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. 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\(^\text{2-3}\), a higher rate of de novo CNVs in cases relative to controls\(^\text{4-6}\) 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\(^\text{7-9}\); 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\(^\text{10-13}\).

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 obtain 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\(^\text{14}\). 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\(^\text{15}\) and iPatten\(^\text{16}\). Data from Affymetrix platforms were analyzed using PennCNV and Birdsuite\(^\text{17}\). Two additional methods, iPattern and C-score\(^\text{18}\), 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\(^\text{2}\). 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; doi:10.1038/ng.3725

NATURE GENETICS  ADVANCE ONLINE PUBLICATION
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, $P = 5.7 \times 10^{-13}$), genes affected (OR = 1.21, $P = 6.6 \times 10^{-21}$) or CNV number (OR = 1.03, $P = 1 \times 10^{-5}$). 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).

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; I500, Illumina 300K; I600, Illumina 610K and Illumina 650W; 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.

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 neurobehavioural 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. In addition to using the same covariates included in genome-wide burden analysis, we controlled for the total number of genes
per subject spanned by rare CNVs to account for signal that merely reflects the global enrichment of CNV burden in cases15. 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

**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 genetic burden BH-FDR 25% (red discs) were used to query GeneMANIA16 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 Frimer et al.17, 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.
A 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, GRIN4), and complexes such as dystrophin and its synaptic interacting proteins (DMD, DTN8, 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–CVN association**

To define specific loci that confer risk for SCZ, we tested CVN 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 CVN loci with gene-based BH-FDR $< 0.05$, two were excluded on the basis of CVN 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 CVN 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 CVN 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 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.29 In this large sample, we observe a high density of ‘non-recurrent’ deletion breakpoints in cases and controls. A snapshot of the breakpoint-association association results from the PGC CNV browser (Online Methods) shows a sawtooth pattern resulting from 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 breakpoint associations, 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 eugenetic 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 (<3%) 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 cases23. 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 \( P = 3.6 \times 10^{-5} \), 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 = 17). 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 sexes23,24. 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 males24. 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 disorders1,25.

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 SCZ26; however, such duplications have been shown to increase risk for developmental delay and a variety of congenital anomalies in pediatric clinical populations27. 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 studies2,6,10,23. 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 postulated29,30 and disputed31 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 4A32. Last, we recognize that a majority of structural variants are not detectable with current genotyping platforms33. 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 CNVs34 and short tandem repeats35.

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 MH094212). 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 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.R., A.C., D.A.C., S.C., A.D., D. E. Domenici, T.E., P.V.G., M.G.H., C.M.H., N.I., A.V.J., E.G.I., K.S.K., G.K., J. Knigl, D.F.L., Q.S.L., I. 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.E.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.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.
CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium

Christian R Manhattan1,175, Daniel P Howrigan2,3, Daniele Merico1,175, Bhooma Thiruvahindrapuram4, Wenting Wu4,5, Douglas S Greer5,6, Danny Antoni5,6, Aniket Shetty5,6, Peter A Holmans6,7, Dalila Pinto6,8,9, Madhusudan Guirala4,5, William M Brandt4,5, Dhearaj Malhotra4,5,10, Zhouzhi Wang1, Karin V Fuentes Fajarado4,5, Michelle S Maile4,5, Stephan Ripke2,3, Ingrid Agartz11,13, Margot Albus14, Madeline Alexander15, Farooq Amin16,17, Joshua Atkins18,19, Silviu A Bacau20, Richard A Belliveau Jr21, Sarah E Bergen4,21, Marcelo Bertalan22,23, Elizabeth Bevacqua3, Tim B Bigdeli20, Donald W Black24, Richard Grugeman25, Nancy G Buccion26, Randy L Buckner27,29, Brendan Bulik-Sullivan2,23, William Byerley30, Wiepke Cahn31, Guiqiu Cai32,33, Murray J Cairns18,33,34, Dominique Campbell35, Rita M Cantor36, Vaughan J Carr33,37, Noa Carrera6, Stanley V Catts33,38, Kimberly D Chambert3, Wei Cheng39, C Robert Cloninger40, David Cohen41, Paul Cormican42, Nick Craddock6,7, Benedicco Crespo-Facorro43,44, James J Crowley45, David Curtis46,47, Michael Davidson48, Kenneth L Davis49, Franziska Degenhardt49,50, Jurgen Del Favero51, Lynn E DeLisi52,53, Dimitris Dikeos54, Timothy Dinan55, Srdjan Djurovic1,56, Gary Donohoe42,45, Elodie Drapeau8, Jubao Dun8,59,58, Frank Dudbridge60, Peter Eichhammer61, Johan Eriksson62,64, Valentina Escott-Price6, Laurent Essioux65, Ayman H Fanous66,69, Kai-How Farhi3, Martillias S Farrell45, Josef Frank70, Lude Franke71, Robert Freedman72, Nelson B Freimer73, Joseph I Friedman8, Andreas J Fraternal8,9, Menachem Fromer3,3,74,75, Giulio Genovese6, Lydmila Georgieva8, Elliot S Gershon59,76, Ina Giegling77,78, Paola Giusti-Rodriguez45, Stephanie Godard79, Jacqueline I Goldstein2,80, Jacob Gratten81, Lieuwe de Haan82, Marian L Hamshere83, Mark Hanssen84, Thomas Hansen22,23, Vahram Haroutunian58,84,85, Annette M Hartmann77, Frans A Henskens53,33,34,86, Stefan Hermus9,50,87, Joel N Hirschhorn80,88,89, Per Hoffmann49,50,87, Andrea Hofman49,50, Hai Liang Huang80, Masashi Ikeda40, Inge Joa91, Anna K Kahler21, René S Kahn31, Luba Kalaydjieva92,93, Juha Karjalainen7, David Kavanagh6, Matthew C Kelle4,94, Brian J Kelly34, James L Kennedy95,97, Yun Jung Kim45, James A Knowles69,98, Bettina Konto77, Claudine Laurent15,59, Phil Lee2,3,75, S Hong Lee81, Sophie E Legge6, Bernard Lerner100, Deborah L Levy53,101, Kung-Yee Liang102, Jeffrey Lieberman103,

1Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada. 2Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. 3Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 4Beyster Center for Psychiatric Genomics, University of California, San Diego, La Jolla, California, USA. 5Department of Psychiatry, University of California, San Diego, La Jolla, California, USA. 6MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. 7National Centre for Mental Health, Cardiff University, Cardiff, UK. 8Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. 9Department 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. 10Department of Neurosurgery, Oregon Health and Science University, Portland, Oregon, USA. 11Department of Genetics and Genomic Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. 12Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia, USA. 13Department of Psychiatry and Behavioral Sciences, Stanford University Medical Center, Stanford, California, USA. 14Department of psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia, USA. 15School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan, New South Wales, Australia. 16Hunter Medical Research Institute, New Lambton, New South Wales, Australia. 17Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA. 18Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. 19Institute of Biological Psychiatry, Mental Health Centre Syd, Hans, Mental Health Services Copenhagen, Copenhagen, Denmark. 20Lundbeck Foundation Initiative for Integrative Psychiatric Research, IPSYCH, Aarhus Denmark. 21Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, Iowa, USA. 22Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. 23School of Nursing, Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA. 24Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA. 25Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA. 26Department of Psychiatry, University of California at San Francisco, San Francisco, California, USA. 27Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, the Netherlands. 28Department of Human Genetics, Icahn School of Medicine at Mount Sinai, New York, New York, USA. 29Schofield Research Institute, Sydney, New South Wales, Australia. 30Priority Centre for Translational Neuroscience and Mental Health, University of Newcastle, Newcastle, New South Wales, Australia. 31Centre Hospitalier du Roussy et INSCIR U1079, Faculty of Medicine, Rouen, France. 32Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. 33School of Psychiatry, University of New South Wales, Sydney, New South Wales, Australia. 34Department of Psychiatry, Royal Brisbane and Women’s Hospital, University of Queensland, Brisbane, Queensland, Australia. 35Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. 36Department of Psychiatry, Washington University St. Louis, Missouri, USA. 37Department of Child and Adolescent Psychiatry, Assistance Publique–Hôpitaux de Paris, Pierre and Marie Curie Faculty of Medicine and Institute for Intelligent Systems and Robotics, Paris, France. 38Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. 39Instituto de Formación e Investigación Marqués de Valdecilla, University Hospital Marqués de Valdecilla, University of Cantabria, Santander, Spain. 40Centre for Investigación Biomédica en Red Salud Mental, Madrid, Spain. 41Department of Genetics, University of
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 28,493 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 publication1.

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, allele 1 – forward, allele 2 – 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 (AFT - aq-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 PennCNV2, iPattern3, Birdsuite4 and C-Score5 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 in 4% (50% reciprocal overlap) frequency in cases and controls combined. We included in the call set CNV’s 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 schizophrenia1 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 study1. 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 x 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 CNV’s 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 CNVs were tested: deletions, duplications and deletions and duplications combined. All analyses were run in PLINK®.

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 26 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 preprocessing 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:

\[ y \sim \text{covariates + global + geneset} \]

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); ‘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 ‘geneset’ 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:

\[ y \sim \text{covariates + global} \]

\[ y \sim \text{covariates + global + geneset} \]

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 (permitted 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:

\[ y \sim \text{covariates + gene} \]

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 \( y \sim \text{covariates} \) and \( y \sim \text{covariates + gene} \).

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, MAGENAI1 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.
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.