The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability

Tarjinder Singh[1], James T. R. Walters[2], Mandy Johnstone[3], David Curtis[4,5], Jaana Suvisaari[6], Minna Torniainen[6], Elliott Rees[2], Conrad Iyegbe[7], Douglas Blackwood[3], Andrew M. McIntosh[8], Georg Kirov[2], Daniel Geschwind[9], Robin M. Murray[7], Marta Di Forti[7], Elvira Bramon[10], Michael Gandal[9], Christina M. Hultman[11], Pamela Sklar[12], INTERVAL Study[13], UK10K Consortium[13], Aarno Palotie[14,15], Patrick F. Sullivan[16,17], Michael C. O'Donovan[2], Michael J. Owen[2], Jeffrey C. Barrett[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1HH, Cambridge, UK. [2]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK. [3]Division of Psychiatry, The University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK. [4]University College London (UCL) Genetics Institute, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK. [5]Centre for Psychiatry, Barts and the London School of Medicine and Dentistry, London, UK. [6]National Institute for Health and Welfare (THL), Helsinki FI-00271, Finland. [7]Institute of Psychiatry, King's College London, 16 De Crespigny Park, London SE5 8AF, UK. [8]Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ , UK. [9]UCLA David Geffen School of Medicine, Los Angeles, California 90095, USA. [10]Division of Psychiatry, University College London, Charles Bell House, Riding House Street, London W1W 7EJ, UK. [11]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-17177 Stockholm, Sweden. [12]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [13]The members of this consortium are listed in the Supplementary Note. [14]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki FI-00014 Finland. [15]Program in Medical and Population Genetics and Genetic Analysis Platform, The Broad Institute of MIT and Harvard, Cambridge MA 02132, USA. [16]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-17177 Stockholm, Sweden. [17]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, 27599-7264, USA.

Correspondence should be addressed to Tarjinder Singh (ts14@sanger.ac.uk) and Jeffrey C. Barrett (barrett@sanger.ac.uk).

Abstract

By meta-analyzing rare coding variants in whole-exome sequences of 4,133 schizophrenia cases and 9,274 controls, de novo mutations in 1,077 trios, and copy number variants from 6,882 cases and 11,255 controls, we show that individuals with schizophrenia carry a significant burden of rare damaging variants in 3,488 genes previously identified as having a near-complete

45   depletion of loss-of-function variants. In schizophrenia patients who also have
46   intellectual disability, this burden is concentrated in risk genes associated with
47   neurodevelopmental disorders. After excluding known neurodevelopmental
48   disorder risk genes, a significant rare variant burden persists in other loss-of-
49   function intolerant genes, and while this effect is notably stronger in
50   schizophrenia patients with intellectual disability, it is also seen in patients who
51   do not have intellectual disability. Together, our results show that rare damaging
52   variants contribute to the risk of schizophrenia both with and without
53   intellectual disability, and support an overlap of genetic risk between
54   schizophrenia and other neurodevelopmental disorders.

55

56   **Introduction**

57

58   Schizophrenia is a common and debilitating psychiatric illness
59   characterized by positive symptoms (hallucinations, delusions, disorganized
60   speech and behaviour), negative symptoms (social withdrawal and diminished
61   emotional expression), and cognitive impairment that result in social and
62   occupational dysfunction[1,2]. Operational diagnostic criteria for the disorder as
63   described in the DSM-V require the presence of at least two of the core
64   symptoms over a period of six months with at least one month of active
65   symptoms[3]. It is increasingly recognized that current categorical psychiatric
66   classifications have a number of shortcomings, in particular that they overlook
67   the increasing evidence for etiological and mechanistic overlap between
68   psychiatric disorders[4].

69

70   A diverse range of pathophysiological processes may contribute to the
71   clinical features of schizophrenia[5]. Indeed, previous studies have suggested a
72   number of hypotheses about schizophrenia pathogenesis, including abnormal
73   pre-synaptic dopaminergic activity[6], postsynaptic mechanisms involved in
74   synaptic plasticity[7], dysregulation of synaptic pruning[8], and disruption to early
75   brain development[9,10]. This complexity is underpinned by the varied nature of
76   genetic contributions to risk of schizophrenia. Genome-wide association studies
77   have identified over 100 independent loci defined by common (minor allele
78   frequency [MAF] > 1%) single nucleotide variants (SNVs)[11], and a recent analysis
79   determined that more than 71% of all one-megabase regions in the genome
80   contain at least one common risk allele[12]. The modest effects of these variants
81   (median odds ratio [OR] = 1.08) combine to produce a polygenic contribution
82   that explains only a fraction (h_g^2 = 0.274) of the overall liability[12]. In addition, a
83   number of rare variants have been identified that have far larger effects on
84   individual risk. These are best exemplified by eleven large, rare recurrent copy
85   number variants (CNVs) but evidence from whole-exome sequencing studies
86   implies that many other rare coding SNVs and *de novo* mutations also confer
87   substantial individual risk[13–17]. There is growing evidence that some of the same
88   genes and pathways are affected by both common and rare variants[7,18]. Pathway
89   analyses of common variants and hypothesis-driven gene set analyses of rare
90   variants have begun to enumerate some of these specific biological processes,
91   including histone methylation, transmission at glutamatergic synapses, calcium
92   channel signaling, synaptic plasticity, and translational regulation by the fragile X
93   mental retardation protein (FMRP)[11,13,14,19,20].

94
95   In addition to exploring the biological mechanisms underlying
96   schizophrenia, genetic analyses can also be used to understand its relationship to
97   other neuropsychiatric and neurodevelopmental disorders. For instance,
98   schizophrenia, bipolar disorder, and autism (ASD) show substantial sharing of
99   common risk variants[21,22]. Sequencing studies of neurodevelopmental disorders
100 suggest that this sharing of genetic risk may extend to rare variants of large
101 effect. In the largest sequencing study of ASD to date, 20 of the 46 genes and all
102 six CNVs implicated (false discovery rate [FDR] < 5%) had been previously
103 described as dominant causes of developmental disorders[23]. Furthermore, an
104 analysis of 60,706 whole exomes led by the ExAC consortium identified 3,230
105 genes with near-complete depletion of protein-truncating variants, and *de novo*
106 loss-of-function (LoF) mutations identified in individuals with ASD or
107 developmental disorders were concentrated in this set of "LoF intolerant"
108 genes[23–25]. Similarly, evidence from rare variants for a broader shared genetic
109 etiology between schizophrenia and neurodevelopmental disorders has begun to
110 emerge. Analyses of whole-exome data provided support for an enrichment of
111 schizophrenia rare variants in intellectual disability genes, and schizophrenia
112 cases were also found to have a higher concentration of ultra-rare disruptive
113 SNVs in the ExAC LoF intolerant genes compared to controls[13,17,26].
114
115   However, the contribution of these rare variants to risk in the wider
116 population of individuals diagnosed with schizophrenia, including those without
117 intellectual disability, remains unclear. Intriguingly, the 11 rare CNVs found to be
118 highly penetrant for schizophrenia also increased risk for intellectual disability
119 and other congenital defects[16,27], and more recently, a meta-analysis of whole-
120 exome sequence data showed that LoF variants in *SETD1A* conferred substantial
121 risk for both schizophrenia and neurodevelopmental disorders[18]. Concurrent
122 analyses of autism whole-exome data found that *de novo* loss-of-function (LoF)
123 mutations identified in ASD probands, particularly those that disrupt genes
124 associated with neurodevelopmental disorders, were disproportionately found
125 in individuals with intellectual disability[23,28]. These emerging results raise the
126 possibility that rare schizophrenia risk variants may be concentrated in a subset
127 of schizophrenia patients with co-morbid intellectual disability. Here, we present
128 the one of the largest accumulation of schizophrenia rare variant data to date,
129 which we jointly analyze with phenotype data on cognitive function. Using this
130 data set, we attempt to identify groups of genes disrupted by schizophrenia rare
131 risk variants, and determine if a subset of patients disproportionately carry
132 these damaging alleles.
133
134 **Results**
135
136 **Study design**
137
138   To maximize our power to detect enrichment of damaging variants in
139 schizophrenia cases in groups of genes, we performed a meta-analysis of three
140 different types of rare coding variant studies: (1) high-quality SNV calls from
141 whole-exome sequences of 4,133 schizophrenia cases and 9,274 matched
142 controls, (2) *de novo* mutations identified in 1,077 schizophrenia parent-proband

143   trios (Figure 1), and (3) CNV calls from genotyping array data of 6,882 cases and
144   11,255 controls. The ascertainment of these samples, data production, and
145   quality control were described previously[18,29]. All *de novo* mutations included in
146   our analysis had been validated through Sanger sequencing, and stringent
147   quality control steps were performed on the case-control data to ensure that
148   sample ancestry and batch were closely matched between cases and controls
149   (Online Methods).
150
151   For each data type, we used appropriate methods to test for an excess of
152   rare variants (Figure 1, Online Methods). In analyses of case-control SNV data,
153   we applied an extension of the variant threshold burden test that corrected for
154   exome-wide differences between cases and controls[30]. We tested all allele
155   frequency thresholds below 0.1% observed in our data, and assessed statistical
156   significance by permutation testing. In analyses of *de novo* SNV data, we
157   compared the observed number of *de novo* mutations to random samples from
158   an expected distribution based on a gene-specific mutation rate model to
159   calculate an empirical *P*-value. For both types of whole-exome sequencing data,
160   we restricted our analyses to loss-of-function variants. Finally, in analyses of
161   case-control CNV data, we used a logistic regression framework that compares
162   the rate of CNVs overlapping a specific gene set while correcting for differences
163   in CNV size and number of genes disrupted[7,19,31]. To ensure our model was well
164   calibrated, we restricted our analyses to small deletions and duplications
165   overlapping fewer than seven genes with MAF < 0.1% (Supplementary Figure 1,
166   Online Methods).
167
168   We tested for an excess of rare damaging variants in schizophrenia
169   patients in 1,766 gene sets (Online Methods, Supplementary Table 1, and
170   detailed results below). Gene set *P*-values were computed using the three
171   methods and variant definitions described above, and then meta-analyzed using
172   Fisher's Method to provide a single *P*-value for each gene set. Because we gave
173   each data type equal weight, gene sets achieving significance typically show at
174   least some signal in all three types of data. We observed a marked inflation in the
175   quantile-quantile (Q-Q) plot of gene set *P*-values (Supplementary Figure 2), so
176   we conducted two analyses to ensure our results were robust and not biased due
177   to methodological or technical artifacts. First, we observed no inflation of *P*-
178   values when testing for enrichment of synonymous variants in our case-control
179   and *de novo* analyses (Supplementary Figure 2). Second, we created random
180   gene sets by sampling uniformly across the genome, and observed null
181   distributions in Q-Q plots regardless of variant class and analytical method
182   (Supplementary Figure 3). These findings suggested that our methods
183   sufficiently corrected for known genome-wide differences in LoF and CNV
184   burden between cases and controls, and other technical confounders like batch
185   and ancestry.
186
187   **Rare, damaging schizophrenia variants are concentrated in LoF intolerant genes**
188
189   We first tested whether rare schizophrenia risk variants were
190   consistently concentrated in genes defined loss-of-function intolerant across
191   study design and variant type. Because some of our schizophrenia exome data

192   was included in the ExAC database, we focused on the subset of 45,376 ExAC
193   exomes without a known psychiatric diagnosis and that were not present in our
194   study. From this subset, 3,488 genes were found to have near-complete
195   depletion of such variants, which we defined as the LoF intolerant gene set. We
196   found that rare damaging variants in schizophrenia cases were enriched in LoF
197   intolerant genes (P < 3.6×10^{-10}, Table 1, Figure 2), with support in case-control
198   SNVs (P < 5×10^{-7}; OR 1.24, 1.16-1.31, 95% CI), case-control CNVs (P =
199   2.6×10^{-4}; OR 1.21, 1.15 – 1.28, 95% CI), and *de novo* mutations (P = 6.7×10^{-3};
200   OR 1.36, 1.1 – 1.68, 95% CI). While this result was consistent with observations
201   in intellectual disability and ASD[24,32] the absolute effect size is smaller (e.g. *de*
202   *novos*, Supplementary Figure 4 and 5). We observed no excess burden of rare
203   damaging variants in the remaining 14,753 genes (Figure 2, Supplementary
204   Figure 5). Furthermore, this signal was spread among many different LoF
205   intolerant genes: if we rank genes by decreasing significance, the enrichment
206   disappears in the case-control SNV analysis (P > 0.05) only after the exclusion of
207   the top 50 genes. This suggests that the contribution of damaging rare variants in
208   schizophrenia is not concentrated in just a handful of genes, but instead spread
209   across many genes.
210
211   **Schizophrenia risk genes are shared with other neurodevelopmental disorders**
212
213   Given the significant enrichment of rare damaging variants in LoF
214   intolerant genes in developmental disorders, autism and schizophrenia, we next
215   asked whether these variants affected the same genes. We found that autism
216   risk genes identified from exome sequencing meta-analyses[23] and genes in which
217   LoF variants are known causes of severe developmental disorders as defined by
218   the DDD study[33,34] were significantly enriched for rare variants in individuals
219   with schizophrenia (P_{ASD} = 9.5×10^{-6}; P_{DD} = 2.3×10^{-6}; Table 1, Online Methods).
220   Previous analyses have shown an enrichment of rare damaging variants in genes
221   whose mRNA are bound by FMRP in both schizophrenia and autism[35,13,32], so we
222   sought to identify further shared biology by testing targets of neural regulatory
223   genes previously implicated in autism[32,36]. We observed enrichment of both
224   such sets: promoter targets of *CHD8* (P = 1.1×10^{-6}) and splice targets of *RBFOX*
225   (P = 1.3×10^{-5}) (Table 1). We noted that some published gene lists attributed to
226   same biological process differed due to choices of assay, cell type, method of
227   sample extraction, and threshold of statistical significance, leading to distinct
228   results in our gene set analyses. For example, we observed a significant
229   enrichment in the published FMRP binding gene set based on mouse brain
230   data[37], but with no signal in one based on a human kidney cell line[38].
231
232   We also tested an additional 1,759 gene sets from databases of biological
233   pathways with at least 100 genes, as we lacked power to detect weak
234   enrichments in smaller sets (Online Methods). We observed enrichment of
235   damaging rare variants in schizophrenia cases at FDR q < 0.05 in 35 of these
236   gene sets (Supplementary Table 1, 2). These included previously implicated gene
237   sets, like the NMDA receptor and ARC complexes[13,14,35,37], as well as novel gene
238   sets, such as genes involved in cytoskeleton (GO: 0007010), chromatin
239   modification (GO:0016568), and chromatin organization (GO: 0006325).
240   Furthermore, the gene sets most significantly enriched (FDR q < 0.01) for

241   schizophrenia rare variants (Table 1) had all been previously linked to autism,
242   intellectual disability, and severe developmental disorders[23,32,33]. Our
243   enrichment results matched some of the findings from a pathway analysis of
244   common risk variants in psychiatric disorders, which also implicated neuronal
245   and chromatin gene sets[20]. However, unlike that study, we found no enrichment
246   of rare variants in immune-related gene sets.
247
248   We noticed that the 1,759 gene sets we tested were collectively enriched
249   with LoF intolerant genes when compared to a random sampling of genes from
250   the genome (Supplementary Figure 6 and 7). For some of the gene sets
251   associated with schizophrenia, this over-representation was quite substantial:
252   67% of the gene targets of FMRP and 74% of the genes associated with severe
253   neurodevelopmental disorders are LoF intolerant. To better understand the
254   consequences of this overlap on our results, we extended the gene set
255   enrichment methods (Online Methods) to condition on LoF intolerance and
256   brain-expression for the 35 gene sets with FDR $q < 0.05$ in the previous analysis
257   (Supplementary Table 2). We first observed that 22 of the 35 gene sets remained
258   significant even after conditioning on brain expression (Supplementary Tables 3,
259   Online Methods), suggesting they represent more specific biological processes
260   involved in schizophrenia. However, only known autism risk genes (P =
261   4.4×10^{-4}) and neurodevelopmental disorder genes ($P = 3 \times 10^{-5}$) had an excess
262   of rare coding variants above the enrichment already observed in LoF intolerant
263   genes (Supplementary Table 3). Thus, in addition to biological pathways
264   implicated specifically in schizophrenia, at least a portion of the schizophrenia
265   risk conferred by rare variants of large effect is shared with childhood onset
266   disorders of neurodevelopment.
267
268   **Schizophrenia patients with intellectual disability have a greater burden of rare**
269   **damaging variants**
270
271   In autism spectrum disorders, the observed excess of rare damaging
272   variants has been shown to be greater in individuals with intellectual disability
273   than those with normal levels of cognitive function[28]. We observed a similar
274   phenomenon in schizophrenia cases carrying *SETD1A* LoF variants[18], so next
275   sought to explore whether this pattern is consistent in gene sets implicated in
276   schizophrenia. We acquired relevant cognitive phenotype data for 2,971 of the
277   4,131 schizophrenia patients with whole-exome sequencing data
278   (Supplementary Figure 8). Of these individuals, 279 were clinically diagnosed
279   with intellectual disability in addition to fulfilling the full diagnostic criteria for
280   schizophrenia (SCZ-ID subgroup, Online Methods). We also identified 1,165
281   individuals for whom we could rule out cognitive impairment (by excluding pre-
282   morbid IQ < 85, fewer than 12 years of schooling or lowest decile of composite
283   cognitive measures, depending on available data, Online Methods). Finally, we
284   identified 1,527 individuals who were not diagnosed with intellectual disability,
285   but in whom some cognitive impairment could not be excluded.
286
287   When stratifying into these three groups (intellectual disability, no
288   intellectual disability but cognitive impairment not excluded, no cognitive
289   impairment), we observed that the burden of rare damaging variants in LoF

290   intolerant genes was significantly greater in the SCZ-ID subgroup than in the
291   remaining schizophrenia cases (P = 2.6×10^{-4}; OR 1.3, 1.12– 1.51, 95% CI) or
292   controls (P < 5×10^{-7}; OR 1.61, 1.37 – 1.89, 95% CI; Figure 3). In the LoF
293   intolerant gene set, 0.27 (0.2 – 0.35, 95% CI) extra singleton (defined as having
294   an allele count of one in our data set) LoF variants were observed per exome in
295   SCZ-ID cases compared to controls, while 0.10 (0.065 – 0.13, 95% CI) extra
296   singleton LoF variants per exome were observed in the remaining schizophrenia
297   cases compared to controls (Online Methods). Furthermore, SCZ-ID individuals
298   had significant enrichment of rare LoF variants in developmental disorder genes
299   compared to the other cases (P = 9×10^{-4}; OR 2.36, 1.41– 3.92, 95% CI) or to
300   controls (P = 9.5×10^{-6}; OR 3.43, 2.01– 5.86, 95% CI; Figure 4). Compared to
301   controls, the SCZ-ID individuals carried 0.045 (0.03 – 0.06, 95% CI) extra
302   singleton LoF variants in developmental disorder genes per exome, suggesting
303   that around 4% of these cases had a LoF variant that is relevant to their clinical
304   presentation. No enrichment in neurodevelopmental disorder genes was
305   observed in schizophrenia patients without intellectual disability, suggesting
306   that these genes were relevant only for that subset of schizophrenia patients
307   (Figure 4, Supplementary Table 4). Notably, even after excluding known
308   developmental disorder genes from the set of LoF intolerant genes, we still
309   observed an enrichment of rare variants in SCZ-ID patients compared to the
310   remaining cases (P = 1×10^{-3}; 1.26, 1.08 – 1.47, 95% CI) or to controls (P
311   < 5×10^{-7}; OR 1.54, 1.31– 1.81, 95% CI; Supplementary Figure 9). Rare variation
312   in these genes contributes more to disease risk in the subset of patients with
313   both schizophrenia and intellectual disability.
314
315   **Rare variants confer risk for schizophrenia in individuals without intellectual**
316   **disability**
317
318   While rare damaging variants in LoF intolerant genes were most enriched
319   in the subset of schizophrenia patients with intellectual disability, we still
320   observed a weaker but significant enrichment in individuals with schizophrenia
321   for whom we could confirm do not have intellectual disability (P = 5.5×10^{-4};
322   1.16, 1.05 – 1.27, 95% CI; Figure 3). Therefore, rare risk variants for
323   schizophrenia follow the pattern previously described in autism: concentrated in
324   individuals with intellectual disability, but not exclusive to that group. To
325   produce a more accurate estimate of the effect of damaging rare variants on
326   schizophrenia conditional on their effects on overall cognition, we recalculated
327   the enrichment of rare variants in LoF intolerant genes in a subset of 2,161
328   schizophrenia cases and 2,398 controls for which data on years of education was
329   available and for whom intellectual disability could be excluded (Supplementary
330   Figure 8). After controlling for differences in educational attainment (Online
331   Methods), individuals with schizophrenia have a 1.26-fold excess of rare variants
332   in LoF intolerant genes (P = 2×10^{-6}; 1.14 – 1.38, 95% CI). This increase in our
333   observed odds ratio is consistent with previous accounts that rare damaging
334   variants also affect educational attainment in controls[39], thus biasing our
335   unconditional estimate.

336   **Discussion**
337

338   Our integrated analysis of thousands of whole-exome sequences
339 demonstrates that rare damaging variants increase risk of schizophrenia both
340 with and without co-morbid intellectual disability. While the identification of
341 individual genes remains difficult at current samples sizes, we show that the
342 burden of damaging *de novo* mutations, rare SNVs and CNVs in schizophrenia is
343 not scattered across the genome but is primarily concentrated in 3,488 genes
344 intolerant of loss-of-function variants. This observation is shared with autism,
345 intellectual disability, and severe neurodevelopmental disorders[32,40]. We
346 recapitulate enrichment in previously published gene sets, including
347 transmission at glutamatergic synapses and translational regulation by FMRP,
348 and implicate other gene sets previously linked to autism, intellectual disability,
349 and severe developmental disorders. However, we find that all of these gene sets
350 share a large number of underlying genes, and are especially enriched with the
351 3,488 genes intolerant of LoF variants. These overlaps among gene sets
352 originating from very different analyses, as well as the subtleties of how they are
353 defined, suggest caution in interpreting biological explanations from observed
354 enrichments.
355
356   We jointly analyzed the case-control SNV data with information on
357 cognitive function for 2,971 patients, and find that LoF variants disrupting genes
358 associated with severe developmental disorders are disproportionately found in
359 individuals with schizophrenia with co-morbid intellectual disability, with 4% of
360 these cases having a single LoF variant that is relevant to their clinical
361 presentation. Even after excluding variants in known developmental disorder
362 genes, rare variants contribute a greater degree to schizophrenia risk in the SCZ-
363 ID subgroup of patients than the remaining schizophrenia population. These
364 results show that some of these genetic perturbations have clear manifestations
365 in childhood, and that rare risk variants in schizophrenia are particularly
366 associated with co-morbid intellectual disability. Our observations are consistent
367 with results in autism in which rare risk variants are associated with intellectual
368 disability[22,23,28]. Notably, a weaker but still significant rare variant burden was
369 observed in schizophrenia patients without cognitive impairment, and this signal
370 persists even after controlling for educational attainment. Together, these results
371 demonstrate that rare variants have different contributions to schizophrenia risk
372 depending on the degree of cognitive impairment. Importantly, they do not
373 simply confer risk for a small subset of patients but contribute to disease
374 pathogenesis more broadly.
375
376   Our study supports the observation that genetic risk factors for
377 psychiatric and neurodevelopmental disorders do not follow clear diagnostic
378 boundaries. Coding variants disrupting the same genes, and quite possibly, the
379 same biological processes, increase risk for a range of phenotypic manifestation.
380 This clinically variable presentation is reminiscent of LoF variants in *SETD1A*
381 and 11 large copy number variant syndromes, previously shown to confer risk
382 for schizophrenia in addition to other prominent developmental defects[16,18]. It is
383 possible that these genes contain an allelic series of variants conferring
384 gradations of risk. A recent schizophrenia GWAS meta-analysis demonstrated
385 that the common variant association signal was similarly enriched in LoF
386 intolerant genes[41], suggesting that schizophrenia risk genes may be perturbed by

387   common variants of subtle effects and disrupted by rare variants of high
388   penetrance in the population. This possibility is also supported by the overlap in
389   at least some of the pathways affected by both rare and common variation, such
390   as chromatin remodeling. However, the most common deletion in the 22q11.2
391   locus and a recurrent two base deletion in *SETD1A* are associated with both
392   schizophrenia and more severe neurodevelopmental disorders, suggesting the
393   same variants can also confer risk for a range of clinical features[18,42,43].
394   Ultimately, it may prove difficult to clearly partition patients genetically into
395   subtypes with similar clinical features, especially if genes and variants
396   previously thought to cause well-characterized Mendelian disorders can have
397   such varied outcomes. This pattern is consistent with the hypothesis that LoF
398   variants in genes under genic constraint result in a spectrum of
399   neurodevelopmental outcomes with the burden of mutations highest in
400   intellectual disability and least in schizophrenia, corresponding to a gradient of
401   neurodevelopmental pathology indexed by the degree of cognitive impairment,
402   age of onset, and severity[4].
403
404   Despite the complex nature of genetic contributions to risk of
405   schizophrenia, it is notable that across study design (trio or case-control) and
406   variant class (SNVs or CNVs), risk loci of large effect are concentrated in a small
407   subset of genes. Previous rare variant analyses in other neurodevelopmental
408   disorders, such as autism, have successfully integrated information across *de*
409   *novo* SNVs and CNVs to identify novel risk loci[23]. As sample sizes increase, meta-
410   analyses leveraging the shared genetic risk across study designs and variant
411   types, including those we did not consider here, such as classical recessive
412   inheritance, will be similarly well powered to identify additional risk genes in
413   schizophrenia.
414

415   **Acknowledgements**
416

Author contributions

T.S., J.C.B conceived and designed the experiments.
T.S performed the statistical analysis.
T.S., J.T.R.W., M.J., D.C., J.S., M.T., E.R., P.F.S analysed the data.
T.S., J.T.R.W., M.J., J.S., M.T., E.R., C.I., D.B., A.M.M., G.K., D.G., R.M.M., M.D.F., E.B., M.G., C.M.H., P.S., A.P., M.C.O., M.J.O., J.C.B contributed reagents/materials/analysis tools.
T.S., D.C., M.J.O., J.C.B wrote the paper

Competing financial interests statement

We have no competing financial interests to declare.

References

1.   van Os, J. & Kapur, S. Schizophrenia. *Lancet* **374,** 635–45 (2009).

484   2.   American Psychiatric Association. *Diagnostic and statistical manual of*
485   *mental disorders (DSM-5{®})*. (American Psychiatric Publishing, 2013).
486   3.   Tandon, R. *et al.* Definition and description of schizophrenia in the DSM-5.
487   *Schizophr. Res.* **150,** 3–10 (2013).
488   4.   Owen, M. J. New approaches to psychiatric diagnostic classification.
489   *Neuron* **84,** 564–571 (2014).
490   5.   Owen, M. J., Sawa, A. & Mortensen, P. B. Schizophrenia. *Lancet* **6736,** 1–12
491   (2016).
492   6.   Howes, O. D. & Kapur, S. The dopamine hypothesis of schizophrenia:
493   version III--the final common pathway. *Schizophr. Bull.* **35,** 549–62 (2009).
494   7.   Pocklington, A. J. *et al.* Novel Findings from CNVs Implicate Inhibitory and
495   Excitatory Signaling Complexes in Schizophrenia. *Neuron* **86,** 1203–1214
496   (2015).
497   8.   Sekar, A. *et al.* Schizophrenia risk from complex variation of complement
498   component 4. *Nature* **530,** 177–183 (2016).
499   9.   Owen, M. J., O'Donovan, M. C., Thapar, A. & Craddock, N.
500   Neurodevelopmental hypothesis of schizophrenia. *Br. J. Psychiatry* **198,**
501   173–5 (2011).
502 10.   Rapoport, J. L., Giedd, J. N. & Gogtay, N. Neurodevelopmental model of
503   schizophrenia: update 2012. *Mol. Psychiatry* **17,** 1228–38 (2012).
504 11.   Schizophrenia Working Group of the Psychiatric Genomics Consortium.
505   Biological insights from 108 schizophrenia-associated genetic loci. *Nature*
506   **511,** 421–7 (2014).
507 12.   Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and
508   other complex diseases using fast variance-components analysis. *Nat.*
509   *Genet.* **47,** 1385–1392 (2015).
510 13.   Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic
511   networks. *Nature* **506,** 179–184 (2014).
512 14.   Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of
513   postsynaptic signalling complexes in the pathogenesis of schizophrenia.
514   *Mol. Psychiatry* **17,** 142–53 (2012).
515 15.   The International Schizophrenia Consortium. Rare chromosomal deletions
516   and duplications increase risk of schizophrenia. *Nature* **455,** 237–41
517   (2008).
518 16.   Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-
519   associated loci. *Br. J. Psychiatry* **204,** 108–14 (2014).
520 17.   Zhu, X., Need, A. C., Petrovski, S. & Goldstein, D. B. One gene, many
521   neuropsychiatric disorders: lessons from Mendelian diseases. *Nat.*
522   *Neurosci.* **17,** 773–781 (2014).
523 18.   Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated
524   with schizophrenia and developmental disorders. *Nat. Neurosci.* **19,** 571–
525   577 (2016).
526 19.   Szatkiewicz, J. P. *et al.* Copy number variation in schizophrenia in Sweden.
527   *Mol. Psychiatry* **19,** 762–773 (2014).
528 20.   Psychiatric Genetics Consortium. Psychiatric genome-wide association
529   study analyses implicate neuronal, immune and histone pathways. *Nat.*
530   *Neurosci.* (2015). doi:10.1038/nn.3922
531 21.   Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders
532   estimated from genome-wide SNPs. *Nat. Genet.* **45,** 984–94 (2013).

533 22. Robinson, E. B. *et al.* Genetic risk for autism spectrum disorders and
534   neuropsychiatric variation in the general population. *Nat. Genet.* **48,** 552–
535   555 (2016).
536 23. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic
537   Architecture and Biology from 71 Risk Loci. *Neuron* **87,** 1215–1233
538   (2015).
539 24. Samocha, K. E. *et al.* A framework for the interpretation of de novo
540   mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
541 25. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706
542   humans. *Nature* **536,** 285–291 (2016).
543 26. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants
544   among 4,877 individuals with schizophrenia. *Nat. Neurosci.* (2016).
545   doi:10.1038/nn.4402
546 27. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia
547   and developmental delay. *Biol. Psychiatry* **75,** 378–85 (2014).
548 28. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism
549   spectrum disorder. *Nature* **515,** 216–21 (2014).
550 29. Rees, E. *et al.* CNV analysis in a large schizophrenia sample implicates
551   deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1.
552   *Hum. Mol. Genet.* **23,** 1669–76 (2014).
553 30. Price, A. L. *et al.* Pooled Association Tests for Rare Variants in Exon-
554   Resequencing Studies. *Am. J. Hum. Genet.* **86,** 832–838 (2010).
555 31. Raychaudhuri, S. *et al.* Accurately assessing the risk of schizophrenia
556   conferred by rare copy-number variation affecting genes with brain
557   function. *PLoS Genet.* **6,** (2010).
558 32. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted
559   in autism. *Nature* **515,** 209–15 (2014).
560 33. Firth, H. V *et al.* DECIPHER: Database of Chromosomal Imbalance and
561   Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84,**
562   524–33 (2009).
563 34. Deciphering Developmental Disorders Study. Prevalence and architecture
564   of de novo mutations in developmental disorders. *Nature* **542,** 433–438
565   (2017).
566 35. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in
567   schizophrenia. *Nature* **506,** 185–90 (2014).
568 36. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates
569   other autism risk genes during human neurodevelopment. *Nat. Commun.*
570   **6,** 6404 (2015).
571 37. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to
572   synaptic function and autism. *Cell* **146,** 247–61 (2011).
573 38. Ascano, M. *et al.* FMRP targets distinct mRNA sequence elements to
574   regulate protein expression. *Nature* **492,** 382–386 (2012).
575 39. Ganna, A. *et al.* Ultra-rare disruptive and damaging mutations influence
576   educational attainment in the general population. *Nat. Neurosci.* **19,** 1563–
577   1565 (2016).
578 40. The Deciphering Developmental Disorders Study. Large-scale discovery of
579   novel genetic causes of developmental disorders. *Nature* **519,** 223–8
580   (2015).
581 41. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in

582   mutation-intolerant genes and maintained by background selection.
583   *bioRxiv* 68593 (2016). doi:10.1101/068593
584 42. Ben-Shachar, S. *et al.* 22q11.2 Distal Deletion: A Recurrent Genomic
585   Disorder Distinct from DiGeorge Syndrome and Velocardiofacial
586   Syndrome. *Am. J. Hum. Genet.* **82,** 214–221 (2008).
587 43. Michaelovsky, E. *et al.* Genotype-phenotype correlation in 22q11.2
588   deletion syndrome. *BMC Med. Genet.* **13,** 122 (2012).

589 **Figure captions**
590
591 **Figure 1:** Analysis workflow. Data sets are shown in blue, statistical methods
592 and analysis steps are shown in green, and results (figures and tables) from the
593 analysis are shown in orange. **A:** Enrichment analyses in 1,766 gene sets using
594 the entire rare variant data set. **B:** Enrichment analyses in LoF intolerant and
595 developmental disorder genes in the subset of cases with information on
596 cognitive function. ID: intellectual disability; SCZ: schizophrenia; SCZ-ID:
597 schizophrenia patients with intellectual disability.

598 **Figure 2:** Enrichment of schizophrenia rare variants in genes intolerant of loss-
599 of-function variants. **A:** Schizophrenia cases compared to controls for rare SNVs
600 and indels; **B:** Rates of *de novo* mutations in schizophrenia probands compared
601 to control probands; **C:** Case-control CNVs. *P*-values shown were from the test of
602 LoF enrichment in **A**, LoF enrichment in **B**, and all CNVs enrichment in **C**. Error
603 bars represent the 95% CI of the point estimate. LoF intolerant: 3,448 genes with
604 near-complete depletion of truncating variants in the ExAC database; Rest: the
605 remaining genes in the genome with pLI < 0.9; Damaging missense: missense
606 variants with CADD phred > 15. Asterisk: $P < 1 \times 10^{-3}$.
607
608 **Figure 3:** Enrichment of rare loss-of-function variants in LoF intolerant genes in
609 schizophrenia cases stratified by information on cognitive function compared to
610 controls. The *P*-values shown were calculated using the variant threshold
611 method comparing LoF burden between the corresponding cases and controls.
612 Error bars represent the 95% CI of the point estimate. Damaging missense:
613 missense variants with CADD phred > 15.
614
615 **Figure 4:** Enrichment of rare loss-of-function variants in known severe
616 developmental disorder genes in schizophrenia cases stratified by information
617 on cognitive function compared to controls. The *P*-values shown were calculated
618 using the variant threshold method comparing LoF burden between the
619 corresponding cases and controls. Error bars represent the 95% CI of the point
620 estimate. Damaging missense: missense variants with CADD phred > 15.
621

Name	N$_{genes}$	Est$_{SNV}$	95% CI of Est$_{SNV}$	P$_{SNV}$	Est$_{DNM}$	95% CI of Est$_{DNM}$	P$_{DNM}$	Est$_{CNV}$	95% CI of Est$_{CNV}$	P$_{CNV}$	P$_{meta}$	Q$_{meta}$
ExAC LoF intolerant genes (pLI > 0.9)	3488	1.24	1.16-1.31	$< 5.0 \times 10^{-7}$	1.36	1.1-1.68	0.0067	1.21	1.15-1.28	0.00026	$< 3.60 \times 10^{-10}$	4.30×10^{-7}
Dominant, diagnostic DDG2P genes, in which LoF variants result in developmental disorders with brain abnormalities	156	1.42	1.07-1.88	0.011	4.18	2.21-8.03	0.00073	1.92	1.54-2.39	0.0016	2.30×10^{-6}	0.00067
Sanders *et al.* autism risk genes (FDR < 10%)	66	1.28	0.97-1.69	0.0095	3.96	1.65-9.94	0.019	2.21	1.75-2.79	0.00033	9.50×10^{-6}	0.0017
Darnell *et al.* targets of FMRP	790	1.24	1.13-1.36	8.5×10^{-6}	1.31	0.83-2.09	0.17	1.32	1.2-1.47	0.0032	9.30×10^{-7}	0.00038
Cotney *et al.* CHD8-targeted promoters (hNSC and human brain tissue)	2920	1.09	1.02-1.16	0.0008	1.77	1.36-2.31	0.00025	1.11	1.05-1.18	0.027	1.10×10^{-6}	0.00038
G2CDB: mouse cortex post-synaptic density consensus	1527	1.20	1.11-1.3	2.5×10^{-6}	1.57	1.06-2.33	0.028	1.04	0.96-1.11	0.32	3.90×10^{-6}	0.00097
Weynvanhentenryck *et al.* CLIP targets of RBFOX	967	1.21	1.11-1.33	4.8×10^{-5}	1.84	1.21-2.8	0.0085	1.07	0.98-1.17	0.2	1.30×10^{-5}	0.002
NMDAR network (defined in Purcell *et al.*)	61	1.66	1.09-2.54	0.0061	5.60	2.06-16.09	0.017	2.46	1.78-3.4	0.0028	3.70×10^{-5}	0.0044
GOBP: chromatin modification (GO:0016568)	519	1.29	1.13-1.49	0.00018	2.26	1.32-3.94	0.0099	1.12	0.99-1.28	0.18	4.20×10^{-5}	0.0046

622 **Table 1**: Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR < 1%. The effect sizes and corresponding
623 *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) are shown for each gene set, along
624 with the Fisher's combined P-value (P$_{meta}$) and the FDR-corrected Q-value (Q$_{meta}$). We only show the most significant gene set if there are
625 multiple ones from the same data set or biological process (see Supplementary Table 1 for all 1,766 gene sets). N$_{genes}$: number of genes
626 in the gene set; Est: effect size estimate and its lower and upper bound assuming a 95% CI; DNM: *de novo* mutation.

Supplementary Table captions

Supplementary Table 1: Full results from enrichment analyses of 1,766 gene sets. The *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) are shown for each gene set, along with the Fisher's combined P-value (P_{meta}) and the FDR-corrected Q-value (Q_{meta}). N_{genes}: number of genes in the gene set; SNV: single nucleotide variants from whole-exome data; DNM: *de novo* mutations.

Supplementary Table 2: Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR < 5%. The effect sizes and corresponding *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) are shown for each gene set, along with the Fisher's combined P-value (P_{meta}) and the FDR-corrected Q-value (Q_{meta}). N_{genes}: number of genes in the gene set; Est: effect size estimate and its lower and upper bound assuming a 95% CI; SNV: single nucleotide variants from whole-exome data; DNM: *de novo* mutations.

Supplementary Table 3: Results from enrichment analyses of FDR < 5% gene sets, conditional on brain-expressed and ExAC LoF intolerant genes. We restrict enrichment analyses to genes that reside in two different background gene sets, one defined on brain-enriched expression in GTeX, and the second on genic constraint (ExAC LoF intolerant genes), and determined if gene sets with FDR < 5% in the meta-analysis still had significance above the specific background. The *P*-values from enrichment tests of each variant type (case-control SNVs, DNM, and case-control CNVs) are shown for each gene set, along with the Fisher's combined P-value (P_{meta}). SNV: single nucleotide variants from whole-exome data; DNM: *de novo* mutations

Supplementary Table 4: Results from enrichment analyses of rare loss-of-function variants in LoF intolerant genes and developmental disorder genes comparing schizophrenia cases stratified by information on cognitive function and matched controls. Each comparison is defined in the Table, and the *P*-values shown were calculated using the variant threshold method comparing LoF burden between the corresponding case and baseline samples. N_{case}: number of case samples; $N_{comparison}$: number of comparison samples; Estimates: effect size estimate and its lower and upper bound assuming a 95% CI.

Online Methods

Sample collections

The ascertainment, data production, and quality control of the schizophrenia case-control whole-exome sequencing data set had been described in detail in an earlier publication[18]. Briefly, the data set was composed of schizophrenia cases recruited as part of eight collections in the UK10K sequencing project, and matched population controls from non-psychiatric arms of the UK10K project, healthy blood donors from the INTERVAL project, and five

673   Finnish population studies. The UK10K data set was combined and analyzed
674   with published data from a Swedish schizophrenia case-control study[35]. The data
675   production, quality control, and analysis of the case-control CNV data set was
676   described in an earlier publication[29]. The schizophrenia cases were recruited as
677   part of the CLOZUK and CardiffCOGS studies, which consisted of both
678   schizophrenia individuals taking the antipsychotic clozapine and a general
679   sample of cases from the UK. Matched controls were selected from four publicly
680   available non-psychiatric data sets. All samples were genotyped using Illumina
681   arrays, and processed and called under the same protocol. Sanger-validated *de*
682   *novo* mutations identified through whole exome-sequencing in seven published
683   studies of schizophrenia parent-proband trios were aggregated and re-annotated
684   for enrichment analyses[13,44–49]. A full description of each trio study, including
685   sequencing and capture technology and sample recruitment was previously
686   described[18].

687 **Sample and variant quality control**
688
689   We jointly called each case data set with its nationality-matched controls,
690   and excluded samples based on contamination, coverage, non-European
691   ancestry, and excess relatedness[18]. A number of empirically derived filters were
692   applied at the variant and genotype level, including filters on GATK VQSR,
693   genotype quality, read depth, allele balance, missingness, and Hardy-Weinberg
694   disequilibrium[18]. After variant filtering, the per-sample transition-to-
695   transversion ratio was ~3.2 across the entire data set, as expected for
696   populations of European ancestry[50]. For the case-control CNV analysis, we
697   similarly excluded samples based on excess relatedness, and only CNVs
698   supported by more than 10 probes and greater than 10 kilobases in size were
699   retained to ensure high quality calls. All *de novo* mutations in our study had been
700   validated using Sanger sequencing.
701
702   We used the Ensembl Variant Effect Predictor (VEP) version 75 to
703   annotate all variants (SNVs and CNVs) according to Gencode v.19 coding
704   transcripts. We defined frameshift, stop gained, splice acceptor, and donor
705   variants as loss-of-function (LoF), and missense or initiator codon variants with
706   the recommended CADD Phred score cut-off of greater than 15 as damaging
707   missense[51]. A gene was annotated as disrupted by a deletion if part of its coding
708   sequence overlapped the copy number event. We more conservatively defined
709   genes as duplicated only if the entire canonical transcript of the gene overlapped
710   with the duplication event.
711
712   Statistical tests of the case-control exome data used case-control
713   permutations within each population (UK, Finnish, Swedish) to generate
714   empirical *P*-values to test hypotheses. No genome-wide inflation was observed in
715   burden tests of individual genes[18]. In the curated set of *de novo* mutations, we
716   observed the expected exome-wide number of synonymous mutations given
717   gene mutation rates from previously validated models[24], suggesting variant
718   calling was generally unbiased across Gencode v.19 coding genes. Lastly, the
719   case-control CNV data set had been previously analyzed for burden of CNVs
720   affecting individual genes, and enrichment analyses in targeted gene sets[7,29].

Rare variant gene set enrichment analyses

Case-control enrichment burden tests For the case-control SNV data set, we
performed permutation-based gene set enrichment tests using an extension of
the variant threshold method[30]. This method assumed that variants with a MAF
below an unknown threshold T were more likely to be damaging than variants
with a MAF above T, and this threshold was allowed to differ for every gene or
pathway tested. To consider different possible values for threshold T, a gene or
gene set test statistic $t(T)$ was calculated for every allowable T, and the
maximum test-statistic, or $t_{\max}$, was selected. The statistical significance of $t_{\max}$
was evaluated by permuting phenotypic labels, and calculating $t_{\max}$ from the
permuted data such that different values of T could be selected following each
permutation. In Price *et al.*, $t(T)$ was defined as the z-score calculated from
regressing the phenotype on the sum of the allele counts of variants in a gene
with MAF $< T$. We extended this method to test for enrichment in gene sets by
regressing schizophrenia status on the total number of damaging alleles in the
gene set of interest with MAF $< T$ ($X_{in,T}$) while correcting for the total number of
damaging alleles genome-wide with MAF $< T$ ($X_{all,T}$). $X_{all,T}$ controlled for
exome-wide differences between schizophrenia cases and controls, ensuring any
significant gene set result was significant beyond baseline differences. $t(T)$ was
defined as the t-statistic testing if the regression coefficient of $X_{in,T}$ deviated
from 0. We then calculated $t(T)$ for all observed thresholds below a minor allele
frequency of 0.1%, and selected the maximum value for the $t_{\max}$ based on the
observed data. To calculate a null distribution for $t_{\max}$, we performed two
million case-control permutations within each population (UK, Finnish, and
Swedish) to control for batch and ancestry, and calculated $t_{\max}$ for each
permuted sample while allowing T to vary. The P-value for each gene set was
calculated as the fraction of the two million permuted samples that had a greater
$t_{\max}$ than what was observed in the unpermuted data. The odds ratio and 95%
confidence interval of each gene set was calculated using a logistic regression
model, regressing schizophrenia status on X_{in} while controlling for total number
of variants genome-wide (X_{all}) and population (UK, Finnish, and Swedish).
Unlike gene set P-values which were calculated using permutation across
multiple frequency thresholds, the odds ratios an d 95% CI were calculated using
only variants observed once in our data set (allele count of 1) to ensure they
were comparable between tested gene sets.

CNV logistic regression We adapted a logistic regression framework described in
Raychaudhuri *et al.* and implemented in PLINK to compare the case-control
differences in the rate of CNVs overlapping a specific gene set while correcting
for differences in CNV size and total genes disrupted[7,19,31]. We first restricted our
analyses to coding deletions and duplications, and tested for enrichment using
the following model:

$$\log\left(\frac{P_{i,\text{case}}}{1-P_{i,\text{case}}}\right) = \beta_0 + \beta_1 s_i + \beta_2 g_{\text{all}} + \beta_3 g_{\text{in}} + \epsilon,$$

where for individual i, p_i is the probability they have schizophrenia i, s_i is the
total length of CNVs, g_{all} is the total number of genes overlapping CNVs, and g_{in} is
the number of genes within the gene set of interest overlapping CNVs. It has been
shown that β_1 and β_2 sufficiently controlled for the genome-wide differences in

767 the rate and size of CNVs between cases and control, while β_3 captured the true
768 gene set enrichment above this background rate[7,19,31]. For each gene set, we
769 reported the one-sided *P*-value, odds ratio, and 95% confidence interval of β_3.

770 **Weighted permutation-based sampling of *de novo* mutations** For each variant
771 class of interest, we first determined the total number of *de novo* mutations
772 observed in the 1,077 schizophrenia trios. We then generated 2 million random
773 samples with the same number of *de novo* mutations, weighting the probability
774 of observing a mutation in a gene by its estimated mutation rate. The baseline
775 gene-specific mutation rates were obtained using the method described in
776 Samocha *et al.* and adapted to produce LoF and damaging missense rates for
777 each Gencode v.19 gene. These mutation rates adjusted for both sequence
778 context and gene length, and were successfully applied in the primary analyses
779 of large-scale exome sequencing of autism and severe developmental disorders
780 with replicable results[23,32,40]. For each gene set, one-sided enrichment *P*-values
781 were calculated as the fraction of two million random samples that had a greater
782 or equal number of *de novo* mutations in the gene set of interest than what is
783 observed in the 1,077 trios. The effect size of the enrichment was calculated as
784 the ratio between the number of observed mutations in the gene set of interest
785 and the average number of mutations in the gene set across the two million
786 random samples. We adapted a method in Fromer *et al*. to calculate 95% credible
787 intervals for the enrichment statistic[13]. We first generated a list of one thousand
788 evenly spaced values between 0 and ten times the point estimate of the
789 enrichment. For each value, the mutation rates of genes in the gene set of
790 interest were multiplied by that amount, and 50,000 random samples of *de novo*
791 mutations were generated using these weighted rates. The probability of
792 observing the number of mutations in the gene set of interest given each effect
793 size multiplier was calculated as the fraction of samples in which the number of
794 mutations in the gene set is the same as the observed number in the 1,077 trios.
795 We normalized the probabilities across the 1,000 values to generate a posterior
796 distribution of the effect size, and calculated the 95% credible interval using this
797 empirical distribution.
798
799 **Combined joint analysis** Gene set *P*-values calculated using the case-control SNV,
800 case-control CNV, and *de novo* data were meta-analyzed using Fisher's combined
801 probability method with *df* = 6 to provide a single test statistic for each gene set.
802 We corrected for the number of gene sets tested in the discovery analysis (n =
803 1,776) by controlling the false discovery rate (FDR) using the Benjamini-
804 Hochberg approach, and reported only results with a *q*-value of less than 5%.
805
806 **Description of gene sets**
807
808 The full list of tested gene sets is found in Supplementary Table 1, and a
809 detailed description is provided in the Supplementary Note. Briefly, we tested all
810 gene sets with more than 100 genes from five public pathway databases. We
811 additionally tested additional gene sets selected based on biological hypotheses
812 about schizophrenia risk, and genome-wide screens investigating rare variants
813 in intellectual disability, autism spectrum disorders, and other
814 neurodevelopmental disorders. All gene identifiers were mapped to the

815 GENCODE v.19 release, and all non-coding genes were excluded. A total of 1,766
816 gene sets were included in our analysis.

Selection of allele frequency thresholds and consequence severity

819   For the case-control whole-exome data, we applied an extension of the
820 variant threshold model (described above). With this method, we tested
821 damaging variants at a number of frequency thresholds without specifying an *a*
822 *priori* MAF cut-off. All thresholds below a MAF of 0.1% observed in our data
823 were tested, and we assessed statistical significance by permutation testing. For
824 all the whole-exome data (case-control and trio data), we restricted our analyses
825 to loss-of-function variants. These variants have a clear and severe predicted
826 functional consequence in that they putatively cause a single-copy loss of a gene.
827 Furthermore, this class of variants had been demonstrated to have the strongest
828 genome-wide enrichment between cases and controls across
829 neurodevelopmental and psychiatric disorders[18,32,40]. When selecting MAF cut-
830 offs for case-control CNVs, we found that while the bulk of the test statistics were
831 not inflated, the tail of gene set *P*-values were dramatically inflated even when
832 testing for enrichment in the random gene sets (Supplementary Figure 1). This
833 inflation in the tail of the Q-Q plot was driven in part by very large (overlapping
834 more than 10 genes), more common (MAF between 0.1% and 1%) CNVs
835 observed mainly in cases or controls. Some of these, such as the known
836 syndromic CNVs, likely harbored true risk genes. However, because these CNVs
837 were highly recurrent in cases and depleted in controls, and disrupted a large
838 number of genes, any gene set that included even a single gene within these
839 CNVs would appear to be significant, even after controlling for total CNV length
840 and genes overlapped. To ensure our model was well calibrated and its *P*-values
841 followed a null distribution for random gene sets, we explored different
842 frequency and size thresholds, and conservatively restricted our analysis to copy
843 number events overlapping less than seven genes (excluding the largest 10% of
844 CNVs) with MAF < 0.1% (Supplementary Figure 1). Our main conclusions
845 remained unchanged even if we selected a more stringent (excluding the largest
846 15% of CNVs) or less stringent (excluding the largest 5% of CNVs) size threshold.

Robustness of enrichment analyses

850   We uniformly sampled genes from the genome (as defined by Gencode
851 v.19) to generate random gene sets with the same size distribution as the 1,776
852 gene sets in our discovery analysis. For each random set, we calculated gene set
853 *P*-values for the case-control SNV data, case-control CNV data, and *de novo* data
854 using the appropriate method and frequency cut-offs across all variant classes. A
855 Q-Q plot was generated using *P*-values from enrichment tests of each data set
856 and variant type. Reassuringly, we observed null distributions in all such Q-Q
857 plots (Supplementary Figure 3).

Comparison of *de novo* enrichment with broader neurodevelopmental disorders

862   We aggregated and re-annotated *de novo* mutations from four studies:
863 1,113 severe DD probands[40], 4,038 ASD probands[23,32], and 2,134 control
864 probands[28,32]. We used the Poisson exact test to calculate differences in *de novo*
865 rates in constrained genes between schizophrenia, ASD, and DD and controls.
866 Counts in each functional class (synonymous, missense, damaging missense, and
867 LoF) were tested separately, and the one-sided *P*-value, rate ratio, and 95% CI of
868 each comparison were reported and plotted in Figure 2, Supplementary Figure 4
869 and 5.
870
871 **Conditional analyses**
872
873   In each of the three methods we used for gene set enrichment, we
874 restricted all variants analyzed to those that reside in the background gene list,
875 and tested for an excess of rare variants in genes shared between the gene set of
876 interest (K) and the background list (B). Brain-enriched genes from GTEx, and
877 the ExAC LoF intolerant genes (pLI > 0.9) were used as backgrounds (see above).
878 For the case-control SNV data, we modified the variant threshold method to
879 regress schizophrenia status on the total number of damaging alleles in genes
880 present in both the gene set of interest and the background gene set ($K \cap B$),
881 while correcting for the total number of damaging alleles in the set of all
882 background genes (B). The logistic regression model for the case-control CNV
883 data was modified to:

884 $$\log\left(\frac{P_{i,\text{case}}}{1-P_{i,\text{case}}}\right) = \beta_0 + \beta_1 s_i + \beta_2 g_B + \beta_3 g_{K \cap B} + \epsilon,$$

885 where g_B is the total number of background genes overlapping a CNV, and $g_{K \cap B}$ is
886 the number of genes in the intersection of the gene set of interest and the
887 background list overlapping a CNV. Finally, we determined the total number of
888 *de novo* mutations within the background gene list observed in the 1,077
889 schizophrenia trios, and generated 2 million random samples with the same
890 number of *de novo* mutations. For each gene set, one-sided enrichment *P*-values
891 were calculated as the fraction of two million random samples that had a greater
892 or equal number of *de novo* mutations in genes in $K \cap B$ than what is observed in
893 the 1,077 trios. Gene set *P*-values were combined using Fisher's method. We
894 restricted our conditional enrichment analysis to gene sets with *q*-value < 5% in
895 the discovery analysis, and adjusted for multiple testing using Bonferroni
896 correction (P = 0.00071, or 0.05/67 tests; see Supplementary Table 3).
897

898 **Rare variants and cognition in schizophrenia**

899   Within the UK10K study, 97 individuals from the MUIR collection were
900 given discharge diagnoses of mild learning disability and schizophrenia (ICD-8
901 and -9). The recruitment guidelines of the MUIR collection were described in
902 detail in a previous publication[52]. In brief, evidence of remedial education was a
903 prerequisite to inclusion, and individuals with pre-morbid IQs below 50 or above
904 70, severe learning disabilities, or were unable to give consent were excluded.
905 The Schizophrenia and Affective Disorders Schedule-Lifetime version (SADS-L)
906 in people with mild learning disability, PANSS, RDC, and DSM-III-R, and St. Louis
907 Criterion were applied to all individuals to ensure that any diagnosis of

908   schizophrenia was robust. Using the clinical information provided alongside the
909   Swedish and Finnish case-control data sets, we identified additional 182
910   schizophrenia individuals who were similarly diagnosed with intellectual
911   disability, for a total of 279 individuals.

912   Cognitive testing and educational attainment data available for a subset of
913   samples were used identify schizophrenia individuals without cognitive
914   impairment. For 502 individuals from the Cardiff collection in the UK10K study,
915   we acquired their pre-morbid IQ as extrapolated from National Adult Reading
916   Test (NART), and identified 412 individuals for analysis after excluding all
917   individuals with predicted pre-morbid IQ of less than 85 (or below one standard
918   deviation of the population distribution for IQ). We additionally acquired
919   information on educational attainment in 54 schizophrenia individuals in the
920   UK10K London collection, and retained 27 individuals without intellectual
921   disability and who completed at least 12 years of schooling. Lastly, the California
922   Verbal Learning Test was conducted on 124 Finnish schizophrenia individuals
923   sequenced as part of UK10K, and a composite score was generated from
924   measures of verbal and visual working memory, verbal abilities,
925   visuoconstructive abilities, and processing speed. All individuals with intellectual
926   disability had been excluded from cognitive testing. Within this set of samples,
927   we additionally excluded any individuals who ranked in the lowest decile in
928   CVLT composite score, and retained 92 individuals for analysis. According to
929   these criteria, we identified 531 of 697 schizophrenia individuals from the UK
930   and Finnish data sets with cognitive data as not having intellectual disability. We
931   additionally acquired data on educational attainment for the Swedish
932   schizophrenia cases and controls from the Swedish National Registry. After
933   excluding individuals with intellectual disability, we identified 1,527
934   schizophrenia individuals who did not complete secondary school (less than 12
935   years of schooling), and 634 schizophrenia individuals who completed at least
936   compulsory and upper secondary schooling (at least 12 years of schooling). The
937   last group with the greatest educational attainment and without intellectual
938   disability was defined as cases without cognitive impairment. In the Swedish
939   sample, 49.4% of control samples had lower educational attainment than the
940   634 individuals with schizophrenia defined as having no cognitive impairment,
941   suggesting that our definition was sufficiently strict. In total, combining the UK,
942   Finnish, and Swedish data, we identified 1,165 schizophrenia individuals without
943   cognitive impairment.

944   Using the variant threshold method, we tested for differences in rare LoF
945   burden between the three case groups (intellectual disability, did not complete
946   secondary school, no cognitive impairment) against controls. We restricted these
947   analyses to three gene sets (LoF intolerant genes, genes in which LoF variants
948   are diagnostic for severe developmental disorders, and LoF intolerant genes
949   after excluding severe developmental disorders genes), and adjusted for multiple
950   testing using Bonferroni correction (P = 0.0038, or 0.05/13 tests).
951   Supplementary Table 4 enumerated all the statistical tests performed. To
952   estimate the per-exome excess of rare singleton (defined as having an allele
953   count of one in our data set) LoF variants in cases compared to controls, we
954   regressed X_{in} (the number of LoF variants in the gene set of interest) on case
955   status (0 or 1) while controlling for X_{all} (the total number of LoF variants

956 genome-wide) and population (UK, Finnish, and Swedish). The effect size and
957 95% CI of the regression coefficient of case status predictor were reported.

958 **Data Availability**
959
960 Sequence data and processed VCFs for the UK10K project were deposited into
961 the European Genome-phenome Archive (EGA) under study accession code
962 EGAO00000000079. The processed VCFs from the Swedish case-control study
963 were deposited in dbGAP under accession code (phs000473.v1.p1). Rare variant
964 counts, and gene-level association results from combining the whole-exome
965 sequencing data sets were described in a previous publication[18] and were made
966 available on the PGC results and download page
967 (https://www.med.unc.edu/pgc/results-and-downloads).
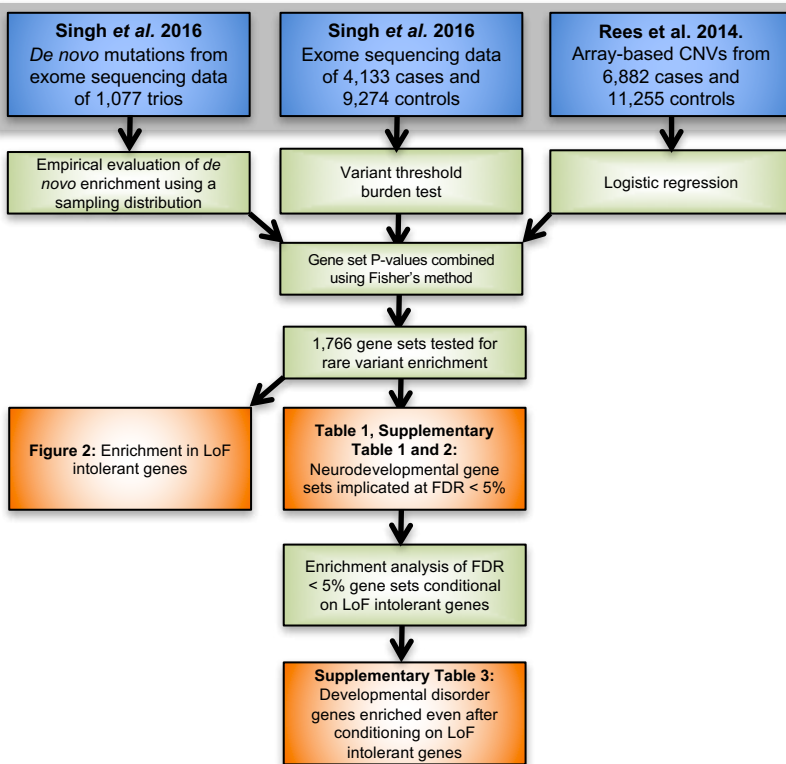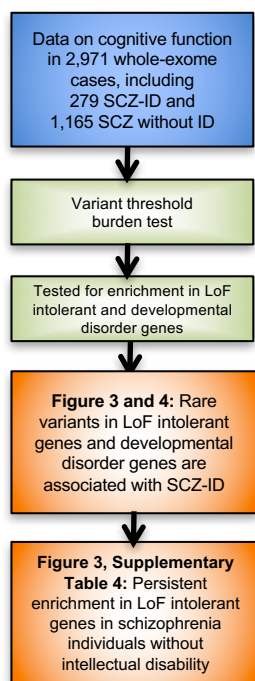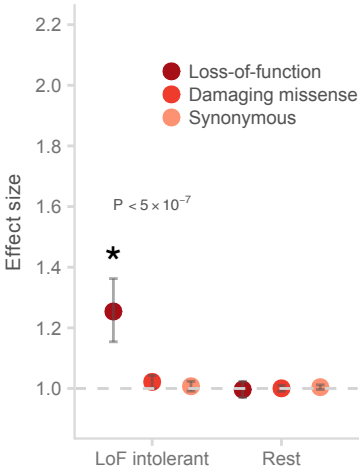968

969 **References for Online Methods**
970

971 44.   Guipponi, M. *et al.* Exome sequencing in 53 sporadic cases of schizophrenia
972   identifies 18 putative candidate genes. *PLoS One* **9,** e112745 (2014).
973 45.   Girard, S. L. *et al.* Increased exonic de novo mutation rate in individuals
974   with schizophrenia. *Nat. Genet.* **43,** 860–3 (2011).
975 46.   McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate
976   chromatin remodeling and support a genetic overlap with autism and
977   intellectual disability. *Mol. Psychiatry* **19,** 652–8 (2014).
978 47.   Takata, A. *et al.* Loss-of-function variants in schizophrenia risk and
979   SETD1A as a candidate susceptibility gene. *Neuron* **82,** 773–80 (2014).
980 48.   Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for
981   schizophrenia. *Nat. Genet.* **43,** 864–8 (2011).
982 49.   Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and
983   neural complexity in schizophrenia. *Nat. Genet.* **44,** 1365–9 (2012).
984 50.   Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles
985   conferring risk for myocardial infarction. *Nature* **518,** 102–106 (2014).
986 51.   Kircher, M. *et al.* A general framework for estimating the relative
987   pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–5 (2014).
988 52.   Doody, G. A., Johnstone, E. C., Sanderson, T. L., Owens, D. G. & Muir, W. J.
989   'Pfropfschizophrenie' revisited. Schizophrenia in people with mild learning
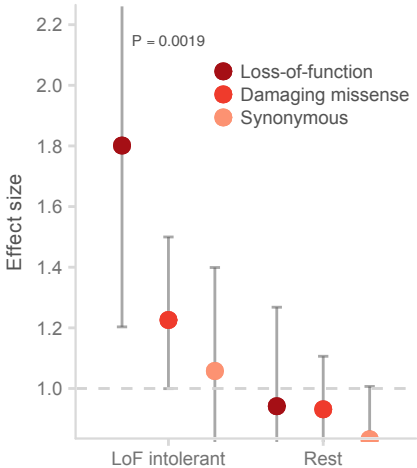990   disability. *Br. J. Psychiatry* **173,** 145–153 (1998).

991
992
993

A

Singh _et al._ 2016 _De novo_ mutations from exome sequencing data of 1,077 trios	**Singh _et al._ 2016** Exome sequencing data of 4,133 cases and 9,274 controls	**Rees et al. 2014.** Array-based CNVs from 6,882 cases and 11,255 controls

Empirical evaluation of _de novo_ enrichment using a sampling distribution

Variant threshold burden test

Logistic regression

Gene set P-values combined using Fisher's method

1,766 gene sets tested for rare variant enrichment

Figure 2: Enrichment in LoF intolerant genes

Table 1, Supplementary Table 1 and 2: Neurodevelopmental gene sets implicated at FDR < 5%

Enrichment analysis of FDR < 5% gene sets conditional on LoF intolerant genes

Supplementary Table 3: Developmental disorder genes enriched even after conditioning on LoF intolerant genes

B

Data on cognitive function in 2,971 whole-exome cases, including 279 SCZ-ID and 1,165 SCZ without ID

Variant threshold burden test

Tested for enrichment in LoF intolerant and developmental disorder genes

Figure 3 and 4: Rare variants in LoF intolerant genes and developmental disorder genes are associated with SCZ-ID

Figure 3, Supplementary Table 4: Persistent enrichment in LoF intolerant genes in schizophrenia individuals without intellectual disability

$P < 5 \times 10^{-7}$

Legend:
- Loss-of-function
- Damaging missense
- Synonymous

Y-axis: Effect size

X-axis: LoF intolerant, Rest
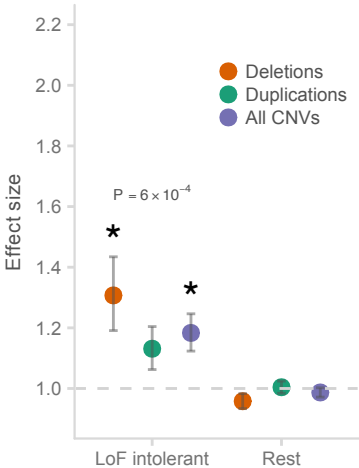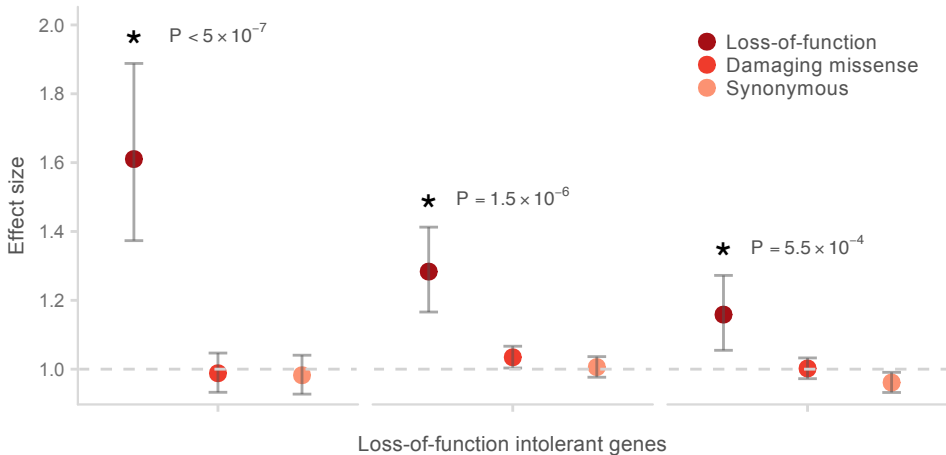
Schizophrenia individuals with intellectual disability v. controls

Schizophrenia individuals who did not complete secondary school v. controls

Schizophrenia individuals without intellectual disability v. controls

Loss-of-function
Damaging missense
Synonymous

$P < 5 \times 10^{-7}$

$P = 1.5 \times 10^{-6}$

$P = 5.5 \times 10^{-4}$

Effect size

Loss-of-function intolerant genes

Schizophrenia individuals with intellectual disability v. controls

Schizophrenia individuals who did not complete secondary school v. controls

Schizophrenia individuals without intellectual disability v. controls

Effect size

$P = 9.5 \times 10^{-6}$

- Loss-of-function
- Damaging missense
- Synonymous

Risk genes for neurodevelopmental disorders