nature genetics

# Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects

CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium*



**Copy number variants (CNVs) have been strongly implicated in the genetic etiology of schizophrenia (SCZ). However, genome-wide investigation of the contribution of CNV to risk has been hampered by limited sample sizes. We sought to address this obstacle by applying a centralized analysis pipeline to a SCZ cohort of 21,094 cases and 20,227 controls. A global enrichment of CNV burden was observed in cases (odds ratio (OR) = 1.11, $P = 5.7 \times 10^{-15}$), which persisted after excluding loci implicated in previous studies (OR = 1.07, $P = 1.7 \times 10^{-6}$). CNV burden was enriched for genes associated with synaptic function (OR = 1.68, $P = 2.8 \times 10^{-11}$) and neurobehavioral phenotypes in mouse (OR = 1.18, $P = 7.3 \times 10^{-5}$). Genome-wide significant evidence was obtained for eight loci, including 1q21.1, 2p16.3 (*NRXN1*), 3q29, 7q11.2, 15q13.3, distal 16p11.2, proximal 16p11.2 and 22q11.2. Suggestive support was found for eight additional candidate susceptibility and protective loci, which consisted predominantly of CNVs mediated by nonallelic homologous recombination.**

Studies of genomic copy number variation (CNV) have established a role for rare genetic variants in the etiology of SCZ[1]. There are three lines of evidence that CNVs contribute to risk for SCZ: genome-wide enrichment of rare deletions and duplications in SCZ cases relative to controls[2,3], a higher rate of *de novo* CNVs in cases relative to controls[4–6] and association evidence implicating a small number of specific loci (**Supplementary Table 1**). All CNVs that have been implicated in SCZ are rare in the population but confer significant risk (ORs 2–60).

To date, CNVs associated with SCZ have largely emerged from mergers of summary data for specific candidate loci[7–9]; yet even the largest genome-wide scans (sample sizes typically <10,000) remain underpowered to robustly confirm genetic association for the majority of pathogenic CNVs reported so far, particularly for those with low frequencies (<0.5% in cases) or intermediate effect sizes (ORs 2–10). It is important to address the low power of CNV studies with larger samples, given that this type of mutation has already proven useful for highlighting some aspects of SCZ-related biology[6,10–13].

The limited statistical power provided by small samples is a substantial obstacle in studies of rare and common genetic variation. In response, global collaborations have been formed to attain large sample sizes, as exemplified by a genome wide association study (GWAS) of SCZ by the Psychiatric Genomics Consortium (PGC) by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC), which identified 108 independent SCZ-associated loci[14]. Recognizing the need for similarly large samples in studies of CNVs for psychiatric disorders, we formed the PGC CNV Analysis Group. Our goal was to enable large-scale analysis of CNVs in psychiatry using centralized and uniform methodologies for CNV calling, quality control and statistical analysis. Here we report the largest genome-wide analysis of CNVs for any psychiatric disorder to date, using data sets assembled by the Schizophrenia Working Group of the PGC.

## RESULTS

### Data processing and meta-analytic methods

Raw intensity data were obtained from 57,577 subjects and 43 separate data sets (**Supplementary Table 2**). After CNV calling and quality control (QC), 41,321 subjects were retained for analysis. We developed a centralized pipeline for systematic calling of CNVs for Affymetrix and Illumina platforms (Online Methods and **Supplementary Fig. 1**). The pipeline included multiple CNV callers run in parallel. Data from Illumina platforms were processed using PennCNV[15] and iPattern[16]. Data from Affymetrix platforms were analyzed using PennCNV and Birdsuite[17].Two additional methods, iPattern and C-score[18], were applied to data from the Affymetrix 6.0 platform. To ensure proper normalization of the X chromosome, male and female subjects were normalized separately. The CNV calls from each program were converted to a standardized format, and a consensus call set was constructed by merging CNV outputs at the sample level. Only CNV segments that were detected by all algorithms were retained. We performed QC at the platform level to exclude samples with poor probe intensity and/or an excessive CNV load (number and length). A final set of rare, high-quality CNVs was defined as those >20 kb in length, encompassing at least 10 probes and <1% minor allele frequency (MAF).

Genetic associations were investigated by case-control tests of CNV burden at four levels: (i) genome-wide (ii) pathways, (iii) genes and (iv) CNV breakpoints. Analyses controlled for SNP-derived principal components, sex, genotyping platform and data quality metrics. Multiple-testing thresholds for genome-wide significance were estimated from family-wise error rates drawn from permutation.

### Genome-wide analysis of CNV burden

An elevated burden of rare CNVs among SCZ cases has been well established[2]. We applied our meta-analytic framework to measure

*A full list of members and affiliations appears at the end of the article.

Received 11 April; accepted 24 October; published online 21 November 2016; corrected online 5 December 2016 (details online); doi:10.1038/ng.3725

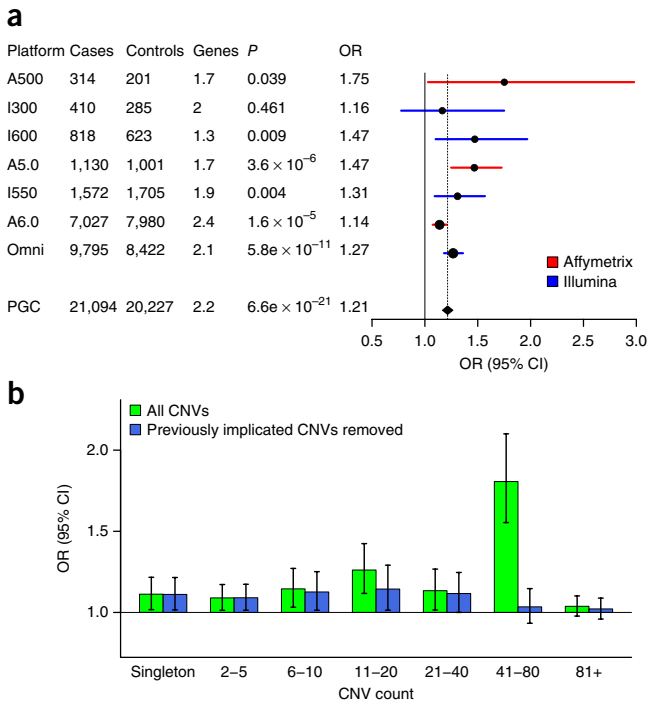


**Figure 1** CNV burden. (**a**) Forest plot of CNV burden (measured as genes affected by CNV), partitioned by genotyping platform, with the full Psychiatric Genomics Consortium (PGC) sample at the bottom. CNV burden is calculated by combining CNV gains and losses. 'Genes' denotes the mean number of genes affected by a CNV in controls. Burden tests use a logistic regression model predicting SCZ case-control status by CNV burden along with covariates (Online Methods). The OR is the exponential of the logistic regression coefficient, and OR > 1 predicts increased SCZ risk. A500, Affymetrix 500; I300, Illumina 300K; I600, Illumina 610K and Illumina 660W; A5.0, Affymetrix 5.0; A6.0, Affymetrix 6.0; Omni, OmniExpress and OmniExpress plus Exome. (**b**) CNV burden partitioned by CNV frequency. For autosomal CNVs, a CNV count of 41 in the sample corresponds to frequency of 0.1% in the full PGC sample. Using the same model as above, each CNV was placed into a single CNV frequency category on the basis of a 50% reciprocal overlap with other CNVs. CNV gene burden with inclusion of all CNVs is shown in green, and burden excluding previously implicated CNV loci is shown in blue.

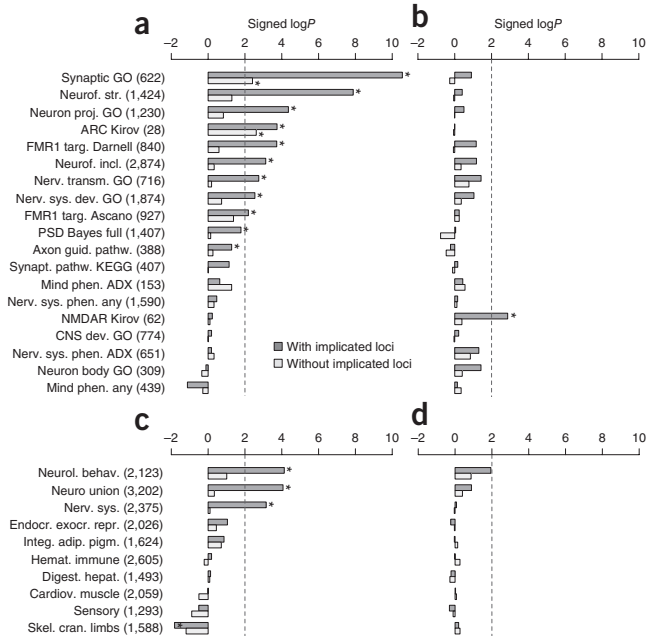


**Figure 2** Gene set burden. (**a**–**d**) Gene set burden test results for rare losses (**a**,**c**) and gains (**b**,**d**) in gene sets for neuronal function, synaptic components, neurological and neurodevelopmental phenotypes in human (**a**,**b**) and gene sets for human homologs of mouse genes implicated in abnormal phenotypes (organized by organ systems) (**c**,**d**), sorted by $-\log_{10}$ of the logistic regression deviance test *P* value multiplied by the beta coefficient sign, obtained for rare losses when including known loci. Asterisk denotes gene sets passing the 10% BH-FDR threshold. Gene sets representing brain expression patterns are not shown, because only a few were significant (losses, 1; gains, 3).

the consistency of overall CNV burden across genotyping platforms and investigate whether a measurable CNV burden persists outside of previously implicated CNV regions. Consistent with previous estimates, the overall CNV burden was significantly greater among SCZ cases when measured as total distance (kb) covered (OR = 1.12, $P = 5.7 \times 10^{-15}$), genes affected (OR = 1.21, $P = 6.6 \times 10^{-21}$) or CNV number (OR = 1.03, $P = 1 \times 10^{-3}$). The burden signal above was driven by CNVs located within genes. Focusing therefore on the number of genes affected by CNV, which was the burden metric with the strongest signal of enrichment in our study, the effect size was consistent across all genotyping platforms (**Fig. 1a**). When we split by CNV type, the effect size for copy number losses (OR = 1.40, $P = 4 \times 10^{-16}$) was greater than for gains (OR = 1.12, $P = 2 \times 10^{-7}$) (**Supplementary Figs. 2** and **3**). Partitioning by CNV frequency (based on 50% reciprocal overlap with the full call set; Online Methods), CNV burden was enriched among cases across a range of frequencies, up to counts of 80 (MAF = 0.4%) in the combined sample (**Fig. 1b**). CNV burden results for individual cohorts are provided in **Supplementary Figure 4**. We observed no enrichment in CNV burden when considering only variants that did not overlap exons (**Supplementary Fig. 5**).

A primary question in this study is how novel loci contribute to excess CNV burden in cases. After removing nine previously implicated CNV loci (where reported *P* values exceed our designated multiple testing threshold) (**Supplementary Table 1**), excess CNV burden in SCZ remained significantly enriched (genes affected OR = 1.11, $P = 1.3 \times 10^{-7}$) (**Fig. 1b**). CNV burden also remained significantly enriched after removal of all reported loci from **Supplementary Table 1**, but the effect size was greatly reduced (OR = 1.08) compared to the enrichment overall (OR = 1.21). When we partitioned CNV burden by frequency, we found that much of the previously unexplained signal was restricted to ultra-rare events (i.e., MAF < 0.1%) (**Fig. 1b**).

## Gene set (pathway) burden

We assessed whether CNV burden was concentrated within defined sets of genes involved in neurodevelopment or neurological function. We evaluated a total of 36 gene sets (**Supplementary Table 3**), including gene sets representing neuronal function, synaptic components and neurological and neurodevelopmental phenotypes in human (19 sets); gene sets based on brain expression patterns (7 sets) and human orthologs of mouse genes whose disruption causes phenotypic abnormalities, including neurobehavioral and nervous system abnormality (10 sets). Genes not expressed in brain (1 set) or associated with abnormal phenotypes in mouse organ systems unrelated to brain (7 sets) were included as negative controls. We mapped CNVs to genes if they overlapped by at least one exonic base pair.

Gene set burden was measured using a logistic regression deviance test[6]. In addition to using the same covariates included in genome-wide burden analysis, we controlled for the total number of genes

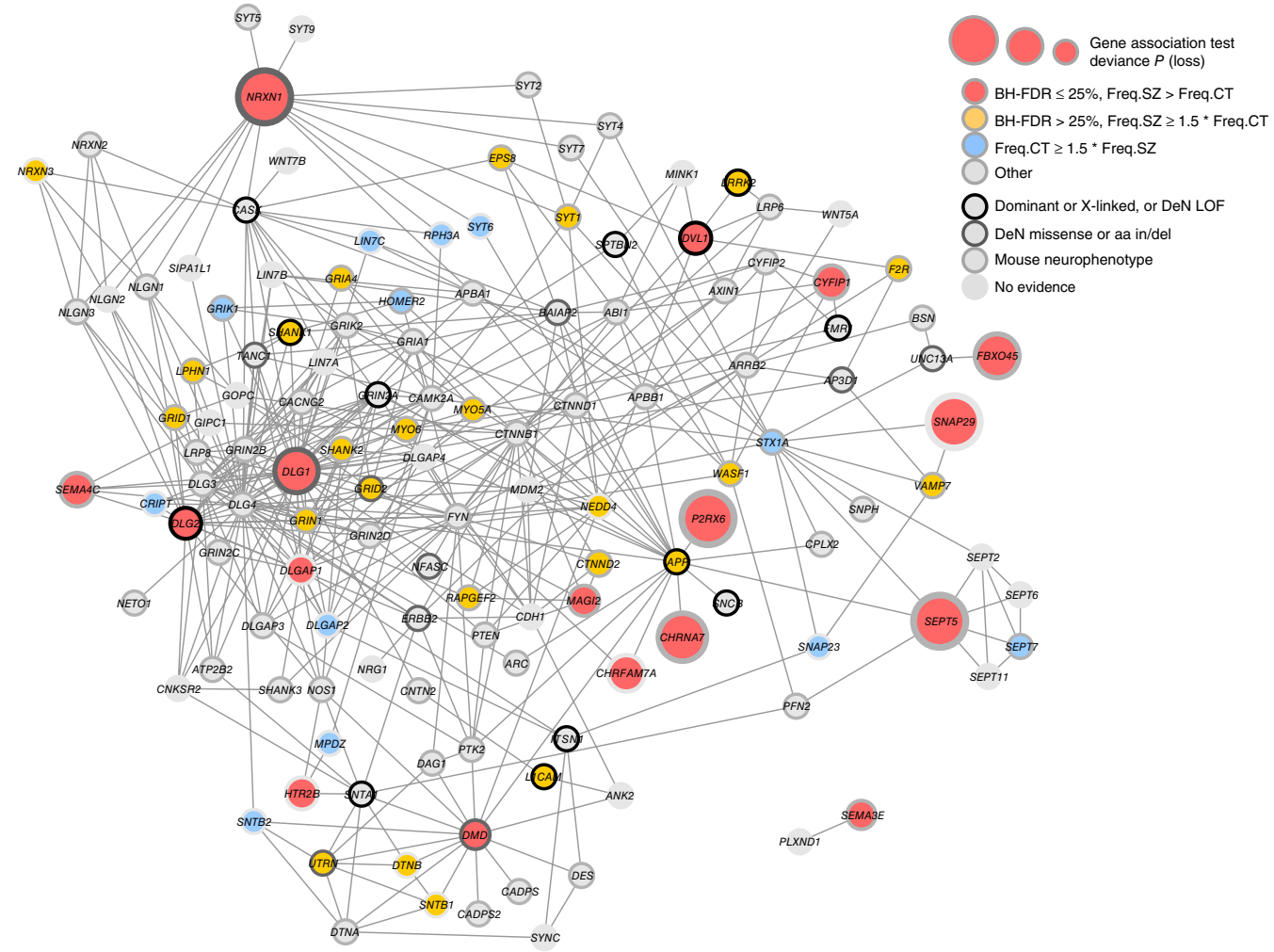


**Figure 3** Encoded-protein interaction network for synaptic genes. Synaptic and ARC-complex genes intersected by a rare loss in at least four case or control subjects and with genic burden BH-FDR 25% (red discs) were used to query GeneMANIA[36] and retrieve additional encoded-protein interaction neighbors, resulting in a network of 136 synaptic genes. Genes are depicted as disks; disk centers are colored on the basis of rare loss frequency being prevalent in cases (Freq.SZ) or controls (Freq.CT); disk borders are colored to mark gene implication in human dominant or X-linked neurological or neurodevelopmental phenotype, *de novo* mutation (DeN) reported by Fromer *et al.*[28], split between loss-of-function (LOF) (frameshift, stop-gain, core splice site) and missense or amino acid insertion or deletion (aa in/del); implication in mouse neurobehavioral abnormality. Genes encoding presynaptic adhesion molecules (*NRXN1*, *NRXN3*), postsynaptic scaffolds (*DLG1*, *DLG2*, *DLGAP1*, *SHANK1*, *SHANK2*) and glutamatergic ionotropic receptors (*GRID1*, *GRID2*, *GRIN1*, *GRIA4*) constitute a highly connected subnetwork with more losses in cases than in controls.



per subject spanned by rare CNVs to account for signal that merely reflects the global enrichment of CNV burden in cases[19]. Multiple-testing correction (Benjamini–Hochberg false discovery rate (BH-FDR)) was performed separately for each gene set group and CNV type (gains, losses). After multiple test correction (BH-FDR ≤ 10%) 15 gene sets were enriched for rare loss burden in cases and 4 for rare gains in cases, none of which were negative control sets (**Fig. 2**).

Of the 15 sets significant for losses, the majority consisted of synaptic or other neuronal components (9 sets); in particular, GO terms synapse and activity-regulated cytoskeleton-associated protein (ARC) complex rank first on the basis of statistical significance and effect size, respectively (**Fig. 2a**). Losses in cases were also significantly enriched for genes involved in nervous system or behavioral phenotypes in mouse but not for gene sets related to other organ system phenotypes (**Fig. 2c**). To account for dependency between synaptic and neuronal gene sets, we retested loss burden following a step-down logistic regression approach, ranking gene sets on the basis of significance or effect size (**Supplementary Table 4**). Only GO terms synapse and ARC complex were significant in at least one of the two step-down analyses, suggesting that burden enrichment in the other neuronal categories is captured mostly by the overlap with synaptic genes. Following the same approach, the mouse neurological and neurobehavioral phenotype set remained nominally significant ($P = 0.01$), suggesting that a portion of this signal was independent of the synaptic gene set. Pathway enrichment was less pronounced for duplications, consistent with the smaller burden effects for this class of CNV. Among synaptic or other neuronal components, duplication burden was significantly enriched only for NMDA receptor complex (**Fig. 2b**); none of the mouse phenotype sets passed the significance threshold for duplications (**Fig. 2d**).

Given that synaptic gene sets were robustly enriched for deletions in cases and showed an appreciable contribution from loci that have not been strongly associated with SCZ previously, we further investigated pathway-level interactions of these sets. A protein-interaction

**Table 1 Significant CNV loci from gene-based association test**

| Chr. | Start | End | Locus (gene) | Status | Putative mechanism | CNV test | Direction | FWER | BH-FDR | CAS | CON | Regional *P* | OR (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 17400000 | 19750000 | 22q11.21 | Previously implicated | NAHR | Loss | Risk | Yes | $3.54 \times 10^{-15}$ | 64 | 1 | $5.70 \times 10^{-18}$ | 67.7 (9.3–492.8) |
| 16 | 29560000 | 30110000 | 16p11.2, proximal | Previously implicated | NAHR | Gain | Risk | Yes | $5.82 \times 10^{-10}$ | 70 | 7 | $2.52 \times 10^{-12}$ | 9.4 (4.2–20.9) |
| 2 | 50000992 | 51113178 | 2p16.3 (*NRXN1*) | Previously implicated | NHEJ | Loss | Risk | Yes | $3.52 \times 10^{-7}$ | 35 | 3 | $4.92 \times 10^{-9}$ | 14.4 (4.2–46.9) |
| 15 | 28920000 | 30270000 | 15q13.3 | Previously implicated | NAHR | Loss | Risk | Yes | $2.22 \times 10^{-5}$ | 28 | 2 | $2.13 \times 10^{-7}$ | 15.6 (3.7–66.5) |
| 1 | 144646000 | 146176000 | 1q21.1 | Previously implicated | NAHR | Loss + gain | Risk | Yes | 0.00011 | 60 | 14 | $1.50 \times 10^{-6}$ | 3.8 (2.1–6.9) |
| 3 | 197230000 | 198840000 | 3q29 | Previously implicated | NAHR | Loss | Risk | Yes | 0.00024 | 16 | 0 | $1.86 \times 10^{-6}$ | INF |
| 16 | 28730000 | 28960000 | 16p11.2, distal | Previously reported | NAHR | Loss | Risk | Yes | 0.0029 | 11 | 1 | $5.52 \times 10^{-5}$ | 20.6 (2.6–162.2) |
| 7 | 72380000 | 73780000 | 7q11.23 | Previously reported | NAHR | Gain | Risk | Yes | 0.0048 | 16 | 1 | $1.68 \times 10^{-4}$ | 16.1 (3.1–125.7) |
| X | 153800000 | 154225000 | Xq28, distal | Novel | NAHR | Gain | Risk | No | 0.049 | 18 | 2 | $3.61 \times 10^{-4}$ | 8.9 (2.0–39.9) |
| 22 | 17400000 | 19750000 | 22q11.21 | Previously reported | NAHR | Gain | Protective | No | 0.024 | 3 | 16 | $4.54 \times 10^{-4}$ | 0.15 (0.04–0.52) |
| 7 | 64476203 | 64503433 | 7q11.21 (*ZNF92*) | Novel | NAHR | Loss + gain | Protective | No | 0.033 | 131 | 180 | $6.71 \times 10^{-4}$ | 0.66 (0.52–0.84) |
| 13 | 19309593 | 19335773 | 13q12.11 (*ZMYM5*) | Novel | NHAR | Gain | Protective | No | 0.024 | 15 | 38 | $7.91 \times 10^{-4}$ | 0.36 (0.19–0.67) |
| X | 148575477 | 148580720 | Xq28 (*MAGEA11*) | Novel | NAHR | Gain | Protective | No | 0.044 | 12 | 36 | $1.06 \times 10^{-3}$ | 0.35 (0.18–0.68) |
| 15 | 20350000 | 20640000 | 15q11.2 | Previously implicated | NAHR | Loss | Risk | No | 0.044 | 98 | 50 | $1.34 \times 10^{-3}$ | 1.8 (1.2–2.6) |
| 9 | 831690 | 959090 | 9p24.3 (*DMRT1*) | Novel | NHEJ | Loss + gain | Risk | No | 0.049 | 13 | 1 | $1.35 \times 10^{-3}$ | 12.4 (1.6–98.1) |
| 8 | 100094670 | 100958984 | 8q22.2 (*VPS13B*) | Novel | NHEJ | Loss | Risk | No | 0.048 | 7 | 1 | $1.74 \times 10^{-3}$ | 14.5 (1.7–122.2) |
| 7 | 158145959 | 158664998 | 7p36.3 (*VIPR2*; *WDR60*) | Previously reported | NAHR | Loss + gain | Risk | No | 0.046 | 20 | 6 | $5.79 \times 10^{-3}$ | 3.5 (1.3–9.0) |

All 17 association signals listed contain at least 1 gene with BH-FDR < 0.05 in the gene-based test, with 8 containing at least 1 gene surpassing the FWER < 0.05. Genomic positions listed are based on hg18 coordinates. For putative CNV mechanisms, NAHR and nonhomologous end joining (NHEJ) are listed as the likely genomic feature driving CNV formation at each locus. Regional *P* values and ORs are from a regional test at each locus, where we combine CNV overlapping the implicated region and run the same test as used for each gene (logistic regression with covariates and deviance test *P* value). CNV losses and gains at 22q11.21 are listed as separate association signals, as CNV losses associate with SCZ risk, whereas CNV gains associate with protection from SCZ. For each association we indicate whether it was previously described in the literature (previously reported) and whether the reported *P* value exceeded the multiple testing correction in this study (previously implicated). CAS, cases; CON, controls.

network was seeded using the synaptic- and ARC complex–associated genes that were intersected by rare deletions in this study (**Fig. 3**). A graph of the network highlights multiple subnetworks of synaptic proteins including presynaptic adhesion molecules (NRXN1, NRXN3), postsynaptic scaffolding proteins (DLG1, DLG2, DLGAP1, SHANK1, SHANK2), glutamatergic ionotropic receptors (GRID1, GRID2, GRIN1, GRIA4), and complexes such as dystrophin and its synaptic interacting proteins (DMD, DTNB, SNTB1, UTRN). A subsequent test of the dystrophin glycoprotein complex (DGC) showed that deletion burden of the synaptic DGC proteins (intersection of GO terms DGC and synapse was enriched in cases (deviance test $P = 0.05$), but deletion burden of the full DGC was not significant ($P = 0.69$).

## Gene–CNV association

To define specific loci that confer risk for SCZ, we tested CNV association at the level of individual genes, using logistic regression deviance test and the same covariates included in genome-wide burden analysis. To correctly account for large CNVs that affect multiple genes, we aggregated adjacent genes into a single locus if their copy number was highly correlated across subjects (more than 50% subject overlap). CNVs were mapped to genes if they overlapped one or more exons. The criterion for genome-wide significance was a family-wise error rate (FWER) < 0.05. The criterion for suggestive evidence was a BH-FDR < 0.05.

Of 18 independent CNV loci with gene-based BH-FDR < 0.05, two were excluded on the basis of CNV calling accuracy or evidence of a batch effect (**Supplementary Note**). The 16 loci that remained after these additional QC steps, comprising 17 separate association signals, are listed in **Table 1**. *P* values for this summary table were obtained by re-running our statistical model across the entire region (**Supplementary Note**). These 16 loci represent a set of novel ($n = 6$), previously reported ($n = 4$) and previously implicated ($n = 7$) regions, with 22q11.21 comprising two separate association signals at the same locus. Manhattan plots of the gene association for losses and gains are shown in **Figure 4**. A permutation-based FDR yielded similar estimates to BH-FDR.

Eight loci attain genome-wide significance, including copy number losses at 1q21.1, 2p16.3 (*NRXN1*), 3q29, 15q13.3, 16p11.2 (distal) and 22q11.2 along with gains at 7q11.23 and 16p11.2 (proximal). An additional eight loci met criteria for suggestive association, including six that have not been reported previously in association with SCZ. On the basis of our estimation of FDR values (BH and permutations), we expect to observe <2 associations meeting suggestive criteria by chance. To further evaluate the six candidate loci identified here, we performed experimental validation of CNV calls in a subset of samples by digital droplet PCR (ddPCR; Online Methods). Validation rates of 100% were obtained for gains of *DMRT1*, *MAGEA11* and distal Xq28, losses of *VPS13B* and gains and losses of *ZNF92* (**Supplementary Table 5**). We obtained a low validation rate at one locus, *ZMYM5* (64%) and therefore do not consider the association at this locus convincing.

## Breakpoint-level CNV association

With our sample size and uniform CNV calling pipeline, many individual CNV loci can be tested with adequate power at the CNV breakpoint level (i.e., the SNP probe defining the start and end of the CNV segment), potentially facilitating discovery at a finer resolution than locus-wide tests. Tests for association were performed at each CNV breakpoint using the residuals of case-control status after controlling for analysis covariates, with significance determined through permutation. Results for losses and gains are shown in **Supplementary**

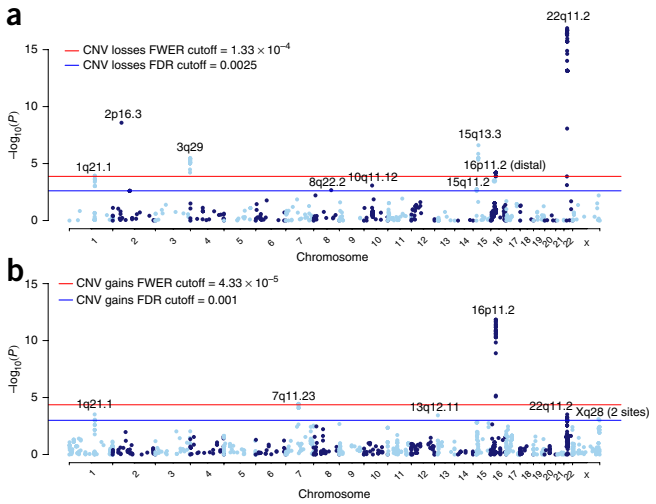


**Figure 4** Gene-based Manhattan plot. (**a**,**b**) Manhattan plot displaying the $-\log_{10}$ deviance *P* value for CNV losses (**a**) and CNV gains (**b**) in the gene-based test. *P* value cutoffs corresponding to FWER < 0.05 and BH-FDR < 0.05 are highlighted in red and blue, respectively. Loci significant after multiple test correction are labeled.

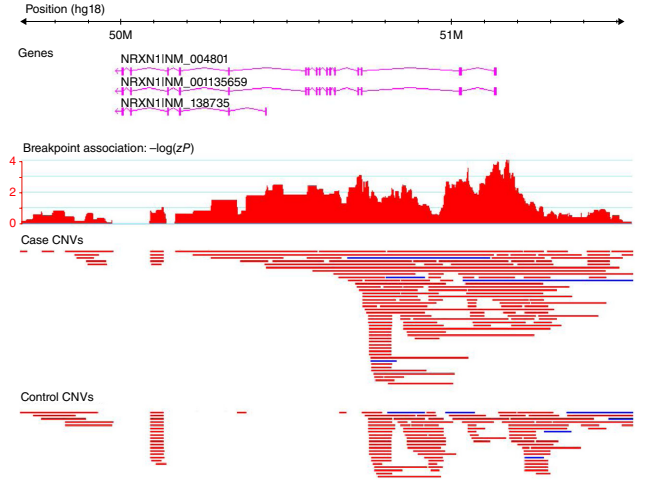


**Figure 5** Manhattan plot of breakpoint-level associations across the *NRXN1* locus. The Manhattan plot (for deletions) represents empirical *P* values at each deletion breakpoint. CNV tracks display duplications (blue) and deletions (red) detected in cases and controls from the PGC SCZ data set.

**Figure 6**. Four independent CNV loci surpass genome-wide significance, all of which were also identified in the gene-based test, including the 15q13.2–13.3 and 22q11.21 deletions, 16p11.2 duplication, and 1q21.1 deletion and duplication. While these loci represent fewer than half of the loci previously implicated in SCZ, we do find support for all loci where the association originally reported meets the criteria for genome-wide correction in this study. We examined association among all previously reported loci showing association to SCZ, including 18 CNV losses and 25 CNV gains (**Supplementary Table 6**); 8 loci have BH-FDR *q*-value < 0.05, 13 loci have BH-FDR *q*-value < 0.1, and 25 of the 42 loci were associated with SCZ at an uncorrected $P < 0.05$.

Associations at some loci become better delineated through breakpoint-level analysis. For instance, *NRXN1* at 2p16.3 is a CNV hot spot, and exonic deletions of this gene are significantly enriched in SCZ[9,20]. In this large sample, we observe a high density of 'nonrecurrent' deletion breakpoints in cases and controls. A snapshot of the breakpoint-associationassociation results from the PGC CNV browser (Online Methods) shows a sawtooth pattern of association. Predominant peaks correspond to exons and transcriptional start sites of *NRXN1* isoforms (**Fig. 5**). This example highlights how, with high diversity of alleles at a single locus, the association peak may become more refined and, in some cases, converge toward individual functional elements. Similarly, visualization of the previously reported SCZ risk loci on 16p13.2 and 8q11.23 showed a high density of duplication breakpoints, which better delineate genes in these regions. It is important, however, to note that CNV breakpoints in the current study are estimated from genotyped SNPs around the true breakpoint and that these breakpoint estimates are limited by the resolution of the genotyping platform and therefore subject to error.

### Novel risk alleles are predominantly NAHR-mediated CNVs

Many CNV loci that have been strongly implicated in human disease are hot spots for nonallelic homologous recombination (NAHR), a process that in most cases is mediated by flanking segmental duplications[21]. We defined a CNV as NAHR when both the start and end breakpoints were located within a segmental duplication. Consistent with the importance of NAHR in generating CNV risk alleles for SCZ, most of the loci in **Table 1** are flanked by segmental duplications. Moreover, after excluding loci implicated in previous studies, the remaining loci with FDR < 0.05 in the gene-base burden test were NAHR enriched (6.03-fold, $P = 0.008$; **Supplementary Fig. 7**) when compared to a null distribution determined by randomizing the genomic positions of associated genes (**Supplementary Note**). These findings suggest that the novel SCZ-associated CNVs are similar to known pathogenic CNVs in that they tend to occur in regions prone to high rates of recurrent mutation.

## DISCUSSION

The present study of the PGC SCZ CNV data set includes the majority of all microarray data that have been generated in genetic studies of SCZ to date. In this we find definitive evidence for eight loci, surpassing strict genome-wide multiple testing correction. We also find evidence for a contribution of novel CNVs conferring either risk or protection to SCZ, with an FDR < 0.05. The complete results, including CNV calls and statistical evidence at the gene or breakpoint level, can be viewed using the PGC CNV browser (Online Methods). Our data suggest that the undiscovered novel risk loci that can be detected with current genotyping platforms lie at the ultra-rare end of the frequency spectrum and still larger samples will be needed to identify them at convincing levels of statistical evidence.

Collectively, the eight SCZ risk loci that surpass genome-wide significance are carried by a small fraction (1.4%) of SCZ cases in the PGC sample. We estimate 0.85% of the variance in SCZ liability is explained by carrying a CNV risk allele within these loci (**Supplementary Note**). As a comparison, 3.4% of the variance in SCZ liability is explained by the 108 genome-wide significant loci identified in the companion PGC GWAS analysis. Combined, the CNV and SNP loci that have been identified to date explain a small proportion (<5%) of heritability. The large data set here provides an opportunity to evaluate the strength of evidence for a variety of loci where an association with SCZ has been reported previously. Of 44 published findings from the recent literature, we find evidence for eight loci (at FDR 5%) and nominal support for an additional 17 loci

(uncorrected $P < 0.05$; **Supplementary Table 6**). Thus, nearly half of the existing candidate loci retain some support in our combined analysis. However, we also find a lack of evidence for many of the previously identified loci, underscoring the value of meta-analytic efforts to assess the validity of such reports. A lack of strong evidence in this data set (which includes samples that overlap with many of the previous studies) may in some cases simply reflect that statistical power is limited for very rare variants, even in large samples. However, it is likely that some of the earlier findings represent chance associations; indeed, the loci that are not supported by our data consist largely of loci for which the original statistical evidence was weak (**Supplementary Table 6**). Thus, our results help to refine the list of promising candidate CNVs. Continued efforts to evaluate the growing number of candidate variants has considerable value for directing future research efforts focused on specific loci.

The novel candidate loci meeting suggestive criteria in this study include two regions on chromosome X. It has been hypothesized that sex-linked loci contribute to SCZ, originally on the basis of the observation of an increased rate of sex chromosome aneuploidy in cases[22]. X-linked loci were not detected in previous CNV studies of SCZ, because none evaluated variants on the sex chromosomes. In the current study, accurate calls were obtained by controlling for sex chromosome ploidy in the normalization and variant calling methods. Notably, duplications of distal Xq28 (regional $P = 3.6 \times 10^{-4}$, OR = 8.9) (**Table 1** and **Supplementary Fig. 8**) appear to confer risk for SCZ in both males and females, and the effect size was greatest in males ($P = 0.01$, OR = $\infty$). Similar patterns consistent with dominant X-linked effects were observed at other loci (**Supplementary Table 7**). Duplications of distal Xq28 have been reported in association with developmental delay in both sexes[23,24]. Notably, of 26 subjects that have been described clinically, nearly half (12/26) have behavioral or psychiatric conditions. Of the five reciprocal deletions that were detected in this study, none were observed in males, consistent with hemizygous loss of distal Xq28 being associated with recessive embryonic lethality in males[24]. Thus, mounting evidence indicates that increased copy number of distal Xq28 is associated with psychiatric illness. These results also provide a further demonstration that CNV risk factors in SCZ overlap with loci that contribute to pediatric developmental disorders[1,25].

We observed multiple 'protective' CNVs that showed a suggestive enrichment in controls, including duplications of 22q11.2 and *MAGEA11* along with deletions and duplications of *ZNF92*. No protective effects were significant after genome-wide correction. Moreover, a rare CNV that confers reduced risk for SCZ may not confer a general protection from neurodevelopmental disorders. For example, microduplications of 22q11.2 appear to confer protection from SCZ[26]; however, such duplications have been shown to increase risk for developmental delay and a variety of congenital anomalies in pediatric clinical populations[27]. It is probable that some of the undiscovered rare alleles affecting risk for SCZ confer protection, but larger sample sizes are needed to determine this unequivocally. If it is true that a proportion of CNVs observed in our control sample represent rare protective alleles, then the heritability of SCZ explained by CNVs may not be fully accounted for by the excess CNV burden in cases.

Our results provide strong evidence that deletions in SCZ are enriched within a highly connected network of synaptic proteins, consistent with previous studies[2,6,10,28]. The large CNV data set here allows a more detailed view of the synaptic network and highlights subsets of genes accounting for the excess deletion burden in SCZ, including those affecting synaptic cell adhesion and scaffolding proteins, glutamatergic ionotropic receptors and protein complexes such as the ARC complex and DGC. Modest CNV evidence implicating dystrophin (DMD) and its binding partners is notable, given that the involvement of certain components of the DGC have been postulated[29,30] and disputed[31] previously. Larger studies of CNV are needed to define a role for this and other synaptic subnetworks in SCZ.

Our current study is well powered to detect CNVs of large effect that occur in >0.1% of cases but is underpowered to detect association to variants with modest effect sizes or to ultra-rare variants regardless of effect size. Furthermore, this study did not assess the contribution of common CNVs to SCZ, one instance of which we know: a recent study demonstrated that the causal variants underlying the strongest common variant association in SCZ include duplications of the gene encoding complement factor 4A[32]. Last, we recognize that a majority of structural variants are not detectable with current genotyping platforms[33]. New technologies for whole-genome sequencing will ultimately provide an assessment of the contribution of a wider array of rare variants, including balanced rearrangements, small CNVs[34] and short tandem repeats[35].

Large-scale collaborations in psychiatric genetics have greatly advanced discovery through genome-wide association studies. Here we have extended this framework to rare CNVs. Our knowledge of the contribution from lower-frequency variants gives us confidence that the application of this framework to large newly acquired data sets has the potential to further the discovery of loci and identification of the relevant genes and functional elements.

## METHODS

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

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

### ACKNOWLEDGMENTS

Core funding for the Psychiatric Genomics Consortium is from the US National Institute of Mental Health (NIMH, U01 MH094421). We thank T. Lehner, A. Addington and G. Senthil (NIMH). The work of the contributing groups was supported by numerous grants from governmental and charitable bodies as well as philanthropic donation. Details are provided in the **Supplementary Note**.

### AUTHOR CONTRIBUTIONS

Management of the study, core analyses and content of the manuscript was the responsibility of the CNV Analysis Group, chaired by J. Sebat and jointly supervised by S.W.S. and B.M.N. together with the Schizophrenia Working Group, chaired by M.C.O'D. Core analyses were carried out by D.P.H., D. Merico, and C.R.M. Data Processing pipeline was implemented by C.R.M., B.T., W.W., D.S.G., M. Gujral, A. Shetty, and W.B. The A custom PGC CNV browser was developed by C.R.M., D.P.H. and B.T. Additional analyses and interpretations were contributed by W.W., D.A. and P.A.H. The individual studies or consortia contributing to the CNV meta-analysis were led by R.A., O.A.A., D.H.R.B., E. Bramon, J.D.B., A.C., D.A.C., S.C., A.D., E. Domenici, T.E., P.V.G., M.G., H.G., C.M.H., N.I., A.V.J., E.G.J., K.S.K., G.K., J. Knight, D.F.L., Q.S.L., J. Liu, S.A.M., A. McQuillin, J.L.M., B.J.M., M.M.N., M.C.O'D., R.A.O., M.J.O., A. Palotie, C.N.P., T.L.P., M.R., B.P.R., D.R., P. Sklar, D.S.C., P.F.S., J.T.R.W. and T.W. The remaining authors contributed to the recruitment, genotyping, or data processing for the contributing components of the meta-analysis. J. Sebat, B.M.N., M.C.O'D., C.R.M., D.P.H., and D. Merico drafted the manuscript, which was shaped by the management group. All other authors saw, had the opportunity to comment on and approved the final draft.

### COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
2. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
3. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
4. Malhotra, D. *et al.* High frequencies of *de novo* CNVs in bipolar disorder and schizophrenia. *Neuron* **72**, 951–963 (2011).
5. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**, 880–885 (2008).
6. 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).
7. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
8. Mulle, J.G. *et al.* Microdeletions of 3q29 confer high risk for schizophrenia. *Am. J. Hum. Genet.* **87**, 229–236 (2010).
9. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet.* **18**, 988–996 (2009).
10. Pocklington, A.J. *et al.* Novel findings from CNVs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron* **86**, 1203–1214 (2015).
11. Horev, G. *et al.* Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc. Natl. Acad. Sci. USA* **108**, 17076–17081 (2011).
12. Golzio, C. *et al. KCTD13* is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485**, 363–367 (2012).
13. Holmes, A.J. *et al.* Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk. *J. Neurosci.* **32**, 18087–18100 (2012).
14. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
15. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
16. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
17. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
18. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene *VIPR2* confer significant risk for schizophrenia. *Nature* **471**, 499–503 (2011).
19. Raychaudhuri, S. *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.* **6**, e1001097 (2010).
20. Kirov, G. *et al.* Comparative genome hybridization suggests a role for *NRXN1* and *APBA2* in schizophrenia. *Hum. Mol. Genet.* **17**, 458–465 (2008).
21. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
22. DeLisi, L.E. *et al.* Schizophrenia and sex chromosome anomalies. *Schizophr. Bull.* **20**, 495–505 (1994).
23. El-Hattab, A.W. *et al.* Int22h-1/int22h-2-mediated Xq28 rearrangements: intellectual disability associated with duplications and in utero male lethality with deletions. *J. Med. Genet.* **48**, 840–850 (2011).
24. El-Hattab, A.W. *et al.* Clinical characterization of int22h1/int22h2-mediated Xq28 duplication/deletion: new cases and literature review. *BMC Med. Genet.* **16**, 12 (2015).
25. Sebat, J., Levy, D.L. & McCarthy, S.E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25**, 528–535 (2009).
26. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol. Psychiatry* **19**, 37–40 (2014).
27. Van Campenhout, S. *et al.* Microduplication 22q11.2: a description of the clinical, developmental and behavioral characteristics during childhood. *Genet. Couns.* **23**, 135–148 (2012).
28. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
29. Zatz, M. *et al.* Cosegregation of schizophrenia with Becker muscular dystrophy: susceptibility locus for schizophrenia at Xp21 or an effect of the dystrophin gene in the brain? *J. Med. Genet.* **30**, 131–134 (1993).
30. Straub, R.E. *et al.* Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am. J. Hum. Genet.* **71**, 337–348 (2002).
31. Mutsuddi, M. *et al.* Analysis of high-resolution HapMap of *DTNBP1* (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. *Am. J. Hum. Genet.* **79**, 903–909 (2006).
32. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
33. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
34. Brandler, W.M. *et al.* Frequency and complexity of de novo structural mutation in autism. *Am. J. Hum. Genet.* **98**, 667–679 (2016).
35. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
36. Zuberi, K. *et al.* GeneMANIA prediction server 2013 update. *Nucleic Acids Res.* **41**, W115–W122 (2013).



**CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium**

**Christian R Marshall[1,175], Daniel P Howrigan[2,3,175], Daniele Merico[1,175], Bhooma Thiruvahindrapuram[1], Wenting Wu[4,5], Douglas S Greer[4,5], Danny Antaki[4,5], Aniket Shetty[4,5], Peter A Holmans[6,7], Dalila Pinto[8,9], Madhusudan Gujral[4,5], William M Brandler[4,5], Dheeraj Malhotra[4,5,10], Zhouzhi Wang[1], Karin V Fuentes Fajarado[4,5], Michelle S Maile[4,5], Stephan Ripke[2,3], Ingrid Agartz[11–13], Margot Albus[14], Madeline Alexander[15], Farooq Amin[16,17], Joshua Atkins[18,19], Silviu A Bacanu[20], Richard A Belliveau Jr[3], Sarah E Bergen[3,21], Marcelo Bertalan[22,23], Elizabeth Bevilacqua[3], Tim B Bigdeli[20], Donald W Black[24], Richard Bruggeman[25], Nancy G Buccola[26], Randy L Buckner[27–29], Brendan Bulik-Sullivan[2,3], William Byerley[30], Wiepke Cahn[31], Guiqing Cai[8,32], Murray J Cairns[18,33,34], Dominique Campion[35], Rita M Cantor[36], Vaughan J Carr[33,37], Noa Carrera[6], Stanley V Catts[33,38], Kimberley D Chambert[3], Wei Cheng[39], C Robert Cloninger[40], David Cohen[41], Paul Cormican[42], Nick Craddock[6,7], Benedicto Crespo-Facorro[43,44], James J Crowley[45], David Curtis[46,47], Michael Davidson[48], Kenneth L Davis[8], Franziska Degenhardt[49,50], Jurgen Del Favero[51], Lynn E DeLisi[52,53], Dimitris Dikeos[54], Timothy Dinan[55], Srdjan Djurovic[11,56], Gary Donohoe[42,57], Elodie Drapeau[8], Jubao Duan[58,59], Frank Dudbridge[60], Peter Eichhammer[61], Johan Eriksson[62–64], Valentina Escott-Price[6], Laurent Essioux[65], Ayman H Fanous[66–69], Kai-How Farh[2], Martilias S Farrell[45], Josef Frank[70], Lude Franke[71], Robert Freedman[72], Nelson B Freimer[73], Joseph I Friedman[8], Andreas J Forstner[49,50], Menachem Fromer[2,3,74,75], Giulio Genovese[3], Lyudmila Georgieva[6], Elliot S Gershon[59,76], Ina Giegling[77,78], Paola Giusti-Rodríguez[45], Stephanie Godard[79], Jacqueline I Goldstein[2,80], Jacob Gratten[81], Lieuwe de Haan[82], Marian L Hamshere[6], Mark Hansen[83], Thomas Hansen[22,23], Vahram Haroutunian[8,84,85], Annette M Hartmann[77], Frans A Henskens[33,34,86], Stefan Herms[49,50,87], Joel N Hirschhorn[80,88,89], Per Hoffmann[49,50,87], Andrea Hofman[49,50], Hailiang Huang[2,80], Masashi Ikeda[90], Inge Joa[91], Anna K Kähler[21], René S Kahn[31], Luba Kalaydjieva[92,93], Juha Karjalainen[71], David Kavanagh[6], Matthew C Keller[94], Brian J Kelly[34], James L Kennedy[95–97], Yunjung Kim[45], James A Knowles[69,98], Bettina Konte[77], Claudine Laurent[15,99], Phil Lee[2,3,75], S Hong Lee[81], Sophie E Legge[6], Bernard Lerer[100], Deborah L Levy[53,101], Kung-Yee Liang[102], Jeffrey Lieberman[103],**

Jouko Lönnqvist[104], Carmel M Loughland[33,34], Patrik K E Magnusson[21], Brion S Maher[105], Wolfgang Maier[106], Jacques Mallet[107], Manuel Mattheisen[23,108–110], Morten Mattingsdal[11,111], Robert W McCarley[52,53], Colm McDonald[112], Andrew M McIntosh[113,114], Sandra Meier[70], Carin J Meijer[82], Ingrid Melle[11,115], Raquelle I Mesholam-Gately[53,116], Andres Metspalu[117], Patricia T Michie[33,118], Lili Milani[117], Vihra Milanova[119], Younes Mokrab[120], Derek W Morris[42,57], Bertram Müller-Myhsok[121–123], Kieran C Murphy[124], Robin M Murray[125], Inez Myin-Germeys[126], Igor Nenadic[127], Deborah A Nertney[128], Gerald Nestadt[129], Kristin K Nicodemus[130], Laura Nisenbaum[131], Annelie Nordin[132], Eadbhard O'Callaghan[133], Colm O'Dushlaine[3], Sang-Yun Oh[134], Ann Olincy[72], Line Olsen[22,23], F Anthony O'Neill[135], Jim Van Os[126,136], Christos Pantelis[33,137], George N Papadimitriou[54], Elena Parkhomenko[8], Michele T Pato[69,98], Tiina Paunio[138], Psychosis Endophenotypes International Consortium[139], Diana O Perkins[140], Tune H Pers[80,89,141], Olli Pietiläinen[138,142], Jonathan Pimm[47], Andrew J Pocklington[6], John Powell[125], Alkes Price[80,143], Ann E Pulver[129], Shaun M Purcell[74], Digby Quested[144], Henrik B Rasmussen[22,23], Abraham Reichenberg[8,85], Mark A Reimers[20], Alexander L Richards[6,7], Joshua L Roffman[28,29], Panos Roussos[74,145], Douglas M Ruderfer[6,74], Veikko Salomaa[63], Alan R Sanders[58,59], Adam Savitz[146], Ulrich Schall[33,34], Thomas G Schulze[70,147], Sibylle G Schwab[148], Edward M Scolnick[3], Rodney J Scott[18,33,149], Larry J Seidman[53,116], Jianxin Shi[150], Jeremy M Silverman[8,151], Jordan W Smoller[3,75], Erik Söderman[13], Chris C A Spencer[152], Eli A Stahl[74,80], Eric Strengman[31,153], Jana Strohmaier[70], T Scott Stroup[103], Jaana Suvisaari[104], Dragan M Svrakic[40], Jin P Szatkiewicz[45], Srinivas Thirumalai[154], Paul A Tooney[18,33,34], Juha Veijola[155,156], Peter M Visscher[81], John Waddington[157], Dermot Walsh[158], Bradley T Webb[20], Mark Weiser[48], Dieter B Wildenauer[159], Nigel M Williams[6], Stephanie Williams[45], Stephanie H Witt[70], Aaron R Wolen[20], Brandon K Wormley[20], Naomi R Wray[81], Jing Qin Wu[18,33], Clement C Zai[95,96], Rolf Adolfsson[132], Ole A Andreassen[11,115], Douglas H R Blackwood[113], Elvira Bramon[160], Joseph D Buxbaum[8,32,85,161], Sven Cichon[49,50,87,162], David A Collier[120,163], Aiden Corvin[42], Mark J Daly[2,3,80], Ariel Darvasi[164], Enrico Domenici[10,165], Tõnu Esko[80,88,89,117], Pablo V Gejman[58,59], Michael Gill[42], Hugh Gurling[47], Christina M Hultman[21], Nakao Iwata[90], Assen V Jablensky[33,93,159,166], Erik G Jönsson[11,13], Kenneth S Kendler[20], George Kirov[6], Jo Knight[95–97], Douglas F Levinson[15], Qingqin S Li[146], Steven A McCarroll[3,88], Andrew McQuillin[47], Jennifer L Moran[3], Bryan J Mowry[81,128], Markus M Nöthen[49,50], Roel A Ophoff[31,36,73], Michael J Owen[6,7], Aarno Palotie[3,75,142], Carlos N Pato[69,98], Tracey L Petryshen[3,29,53,167], Danielle Posthuma[168–170], Marcella Rietschel[70], Brien P Riley[20], Dan Rujescu[77,78], Pamela Sklar[74,85,145], David St Clair[171], James T R Walters[6], Thomas Werge[22,23,172], Patrick F Sullivan[21,45,140], Michael C O'Donovan[6,7,176], Stephen W Scherer[1,173,176], Benjamin M Neale[2,3,75,80,176] & Jonathan Sebat[4,5,174,176]

[1]Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Beyster Center for Psychiatric Genomics, University of California, San Diego, La Jolla, California, USA. [5]Department of Psychiatry, University of California, San Diego, La Jolla, California, USA. [6]MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [7]National Centre for Mental Health, Cardiff University, Cardiff, UK. [8]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [9]Department of Genetics and Genomic Sciences, Seaver Autism Center, Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [10]Neuroscience Discovery and Translational Area, Pharma Research and Early Development, F. Hoffmann-La Roche, Ltd, Basel, Switzerland. [11]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [12]Department of Psychiatry, Diakonhjemmet Hospital, Oslo, Norway. [13]Department of Clinical Neuroscience, Psychiatry Section, Karolinska Institutet, Stockholm, Sweden. [14]State Mental Hospital, Haar, Germany. [15]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California, USA. [16]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia, USA. [17]Department of Psychiatry and Behavioral Sciences, Atlanta Veterans Affairs Medical Center, Atlanta, Georgia, USA. [18]School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan, New South Wales, Australia. [19]Hunter Medical Research Institute, New Lambton, New South Wales, Australia. [20]Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA. [21]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [22]Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, Copenhagen, Denmark. [23]Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus Denmark. [24]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, Iowa, USA. [25]Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [26]School of Nursing, Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA. [27]Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA. [28]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA. [29]Athinoula A. Martinos Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [30]Department of Psychiatry, University of California at San Francisco, San Francisco, California, USA. [31]Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, the Netherlands. [32]Department of Human Genetics, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [33]Schizophrenia Research Institute, Sydney, New South Wales, Australia. [34]Priority Centre for Translational Neuroscience and Mental Health, University of Newcastle, Newcastle, New South Wales, Australia. [35]Centre Hospitalier du Rouvray and INSERM U1079, Faculty of Medicine, Rouen, France. [36]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [37]School of Psychiatry, University of New South Wales, Sydney, New South Wales, Australia. [38]Department of Psychiatry, Royal Brisbane and Women's Hospital, University of Queensland, Brisbane, Queensland, Australia. [39]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [40]Department of Psychiatry, Washington University, St. Louis, Missouri, USA. [41]Department of Child and Adolescent Psychiatry, Assistance Publique–Hospitaux de Paris, Pierre and Marie Curie Faculty of Medicine and Institute for Intelligent Systems and Robotics, Paris, France. [42]Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. [43]Instituto de Formación e Investigación Marqués de Valdecilla, University Hospital Marqués de Valdecilla, University of Cantabria, Santander, Spain. [44]Centro Investigación Biomédica en Red Salud Mental, Madrid, Spain. [45]Department of Genetics, University of

North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [46]Department of Psychological Medicine, Queen Mary University of London, London, UK. [47]Molecular Psychiatry Laboratory, Division of Psychiatry, University College London, London, UK. [48]Department of Psychiatry, Sheba Medical Center, Tel Hashomer, Israel. [49]Institute of Human Genetics, University of Bonn, Bonn, Germany. [50]Department of Genomics, Life and Brain Center, Bonn, Germany. [51]Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, University of Antwerp, Antwerp, Belgium. [52]Virginia Boston Health Care System, Brockton, Massachusetts, USA. [53]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA. [54]First Department of Psychiatry, University of Athens Medical School, Athens, Greece. [55]Department of Psychiatry, University College Cork, Cork, Ireland. [56]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. [57]Cognitive Genetics and Therapy Group, School of Psychology and Discipline of Biochemistry, National University of Ireland Galway, Galway, Ireland. [58]Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, Illinois, USA. [59]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois, USA. [60]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [61]Department of Psychiatry, University of Regensburg, Regensburg, Germany. [62]Folkhälsan Research Center and Biomedicum Helsinki, Helsinki, Finland. [63]National Institute for Health and Welfare, Helsinki, Finland. [64]Department of General Practice, Helsinki University Central Hospital, University of Helsinki, Helsinki, Finland. [65]Translational Technologies and Bioinformatics, Pharma Research and Early Development, F. Hoffman-La Roche, Basel, Switzerland. [66]Mental Health Service Line, Washington Virginia Medical Center, Washington, DC, USA. [67]Department of Psychiatry, Georgetown University, Washington, DC, USA. [68]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA. [69]Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, California, USA. [70]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany. [71]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [72]Department of Psychiatry, University of Colorado Denver, Aurora, Colorado, USA. [73]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, California, USA. [74]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [75]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [76]Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. [77]Department of Psychiatry, University of Halle, Halle, Germany. [78]Department of Psychiatry, University of Munich, Munich, Germany. [79]Departments of Psychiatry, and Human and Molecular Genetics, INSERM, Institut de Myologie, Hôpital de la Pitiè-Salpêtrière, Paris, France. [80]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [81]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [82]Department of Psychiatry, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands. [83]Illumina, Inc., La Jolla, California, USA. [84]J.J. Peters Virginia Medical Center, Bronx, New York, USA. [85]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [86]School of Electrical Engineering and Computer Science, University of Newcastle, Newcastle, New South Wales, Australia. [87]Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland. [88]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [89]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, USA. [90]Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Japan. [91]Regional Centre for Clinical Research in Psychosis, Department of Psychiatry, Stavanger University Hospital, Stavanger, Norway. [92]Centre for Medical Research, University of Western Australia, Perth, Western Australia, Australia. [93]Perkins Institute for Medical Research, University of Western Australia, Perth, Western Australia, Australia. [94]Department of Psychology, University of Colorado Boulder, Boulder, Colorado, USA. [95]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [96]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [97]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada. [98]Zilkha Neurogenetics Institute, Keck School of Medicine at the University of Southern California, Los Angeles, California, USA. [99]Department of Child and Adolescent Psychiatry, Pierre and Marie Curie Faculty of Medicine, Paris, France. [100]Department of Psychiatry, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [101]Psychology Research Laboratory, McLean Hospital, Belmont, Massachusetts, USA. [102]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA. [103]Department of Psychiatry, Columbia University, New York, New York, USA. [104]Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, Helsinki, Finland. [105]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [106]Department of Psychiatry, University of Bonn, Bonn, Germany. [107]CNRS, Laboratoire de Génétique Moléculaire de la Neurotransmission et des Processus Neurodégénératifs, Hôpital de la Pitié-Salpêtrière, Paris, France. [108]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [109]Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark. [110]Department of Genomics Mathematics, University of Bonn, Bonn, Germany. [111]Research Unit, Sørlandet Hospital, Kristiansand, Norway. [112]Department of Psychiatry, National University of Ireland Galway, Galway, Ireland. [113]Division of Psychiatry, University of Edinburgh, Edinburgh, UK. [114]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK. [115]Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [116]Massachusetts Mental Health Center Public Psychiatry Division of the Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [117]Estonian Genome Center, University of Tartu, Tartu, Estonia. [118]School of Psychology, University of Newcastle, Newcastle, New South Wales, Australia. [119]First Psychiatric Clinic, Medical University, Sofia, Bulgaria. [120]Eli Lilly and Company, Ltd., Windlesham, UK. [121]Max Planck Institute of Psychiatry, Munich, Germany. [122]Institute of Translational Medicine, University of Liverpool, Liverpool, UK. [123]Cluster for Systems Neurology (SyNergy), Munich, Germany. [124]Department of Psychiatry, Royal College of Surgeons in Ireland, Dublin, Ireland. [125]Institute of Psychiatry, King's College London, London, UK. [126]Maastricht University Medical Centre, South Limburg Mental Health Research and Teaching Network, EURON, Maastricht, the Netherlands. [127]Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany. [128]Queensland Centre for Mental Health Research, University of Queensland, Brisbane, Queensland, Australia. [129]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [130]Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. [131]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana, USA. [132]Department of Clinical Sciences, Psychiatry, Umeå University, Umeå, Sweden. [133]DETECT Early Intervention Service for Psychosis, Blackrock, Ireland. [134]Lawrence Berkeley National Laboratory, University of California at Berkeley, Berkeley, California, USA. [135]Centre for Public Health, Institute of Clinical Sciences, Queen's University Belfast, Belfast, UK. [136]Institute of Psychiatry, King's College London, London, UK. [137]Melbourne Neuropsychiatry Centre, University of Melbourne and Melbourne Health, Melbourne, Victoria, Australia. [138]Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland. [139]A list of members and affiliations appears in the **Supplementary Note**. [140]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [141]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Denmark. [142]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Helsinki, Finland. [143]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [144]Department of Psychiatry, University of Oxford, Oxford, UK. [145]Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [146]Neuroscience Therapeutic Area, Janssen Research and Development, Raritan, New Jersey, USA. [147]Department of Psychiatry and Psychotherapy, University of Göttingen, Göttingen, Germany. [148]Psychiatry and Psychotherapy Clinic, University of Erlangen, Erlangen, Germany. [149]Hunter New England Health Service, Newcastle, New South Wales, Australia. [150]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA. [151]Research and Development, Bronx Veterans Affairs Medical Center, New York, New York, USA. [152]Wellcome Trust Centre for Human Genetics, Oxford, UK. [153]Department of Medical Genetics, University Medical Centre Utrecht, Utrecht, the Netherlands. [154]Berkshire Healthcare NHS Foundation Trust, Bracknell, UK. [155]Department of Psychiatry, University of Oulu, Oulu, Finland. [156]Department of Psychiatry, University Hospital of Oulu, Oulu, Finland. [157]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin, Ireland. [158]Health Research Board, Dublin, Ireland. [159]School of Psychiatry and Clinical Neurosciences, University of Western Australia, Perth, Western Australia, Australia. [160]Division of Psychiatry, University College London, London, UK. [161]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [162]Institute of Neuroscience and Medicine (INM-1), Research Center Juelich, Juelich, Germany. [163]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK. [164]Department of Genetics, Hebrew University of Jerusalem, Jerusalem, Israel. [165]Centre for Integrative Biology, University of Trento, Trento, Italy. [166]Centre for Clinical Research in Neuropsychiatry, School of Psychiatry and Clinical Neurosciences, University of Western Australia, Perth, Western Australia, Australia. [167]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [168]Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, Amsterdam, the Netherlands. [169]Department of Complex Trait Genetics, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam, the Netherlands. [170]Department of Child and Adolescent Psychiatry, Erasmus University Medical Centre, Rotterdam, the Netherlands. [171]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK. [172]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [173]Department of Molecular Genetics and McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada. [174]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California, USA. [175]These authors contributed equally to this work. [176]These authors jointly directed this work. Correspondence should be addressed to J. Sebat (sebat@ucsd.edu).

## ONLINE METHODS

**Overview.** We assembled a CNV analysis group with the goal of leveraging the extensive expertise within the Psychiatric Genomics Consortium (PGC) to develop a fully automated centralized pipeline for consistent and systematic calling of CNVs for both Affymetrix and Illumina platforms. An overview of the analysis pipeline is shown in **Supplementary Figure 1**. After an initial data formatting step, we constructed batches of samples for processing using PennCNV, iPattern, C-score (GADA and HMMSeg) and Birdsuite for Affymetrix 6.0. For Affymetrix 5.0 data we used Birdsuite and PennCNV; for Affymetrix 500 we used PennCNV and C-score; and for all Illumina arrays we used PennCNV and iPattern. We then constructed a consensus CNV call data set by merging data at the sample level and further filtered calls to make a final data set (**Supplementary Table 2**). Prior to any filtering, we processed raw genotype calls for a total of 57,577 individuals, including 28,684 SCZ cases and 28,893 controls.

**Study sample.** A complete list of data sets included in the current study can be found in **Supplementary Table 2**. A more detailed description of the original studies can be found in a previous publication[1].

**CNV Analysis pipeline architecture and sample processing.** All aspects of the CNV analysis pipeline were built on the Genetic Cluster Computer (GCC) (see below).

*Input acceptance and preprocessing.* For Affymetrix we used CEL files (all converted to the same format) as input, whereas for Illumina we required Genome or Beadstudio exported TXT files with the following values: sample ID, SNP name, chr, position, allele1 – forward, allele2 – forward, X, Y, B allele freq and log R ratio. Samples were then partitioned into 'batches' to be run through each pipeline. For Affymetrix samples, we created analysis batches on the basis of the plate ID (if available) or genotyping date. Each batch had approximately 200 samples. Each batch included at least 50 subjects of each sex. Affymetrix Power Tools (APT - apt-copynumber-workflow) was then used to calculate summary statistics about chips analyzed. Gender mismatches were identified and excluded as were experiments with MAPD > 0.4. For Illumina data, we first determined the genome build and converted to hg18 if necessary and created analysis batches on the basis of the plate ID or genotyping date.

*Composite pipeline.* The composite pipeline comprises CNV callers PennCNV[2], iPattern[3], Birdsuite[4] and C-Score[5] organized into component pipelines. We used all four callers for Affymetrix 6.0 data, and we used PennCNV and C-Score for Affymetrix 500. Probe annotation files were preprocessed for each platform. Once the array design files and probe annotation files were preprocessed, each individual pipeline component pipeline was run in two steps: (i) processing the intensity data by the core pipeline process to produce CNV calls and (ii) parsing the specific output format of the core pipeline and converting the calls to a standard form designed to capture confidence scores, copy number states and other information computed by each pipeline.

Merging of CNV data and QC filtering is described in detail in the **Supplementary Note**. Briefly, for each subject, CNV calls were made using multiple algorithms. Only CNV calls that were made using multiple algorithms were included in the call set. Sample level QC filtering was performed by removing arrays with excessive probe variance or GC bias and removal of samples with mismatches in gender or ethnicity or chromosomal aneuploidies. The final filtered CNV data set was annotated with Refseq genes (transcriptions and exons). After this stage of QC, we had a total of 52,511 individuals, with 27,034 SCZ cases and 25,448 controls. To make our final data set of rare CNVs for all subsequent analysis we filtered out variants that were present at ≥1% (50% reciprocal overlap) frequency in cases and controls combined. We included in the call set CNVs that were ≥20 kb and ≥10 probes in length and overlapped <50% with regions tagged as copy number polymorphic on any other platform.

To minimize the impact of technical artifacts and potential confounds on CNV association results, we removed from the data set individuals that did not pass QC filtering from the companion PGC GWAS study of schizophrenia[1] as well as well as case or control samples that could not be matched by array platform or reconciled by using a common set of probes.

**Statistics.** *Regression of potential confounds on case-control ascertainment.* The PGC cohorts are a combination of many data sets drawn from the US and Europe, and it is important to ensure that any bias in sample ascertainment does not drive spurious association to SCZ. In order to ensure the robustness of the analysis, burden and gene set analyses included potential confounding variables as covariates in a logistic regression framework. Owing to the number of tests run at breakpoint-level association, we employed a step-wise logistic regression approach to allow for the inclusion of covariates in our case-control association, which we term the SCZ residual phenotype.

Covariates included sex, genotyping platform and ancestry principal components derived from SNP genotypes on the same samples in a previous study[1]. Control for population stratification is described in the **Supplementary Note**. We were unable to control for data set or genotyping batch, as a subset of the contributing data sets are fully confounded with case-control status. Only principal components that showed a significant association to small CNV burden were used (small CNV being defined as autosomal CNV burden with CNV < 100 kb). Among the top 20 principal components, only the first, second, third, fourth and eighth principal components showed association with small CNV burden (with $P < 0.01$ used as the significance cutoff).

Last, in order to control for case-control differences in CNV ascertainment due to data quality we sought to identify data quality metrics that were confounded with case status. Affymetrix (MAPD and waviness-sd) and Illumina (LRRSD, BAFSD, GCWF) QC metrics were re-examined across studies to assess whether any additional outliers were present. Three outliers were removed, as their mean B allele (or minor allele) frequency deviated significantly from 0.5. Many CNV metrics are autocorrelated, as they measure similar patterns of variation in the probe intensity. Thus, we focused on the primary measure of probe variance—MAPD and LRRSD. Among Affymetrix 6.0 data sets, MAPD did not differ between in cases and controls ($t = 1.14$, $P = 0.25$). However, among non-Affymetrix 6.0 data sets, LRRSD showed significant differences between cases and controls ($t = -35.3$, $P < 2 \times 10^{-16}$), with controls having a higher standardized mean LRRSD (0.227) than cases (−0.199). Thus, to control for any spurious associations driven by CNV calling quality, we included MAPD (for Affymetrix platforms) or LRRSD (for Illumina platforms) as covariates in downstream analysis, which we designate as our CNV metric covariate for each individual. Prior to inclusion in the combined data set, the CNV metric variable was normalized within each respective genotyping platform.

To calculate the SCZ residual phenotype, we first fit a logistic regression model of covariates to affection status, and then extracted the Pearson residual values for use in a quantitative association design for downstream analyses. Residual phenotype values in cases are all above 0, and controls are below 0 and are graphed against overall kilobase burden in **Supplementary Figure 9**. We removed three individuals with an SCZ residual phenotype >3 (or −3 in controls). After the post-processing round of QC, we retained a data set with a total of 41,321 individuals comprising 21,094 SCZ cases and 20,227 controls.

**CNV burden analysis.** We analyzed the overall CNV burden in a variety of ways to discern which general properties of CNV are contributing to SCZ risk. Overall individual CNV burden was measured in three distinct ways: (i) kilobase burden of CNVs, (ii) number of genes affected by CNVs and (iii) number of CNVs. Genes were counted only if the CNV overlapped a coding exon. We also partitioned our analyses by CNV type, size and frequency. CNV type is defined as copy number losses (or deletions), copy number gains (or duplications) or both copy number losses and gains. To assign a specific allele frequency to a CNV, we used the–cnv-freq-method2 command in PLINK, whereby the frequency is determined as the total number of CNVs overlapping the target CNV segment by at least 50%. This method differs from other methods that assign CNV frequencies by genomic region, whereby a single CNV spanning multiple regions may be included in multiple frequency categories.

For **Figure 1**, and **Supplementary Figures 2** and **3**, we partitioned CNV burden by genotyping platform. Owing to the small sample size of the Omni 2.5 array (28 cases and 10 controls), they were excluded from presentation in the figures but are included in all burden analyses with the total PGC sample. Using a logistic regression framework with the inclusion of covariates detailed

above, we predicted SCZ status using CNV burden as an independent predictor variable, thus enabling an accurate estimate of the contribution of CNV burden. In addition, to determine the proportion of CNV burden risk that is attributable to loci that have not been implicated in previous studies of SCZ, we ran all burden analyses after removing CNVs that overlapped previously implicated CNV boundaries by more than 10%.

**CNV breakpoint-level association.** Association was tested at each respective CNV breakpoint. Three categories of CNV were tested: deletions, duplications and deletions and duplications combined. All analyses were run in PLINK[6].

We ran breakpoint-level association using the SCZ residual phenotype as a quantitative variable, with significance determined through permutation of phenotype residual labels. An additional *z*-scoring correction, explained below, was used to control for any extreme values in the SCZ residual phenotype and efficiently estimate two-sided empirical *P* values for highly significant loci. To ensure against the potential loss of power from the inclusion of covariates, we also ran a single degree of freedom Cochran-Mantel-Haenzel (CMH) test stratified by genotyping platform, with a 2 (CNV carrier status) × 2 (phenotype status) × N (genotyping platform) contingency matrix. Although the CMH test does not account for more subtle biases that could drive false positive signals, it is robust to signals driven by a single platform and allows for each CNV carrier to be treated equally. Loci that surpassed genome-wide correction in either test were followed up for further evaluation.

*z-score recalibration of empirical testing.* Breakpoint-level association *P* values from the SCZ residual phenotype were initially obtained by performing 1 million permutations at each CNV position, wherein each permutation shuffles the SCZ residual phenotype among all samples and retains the SCZ residual mean for CNV carriers and noncarriers. For extremely rare CNVs, however, CNV carriers at the extreme ends of the SCZ residual phenotype can produce highly significant *P* values. Although we understand that such rare events are unable to surpass strict genome-wide correction, we wanted to retain all tests to help delineate the potential fine-scale architecture within a single region of association. To properly account for the increased variance when only a few individuals are tested, we applied an empirical *z*-score correction to the CNV carrier mean. In order to get an empirical estimate of the variance for each test, we calculated the s.d. of residual phenotype mean differences in CNV carriers and noncarriers from 5,000 permutations. *z*-scores are calculated as the observed case-control mean difference divided by the empirical s.d., with corresponding *P* values calculated from the standard normal distribution. Concordance of the initial empirical and *z*-score *P* values are close to unity for association tests with ≥6 CNVs, whereas *z*-score *P* values are more conservative among tests with <6 CNVs. Furthermore, the *z*-score method naturally provides an efficient manner to estimate highly significant empirical *P* values that would involve hundreds of millions of permutations to achieve. Genome-wide correction for multiple testing was determined as described in the **Supplementary Note**.

**Gene set burden enrichment analysis: gene sets.** Gene sets with an *a priori* expectation of association to neuropsychiatric disorders were compiled, and CNV calls were preprocessed as described in the **Supplementary Note**.

For each gene set, we fit the following logistic regression model (as implemented by the R function glm of the stats package), where subjects are statistical sampling units:

$$y \sim \text{covariates} + \text{global} + \text{gene set}$$

Where *y* is the dicotomic outcome variable (schizophrenia = 1, control = 0); 'covariates' is the set of variables used as covariates also in the genome-wide burden and breakpoint-association analysis (sex, genotyping platform, CNV metric, and CNV associated principal components); 'global' is the measure of global genic CNV burden (this covariate accounts for nonspecific association signal that could be merely reflective of an overall difference CNV burden between cases and controls. For the results in the main text, we used the total gene number (*U*) from universal gene set count; we also calculated results for total length (*TL*) and variant number plus variant mean length (*CNML*)); and 'gene set 'is the gene set gene count. The gene set burden enrichment was assessed by performing a chi-square deviance test (as implemented by the R function anova.glm of the stats package) comparing these two regression models:

$$y \sim \text{covariates} + \text{global}$$

$$y \sim \text{covariates} + \text{global} + \text{gene set}$$

We reported the following statistics: coefficient beta estimate (*Coeff*); *t*-student distribution-based coefficient significance *P* value (as implemented by the R function summary.glm of the stats package, abbreviated as *Pvalue_glm*); deviance test *P* value (*Pvalue_dev*); gene set size (i.e., number of genes is the gene set, regardless of CNV data); BH-FDR; percentage of schizophrenia and control subjects with at least *n* genes affected by a CNV of the desired type (loss or gain) in the gene set (*SZ_g1n, SZ_g2n, … CT_g1n, …*). By performing simple simulation analyses, we realized that *Pvalue_glm* can be extremely over-conservative in presence of very few gene set counts different from 0, whereas *Pvalue_dev* tends to be slightly under-conservative. Although the two *P* values tended to agree well for gene set analysis, *Pvalue_glm* is systematically over-conservative for gene analysis, as smaller counts are typically available for single genes.

**Gene-association analysis.** Subjects were restricted to the ones with at least one rare CNV. Only genes with at least a minimum number of subjects affected by CNV were tested; this threshold was picked by comparing the BH-FDR to the permutation-based FDR and ensuring limited FDR inflation (permuted FDR < 1.65 × BH-FDR at BH-FDR threshold = 5%) while maximizing power. For gains, the threshold was set to 12 counts, and for losses it was set to 8 counts.

For each gene, we fit the following logistic regression model (as implemented by the R function glm of the stats package), where subjects are statistical sampling units:

$$y \sim \text{covariates} + \text{gene}$$

Where *y* is the dichotomous outcome variable (schizophrenia = 1, control = 0), 'covariates' is the set of variables used as covariates also in the genome-wide burden and breakpoint-association analysis (sex, genotyping platform, CNV metric, and CNV associated principal components) and 'gene' is the binary indicator for the subject having or not having a CNV of the desired type (loss or gain) mapped to the gene. The gene burden was assessed by performing a chi-square deviance test (as implemented by the R function anova.glm of the stats package) comparing the regression models *y ~ covariates* and *y ~ covariates + gene*.

Genome-wide correction for multiple testing was determined as described in the **Supplementary Note**.

**Experimental validation of CNV calls by digital droplet PCR.** For 6 novel candidate loci that were identified in this study, we sought to confirm CNV calling accuracy by experimental validation of CNV calls in a subset of study samples. Within each association peak, a segment was defined that overlapped a majority of calls. Appropriate digital droplet assays were then selected from Bio-Rad. A single FAM-labeled probe was designed for *DMRT1*, *ZMYM5*, *ZNF92*, *MAGEA11* and distal Xq28. Because some deletions of the *VPS13B* gene were nonoverlapping, two different probes were selected for this locus. CNV calls (up to a maximum of 17) were selected from the core target region. Probe details, CNV calls and validation results can be found in **Supplementary Table 5**. Study samples were then obtained from the respective PGC studies and four population control samples were obtained from Coriell Cell repositories (ND00745, ND01936, ND00689, ND01317) to be used as negative controls for ddPCR assays. EcoRI-digested samples (10 ng genomic DNA) were analyzed in triplicate by ddPCR using the FAM-labeled CNV probe and HEX-labeled reference probe M0005 RPP30-HEX (**Supplementary Table 5**) in the UCSD CFAR Genomics and Sequencing Core. PCR droplets were generated using a Bio-Rad QX100 Droplet Generator, then quantitative PCR was performed




doi:10.1038/ng.3725

using the GeneAmp PCR system 9700 (Applied Biosystems) instrument according to manufacturer's protocols (40 cycles at 94 °C for 30 s and 60 °C for 1 min). PCR droplets were read and analyzed on Bio-Rad QX100 Droplet Reader with QuantaSoft software.

**Data availability.** Visualization of 16p13.2 is available at http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18/?name=chr16:8607047..9607046; visualization of 8q11.23 locus is available at http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18/?name=chr8:53243575..54243574; gene reviews are available at https://www.ncbi.nlm.nih.gov/books/NBK349624/. the Genetic Cluster Computer (GCC) is available at https://userinfo.surfsara.nl/systems/lisa. The PGC CNV resource is available through a custom browser at http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18/, and the rare CNV call set can be obtained from the European Genome-Phenome Archive (accession number EGAS00001001960).

# Erratum: Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects

**CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium**
***Nat. Genet.;* doi:10.1038/ng.3725; corrected online 5 December 2016**

In the version of this article initially published online, author Daniel P. Howrigan was not listed as having contributed equally to this work. The error has been corrected for the print, PDF and HTML versions of this article.