CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/125988/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Szatkiewicz, Jin P., Fromer, Menachem, Nonneman, Randal J., Ancalade, NaEshia, Johnson, Jessica S., Stahl, Eli A., Rees, Elliott , Bergen, Sarah, Hultman, Christina, Kirov, George , O'Donovan, Michael , Owen, Michael , Holmans, Peter , Sklar, Pamela, Sullivan, Patrick F., Purcell, Shaun M., Crowley, James J. and Ruderfer, Douglas M. 2020. Characterization of single gene copy number variants in schizophrenia. Biological Psychiatry 87 (8) , pp. 736-744. 10.1016/j.biopsych.2019.09.023

Publishers page: http://dx.doi.org/10.1016/j.biopsych.2019.09.023

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Characterization of single gene copy number variants in schizophrenia

Jin P. Szatkiewicz[1], Menachem Fromer[2], Randal J. Nonneman[1], NaEshia Ancalade[1], Jessica S. Johnson[2], Eli A. Stahl[2], Elliott Rees[3], Sarah Bergen[4], Christina Hultman[4], George Kirov[3], Michael O'Donovan[3], Michael Owen[3], Peter Holmans[3], Pamela Sklar[2], Patrick F. Sullivan[1,4], Shaun M. Purcell[5], James J. Crowley[1,6], Douglas M. Ruderfer[7*]

1. Center for Psychiatric Genomics, Department of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA
2. Division of Psychiatric Genomics, Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, NYC, NY
3. MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, School of Medicine, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK
4. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
5. Department of Psychiatry, Brigham & Women's Hospital, Harvard Medical School, Boston, MA
6. Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
7. Division of Genetic Medicine, Departments of Medicine, Psychiatry and Biomedical Informatics, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN USA

Corresponding author:
Douglas Ruderfer
515B Light Hall
Vanderbilt University Medical Center
Nashville, TN 37232
douglas.ruderfer@vanderbilt.edu

Short title: Single gene copy number variants in schizophrenia

Keywords: schizophrenia, copy number variation, exome-sequencing, single-gene, calcium channel, genetics

Abstract

Background: Genetic studies of schizophrenia have implicated numerous risk loci including several copy number variants (CNVs) of large effect and hundreds of loci of small effect. In only a few cases has a specific gene been clearly identified. Rare CNVs affecting a single gene offer a potential avenue to discovering schizophrenia risk genes.

Methods: CNVs were generated from exome-sequencing of 4,913 schizophrenia cases and 6,188 controls from Sweden. We integrated multiple CNV calling methods (XHMM and ExomeDepth) to expand our set of single-gene CNVs and leveraged two different approaches for validating these variants (qPCR and Nanostring).

Results: We found a significant excess of all rare CNVs (deletions p=0.0004, duplications p=0.0006) and single-gene CNVs (deletions p=0.04, duplications p=0.03) in schizophrenia cases compared to controls. An expanded set of CNVs generated from integrating multiple approaches showed a significant burden of deletions in 11/21 gene-sets previously implicated in schizophrenia and across all genes in those sets (p=0.008), although no tests survived correction. We performed an extensive validation of all deletions in the significant set of voltage-gated calcium channels among CNVs called from both exome-sequencing and genotyping arrays. In total, 4 exonic, single-gene deletions validated in cases and none in controls (p=0.039), of which all were identified by exome-sequencing.

Conclusions: These results point to the potential contribution of single-gene CNVs to schizophrenia, that the utility of exome-sequencing for CNV calling has yet to be maximized and single-gene CNVs should be included in gene focused studies using other classes of variation.

Introduction

Schizophrenia (SCZ) is a heritable psychiatric disorder that causes substantial morbidity, mortality, and personal and societal costs(1–4). Identifying genetic variation influencing risk will improve our biological understanding of SCZ. Copy number variants (CNVs) are appealing as they directly alter gene dosage providing an interpretable effect on gene function. SCZ cases carry a burden of large and rare CNVs (>100 kb and <1%)(5,6) and multiple rare recurrent CNVs with

substantial effects on risk (genotypic relative risks 4-20) have been identified (e.g., 16p11.2 and 22q11.21)(6–11). Most of these known CNVs are megabase-sized and affect the dosages of many genes, but if specific genes contributing to risk could be identified it would aid our understanding of the neurobiology of the disorder. Thus far, only a few individual genes from genetic studies of CNVs and SNVs have been implicated: *NRXN1*(12)*, TOP3B*(13), *RBM12*(14) and *SETD1A*(15), all of which provided novel insights into SCZ pathophysiology. Therefore, gene-focused CNV evaluation in large samples with high resolution capture is needed.

The majority of CNVs have a small genomic footprint(16–19) and, due to technological limitations or cost, their contribution to SCZ remains unknown(20). Commercial microarrays are limited in resolution by probe density and are largely incapable of detecting CNVs below 10 kb while also having low specificity for CNVs between 10 and 100 kb(21). CNV detection from whole genome sequencing offers a substantial improvement, but remains expensive and is currently infeasible for large samples. Whole exome sequencing (WES) can be used to identify CNVs impacting exons(22). These data, while noisy from dependence on read depth and lacking exact breakpoints from the discrete nature of exons, can be used to identify smaller CNVs affecting single genes that may be more interpretable in their contribution to SCZ risk.

Here, we performed a comprehensive analysis of CNVs from WES data in the Swedish Schizophrenia Study of 4,978 schizophrenia cases and 6,256 controls(23). Our goals were to evaluate the impact of single-gene CNVs on SCZ risk, and to discover copy number changes in specific genes that could lead to improved mechanistic understanding of SCZ risk. All samples also have GWAS genotyping arrays and Illumina exome arrays(24,25) providing additional data to follow up and validate CNVs.

Methods

Sample description

We extracted DNA from venous blood samples from 11,234 Swedish participants (4,978 SCZ cases, 6,256 controls, mean age at sample collection: 55 years). An additional 1,172 samples were included in generating and cleaning exome-sequencing CNVs to improve estimates of copy number and frequency but were removed before analyses (total N: 12,384). All procedures were approved by ethical committees in Sweden and the US, and all subjects provided informed written consent. Genomic investigation of each subject was done using independent technologies including GWAS genotyping(24), exome array genotyping(20), and exome sequencing(23,26). Genotyping and sequencing were conducted at the Broad Institute. Rare CNVs from GWAS arrays and exome genotyping arrays had been previously generated(20,25), and is briefly described in the supplementary material. Exome-sequencing based CNVs were generated for this analysis and have not previously been reported. Individuals already known to be carrying large CNV were included in all analyses. All genomic locations are given in NCBI build 37/UCSC hg19 coordinates.

CNV calling and QC using XHMM

We ran XHMM (eXome-Hidden Markov Model) as previously described(22,27), including calculating mean per-base coverage across 189,894 targets (sequences designed for capture, predominantly exons) using GATK DepthOfCoverage. A total of 14,555 targets were excluded before CNV calling due to: mean sequencing depth <10x, low complexity sequence (as defined by RepeatMasker) in >25% of its span, GC content <10% or >90%, and spanning <10 bp or >10 kb. The resulting sample-by-target read depth matrix was scaled by mean-centering the targets, after which principal component analysis (PCA) of the matrix was performed. To normalize the data,

the top 109 principal components (those with variance >70% of the mean variance across all components) were removed from the data to account for systematic biases at the target- or sample-level, such as GC content or sequencing batch effects. Additional targets (n=37) were removed if variance in read depth remained high after normalization (standard deviation >50). CNVs were called using the Viterbi hidden Markov model (HMM) with default XHMM parameters, and XHMM CNV quality scores (SQ) were calculated using the forward-backward HMM. For any CNV detected in at least one individual, we statistically genotyped all samples using the same XHMM quality scores and outputted as a single VCF. Twenty-two samples failed CNV calling due to low overall read depth, A total of 175,303 targets were used to call CNVs across 12,384 samples after all filtering. CNVs from sex chromosomes would be inaccurately called since males and females were run together and so were removed from analyses.

There were 494,403 autosomal CNVs called by XHMM before any filtering. We removed 115 individuals (56 cases, 59 controls) with > 3 standard deviations from the mean in total number of CNVs (71.5) or total genomic content affected by CNVs (6,529 kb). After sample outlier removal, 484,940 CNV (SQ > 0) were used to develop a frequency filter, and we retained only CNVs present in less than 1% of individuals (<0.5% minor allele frequency). To account for the discrete nature of exons, each target was numbered sequentially based on genomic coordinates and frequency filtering was done using the sequential target information before mapping targets to genomic positions. After frequency filtering, there were 51,812 CNVs with a per individual mean of 4.3 ranging from 1-107. After quality filtering (SQ ≥ 60), 14,243 CNVs remained (we refer to this dataset going forward as the “exome QC” dataset). The median CNV length was 22,991bp and 77% (n=10,950) were below 100kb which is a typical cutoff for array-based CNVs. We note,

however, that the lengths of CNVs generated from exome-sequencing are often inaccurate due to the discrete nature in which breakpoints are determined.

Expanded single-gene CNV dataset integrating XHMM and ExomeDepth

Increasing XHMM quality scores (SQ) disproportionately removes shorter CNVs. In an effort to quantify the proportion of shorter CNVs with lower quality scores that are true, we used exome-sequencing data from 624 trios(28) and calculated transmission as a function of quality score and minimum number of targets required per CNV. These data were processed as described above. We focused on rare CNVs (< 0.1%) to avoid counting transmissions arbitrarily. At default filtering thresholds (SQ ≥ 60, >= 3 exons) we calculated a transmission rate of 0.42 (64 maternal CNVs, 26 transmitted; 71 paternal CNVs, 30 transmitted). CNVs having a single supporting exon and no minimum SQ (i.e. all CNVs) were substantially more frequent but had reduced transmission rate of 0.114 (449 maternal CNVs, 55 transmitted; 575 paternal CNVs, 62 transmitted). However, this transmission rate suggests that potentially 20+% of these "low quality" events may be real. As a method to retain the true shorter CNVs while removing as many of the false positive CNVs as possible we required additional support from an independent approach ExomeDepth(29). Briefly, ExomeDepth selects a reference set of individuals having similar sequencing properties independently for CNV inference of each sample. We called CNVs within experimental plates of 96 individuals that were processed and sequenced at the same time in order to provide the most comparable reference set for each sample and reduce batch effects. In total, we called CNVs for 12,313 samples totaling 1,915,300 CNV with a mean of 155.5 per individual and ranging from 1-811.

We retained all XHMM calls with SQ ≥ 60 and any CNV called by both ExomeDepth and XHMM regardless of quality score (referred to as the "expanded exome" dataset). For comparison, 92% of calls from the exome QC dataset were called by ExomeDepth whereas only 20% of CNVs affecting a single exon and low SQ (< 30) were called across both methods. In total, the expanded exome dataset had 24,843 CNVs (10,600 added to exome QC). To further assess quality, we compared the additional CNVs to a high confident set of CNVs (>100kb) from genotyping arrays of the same individuals(25). While the vast majority of these calls (90%, 3,254 out of 3,597) are identified in our exome QC dataset, there are still 343 that that are called by XHMM but that do not surpass the filtering threshold. Only 28% (10,600/37,569) of the possible XHMM calls were added to our expanded exome dataset yet 88% (300/343) of the remaining high-quality genotyping CNVs were included. Using the union of the two approaches allows us to expand our set of shorter CNVs while retaining only those with the most support.

CNV burden and association analyses

We performed burden and association analyses using Plink(30), employing empirical permutation (n=10,000) of case/control label where permutation was performed within sequencing batch to account for any batch effects. CNVs were considered to affect a gene if there was any overlap of the genomic coordinates of the CNVs and the gene. For gene-set tests, we used a regression framework built into Plink(31) that tests whether cases carry more CNVs in the set of genes compared to all genes after covarying for number and amount of CNVs.

Incorporating CNVs from previously run genotyping arrays of the same individuals

To maximize the sensitivity to detect gene/exon level CNVs, we constructed a union call set by combining the data from GWAS array, exome array, and our expanded exome CNV dataset. We first created a database of all non-redundant CNVs, where, for each CNV record, we indicated (1) how many platform(s) had identified the CNV; (2) which specific platform(s) had identified the CNV; (3) the coordinates of CNV from each platform. We considered two CNVs redundant if they had the same direction of the copy number change and they overlapped more than 50% of their lengths. Details for this "exome plus array" dataset are described in supplemental materials (**Tables S3-S4**, **Figure S2**).

Validation of CNV

We attempted validation of 55 deletions from the exome plus array dataset that affected any calcium channel gene (N = 26 genes) using a combination of both quantitative PCR (qPCR) and NanoString nCounter technology. First, qPCR was used to verify 6 CNVs detected in calcium channel genes *CACNA2D3*, *CACNA1B*, *CACNA2D4* and *CACNG2* (**Table S5**). Several predesigned TaqMan Copy Number Assays were run in quadruplicate along with the internal RNase P Copy Number Reference Assay according to manufacturer's instructions (Applied Biosystems, Foster City, CA). Briefly, 20 µl reactions containing 1 µl DNA (5 ng), 10 µl of 2X Taqman Genotyping Master Mix, 1 µl of one target CNV assay and 1 µl of RNase P reference assay were mixed. All qPCR reactions were run on a Life Technologies StepOnePlus machine with the following thermal cycling conditions: 95°C for 10 min, followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. Samples included all suspected CNV carriers for each gene, regardless of case or control status, as well as four presumed two-copy controls per gene.

Second, for a larger scale validation, we used Nanostring nCounter technology. For each CNV, two probes were designed and analyses were performed according to manufacturer instructions. In brief, a spike-in plasmid of known amount was used to control for variability in DNA quantity across all samples and additional controls ensured optimal hybridization and purification efficiency. After hybridization and removal of excess probes, the probe/target complexes were aligned and immobilized in the nCounter Cartridge, and imaged in the nCounter Digital Analyzer for detection of CNVs. In a previous study, we examined nCounter's CNV calling accuracy by testing 37 known CNVs in 384 samples and found 97% concordance. We were able to successfully attempt validation for 48 of the 55 deletions.

Results

Exome-sequencing CNVs demonstrate high concordance with genotyping array based CNVs while contributing substantial numbers of novel variants

We generated CNVs using XHMM for 4,913 SCZ cases and 6,188 controls resulting in a total of 14,243 rare (present in less than 1% of individuals) and high quality (SQ $\geq$ 60) CNVs ("exome QC dataset"). In a comparison to previously published CNVs from genotyping arrays on these individuals(25) (see Supplementary Methods) we identified 78% of the array-based CNVs in the exome QC dataset. More interestingly, 75% of the exome QC calls were not seen in the array-based call set. Individuals carried, on average, 2.2 times more CNVs in the exome QC dataset than in the array-based call set (1.28 versus 0.59 CNVs). This comparison is described in more detail(22). Specific to this work, 53% of exome QC CNVs overlapped a single protein coding gene

(94% had length < 100kb) and, of those, only 12.6% were included in the previous work on this sample leaving 87.4% or 6,622 single-gene CNVs to be analyzed for the first time here.

Significant burden of exome-sequencing based CNVs in SCZ including among single-gene CNVs

We first assessed the burden of all CNVs in the exome QC dataset to SCZ. Utilizing empirical permutation of case/control label (see Methods) we identified a significant increase in the numbers of deletions (case rate: 0.56, control rate: 0.51, p = 0.0004) and duplications (case rate: 0.78, control rate: 0.72, p = 0.0006) in SCZ cases compared to controls as seen previously in this sample(25). To identify the contribution of the novel CNVs in our exome QC dataset, we performed the same burden test using only CNVs new to this analysis and not called by arrays in previous work. Here, we again saw significant burden in cases for both deletions (case rate: 0.48, control rate: 0.45, p = 0.0114) and duplications (case rate: 0.56, control rate: 0.51, p = 0.0003). The exome QC CNVs are substantially shorter and therefore more likely to affect only a single gene. We tested whether burden of CNVs was primarily driven by larger events affecting multiple genes or if single-gene CNVs were contributing. We identified a significant but modest burden of single-gene deletions (case rate: 0.36, control rate: 0.34, p = 0.0395) and duplications (case rate: 0.34, control rate: 0.32, p = 0.0332) in SCZ cases compared to controls (***Figure 1***, ***Table 1***). These results were not driven by CNV length in deletions (p < 100kb = 0.071, p > 100kb = 0.072) or duplications (p < 100kb = 0.053, p > 100kb = 0.181).

Expanding the set of potential single-gene CNVs and testing for excess in specific genes

We next sought to test whether CNVs could implicate specific genes using both the exome QC dataset as well as an expanded exome dataset created to increase the proportion of shorter CNVs which our QC filters were disproportionately removing (see Methods). Briefly, we integrated CNVs called from both XHMM and ExomeDepth(29) retaining CNVs if detected by both methods regardless of XHMM quality scores or if detected only by XHMM at our previous filtering threshold (SQ ≥ 60). In total, our "expanded exome dataset" included an additional 10,600 CNVs (total: 24,843) substantially increasing the proportion of shorter events (Figure S1). Individual genes were tested for excess of deletions or duplications using empirical permutation. After 10,000 permutations in our exome QC dataset, 21 genes were significantly enriched for duplications and 40 genes were significantly enriched for deletions in cases compared to controls after multiple test correction. All significant genes fell into two genomic regions of already known large SCZ risk CNVs, 16p11.2 (duplications) and 22q11.2 (deletions) leaving no novel genes identified (***Figure 2***). Finally, using our expanded exome dataset we again tested for enrichment of deletions and duplications in specific genes. No gene was significant after correction with the most significant genes again being driven by the larger 16p11.2 or 22q11.2 CNVs.

Testing contribution of only single-gene CNVs to previously implicated SCZ gene sets

In the absence of any novel genes being identified above, we tested whether single-gene CNVs were enriched among previously implicated gene sets. In the expanded exome dataset, there were 14,091 CNVs affecting only a single protein-coding gene (7,423 deletions, 6,668 duplications) and 7,703 affecting multiple genes (2,443 deletions, 5,260 duplications). The sets tested included genes previously implicated in other SCZ studies (GWAS loci(32), *de novo* variants(33), CNV regions(6)), synaptic function(5) (ARC, mGluR5, NMDAR, PSD95), calcium channels(26)

(CAV2, Voltage-gated), secondary sets (FMRP targets(23,26), ASD/DD/ID *de novo*(33), essential genes(34), constrained genes(35), RBFOX related genes(23) and antipsychotic targets(36)). The combined set of genes across all sets (n = 8,970) showed significant excess in cases for single-gene deletions (p = 0.008) but not duplications (p=0.186). We identified nominally significant enrichment of single-gene deletions in over half (11 out of 21) of the sets (***Table 2***), however no gene set surpassed a Bonferroni corrected p-value of 0.0005 for the 88 tests performed. For comparison, multi-gene deletions also showed nominally significant enrichment in 11 of the 21 sets, including 6 that surpassed Bonferroni correction (all genes, DD *de novo*, ID *de novo*, constrained, essential and SCZ deletion regions). Multi-gene duplications were significantly enriched in 7 of the 11 sets but none survived correction including all genes (p=0.007).

Broad scale exploration of CNVs in calcium channel genes combining both the expanded exome dataset and array-based calls

Among the most significant gene sets, we selected voltage-gated calcium channels for a full-scale validation since it represented an approachable number of CNVs to validate comprehensively and had significant prior supporting literature. Across the 26 genes, we identified 6 deletions in cases and 0 in controls from our expanded exome dataset (***Figure 3***). Since validation with an independent technology is considered the gold standard for CNV work, we attempted to validate these deletions using quantitative PCR (qPCR). Four of the deletions validated, two identical single exon deletions in *CACNA2D3* did not validate (these two did not surpass filtering thresholds to be included in the exome QC dataset). Since we had additional CNV data from genotyping arrays, we wanted to validate a larger set of calcium channel deletions to more comprehensively catalog the contribution of deletions in these genes to risk of SCZ in this sample. We identified a set of

deletions across our exome plus array dataset (see Methods and Supplementary Methods) overlapping any voltage-gated calcium gene. In total, we identified 55 deletions in 55 different samples of which 48 could be tested using NanoString nCounter technology (see Methods). Of these, 34 were located over three common, intronic copy number polymorphisms (all of which validated). Of the 21-remaining rare-variant calls, 6 validated (see ***Table 3***). The low validation rate is representative of our decision to take all CNVs with limited evidence and not filter on confidence. Nearly all of the CNVs that did not validate were low quality calls from the genotyping arrays. Four of the validated deletions were single-gene, in cases and identified in the expanded exome dataset. The remaining two validated deletions were a non-exonic deletion and a multi-gene deletion in a control identified from genotyping arrays. After validation, we were left with 4 single-gene deletions in cases and 0 in controls ($p = 0.039$).

Discussion

This study represents an evaluation of smaller CNVs in a large SCZ sample. We found that, independent of larger events, deletions of single genes may contribute to schizophrenia risk through a number of biological pathways previously identified for SCZ. In particular, we identify and validate a small number of deletions in voltage-gated calcium channels that are enriched in SCZ cases compared to controls. We also demonstrate the utility of exome-sequencing to identify shorter, single-gene CNVs and the potential to improve the resolution of those events through combining multiple methods for further study.

To date, the contribution of CNVs to SCZ risk has been predominantly from large (>100kb) and rare CNVs both in specific loci and in aggregate across the genome(6). The ability to determine the contribution from shorter CNVs has been both technologically limited by the use of genotyping arrays but also biologically up for debate as few single genes have been implicated in SCZ risk

and nearly all risk increasing CNVs affect many genes. Here, we point to the potential contribution of single-gene CNVs to risk for SCZ. This contribution can be identified both genome-wide and within genes having been previously implicated from other studies of genetic variation including synaptic genes, genes having *de novo* mutations in SCZ, DD, ASD or ID, conserved genes and gene targets of antipsychotics. In comparison, multi-gene CNVs showed more significant enrichment among these sets including 6 surpassing correction but the majority of significant sets were shared between single-gene and multi-gene CNVs. Many of these gene sets were discovered from large CNV analyses making these results already known. Also, many of the large CNVs implicated in SCZ also contribute to other related phenotypes such as DD and ID. More interestingly, there were 4 sets that showed significant enrichment only in single-gene CNVs that included calcium channels and SCZ LoF *de novo* variants pointing to potential examples of variants of large effect on SCZ risk that have not yet been seen in the larger CNVs. This work points to a confluence of evidence that these gene sets are relevant for schizophrenia biology. We did not identify any specific gene that was significantly associated after correction for multiple testing. Given other studies of rare variation in complex diseases with similar sample sizes, this is not surprising(25) but our results suggest that combining CNV data with SNV data could improve power to implicate specific genes and robust approaches to combine these classes of variation are needed. For these approaches, leveraging knowledge of how intolerant a gene is to variation, thereby weighting variants by their potential impact may also improve discovery. Further, while the overall contribution of CNVs to SCZ risk is modest and the contribution from single-gene CNVs is even less, the addition of CNV burden to measures of individual risk such as polygenic risk scores could offer improvements in risk stratification and should be fully assessed.

Calcium channel genes have been implicated in psychiatric disease risk, including in SCZ for many years. Studies to date from the genetics of SCZ have implicated both particular loci and the geneset as a whole. Here, we show an excess of single-gene CNVs in calcium channels among SCZ patients that remains after qPCR validation. Given the importance of this gene set and the relative size, we also performed a larger validation of deletions using a higher throughput method that again confirmed 4 qPCR validated single-gene deletions in cases as well as validating several common CNVs, one >2Mb deletion in a control and one deletion that did not overlap an exon in a control. Our results suggest that deleting a single calcium channel gene may be relevant for SCZ risk however substantially more data will be required to confirm this finding.

We show that exome-sequencing can identify a substantial number of novel CNVs that are not captured by genotyping arrays and are predominantly affecting only a single gene. Further, this work points to the existence of many real single-gene CNVs that are filtered out by default filtering criteria and by combining multiple currently existing approaches we can capture an expanded set of true calls. While exome-sequencing can substantially improve resolution of CNV calling it is not without its weaknesses and limitations that become even clearer as CNVs get smaller. Whole-genome sequencing will offer the best resolution to confidently identify single-gene CNVs but is still prohibitively expensive for most labs and hundreds of thousands of exome sequences currently exist, and many more are being generated, making CNV calling from exome-sequencing still important. We believe there are opportunities to improve the ability to call shorter CNVs from exome-sequencing that are more sophisticated than merging call sets from multiple approaches and there is continued effort in this area that we anticipate will provide additional value to CNV calling from exome-sequencing.

Here, we demonstrate a potential role for single-gene deletions to contribute to SCZ risk through similar pathways as previously implicated. We perform a comprehensive validation of deletions in voltage-gated calcium channel genes and show an enrichment of these deletions in SCZ cases compared to controls. Finally, we demonstrate further utility for CNV generated from exome-sequencing and the ability to improve resolution of shorter events which could improve our ability to identify biological causes of diseases like SCZ.

Acknowledgements

This work was supported by NIMH R01 MH111776 (DMR) and R21 MH104831 (JPS, JJC). PFS gratefully acknowledges support from the Swedish Research Council (Vetenskapsrådet, award D0886501). The Sweden Schizophrenia Study was supported by NIMH R01 MH077139.

Declaration of Conflicts of Interest

The authors declare no conflicts of interest.

Figure Legends

Figure 1. Burden tests across all high confident exome-seq CNVs (all), those not previously analyzed from genotyped arrays (new), those previously published (published) and only those CNVs affecting a single protein coding gene (single-gene). Deletions are in red (left) and duplications are in blue (right). Significance is represented as $p < 0.05$ (*), $p < 0.001$ (**).

Figure 2. Gene-based Manhattan plot of duplications in blue (top) and deletions in red (bottom). Genes in most significant regions are labeled by known CNVs in that region.

Figure 3. Gene model plots for each of the 4 genes and 6 deletions identified in voltage-gated calcium channel genes. Upper grey bars portray deletion in genomic space, below that is the gene model in genomic space. The bottom bars represent the exons as transcribed, red indicates exons that were deleted. All deletions replicated except the two shown in panel B.

		Deletions				*Duplications*			
		N	**Case rate**	**Control rate**	**P**	**N**	**Case rate**	**Control rate**	**P**
All	*All*	5900	0.56	0.51	**0.0004**	8343	0.78	0.73	**0.0006**
	New	5101	0.48	0.45	**0.0114**	5925	0.56	0.51	**0.0003**
	Previously called by genotyping arrays	799	0.08	0.06	**0.0002**	2418	0.23	0.21	0.1396
Single gene	*All*	3894	0.36	0.34	**0.0395**	3680	0.34	0.32	**0.0332**
	New	3530	0.33	0.31	0.0543	3092	0.29	0.27	**0.0162**
	Previously called by genotyping arrays	364	0.03	0.03	0.1998	588	0.05	0.05	0.6180
Multiple genes	*All*	1773	0.18	0.14	**0.0001**	4516	0.42	0.39	**0.0030**
	New	1339	0.13	0.11	**0.0174**	2695	0.25	0.23	**0.0076**
	Previously called by genotyping arrays	434	0.05	0.03	**0.0001**	1821	0.17	0.16	0.0756

Table 1. CNV burden results stratified by CNV type (deletions, duplications), number of genes affected (all, single gene or multiple genes) and whether the CNV was unique to our exome-sequencing call set or was identified in previous array-based CNV work. Bolded p-values are less than 0.05.

			Single gene CNV		*Multi gene CNV*	
Group	**Set**	**N genes**	**P del**	**P dup**	**P del**	**P dup**
SCZ sets	PGC2 SCZ 108 loci	329	0.4199	0.7021	0.6372	**0.0260**
	SCZ de novo LoF	87	**0.0296**	0.0911	0.4697	0.6710
	SCZ de novo NS	611	0.2761	0.9024	0.0857	0.5107
	PGC2 16 CNV	175	**0.0055**	0.9989	**0.0451**	**0.0212**
	PGC2 16 CNV (deletions)	78	0.2594	0.9289	**0.0001**	0.9602
	PGC2 16 CNV (duplications)	111	**0.0077**	0.5790	0.0757	**0.0082**
Synaptic Sets	ARC	28	1.0000	0.3543	0.4426	0.5778
	mGluR5	39	0.2425	0.8117	0.2168	0.2144
	NMDAR_network	61	**0.0335**	0.8099	**0.0180**	**0.0243**
	PSD-95_(core)	65	0.0697	0.8984	**0.0009**	0.2630
Calcium Channel Sets	CAV2	206	**0.0347**	0.4065	0.1624	0.4822
	CAV2 Ion	44	0.1738	0.2395	0.3951	0.6500
	Voltage-gated_Calcium_Channel_Genes	26	**0.0082**	0.2792	0.9953	0.8284
Secondary Sets	FMRP-targets	788	**0.0205**	0.5018	**0.0216**	**0.0123**
	ASD de novo	1080	0.0521	0.4664	0.0796	0.1123
	DD de novo	1271	**0.0161**	0.9660	**0.0003**	0.1308
	ID de novo	350	0.0898	0.5774	**0.0001**	0.5938
	Antipsychotic targets	347	**0.0268**	0.9338	**0.0342**	0.7322
	Essential genes	3915	0.0622	0.1975	**0.0001**	0.0987
	Rbfox	2737	**0.0058**	0.2056	**0.0087**	**0.0200**
	LoF intolerant (pLI > 0.9)	3488	**0.0163**	0.2430	**0.0001**	**0.0041**

Table 2. Geneset CNV results for single-gene and multi-gene CNVs in expanded dataset. Bold represents pathways with p-value < 0.05.

Status	Chr	Start	Stop	Size (bp)	Gene	Single-gene?	Exonic?
case	9	140866027	141004005	137978	CACNA1B	yes	yes
case	9	140846726	141016451	169725	CACNA1B	yes	yes
case	12	1949932	1965357	15425	CACNA2D4	yes	yes
case	22	36960396	36960935	539	CACNG2	yes	yes
control	3	54262746	54316431	53685	CACNA2D3	yes	**no**
control	7	79818265	82072777	2254512	CACNA2D1	**no**	yes

Table 3. List of all rare deletions overlapping the 26 voltage-gated calcium channel genes that validated including one that did not overlap an exon and one that was not single-gene. Four of the 6 single-gene deletions identified in the geneset analyses and in figure 3 validated (CACNA2D3 did not).

References

1. Saha S, Chant D, McGrath J. A Systematic Review of Mortality in Schizophrenia: Is the Differential Mortality Gap Worsening Over Time? Arch Gen Psychiatry. 2007 Oct 1;64(10):1123–31.

2. WHO | The global burden of disease: 2004 update [Internet]. WHO. [cited 2018 Mar 6]. Available from: http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/

3. Knapp M, Mangalore R, Simon J. The global costs of schizophrenia. Schizophr Bull. 2004;30(2):279–93.

4. Saha S, Chant D, Welham J, McGrath J. A Systematic Review of the Prevalence of Schizophrenia. PLOS Med. 2005 May 31;2(5):e141.

5. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol Psychiatry. 2012 Feb;17(2):142–53.

6. CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017 Jan;49(1):27–35.

7. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet. 2012 Aug;13(8):537–51.

8. Malhotra D, Sebat J. CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. Cell. 2012 Mar 16;148(6):1223–41.

9. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J, et al. Copy Number Variants in Schizophrenia: Confirmation of Five Previous Findings and New Evidence for 3q29 Microdeletions and VIPR2 Duplications. Am J Psychiatry. 2011 Mar;168(3):302–16.

10. Guha S, Rees E, Darvasi A, Ivanov D, Ikeda M, Bergen SE, et al. Implication of a Rare Deletion at Distal 16p11.2 in Schizophrenia. JAMA Psychiatry. 2013 Mar 1;70(3):253–60.

11. Liao H-M, Chao Y-L, Huang A-L, Cheng M-C, Chen Y-J, Lee K-F, et al. Identification and characterization of three inherited genomic copy number variations associated with familial schizophrenia. Schizophr Res. 2012 Aug 1;139(1):229–36.

12. Kirov G, Rujescu D, Ingason A, Collier DA, O'Donovan MC, Owen MJ. Neurexin 1 (NRXN1) Deletions in Schizophrenia. Schizophr Bull. 2009 Sep;35(5):851–4.

13. Stoll G, Pietiläinen OPH, Linder B, Suvisaari J, Brosi C, Hennah W, et al. Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. Nat Neurosci. 2013 Sep;16(9):1228–37.

14. Steinberg S, Gudmundsdottir S, Sveinbjornsson G, Suvisaari J, Paunio T, Torniainen-Holm M, et al. Truncating mutations in RBM12 are associated with psychosis. Nat Genet [Internet]. 2017 Jun 19 [cited 2017 Jun 20];advance online publication. Available from: https://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3894.html

15. Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, et al. Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. Nat Neurosci. 2016 Apr;19(4):571–7.

16. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct;526(7571):75–81.

17. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011 Feb;470(7332):59–65.

18. Stankiewicz P, Lupski JR. Structural Variation in the Human Genome and its Role in Disease. Annu Rev Med. 2010;61(1):437–55.

19. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of Human Copy Number Variation and Multicopy Genes. Science. 2010 Oct 29;330(6004):641–6.

20. Szatkiewicz JP, Neale BM, O'Dushlaine C, Fromer M, Goldstein JI, Moran JL, et al. Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. Mol Psychiatry. 2013 Nov;18(11):1178–84.

21. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011 May;12(5):363–76.

22. Ruderfer DM, Hamamsy T, Lek M, Karczewski KJ, Kavanagh D, Samocha KE, et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat Genet. 2016 Oct;48(10):1107–11.

23. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landén M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci. 2016 Nov;19(11):1433.

24. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013 Oct;45(10):1150–9.

25. Szatkiewicz JP, O'Dushlaine C, Chen G, Chambert K, Moran JL, Neale BM, et al. Copy number variation in schizophrenia in Sweden. Mol Psychiatry. 2014 Jul;19(7):762–73.

26. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014 Feb 13;506(7487):185–90.

27. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012 Oct 5;91(4):597–607.

28. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. *De novo* mutations in schizophrenia implicate synaptic networks. Nature. 2014 Feb;506(7487):179–84.

29. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012 Nov 1;28(21):2747–54.

30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep 1;81(3):559–75.

31. Raychaudhuri S, Korn JM, McCarroll SA, Consortium TIS, Altshuler D, Sklar P, et al. Accurately Assessing the Risk of Schizophrenia Conferred by Rare Copy-Number Variation Affecting Genes with Brain Function. PLOS Genet. 2010 Sep 9;6(9):e1001097.

32. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul 24;511(7510):421–7.

33. Nguyen HT, Bryois J, Kim A, Dobbyn A, Huckins LM, Munoz-Manchado AB, et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Med. 2017 Dec 20;9:114.

34. Ji X, Kember RL, Brown CD, Bućan M. Increased burden of deleterious variants in essential genes in autism spectrum disorder. Proc Natl Acad Sci. 2016 Dec 27;113(52):15054–9.

35. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.

36. Ruderfer DM, Charney AW, Readhead B, Kidd BA, Kähler AK, Kenny PJ, et al. Polygenic overlap between schizophrenia risk and antipsychotic response: a genomic medicine approach. Lancet Psychiatry. 2016 Apr 1;3(4):350–7.