De novo mutations identified by exome sequencing implicate rare missense variants in *SLC6A1* in schizophrenia

Supplementary Material

Table of Contents

1. Ethics statement

All research conducted as part of this study was approved by ethical bodies and consistent with regulatory and ethical guidelines.

2. Sample description

Bulgarian trios

We sequenced 77 proband-parent trios recruited from Bulgaria whose ascertainment and diagnosis are as described previously[1]. These trios were independent from a previous exome-sequencing study of Bulgarian samples from our group[2]. Briefly, all cases had been hospitalised and met DSM-IV criteria [3] for schizophrenia or schizoaffective disorder based upon SCAN (Schedules for Clinical Assessment in Neuropsychiatry)[4] interview by psychiatrists, and review of case notes. Cases were recruited from general adult psychiatric services and were typical of those attending those services. All participants provided informed consent.

German trios

The German sample included 337 parent-proband trios. Patients were identified through hospital records or during inpatient stays or outpatient clinics. All research subjects and, where applicable, their legal guardians provided a written informed consent to participate in the study. The ethical committee of Wuerzburg reviewed and approved the study. Patients were diagnosed according to ICD-10 criteria, whereby a consensus diagnosis was made by at least two independent, trained raters based on all available clinical information standardized by the AMDP-System (Manual for Assessment and Documentation of Psychopathology in Psychiatry). DNA samples of the participants were extracted from peripheral blood.

Russian trios

The sample included 83 trios. Probands were inpatients at the psychiatric units of the Mental Health Research Centre, Moscow, Russia. All patients were diagnosed with schizophrenia or schizoaffective disorder. The diagnosis was made by two psychiatrists according to diagnostic criteria of ICD-10 and was based on medical records and a semi-structured interview (MINI, SADS). Interviews were conducted by trained researchers. All participants provided a written informed consent to molecular-genetic research. DNA was extracted from peripheral blood.

Spanish trios

The Spanish sample included 37 schizophrenia trios. Patients were diagnosed at the Hospital Gregorio Marañón, and were diagnosed with Schizophrenia or Schizophreniform disorder. Diagnoses were determined by clinical psychiatrists or psychologists, according to DSM-IV criteria with the Structured Clinical Interview for DSM I and II (SCID-I and II) for adults, and the Kiddie-Schedule for Affective Disorders & Schizophrenia, Present & Lifetime Version (K-SADS-PL) for participants aged under 18 years. The diagnostic interviews were administered both at baseline and at 2-years follow-up. DNA was extracted from peripheral blood.

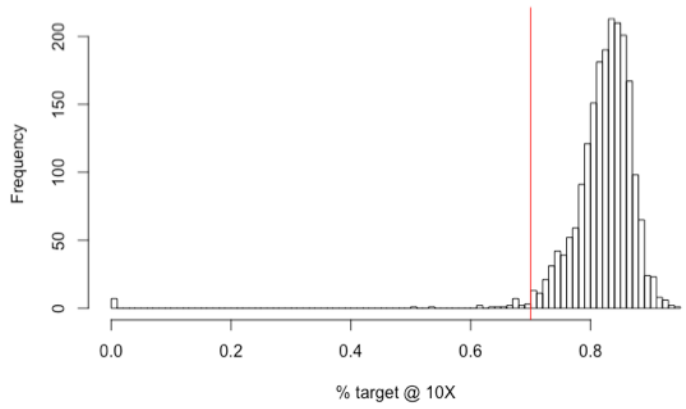
UK trios

The schizophrenia families from the UK were recruited as part of sib-pair and case–control collections. This cohort has been described in detail elsewhere[5]. All probands had received a DSM-IV diagnosis of schizophrenia or schizoaffective disorder, where a consensus diagnosis was made by two independent, trained raters based on all available clinical information including a semi-structured interview [PSE-9 or Assessment of Symptoms and History or Schedules for Clinical Assessment for Neuropsychiatry (SCAN)[4]], examination of case notes and information from relatives and mental health professionals. All interviews were conducted by psychiatrists and psychologists after written consent was obtained following local ethical approval guidelines.

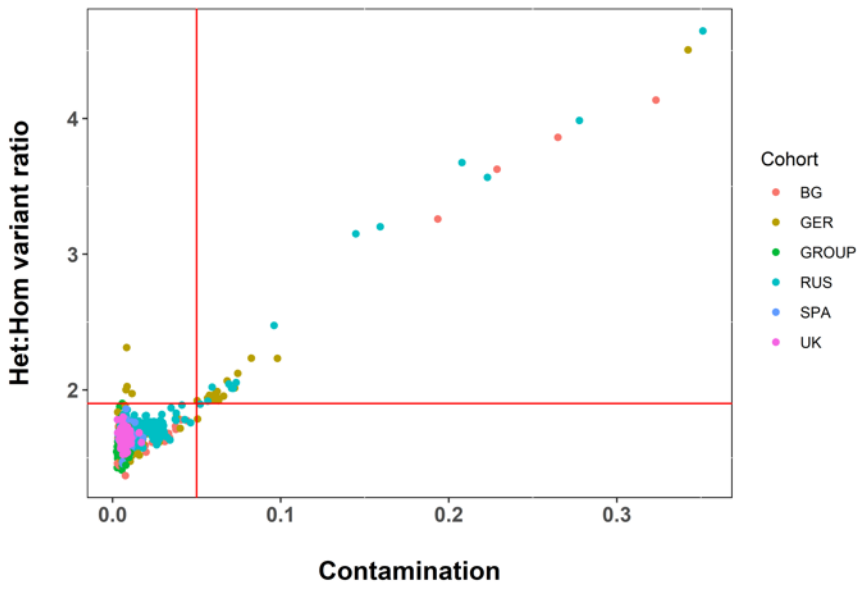
GROUP trios

The 91 GROUP families were recruited from several sites across the Netherlands. Cases were between 16 and 50 years of age, and had received a diagnosis of schizophrenia according to DSM-IV criteria. To assess DSM-IV diagnosis, the Comprehensive Assessment of Symptoms and History (CASH)[6] or SCAN interviews[4] were used. The study protocol was approved centrally by the Ethical Review Board of the University Medical Centre Utrecht and subsequently by local review boards of each participating institute. A detailed description of the GROUP cohort can be found here in Korver et al 2012[7].

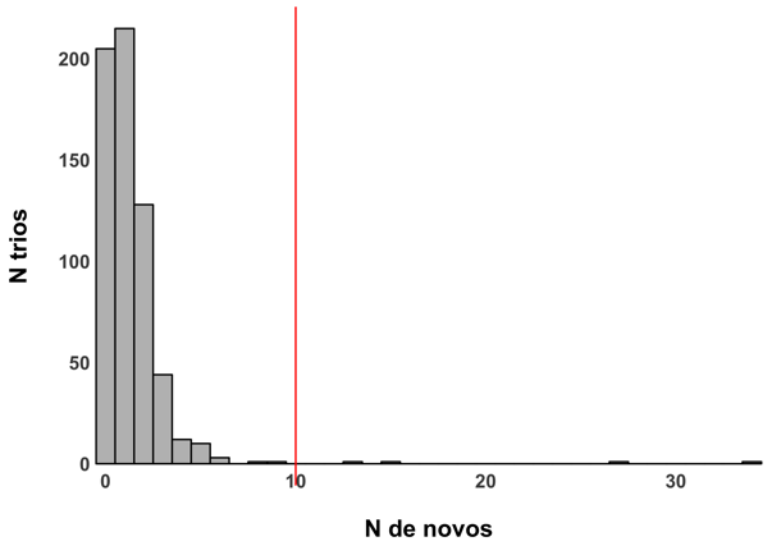
3. Quality control supplementary figures



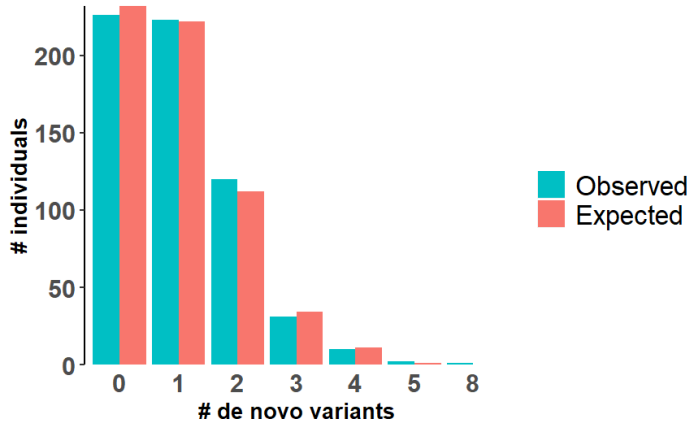
Supplementary Figure 1. Proportion of exome target sequenced to ≥10X coverage in each sample. Red line indicates our cut off (70% of exome target) for excluding samples due to low coverage.



Supplementary Figure 2. Exclusion thresholds for contamination and/or heterozygosity. Contamination was estimated using the FREEMIX sequence only estimate of contamination method[8].



Supplementary Figure 3. Number of *de novo* variants per trio. Red vertical line indicates the threshold used to exclude probands as outliers for number of *de novo* variants.



Supplementary Figure 4. Distribution of number of coding variants per trio.

4. CNV analysis

For the German, Spanish, GROUP and Russian cohorts from the new trio sample, we used Log R ratio and B-allele frequency information from the SNP genotyping data to call CNVs using the PennCNV algorithm, following standard protocol and adjusting for GC content[9]. We performed our standard pipeline for CNV quality control[10] to exclude samples that were outliers for LRR standard deviation, B-allele frequency drift, wave factor and total number of CNVs called per person. As part of our initial QC, individual CNV calls were excluded if they were < 10kb, covered by < 10 probes, > 50% of their length overlapped low copy repeats or the CNV was observed in more than 1% of the sample. This moderately low QC threshold for CNV size and probe coverage was used to maximise our sensitivity to detect *de novo* CNVs. Importantly, to exclude false positive CNVs, we subsequently manually inspected the raw LRR standard deviation and B-allele frequency traces in both parents and child for all putative *de novo* CNVs, defined as CNVs observed in the proband and no overlapping CNV in either parent. We note that all four *de novo* deletions in our new sample that intersect a LoF intolerant gene, and thus contribute to our primary pTDT analysis, were over 200kb in size (mean size = 2.6Mb, Supplementary Table S6).

For the Bulgarian trios, we analysed *de novo* CNVs that have been previously published by our group[1].

5. Identifying sample ancestry through principal component analysis

We performed a principal component (PC) analysis using SNP genotype data from the 1000 genomes project, and then projected our proband-parent trios onto these PCs using EIGENSOFT smartPCA[11]. We did this separately for samples genotyped on Illumina psychchips (Supplementary Figure S5) and Affymetrix 6.0 arrays (Supplementary Figure S6), noting that all members of a trio were genotyped on the same array. For the purpose of excluding trios with non-European ancestry in the polygenic-transmission disequilibrium analysis, we defined European samples as those falling within the box surrounded by black lines shown in the bottom panels of Supplementary Figures S5 and S6. For samples contributing to the pTDT analysis, this resulted in 128/1122 (11.4%) trios being excluded for having non-European ancestry.

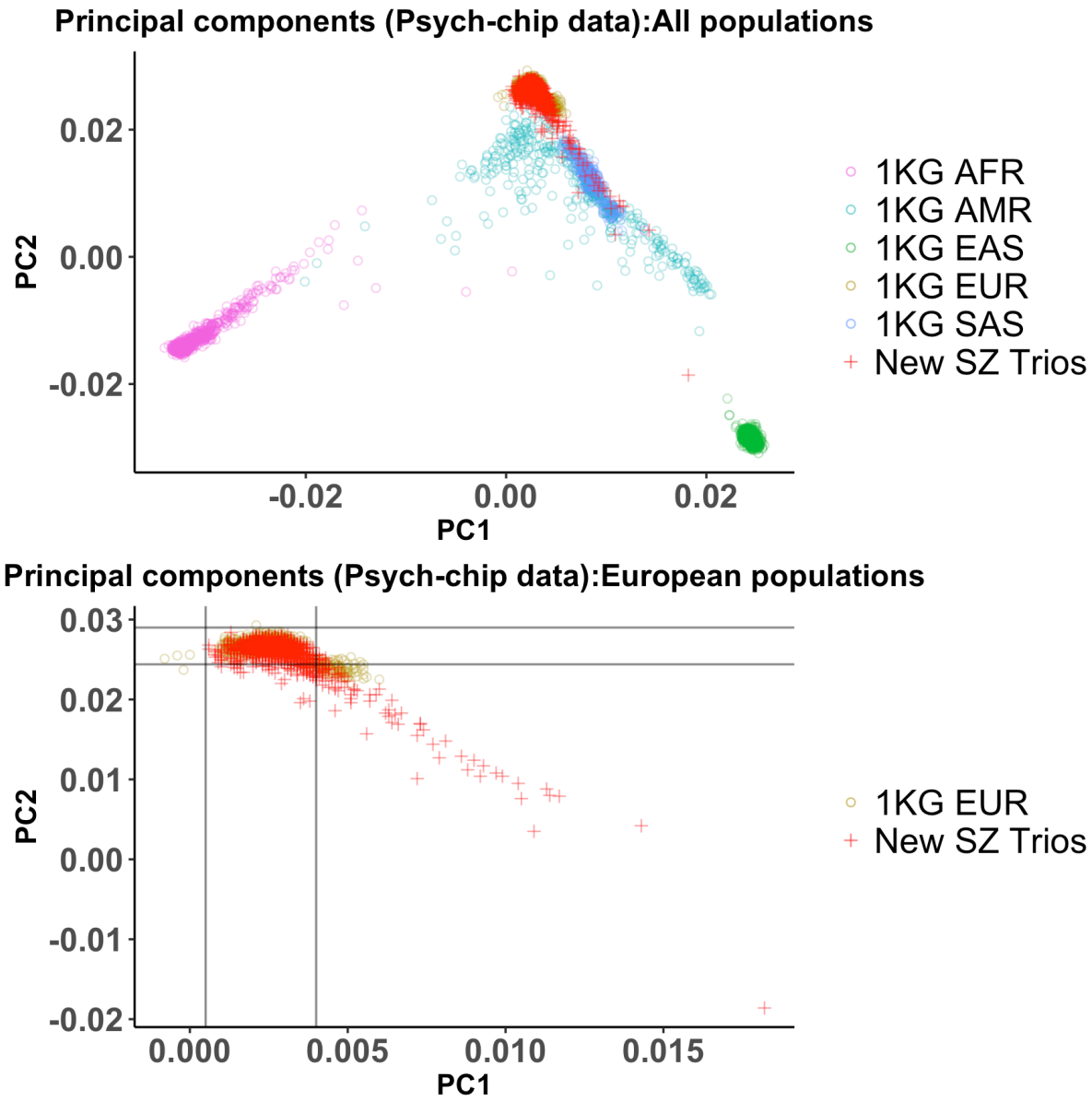


Figure S5. Sample ancestry for 469 trios genotyped on Illumina psych-chip arrays. **Top panel:** Principal components 1 and 2 are shown for the new trios that were genotyped on Illumina psych-chip arrays alongside 5 super-populations from the 1000-genomes dataset (AFR = African, AMR = Ad mixed American, EAS = East Asian, EUR = European, SAS = South Asian). **Bottom panel:** The bottom panel shows the same data as the top panel magnified for European samples. European samples are defined as those falling within the box marked by black lines. Out of the 469 trios genotyped on Illumina psych-chip arrays, 402 had European ancestry.

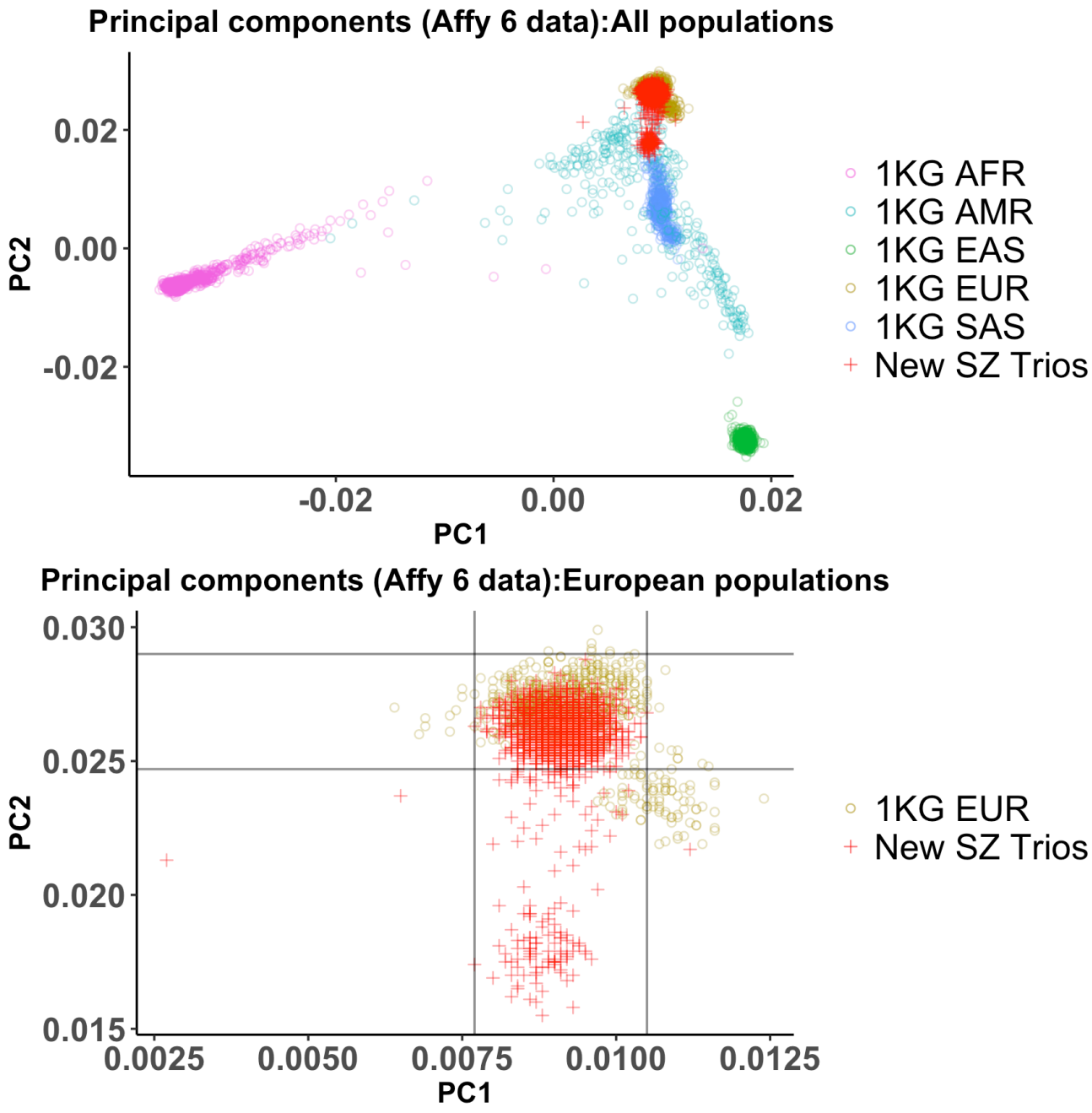


Figure S6. Sample ancestry for 653 trios genotyped on Affymetrix 6.0 arrays. **Top panel:** Principal components 1 and 2 are shown for the new trios that were genotyped on Affymetrix 6.0 arrays alongside 5 super-populations from the 1000-genomes dataset (AFR = African, AMR = Ad mixed American, EAS = East Asian, EUR = European, SAS = South Asian). **Bottom panel:** The bottom panel shows the same data as the top panel magnified for European samples. European samples are

defined as those falling within the box marked by black lines. Out of the 653 trios genotyped on Affymetrix 6.0 arrays, 592 had European ancestry.

6. Sequencing coverage

In our new exome sequencing data set, the median proportion of the exome target that was covered at ≥ 10X across all samples was 83%. This level of coverage is modestly reduced compared with our previous schizophrenia study[2], which had a median of 93% of the exome target covered at ≥ 10X across all samples. Significantly low coverage would result in *de novo* variants in some true carriers being missed; however, as we show in Table S1, the rate of different classes of *de novo* variant in our new trios does not significantly differ to that reported in previous studies of schizophrenia, suggesting coverage has not impacted variant discovery in the present study any more than in previous ones. It is most important to note that coverage issues cannot explain our results; almost all *de novo* variants used to define 'carriers' in our main pTDT analysis have been validated either in the current study or in our previous study by Fromer et al 2014[2]. Additionally, should low coverage mean we miss *de novo* variants in some true carriers, and possibly falsely assign *de novos* to others, the effect of this misclassification would be to blur the distinction between carriers and non-carriers of mutations, and therefore obscure potential pTDT differences. Thus, our finding of a pTDT difference cannot be attributed to low coverage. Our other main finding relates to *SLC6A1*; all mutations in that gene have been validated by Sanger sequencing. Thus, coverage cannot explain our main findings, indeed if there was any failure to detect true mutations in that gene, ours is a conservative estimate of the association signal.

7. Enrichment of *de novo* variants in alternative definitions of LoF intolerant genes

In our primary analysis, we defined LoF intolerant genes as genes with a pLi score ≥ 0.9, using pLi metrics generated from the non-psychiatric component of ExAC. Following review, we have tested the enrichment of LoF *de novo* variants in LoF intolerant genes using the following alternative definitions of LoF intolerance:

1) **gnomAD pLi:** Genes with pLi scores ≥ 0.9, based on constraint metrics generated from the gnomAD dataset[12].

2) **gnomAD observed/expected ratio**: Genes with a loss-of-function observed/expected upper bound fraction score ≤ 0.35, which is the new gnomAD constraint score and recommended threshold for defining LoF intolerance (see [12] and https://macarthurlab.org/2018/10/17/gnomad-v2-1/ for further details).

The degree of enrichment of LoF *de novo* variants in schizophrenia is similar regardless of the definition of LoF intolerant genes, with all 95% CIs overlapping (Table S15).

LoF intolerant gene Set (n genes)	**Rate ratio (95% CI)**	***P***
ExAC non-psych pLi (3,471)	1.58 (1.28, 1.96)	2.5×10^{-5}
gnomAD pLi (3,117)	1.59 (1.27, 1.98)	3.41×10^{-5}
gnomAD observed/expected ratio (2,977)	1.72 (1.38, 2.13)	9.0×10^{-7}

Table S15. Enrichment of LoF *de novo* variants in all schizophrenia trios (n=3,444) in different definitions of LoF intolerant genes. ExAC non-psych pLi = genes with pLi score ≥ 0.9 based on the non-psychiatric component of ExAC; gnomAD pLi = Genes with pLi scores ≥ 0.9, based on constraint metrics generated from the gnomAD dataset; gnomAD observed/expected ratio = Genes with a loss-of-function observed/expected upper bound fraction score ≤ 0.35, based on gnomAD. Enrichment P-values were generated using a two-sided two-sample Poisson rate ratio test. P-values are not adjusted for multiple comparisons.

Moreover, changing our definition of LoF intolerant genes to gnomAD makes no difference to our pTDT results (ExAC pLi: n carriers = 48, mean carrier pTDT = 0.1 (-0.15, 0.35); gnomAD pLi: n carriers = 46, mean carrier pTDT = 0.092 (-0.17, 0.35); gnomAD observed/expected ratio: n carriers = 47, mean carrier pTDT = 0.085 (-0.17, 0.34)). To avoid *post-hoc* effects, we present in the main text our primary analysis based ExAC non-psychiatric pLi scores rather than cherry pick the 'best' result.

8. References

1 Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142-153 (2012).

2 Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).

3 American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR®*. (American Psychiatric Pub, 2000).

4 Wing, J. K. *et al.* SCAN: Schedules for Clinical Assessment in Neuropsychiatry. *Arch. Gen. Psychiatry.* **47**, 589-593 (1990).

5 Williams, N. M. *et al.* A Two-Stage Genome Scan for Schizophrenia Susceptibility Genes in 196 Affected Sibling Pairs. *Hum Mol Genet* **8**, 1729-1739 (1999).

6 Andreasen, N. C., Flaum, M. & Arndt, S. The Comprehensive Assessment of Symptoms and History (CASH): An Instrument for Assessing Diagnosis and Psychopathology. *Arch Gen Psychiatry* **49** (1992).

7 Korver, N. *et al.* Genetic Risk and Outcome of Psychosis (GROUP), a multi site longitudinal cohort study focused on gene–environment interaction: objectives, sample characteristics, recruitment and assessment methods. *Int J Methods Psychiatr Res* **21**, 205-221 (2012).

8 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848 (2012).

9 Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17** (2007).

10 Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* **204**, 108-114 (2014).

11 Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet* **2** (2006).

12 Karczewski, K. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. doi:10.1101/531210 (2019).

13 Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat Genet* **51**, 88-95 (2019).

14 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241 (2016).

15 Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**, 1161-1170 (2018).

16 Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci* **19**, 1433-1441 (2016).