nature neuroscience

Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia

Giulio Genovese[1–3], Menachem Fromer[4,5], Eli A Stahl[4,5], Douglas M Ruderfer[4,5], Kimberly Chambert[1], Mikael Landén[6], Jennifer L Moran[1], Shaun M Purcell[4,5], Pamela Sklar[4,5], Patrick F Sullivan[7,8], Christina M Hultman[8] & Steven A McCarroll[1–3]

By analyzing the exomes of 12,332 unrelated Swedish individuals, including 4,877 individuals affected with schizophrenia, in ways informed by exome sequences from 45,376 other individuals, we identified 244,246 coding-sequence and splice-site ultra-rare variants (URVs) that were unique to individual Swedes. We found that gene-disruptive and putatively protein-damaging URVs (but not synonymous URVs) were more abundant among individuals with schizophrenia than among controls ($P = 1.3 \times 10^{-10}$). This elevation of protein-compromising URVs was several times larger than an analogously elevated rate for *de novo* mutations, suggesting that most rare-variant effects on schizophrenia risk are inherited. Among individuals with schizophrenia, the elevated frequency of protein-compromising URVs was concentrated in brain-expressed genes, particularly in neuronally expressed genes; most of this elevation arose from large sets of genes whose RNAs have been found to interact with synaptically localized proteins. Our results suggest that synaptic dysfunction may mediate a large fraction of strong, individually rare genetic influences on schizophrenia risk.

Schizophrenia is a psychiatric disorder with a lifetime risk of about 0.7%[1] and a heritability of 60–80%[2,3] despite greatly reduced reproductive fecundity[4,5]. Because individuals affected with schizophrenia have fewer offspring, purifying selection is expected to prevent high-risk alleles from reaching even modest allele frequencies[6]. Indeed, estimates of selection (when based only on the reproductive costs of schizophrenia) may underestimate the actual selective pressure against such alleles given emerging evidence that such alleles have multiple adverse effects: for example, rare copy number variations (CNVs), with penetrances ranging from 2–30% (the latter observed for 22q11.2 deletions), negatively affect cognition and fecundity even in their more-typical presentation without schizophrenia[7]. To the extent that such observations condition expectations for rare single-nucleotide variants, variants with a large effect on schizophrenia risk are likely to be rare in populations, requiring sequencing to find them.

Distinguishing those variants that are extremely rare from variants that are segregating in a population is ideally informed by sequencing very-large numbers of individuals from the same population. We therefore analyzed the sequences of 12,332 unrelated individuals (4,946 affected with schizophrenia, 6,242 unaffected controls and 1,144 with other psychiatric illnesses whose analysis is beyond the scope of the current study) from Sweden (Online Methods). We further informed this analysis with a much larger set of exome sequencing data from 45,376 individuals from multiple non-psychiatric cohorts ascertained by the Exome Aggregation Consortium[8]. This made it possible to identify among the Swedish research participants 244,246 coding-sequence and splice-site URVs that were present in single individuals, a set of variants that is greatly enriched for recent mutations, relative to the vastly larger fraction of heterozygosity that is a result of less-rare variants (**Fig. 1a**). This large set of variants made it possible to identify broad biological patterns among an excess of more than 1,000 protein-damaging URVs that we found in the exomes of 4,877 individuals affected with schizophrenia.

RESULTS

Exome-wide enrichment of URVs

We analyzed the protein-coding sequences (the exomes) of 12,332 unrelated Swedish individuals, including 4,946 affected individuals with schizophrenia (2,951 males and 1,995 females), 6,242 unaffected controls (3,182 males and 3,060 females) and 1,144 individuals affected with other disorders (443 males and 701 females, used for population genetic analyses, but not as cases or controls). After removing 119 individuals for quality control reasons (mostly because of divergent ancestry, Online Methods), we identified 244,246 coding-sequence and splice-site URVs (among 4,877 schizophrenia cases and

[1]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [4]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [5]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [6]Department of Psychiatry and Neurochemistry, Institute of Neuroscience a Physiology at Sahlgrenska Academy at University of Gothenburg, Göteborg, Sweden. [7]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina, USA. [8]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to G.G. (giulio.genovese@gmail.com) or S.A.M. (mccarroll@genetics.med.harvard.edu).

Received 27 May; accepted 6 September; published online 3 October 2016; doi:10.1038/nn.4402

6,203 controls) that were present in only 1 of the 12,332 unrelated Swedish exomes analyzed and never seen in the Exome Aggregation Consortium (ExAC) cohort (which numbered 45,376 individuals after excluding the subjects from this cohort and other subjects ascertained for psychiatric disorders).

We focused on URVs in most analyses because such variants, although comprising a tiny fraction (less than 0.2%) of the heterozygous sites in an individual, will be greatly enriched for recent mutations, and will therefore have been exposed to fewer generations of purifying selection. The size of the Swedish cohort analyzed, and the additional sequence data from ExAC, allowed us to greatly refine the identification of URVs; for example, among 5,092 (of the 12,332) individuals who were also part of an earlier sequencing study[9], the additional data allowed us to re-classify ~66% of variants that had been 'singletons' as segregating variants (not URVs in the current analysis). This may have been particularly helpful for refining analyses of challenging-to-interpret missense variants, as we describe below.

We classified coding-sequence and splice-site variants into four groups (**Fig. 1b**): synonymous, which were exonic variants not predicted to change the encoded protein (63,230 URVs); missense non-damaging, which were missense variants not predicted to damage protein function (by the criteria below) (134,100 URVs); damaging, which were missense variants predicted (by an algorithm) to compromise protein function (Online Methods), in-frame indels or variants affecting protein-protein-contact domains (27,390 URVs); and disruptive, which were variants that truncated or abrogated the encoded protein in a way that was readily classified as loss of function[10] or as triggering nonsense-mediated decay (NMD)[11], including nonsense, frame shift, splice-site and, very rarely, read-through variants. (19,526 URVs).

The terms protein-damaging and gene-disruptive refer to predicted effects on individual gene copies and the encoded proteins, rather than to effects on phenotypes; effects on phenotypes can be inferred only from association analysis.

Missense damaging URVs accounted for approximately 15% of all missense URVs (**Supplementary Fig. 1**). There was a median of two disruptive and two damaging URVs per individual (four total) (**Supplementary Fig. 2**).

To assess whether schizophrenia was associated with an increased number of coding-sequence and splice-site URVs (in specific genes, across the exome, or in sets of genes), we used linear regression to control for possible confounding variables, including each individual's overall number of detected URVs (including non-coding URVs), sex, birth year, the hybrid selection kit used for exome enrichment, and the first 20 principal components estimated from exome-wide single-nucleotide polymorphism (SNP) and indel genotypes (**Supplementary Table 1**).

An important negative control, to address the possibility that analyses could be affected by population structure, differences in average relatedness within the case and control groups, or by technical variation, was to ask whether functionally neutral forms of variation showed any apparent differences in frequency between case and control groups. We did not observe a significant difference ($P = 0.81$) in the rate of synonymous URVs between schizophrenia cases and controls (**Fig. 1c**). We also did not observe a significant difference ($P = 0.55$) for non-coding URVs (**Supplementary Fig. 3a**).

In contrast, we observed significant case-control differences in the rates of disruptive URVs (a difference of 0.12 variants per person; 95% confidence interval (CI) = 0.07–0.17; $P = 1.8 \times 10^{-6}$) and damaging URVs (0.12 variants per person; 95% CI = 0.07–0.18; $P = 3.4 \times 10^{-5}$). P values determined by permuting the phenotype data 10 million

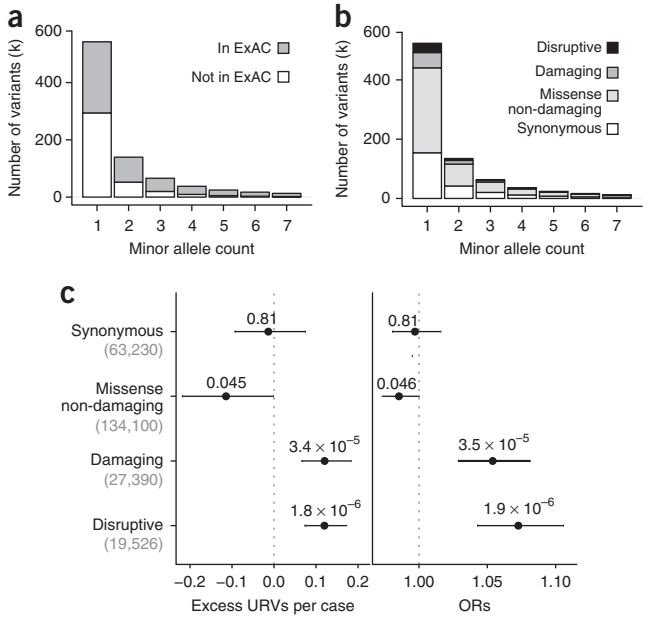

Figure 1 URV distribution and association with schizophrenia. (**a**,**b**) Counts across coding-sequence and splice-site rare variants stratified by minor allele count across exome-sequencing data from 12,332 individuals indicating how many variants were observed in the ExAC cohort (**a**) and how many variants were classified as disruptive, damaging, missense non-damaging and synonymous (**b**). (**c**) Observed enrichment in schizophrenia cases compared with controls for coding-sequence and splice-site URVs across the main four annotation types. Enrichment and P values were computed using a linear regression model (left) and a logistic regression model (right). Horizontal bars indicate 95% confidence intervals.

times agreed with P values from the linear regression analysis ($P = 0.81$ for synonymous, $P = 0.045$ for missense non-damaging, $P = 3.5 \times 10^{-5}$ for damaging, $P = 1.8 \times 10^{-6}$ for disruptive); this suggests that the P values from the regression model are well-calibrated. Damaging and disruptive URVs showed similarly elevated frequencies in cases and were therefore combined into a single category termed dURVs (disruptive and damaging ultra-rare variants) for subsequent analyses.

Adjusting for covariates, there were 7% more dURVs in affected individuals than in controls (odds ratios (OR) = 1.07; 95% CI = 1.05–1.09; $P = 1.5 \times 10^{-10}$), as the case-associated elevation in dURVs (of about 0.25 variants per patient; 95% CI = 0.17–0.32) occurred on a background of about four dURVs per patient. The elevated frequency of dURVs among individuals affected with schizophrenia appeared to arise from multiple types of dURVs, including in-frame indels, protein-protein-contact, splice-acceptor, splice-donor, stop-gained and frame-shift variants (**Supplementary Fig. 3b**).

To assure that this result was not the result of population stratification in Sweden, we further estimated the enrichment in a more genetically homogeneous subset of the Swedish cohort (3,554 schizophrenia cases and 5,164 controls) that excluded individuals with significant amounts of Finnish or Northern Sweden ancestry. Individuals with schizophrenia showed a similar dURV excess in this more genetically homogeneous group (excess of 0.25 dURVs per case, 95% CI = 0.16–0.34; $P = 2.2 \times 10^{-8}$).

We next estimated the extent to which dURVs tend to be inherited or *de novo*. Although parental DNA would be necessary to directly ascertain which specific dURVs are *de novo* mutations (DNMs), we were able to compare the schizophrenia-associated elevation

in dURVs (~0.25 per exome) to an analogous elevation in DNMs detected in earlier studies of 617 affected and 1,911 unaffected father-mother-offspring trios[12,13]. Using data from the trios, we estimated the frequencies of DNMs that were protein damaging or protein disruptive (dDNMs) by the same criteria that we used to identify dURVs (including restricting to variants not previously observed in ExAC)[8]. These data yielded an elevation of about 0.03 such DNMs per exome, based on the difference between rates of 0.185 (95% CI = 0.151–0.219) for individuals with schizophrenia[12] and 0.156 (95% CI = 0.139–0.174) for unaffected individuals[13]. This estimate (0.03 per exome) was several times smaller than the elevation of dURVs in affected individuals in our population-based study (0.25 per exome). We note that such a comparison requires the imperfect assumption of uniform technical ascertainment across the sequencing studies; even under plausible relaxations of this assumption, the dURV excess greatly exceeded the dDNM excess. In addition, when estimated by this same approach, rates of synonymous and non-damaging DNMs were similar, 0.475 in affected and 0.459 in unaffected individuals, suggesting that the analysis is well-calibrated. We conclude that the great majority of the dURVs driving the elevated rates in schizophrenia were inherited rather than *de novo*, although the very-low allele frequency of these variants suggests that they are on average just a few generations old.

Although the elevated frequency of dURVs among affected individuals was statistically significant ($P = 1.4 \times 10^{-10}$), it was still only a modest increase of 0.25 dURVs on a background of about four dURVs per individual. This excess could in principle be concentrated in individual genes or in sets of functionally related genes.

Single gene burden analysis

Joint analysis of many rare variants that affect the same gene or sets of genes can increase power to identify genes whose disruption increases the risk of schizophrenia. To find individual genes that had significantly more rare variants in cases or controls, we performed a burden test using the SNP-set (Sequence) Kernal Association Test software[14] adjusting for previously defined covariates (Online Methods). We tested for disruptive, damaging, disruptive and damaging, and missense variants that were either ultra-rare, singletons (in the Sweden cohort), had a minor allele count ≤5 (minor allele frequency <0.02%), had a minor allele count ≤10 (minor allele frequency <0.05%), had a minor allele frequency <0.1%, or had a minor allele frequency <0.5%.

Given the sample size, our analysis would have >90% power (at $\alpha = 2.5 \times 10^{-6}$) to detect any gene for which rare, disruptive and damaging variants were present in 1% of schizophrenia cases, even if such variants had only a relatively modest effect size[15] (odds ratio of at least 3, about 2% penetrance), and still greater power if effect sizes were larger. No individual gene surpassed exome-wide significance in this analysis (**Supplementary Fig. 4**). The individual gene with the strongest enrichment was *KL* (*klotho*) (**Supplementary Table 2**), in which we found eight different dURVs in cases and none in controls ($P = 3.7 \times 10^{-4}$), but this result was not significant given the number of genes tested. Other models, based on higher levels of polygenicity, therefore appear to be more plausible: in a model in which a hypothetical gene is affected in 0.1% of schizophrenia cases, we would have only ~4% power to conclusively find this effect at exome-wide significance, and a far-larger sample would be required. The finding that no individual gene surpassed exome-wide significance in this analysis suggests that no one gene is likely to have rare variants that explain even 1% of schizophrenia cases.

Among genes previously reported to have potential connections between rare variants and schizophrenia, we identified in schizophrenia cases an ultra-rare splice donor variant in *TAF13* (ref. 12), an ultra-rare nonsense variant in *SETD1A* (refs. 16,17) and a single ultra-rare nonsense variant in *NRXN1*, a gene in which exonic deletions are associated with schizophrenia[18]. We did not find any evidence of enrichment of dURVs in *DPYD* (ref. 19) (in which we found two dURVs in cases and six in controls), nor in *DISC1* (ref. 20) (one dURV among cases and two in controls) (**Supplementary Table 3**).

With this high level of polygenicity, foreshadowed by earlier results[9,16], it appears that definitive implication of individual genes will require sequencing still-larger numbers of exomes or whole genomes[6]. We therefore focused on sets of genes with plausibly overlapping biological functions as a way of concentrating a diffuse genetic signal.

Enrichment of variants from cases in constrained genes

We tested gene sets for an enrichment of dURVs (in cases relative to controls) by comparing each gene set's enrichment level to that of the average gene (Online Methods). We made this stringent correction to account for the fact that any large gene set is more likely to encompass the exome-wide excess of dURVs that we see in the genomes of individuals with schizophrenia. Our practice greatly deflates the resulting *P* values.

Subsets of human genes have been previously identified as 'missense constrained' (based on a lack of functional coding variation in controls) or 'loss-of-function intolerant' (based on a smaller-than-expected number of loss-of-function mutations in population-scale data)[13,21]. Similar to recent findings in autism, we observed a significant enrichment (in cases relative to controls) of dURVs in missense-constrained genes[22] (OR = 1.28; 95% CI = 1.20–1.37; $P = 3.2 \times 10^{-8}$) and loss-of-function intolerant genes[8] (OR = 1.17; 95% CI = 1.12-1.21; $P = 1.7 \times 10^{-8}$) (**Fig. 2**). Both missense-constrained and loss-of-function intolerant genes were enriched for disruptive variants relative to damaging variants (**Supplementary Fig. 5**); for the latter set, this may reflect that these genes were ascertained specifically for intolerance to disruptive mutations.

In contrast, genes not meeting earlier criteria for loss-of-function intolerance or missense constraint were much less enriched for dURVs (**Supplementary Fig. 6**). This important negative control confirms that the schizophrenia-associated elevation that we observed (for constrained genes) is not a result of false positives that are disproportionally represented across disruptive and damaging variants in cases. The observed enrichment was consistent across data from previously analyzed exomes[9] and newly generated data (**Supplementary Fig. 5**); given that the previously analyzed exomes were sequenced across randomized batches with equal number of cases and controls in each batch, this provides additional evidence that the enrichment is not a result of technical effects.

Tissues and cell types

The excess of dURVs could in principle be concentrated in genes expressed in specific tissues. Distinct tissues have both shared and tissue-specific sets of expressed genes. We found that a set of 2,647 genes expressed specifically in brain tissue[23] was strongly enriched for dURVs (OR = 1.17; 95% CI = 1.11–1.23; $P = 1.2 \times 10^{-4}$), whereas sets of genes with expression specific to other tissues (including immune cells) were not (**Fig. 3a** and **Supplementary Fig. 7**). At the same time, the 'brain-specific' genes explained only part of this signal, whereas a larger set of brain-expressed genes explained most of it, suggesting that much of the signal may have come from genes that are expressed in brain as well as other tissues (**Fig. 3a**). This result aligns with earlier findings that SNP haplotypes that have been implicated in schizophrenia genome-wide association studies (GWAS) tend to overlap

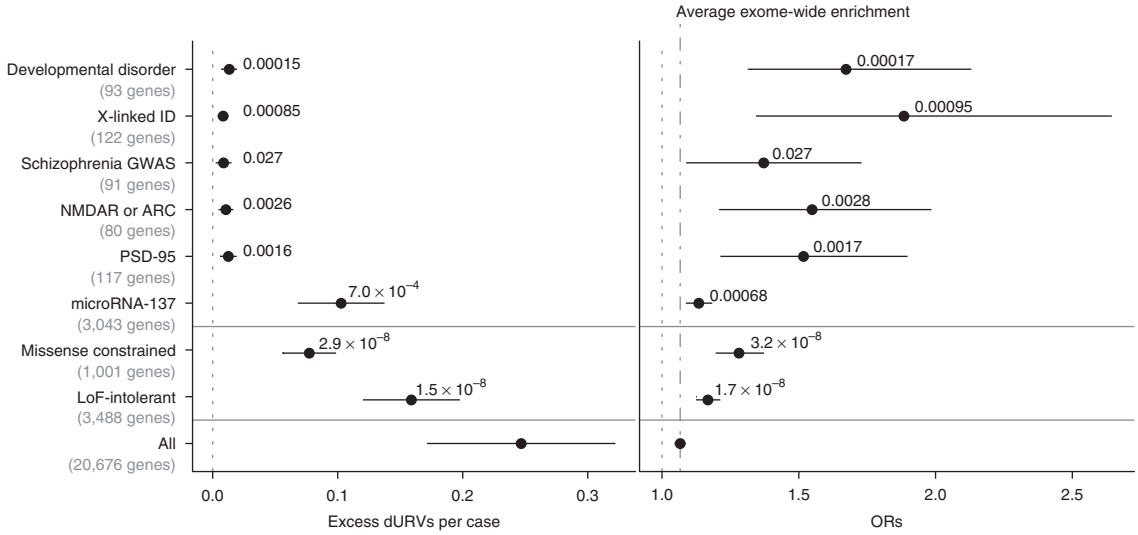

Figure 2 dURV enrichment in schizophrenia cases across selected gene sets. Shown are excess per case and ORs for dURVs across loss-of-function intolerant (LoF-intolerant) genes, missense-constrained genes, protein complexes genes, genes associated through common variants, predicted miRNA-137 targets and intellectual disability genes. Enrichment and *P* values were computed using a linear regression model (left) and a logistic regression model (right) using exome-wide dURV count as a covariate to correct for average exome-wide burden (dot-dashed line). Horizontal bars indicate 95% confidence intervals.

(to a non-random degree) with sequences identified as putative enhancers in chromatin-profiling experiments on brain tissue[24,25].

The brain contains a complex mixture of cell types, each of which expresses different, and only partially overlapping, sets of genes. To identify cell types through which rare variants might act to affect risk of schizophrenia, we evaluated (for enrichment of dURVs in affected relative to unaffected individuals) sets of genes that were identified as being specific to neurons, astrocytes and oligodendrocytes by earlier cell sorting and transcriptional profiling experiments[26]. A set of 3,388 neuron-specific genes had a strong enrichment of mutations in schizophrenia cases (OR = 1.17; 95% CI = 1.12–1.22; $P = 1.9 \times 10^{-7}$), comparable to that observed for genes specific to brain tissue itself. Genes specifically expressed in other brain cell types, such as astrocytes and oligodendrocytes, were no more enriched than the average gene (**Fig. 3b** and **Supplementary Fig. 8a**). These results nominate neurons as the central nervous system (CNS) cell type in which genetic perturbations most affect schizophrenia risk, although they do not exclude more-modest contributions from other CNS cell types.

Neurons are broadly classified into excitatory and inhibitory classes. The case-control excess of dURVs showed a similar degree of concentration in genes expressed in excitatory and inhibitory neurons (**Supplementary Fig. 8b**). The small number of genes that were specific to excitatory or inhibitory neurons (relative to the other class) were insufficient to concentrate this genetic signal, which appeared to reside primarily in genes that were expressed in both neuronal classes (**Supplementary Fig. 8b**).

Synaptic mRNAs

A strong and consistent finding in exome-sequencing studies of schizophrenia involves an excess of variants in genes whose mRNAs are bound by the fragile X mental retardation protein (FMRP)[9,12,27]. The large excess of dURVs that we ascertained in the current set of schizophrenia cases is evidence for this relationship (OR = 1.23; 95% CI = 1.17–1.30; $P = 8.2 \times 10^{-9}$).

The enrichment of dURVs among genes that encode FMRP-bound transcripts could have multiple potential biological explanations. One potential explanation could involve the translational-inhibition capacity of FMRP, as implied by the common description of such genes as FMRP 'targets'. Another potential interpretation is that it is in fact the localization of these RNAs to neuronal processes and synapses by FMRP, its shuttling activity, that defines the important biological commonality among these genes. Yet a third possibility is that FMRP-binding experiments have simply been effective ways of ascertaining neuronally expressed genes.

To evaluate these possibilities, we first considered a different set of genes whose mRNAs are carried to synapses by a different shuttling protein, CELF4 (ref. 28). The genes encoding CELF4-bound mRNAs also showed an enrichment of dURVs in schizophrenia cases; this enrichment was greater than that of the average gene (OR = 1.14; 95% CI = 1.09–1.19; $P = 6.6 \times 10^{-4}$), although less strong than that of genes encoding FMRP-bound RNAs.

We also investigated whether genes encoding mRNAs that are bound by RBFOX splicing factors, which are known to regulate synaptic genes[29] and have been observed at synapses[30], could explain a substantial fraction of the dURVs. Earlier experimental work (based on the HITS-CLIP technique for identifying RNAs bound to proteins of interest) has defined constellations of genes whose RNAs are bound by RBFOX1, RBFOX2 or RBFOX3 (we considered RBFOX1 and RBFOX3 together because of their largely overlapping sets of bound genes[31]). Genes whose transcripts are bound by RBFOX1 or RBFOX3 were enriched in dURVs (OR = 1.16; 95% CI = 1.11–1.21; $P = 6.7 \times 10^{-7}$). A somewhat stronger enrichment was apparent for genes whose RNAs are bound by RBFOX2 (OR = 1.21; 95% CI = 1.16–1.26; $P = 6.3 \times 10^{-12}$).

We also observed enrichment in synaptic genes, as defined by the SynaptomeDB[32] (OR = 1.14; 95% CI = 1.09–1.20; $P = 0.0022$), although this smaller set of genes explained a smaller fraction of the case-control difference in dURVs (**Fig. 3c**).

We were concerned that the enrichment for dURVs in genes with synaptically localized transcripts could, in principle, simply be a result of these experiments having been highly effective at isolating transcripts that are present in neurons (which strongly express FMRP,

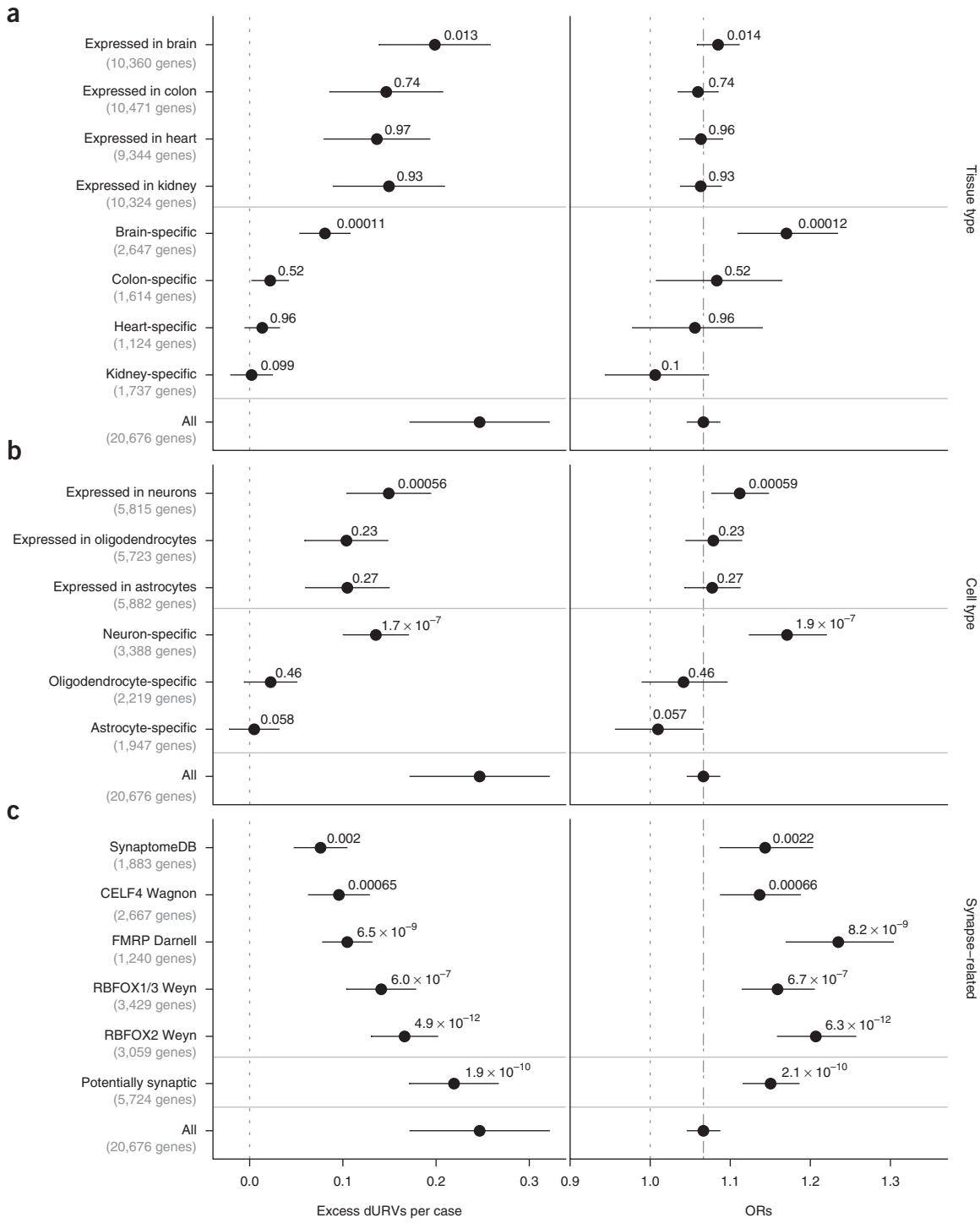

Figure 3 dURVs enrichment in schizophrenia cases across tissue, brain cell type and synaptic gene sets. Excess per case and ORs for dURVs across genes with higher expression in a given tissue (**a**), genes with higher expression in a given cell type (**b**) and genes expected to localize to synapses (**c**). Enrichment and *P* values were computed using a linear regression model (left) and a logistic regression model (right) using exome-wide dURV count as a covariate to correct for average exome-wide burden (dot-dashed line). Horizontal bars indicate 95% confidence intervals.

CELF4 and RBFOX1/2/3); in this case, the importance of synaptic localization would be uncertain. To address this possibility, we identified, from earlier experimental data, sets of genes expressed in brain tissue[23], neurons[26], excitatory neurons and inhibitory neurons[33]. In each set, we defined a gene as being 'potentially synaptic' if it was in any of the previously constructed FMRP, CELF4, RBFOX2 or SynaptomeDB gene sets, and then stratified each of the neuronal/brain expression gene sets on the basis of whether or not the genes were potentially synaptic (**Fig. 4**). No matter how we defined neuronally expressed genes, we observed that this tendency to contain an excess of dURVs in schizophrenia cases distinguished the potentially synaptic genes (which showed elevated rates of dURVs in schizophrenia)

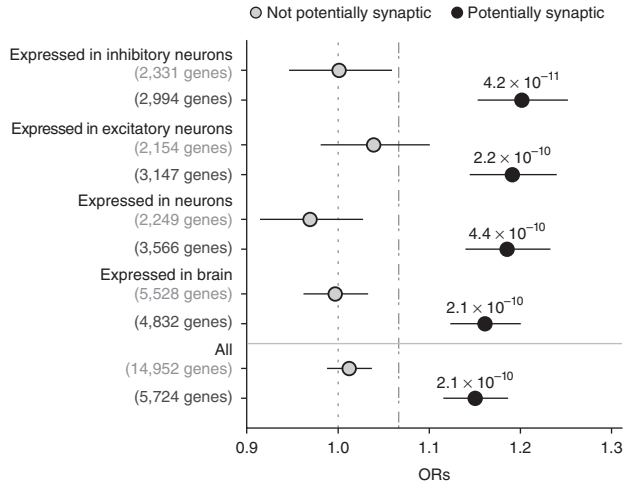

Figure 4 dURVs enrichment in schizophrenia cases across brain cell type gene sets stratified by synaptic localization. ORs for enrichment of dURVs across genes expressed in brain tissue, neuronal cells, inhibitory neurons and excitatory neurons, stratified between genes recognized as synaptic and genes recognized as non-synaptic. Synaptic genes were defined as genes that were part of the FMRP, RBFOX2, CELF4 or SynaptomeDB gene sets. Enrichment and *P* values were computed using a logistic regression model using exome-wide dURV count as a covariate to correct for average exome-wide burden (dot-dashed line). Horizontal bars indicate 95% confidence intervals. Across each gene set, synaptic genes were clearly more enriched for variants in schizophrenia cases than non-synaptic genes.

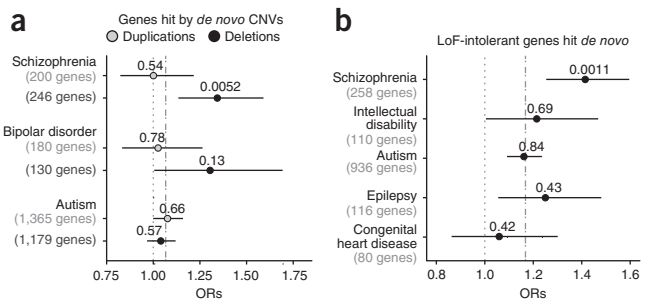

Figure 5 dURVs enrichment in schizophrenia cases across genes previously observed as being affected by *de novo* mutations. ORs for enrichment of dURVs across genes overlapping *de novo* deletions and duplications in schizophrenia, bipolar disorder and autism trios (**a**), and across (**b**) loss-of-function intolerant genes with observed *de novo* mutations in schizophrenia, intellectual disability, congenital heart disease, epilepsy and autism trios (**b**). Enrichment and *P* values were computed using a logistic regression model using exome-wide dURV count (**a**) and dURV count across loss-of-function intolerant genes (**b**) as a covariate to correct for average burden (dot-dashed line). Horizontal bars indicate 95% confidence intervals.

from other neuronally expressed genes (which did not) (**Fig. 4**). These large constellations of potentially synaptic genes appeared to explain a large fraction (collectively more than 70%) of the exome-wide enrichment in dURVs (**Fig. 4**).

Protein complexes

Protein complexes have been used to define sets of genes with aligned activities, offering potentially meaningful ways to group genes for genetic analysis. We focused on genes encoding proteins that have been detected at synaptic complexes by co-immunoprecipitation with known synaptic components followed by mass spectrometric proteomic analyses. These gene sets have been the source of primary enrichment results in earlier studies of CNVs and rare and *de novo* SNVs in schizophrenia patients[9,12,34]. We observed case-control enrichment of dURVs among genes thus defined as encoding interactors with PSD-95 (OR = 1.52; 95% CI 1.21–1.90; $P = 0.0017$), ARC and NMDA receptors (NMDARs)[35] (OR = 1.55; 95% CI = 1.21–1.98; $P = 0.0028$) (**Fig. 2**). Despite these elevated levels of enrichment, these smaller gene sets explained much smaller fractions (collectively 4–12%) of the case-control enrichment in dURVs, perhaps reflecting that these gene sets include just a fraction of the proteins that are present at synapses.

More-complete ascertainment of the protein components of synaptic structures is an important future research direction that might advance functional analysis and interpretation of larger constellations of rare variants.

Overlap with GWAS genes and intellectual disability

We tested whether genes in the 108 GWAS loci recently identified in schizophrenia also contain an excess of dURVs. We observed a nominally significant enrichment in genes overlapping regions near common variants associated with schizophrenia[24] (OR = 1.37; 95% CI = 1.09–1.73; $P = 0.027$; **Fig. 2**) hinting at some degree of convergence. This overlap was greater than could be explained by any individual gene or small set of genes. Predicted targets of miRNA-137 (ref. 36), which was previously identified as being localized near common SNPs associated with schizophrenia[37], were also significantly enriched for dURVs (OR = 1.13; 95% CI = 1.09–1.18; $P = 6.8 \times 10^{-4}$; **Fig. 2**).

Mutations associated with intellectual disability and developmental disorders are often also substantial risk factors for syndromic forms of autism and perhaps schizophrenia[17,38–40]. We observed that the dURV elevation was concentrated in X-linked intellectual disability (XLID) genes[41,42] (OR = 1.88; 95% CI = 1.34–2.64; $P = 9.5 \times 10^{-4}$) and in developmental disorder (DD) genes[43] (OR = 1.67; 95% CI = 1.31–2.13; $P = 1.6 \times 10^{-4}$) (Online Methods). Of potential interest, we identified four dURVs in schizophrenia cases (and none in controls) in XLID gene *KDM5C*, an H3K4 methylation eraser gene[44]; 11 dURVs in cases (and 2 in controls) in DD gene *KDM5B*, another H3K4 methylation eraser gene[45]; and 11 dURVs in cases (and 3 in controls) in DD gene *ITPR1*, which encodes an inositol triphosphate receptor[46] (**Supplementary Tables 2** and **3**). The enrichment of XLID variants was not different between female and male cases (**Supplementary Fig. 9**).

Overlap with *de novo* mutations ascertained in trios

We further tested for enrichment of dURVs in genes overlapping *de novo* copy number variants (CNVs) previously found in individuals with schizophrenia, bipolar disorder and autism (**Supplementary Tables 4** and **5**, see Online Methods), and genes in which *de novo* non-synonymous mutations were previously ascertained in individuals with autism, congenital heart disease, epilepsy, intellectual disability and schizophrenia (**Supplementary Tables 6** and **7**, see Online Methods). Because *de novo* non-synonymous mutations have been ascertained in such a large number of genes, we sought to increase specificity by restricting this analysis to loss-of-function intolerant genes, as previously defined[8]. We observed a significant enrichment in genes in *de novo* deletions that were previously ascertained in schizophrenia cases (OR = 1.34; 95% CI = 1.13–1.59; $P = 0.0052$) (**Fig. 5a**), as well as an enrichment in loss-of-function intolerant genes with *de novo* non-synonymous mutations in schizophrenia cases (OR = 1.41; 95% CI = 1.25–1.60; $P = 0.0011$) (**Fig. 5b**).

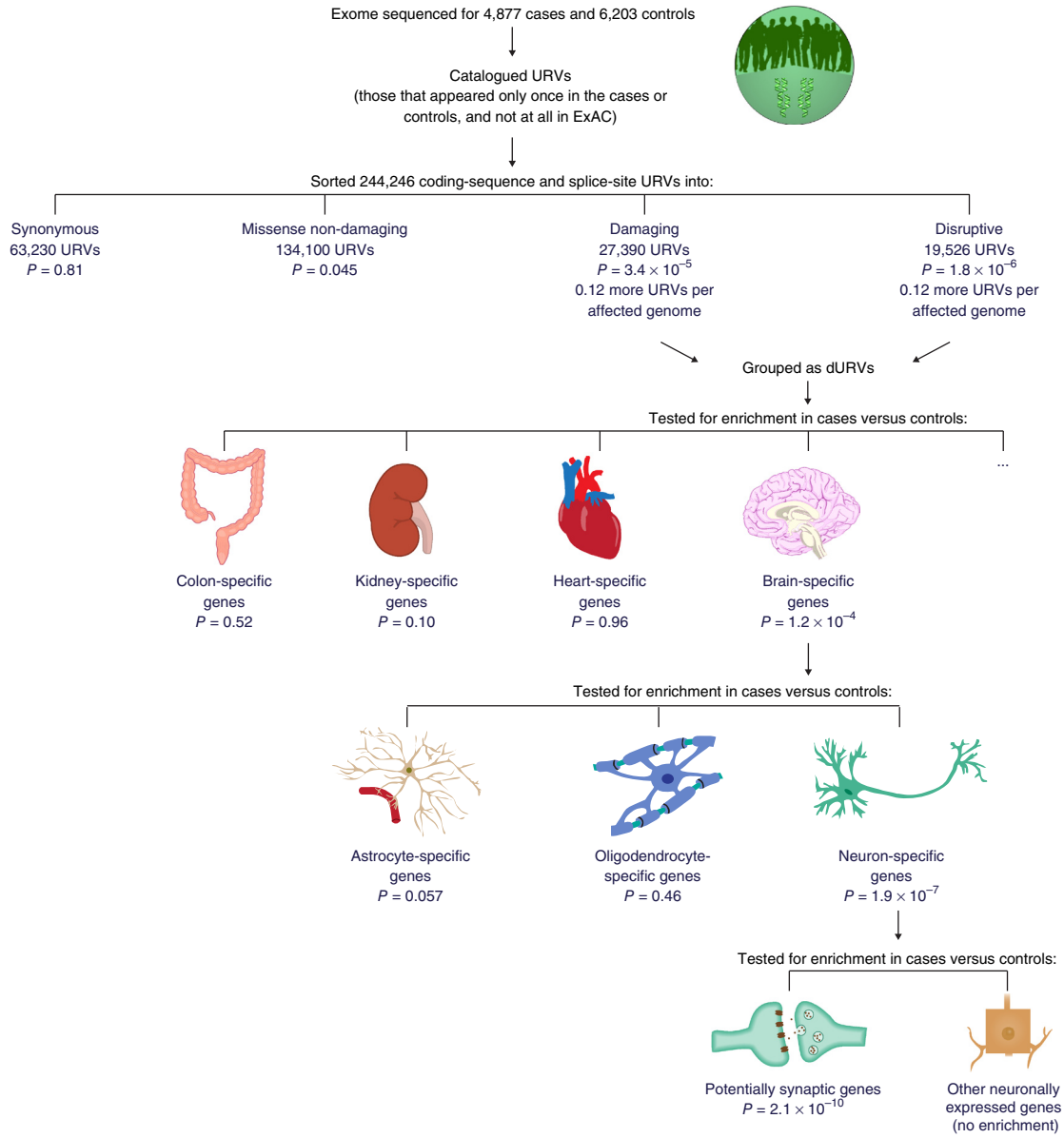

Figure 6 Dissection of the dURVs enrichment in schizophrenia cases. An enrichment of URVs in the exomes of individuals affected with schizophrenia (relative to variants in control exomes) was observed exclusively in dURVs. After correcting for exome-wide dURV count, this enrichment was observed as being concentrated in brain-specific genes, but not in other tissue-specific genes; in neuron-specific genes, but not in other brain cell type specific genes; and finally in potentially synaptic genes, but not in other neuronally expressed genes.

DISCUSSION

By sequencing the exomes of 12,332 unrelated individuals from Sweden, including 4,946 affected with schizophrenia, we found an exome-wide burden of dURVs in individuals affected with schizophrenia. This excess rare-variant burden, approximately 0.25 such variants per person (on a background of four such variants), was several times greater than the schizophrenia-associated elevation in rates of gene-disruptive and protein-damaging *de novo* mutations, suggesting that the observed excess arose mostly from inherited variants. For less-rare (segregating) variants of even modest allele frequencies, we were unable to detect any excess in affected relative to unaffected individuals (Online Methods), consistent with a previous analysis of non-ultra-rare exonic variants in other cohorts[47].

The excess of dURVs in schizophrenia cases largely resided in brain-expressed genes, and more specifically in genes that are expressed in neurons, rather than in other CNS cell types (**Fig. 6**). It is possible that earlier associations to small, protein-interaction-defined gene sets (such as PSD-95, NMDAR and ARC)[9,12,34], which appear to explain combined a much-smaller fraction of the exome-wide dURV burden in schizophrenia (collectively 4–12%), have been proxies for a far-wider set of rare-variant effects at synapses.

Most of the excess of dURVs in affected individuals' exomes appeared to be concentrated in a larger set of genes encoding potentially synaptic proteins. Genes whose transcripts are bound by FMRP or CELF4, which transport a subset of neuronal RNAs to neuronal processes and synapses, or RBFOX2, which regulates many synaptic RNAs and has been observed at synapses, explained considerably larger fractions (collectively more than 70%) of the global rare-variant enrichment that we observed in cases. Genes encoding RNAs bound by FMRP or RBFOX proteins have been shown to be enriched for

mutations in subjects with autism and/or schizophrenia[9,12,13,48,49], although it has been unclear whether such potential effects are a small or a large fraction of strongly risk-increasing variants. Although it is tempting to attribute the association of schizophrenia with dURVs in FMRP-associated, CELF4-associated and RBFOX2-associated genes to the specific biological activities of these proteins, we propose that their association may simply reflect the synaptic localization and function of the transcripts and proteins encoded by these genes.

We observed a significant overlap of the dURV excess with genes in which *de novo* non-synonymous mutations and deletions have been found in schizophrenia cases. We also observed a significant enrichment across intellectual disability genes on the X chromosome and in developmental disorder genes. This enrichment is compatible with observations of the role of intellectual disability genes in some cases of autism[38–40] and schizophrenia[17], although the penetrance of such mutations for schizophrenia may be much less than their penetrance for intellectual disability, and they may reside primarily in syndromic cases in which schizophrenia is preceded by other developmental disorders[17].

The fact that an analysis of the current scale (4,877 cases, 6,203 controls, and 45,376 other genomes used to help identify ultra-rare variants) did not implicate individual genes of large effect in an unbiased exome-wide search, although it documented a very clear exome-wide elevation of hundreds of pathogenic variants across 4,877 individuals affected with schizophrenia relative to controls, lends further support to the emerging impression that the high polygenicity of schizophrenia extends to both rare and common variants[9,16]. Because of the rareness of these variants and the infrequency with which any individual gene is affected by them, even among schizophrenia cases, the sequencing of much larger cohorts will be needed to identify the specific individual genes in which rare variants shape risk for schizophrenia.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Accession codes. dbGaP: phs000473.v2.p2.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Usher for comments on the manuscript and work on the figures. This study was supported by grants from the National Human Genome Research Institute (U54 HG003067, R01 HG006855 to S.A.M.), the National Institute of Mental Health (R01 MH077139 to P.F.S., R01 MH095034 to P.S., and RC2 MH089905 to S.M.P. and P.S.), the Stanley Center for Psychiatric Research, the Alexander and Margaret Stewart Trust, and the Sylvan C. Herman Foundation.

AUTHOR CONTRIBUTIONS

G.G. and S.A.M. designed the analyses and wrote early drafts of the manuscript. G.G. performed the analyses. M.F. contributed to analyses of *de novo* mutated genes, D.M.R. and E.A.S. contributed with the specific design of the analyses. K.C. contributed with sample processing and data management. M.L., J.L.M., S.M.P., P.S., P.F.S. and C.M.H. contributed with sample and phenotype collection. All of the authors contributed to interpretation of the findings and revisions of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30**, 67–76 (2008).
2. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
3. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).
4. Bundy, H., Stahl, D. & MacCabe, J.H. A systematic review and meta-analysis of the fertility of patients with schizophrenia and their unaffected relatives. *Acta Psychiatr. Scand.* **123**, 98–106 (2011).
5. Power, R.A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22–30 (2013).
6. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111**, E455–E464 (2014).
7. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
8. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
9. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
10. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
11. Maquat, L.E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* **5**, 89–99 (2004).
12. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
13. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
14. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92**, 841–853 (2013).
15. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
16. Takata, A. *et al.* Loss-of-function variants in schizophrenia risk and *SETD1A* as a candidate susceptibility gene. *Neuron* **82**, 773–780 (2014).
17. Singh, T. *et al.* Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
18. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet.* **18**, 988–996 (2009).
19. Xu, B. *et al. De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
20. Millar, J.K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* **9**, 1415–1423 (2000).
21. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
22. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
23. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
24. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
25. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
26. Cahoy, J.D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
27. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
28. Wagnon, J.L. *et al.* CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS Genet.* **8**, e1003067 (2012).
29. Lee, J.-A. *et al.* Cytoplasmic Rbfox1 regulates the expression of synaptic and autism-related genes. *Neuron* **89**, 113–128 (2016).
30. Hamada, N. *et al.* Biochemical and morphological characterization of A2BP1 in neuronal tissue. *J. Neurosci. Res.* **91**, 1303–1311 (2013).
31. Weyn-Vanhentenryck, S.M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6**, 1139–1152 (2014).
32. Pirooznia, M. *et al.* SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **28**, 897–899 (2012).
33. Mo, A. *et al.* Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86**, 1369–1384 (2015).
34. 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).
35. Bayés, A. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* **14**, 19–21 (2011).

36. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, R90 (2010).
37. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
38. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of *de novo* and familial influences. *Proc. Natl. Acad. Sci. USA* **111**, 15161–15165 (2014).
39. Robinson, E.B., Neale, B.M. & Hyman, S.E. Genetic research in autism spectrum disorders. *Curr. Opin. Pediatr.* **27**, 685–691 (2015).
40. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552–555 (2016).
41. Moeschler, J.B. Genetic evaluation of intellectual disabilities. *Semin. Pediatr. Neurol.* **15**, 2–9 (2008).
42. Gécz, J., Shoubridge, C. & Corbett, M. The genetic landscape of intellectual disability arising from chromosome X. *Trends Genet.* **25**, 308–316 (2009).
43. McRae, J.F. *et al.* Prevalence, phenotype and architecture of developmental disorders caused by *de novo* mutation. *bioRxiv* http://dx.doi.org/10.1101/049056 (2016).
44. Jensen, L.R. *et al.* Mutations in the *JARID1C* gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. *Am. J. Hum. Genet.* **76**, 227–236 (2005).
45. Xiang, Y. *et al.* JARID1B is a histone H3 lysine 4 demethylase up-regulated in prostate cancer. *Proc. Natl. Acad. Sci. USA* **104**, 19226–19231 (2007).
46. Matsumoto, M. *et al.* Ataxia and epileptic seizures in mice lacking type 1 inositol 1,4,5-trisphosphate receptor. *Nature* **379**, 168–171 (1996).
47. Richards, A.L. *et al.* Exome arrays capture polygenic rare variant contributions to schizophrenia. *Hum. Mol. Genet.* **25**, 1001–1007 (2016).
48. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
49. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).

ONLINE METHODS

Sample collection and sequencing. A total of 12,384 blood-derived DNA samples from Swedish research participants were collected from 2005 to 2013. Psychiatric cases with a diagnosis of schizophrenia or bipolar disorder were ascertained from the Swedish National Hospital Discharge Register as described in previous studies[9,50], which captures all inpatient hospitalizations. Controls were randomly selected from population registers. Excluding subjects with bipolar disorder, age information at the time of DNA sampling was available for each individual. All subjects provided informed consent; institutional human subject committees approved the research (UNC IRB # 04-1465). All procedures were approved by the ethical committees in Sweden and in the US.

The 12,384 samples collected were sequenced in twelve separate waves. The first wave employed an earlier version of the hybrid-capture procedure (Agilent SureSelect Human All Exon Kit), which targets ~28 million base pairs of the human genome, partitioned in ~160,000 intervals, whereas the samples from the other waves used a newer version (Agilent SureSelect Human All Exon v.2 Kit), which targets ~32 million base pairs of the human genome, partitioned in ~190,000 intervals. The first wave was sequenced using Illumina GAII instruments and the remaining waves were sequenced using Illumina HiSeq 2000 and HiSeq 2500 instruments, with pair ended sequencing reads of 76 base pairs across all waves. Sequencing was performed at the Broad Institute of MIT and Harvard across the period of time from 2010 to 2013. With the exception of the first wave, we did not observe significant differences across waves and cases status beyond what could be explained by ancestry (**Supplementary Fig. 10**).

This cohort has been previously analyzed in relation to schizophrenia for common variants[24,37,50–52] and copy number variants[52,53], and in relation to somatic mosaic mutations[54]. Exome sequence data for approximately half of the individuals in the cohort had already been analyzed in relation to schizophrenia phenotype in a previous study[9] and in a more recent study[17].

Preliminary quality control for individuals. Exome sequence data from 12,384 samples was aligned against the GRCh37 human genome reference with bwa aln 0.5.9 (ref. 55) and further processed using the GATK framework[56]. Genotype calls were generated using GATK Haplotype Caller version 3.1-144-g00f68a3 and best practices[57,58]. Variants filtered out by the GATK Variant Quality Score Recalibration (VQSR) tool were excluded. Genotypes over sites with less than 10x sequencing coverage were set to missing. We identified and removed 4 duplicate individuals and 48 individuals with a first degree relationship (**Supplementary Fig. 11**) with other individuals in the cohort using the plink[59,60] software. We then computed the number of ultra-rare SNPs and indels never observed in ExAC[8] for each of the remaining 12,332 samples and identified one individual with 1,757 ultra-rare SNPs and 22 ultra-rare indels from the sixth sequencing wave (see Supplementary Table S5F from ref. 54), four individuals with between 92 and 127 ultra-rare indels from the first sequencing wave, five individuals from waves 11 and 12 with between 410 and 496 ultra-rare SNPs due to African ancestry. These ten individuals were excluded from further analysis. The resulting individuals had a range of 5–259 ultra-rare SNPs and 0–19 ultra-rare indels (**Supplementary Fig. 12a**). We further removed 15 individuals for whom the reported sex and the inferred sex from inbreeding coefficients on the X chromosome mismatched (**Supplementary Fig. 13**), including 7 individuals with 47, XXY karyotype (Klinefelter syndrome), and 94 individuals with more than 100 URVs.

Association with common variants. A logistic regression model was used to estimate association between single variants and schizophrenia phenotype correcting for sex and the first five principal components using plink[59,60]. We identified two loci (**Supplementary Fig. 14**) with statistically significant associations ($P < 10^{-6}$) replicating a couple of common variants associations previously observed for this cohort[50]: a single variant rs281766 on chromosome 2 in the UTR5 of genes *TYW5* and *C2orf47*, a variant in strong linkage disequilibrium with the seventh strongest independently associated variant in the largest meta-analysis for schizophrenia[24], and seven variants in the MHC region around the HLA genes, also a region with extensive linkage to known causal variants associated with schizophrenia[61].

Variant annotation. We annotated all genotyped variants with SnpEff 4.2 (build 2015-12-05)[62] using Ensembl gene models from database GRCh37.75. We further annotated variants with SnpSift 4.2 (build 2015-12-05)[63] using annotations from database dbNSFP 2.9 (refs. 64,65). Variants identified within transcripts from UCSC known genes[66] were further classified into four groups:

- synonymous: whenever classified with synonymous effect by SnpEff
- missense non-damaging: whenever classified with missense effect but not classified as *damaging* (by the criteria below)
- putatively protein-damaging: whenever classified with MODERATE impact by SnpEff and further predicted as damaging by each among SIFT[67], PolyPhen-2 (ref. 68), LRT[69], Mutation Taster[70], Mutation Assessor[71], and PROVEAN[72] algorithms or classified as either in-frame indels or protein-protein-contact variants[73]
- gene disruptive: whenever classified with HIGH impact by SnpEff with the exclusion of protein-protein-contact variants

Notice that FATHMM[74] predictions included in dbNSFP were not used due to poor performance with respect to minor allele count (**Supplementary Fig. 1**) and a small number of variants defined as damaging by the predictor (**Supplementary Fig. 3b**). The final predictor performed better than all other individual predictors (**Supplementary Fig. 3b**), but it was not overfit as to be the best predictor for this cohort (**Supplementary Fig. 15**).

Estimation of principal components. Out of a total of 1,753,312 variants passing VQSR filters, 66,874 were identified as in common with variants from the 1000 Genomes project phase 1 data set[75] and included as part of the Omni2.5 genotype array. We used this subset of highly confident variants to estimate population stratification. We selected exclusively Omni2.5 polymorphic sites because more robust in the 1000 Genomes data set to artifacts due to the heterogeneity of the sequencing technologies used within the 1000 Genomes project. We then further restricted to variants with minor allele frequency larger than 1% in both the Sweden and the 1000 Genomes data set and we pruned for variants in linkage disequilibrium using plink[59,60] (with command line '--indep 50 5 2'). We then merged the Sweden and the 1000 Genomes data set and computed principal components using plink and GCTA[76] (**Supplementary Fig. 12c**,**d**). Estimated third and fifth principal components corresponded to previously observed Finnish and Northern-Southern Sweden clines[12] (**Supplementary Fig. 12d**), while first, second and fourth principal components corresponded to the three main principal components in the 1000 Genomes project phase 1 distinguishing African, East Asian, and Native American ancestry. While principal components did correlate with overall amounts of URVs (**Supplementary Fig. 12e**,**f**), rather than removing individuals with exotic ancestry based on principal components loading, we simply removed individuals with more than 100 URVs and we included sex, year of birth (**Supplementary Fig. 12b**), exome capturing kit, the first 20 principal component loadings, and the total number of URVs for each individual as covariates in all statistical analyses involving URVs and dURVs.

Quality control for common variants. Variants were excluded whether failing the GATK VQSR tool (117,629 variants), having inbreeding coefficient less than zero (that is, more observed heterozygotes than expected) while at the same time failing a Hardy-Weinberg equilibrium test with a false discovery rate of 10^{-6} (8,306 additional variants), or whether associating with any of the 146 batches among which the cohort was split for sequencing in the sequencing facility at the Broad Institute (3,700 additional variants). Due to prevalent population stratification within batches, we estimated unusual associations with a logistic regression model including sex and the first 20 principal components loadings as covariates.

Excess of dURVs. For each gene set, to estimate the excess of dURVs in cases with schizophrenia (or the odds ratios for schizophrenia phenotype) we used a linear (or logistic) regression model correcting for sex, overall URV count, birth year, the hybrid selection kit used to enrich for exome sequence, and the first 20 principal components estimated from exome-wide SNP genotypes. When estimating P values, to estimate the importance of each gene set with respect to the observed exome-wide enrichment, we further corrected for exome-wide dURV count. This expedient allows to better estimate the importance of each gene set irrespectively of its size and to answer the more precise question of whether a gene set concentrates the exome-wide dURV enrichment better than the average gene.

doi:10.1038/nn.4402

Construction of gene sets. We used different resources to build the gene sets for which the burden of dURVs was computed:

1. For missense constrained genes we used genes from Supplementary Table 2 of ref. 22.
2. For loss-of-function intolerant genes we used genes from ref. 8, available at ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/functional_gene_constraint/.
3. For genes with expression specific to the brain, we used expression table from Supplementary Data Set 1 of ref. 23 and we selected genes for which expression in brain was four times higher than the median expression across all 27 different tissues (**Supplementary Fig. 7**).
4. For brain genes with expression specific to neurons, we used expression table from Supplementary Table S3b of ref. 26 and we selected genes for which log-expression in Neurons P7n cell type was 0.5 greater than the median log-expression across 11 CNS cell types (**Supplementary Fig. 8a**).
5. For RBFOX2 and RBFOX1/3 gene sets we selected genes from Supplementary Table S1 of ref. 31 for which at least, respectively, one Rbfox2 tag count was greater than or equal to 4, and one Rbfox1 and Rbfox3 tag counts sum was greater than or equal to 12. A single gene set was generated for RBFOX1 and RBFOX3 due to high correlation between tag counts for the two genes.
6. Instead of using the classical FMRP Darnell gene set of 842 mouse genes from Supplementary Table S2A of ref. 27 including all genes with FDR < 0.01, we used a larger gene set of 1,285 mouse genes from Supplementary Table S2C of ref. 27 including genes with FDR < 0.1.
7. For CELF4 we used genes with "iCLIP occupancy" greater than 0.2 from Supplementary Table S4 of ref. 28.
8. To create a gene set with synaptic genes we included 1,887 genes from the SynaptomeDB[32] from the presynaptic proteins, presynaptic active zone, synaptic vesicles, and postsynaptic density categories.
9. To create a set of genes expressed in brain, we used expression table from Supplementary Data Set 1 of ref. 23 and we selected genes for which fragments per kilobase of transcript per million (FPKM) in brain was larger than 5.
10. To create a set of genes expressed in neurons, we used expression table from Supplementary Table S3b of ref. 26 and we selected genes for which log-expression in Neurons P7n cell type was larger than 9.
11. To create sets of genes expressed in excitatory and inhibitory neurons, we used expression table from Table S2 of ref. 33 and we selected genes for which the average transcripts per million (TPM) of, respectively, excitatory pyramidal neurons and inhibitory neurons, the latter including parvalbumin (PV)-expressing fast-spiking or vasoactive intestinal peptide (VIP)-expressing interneurons, was larger than 50. Similarly for sets of genes specific for each neuron type, we selected genes expressed more than 5 times the minimum expression observed across all types (**Supplementary Fig. 8b**).
12. To generate a list of predicted targets of microRNA-137, we used human targets with good mirSVR score from ref. 39, available at http://www.microrna.org/.
13. To generate PSD-95 complex gene sets, we used a gene list generated from human cortex biopsy data[35] available at http://www.genes2cognition.org/db/GeneList/L00000049.
14. To compute a combined NMDAR and ARC complexes gene set, we used genes from Table S9 of ref. 34.
15. For genes implicated in common variant association studies, we used genes overlapping 62 regions from the 108 regions known to be associated with schizophrenia[24], for which the overlap yielded at most four genes.
16. To generate genes involved in X-linked intellectual disability (XLID) we used gene lists available online (see next section).
17. To generate genes involved in developmental disorder, we selected genes from Supplementary Table 3 of ref. 43.
18. For genes implicated in *de novo* CNV studies, we used genes overlapping *de novo* deletions and duplications identified in autism[77–84] and identified in bipolar disorder and schizophrenia[34,85–88] cases (**Supplementary Tables 4** and **5**).
19. For genes implicated in *de novo* nonsynonymous mutations from exome sequencing studies, we used genes identified as mutated in autism[13,49,89,90], epilepsy[91,92], congenital heart disease[93], intellectual disability[94–97], and schizophrenia[12,19,98–100] (**Supplementary Tables 6** and **7**).

Enrichment in XLID genes. We used three different resources available online to define XLID genes:

- XLID OMIM genes were defined as those genes causing mental retardation phenotype in the OMIM database[101] (http://omim.org/geneMap/X).
- XLID GCC genes were defined as those genes tested by the Greenwood Genetic Center[102] (http://www.ggc.org/diagnostic/tests-costs/test-finder/test-finder.html?id=242).
- XLID Chicago genes were those tested by the Genetic Services Laboratories of the university of Chicago[41,42,103,104] (http://dnatesting.uchicago.edu/tests/x-linked-non-specific-intellectual-disability-sequencing-panel).

We tested for enrichment of dURVs in the above gene sets and in genes including all of the above gene sets, and separately genes believed to escape and not escape X-inactivation in humans[105], genes from OMIM including autosomal genes causing intellectual disability, and genes involved in developmental disorders through *de novo* mutations[43]. We tested for enrichment separately in males and females, as well as combined (**Supplementary Fig. 9**).

While XLID genes and developmental disorders genes were strongly enriched in schizophrenia cases, autosomally linked intellectually disability genes were not. This discrepancy might reflect a better characterization of intellectual disability genes on the X chromosome due to a more straightforward study design for how these genes were discovered. We also observed that XLID genes which escape X-inactivation were more enriched than other XLID genes. This might reflect a disproportionate contribution to intellectual disability and psychosis from dosage sensitive brain-related genes on the X chromosome[106–108]. We did not observe a difference in effect sizes between males and females.

Similarly to rare variants enrichment, common variants associated with schizophrenia are localized near XLID genes *CNKSR2* and *NLGN4X*, both of which escape X inactivation, as well as non-XLID gene *PJA* 1 (ref. 24).

No detectable enrichment of less-rare variants. Given the strong case-control enrichment of dURVs in potentially synaptic genes, we used these genes to perform a sensitive evaluation of whether we could observe an increased burden of less-rare disruptive and damaging variants in the same set. Using a standard burden test for non-ultra-rare variants with minor allele count 10 or less and controlling for covariates, we observed no statistically significant enrichment of disruptive and damaging variants in schizophrenia cases compared to controls ($P = 0.59$), whereas the same test was highly significant when restricted to URVs ($P = 1.7 \times 10^{-19}$, without controlling for exome-wide enrichment) (**Supplementary Table 8**).

Variance explained. While a predictor based on common variants[24,109] explained 15% of the variance in schizophrenia liability in this cohort, a predictor based on the cumulative burden of dURVs in all genes explained only 0.48% ($P = 1.5 \times 10^{-10}$), and a similar predictor in potentially synaptic genes explained only 0.92% (Nagelkerke's coefficient of determination) ($P = 6.3 \times 10^{-19}$). We also attempted to generate a polygenic score based on the cumulative number of dURVs in genes that had burden of dURVs in cases greater than or equal to controls. Using a leave-one-out strategy, the resulting predictor explained 0.47% ($P = 2.3 \times 10^{-10}$) of the phenotypic variability. These estimates are naturally lower bounds on the effect of rare variants; knowledge of the correct effect size of each variant would significantly increase the predictive value of dURVs, though obtaining such knowledge will require sequencing a vastly larger number of exomes.

Data availability. Scripts used to perform all of the analyses are available at https://github.com/freeseek/gwaspipeline and data are available through dbGAP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473.v2.p2.

50. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
51. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

52. Bergen, S.E. *et al.* Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol. Psychiatry* **17**, 880–886 (2012).
53. Szatkiewicz, J.P. *et al.* Copy number variation in schizophrenia in Sweden. *Mol. Psychiatry* **19**, 762–773 (2014).
54. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
56. McKenna, A. *et al.* The Genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
57. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
58. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 1–33 (2013).
59. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
60. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
61. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
62. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
63. Cingolani, P. *et al.* Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, 35 (2012).
64. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).
65. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013).
66. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006).
67. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
68. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
69. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
70. Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
71. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
72. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
73. Berman, H.M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
74. Shihab, H.A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
75. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
76. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
77. Szatmari, P. *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**, 319–328 (2007).
78. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
79. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
80. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
81. Itsara, A. *et al. De novo* rates and selection of large copy number variation. *Genome Res.* **20**, 1469–1481 (2010).
82. Sanders, S.J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
83. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886–897 (2011).
84. Gilman, S.R. *et al.* Rare *de novo* variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898–907 (2011).
85. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**, 880–885 (2008).
86. Malhotra, D. *et al.* High frequencies of *de novo* CNVs in bipolar disorder and schizophrenia. *Neuron* **72**, 951–963 (2011).
87. Noor, A. *et al.* Copy number variant study of bipolar disorder in Canadian and UK populations implicates synaptic genes. *Am. J. Med. Genet.* **165B**, 303–313 (2014).
88. Georgieva, L. *et al. De novo* CNVs in bipolar affective disorder and schizophrenia. *Hum. Mol. Genet.* **23**, 6677–6683 (2014).
89. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
90. Jiang, Y.H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
91. Allen, A.S. *et al. De novo* mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
92. EuroEPINOMICS-RES Consortium; Epilepsy Phenome/Genome Project; Epi4K Consortium. *De novo* mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–370 (2014).
93. Zaidi, S. *et al. De novo* mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).
94. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
95. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
96. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
97. Hamdan, F.F. *et al. De novo* mutations in moderate or severe intellectual disability. *PLoS Genet.* **10**, e1004772 (2014).
98. Girard, S.L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat. Genet.* **43**, 860–863 (2011).
99. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
100. McCarthy, S.E. *et al. De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol. Psychiatry* **19**, 652–658 (2014).
101. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
102. Stevenson, R.E., Schwartz, C.E., Rogers, R.C. & Rogers, R.C. *Atlas of X-linked Intellectual Disability Syndromes* (Oxford University Press, 2012).
103. Moeschler, J.B., Shevell, M. & American Academy of Pediatrics Committee on Genetics. Clinical genetic evaluation of the child with mental retardation or developmental delays. *Pediatrics* **117**, 2304–2316 (2006).
104. Rauch, A. *et al.* Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. *Am. J. Med. Genet. A.* **140**, 2063–2074 (2006).
105. Cotton, A.M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.* **14**, R122 (2013).
106. Crow, T.J. The XY gene hypothesis of psychosis: origins and current status. *Am. J. Med. Genet.* **162B**, 800–824 (2013).
107. Ji, B., Higa, K.K., Kelsoe, J.R. & Zhou, X. Over-expression of XIST, the master gene for X chromosome inactivation, in females with major affective disorders. *EBioMedicine* **2**, 909–918 (2015).
108. Crow, T.J. Is psychosis a disorder of XY epigenetics? *EBioMedicine* **2**, 794–795 (2015).
109. Vilhjálmsson, B.J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).