Universiteit Antwerpen

This item is the archived peer-reviewed author-version of:

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

Reference:
Marshall Christian R., Howrigan Daniel P., Merico Daniele, Thiruvahindrapuram Bhooma, Wu Wenting, Greer Douglas S., Antaki Danny, Shetty Aniket, Holmans Peter A., Pinto Dalila, ...- Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects
Nature genetics - ISSN 1061-4036 - 49:1(2017), p. 27-35
Full text (Publisher's DOI): https://doi.org/10.1038/NG.3725
To cite this reference: https://hdl.handle.net/10067/1415190151162165141

uantwerpen.be
Institutional repository IRUA

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 December 20.

Published in final edited form as:

Nat Genet. 2017 January ; 49(1): 27–35. doi:10.1038/ng.3725.

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

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

Abstract

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 (OR=1.11, $P=5.7\times10^{-15}$), which persisted after excluding loci implicated in previous studies (OR=1.07, $P=1.7\times10^{-6}$). CNV burden was enriched for genes associated with synaptic function (OR = 1.68, $P = 2.8\times10^{-11}$) and neurobehavioral phenotypes in mouse (OR = 1.18, $P= 7.3\times10^{-5}$). 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 non-allelic homologous recombination.

Correspondence: jsebat@ucsd.edu.

*these authors contributed equally

†these authors co-supervised the study

URLs

PGC CNV browser, http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18.

Visualization 16p13.2: http://bit.ly/1NPgIuq

Visualization of 8q11.23 locus: http://bit.ly/1PwdYTt

Xq28 gene reviews: http://bit.ly/2au9QGb

Genetic Cluster Computer (GCC): https://userinfo.surfsara.nl/systems/lisa

Data Availability

The PGC CNV resource is now publicly 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 (Study accession #EGAS00001001960).

Author Contributions

Management of the study, core analyses and content of the manuscript was the responsibility of the CNV Analysis Group chaired by J.S. 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.M., and C.R.M. Data Processing pipeline was implemented by C.R.M., B.T., W.W., D.G., M.G., A.S. 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., A.D.B., E. Bramon, J.D.B., A.C., D.A.C., S.C., A.D., E. Domenici, H.E., 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, T. Lencz, D.F.L., Q.S.L., J. Liu, A.K.M., S.A.M., A. McQuillin, J.L.M., P.B.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.C.S, P. Sklar. D.St.C., P.F.S., D.R.W., J.R.W., 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.S., B.M.N, M.C.O'D, C.R.M, D.P.H., and D.M. 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 Interest

J.S. is a co-inventor on patents granted (8554488) and pending (20140171371) on genetic methods for the diagnosis of psychiatric disorders. Several of the authors are employees of the following pharmaceutical companies: F. Hoffman-La Roche (E.D., L.E.), Eli Lilly (D.A.C., Y.M., L.N.) and Janssen (A.S., Q.S.L). None of these companies influenced the design of the study, the interpretation of the data, the amount of data reported, or financially profit by publication of the results, which are pre-competitive. The other authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Introduction

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 (odds ratios 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 under-powered 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 (odds ratios 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 significant obstacle in studies of rare and common genetic variation. In response, global collaborations have been formed in order to attain large sample sizes, as exemplified by a study by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC) which identified 108 independent schizophrenia 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 analyses 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 datasets assembled by the Schizophrenia Working Group of the PGC.

Data processing and meta-analytic methods

Raw intensity data were obtained from 57,577 subjects from 43 separate datasets (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. (**Methods** and Supplementary Figure 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. In order 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 >20kb in length, at least 10 probes, and <1% MAF.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Genetic associations were investigated by case-control tests of CNV burden at four levels: (1) genome-wide (2) pathways, (3) genes, and (4) 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 the consistency of overall CNV burden across genotyping platforms, and whether a measurable amount of 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 Kb covered (OR=1.12, $p = 5.7\times10^{-15}$), genes affected (OR=1.21, $p = 6.6\times10^{-21}$), or CNV number (OR=1.03, $p = 1\times10^{-3}$). The burden signal above was driven by CNVs located within genes. Focusing heretofore on the number of genes affected by CNV, the burden metric with the strongest signal of enrichment in our study, the effect size was consistent across all genotyping platforms (Figure 1a). When we split by CNV type, the effect size for copy number losses (OR=1.40, $p = 4\times10^{-16}$) was greater than for gains (OR=1.12, $p = 2\times10^{-7}$) (Supplementary Figures 2 and 3). Partitioning by CNV frequency (based on 50% reciprocal overlap with the full call set, **Methods**), CNV burden was enriched among cases across a range of frequencies, up to counts of 80 (MAF = 0.4%) in the combined sample (Figure 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 Figure 5)

A primary question in this study is the contribution of novel loci to the 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, $p = 1.3\times10^{-7}$, Figure 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%, Figure 1b).

Gene-set (pathway) burden

We assessed whether CNV burden was concentrated within defined sets of genes involved in neurodevelopment or neurological function. A total of 36 gene-sets were evaluated (for a description see Supplementary Table 3), consisting of 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sets) were included as negative controls. We mapped CNVs to genes if they overlapped by at least one exonic basepair.

Gene-set burden was tested using 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 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 (Benjamini-Hochberg FDR ≤ 10%) 15 gene-sets were enriched for rare loss burden in cases and 4 for rare gains in cases, none of which are negative control sets (Figure 2).

Of the 15 sets significant for losses, the majority consisted of synaptic or other neuronal components (9 sets); in particular, “GO synaptic” (GO:0045202) and the activity-regulated cytoskeleton-associated protein complex, or “ARC complex”, rank first based on statistical significance and effect-size respectively (Figure 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 (Figure 2c). To account for dependency between synaptic and neuronal gene-sets, we re-tested loss burden following a step-down logistic regression approach, ranking gene-sets based on significance or effect size (Supplementary Table 4). Only GO synaptic 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 mostly captured by the overlap with synaptic genes. Following the same approach, the mouse neurological/neurobehavioral phenotype set remained nominally significant, 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; (Figure 2b); none of the mouse phenotype sets passed the significance threshold for duplications (Figure 2d).

Given that synaptic gene sets were robustly enriched for deletions in cases, and with an appreciable contribution from loci that have not been strongly associated with SCZ previously, pathway-level interactions of these sets were further investigated. A protein-interaction network was seeded using the synaptic and ARC complex genes that were intersected by rare deletions in this study (Figure 3). A graph of the network highlights multiple subnetworks of synaptic proteins including pre-synaptic adhesion molecules (NRXN1, NRXN3), post-synaptic 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) revealed that deletion burden of the synaptic DGC proteins (intersection of “GO DGC” GO:0016010 and “GO synapse” GO:0045202) was enriched in cases (Deviance test P = 0.05), but deletion burden of the full DGC was not significant (P = 0.69).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 used the Family-Wise Error Rate (FWER) < 0.05. The criterion for suggestive evidence used a Benjamini-Hochberg False Discovery Rate (BH-FDR) < 0.05.

Of eighteen independent CNV loci with gene-based BH-FDR < 0.05, two were excluded based on CNV calling accuracy or evidence of a batch effect (Supplementary Note). The sixteen loci that remain after these additional QC steps, comprising seventeen 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 sixteen 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 provided in Figure 4. A permutation-based false discovery rate 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 meet criterion for suggestive association, including six that have not been reported previously in association with SCZ. Based on our estimation of False Discovery Rates (BH and permutations), we expect to observe less than two associations meeting suggestive criteria by chance. In order to further evaluate the six new candidate loci identified here, we performed experimental validation of CNV calls in a subset of samples by digital droplet PCR (ddPCR, see **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 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 previously implicated SCZ loci, we do find support for all loci where the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 $p < .05$.

Associations at some loci become better delineated through breakpoint-level analysis. For instance, *NRXN1* at 2p16.3 is a CNV hotspot, and exonic deletions of this gene are significantly enriched in SCZ[9, 20]. In this large sample, we observe a high density of "non-recurrent" deletion breakpoints in cases and controls. A snapshot of the breakpoint association results from the PGC CNV browser (see URLs) reveals a saw-tooth pattern of association. Predominant peaks correspond to exons and transcriptional start sites of *NRXN1* isoforms (Figure 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 reveals 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 hotspots for non-allelic homologous recombination (NAHR), a process which in most cases is mediated by flanking segmental duplications [21]. We defined a CNV as "NAHR" when both the start and end breakpoint is located within a segmental duplication. Consistent with the importance of NAHR in generating CNV risk alleles for schizophrenia, most of the loci in Table 1 are flanked by segmental duplications. Moreover, after excluding loci that have been 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 Figure 7), when compared to a null distribution determined by randomizing the genomic positions of associated genes (Supplemental Note). These findings suggest that the novel SCZ CNVs share similar characteristics 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 dataset includes the majority of all microarray data that has 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 (URLs). Our data suggest that the undiscovered novel risk loci that can be detected with current genotyping

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 dataset 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 8 loci at a false discovery rate of 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 dataset (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, based originally on the observation of an increased rate of sex chromosome aneuploidy in cases [22]. X-linked loci could not be detected in previous CNV studies of SCZ, because none to date 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 $P = 3.6\times10^{-4}$, OR = 8.9, Table 1 and Supplementary Figure 8) appear to confer risk for SCZ in both males and females, and the effect size was greatest in males (P = 0.01, OR = ∞). 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 schizophrenia 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 dataset here allows a more detailed view of the synaptic network and highlights subsets of genes account for the excess deletion burden in SCZ, including 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 intriguing 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 sub-networks 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 has demonstrated that the causal variants underlying the strongest common variant association in SCZ include duplications of Complement factor 4A [35]. Lastly, we recognize that a majority of structural variants (SVs) are not detectable with current genotyping platforms [32]. 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 [33] and short tandem repeats [34].

This study represents a milestone. 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 datasets has the potential to further the discovery of loci and identification of the relevant genes and functional elements.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

an initial data formatting step we constructed batches of samples for processing using four different methods, 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 dataset by merging data at the sample level and further filtered calls to make a final dataset 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 datasets that were 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]

Copy Number Variant Analysis Pipeline Architecture and Sample Processing

All aspects of the CNV analysis pipeline were built on the Genetic Cluster Computer (GCC) in the Netherlands https://userinfo.surfsara.nl/systems/lisa.

Input Acceptance and Preprocessing—For Affymetrix we used the *.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 based on 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 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 based on 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 pre-processed, each individual pipeline component pipeline was run in two steps: 1) processing the intensity data by the core pipeline process to produce CNV calls, 2) 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 Quality control (QC) filtering—*Merging of CNV data and Quality control (QC) filtering* is described in detail in the supplementary material. 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 dataset was annotated with Refseq genes (transcriptions and exons). After this stage of quality control (QC), we had a total of **52,511** individuals, with **27,034** SCZ cases and **25,448** controls. To make our final dataset 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.

In order to minimize the impact of technical artifacts and potential confounds on CNV association results, we removed from the dataset 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 datasets 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. Due 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 material. We were unable to control for dataset or genotyping batch, as a subset of the contributing datasets 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 in size). Among the top 20 principal components, only the 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, and 8^{th} principal component showed association with small CNV burden (with $p < 0.01$ used as the significance cutoff).

Lastly, 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 if any additional outliers were present. Only three outliers were removed as their mean B allele (or minor allele) frequency deviated significantly from 0.5. Many CNV metrics are auto-correlated, 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 datasets, MAPD did not differ between in cases and controls (t=1.14, $p = 0.25$). However, among non-Affymetrix 6.0 datasets, LRRSD showed significant differences between cases and controls (t=−35.3, $p <$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$2e^{-16}$), 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 dataset, 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 zero, and controls below zero, and are graphed against overall kb burden in Supplementary Figure 9. We removed three individuals with an SCZ residual phenotype greater than three (or negative three in controls). After the post-processing round of QC, we retained a dataset 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 3 distinct ways – 1) Kb burden of CNVs, 2) Number of genes affected by CNVs, and 3) 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), and 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 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, and the abbreviations for each platform are expanded below:

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

Due to the small sample size of the Omni 2.5 array (28 cases and 10 controls), they were excluded from presentation in the figure, 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

variable, thus allowing us to get 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 using 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, is 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. While 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 the surpassed genome-wide correction in either test was 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 one million permutations at each CNV position, whereby each permutation shuffles the SCZ residual phenotype among all samples, and retains the SCZ residual mean for CNV carriers and non-carriers. For extremely rare CNV, however, CNV carriers at the extreme ends of the SCZ residual phenotype can produce highly significant *p*-values. While 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 standard deviation of residual phenotype mean differences in CNV carriers and non-carriers from 5,000 permutations. Z-scores are calculated as the observed case-control mean difference divided by the empirical standard deviation, 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 six or more CNV, whereas Z-score *p*-values are more conservative among tests with less than six CNV. 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 material.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 ~ covariates + global + 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 non-specific 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 (abbreviated as *U* from universe gene-set count); we also calculated results for total length (abbreviated as *TL*) and variant number plus variant mean length (abbreviated as *CNML*)
- *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 ~ covariates + global

y ~ covariates + global + gene-set

We reported the following statistics:

- coefficient beta estimate (abbreviated as *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 (abbreviated as *Pvalue_dev*)
- gene-set size (i.e. number of genes is the gene-set, regardless of CNV data)
- BH-FDR (Benjamini-Hochberg False Discovery rate)
- percentage of schizophrenia and control subjects with at least 1 gene, 2 genes, etc… impacted by a CNV of the desired type (loss or gain) in the gene-set (abbreviated as *SZ_g1n, SZ_g2n, … CT_g1n, …*)

Please note that, by performing simple simulation analyses, we realized that *Pvalue_glm* can be extremely over-conservative in presence of very few gene-set counts different than 0, while *Pvalue_dev* tends to be slightly under-conservative. While the two p-values tend to agree well for gene-set analysis, *Pvalue_glm* is systematically over-conservative for gene analysis since smaller counts are typically available for single genes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 impacted 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, while 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 ~ 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)
- *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 these two regression models:

- *y ~ covariates*
- *y ~ covariates + gene*

Genome wide correction for multiple testing was determined as described in the supplementary material.

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 we a defined a segment was defined that overlapped a majority of calls. Appropriate digital droplet assays were then selected from the BioRad catalog. A single FAM-labeled probe was designed for *DMRT1, ZMYM5, ZNF92, MAGEA11* and Distal Xq28. Because some deletions of the *VPS13B* gene were non-overlapping, 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 two studies (Sweden and CLOZUK) and 4 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 of 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 & 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

according to manufacturer's protocols (40 cycles at 94°C for 30 sec and 60°C for 1 min). PCR droplets were read & analyzed on Bio-Rad QX100 Droplet Reader with QuantaSoft software.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Christian R. Marshall[1,*], Daniel P. Howrigan[2,3,*], Daniele Merico[1,*], 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[11,12,13], 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[27,28,29], 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[62,63,64], Valentina Escott-Price[6], Laurent Essioux[65], Ayman H. Fanous[66,67,68,69], 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[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[95,96,97], Yunjung Kim[45], James A. Knowles[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[23,108,109,110], 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],

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Vihra Milanova[119], Younes Mokrab[120], Derek W. Morris[42,57], Bertram Müller-Myhsok[121,122,123], 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[98], Tiina Paunio[138], Psychosis Endophenotypes International Consortium, Diana O. Perkins[139], Tune H. Pers[80,89,140], Olli Pietiläinen[138,141], Jonathan Pimm[47], Andrew J. Pocklington[6], John Powell[125], Alkes Price[80,142], Ann E. Pulver[129], Shaun M. Purcell[74], Digby Quested[143], 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[74,144], Douglas M. Ruderfer[6,74], Veikko Salomaa[63], Alan R. Sanders[58,59], Adam Savitz[145], Ulrich Schall[33,34], Thomas G. Schulze[70,146], Sibylle G. Schwab[147], Edward M. Scolnick[3], Rodney J. Scott[18,33,148], Larry J. Seidman[53,116], Jianxin Shi[149], Jeremy M. Silverman[8,150], Jordan W. Smoller[3,75], Erik Söderman[13], Chris C.A. Spencer[151], Eli A. Stahl[74,80], Eric Strengman[31,152], Jana Strohmaier[70], T. Scott Stroup[103], Jaana Suvisaari[104], Dragan M. Svrakic[40], Jin P. Szatkiewicz[45], Srinivas Thirumalai[153], Paul A. Tooney[18,33,34], Juha Veijola[154,155], Peter M. Visscher[81], John Waddington[156], Dermot Walsh[157], Bradley T. Webb[20], Mark Weiser[48], Dieter B. Wildenauer[158], 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], Wellcome Trust Case-Control Consortium, Rolf Adolfsson[132], Ole A. Andreassen[11,115], Douglas H.R. Blackwood[113], Elvira Bramon[159], Joseph D. Buxbaum[8,32,85,160], Sven Cichon[49,50,87,161], David A Collier[120,162], Aiden Corvin[42], Mark J. Daly[2,3,80], Ariel Darvasi[163], Enrico Domenici[10,164], 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[33,93,158,165], Erik G Jönsson[11,13], Kenneth S Kendler[20], George Kirov[6], Jo Knight[95,96,97], Douglas F. Levinson[15], Qingqin S Li[145], 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[3,75,141], Carlos N. Pato[98], Tracey L. Petryshen[3,53,166], Danielle Posthuma[167,168,169], Marcella Rietschel[70], Brien P. Riley[20], Dan Rujescu[77,78], Pamela Sklar[74,85,144], David St. Clair[170], James T.R. Walters[6], Thomas Werge[22,23,171], Patrick F. Sullivan[21,45,139], Michael C O'Donovan[6,7,†], Stephen W. Scherer[1,172,†], Benjamin M. Neale[2,3,75,80,†], Jonathan Sebat[4,5,173,†], and the CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium

Affiliations

[1]The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA [4]Beyster Center for Psychiatric Genomics,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

University of California, San Diego, La Jolla, CA 92093, USA [5]Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093, USA [6]MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, CF24 4HQ, UK [7]National Centre for Mental Health, Cardiff University, Cardiff, CF24 4HQ, UK [8]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [9]Department of Genetics and Genomic Sciences, Seaver Autism Center, The Mindich Child Health & Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [10]Neuroscience Discovery and Translational Area, Pharma Research & Early Development, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland [11]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, 0424 Oslo, Norway [12]Department of Psychiatry, Diakonhjemmet Hospital, 0319 Oslo, Norway [13]Department of Clinical Neuroscience, Psychiatry Section, Karolinska Institutet, SE-17176 Stockholm, Sweden [14]State Mental Hospital, 85540 Haar, Germany [15]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California 94305, USA [16]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia 30322, USA [17]Department of Psychiatry and Behavioral Sciences, Atlanta Veterans Affairs Medical Center, Atlanta, Georgia 30033, USA [18]School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan NSW 2308, 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 23298, USA [21]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-17177, Sweden [22]Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, DK-4000, Denmark [23]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark [24]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, Iowa 52242, USA [25]University Medical Center Groningen, Department of Psychiatry, University of Groningen, NL-9700 RB, The Netherlands [26]School of Nursing, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, USA [27]Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA [28]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, USA [29]Athinoula A. Martinos Center, Massachusetts General Hospital, Boston, Massachusetts 02129, USA [30]Department of Psychiatry, University of California at San Francisco, San Francisco, California, 94143 USA [31]University Medical Center Utrecht, Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, 3584 Utrecht, The Netherlands [32]Department of Human Genetics, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [33]Schizophrenia Research Institute, Sydney NSW 2010, Australia [34]Priority Centre for Translational Neuroscience and Mental Health, University of Newcastle, Newcastle NSW 2300, Australia [35]Centre Hospitalier du Rouvray and INSERM U1079 Faculty of Medicine, 76301 Rouen,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

France [36]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA [37]School of Psychiatry, University of New South Wales, Sydney NSW 2031, Australia [38]Royal Brisbane and Women's Hospital, University of Queensland, Brisbane QLD 4072, Australia [39]Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina 27514, USA [40]Department of Psychiatry, Washington University, St. Louis, Missouri 63110, 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, 75013, France [42]Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Dublin 8, Ireland [43]University Hospital Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, University of Cantabria, E-39008 Santander, Spain [44]Centro Investigación Biomédica en Red Salud Mental, Madrid, Spain [45]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA [46]Department of Psychological Medicine, Queen Mary University of London, London E1 1BB, UK [47]Molecular Psychiatry Laboratory, Division of Psychiatry, University College London, London WC1E 6JJ, UK [48]Sheba Medical Center, Tel Hashomer 52621, Israel [49]Institute of Human Genetics, University of Bonn, D-53127 Bonn, Germany [50]Department of Genomics, Life and Brain Center, D-53127 Bonn, Germany [51]Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, University of Antwerp, B-2610 Antwerp, Belgium [52]VA Boston Health Care System, Brockton, Massachusetts 02301, USA [53]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts 02115, USA [54]First Department of Psychiatry, University of Athens Medical School, Athens 11528, Greece [55]Department of Psychiatry, University College Cork, Co. Cork, Ireland [56]Department of Medical Genetics, Oslo University Hospital, 0424 Oslo, Norway [57]Cognitive Genetics and Therapy Group, School of Psychology and Discipline of Biochemistry, National University of Ireland Galway, Co. Galway, Ireland [58]Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, Illinois 60201, USA [59]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois 60637, USA [60]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK [61]Department of Psychiatry, University of Regensburg, 93053 Regensburg, Germany [62]Folkhälsan Research Center, Helsinki, Finland, Biomedicum Helsinki 1, Haartmaninkatu 8, FI-00290, Helsinki, Finland [63]National Institute for Health and Welfare, P.O. BOX 30, FI-00271 Helsinki, Finland [64]Department of General Practice, Helsinki University Central Hospital, University of Helsinki P.O. BOX 20, Tukholmankatu 8 B, FI-00014, Helsinki, Finland [65]Translational Technologies and Bioinformatics, Pharma Research and Early Development, F.Hoffman-La Roche, CH-4070 Basel, Switzerland [66]Mental Health Service Line, Washington VA Medical Center, Washington DC 20422, USA [67]Department of Psychiatry, Georgetown University, Washington DC 20057, USA [68]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

23298, USA [69]Department of Psychiatry, Keck School of Medicine at University of Southern California, Los Angeles, California 90033, USA [70]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, D-68159 Mannheim, Germany [71]Department of Genetics, University of Groningen, University Medical Centre Groningen, 9700 RB Groningen, The Netherlands [72]Department of Psychiatry, University of Colorado Denver, Aurora, Colorado 80045, USA [73]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, California 90095, USA [74]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [75]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA [76]Departments of Psychiatry and Human Genetics, University of Chicago, Chicago, Illinois 60637 USA [77]Department of Psychiatry, University of Halle, 06112 Halle, Germany [78]Department of Psychiatry, University of Munich, 80336, Munich, Germany [79]Departments of Psychiatry and Human and Molecular Genetics, INSERM, Institut de Myologie, Hôpital de la Pitiè-Salpêtrière, Paris, 75013, France [80]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA [81]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia [82]Academic Medical Centre University of Amsterdam, Department of Psychiatry, 1105 AZ Amsterdam, The Netherlands [83]Illumina, La Jolla, California, California 92122, USA [84]J.J. Peters VA Medical Center, Bronx, New York, New York 10468, USA [85]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [86]School of Electrical Engineering and Computer Science, University of Newcastle, Newcastle NSW 2308, Australia [87]Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, CH-4058, Switzerland [88]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA [89]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children’s Hospital, Boston, Massachusetts 02115, USA [90]Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Aichi, 470-1192, Japan [91]Regional Centre for Clinical Research in Psychosis, Department of Psychiatry, Stavanger University Hospital, 4011 Stavanger, Norway [92]Centre for Medical Research, The University of Western Australia, Perth, WA 6009, Australia [93]The Perkins Institute for Medical Research, The University of Western Australia, Perth, WA 6009, Australia [94]Department of Psychology, University of Colorado Boulder, Boulder, Colorado 80309, USA [95]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, M5T 1R8, Canada [96]Department of Psychiatry, University of Toronto, Toronto, Ontario, M5T 1R8, Canada [97]Institute of Medical Science, University of Toronto, Toronto, Ontario, M5S 1A8, Canada [98]Department of Psychiatry and Zilkha Neurogenetics Institute, Keck School of Medicine at University of Southern California, Los Angeles, California 90089, USA [99]Department of Child and Adolescent Psychiatry, Pierre and Marie Curie Faculty of Medicine,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Paris 75013, France [100]Department of Psychiatry, Hadassah-Hebrew University Medical Center, Jerusalem 91120, Israel [101]Psychology Research Laboratory, McLean Hospital, Belmont, MA [102]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland 21205, USA [103]Department of Psychiatry, Columbia University, New York, New York 10032, USA [104]Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, P.O. BOX 30, FI-00271 Helsinki, Finland [105]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA [106]Department of Psychiatry, University of Bonn, D-53127 Bonn, Germany [107]Centre National de la Recherche Scientifique, Laboratoire de Génétique Moléculaire de la Neurotransmission et des Processus Neurodégénératifs, Hôpital de la Pitié Salpêtrière, 75013, Paris, France [108]Department of Biomedicine, Aarhus University, DK-8000 Aarhus C, Denmark [109]Centre for Integrative Sequencing, iSEQ, Aarhus University, DK-8000 Aarhus C, Denmark [110]Department of Genomics Mathematics, University of Bonn, D-53127 Bonn, Germany [111]Research Unit, Sørlandet Hospital, 4604 Kristiansand, Norway [112]Department of Psychiatry, National University of Ireland Galway, Co. Galway, Ireland [113]Division of Psychiatry, University of Edinburgh, Edinburgh EH16 4SB, UK [114]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH16 4SB, UK [115]Division of Mental Health and Addiction, Oslo University Hospital, 0424 Oslo, Norway [116]Massachusetts Mental Health Center Public Psychiatry Division of the Beth Israel Deaconess Medical Center, Boston, Massachusetts 02114, USA [117]Estonian Genome Center, University of Tartu, Tartu 50090, Estonia [118]School of Psychology, University of Newcastle, Newcastle NSW 2308, Australia [119]First Psychiatric Clinic, Medical University, Sofia 1431, Bulgaria [120]Eli Lilly and Company Limited, Erl Wood Manor, Sunninghill Road, Windlesham, Surrey, GU20 6PH UK [121]Max Planck Institute of Psychiatry, 80336 Munich, Germany [122]Institute of Translational Medicine, University of Liverpool, Liverpool L69 3BX, UK127Munich [123]Cluster for Systems Neurology (SyNergy), 80336 Munich, Germany [124]Department of Psychiatry, Royal College of Surgeons in Ireland, Dublin 2, Ireland [125]King's College London, London SE5 8AF, UK [126]Maastricht University Medical Centre, South Limburg Mental Health Research and Teaching Network, EURON, 6229 HX Maastricht, The Netherlands [127]Department of Psychiatry and Psychotherapy, Jena University Hospital, 07743 Jena, Germany [128]Queensland Centre for Mental Health Research, University of Queensland, Brisbane QLD 4076, Australia [129]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA [130]Department of Psychiatry, Trinity College Dublin, Dublin 2, Ireland [131]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, 46285 Indiana, USA [132]Department of Clinical Sciences, Psychiatry, Umeå University, SE-901 87 Umeå, Sweden [133]DETECT Early Intervention Service for Psychosis, Blackrock, Co. Dublin, Ireland [134]Lawrence Berkeley National Laboratory, University of California at Berkeley, Berkeley, California 94720, USA [135]Centre for Public Health, Institute of Clinical Sciences, Queen's University Belfast, Belfast BT12 6AB,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

UK [136]Institute of Psychiatry, King's College London, London SE5 8AF, UK [137]Melbourne Neuropsychiatry Centre, University of Melbourne & Melbourne Health, Melbourne VIC 3053, Australia [138]Public Health Genomics Unit, National Institute for Health and Welfare, P.O. BOX 30, FI-00271 Helsinki, Finland [139]Department of Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-7160, USA [140]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800, Denmark [141]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, P.O. BOX 20 FI-00014, Helsinki, Finland [142]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA [143]Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK [144]Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [145]Neuroscience Therapeutic Area, Janssen Research and Development, Raritan, New Jersey 08869, USA [146]Department of Psychiatry and Psychotherapy, University of Göttingen, 37073 Göttingen, Germany [147]Psychiatry and Psychotherapy Clinic, University of Erlangen, 91054 Erlangen, Germany [148]Hunter New England Health Service, Newcastle NSW 2308, Australia [149]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA [150]Research and Development, Bronx Veterans Affairs Medical Center, New York, New York 10468, USA [151]Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, UK [152]Department of Medical Genetics, University Medical Centre Utrecht, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands [153]Berkshire Healthcare NHS Foundation Trust, Bracknell RG12 1BQ, UK [154]Department of Psychiatry, University of Oulu, P.O. BOX 5000, 90014, Finland [155]University Hospital of Oulu, P.O. BOX 20, 90029 OYS, Finland [156]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin 2, Ireland [157]Health Research Board, Dublin 2, Ireland [158]School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, WA 6009, Australia [159]University College London, London WC1E 6BT, UK [160]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [161]Institute of Neuroscience and Medicine (INM-1), Research Center Juelich, 52428 Juelich, Germany [162]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, SE5 8AF, UK [163]Department of Genetics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel [164]Centre for Integrative Biology, University of Trento, 38122 Trento, Italy [165]Centre for Clinical Research in Neuropsychiatry, School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Medical Research Foundation Building, Perth WA 6000, Australia [166]Center for Human Genetic Research and Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, USA [167]Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, Amsterdam 1081, The Netherlands [168]Department of Complex Trait Genetics, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam 1081, The Netherlands [169]Department of Child and Adolescent Psychiatry, Erasmus University

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Medical Centre, Rotterdam 3000, The Netherlands [170]University of Aberdeen, Institute of Medical Sciences, Aberdeen, AB25 2ZD, UK [171]Department of Clinical Medicine, University of Copenhagen, Copenhagen 2200, Denmark [172]Department of Molecular Genetics and McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada [173]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093, USA

Acknowledgments

Core funding for the Psychiatric Genomics Consortium is from the US National Institute of Mental Health (U01 MH094421). We thank Thomas Lehner, Anjene Addington and Geetha 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 Notes. Membership of the Wellcome Trust Case Control Consortium and Psychosis Endophenotype International Consortium are provided in the Supplementary Notes.

References

1. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012; 148:1223–41. [PubMed: 22424231]
2. Walsh T, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008; 320:539–43. [PubMed: 18369103]
3. The International Schizophrenia, C. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008; 455:237–241. [PubMed: 18668038]
4. Malhotra D, et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron. 2011; 72:951–63. [PubMed: 22196331]
5. Xu B, et al. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet. 2008; 40:880–5. [PubMed: 18511947]
6. Kirov G, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Molecular psychiatry. 2012; 17:142–53. [PubMed: 22083728]
7. McCarthy SE, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009; 41:1223–7. [PubMed: 19855392]
8. Mulle JG, et al. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet. 2010; 87:229–36. [PubMed: 20691406]
9. Rujescu D, et al. Disruption of the neurexin 1 gene is associated with schizophrenia. Hum Mol Genet. 2008
10. Pocklington AJ, et al. Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia. Neuron. 2015; 86:1203–14. [PubMed: 26050040]
11. Horev G, et al. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. Proc Natl Acad Sci U S A. 2011; 108:17076–81. [PubMed: 21969575]
12. Golzio C, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. 2012; 485:363–7. [PubMed: 22596160]
13. Holmes AJ, et al. Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk. J Neurosci. 2012; 32:18087–100. [PubMed: 23238724]
14. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–7. [PubMed: 25056061]
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. 2007; 17:1665–74. [PubMed: 17921354]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

16. Pinto D, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010; 466:368–72. [PubMed: 20531469]
17. Korn JM, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat.Genet. 2008; 40:1253–1260. [PubMed: 18776909]
18. Vacic V, et al. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature. 2011; 471:499–503. [PubMed: 21346763]
19. Raychaudhuri S, et al. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet. 2010; 6
20. Kirov G, et al. Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. Hum Mol Genet. 2008; 17:458–65. [PubMed: 17989066]
21. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. 1998; 14:417–22. [PubMed: 9820031]
22. DeLisi LE, et al. Schizophrenia and sex chromosome anomalies. Schizophr Bull. 1994; 20:495–505. [PubMed: 7973466]
23. El-Hattab AW, et al. Int22h-1/int22h-2-mediated Xq28 rearrangements: intellectual disability associated with duplications and in utero male lethality with deletions. J Med Genet. 2011; 48:840–50. [PubMed: 21984752]
24. El-Hattab AW, et al. Clinical characterization of int22h1/int22h2-mediated Xq28 duplication/deletion: new cases and literature review. BMC Med Genet. 2015; 16:12. [PubMed: 25927380]
25. Sebat J, Levy DL, McCarthy SE. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. Trends Genet. 2009; 25:528–35. [PubMed: 19883952]
26. Rees E, et al. Evidence that duplications of 22q11.2 protect against schizophrenia. Mol Psychiatry. 2014; 19:37–40. [PubMed: 24217254]
27. Van Campenhout S, et al. Microduplication 22q11.2: a description of the clinical, developmental and behavioral characteristics during childhood. Genet Couns. 2012; 23:135–48. [PubMed: 22876571]
28. Fromer M, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014; 506:179–84. [PubMed: 24463507]
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. 1993; 30:131–4. [PubMed: 8445617]
30. Straub RE, 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. 2002; 71:337–48. [PubMed: 12098102]
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. 2006; 79:903–9. [PubMed: 17033966]
32. Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75–81. [PubMed: 26432246]
33. Brandler WM, et al. Frequency and Complexity of De Novo Structural Mutation in Autism. Am J Hum Genet. 2016; 98:667–79. [PubMed: 27018473]
34. Gymrek M, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016; 48:22–9. [PubMed: 26642241]
35. Sekar A, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016; 530:177–83. [PubMed: 26814963]
36. Zuberi K, et al. GeneMANIA prediction server 2013 update. Nucleic Acids Res. 2013; 41:W115–22. [PubMed: 23794635]
37. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–7. [PubMed: 25056061]
38. 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. 2007; 17:1665–74. [PubMed: 17921354]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

39. Pinto D, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010; 466:368–72. [PubMed: 20531469]

40. Korn JM, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat.Genet. 2008; 40:1253–1260. [PubMed: 18776909]

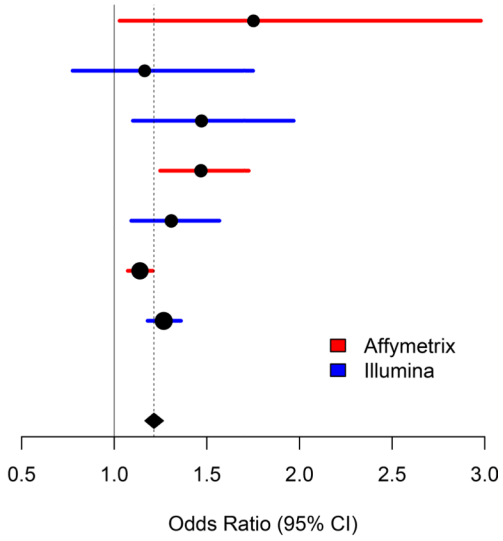
41. McCarthy SE, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009; 41:1223–7. [PubMed: 19855392]

42. Purcell S, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007; 81:559–75. [PubMed: 17701901]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A

Platform	Cases	Controls	Genes	pval	OR
A500	314	201	1.7	0.039	1.75
I300	410	285	2	0.461	1.16
I600	818	623	1.3	0.009	1.47
A5.0	1130	1001	1.7	3.6e−06	1.47
I550	1572	1705	1.9	0.004	1.31
A6.0	7027	7980	2.4	1.6e−05	1.14
omni	9795	8422	2.1	5.8e−11	1.27
PGC	**21094**	**20227**	**2.2**	**6.6e−21**	**1.21**



B

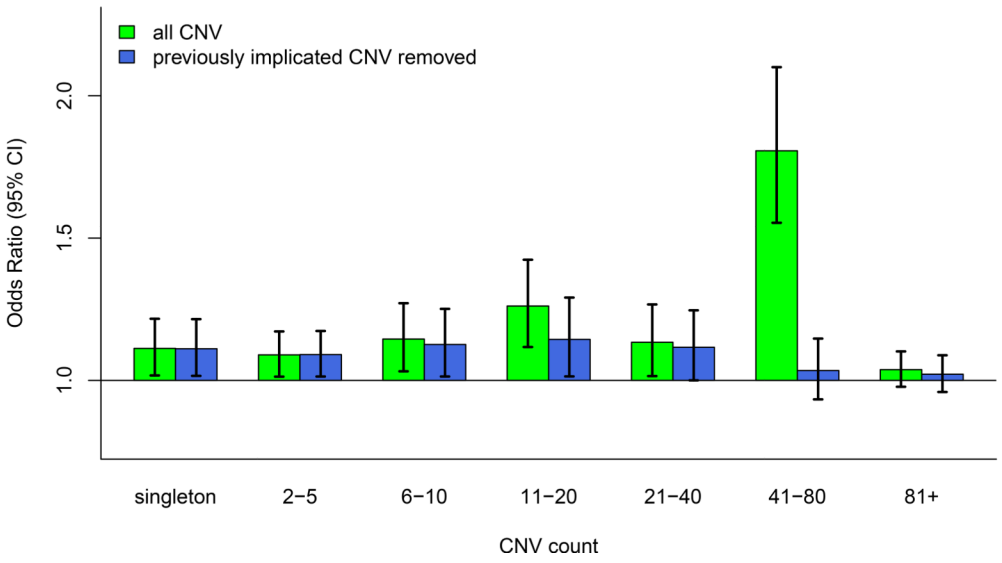


Figure 1. CNV Burden

(A) Forest plot of CNV burden (measured here as genes affected by CNV), partitioned by genotyping platform, with the full PGC sample at the bottom. CNV burden is calculated by combining CNV gains and losses. Numbers of case and controls for each platform are listed, and “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 (see methods). The odds ratio is the exponential of the logistic regression coefficient, and odds ratios above one predict increased SCZ risk. **(B)** CNV burden partitioned by CNV frequency. For reference, for autosomal CNVs, a CNV count of 41 in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 based on a 50% reciprocal overlap with other CNVs. CNV gene burden with inclusion of all CNVs are shown in green, and burden excluding previously implicated CNV loci are shown in blue.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

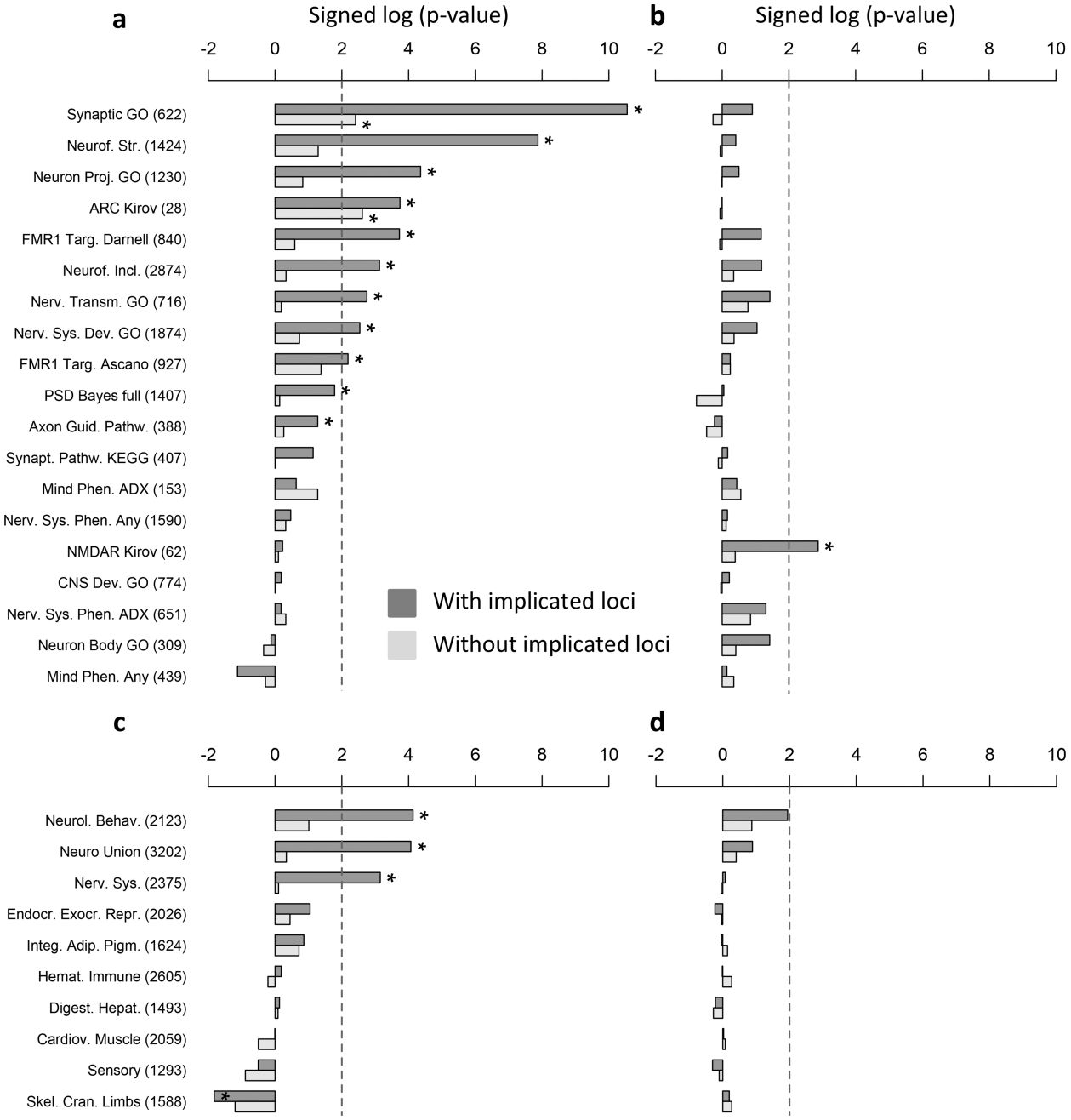


Figure 2. Gene-set Burden

Gene-set burden test results for rare losses **(a, c)** and gains (**b, d**); frames a–b display gene-sets for neuronal function, synaptic components, neurological and neurodevelopmental phenotypes in human; frames **c–d** display gene-sets for human homologs of mouse genes implicated in abnormal phenotypes (organized by organ systems); both are 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. Gene-sets passing the 10% BH-FDR threshold are marked with "*". Gene-sets representing brain expression patterns were omitted from the figure because only a few were significant (losses: 1, gains: 3).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

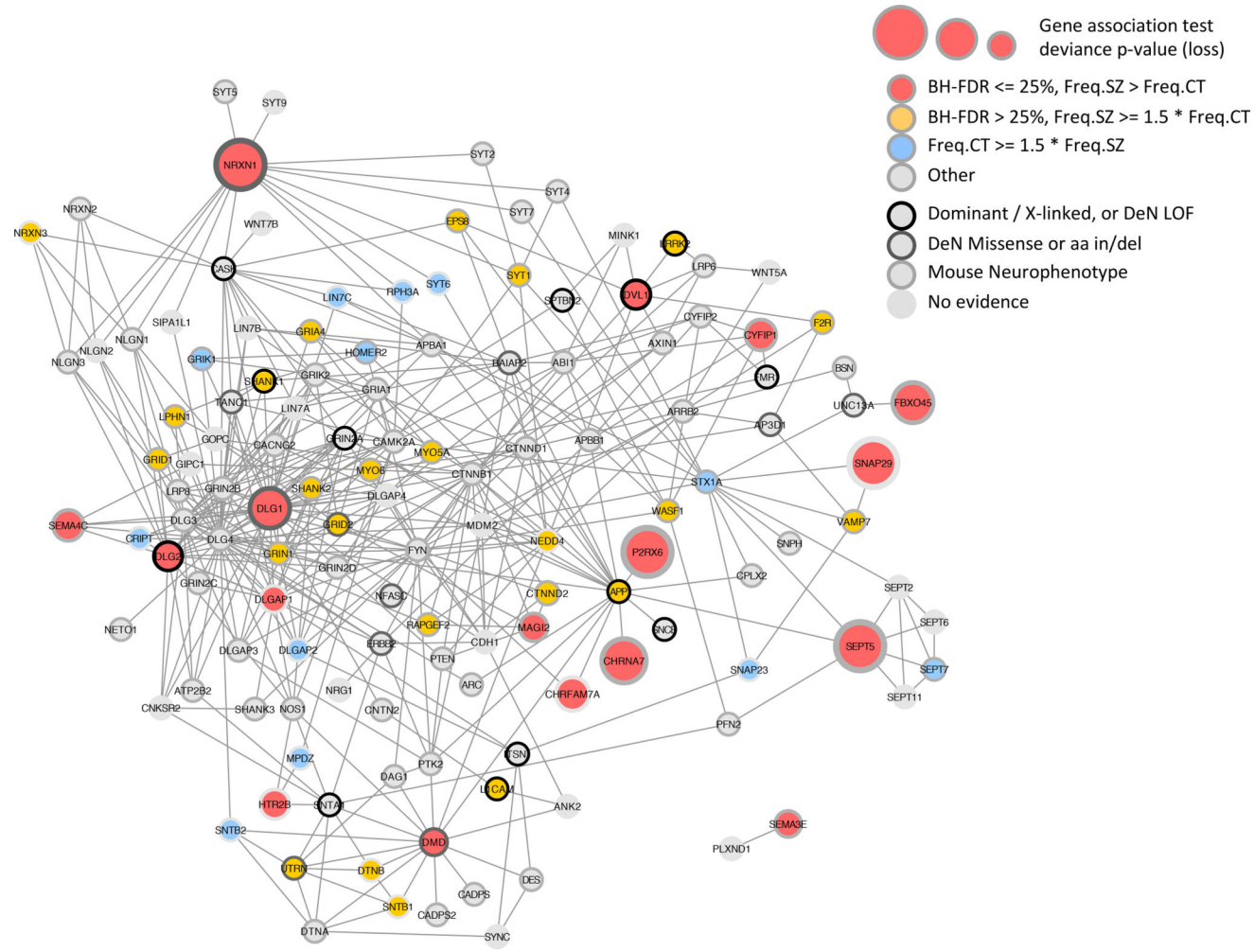


Figure 3. Protein Interaction Network for Synaptic Genes

Synaptic and ARC-complex genes intersected by a rare loss in at least 4 case or control subjects and with genic burden Benjamini-Hochberg FDR <= 25% (red discs) were used to query GeneMANIA[36] and retrieve additional protein interaction neighbors, resulting in a network of 136 synaptic genes. Genes are depicted as disks; disk centers are colored based on rare loss frequency (Freq.SZ and Freq.CT) being prevalent in cases or controls; disk borders are colored to mark (i) gene implication in human dominant or X-linked neurological or neurodevelopmental phenotype, (ii) *de novo* mutation (DeN) reported by Fromer et al. [28], split between LOF (frameshift, stop-gain, core splice site) and missense or amino acid insertion / deletion, (iii) implication in mouse neurobehavioral abnormality. Pre-synaptic adhesion molecules (NRXN1, NRXN3), post-synaptic 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 controls.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

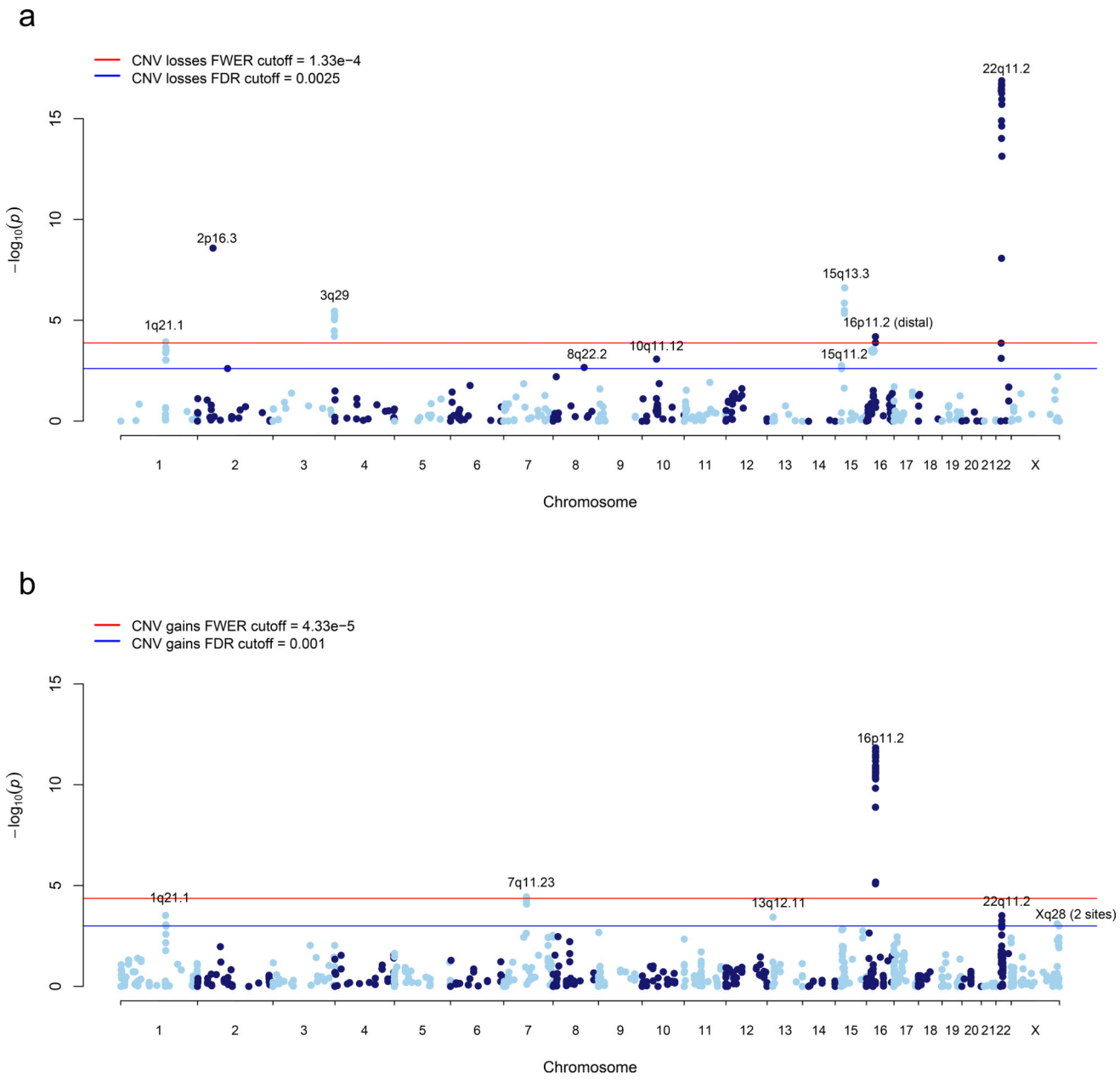


Figure 4. Gene Based Manhattan Plot

Manhattan plot displaying the –log10 deviance *p*-value for **(a)** CNV losses and **(b)** CNV gains 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

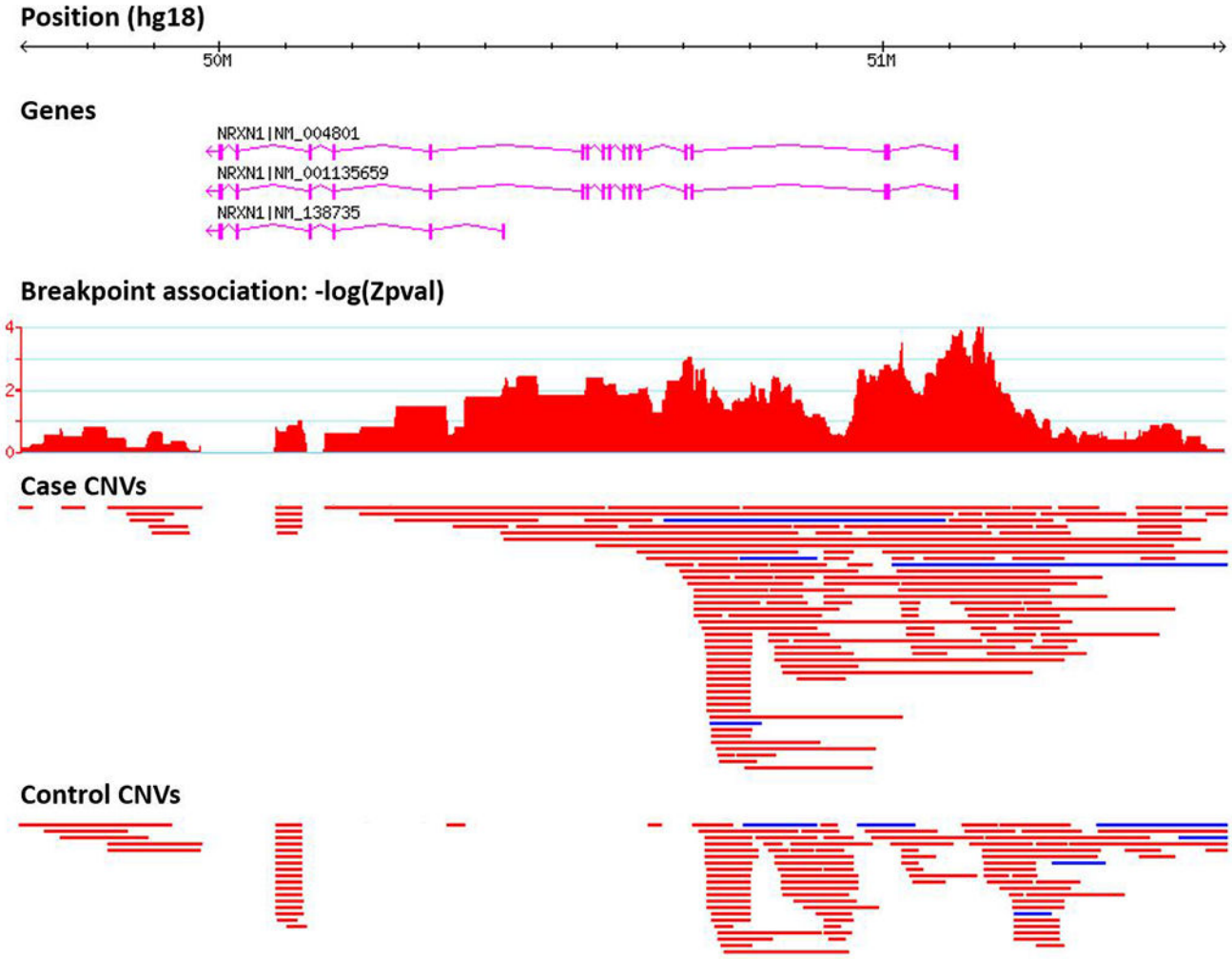


Figure 5. Manhattan plot of breakpoint-level associations across the Neurexin-1 locus

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 dataset.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Nat Genet. Author manuscript; available in PMC 2017 December 20.

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-value	Odds Ratio [95% CI]
22	17,400,000	19,750,000	22q11.21	Previously Implicated	NAHR	loss	risk	yes	3.54E-15	64	1	5.70E-18	67.7 [9.3–492.8]
16	29,560,000	30,110,000	16p11.2 (proximal)	Previously Implicated	NAHR	gain	risk	yes	5.82E-10	70	7	2.52E-12	9.4 [4.2–20.9]
2	50,000,992	51,113,178	2p16.3 **NRXN1**	Previously Implicated	NHEJ	loss	risk	yes	3.52E-07	35	3	4.92E-09	14.4 [4.2–46.9]
15	28,920,000	30,270,000	15q13.3	Previously Implicated	NAHR	loss	risk	yes	2.22E-05	28	2	2.13E-07	15.6 [3.7–66.5]
1	144,646,000	146,176,000	1q21.1	Previously Implicated	NAHR	loss+gain	risk	yes	0.00011	60	14	1.50E-06	3.8 [2.1–6.9]
3	197,230,000	198,840,000	3q29	Previously Implicated	NAHR	loss	risk	yes	0.00024	16	0	1.86E-06	INF
16	28,730,000	28,960,000	16p11.2 (distal)	Previously Reported	NAHR	loss	risk	yes	0.0029	11	1	5.52E-05	20.6 [2.6–162.2]
7	72,380,000	73,780,000	7q11.23	Previously Reported	NAHR	gain	risk	yes	0.0048	16	1	1.68E-04	16.1 [3.1–125.7]
X	153,800,000	154,225,000	Xq28 (distal)	Novel	NAHR	gain	risk	no	0.049	18	2	3.61E-04	8.9 [2.0–39.9]
22	17,400,000	19,750,000	22q11.21	Previously Reported	NAHR	gain	protective	no	0.024	3	16	4.54E-04	0.15 [0.04–0.52]
7	64,476,203	64,503,433	7q11.21 **ZNF92**	Novel	NAHR	loss+gain	protective	no	0.033	131	180	6.71E-04	0.66 [0.52–0.84]
13	19,309,593	19,335,773	13q12.11 **ZMYM5**	Novel	NHAR	gain	protective	no	0.024	15	38	7.91E-04	0.36 [0.19–0.67]
X	148,575,477	148,580,720	Xq28 **MAGEA11**	Novel	NAHR	gain	protective	no	0.044	12	36	1.06E-03	0.35 [0.18–0.68]
15	20,350,000	20,640,000	15q11.2	Previously Implicated	NAHR	loss	risk	no	0.044	98	50	1.34E-03	1.8 [1.2–2.6]
9	831,690	959,090	9p24.3 **DMRT1**	Novel	NHEJ	loss+gain	risk	no	0.049	13	1	1.35E-03	12.4 [1.6–98.1]
8	100,094,670	100,958,984	8q22.2 **VPS13B**	Novel	NHEJ	loss	risk	no	0.048	7	1	1.74E-03	14.5 [1.7–122.2]
7	158,145,959	158,664,998	7p36.3 **VIPR2 WDR60**	Previously Reported	NAHR	loss+gain	risk	no	0.046	20	6	5.79E-03	3.5 [1.3–9.0]

All seventeen association signals listed contain at least one gene with Benjamini-Hochberg false discovery rate (BH-FDR) < 0.05 in the gene-based test, with eight containing at least one gene surpassing the family-wise error rate (FWER) < 0.05. Genomic positions listed are using hg18 coordinates. For putative CNV mechanisms, non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ) are listed as the likely genomic feature driving CNV formation at each locus. Regional *p*-values and odds ratios listed 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 the 22q11.21 locus are listed as separate association signals, as CNV losses associate with SCZ risk while CNV gains associate with protection from SCZ. For each association we indicate whether it was previously described in the literature (Previously Reported) and if the reported P-value exceeded the multiple testing correction in this study (Previously Implicated).