10 Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl, New York, NY 10029

11 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl, New York, NY, 10029

12 Department of Pathology and Cell Biology, Columbia University Medical Center, 1150 St Nicholas Ave, New York, NY, 10032

13 Taub Institute for Research of Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, 1150 St Nicholas Ave, New York, NY, 10032

14 Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl, New York, NY 10029

15 Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY, 10065

16 Departments of Medicine, Genetic Medicine and Surgery, Weill Cornell Medical College, New York, NY

17 Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY, 10065

18 Department of Pathology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461

19 Department of Neurology, Massachusetts General Hospital, 16th St, Charlestown, MA 02129

20 Department of Human Biology, Haifa University, Haifa, Israel

21 Center for Computational Biology and Bioinformatics, Columbia University, 1130 St Nicholas Ave, New York, NY, 10032

*** These two authors contributed equally**

§ Deceased

† Corresponding authors: tlencz@northwell.edu; itsik@cs.columbia.edu

Abstract

IMPORTANCE: Schizophrenia is a serious mental illness with high heritability. While common genetic variants account for a portion of the heritability, identification of rare variants associated with the disorder has proven challenging.

OBJECTIVE: To identify genes and gene sets associated with schizophrenia in a founder population (Ashkenazi Jewish), and to determine the relative power of this population for rare variant discovery.

DESIGN, SETTING, AND PARTICIPANTS: Data on exonic variants were extracted from whole genome sequences drawn from 786 patients with schizophrenia and 463 healthy control subjects, all drawn from the Ashkenazi Jewish population. Variants observed in two large publicly available datasets (total n≈153,000, excluding neuropsychiatric patients) were filtered out, and novel ultra-rare variants (URVs) were compared in cases and controls.

MAIN OUTCOMES AND MEASURES: The number of novel URVs and genes carrying them were compared across cases and controls. Genes in which only cases or only controls carried novel, functional URVs were examined using gene set analyses.

RESULTS: Cases had a higher frequency of novel missense or loss of function (MisLoF) variants compared to controls, as well as a greater number of genes impacted by MisLoF variants. Characterizing 141 "case-only" genes (in which ≥ 3 AJ cases in our dataset had MisLoF URVs with none found in our AJ controls), we replicated prior findings of both enrichment for synaptic gene sets, as well as specific genes such as *SETD1A* and *TRIO*. Additionally, we identified cadherins as a novel gene set associated with schizophrenia including a recurrent mutation in *PCDHA3*. Several genes associated with autism and other neurodevelopmental disorders including *CACNA1E*, *ASXL3*, *SETBP1*, and *WDFY3*, were also identified in our case-only gene list, as was *TSC2*, which is linked to tuberous sclerosis. Modeling the effects of purifying selection demonstrated that deleterious rare variants are greatly over-represented in a founder population with a tight bottleneck and rapidly expanding census, resulting in enhanced power for rare variant association studies.

CONCLUSIONS AND RELEVANCE: Identification of cell adhesion genes in the cadherin/protocadherin family is consistent with evidence from large-scale GWAS in schizophrenia, helps specify the synaptic abnormalities that may be central to the disorder, and suggests novel potential treatment strategies (e.g., inhibition of protein kinase C). Study of founder populations may serve as a cost-effective way to rapidly increase gene discovery in schizophrenia and other complex disorders.

Twin studies and other family-based designs have long demonstrated that schizophrenia (SCZ) is highly heritable (h^2≈.6-.85)[1–3]. While large-scale genome-wide association studies (GWAS) have discovered increasing numbers of common (minor allele frequency > 1%) variants associated with illness[4–6], the cumulative effect of such variants accounts for only about a third of the total heritability of SCZ[7,8]. It is therefore likely that rare genetic variants contribute substantially to the heritability of SCZ[9,10], and such rare variants might have considerably higher effect sizes (odds ratios) relative to common variants[11]. For example, several rare (frequency<<1% in the general population) copy number variants have been reliably associated with SCZ, with odds ratios ranging from 5-20 or higher[12].

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)[13]; 2) at any given gene, a variety of different alleles may have deleterious effects (allelic heterogeneity)[14]; 3) deleterious rare variants are generally driven to extremely low frequencies due to purifying selection[15]; and 4) the background rate of benign rare variation across the population is very high[16]. To date, only very large international consortia efforts have identified any schizophrenia-associated SNV's. The largest such effort, the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium with 25,000 cases and nearly 100,000 controls, identified only 10 exome-wide significant genes.[17]

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[18,19]. For example, the Ashkenazi Jewish (AJ) population, currently numbering more than 10 million individuals worldwide, derives effectively from a mere ~300 founders approximately 750 years ago[20,21]. While the AJ population is well known to be enriched for deleterious variants leading to rare recessive disorders[22], AJ also demonstrate 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[23] and the *BRCA1* c.66_67AG allele associated with breast cancer[24]. 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[25].

In the present study, we examined rates of protein-altering rare variants in AJ cases with schizophrenia compared with AJ controls. Based on prior research[9,26–28], we hypothesized that cases would be enriched for rare deleterious variants, especially in genes expressed at the neuronal synapse. We sought to extend these results to additional categories of genes that might be detectable due to the greater frequency of rare variants observed in the AJ population. Additionally, we attempted to replicate the schizophrenia risk genes identified by the SCHEMA consortium. Finally, we modelled the process of purifying selection in a rapidly expanding, bottlenecked population, in order to quantify the relative power of AJ for rare variant discovery.

Methods

Subjects

Sequenced samples (n=1346) were derived from subjects described previously from multiple case-control cohorts summarized in Supplementary Table 1. All samples were self-reported to be Ashkenazi Jewish, and also verified as AJ by principal components analysis of previously collected SNP array data as described in our prior publications[29,30]. Informed consent was obtained in accordance with institutional policies and the studies were approved by the corresponding institutional review boards.

Patients with schizophrenia were recruited from hospitalized inpatients at seven medical centers in Israel as described previously[31,32]. 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[32,33], but also including a small number of subjects ascertained for non-psychiatric disorders (Inflammatory Bowel Disease or Dystonia)[34,35].

Sequencing and variant calling pipeline

All samples were sequenced on the Illumina HiSeqX platform, using methods described previously[36]. 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)[37] 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)[38]. 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[39]. After the GATK pipeline, we also filtered the

SNP and small INDEL on LCR regions[40] and 1000G masked difficult regions[41], since these regions are enriched with calling errors that cannot be filtered effectively by the VQSR model, as we demonstrated previously[36]. 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[42]). 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[43] freeze 5 release (with coordinates lifted over to hg19). We identified a small set of TAGC samples (n = 34; 21 cases / 13 controls) with excessive number of exonic URVs (>=50), shown as outliers in the distribution (Supplementary Figure 1). 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 down-stream analyses (Supplementary Table 1).

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 immediately flanking exons), 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.

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. For each category of URV (MisLoF and non-MisLoF), we focused on genes in which only cases or only controls harbored a 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: (Case only genes – Control only genes))/(Case only genes + Control 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 <.05 ($p<7.9x10^{-5}$).[44] 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). To guard against potential confound by gene size, permutations were performed to match the size distribution of our case-only gene set. A permuted gene was randomly sampled from a window of ±25 in the order of coding sequence size for each gene in the

set to be matched. A total of 10000 iterations were performed, and the empirical p-value was defined by the proportion of the permuted gene set with overlap >= 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)[45], a large scale autism spectrum disorder(ASD) exome study[46], and the integration of the ASD set with a large-scale study of ASD, developmental disorder(DD), and intellectual disability(ID) exome sequencing studies[47]; 2. Genes known to encode proteins of the synapse aggregated in SynaptomeDB[48], and genes regulated by known neuronal RNA-binding proteins, include CELF4[49], FMRP[50], Rbfox1/2/3[51]; 3. Genes constrained by missense[52] and LoF variants with pLi>0.9[42]. Since X and Y chromosome genes of AJ samples are not included, we adjusted the number of genes in each gene set and the total number of 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 between AJ case-only and control-only MisLoF genes in the prior gene sets using the chi-squared 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.

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 $R = exp(ln(N_{max}/N_B)/G)$. 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 a schizophrenic individual will have reduced relative fecundity based on their sex. Within our simulated population, we track the number of variant carriers while varying the population bottleneck size N_B, number of generations G, relative fecundity of schizophrenic subjects, and 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.

Results

Greater rates of damaging variants in cases

After QC procedures, a total of 786 SCZ cases and 463 controls were available for final analysis. Groups did not significantly differ in total number of variants called in their whole genome (~3.68M) or exome (~49K) (Table 1). After filtering on all variants observed in the TOPMED and gnomAD (v2.1.1, non-neuro)

datasets, cases and controls did not significantly differ on total number of novel variants observed genome-wide (~5K). However, cases had significantly more novel exome-wide variants, exclusively limited to singletons (17.76±6.24 vs 15.44±6.42, $p=6.13x10^{-10}$; Table 1).

Against this backdrop, cases and controls were compared on the number of functional vs silent variants within the exome, in two ways: variant-based tests and gene-based tests. First, 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 MisLoF and non-MisLoF URVs at the gene level. 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).

Characterizing case-only genes

Given that cases demonstrated significant elevation in genes carrying MisLoF URVs, we next sought to characterize the case-only MisLoF genes. As shown in Supplementary Table 2, eight genes had Case ≥ 5 and Control = 0. Notably, one of these was *SETD1A*, a methyltransferase gene that was the first to reach genome-wide significance in a schizophrenia rare variant study[53]. We next tested the set of 9 autosomal exome-wide-significant schizophrenia genes identified by the SCHEMA consortium[44] 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). Three of 9 autosomal exome-wide significant SCHEMA genes (*SETD1A*, *TRIO*, and *XPO7*) were among the 141 case-only genes, a 47-fold over-representation relative to chance (hypergeometric test $p = 3.03x10^{-7}$). Results remained significant when permutation tests controlling for gene size were performed (empirical $p=5.8x10^{-3}$). Similar results were obtained examining overlap of our 141 case-only genes with the set of 29 autosomal SCHEMA genes that met the criteria of FDR<.05; in addition to the three genes above, *STAG1* was also shared between our case-only list and SCHEMA (4/29 genes; hypergeometric $p=1.77x10^{-6}$; empirical $p=4.6x10^{-3}$ using permutations controlling for gene size).

More broadly, we utilized gene set analyses to characterize our 141 case-only genes, as compared to a similarly-sized set of 148 "control-only" genes in which ≥ 2 AJ controls had MisLoF URVs (with none found in our AJ cases). First, we tested sets of genes selected *a priori* based on prior literature; specifically, previous case-control exome studies in schizophrenia have identified: 1) overlaps with other developmental brain disorders (DBD) 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 missense and/or loss of function variants than average, presumably due to purifying selection)[9,26–28,45]. As shown in Table 2, each of these gene sets demonstrated significant (by hypergeometric test) overlap with the case-only, but not the control-only, gene lists; moreover, the difference in enrichment between the case-only and control-only gene sets was statistically significant in all cases. Permutation tests accounting for gene size demonstrated similar results, with the exception that the overlap with ASD/ID gene sets was no longer significant (Supplementary Table 3).

Next, we examined enrichment of our case-only gene list across all GO categories, Panther protein classes, and synaptic components (annotated by SYNGO[54]). As shown in Table 3, novel categories of enrichment for the 141 schizophrenia-associated genes were observed for biological processes related to cell adhesion, and specifically to the cadherin class of proteins. GO cellular components analysis revealed the expected enrichment for synaptic and related compartments such as neuron projection and vesicle (Supplementary Table 4), while SYNGO analysis demonstrated that enrichment extended across both presynaptic and postsynaptic genes (Supplementary Table 5). By contrast, the 148 control-only genes showed no synaptic enrichment (all q-value>.75), no enriched GO categories (all FDR>.05), and only one enriched Panther protein class, which was in a very small gene set (Hsp90 family chaperone, overlap of 3/9 genes; $p=3.59x10^{-5}$, $FDR=7.00x10^{-3}$).

Identifying recurring, damaging URVs

The foregoing analyses examined singleton URVs only, which have been the primary focus of exome studies in schizophrenia to date; indeed, the SCHEMA dataset of case-only variants contains ~95% singletons, and <1% of all case-only variants in SCHEMA are observed 3 or more 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 schizophrenia cases (n=869) and controls (n=2415) from SCHEMA, in order to identify individual URVs that were observed ≥3 times in cases (i.e., at least ~1/1000 allele frequency in the 3,310 AJ case chromosomes available). For further filtering, we included exome data from an additional 1,587 AJ controls from a separate study of longevity, resulting in 8930 AJ control chromosomes available). As shown in Supplementary Table 6, 17 MisLoF URVs were observed in ≥3 cases and zero controls (nominal p<.05 by Fisher's exact test). Notably, the most common variant (observed 5 times, case frequency = .15%) is a putatively damaging (CADD score = 23.4) missense variant in *PCDHA3*, part of the protocadherin cluster on chromosome 5.

Modeling the rise of damaging rare variants 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[20,21], and population-based estimates[55] of reduced fecundity in schizophrenia (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 (Supplementary Table 7) in order 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 schizophrenia. As shown in Figure 2, between 30-50% of such variants escape extinction within the range of penetrance expected, given the genetic architecture of the disorder[11]. Based on these results, and given rough estimates of the AJ population today (~10M) and the prevalence of schizophrenia (~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 the present study to detect a given variant at exome-wide significance (Supplementary Figure

2), which generally ranged between 5% and 20%. However, we also calculated that a slightly larger study, of 5000 cases and 9000 controls, would have power of ~20-40% to detect any individual variant in the range of realistic penetrance (Supplementary Figure 3), and would therefore have >80% power to detect *at least* one 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 1000 genes or more[9,27], the significant findings documented in Table 2, and the long list of variants at greater than doubleton frequency documented in Supplementary Table 6.

Discussion

The present study demonstrates the enhanced power available to genetic studies performed in populations enriched for rare variants, consistent with recent work in schizophrenia[26] and other phenotypes[18,25,56]. We further reduced background heterogeneity by utilizing a strict filter against all variants reported in non-neuropsychiatric samples across the two largest publicly available sequencing datasets, gnomAD[42] and TOPMED[43]. Thus, despite relatively modest sample sizes, the present study was able to replicate several previously identified schizophrenia-associated genes (*SETD1A, TRIO, XPO7*)[44,53] and gene sets (synaptic, DBD-related, and constrained genes)[9,26–28,45], with these analyses serving as a positive control for our approach. Beyond these replications, we were also able to make several novel discoveries, as described below.

First, we identified several novel gene sets associated with schizophrenia. The strongest statistical signal was observed for cell adhesion processes, including 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[57]. Surprisingly, cadherins have not received much attention in the schizophrenia genetics literature, despite the considerable recent focus on both calcium activity and synaptic proteins[58]. While there are more than 100 different proteins in the cadherin superfamily[57], it is noteworthy that three of the four 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[59,60]. Homozygous mutations in *FAT4* cause Van Maldergem syndrome, a recessive intellectual disability marked by periventricular neuronal heterotopia[61], while mutations in *FAT1* have been observed in autism[62].

Relatedly, a single missense variant in a protocadherin gene (*PCDHA3*) was observed at higher rate of recurrence (5 observations) in cases than any other ultra-rare (i.e., not in healthy individuals) variant in the published schizophrenia literature (although it should be noted that one splice acceptor variant in *SETD1A* appears six times in the SCHEMA database). *PCDHA3* is one of several protocadherins, clustered at a single locus on chromosome 5, which serve as a "molecular barcode" on the neuronal cell surface, guiding neurites away from forming synapses with other neurites from the same cell[63]. Altered expression of protocadherins (including *PCDHA3)* in schizophrenia has been implicated by a recent transcriptome-wide association study of both prefrontal cortex and hippocampus[64], and cortical interneurons derived from induced pluripotent stem cells (iPSCs) of patients with schizophrenia showed

reduced *PCDHA3* expression compared to similarly derived interneurons from controls[65]. 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 glutamatergic neurons), and that these deficits could be reversed by treatment with an inhibitor of protein kinase C[65].

When our samples were combined with Ashkenazi patients from the SCHEMA database, the *PCDHA3* missense variant was observed in 0.3% of all Ashkenazi cases in the present study. While unusually high for a schizophrenia-associated URV, this carrier rate is low compared to the ~4% rate observed for the most common *BRCA1* founder variant in Ashkenazi breast cancer cases[66], and the ~15% rate of the *LRRK2* G2019S variant amongst Ashkenazi patients with Parkinson's disease[67]. 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 schizophrenia[5], a disorder which results in markedly reduced fecundity[55]. Nevertheless, our simulations demonstrated the limits of purifying selection in a founder population with a tight bottleneck. One-third to one-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[19]. Consequently, ascertainment of additional samples from founder populations can be a highly cost-effective way of rapidly enhancing power of rare variant studies[18].

The overlap of our schizophrenia case-only gene list with gene sets derived from autism and other developmental brain disorders was notable, insofar as exome studies in these disorders have been more well-powered than schizophrenia studies to date[68]. Consequently, the overlapping genes indicated in the first three rows of Table 2 have a strong prior probability of association, especially given prior evidence that rare single nucleotide variants (e.g., *SETD1A*)[53] and copy number variants[69] tend to be shared across schizophrenia and other neurodevelopmental disorders. In the present study, case-only variants in these overlapping genes (*ASXL3*, *BIRC6*, *CACNA1E*, *DIP2A*, *DST*, *LAMA2*, *NSD1*, *PCDH15*, *SETBP1*, *SETD1A*, *STARD9*, *TRIO*, *WDFY3*) were overwhelmingly (10:1 ratio) missense rather than loss of function; by contrast, many of the documented ASD/ID/DD variants in these same genes tend to be loss of function. Thus, it is possible that our findings represent allelic series at these genes, in which more damaging variants are associated earlier-onset, more severe clinical phenotypes[70]. Similarly, we identified 5 cases (and no controls) with novel missense variants in *TSC2*, a gene in which mutations (primarily loss of function) are known to cause tuberous sclerosis (TS). TS is an autosomal dominant disorder marked by hamartomas across multiple organs, potentially including the brain[71]. Case reports of psychotic features in TS patients have proliferated for decades[72]; a recent survey of a large international cohort of TS patients identified psychosis in 11% of adults[73]. Since the affected cases in the present study were not noted in their medical report to have TS, our results suggest that schizophrenia can be the primary presenting feature of *TSC2* mutations.

Conclusions

Despite a relatively small sample size for a genomic study, we demonstrate that a founder population can have enhanced power for detecting rare variants associated with schizophrenia. As a positive control, our analysis of genes marked by case-only novel functional URVs replicated several known

findings, including enrichment of gene sets involved with synaptogenesis and other neurodevelopmental disorders. Additionally, we identified several novel associated gene sets, particularly those related to cellular adhesion functions of the cadherin gene family. Given the multiple sources of heterogeneity that complicate genetic sequencing studies, examination of founder populations may be a cost-effective approach to identify novel disease genes and pathways.

References

1. McGue M, Gottesman II, Rao DC. The transmission of schizophrenia under a multifactorial threshold model. *Am J Hum Genet*. 1983;35(6):1161-1178.

2. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry*. 2003;60(12):1187-1192. doi:10.1001/archpsyc.60.12.1187

3. Hilker R, Helenius D, Fagerlund B, et al. Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry*. 2018;83(6):492-498. doi:10.1016/j.biopsych.2017.08.017

4. Biological insights from 108 schizophrenia-associated genetic loci. - PubMed - NCBI. Accessed March 3, 2020. https://www.ncbi.nlm.nih.gov/pubmed/25056061

5. Pardiñas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet*. 2018;50(3):381-389. doi:10.1038/s41588-018-0059-2

6. Lam M, Chen C-Y, Li Z, et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet*. 2019;51(12):1670-1678. doi:10.1038/s41588-019-0512-x

7. Lee SH, DeCandia TR, Ripke S, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet*. 2012;44(3):247-250. doi:10.1038/ng.1108

8. Loh P-R, Bhatia G, Gusev A, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet*. 2015;47(12):1385-1392. doi:10.1038/ng.3431

9. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*. 2014;506(7487):185-190. doi:10.1038/nature12975

10. Ganna A, Satterstrom FK, Zekavat SM, et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am J Hum Genet*. 2018;102(6):1204-1211. doi:10.1016/j.ajhg.2018.05.002

11. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet*. 2012;13(8):537-551. doi:10.1038/nrg3240

12. Marshall CR, Howrigan DP, Merico D, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet*. 2017;49(1):27-35. doi:10.1038/ng.3725

13. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci*. 2014;17(6):782-790. doi:10.1038/nn.3708

14. Li B, Leal SM. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet*. 2009;5(5):e1000481. doi:10.1371/journal.pgen.1000481

15. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*. 2007;80(4):727-739. doi:10.1086/513473

16. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-69. doi:10.1126/science.1219240

17. KaiserOct. 25 J, 2019, Am 10:00. Intensive DNA search yields 10 genes tied directly to schizophrenia. Science | AAAS. Published October 25, 2019. Accessed April 17, 2020. https://www.sciencemag.org/news/2019/10/intensive-dna-search-yields-10-genes-tied-directly-schizophrenia

18. Locke AE, Steinberg KM, Chiang CWK, et al. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*. 2019;572(7769):323-328. doi:10.1038/s41586-019-1457-z

19. Wang SR, Agarwala V, Flannick J, et al. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. *Am J Hum Genet*. 2014;94(5):710-720. doi:10.1016/j.ajhg.2014.03.019

20. Palamara PF, Lencz T, Darvasi A, Pe'er I. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*. 2012;91(5):809-822. doi:10.1016/j.ajhg.2012.08.030

21. Carmi S, Hui KY, Kochav E, et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun*. 2014;5:4835. doi:10.1038/ncomms5835

22. Baskovich B, Hiraki S, Upadhyay K, et al. Expanded genetic screening panel for the Ashkenazi Jewish population. *Genet Med*. 2016;18(5):522-528. doi:10.1038/gim.2015.123

23. Ozelius LJ, Senthil G, Saunders-Pullman R, et al. LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews. *N Engl J Med*. 2006;354(4):424-425. doi:10.1056/NEJMc055509

24. Friedman LS, Szabo CI, Ostermeyer EA, et al. Novel inherited mutations and variable expressivity of BRCA1 alleles, including the founder mutation 185delAG in Ashkenazi Jewish families. *Am J Hum Genet*. 1995;57(6):1284-1297.

25. Rivas MA, Avila BE, Koskela J, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet*. 2018;14(5):e1007329. doi:10.1371/journal.pgen.1007329

26. Gulsuner S, Stein DJ, Susser ES, et al. Genetics of schizophrenia in the South African Xhosa. *Science*. 2020;367(6477):569-573. doi:10.1126/science.aay8833

27. Nguyen HT, Bryois J, Kim A, et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med*. 2017;9(1):114. doi:10.1186/s13073-017-0497-y

28. Genovese G, Fromer M, Stahl EA, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci*. 2016;19(11):1433-1441. doi:10.1038/nn.4402

29. Atzmon G, Hao L, Pe'er I, et al. Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am J Hum Genet*. 2010;86(6):850-859. doi:10.1016/j.ajhg.2010.04.015

30. Guha S, Rosenfeld JA, Malhotra AK, et al. Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. *Genome Biol*. 2012;13(1):R2. doi:10.1186/gb-2012-13-1-r2

31. Guha S, Rees E, Darvasi A, et al. Implication of a rare deletion at distal 16p11.2 in schizophrenia. *JAMA Psychiatry*. 2013;70(3):253-260. doi:10.1001/2013.jamapsychiatry.71

32. Lencz T, Guha S, Liu C, et al. Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat Commun*. 2013;4:2739. doi:10.1038/ncomms3739

33. Walter S, Atzmon G, Demerath EW, et al. A genome-wide association study of aging. *Neurobiol Aging*. 2011;32(11):2109.e15-28. doi:10.1016/j.neurobiolaging.2011.05.026

34. Kenny EE, Pe'er I, Karban A, et al. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet*. 2012;8(3):e1002559. doi:10.1371/journal.pgen.1002559

35. Risch NJ, Bressman SB, Senthil G, Ozelius LJ. Intragenic Cis and Trans modification of genetic susceptibility in DYT1 torsion dystonia. *Am J Hum Genet*. 2007;80(6):1188-1193. doi:10.1086/518427

36. Lencz T, Yu J, Palmer C, et al. High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. *Hum Genet*. 2018;137(4):343-355. doi:10.1007/s00439-018-1886-z

37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

38. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

39. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

40. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 2014;30(20):2843-2851. doi:10.1093/bioinformatics/btu356

41. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

42. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

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

44. SCHEMA Consortium. Exome meta-analysis results. Accessed May 11, 2020. https://schema.broadinstitute.org/results

45. Gonzalez-Mantilla AJ, Moreno-De-Luca A, Ledbetter DH, Martin CL. A Cross-Disorder Method to Identify Novel Candidate Genes for Developmental Brain Disorders. *JAMA Psychiatry*. 2016;73(3):275-283. doi:10.1001/jamapsychiatry.2015.2692

46. Satterstrom FK, Kosmicki JA, Wang J, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*. 2020;180(3):568-584.e23. doi:10.1016/j.cell.2019.12.036

47. Coe BP, Stessman HAF, Sulovari A, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature Genetics*. 2019;51(1):106-116. doi:10.1038/s41588-018-0288-4

48. Pirooznia M, Wang T, Avramopoulos D, et al. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics*. 2012;28(6):897-899. doi:10.1093/bioinformatics/bts040

49. Wagnon JL, Briese M, Sun W, et al. CELF4 Regulates Translation and Local Abundance of a Vast Set of mRNAs, Including Genes Associated with Regulation of Synaptic Function. *PLOS Genetics*. 2012;8(11):e1003067. doi:10.1371/journal.pgen.1003067

50. Darnell JC, Van Driesche SJ, Zhang C, et al. FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell*. 2011;146(2):247-261. doi:10.1016/j.cell.2011.06.013

51. Weyn-Vanhentenryck SM, Mele A, Yan Q, et al. HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports*. 2014;6(6):1139-1152. doi:10.1016/j.celrep.2014.02.005

52. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*. 2014;46(9):944-950. doi:10.1038/ng.3050

53. Singh T, Kurki MI, Curtis D, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci*. 2016;19(4):571-577. doi:10.1038/nn.4267

54. Koopmans F, van Nierop P, Andres-Alonso M, et al. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron*. 2019;103(2):217-234.e4. doi:10.1016/j.neuron.2019.05.002

55. Power RA, Kyaga S, Uher R, et al. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry*. 2013;70(1):22-30. doi:10.1001/jamapsychiatry.2013.268

56. Selvan ME, Zauderer MG, Rudin CM, et al. Inherited rare, deleterious variants in ATM increase lung adenocarcinoma risk. *medRxiv*. Published online March 23, 2020:2020.03.19.20034942. doi:10.1101/2020.03.19.20034942

57. Friedman LG, Benson DL, Huntley GW. Cadherin-based transsynaptic networks in establishing and modifying neural connectivity. *Curr Top Dev Biol*. 2015;112:415-465. doi:10.1016/bs.ctdb.2014.11.025

58. Nanou E, Catterall WA. Calcium Channels, Synaptic Plasticity, and Neuropsychiatric Disease. *Neuron*. 2018;98(3):466-481. doi:10.1016/j.neuron.2018.03.017

59. Fulford AD, McNeill H. Fat/Dachsous family cadherins in cell and tissue organisation. *Current Opinion in Cell Biology*. 2020;62:96-103. doi:10.1016/j.ceb.2019.10.006

60. Avilés EC, Goodrich LV. Configuring a robust nervous system with Fat cadherins. *Seminars in Cell & Developmental Biology*. 2017;69:91-101. doi:10.1016/j.semcdb.2017.06.001

61. Cappello S, Gray MJ, Badouel C, et al. Mutations in genes encoding the cadherin receptor-ligand pair DCHS1 and FAT4 disrupt cerebral cortical development. *Nat Genet*. 2013;45(11):1300-1308. doi:10.1038/ng.2765

62. Cukier HN, Dueker ND, Slifer SH, et al. Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Molecular Autism*. 2014;5(1):1. doi:10.1186/2040-2392-5-1

63. Canzio D, Maniatis T. The generation of a protocadherin cell-surface recognition code for neural circuit assembly. *Current Opinion in Neurobiology*. 2019;59:213-220. doi:10.1016/j.conb.2019.10.001

64. Collado-Torres L, Burke EE, Peterson A, et al. Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron*. 2019;103(2):203-216.e8. doi:10.1016/j.neuron.2019.05.013

65. Shao Z, Noh H, Bin Kim W, et al. Dysregulated protocadherin-pathway activity as an intrinsic defect in induced pluripotent stem cell-derived cortical interneurons from subjects with schizophrenia. *Nat Neurosci*. 2019;22(2):229-242. doi:10.1038/s41593-018-0313-z

66. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2 | Science. Accessed May 14, 2020. https://science.sciencemag.org/content/302/5645/643.full

67. Correia Guedes L, Ferreira JJ, Rosa MM, Coelho M, Bonifati V, Sampaio C. Worldwide frequency of G2019S LRRK2 mutation in Parkinson's disease: a systematic review. *Parkinsonism Relat Disord*. 2010;16(4):237-242. doi:10.1016/j.parkreldis.2009.11.004

68. Myers SM, Challman TD, Bernier R, et al. Insufficient Evidence for "Autism-Specific" Genes. *Am J Hum Genet*. 2020;106(5):587-595. doi:10.1016/j.ajhg.2020.04.004

69. Kirov G. CNVs in neuropsychiatric disorders. *Hum Mol Genet*. 2015;24(R1):R45-49. doi:10.1093/hmg/ddv253

70. Shohat S, Ben-David E, Shifman S. Varying Intolerance of Gene Pathways to Mutational Classes Explain Genetic Convergence across Neuropsychiatric Disorders. *Cell Reports*. 2017;18(9):2217-2227. doi:10.1016/j.celrep.2017.02.007

71. Henske EP, Jóźwiak S, Kingswood JC, Sampson JR, Thiele EA. Tuberous sclerosis complex. *Nat Rev Dis Primers*. 2016;2:16035. doi:10.1038/nrdp.2016.35

72. Herkert EE, Wald A, Romero O. Tuberous sclerosis and schizophrenia. *Dis Nerv Syst*. 1972;33(7):439-445.

73. de Vries PJ, Belousova E, Benedik MP, et al. TSC-associated neuropsychiatric disorders (TAND): findings from the TOSCA natural history study. *Orphanet Journal of Rare Diseases*. 2018;13(1):157. doi:10.1186/s13023-018-0901-8

.

Acknowledgements

The authors are extremely grateful to Soren Germer, Ph.D. and his team at the New York Genome Center for performing the Illumina sequencing. We acknowledge financial support from the Human Frontier Science Program (SC); NIH research grants AG042188 (GA), DK62429, DK062422, DK092235 (JHC), NS050487, NS060113 (LNC), AG021654, AG027734 (NB), MH089964, MH095458, MH084098 (TL), and CA121852 (computational infrastructure, IPe'er); NSF research grants 08929882 and 0845677 (IPe'er); Rachel and Lewis Rudin Foundation (HE); Northwell Health Foundation (TL); Brain & Behavior Foundation (TL); US-Israel Binational Science Foundation (TL, AD); LUNGevity Foundation (ZHG); New York Crohn's Disease Foundation (IPeter); Edwin & Caroline Levy and Joseph & Carol Reich (SB); the Parkinson's Disease Foundation (LNC); the Sharon Levine Corzine Cancer Research Fund (KO); and the Andrew Sabin Family Research Fund (KO).

Author Contributions

TL and IP led the analysis, and TL led the writing of the manuscript. JY and RRK conducted the primary analyses, with assistance from ML and SC. TL led the funding of the study. TL, AD, GA, DB, NB, and LNC provided samples and conducted lab work. TL, IP, NB, SB, AD, JHC, LNC, ZHG, VJ, RK, SL, KO, HO, LJO, IP, AMK, and GA initiated and designed the study, and provided funding.

Competing financial interests: The authors declare no competing financial interests.

Table 1.

		per SCZ case (n=786)	**s.d.**	**per Control (n=463)**	**s.d.**	**p-value**	**95% CI of differences (case - control)**
Post-QC	variants per genome	3676280	17055	3676451	18899	0.87	[-2268, 1925]
	variants per exome	49437	356	49454	387	0.45	[-60, 27]
Post-Filter	variants per genome	5028	427	5034	322	0.81	[-47, 37]
	variants per exome	30.64	6.14	28.66	7.15	1.55×10^{-6}	[1.18, 2.79]
	non-singleton variants per exome	12.88	2.70	13.23	3.01	0.043	[-0.68, -0.012]
	singleton variants (URV) per exome	17.76	6.24	15.44	6.42	6.13×10^{-10}	[1.59, 3.05]
	non-MisLoF	9.86	4.13	8.81	3.97	8.42×10^{-6}	[0.59, 1.52]
	MisLoF	7.90	3.54	6.63	3.51	1.01×10^{-9}	[0.87, 1.68]

TABLE 2.	**Genes in Gene_set**	**Case-only overlap**	**Control-only overlap**	**P(Case)**	**P(control)**	**Chi^2**	**Chi^2 p-value**
DBD[45]	241	11	1	1.62×10^{-7}	5.40×10^{-1}	9.21	2.40×10^{-3}
ASD[46]	102	4	0	2.01×10^{-2}	1.00	4.26	3.91×10^{-2}
ASD_ID_DD[46,47]	274	8	2	6.88×10^{-3}	7.93×10^{-1}	4.04	4.45×10^{-2}
synaptome[48]	1828	26	15	4.59×10^{-2}	8.92×10^{-1}	4.09	4.31×10^{-2}
celf4[49]	2504	41	23	8.50×10^{-4}	8.15×10^{-1}	7.68	5.60×10^{-3}
fmrp[50]	1210	35	19	9.06×10^{-9}	5.62×10^{-2}	6.83	8.98×10^{-3}
rbfox2[51]	2911	48	27	2.04×10^{-4}	8.20×10^{-1}	9.38	2.19×10^{-3}
rbfox13[51]	3255	50	27	8.13×10^{-4}	9.48×10^{-1}	10.95	9.35×10^{-4}
Missense constrained[52]	961	23	8	9.86×10^{-5}	8.12×10^{-1}	8.97	2.74×10^{-3}
LoF constrained (pLI>0.9)[42]	3264	62	35	6.09×10^{-8}	5.16×10^{-1}	13.37	2.55×10^{-4}

TABLE 3.

Biological Processes	**total genes**	**overlap**	**expected**	**raw p**	**FDR**
homophilic cell adhesion via plasma membrane adhesion molecules	165	10	1.11	2.17×10^{-7}	3.45×10^{-3}
cell adhesion	922	19	6.19	1.46×10^{-5}	4.65×10^{-2}
biological adhesion	928	20	6.23	4.42×10^{-6}	3.51×10^{-2}
neurogenesis	1664	27	11.17	1.58×10^{-5}	4.19×10^{-2}
Panther Proteins					
cadherin	18	4	0.12	7.75×10^{-6}	1.51×10^{-3}
cell adhesion molecule	90	5	0.6	3.85×10^{-4}	1.88×10^{-2}
intermediate filament binding protein	15	3	0.1	1.55×10^{-4}	1.51×10^{-2}
intermediate filament	15	3	0.1	1.55×10^{-4}	1.01×10^{-2}
extracellular matrix protein	166	6	1.11	9.63×10^{-4}	3.76×10^{-2}

Figure 1

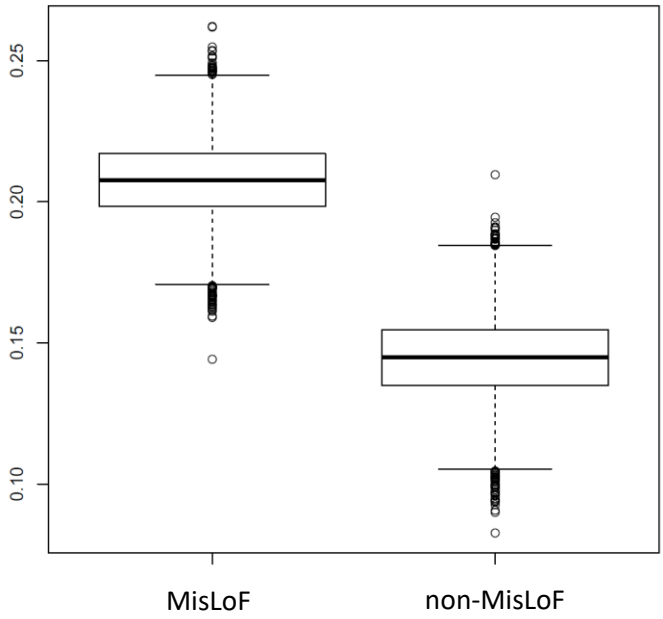
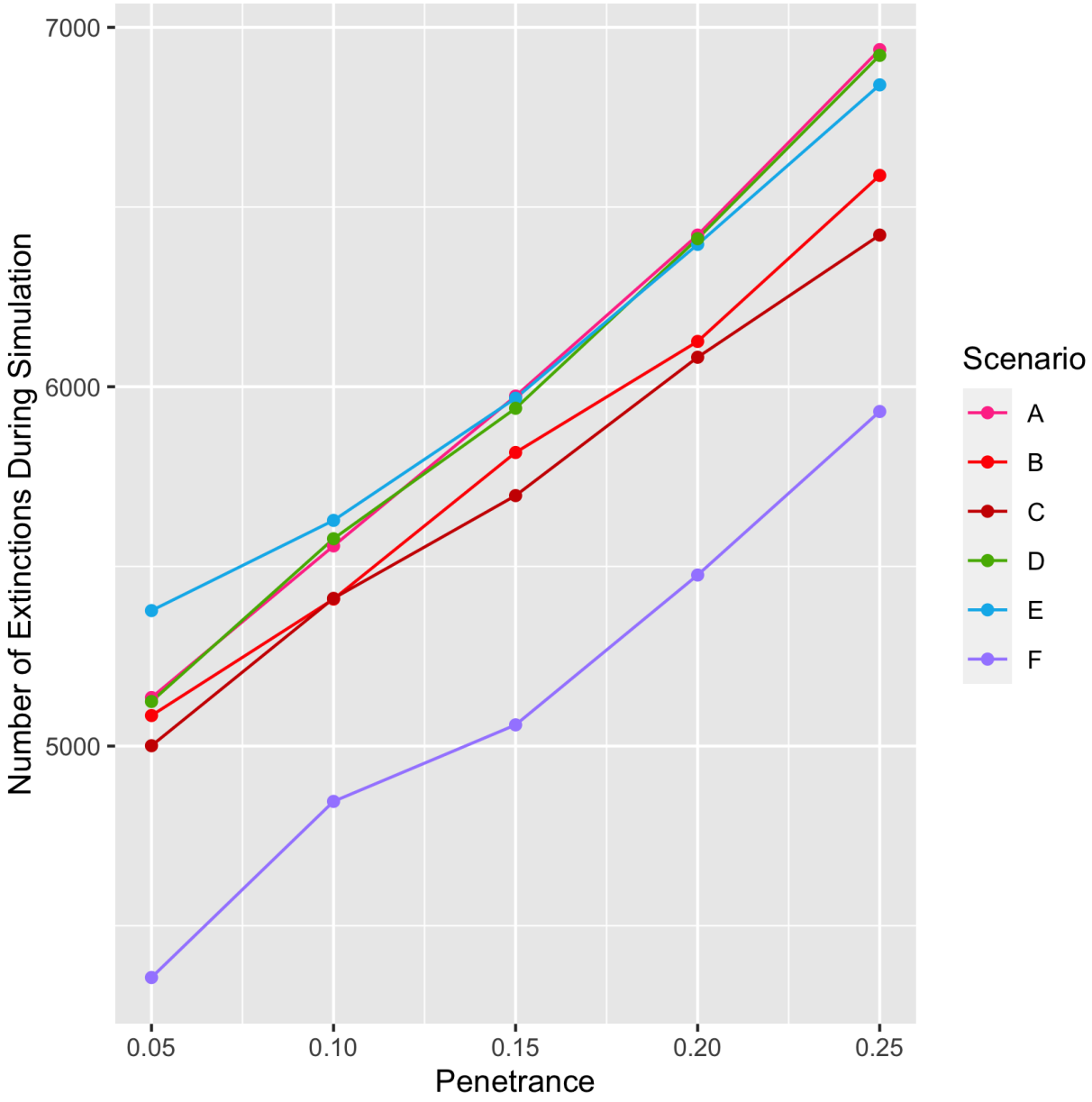


Figure 2.



Supplementary Table 1.

Investigator(s)	All sequenced	Passed for joint-calling	After removing related	Included in final analysis	Female	Male	Phenotypes	Paper describing samples (PMID)
Todd Lencz, Ariel Darvasi	815	807	807	786	283	503	Schizophrenia cases	24253340
Todd Lencz, Ariel Darvasi	132	132	131	130	83	47	controls	24253340
Gil Atzmon	331	304	301	298	178	120	controls	21782286
Judy Cho, Inga Peter	34	33	23	16	6	10	IBD cases	22412388
Laurie Ozelius	34	34	21	19	10	9	Dystonia cases	17503336
TOTAL	1346	1310	1283	1249	560	689		

Supplementary Table 3.

	Genes in Gene_set	Case-only overlap	Control-only overlap	Permutation P(Case)	Permutation P(control)
DBD[45]	241	11	1	0.0034	0.9534
ASD[46]	102	4	0	0.1404	1
ASD_ID_DD[46,47]	274	8	2	0.1429	0.7887
synaptome[48]	1828	26	15	3.00×10^{-4}	0.2102
celf4[49]	2504	41	23	$<1 \times 10^{-4}$	0.1723
fmrp[50]	1210	35	19	3.00×10^{-4}	0.2719
rbfox2[51]	2911	48	27	$<1 \times 10^{-4}$	0.2131
rbfox13[51]	3255	50	27	$<1 \times 10^{-4}$	0.325
Missense constrained[52]	961	23	8	0.0092	0.8221
LoF constrained (pLI>0.9)[42]	3264	62	35	$<1 \times 10^{-4}$	0.1974

Supplementary Table 4.

GO cellular component	total genes	overlap	expected	raw p	FDR
neuron projection (GO:0043005)	1381	23	9.27	5.00×10^{-5}	2.51×10^{-2}
synapse (GO:0045202)	1373	22	9.22	1.30×10^{-4}	2.60×10^{-2}
cell junction (GO:0030054)	2103	31	14.12	2.19×10^{-5}	2.19×10^{-2}
plasma membrane bounded cell projection (GO:0120025)	2208	31	14.83	5.55×10^{-5}	2.23×10^{-2}
cell projection (GO:0042995)	2296	32	15.42	4.70×10^{-5}	3.14×10^{-2}
cytoskeleton (GO:0005856)	2278	31	15.3	9.91×10^{-5}	2.21×10^{-2}
vesicle (GO:0031982)	3886	45	26.09	9.19×10^{-5}	2.30×10^{-2}
organelle (GO:0043226)	13720	117	92.12	2.14×10^{-6}	4.28×10^{-3}
cellular anatomical entity (GO:0110165)	18633	137	125.11	1.32×10^{-4}	2.41×10^{-2}
cellular_component (GO:0005575)	18824	138	126.39	7.80×10^{-5}	2.61×10^{-2}
Unclassified (UNCLASSIFIED)	2027	2	13.61	7.80×10^{-5}	2.23×10^{-2}

Supplementary Table 5.

Cellular Component	**Gene count**	**p-value**	**q-value**
Synapse	18	6.19×10^{-4}	2.25×10^{-3}
presynaptic active zone	5	7.49×10^{-4}	2.25×10^{-3}
Presynapse	10	4.72×10^{-3}	9.45×10^{-3}
integral component of postsynaptic density membrane	3	2.84×10^{-2}	4.27×10^{-2}
Biological Process			
Synapse	18	6.19×10^{-4}	2.25×10^{-3}
presynaptic active zone	5	7.49×10^{-4}	2.25×10^{-3}
Presynapse	10	4.72×10^{-3}	9.45×10^{-3}
integral component of postsynaptic density membrane	3	0.0284	0.0427
Postsynapse	8	0.0935	0.1122

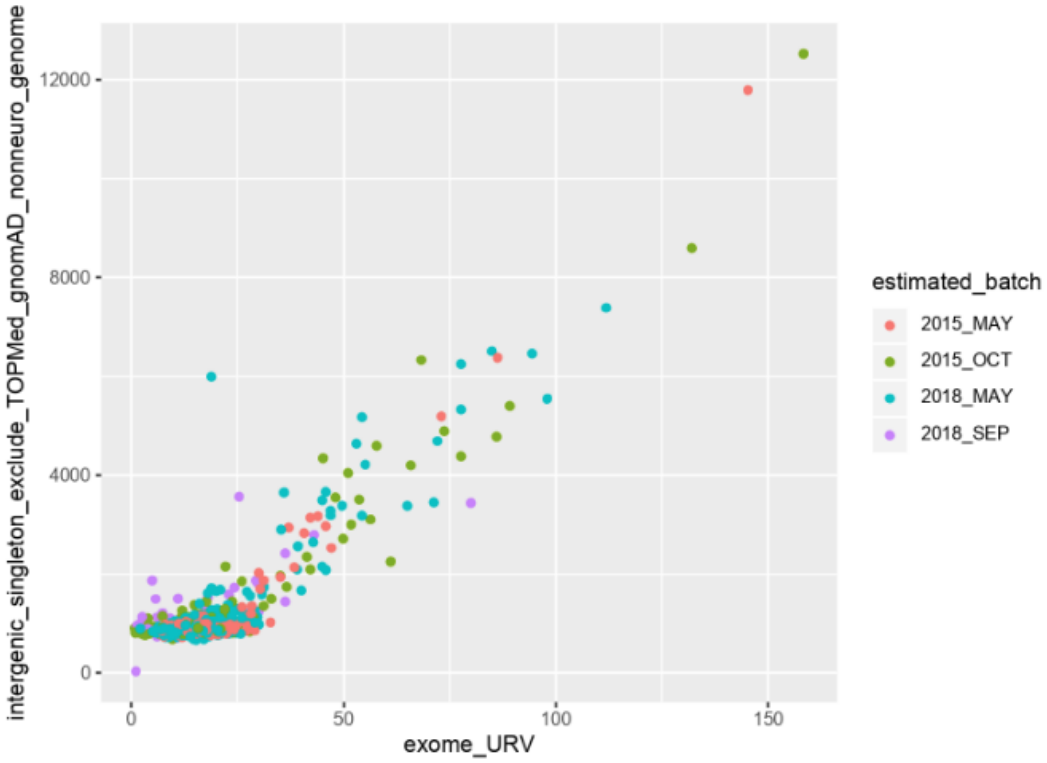
Supplementary Table 6.

N Cases	Gene	Chr	Pos	Ref	Alt	Impact
5	*PCDHA3*	5	140182458	A	G	missense
4	*ACBD6*	1	180382593	T	G	missense
4	*IGF1R*	15	99251324	C	T	missense
3	*ATAD3C*	1	1389777	CA	C	frameshift
3	*ACOX3*	4	8412048	G	A	missense
3	*FIGNL1*	7	50513662	G	A	missense
3	*CEP104*	1	3740011	T	C	missense
3	*UBR4*	1	19426131	A	G	missense
3	*KLHL30*	2	239049577	T	C	missense
3	*TBX18*	6	85457687	G	C	missense
3	*CACNA2D1*	7	81599254	A	C	missense
3	*TENM4*	11	78369465	A	T	missense
3	*TNFRSF1A*	12	6440026	C	A	missense
3	*BCAT1*	12	25002856	T	C	missense
3	*LCP1*	13	46722531	C	T	missense
3	*PCSK2*	20	17240934	A	T	missense
3	*PTK6*	20	62164958	A	G	missense

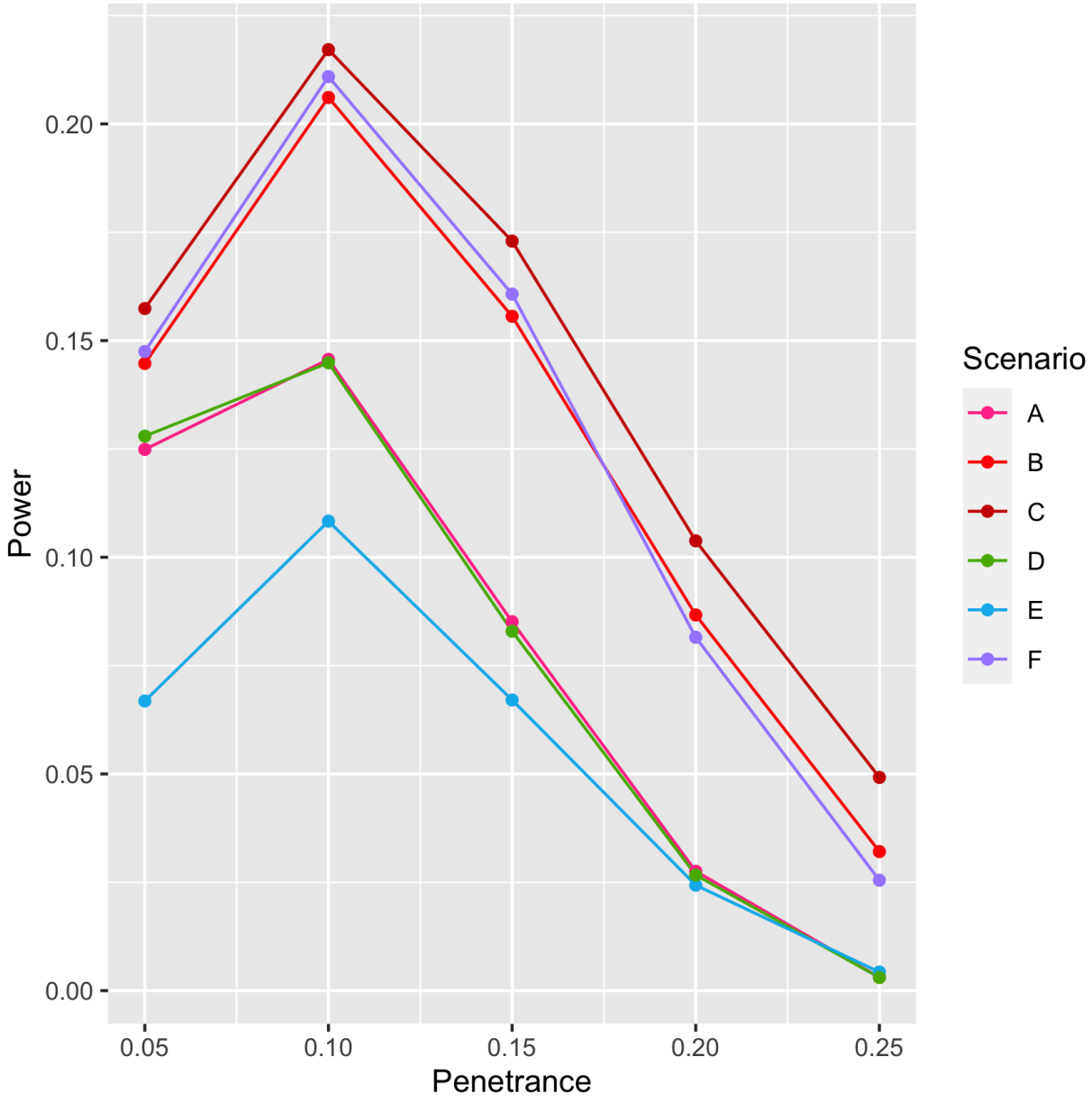
Supplementary Table 7.

Scenario	A	B	C	D	E	F
N Generations	30	30	30	30	35	25
Bottleneck N	300	300	300	500	300	300
Fecundity Ratio (F)	0.5	0.75	0.6	0.5	0.5	0.5
Fecundity Ratio (M)	0.25	0.25	0.4	0.25	0.25	0.25

Supplementary Figure 1.



Supplementary Figure 2.



Supplementary Figure 3.