nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-019-0565-2

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

Elliott Rees[1], Jun Han[1], Joanne Morgan[1], Noa Carrera[1], Valentina Escott-Price[1], Andrew J. Pocklington[1], Madeleine Duffield[1], Lynsey S. Hall[1], Sophie E. Legge[1], Antonio F. Pardiñas[1], Alexander L. Richards[1], Julian Roth[2], Tatyana Lezheiko[3], Nikolay Kondratyev[3], Vasilii Kaleda[3], Vera Golimbet[3], Mara Parellada[4], Javier González-Peñas[4], Celso Arango[4], GROUP Investigators[17], Micha Gawlik[2], George Kirov[1], James T. R. Walters[1], Peter Holmans[1], Michael C. O'Donovan[1*] and Michael J. Owen[1*]

Schizophrenia is a highly polygenic disorder with important contributions from both common and rare risk alleles. We analyzed exome sequencing data for de novo variants (DNVs) in a new sample of 613 schizophrenia trios and combined this with published data to give a total of 3,444 trios. In this new data, loss-of-function (LoF) DNVs were significantly enriched among 3,471 LoF-intolerant genes, which supports previous findings. In the full dataset, genes associated with neurodevelopmental disorders ($n = 159$) were significantly enriched for LoF DNVs. Within these neurodevelopmental disorder genes, *SLC6A1*, which encodes a γ-aminobutyric acid transporter, was associated with missense-damaging DNVs. In 1,122 trios for which genome-wide common variant data were available, schizophrenia and bipolar disorder polygenic risk were significantly overtransmitted to probands. Probands carrying LoF or deletion DNVs in LoF-intolerant or neurodevelopmental disorder genes had significantly less overtransmission of schizophrenia polygenic risk than did non-carriers, which provides a second robust line of evidence that these DNVs increase liability to schizophrenia.

Genetic liability to schizophrenia involves a combination of rare and common risk alleles that are distributed across the genome[1]. Common schizophrenia risk alleles with odds ratios of <1.3 account for at least one third of genetic liability[2–4], although only a small fraction of this is captured by the 145 genome-wide significant loci that were implicated in the largest published genome-wide association study (GWAS) of the disorder[5]. At the other end of the frequency spectrum, rare copy number variants (CNVs) and rare coding variants, both of which sometimes occur as DNVs, have been implicated in the disorder[6–8]. Although CNVs and rare coding variants are enriched in schizophrenia, not all rare variants that are observed in individuals with schizophrenia, including those occurring de novo, are expected to be aetiologically relevant, as there is a baseline burden of these variants in the general population.

In people with other neurodevelopmental disorders in which CNVs and rare coding variants have a role, particularly autism spectrum disorder (ASD)[9,10] and developmental delay[11,12], the enrichment for rare coding variants is greatest in genes that are classified as intolerant to LoF variants (that is, variants that introduce premature stop codons or frameshifts in the encoded protein, or that are predicted to disrupt messenger RNA splicing). This indicates that rare coding variants in these genes are more likely to be pathogenic for those disorders than are rare coding variants occurring elsewhere in the genome. Moreover, greater enrichment is found for LoF DNVs than for missense DNVs that change an encoded amino acid, which indicates that the former class of mutation is particularly likely to be pathogenic. Similar observations have been made in schizophrenia, in which an excess of LoF DNVs was found to be largely restricted to LoF-intolerant genes[7], although the degree of enrichment is lower than that for ASD or developmental disorders.

In studies of ASD and developmental disorders, a marked excess of rare coding variants has been observed for 99 and 93 genes, respectively, 33 of these genes overlapping between these disorders[9,11]. Only two genes, *SETD1A* (ref. [13]) and *RBM12* (ref. [14]), are currently associated with rare coding variants in schizophrenia. This is partly because of lower statistical power, as the number of trios that have been exome sequenced in studies of schizophrenia ($n = 2,831$) is lower than those that have been exome sequenced in equivalent studies of developmental disorders ($n = 7,580$) (ref. [11]) and ASD ($n = 6,430$) (ref. [9]), but it also reflects the weaker enrichment in schizophrenia for this type of variant. As a set, genes disrupted by DNVs in neurodevelopmental disorders are also enriched for DNVs in schizophrenia[15,16], and therefore some of the genes that are implicated in ASD and developmental disorders by rare coding variants may also be involved in the etiology of schizophrenia.

[1]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [2]Department of Psychiatry and Psychotherapy, University of Würzburg, Würzburg, Germany. [3]Clinical Genetics Laboratory, Mental Health Research Centre, Moscow, Russia. [4]Department of Child and Adolescent Psychiatry, Hospital General Universitario Gregorio Marañón, IisGM, School of Medicine, Universidad Complutense, CIBERSAM, Madrid, Spain. [17]A full list of authors and affiliations appears at the end of the paper.
*e-mail: odonovanmc@cardiff.ac.uk; owenmj@cardiff.ac.uk

With the aim of contributing to the schizophrenia rare variant discovery effort, we have undertaken exome sequencing in a new sample of 613 schizophrenia trios, and have combined our data with published data from 2,831 trios, which includes 617 trios that were sequenced previously by our group[15], to provide the largest analysis of coding DNVs in schizophrenia to date. Given the anticipated modest power even of this sample, as we have done before for CNV analysis[17], we exploited the well documented overlap in the genetic aetiologies of schizophrenia, ASD and developmental disorders, to undertake a hypothesis-focused analysis of neurodevelopmental disorder genes in schizophrenia, which highlights *SLC6A1* as a novel risk gene.

The involvement of common variant polygenic risk in schizophrenia is already established[2,4,18], but few existing studies have empirically examined the relationships between different classes of rare and common variants. An early case-control exome sequencing study of schizophrenia found evidence for independent additive effects for common alleles, rare CNVs and rare coding variants when cases were compared with controls, but found no within-case correlation between the burden of each type[19]. More recent evidence indicates a negative correlation within cases for schizophrenia-associated CNV carrier status and common risk variant burden, which is consistent with the hypothesis that the common and rare alleles act together[20,21]. Thus, compared to controls, affected carriers of schizophrenia-associated CNVs have an increased burden of common schizophrenia risk alleles as measured by the polygenic risk score (PRS)[21], but in a within-case analysis, this burden is inversely proportional to the estimated effect size of the implicated CNV[20]. In ASD and developmental disorders, common variant polygenic risk has been shown to be overtransmitted from parents to probands, but no difference has been reported between those that do or do not carry a disorder-associated DNV[22,23]. As yet, the relationship between de novo mutations and common allele risk has not been studied in schizophrenia. Here, we examine this relationship using the polygenic transmission disequilibrium test (pTDT)[23]. Specifically, we show that people with schizophrenia who are carriers of DNVs in gene sets that are proposed to be relevant to schizophrenia have a lower common risk allele burden than do people with schizophrenia who are not carriers.

Results

De novo mutation rates. After sample and variant quality control was carried out (see Methods and Supplementary Figs. 1–3), 606 coding DNVs were observed in 613 probands (433 males and 180 females), which corresponds to a rate of 0.99 (s.e.m. = 0.041) events per proband; this is not significantly different to the rate that was observed in a sample of 2,831 schizophrenia trios that was published previously (previous de novo rate, 1.004; rate ratio (95% confidence interval (CI)) = 0.98 (0.9, 1.08); $P = 0.74$; Supplementary Table 1). Of the coding DNVs, 154 were synonymous, 372 were missense, 15 were inframe indels, 2 were start-loss, 1 was a stop-loss and 62 were LoF (19 stop-gain, 13 splice and 30 frameshift indels). The number of coding DNVs that was observed per trio followed the expected Poisson distribution (Supplementary Fig. 4).

De novo variant enrichment tests. In the new dataset, a significant excess of LoF DNVs was observed among LoF-intolerant genes (Fig. 1, rate ratio (95% CI) = 2.21 (1.3, 3.75); $P = 2.3 \times 10^{-3}$; Supplementary Table 2). Consistent with previous reports, no evidence for DNV enrichment was found in the following negative control gene set tests: LoF DNVs in LoF-tolerant genes (Fig. 1), synonymous DNVs in LoF-intolerant genes, and synonymous DNVs in LoF-tolerant genes (Supplementary Table 3). After the new trio data were combined with data from 2,831 trios that had been published previously, LoF DNVs were found to be enriched in LoF-intolerant genes with a rate ratio (95% CI) of 1.58 (1.28, 1.96) ($P = 2.5 \times 10^{-5}$)

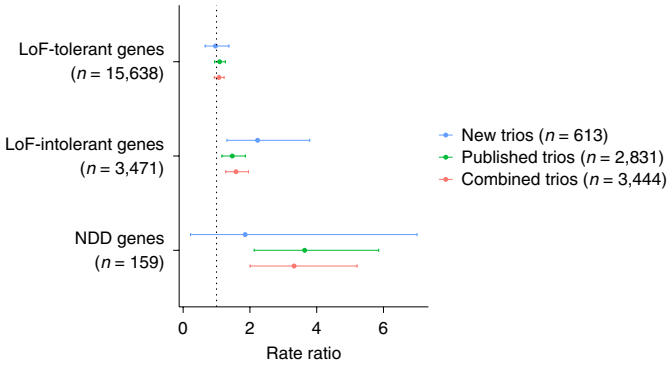

Fig. 1 | Gene set enrichment for loss-of-function de novo variants. LoF DNVs were tested in LoF-intolerant genes and in neurodevelopmental disorder genes. For LoF-intolerant and neurodevelopmental disorder gene sets, rate ratios and 95% CIs are relative to the baseline DNV rate, which is defined as the LoF DNV enrichment observed for all genes outside the given set. LoF DNV enrichment for LoF-tolerant genes is shown as a negative control. A breakdown of the LoF-intolerant and neurodevelopmental disorder gene set results is provided in Supplementary Tables 2 and 3. NDD, neurodevelopmental disorder. The dashed vertical line indicates a rate ratio of 1 (i.e., no enrichment of de novo variants).

(Fig. 1 and Supplementary Table 2). Following the manuscript peer-review process, alternative definitions of LoF-intolerant genes were tested, based on constraint metrics generated from the Genome Aggregation Database (gnomAD) dataset[24]; the degree of enrichment of LoF DNVs in schizophrenia is similar regardless of the definition of LoF-intolerant genes (see Supplementary Material for full results).

In the combined trio data, no individual gene was significantly enriched for LoF DNVs after correction for all genes tested ($n = 19{,}109$). The novel gene that showed the most significant result was *CUL1*, which had two LoF DNVs in the new trios and one additional LoF DNV in the published trios (Table 1).

We have shown previously that rare CNVs that increase the risk of developing schizophrenia are effectively confined to those that also influence other neurodevelopmental disorders[17]. When neurodevelopmental disorder genes are defined as those that are significantly enriched for rare coding variants in recent large studies of ASD[9] and developmental disorders[11] ($n = 159$), neurodevelopmental disorder genes were significantly enriched for LoF DNVs in the combined trio data (Fig. 1; rate ratio (95% CI) = 3.3 (2.0, 5.17); $P = 8.2 \times 10^{-6}$; Supplementary Table 2), and this enrichment was significantly greater than for LoF-intolerant genes (rate ratio (95% CI) = 2.37 (1.41, 3.8); $P = 8.8 \times 10^{-4}$). In the full sample of trios, no enrichment of missense-damaging DNVs was observed for subgenic regions that have been identified as depleted for missense variation[25] (rate ratio (95% CI) = 1.004 (0.85, 1.18); $P = 0.9$). The rate of missense-damaging DNVs in neurodevelopmental disorder genes was elevated compared with the background rate (rate ratio (95% CI) = 1.53 (0.79, 2.7)), but this is not significant ($P = 0.16$), which possibly reflects the small number of DNVs in neurodevelopmental disorder genes ($n = 13$).

To exploit the strong enrichment among neurodevelopmental disorder genes for DNVs in schizophrenia, a focused analysis of genes in this set was undertaken, with the aim of identifying high probability schizophrenia risk genes. As highlighted in the study of ASD[9], association of the disorder with some neurodevelopmental disorder genes is driven by LoF variants alone, by a combination of LoF variants and missense variants, and in some cases, primarily by missense variants. Therefore, all of those classes of mutation were considered in our analysis. All LoF or missense-damaging DNVs

Table 1 | Genes disrupted by two or more loss-of-function de novo variants

Gene	New trios ($n=613$)		Published trios ($n=2,831$)		All trios ($n=3,444$)	
	LoF DNVs	*P*	LoF DNVs	*P*	LoF DNVs	*P*
SETD1A	0	1	3	1.90×10^{-6}	3	3.00×10^{-6}
CUL1	2	3.60×10^{-5}	1	0.04	3	2.00×10^{-5}
TAF13	0	1	2	2.40×10^{-5}	2	3.30×10^{-5}
GALNT9	0	1	2	2.90×10^{-5}	2	4.20×10^{-5}
HENMT1	0	1	2	5.50×10^{-5}	2	7.90×10^{-5}
PAF1	1	0.0028	1	0.013	2	0.00013
SV2B	0	1	2	0.00016	2	0.00023
NRXN3	0	1	2	0.00022	2	0.0003
HIVEP3	0	1	2	0.00026	2	0.00035
RB1CC1	0	1	2	0.00046	2	0.00065
SMARCC2	0	1	2	0.0005	2	0.00068
MKI67	0	1	2	0.00085	2	0.0012
CHD8	0	1	2	0.0009	2	0.0013
TENM1	1	0.0077	1	0.046	2	0.0014
TRIO	0	1	2	0.0012	2	0.0016
SCN2A	1	0.012	1	0.057	2	0.0024
DNAH9	0	1	2	0.0018	2	0.0026
KMT2C	0	1	2	0.0086	2	0.012
KIAA1109	0	1	2	0.01	2	0.015
TTN	1	0.16	2	0.22	3	0.092

Individual gene enrichment *P* values were generated using a one-sided Poisson test. The gene that showed the most significant results, *SETD1A*, has been identified previously as a schizophrenia risk gene[13].

that were observed in neurodevelopmental disorder genes and, where available, the phenotypes that were observed in these carriers are presented in Supplementary Table 4.

Significant association of *SLC6A1* with missense-damaging DNVs was observed in our new trio data after correction for three classes of mutation (LoF, missense-damaging and LoF plus missense-damaging) and for 159 neurodevelopmental disorder genes (2 missense-damaging DNVs; $P=7.46\times10^{-5}$; $P_{corrected}=0.036$). This finding was supported in our analysis of all trio data, in which one additional missense-damaging DNV was observed (Table 2; 3 missense-damaging DNVs; $P=5.2\times10^{-5}$, $P_{corrected}=0.025$). It is striking that in the study of ASD[9], association to *SLC6A1* was also driven by missense variants ($n=8$) rather than by LoF variants ($n=1$). The rationale outlined by the Deciphering Developmental Disorders Study[26] was followed as we undertook a combined analysis of schizophrenia and ASD DNVs; the evidence for enrichment of missense-damaging DNVs (MPC ≥ 2; MPC is a missense deleteriousness metric) in *SLC6A1* was more than three orders of magnitude stronger than for ASD alone, which supports the hypothesis that missense variants in this gene contribute to both disorders (combined, $P=1.6\times10^{-14}$; ASD alone, $P=8.0\times10^{-11}$).

Polygenic transmission disequilibrium tests. Schizophrenia and bipolar disorder PRS values were significantly overtransmitted from parents to probands (Fig. 2 and Supplementary Table 5). These results did not differ when the analysis was restricted to trios with European ancestry (as defined by principal component analysis; Supplementary Table 5).

Under a liability threshold model, probands carrying DNVs of large effect size should require less transmission of polygenic risk than probands without such a variant. To test this, the mean pTDT was compared between carriers of candidate schizophrenia-related DNVs and the remainder of the sample. Candidate schizophrenia-related DNVs are defined here as LoF DNVs in a LoF-intolerant gene or in a neurodevelopmental disorder gene. Given that CNV deletions that disrupt LoF-intolerant genes are associated with schizophrenia[7], we also included de novo CNV deletions that disrupt one of these genes as candidate schizophrenia-related DNVs (CNVs contributing to this analysis are presented in Supplementary Table 6; the CNV calling procedure is outlined in the Supplementary Material).

Probands carrying candidate schizophrenia-related DNVs had a significantly lower mean pTDT than that of probands who did not carry one of these DNVs (carrier mean pTDT (95% CI) = 0.07 (−0.15, 0.29); non-carrier mean pTDT (95% CI) = 0.48 (0.43, 0.54); $P=3.5\times10^{-4}$; Fig. 3). Based on the mean pTDT point estimates, the overtransmission of common risk alleles from parents is about sevenfold greater to non-carriers than to carriers of candidate schizophrenia-related DNVs, although this estimate is imprecise given the width of the CIs (Fig. 3). Similar patterns were observed when LoF and deletion DNVs were tested separately (Fig. 3). In a negative control test, the mean pTDT did not significantly differ between probands carrying a synonymous DNV in either a LoF-intolerant or a neurodevelopmental disorder gene, and non-carriers (Fig. 3).

The finding that the mean pTDT deviation for the schizophrenia PRS was significantly different between probands carrying candidate schizophrenia-related DNVs and non-carrying probands was consistent across the schizophrenia PRS training data *P* value thresholds (Supplementary Table 7). Although the pTDT method is expected to be robust to population stratification, the efficacy of PRS as a measure of relative liability varies with the extent to which the ancestry of the sample from which risk alleles are derived (the source GWAS) matches the ancestry of those being tested (in our case the trios). Given that the source GWAS is primarily

Table 2 | Neurodevelopmental disorder genes with at least one loss-of-function or missense-damaging de novo variant observed in schizophrenia

Gene	Observed DNVs			*P* (uncorrected)		
	$Miss_{dam}$	LoF	$Miss_{dam}$ + LoF	$Miss_{dam}$	LoF	$Miss_{dam}$ + LoF
SLC6A1	3	0	3	5.2×10^{-5}*	1	7.9×10^{-5}*
SCN2A	1	2	3	0.15	0.0024	0.0019
SMARCC2	0	2	2	1	0.00068	0.0019
PUF60	1	1	2	0.056	0.022	0.003
MED13L	1	1	2	0.048	0.064	0.0062
DEAF1	0	1	1	1	0.0082	0.011
TRIO	0	2	2	1	0.0016	0.014
CHD8	0	2	2	1	0.0013	0.023
CHD4	1	1	2	0.2	0.04	0.03
KMT2C	0	2	2	1	0.012	0.04
PTEN	1	0	1	0.029	1	0.044
GNAO1	1	0	1	0.042	1	0.052
TEK	0	1	1	1	0.025	0.057
AUTS2	0	1	1	1	0.03	0.057
CSNK2A1	1	0	1	0.052	1	0.064
POGZ	0	1	1	1	0.048	0.066
NACC1	1	0	1	0.08	1	0.085
KDM5B	0	1	1	1	0.084	0.092
TLK2	1	0	1	0.075	1	0.1
KDM6B	0	1	1	1	0.025	0.13
GRIN2B	1	0	1	0.15	1	0.17
SYNGAP1	0	1	1	1	0.026	0.21

Enrichment *P* values were generated using a one-sided Poisson test from the analysis of all schizophrenia trios ($n = 3{,}444$). $Miss_{dam}$, missense-damaging (MPC score ≥2). *P values that survive correction for 159 neurodevelopmental disorder genes and 3 mutation classes (LoF, missense-damaging and LoF plus missense-damaging).

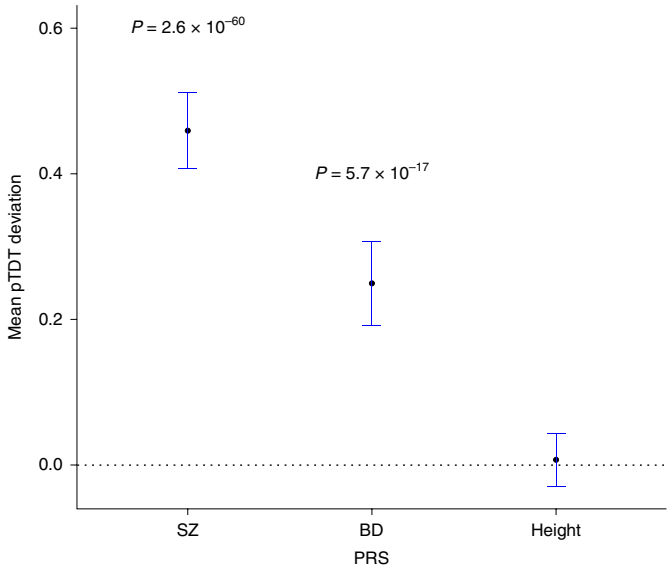

Fig. 2 | Mean pTDT deviation and 95% confidence intervals for schizophrenia, bipolar disorder and height polygenic risk scores. One-sided one-sample *t*-tests were used to evaluate polygenic overtransmission in 1,122 schizophrenia (SZ) proband–parent trios. Polygenic risk for schizophrenia and bipolar disorder (BD) is significantly overtransmitted to schizophrenia probands.

of European ancestry, we confirmed that our findings remained the same when our analysis was restricted to trios with European ancestry (Supplementary Figs. 5,6) despite the smaller sample size (all results for European-only trios are presented in Supplementary Tables 7,8).

The mean pTDT in carriers of candidate schizophrenia-related DNVs was not significantly greater than the null (Fig. 3). Based on the pTDT standard deviation that was observed for the schizophrenia PRS in all trios (0.89), the test had only 80% power to detect a significant ($\alpha = 0.05$) mean pTDT of 0.4 in the 63 carriers of candidate schizophrenia-related DNVs. Therefore, although we can be confident that the overtransmission to candidate DNV carriers is less than to non-carriers, power limitations mean that it cannot be concluded that candidate DNV carriers have no contribution from common alleles.

Following the manuscript peer-review process, an exploratory analysis was performed to evaluate whether pTDT was lower in carriers of additional classes of DNV. Despite testing a wide range of alternative variant filters (for example, excluding DNVs observed in gnomAD, or, as it was previously known, the Exome Aggregation Consortium (ExAC)), missense annotations (for example, MPC scores and constrained coding regions), and CNVs intersecting only LoF-tolerant genes, no set of DNV carriers had a significantly greater reduction in pTDT than that observed for our primary candidate schizophrenia-related set of DNVs defined above (see Supplementary Table 8 for all results).

Discussion

Proband–parent trio studies have identified large numbers of genes associated with DNVs in ASD and developmental disorders[9,11]. Although similar studies in schizophrenia have revealed general pathophysiological insights into the disorder, such as a role for proteins involved in postsynaptic signaling complexes[15,27], schizophrenia gene discovery through DNV analysis has been hindered by small sample sizes. To add to efforts to overcome this limitation, exome sequencing was performed on a new sample of 613 schizophrenia trios. We confirmed previous work that showed significant enrichment of schizophrenia LoF DNVs among a set of 3,471 genes that are intolerant to this class of mutation, and we identified a greater enrichment of DNVs in a smaller set of 159 genes that are associated with rare coding variants in neurodevelopmental disorders. GWAS data suggest that common risk alleles are under negative selection[28] and are enriched in highly conserved genes[5], but that they are nevertheless maintained by population mechanisms related to background selection and genetic drift[5,29]. The findings from rare mutations, both CNVs[7,30] and rare coding variants[7,31], also support a role for deleterious point mutations in schizophrenia that are under more intense negative selection than are alleles of weak effect. Despite this, the population burden of schizophrenia risk alleles appears to be maintained by mutation-selection balance; for CNVs, strong selection is balanced by their relatively high mutation rates[30], whereas for exonic mutations, in the face of a low per-base mutation rate, balance is likely to be maintained by the large size of the mutational target.

In our analysis of all schizophrenia trios, no novel gene was unequivocally associated with DNVs after correction for all genes tested. Despite conducting the largest analysis of DNVs in schizophrenia so far, it is clear that even larger samples are required to identify specific risk genes with genome-wide levels of significance. However, our exploitation of the observation here of strong enrichment for DNVs in known neurodevelopmental disorder genes,

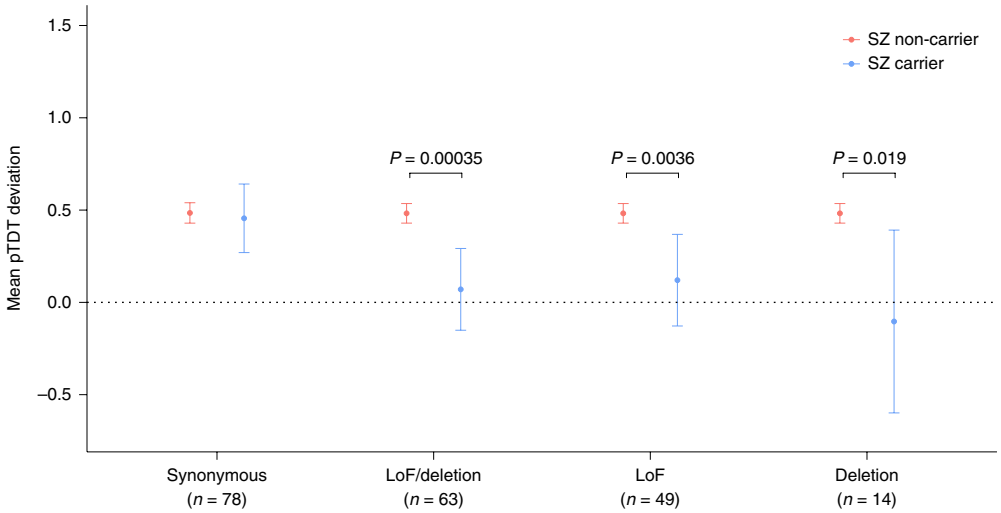

Fig. 3 | Mean pTDT deviation and 95% confidence interval for schizophrenia PRS. Results are shown for probands carrying various classes of DNV in a LoF-intolerant gene or in a neurodevelopmental disorder gene; our primary analysis defined schizophrenia (SZ) carriers as probands with a LoF or deletion DNV in a LoF-intolerant gene or in a neurodevelopmental disorder gene (LoF/deletion). Results are also shown separately for carriers of LoF and deletion DNVs, respectively. A one-sided two-sample *t*-test was used to compare the mean pTDT deviation scores across groups of trios ($n = 1{,}122$ trios).

provides evidence for association between *SLC6A1*, which encodes a sodium-dependent GABA transporter (also known as GAT1), and missense-damaging DNVs, and our approach is based on the wealth of data that shows that rare CNVs that increase the risk of developing schizophrenia are effectively confined to those that also influence other neurodevelopmental disorders[17]. *SLC6A1* is involved in reuptake of the inhibitory neurotransmitter GABA from the synaptic cleft; our finding therefore adds to the evidence for perturbation of GABAergic neuronal signaling in genetic risk for schizophrenia[32]. Consistent with our findings, the largest study of rare coding variants in ASD found *SLC6A1* to be the most significant of only 4 genes for which the association signal was driven by missense-damaging variants (8 missense and 1 LoF DNVs)[9]. In myoclonic atonic epilepsy and other developmental disorders, LoF variants account for 54% and 30% of the observed nonsynonymous DNVs, respectively[9]. Given the strong convergent evidence for *SLC6A1*, and specifically for a role for missense mutations, with that from other neurodevelopmental disorders, this gene is also highly likely to be involved in schizophrenia. This conclusion is further supported by the result of the DNV missense meta-analysis of ASD and schizophrenia, in which the combined evidence for association is more than three orders of magnitude stronger than the (already strong) evidence for association with ASD alone, and surpasses genome-wide significance by eight orders of magnitude. Given the small number of DNVs in *SLC6A1*, it will be important to extend our finding in other samples, and clearly, a greater number of DNVs will be required to establish that risk is conferred largely by missense rather than by LoF mutations.

The role of polygenic risk in schizophrenia has been widely studied using large case-control samples, but, to our knowledge, the pTDT method has not been used to investigate polygenic risk in schizophrenia. The pTDT method has several advantages over case-control PRS studies as it is not confounded by ancestry or ascertainment bias, or by the possibility of effects arising from 'super-healthy' controls in discovery GWAS and subsequent PRS test samples[23]. Our results provide strong refutation that such effects might explain the PRS effects that have been widely publicized in the literature, including that of overlap in risk between schizophrenia and bipolar disorder.

More importantly in the present context, our finding that carriers of LoF DNVs in genes that are defined by LoF intolerance, or in a known neurodevelopmental disorder gene, have significantly lower distortion of transmission of polygenic liability from the mean parental PRS than do non-carriers, provides statistically independent evidence that a substantial proportion of this class of DNV contributes to schizophrenia pathogenesis. This is an important observation given the possibility that gene set enrichments documented previously in cases of these variants could have been driven by errors in the calibration of the expected mutation rate, or by technical issues arising from comparing cases and controls (or case-control trios) that are often derived opportunistically from different studies.

Our limited sample size does not allow accurate estimation of the magnitude of difference in the transmission distortion between probands that carry candidate schizophrenia-related DNVs and those that do not, but the point estimate is that the distortion in non-carriers is about sevenfold compared to that in carriers (and almost tenfold when restricted to those of European ancestry). This suggests that, on average, the candidate DNVs contribute a substantial amount of liability in those who carry them. Indeed, in the present study, carriers of candidate schizophrenia-related DNVs did not significantly over-inherit a common allele burden from their parents, which is consistent with the idea that DNVs in LoF-intolerant genes act as monogenic risk factors in those who carry them. However, it is important to stress that the latter finding is also consistent with limited power (as discussed in the results) rather than with the absence of a role for common variation in the carriers, and it should be noted that the point estimate for the pTDT in candidate DNV carriers is greater than zero. It will be important for future larger studies to determine whether differences in coaction between common and rare risk alleles exist between schizophrenia and neurodevelopmental disorders. Meanwhile, with respect to the genetic architecture of schizophrenia, together with previous findings from CNVs alone[20,21], our data can be interpreted as consistent with a polygenic liability threshold model of schizophrenia[33].

In conclusion, this study provides further evidence that certain classes of DNV are associated with increased risk for schizophrenia. We highlight strong evidence that mutations in *SLC6A1*, a gene known to be associated with ASD, developmental disorders and epilepsy, confer a high risk for schizophrenia. By combining exome sequencing and GWAS data, we show that carriers of candidate schizophrenia-related DNVs inherit significantly fewer common

risk alleles than do non-carrying cases, which provides strong, statistically independent, evidence that these DNVs contribute to schizophrenia liability.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-019-0565-2.

Received: 16 April 2019; Accepted: 22 November 2019;
Published online: 13 January 2020

References

1. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
2. Ripke, S. et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
3. Lee, S. H. et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
4. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
5. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
6. Rees, E., O'Donovan, M. C. & Owen, M. J. Genetics of schizophrenia. *Curr. Opin. Behav. Sci.* **2**, 8–14 (2015).
7. Singh, T. et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat. Genet.* **49**, 1167–1173 (2017).
8. Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
9. Satterstrom, F. K. et al. Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. Preprint at *bioRxiv* https://doi.org/10.1101/484113 (2018).
10. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
11. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
12. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
13. Singh, T. et al. Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
14. Steinberg, S. et al. Truncating mutations in *RBM12* are associated with psychosis. *Nat. Genet.* **49**, 1251–1254 (2017).
15. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
16. Howrigan, D. et al. Schizophrenia risk conferred by protein-coding de novo mutations. Preprint at *bioRxiv* https://doi.org/10.1101/495036 (2018).
17. Rees, E. et al. Analysis of intellectual disability copy number variants for association with schizophrenia. *JAMA Psychiatry* **73**, 963–969 (2016).
18. International Schizophrenia Consortium. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
19. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
20. Bergen, S. E. et al. Joint contributions of rare copy number variants and common SNPs to risk for schizophrenia. *Am. J. Psychiatry* **176**, 29–35 (2018).
21. Tansey, K. E. et al. Common alleles contribute to schizophrenia in CNV carriers. *Mol. Psychiatry* **21**, 1085–1089 (2015).
22. Niemi, M. E. K. et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).
23. Weiner, D. J. et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).
24. 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. Preprint at *bioRxiv* https://doi.org/10.1101/531210 (2019).
25. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/148353 (2017).
26. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
27. 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).
28. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
29. Keller, M. C. Evolutionary perspectives on genetic and environmental risk factors for psychiatric disorders. *Annu. Rev. Clin. Psychol.* **14**, 471–493 (2018).
30. Rees, E., Moskvina, V., Owen, M. J., O'Donovan, M. C. & Kirov, G. De novo rates and selection of schizophrenia-associated copy number variants. *Biol. Psychiatry* **70**, 1109–1114 (2011).
31. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
32. Pocklington, A. J. et al. Novel findings from CNVs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron* **86**, 1203–1214 (2015).
33. Gottesman, I. I. & Shields, J. A polygenic theory of schizophrenia. *Proc. Natl Acad. Sci. USA* **58**, 199–205 (1967).

GROUP Investigators

Behrooz Z. Alizadeh[6,7], Therese van Amelsvoort[8], Richard Bruggeman[6,9], Wiepke Cahn[10,11], Lieuwe de Haan[12,13], Jurjen J. Luykx[10,14], Bart P. F. Rutten[8], Jim van Os[10,15] and Ruud van Winkel[8,16]

[6]University of Groningen, University Medical Center Groningen, University Center for Psychiatry, Rob Giel Research Center, Groningen, The Netherlands. [7]Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands. [8]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, The Netherlands. [9]Department of Clinical and Developmental Neuropsychology, University of Groningen, Groningen, The Netherlands. [10]Department of Psychiatry, Brain Centre Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands. [11]Altrecht General Mental Health Care, Utrecht, The Netherlands. [12]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [13]Arkin Institute for Mental Health, Amsterdam, The Netherlands. [14]Department of Translational Neuroscience, University Medical Center Utrecht, Brain Center Rudolf Magnus, Utrecht, The Netherlands. [15]Department of Psychosis Studies, Institute of Psychiatry, King's College London, King's Health Partners, London, UK. [16]Research Group Psychiatry, Department of Neuroscience, KU Leuven, Leuven, Belgium.

Methods

Sample overview. Exome sequencing was carried out on 674 schizophrenia proband–parent trios, consisting of 2,000 individuals, on the Illumina HiSeq 4000 platform. The proband–parent trios were composed of 653 trios, 9 quads (two affected children) and one family with 3 affected children. None of these samples have been exome sequenced previously. The families were recruited by six independent groups (Supplementary Table 9), and were ascertained from general psychiatric wards or outpatient clinics. Proband–parent trios were ascertained blind to any genome analysis. Randomization of experimental groups was not applicable to this study (see Life Sciences Reporting Summary for more detail). All probands had received a DSM-IV (Diagnostic and Statistical Manual of Mental Disorders; fourth edition) or ICD-10 (International Statistical Classification of Diseases and Related Health Problems; 10th revision) diagnosis of schizophrenia or schizoaffective disorder. Individuals with a known diagnosis of intellectual disability or other neurodevelopmental disorder were not included. For proband samples that passed quality control (quality control procedure described below), information on family history of schizophrenia or psychosis was available for 552 trios; 66% of probands were recorded as having no family history of these disorders. Further details on the recruitment and diagnostic criteria for each cohort are provided in the Sample description section of the Supplementary Material.

Exome sequence generation. Exome sequence was generated using the Nextera DNA Exome capture kit, the HiSeq 3000/4000 PE Cluster Kit and the HiSeq 3000/4000 SBS Kit (both Illumina). Raw sequencing reads were processed according to the genome analysis toolkit (GATK) best practice guidelines[34,35]. Reads were aligned to the human reference genome (GRCh37) using the Burrow–Wheeler Aligner (bwa) v0.7.15 (ref. [36]). Variants were called using GATK haplotype caller (v3.4) and filtered using the GATK variant quality score recalibration tool, VQSR. For all samples that passed quality control (criteria outlined below), sequence data were generated for a median of 83% of the exome target at ≥10× coverage. Sequencing coverage is discussed further in the Supplementary Material. For future users of our new dataset, the median proportion to which each gene is covered at ≥10× coverage is provided in Supplementary Table 10.

Sample quality control. Trios ($n=27$) were excluded for low sequencing coverage if less than 70% of the exome target achieved ≥10× coverage in the proband or in either parent (Supplementary Fig. 1). An additional 27 trios were excluded for excess heterozygosity (heterozygote/homozygote ratio >1.9) or evidence of cross-sample contamination (as measured by the FREEMIX sequence-only estimate of contamination[37]) (Supplementary Fig. 2). The last two metrics are highly correlated. Identity-by-descent (IBD) analysis (PLINK v1.9) to ensure that proband–parent relationships were as expected resulted in exclusion of three trios. Four additional trios were excluded as outliers for the number of DNVs (Supplementary Fig. 3). Following implementation of all the above sample quality control steps, 613 proband–parent trios were retained for DNV analysis.

Variant quality control. In each of our newly sequenced samples, genotypes were excluded if they did not meet the following criteria: depth ≥10×; genotype quality score ≥30; allele balance ≤0.1 and ≥0.9 for homozygous genotypes for the reference and alternative allele, respectively; allele balance between 0.2 and 0.8 for heterozygous genotypes. For samples and variants that passed quality control, no difference was observed in the number of heterozygous variants that were transmitted or not transmitted from parents to probands (transmission disequilibrium test $P=0.53$), indicating high data quality.

De novo variant calling. Putative DNVs in the new trios were identified as sites that were heterozygous in the proband and homozygous for the reference allele in both parents. All trio members were required to pass genotype quality control as described above. We considered putative DNVs to be (1) those in which there were no reads for the mutant allele in either parent, and for which the mutant allele was not called in any other sample of the new trios (parent or proband) and (2) those in which the mutant allele count was ≤3 in all newly sequenced samples, there were no mutant allele variant reads in either parent, and that had at least 5 reads of the mutant allele in the proband. Read alignments for all putative DNVs were manually inspected using the Integrative Genomics Viewer (http://software.broadinstitute.org/software/igv) and variants were reassigned as high or low confidence if there was no evidence for read misalignment or evidence for read misalignment, respectively.

Sanger sequencing was used to perform a validation experiment, where DNA was available and primers could be designed, on all high-confidence LoF DNVs, as well as on additional putative DNVs. In total, primers were successfully designed for 205 putative DNVs. We observed high validation rates for high-confidence DNVs (95.5%) and low rates (3.4%) for low-confidence DNVs (Supplementary Table 11). Following these results, in our new trios all high-confidence DNVs ($n=606$ coding DNVs; Supplementary Table 12) were included in the downstream analyses.

Adding published de novo data. To increase the power of our analysis, DNVs from 2,831 schizophrenia trios that had been published previously were included. When combined with our new trios, this resulted in a sample size of 3,444 schizophrenia trios. No statistical methods were used to predetermine sample sizes but this trio sample is the largest reported to date and consists of all publicly available data from exome sequencing studies of DNVs in schizophrenia[15,16]. It is of note that no DNV from our new trios was also observed among the schizophrenia de novo data that were published previously, thus confirming the independence of our new trio dataset. A summary of the published data can be found in Supplementary Table 13.

Statistics. *De novo variant analysis.* To test whether DNVs were enriched in single genes or in sets of genes the statistical framework described in ref. [38] was used. Here, for a given set of genes, the number of DNVs expected in our new sample was estimated using per-gene mutation rates[39], adjusted for sequencing coverage. When the number of expected DNVs was estimated in the trios that were published previously, we did not adjust per-gene mutation rates for coverage as coverage metrics were not available for all samples; the use of unadjusted per-gene mutation would overestimate the expected number of DNVs in these trios, and produce more conservative enrichment results. For our gene set analysis, LoF-intolerant genes are defined as genes with a pLi score of ≥ 0.9, using pLi metrics generated from the non-psychiatric component of the ExAC dataset[31] (available from https://gnomad.broadinstitute.org/downloads#exac-variants). For single genes, a one-sided Poisson test (implemented in R) was used to test whether the overall burden of DNVs was significantly greater than that which was expected. For our primary de novo gene set analysis, background de novo rates were controlled for by using a two-sample Poisson rate ratio test, which compared the DNV enrichment observed for genes in the set to that in genes outside the set.

DNVs from both the new trios and from de novo data that were published previously were annotated using the Ensemble Variant Effect Predictor (v96) (ref. [40]). LoF variants are defined in this study as stop-gain, splice-acceptor, splice-donor and frameshift mutations. Although a small number of start-loss and stop-loss DNVs were observed, these were not included in our LoF annotation as mutation rates are not available for these variants. We classify missense-damaging variants as missense variants with an MPC score of ≥2, as this metric has proven to be effective in identifying variants that are associated with ASD[9,25]. Missense-damaging mutation rates for individual genes were calculated by summing the tri-nucleotide mutation probabilities for all sites with an MPC score of ≥2. In line with previous work by us and others[15,16], if an individual carried multiple DNVs in the same gene, these were conservatively considered to be the result of a single mutation event, and only the variant predicted to be most deleterious was retained for analysis .

Polygenic risk scores. Where trios were available ($n=1,122$), single nucleotide polymorphism (SNP) genotype data were used to generate polygenic risk scores. Genotype and exome sequence data were confirmed to belong to the same individual through IBD analysis (PLINK v1.9). A summary of the datasets for which both exome sequencing and SNP genotype data were available can be found in Supplementary Table 14. To derive PRS values for schizophrenia, bipolar disorder and height, we used the largest available GWAS summary statistics that were independent from our trio test data. Given the Bulgarian trio cohort overlapped with that used by the Psychiatric Genomics Consortium (PGC2), we computed schizophrenia PRS in these trios using custom PGC2 GWAS summary statistics that omitted the overlapping samples. The bipolar disorder PRS was used because previous studies have shown that common variant liability to schizophrenia and bipolar disorder is shared substantially[41]. The height PRS was used as a negative control. A summary of the training data that were used to generate PRS values can be found in Supplementary Table 14.

For quality control purposes, SNP genotype data were first harmonized to the Haplotype Reference Consortium panel using the Genotype Harmonizer package[42] and were then subjected to standard quality control, which included exclusion of samples with a call rate of <95%, SNPs with a minor allele frequency of <0.1, SNPs with >1% missingness, or SNPs with a Hardy–Weinberg equilibrium exact test P value of $<1\times10^{-6}$. PRS values were generated using PRSice 2 software[43], in which SNPs were clumped based on a window of 250 kb and a maximum r^2 of 0.2. PRS values were generated across a range of training data P value thresholds ($P<0.5$, 0.1, 0.05 and 0.001).

pTDT deviation. To test for overtransmission of polygenic risk, the pTDT was used as described in ref. [23] Here, pTDT deviation scores were generated for each trio by subtracting the mean parental PRS from the child PRS (equation 1). pTDT deviation scores were standardized by dividing them by the cohort-specific mean parental PRS s.d.

$$\text{pTDT deviation} = \frac{\text{PRS}_{\text{proband}} - \text{PRS}_{\text{parental mean}}}{\text{s.d.}\left(\text{PRS}_{\text{parental mean}}\right)} \quad (1)$$

To test whether the mean pTDT deviation was significantly greater than zero, which represents an overtransmission of polygenic risk, a one-sided one-sample t-test was used. A one-sided two-sample t-test was used to compare the mean pTDT deviation scores across groups of trios.

The primary pTDT results were produced using a PRS generated with a *P* threshold of 0.05, as this threshold explained the most case-control variance in the 2014 schizophrenia PGC analysis[4]. However, pTDT results obtained for PRS values that were generated across different *P* value thresholds are also presented in Supplementary Table 5 .

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

Data availability

DNVs discovered from the new trios are published in Supplementary Table 12. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

A description of the R functions used for statistical analysis can be found in the Life Sciences Reporting Summary.

References

34. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
35. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. 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).
38. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
39. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–7.25.15 (2015).
40. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
41. The Brainstorm Consortium et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
42. Deelen, P. et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
43. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31**, 1466–1468 (2015).

Acknowledgements

The work at Cardiff University was supported by Medical Research Council Centre Grant no. MR/L010305/1 (M.J.O.) and Program Grant no. G0800509 (M.J.O, M.C.O'D., J.T.R.W., V.E.-P., P.H. and A.J.P.), European Community Seventh Framework Programme Grant no. HEALTH-F2-2010-241909 (Project EU-GEI, M.C.O'D.), and European Union Seventh Framework Programme for research, technological development, and demonstration Grant no. 279227 (CRESTAR Consortium, M.C.O'D. and J.T.R.W.). We acknowledge L. Bates and L. Hopkins, at Cardiff University, for laboratory sample management. We acknowledge M. Einon, at Cardiff University, for support with the use and setup of computational infrastructures.

Author contributions

M.C.O'D., M.J.O., J.T.R.W., P.H. and E.R. conceived and designed the research. E.R. analyzed the data. J.H., J.M. and N.C. performed and managed the sequencing experiments. J.H. and M.D. performed the Sanger sequencing validation experiment. V.E.-P., A.J.P., L.H., S.E.L., A.F.P. and A.L.R. contributed to the interpretation of the results. T.L., N.K., V.K., V.G., M.P., J.G.-P., C.A., GROUP Investigators, M.G., J.R., G.K., J.T.R.W., M.C.O'D. and M.J.O. led the acquisition of the clinical samples. E.R., M.C.O'D. and M.J.O. wrote the manuscript, which was read, edited and approved by all authors.

Competing interests

M.C.O'D., M.J.O., P.H., J.T.R.W. and A.J.P. are supported by a collaborative research grant from Takeda. Takeda played no part in the conception, design, implementation, funding or interpretation of this study. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0565-2.

Correspondence and requests for materials should be addressed to M.C.O. or M.J.O.

Peer review information *Nature Neuroscience* thanks Ryan Yuen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Michael O'Donovan & Michael Owen

Last updated by author(s): 2019/10/23

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Raw sequencing data was produced on Illumina HiSeq 4000 instruments. Sequencing reads were processed according to GATK best practice guidelines. Reads were aligned to the human reference genome (GRCh37) using bwa version 0.7.15. Variants were called using GATK haplotype caller (v3.4) and filtered using the GATK Variant Quality Score Recalibration (VQSR) tool. Sequence quality metrics (such as coverage) were generated using PICARD tools (v 1.97). Contamination was estimated using VerifyBamID (https://genome.sph.umich.edu/wiki/VerifyBamID).
Data analysis	Statistical analyses were conducted in R software (v 3.2.4). Standard R packages were used to perform t tests (t.test), Poisson tests (ppois()) and two sample poisson rate tests (poisson.test()). Ensemble Variant Effect Predictor (version 96) was used to annotate variants. PRSice software (version 2) was used to generate polygenic risk scores. EIGENSOFT smartPCA (version 6.0.1) was used to perform a Principal Component Analysis.

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

Data

Policy information about availability of data

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

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

De novo mutations discovered from the new trios are published in Supplementary Table S12. The data that support the findings of this study are available from the corresponding author upon request.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No statistical methods were used to pre-determine sample sizes. Our schizophrenia trio sample is the largest reported to date, and consists of all publicly available data from exome-sequencing studies of de novo variants in schizophrenia.
Data exclusions	To minimise the false positive and false negative rate for de novo variants identified in our new schizophrenia proband-parent trio sample, we performed the following sample quality control: Samples were excluded if <70% of the exome target was covered at >= 10X; this exclusion criteria was based on that used in the UK10K exome sequencing study (Singh et al 2016). Samples were excluded if they had evidence of contamination (>5%) and/or excess heterozygosity (heterozygote:homozygote ratio > 1.9) and/or an excess of de novo variants (>10); these exclusion criteria were based on the distribution of these metrics in our dataset (see Supplementary Figure S2 and Supplementary Figure S3).
Replication	Results from our de novo enrichment analysis are presented using data from our new trios as well as in independent, previously published data sets. We also present results from a combined analysis of new trios and previously published trios. Findings from our gene-set analysis of LoF intolerant genes and neurodevelopmental disorder genes, and our finding for SLC6A1, were supported in both our new trios and the independent previously published trios.
Randomization	Randomization of experimental groups is not applicable to this study. Probands were allocated to the case group on the basis of having a DSM-IV or ICD-10 diagnosis of schizophrenia or schizoaffective disorder. To test the enrichment of de novo variants in a given gene, the observed de novos rate was compared to the expected de novo rate, which is based on known per-gene mutation rates, and thus does not require randomization.
Blinding	Proband-parent trios were ascertained blind to any genome analysis. Common variant polygenic scores were assigned blind to de novo mutation status.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	All probands included in the current study had received a DSM-IV or ICD-10 diagnosis of schizophrenia. As we used a withinfamily study design, our results are controlled for population and ascertainment biases. However, to support the robustness of our findings, we report results from our pTDT analysis separately for samples with European ancestry (994/1122 trios).
Recruitment	Bulgarian trios We sequenced 77 proband-parent trios recruited from Bulgaria whose ascertainment and diagnosis are as described previously (Kirov et al 2014). All cases had been hospitalised and met DSM-IV criteria for schizophrenia or schizoaffective disorder based upon SCAN (Schedules for Clinical Assessment in Neuropsychiatry) 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 proband-parent 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. 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)], 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) or SCAN interviews 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.

Ethics oversight

Research Ethics Committee for Wales

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