January 2020

Identifying Rare Genetic Variation In Obsessive-Compulsive Disorder

Sarah Abdallah

Identifying Rare Genetic Variation in Obsessive-Compulsive Disorder

A Thesis Submitted to the Yale University School of Medicine
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Medicine

by
Sarah Barbara Abdallah
2020

ABSTRACT

IDENTIFYING RARE GENETIC VARIATION IN OBSESSIVE-COMPULSIVE DISORDER

Sarah B. Abdallah, Carolina Cappi, Emily Olfson, and Thomas V. Fernandez. Child Study Center, Yale University School of Medicine, New Haven, CT

Obsessive-compulsive disorder (OCD) is a neuropsychiatric developmental disorder with known heritability (estimates ranging from 27%-80%) but poorly understood etiology. Current treatments are not fully effective in addressing chronic functional impairments and distress caused by the disorder, providing an impetus to study the genetic basis of OCD in the hopes of identifying new therapeutic targets. We previously demonstrated a significant contribution to OCD risk from likely damaging *de novo* germline DNA sequence variants, which arise spontaneously in the parental germ cells or zygote instead of being inherited from a parent, and we successfully used these identified variants to implicate new OCD risk genes. Recent studies have demonstrated a role for DNA copy-number variants (CNVs) in other neuropsychiatric disorders, but CNV studies in OCD have been limited. Additionally, studies of autism spectrum disorder and intellectual disability suggest a risk contribution from post-zygotic variants (PZVs) arising *de novo* in multicellular stages of embryogenesis, suggesting these mosaic variants can be used to study other neuropsychiatric disorders. In the studies presented here, we aim to characterize the contribution of PZVs and rare CNVs to OCD risk.

We examined whole-exome sequencing (WES) data from peripheral blood of 184 OCD trio families (unaffected parents and child with OCD) and 777 control trios that passed quality control measures. We used the bioinformatics tool MosaicHunter to

identify low–allele frequency, potentially mosaic single-nucleotide variants (SNVs) in probands (OCD cases) and in control children. We then applied the XHMM tool to 101 of the OCD trio families and to the 777 control trio families, all generated with the same capture library and platform, to identify CNVs.

The rate of all single-nucleotide PZVs per base pair was not significantly different between OCD probands (4.90 x 10^{-9}) and controls (4.93 x 10^{-9}), rate ratio = 0.994, p = 1. The rate of likely-damaging PZVs (those altering a stop codon or splice site) also is not significantly different in OCD probands (1.45 x 10^{-9}) than in controls (1.09 x 10^{-9}), rate ratio = 1.33, p = 0.653.

When examining CNVs, the proportion of children with at least one rare duplication or deletion is not significantly different between OCD cases (0.869) and controls (0.796), chi-square = 2.97, p = 0.0846. However, when considering deletions separately from duplications, the proportion of children with at least one rare deletion is higher in OCD trios (0.606) than in controls (0.448), chi-square = 8.86, p = 0.00292.

Although we did not detect a higher burden of PZVs in blood in individuals with OCD, further studies may benefit from examining a larger sample of families or from looking for PZVs in other tissues. The higher rate of *de novo* deletions in cases vs. controls suggests they may contribute to OCD risk, but further work is needed to experimentally validate the detected CNVs. We hope to eventually use these CNVs to identify OCD risk genes that could provide jumping-off points for future studies of molecular disease mechanisms.

Acknowledgements: Special thanks to my research mentor, Tom Fernandez, for supervising this thesis and to Emily Olfson for her advice and contributions to this work. They have been lovely, brilliant, and encouraging people to work with. I also have appreciated the encouragement from other members of the Child Study Center and their efforts to create a welcoming work environment. Thanks to my parents and friends for supporting my efforts to pursue this sort of work and helping me through the growing pains. Additional thanks to the Yale Office of Student Research for their support.

The research included in this thesis was funded by grants from the Allison Family Foundation, Brain and Behavior Research Foundation (NARSAD), and the National Institute of Mental Health under award number R01MH114927 (TVF) and by research fellowship funding from the Howard Hughes Medical Institute, American Society for Human Genetics, and American Academy of Child and Adolescent Psychiatry (SBA).

Table of Contents

INTRODUCTION

Features of Obsessive-Compulsive Disorder

Obsessive-compulsive disorder (OCD) is a developmental neuropsychiatric disorder with estimated prevalence of 1-3% worldwide. It is characterized by disabling obsessions (intrusive, unwanted thoughts, sensations, or urges) and compulsions (ritualized, repetitive behaviors that are difficult to control) (1). These symptoms can cause distress, significantly compromise the affected individual's social and occupational functioning, and lead to increased risk of mortality, such that the World Health Organization has named OCD among the ten most disabling medical conditions worldwide (2). Although serotonergic antidepressants have been used in the treatment of OCD for several decades, these pharmacologic treatments are not completely effective, producing 30-50% reduction of symptoms in 60-80% of patients, and untreated OCD tends to persist and become chronic (2, 3). The main barrier to developing more effective therapeutic options for OCD is a poor understanding of its underlying etiology. For this reason, there is great incentive to study the molecular basis of the disorder in the hopes of identifying new therapeutic targets.

Like many neuropsychiatric disorders, OCD has high clinical heterogeneity, with a wide range of possible symptoms and severity, such that different patients with the disorder may have little to no phenotypic overlap. Efforts to better understand this heterogeneity have used factor-analytic and clustering approaches to identify symptom dimensions or subtypes in OCD (4-6). However, large-scale genetic studies generally group together phenotypically divergent patients, potentially diluting genetic signals that may be specific to a subgroup of patients. Further complicating efforts, OCD often is comorbid with other neuropsychiatric disorders, namely tic disorders, creating the potential for confounding signals in genetic studies (5, 6).

OCD is thought to arise from a combination of genetic and environmental factors. Twin and family studies have demonstrated substantial heritability of OCD, with estimates around 27-47% for adult-onset cases and 40-80% for early-onset (childhood) OCD (1, 7-15). Despite evidence for a significant genetic contribution to OCD pathogenesis, risk gene discovery efforts have had little success so far, and the underlying genetic basis of the disorder remains poorly understood. It is challenging to identify these responsible genetic variants and genes because OCD is highly polygenic, meaning many genes contribute to the disorder, and the combination of genetic factors contributing OCD risk differs between patients (15-17). Current prevailing wisdom suggests a combination of small-effect common variants and large-effect rare variants, either inherited from parents or arising spontaneously, in hundreds of genes and within the intergenic space contribute to OCD pathogenesis (16, 17). This complexity requires geneticists to draw from different types of genetic information and methods of analysis to statistically implicate risk genes.

Approaches to Studying OCD Genetics

Investigations into the genetic basis of OCD have taken several approaches to uncovering the relevant genes, types of variation, and biological pathways involved in the disorder (7, 15). The following section examines the relative success and findings of these approaches to date.

Association Studies

To date, few genome-wide association studies (GWAS) exploring the contribution of common genetic variation to OCD have been conducted. Stewart et al. (18) performed a meta-analysis of 1,465 cases, 5,557 ancestry-matched controls, and 400 parent-child trios, while Mattheisen et al. (19) examined 1,406 individuals with OCD from 1,065 families. In the individual studies and a meta-analysis of both by the International OCD Foundation (20), no loci reached genome-wide statistical significance ($p < 5 \times 10^{-8}$) in the final analyses. While GWAS overall have been unsuccessful in identifying reproducible genetic associations with OCD, common variants of small effect sizes are thought to contribute partially to OCD heritability, and the lack of success with GWAS so far may be due to insufficient sample sizes (16, 18, 19, 21). One would expect that a relatively large proportion of loci approaching genome-wide significance would cross the significance threshold in future GWAS with larger sample sizes. By this supposition, overall trends or pathway enrichment among genes in these loci may still point to relevant biology.

In contrast with the hypothesis-free nature of GWAS, candidate gene association studies focus on single nucleotide polymorphisms (SNPs) within a preselected gene hypothesized to be biologically relevant to a disease. While over 100 of these studies have been conducted in OCD, few consistent findings have been reported (1, 8). Due to issues of publication bias and failure to account for environmental and genetic background of participants, among other factors, candidate gene studies are prone to false positive results that largely have not been replicated (22-27). Further, many lack the sample size needed to detect the small effects expected for complex disorders like OCD

(26, 28). A meta-analysis of 230 polymorphisms from 113 candidate association studies found a statistically significant association between OCD and alleles of two serotonergic genes (*5-HTTLPR* and *HTR2A*) among all patients; among males only, it found a significant association between OCD and *COMT* and *MAOA* alleles (28). Since the publication of this meta-analysis, replicability of these results has been mixed, with successful replication of the association with OCD for the common L_A allele of *5-HTTLPR* but not for gene polymorphisms of *HTR2A, COMT*, and *MAOA* (29-31). Unfortunately, because the genes or loci of interest are selected based on presupposition, candidate gene studies are less useful in uncovering novel biology underlying disease pathogenesis.

Rare Variation in Psychiatric Disease

While the aforementioned association studies attempt to pinpoint common variation contributing to disease risk, other study designs leverage information about rare variation to infer biology underlying disease. Investigation of rare variation in autism spectrum disorder (ASD) has successfully associated several genes with ASD risk and implicated specific brain regions and developmental timepoints in its pathogenesis (32), suggesting these approaches hold promise.

Linkage Studies of Rare Inherited Variants

Because a child inherits about four to five million rare variants from their parents, there is low statistical power to detect which of these variants fall in disease risk genes and are contributing to disease risk in a patient cohort. Further, because inherited variants are subject to natural selection pressure while passing through generations, those that persist are unlikely to have high damaging capacity (33). Thus, the utility of these variants in implicating disease risk genes is limited to cases of families with multiple affected individuals carrying very rare, large-effect inherited variants. In these families, linkage studies can identify putative causal variants that associate with affected status within the family (34). While several genome-wide linkage studies have been conducted in OCD, few loci have reached genome-wide statistical significance and none have been replicated (35-39).

De Novo Variation

De novo variants arise spontaneously in the child due to DNA replication errors and are not inherited from parents. In contrast to inherited variants, *de novo* single-nucleotide variants arising in the germline (egg or sperm) or zygote are infrequent, occurring on average 44-82 times throughout a person's genome and only once or twice in the coding regions, or exome (33). This rarity makes them much more useful for detecting disease risk genes across cohorts. Genetic studies of other psychiatric disorders have successfully harnessed *de novo* variants as a powerful means of identifying disease risk genes (40-43). Recently, our group has applied this approach to OCD (see preliminary studies) with success (44).

Post-Zygotic Variants

Post-zygotic variants (PZVs), *de novo* variants arising soon after conception rather than in the parental germ cells, produce a mosaic child with the variant in only a fraction of cells throughout the body. Figure 1 depicts the different developmental timepoints at which germline *de novo* variants and PZVs arise. In contrast to oncogenic somatic mutations that can accumulate over an individual's lifetime, PZVs occur in early embryogenesis and theoretically should appear in multiple cell and tissue types descended from the original embryonic cell. With high depths of coverage, next-generation sequencing allows for detection of potential mosaic variants based on the observed mutant allele fraction, or the fraction of DNA segments with the variant allele at a genomic position. Germline *de novo* variants theoretically should have a mutant allele fraction of 50%, so any variants below a certain cutoff (e.g. 30%) are discarded as likely technical artifacts (45). However, PZVs should have a mutant allele fraction far below 50% and likely produce true signal buried among these discarded variants.

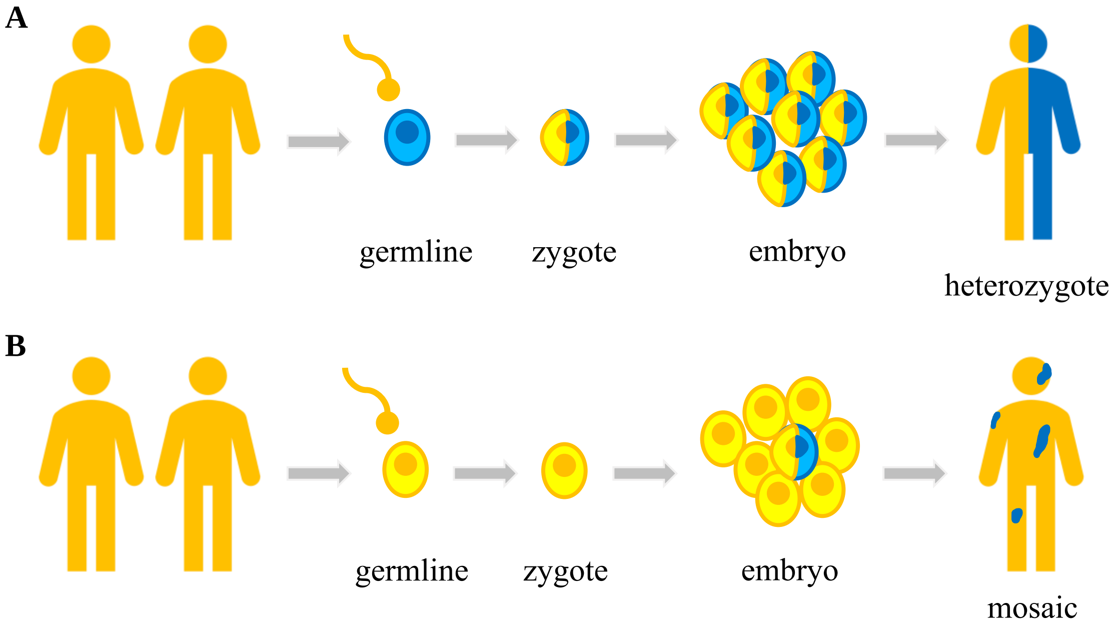

Figure 1. Consequences of spontaneous variants in offspring. **(A)** A germline *de novo* variant arises in one parental germ cell and propagates through all cells of the child's

body, producing a child who is heterozygous for the variant. **(B)** After the zygote has split into a multicellular embryo, a PZV arises in one of the cells and propagates through the cell's descendants, producing a child who is mosaic for the variant.

PZVs have been of recent interest in the study of several neuropsychiatric disorders but are poorly understood within the context of these disorders. Recent studies looking at previously identified de novo variants in ASD (46-49) and intellectual disability (50) have shown that 5.8% and 6.5%, respectively, were in fact post-zygotic rather than germline mutations. Several studies found that PZVs were enriched (more frequent) in ASD probands (clinically affected individuals with unaffected parents and siblings) compared to their unaffected siblings, and by one estimate the detected PZVs contributed to 5.1% of ASD diagnoses, suggesting a role for somatic mosaicism in ASD (46-49). These findings suggest that mosaic variation may provide a fruitful avenue to examine the genetic underpinnings of neuropsychiatric disorders and may contribute clinically meaningful genetic risk that previously was overlooked.

Structural (Copy Number) Variation

Examination of chromosomal structural variation, defined as variation in DNA segments over one kilobase (kb) in length, has suggested a role in OCD pathogenesis. Early cytogenetic and locus-specific studies of OCD cases identified inversions or translocations of large DNA segments that converged on overlapping chromosomal locations (15, 51). DNA microarrays, which provide better genome-wide resolution than older cytogenetic techniques such as karyotyping, have improved detection of copy-number variants (CNVs; deletions or duplications of DNA sequences over one kb in length) in recent years. Three microarray studies of CNVs in OCD found no overall increased rate compared to controls. However, one study found that OCD cases harbored a significantly higher rate of large deletions overlapping regions implicated in other neurodevelopmental disorders, and the other two found a significantly higher rate of rare CNVs affecting genes related to neurological function (11, 51, 52).

While microarrays have improved resolution compared to older techniques like karyotyping and fluorescence *in situ* hybridization (FISH), they still are best at detecting larger CNVs with a lower limit of about 30 kb in size. In contrast, high-throughput sequencing approaches like WES can be used to more accurately detect small- to medium-sized CNVs, which are more frequent in number compared to large CNVs (33, 53). Rare exonic deletions of 1-30 kb size have been estimated to contribute to disease risk in up to 7% of ASD cases. Further, unlike large CNVs that typically contain multiple genes, small exonic CNVs typically affected just one gene, making them useful for risk gene discovery and pathway analysis (53). It is possible rare, smaller CNVs impart a previously undetected contribution to OCD pathogenesis as well and can provide new insights into underlying biology.

Preliminary Studies

Our group recently published the first analysis of rare inherited and germline *de novo* single-nucleotide variants (SNVs) and insertion-deletion variants (indels) in patients with OCD. The cohort collected for this study exclusively contained simplex probands (affected individuals with no known affected first-degree relatives) to increase the likelihood of detecting *de novo* variants. After quality control, analyses were conducted on whole-exome sequencing (WES) from peripheral blood in 184 OCD parent-proband trios (families comprising two unaffected parents and one affected child) and in 777 control trios (unaffected parents and child). Among this cohort, likely-damaging germline *de novo* variants were enriched in OCD probands compared to controls. These damaging variants include likely gene-disrupting variants (LGD; nonsense, frameshift, or splice site mutations) and missense mutations predicted to be damaging by the software PolyPhen2 (Mis-D). The study also estimated that *de novo* variants found within 335 genes contributed to risk in 22% of cases (44). These findings suggest a significant contribution of *de novo* SNVs and indels to OCD risk. Identification of these variants implicated two new OCD risk genes, *CHD8* and *SCUBE1*, based on gene-level recurrence, i.e. the presence of at least two damaging (LGD or Mis-D) *de novo* variants in the same gene in two unrelated probands.

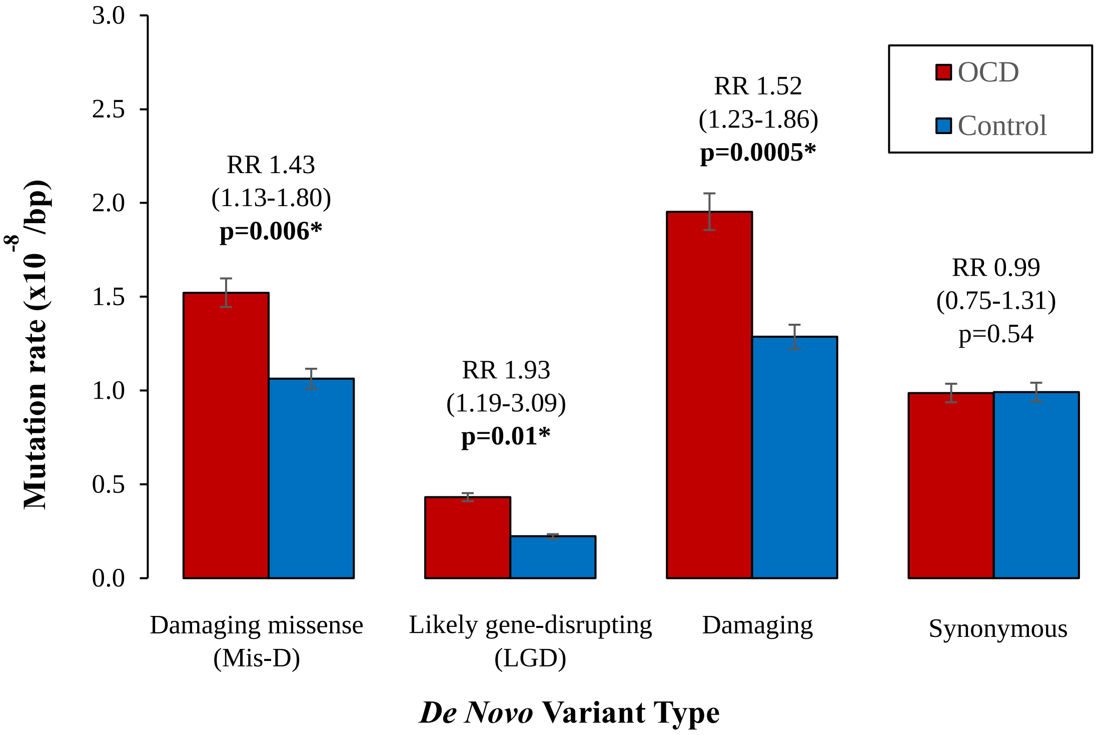

Figure 2. Germline *de novo* SNVs and indels in OCD probands vs. controls. Compared to control children, OCD probands have significantly higher rates of Mis-D, LGD, and total damaging germline *de novo* variants compared to controls. In contrast, synonymous variants, which do not affect a gene’s protein product, are not expected to contribute to OCD pathogenesis and are not more frequent in cases compared to controls. Figure modified from Cappi et al. (44).

With an increased sample size of trios, we expect to identify additional risk genes, particularly among the set of genes with one identified damaging variant to date. These studies are underway. In the meantime, we can extend the value of our current sample by identifying different types of genetic variants within our WES data. These variants may account for some missing information about OCD's genetic basis and can provide additional information to use in risk gene analyses.

Statement of Purpose and Specific Aims

We intend to build on our previous work using rare genetic variation detected in WES of OCD trios to gain insights into the underlying biology of OCD. The overarching purpose is to implement tools to identify two additional types of genetic variation from our WES data, characterize the contribution of that variation to OCD risk, and use those variants in statistical analyses to identify new potential OCD risk genes. These approaches have not yet been described in the literature and could provide promising new avenues to elucidate the genetic basis of OCD. This project will serve to fill a large knowledge gap by providing insight into OCD genetics, paving the way for further molecular and mechanistic studies of the disorder.

Aim 1: Characterize the Contribution of PZVs to OCD

The potential role of mosaic variation has not yet been described in the OCD literature but could add to our understanding of the genetic etiology of OCD. We aim to implement and optimize a computational approach to detect PZVs from WES data and to characterize the burden of PZVs in OCD cases versus control probands. With our depth of sequencing coverage in cases (76 reads per position on average) we can expect to detect over 95% of SMVs with a mutant allele fraction of at least 20% and over 90% of SMVs with a mutant allele fraction of at least 10% (54). Like our finding for damaging germline *de novo* variants, we hypothesize that PZVs predicted to be damaging will have an increased burden (occur at a greater frequency) in OCD probands compared to controls, suggesting a role for PZVs in OCD pathogenesis.

Aim 2: Characterize the Contribution of CNVs to OCD

The few studies that have explored the role of CNVs in OCD have used microarray data, which has limited resolution compared to sequencing. We anticipate we will be able to detect more CNVs from our WES data for OCD families. While WES covers only the exome (the coding region of the genome) and cannot be used to detect portions of CNVs in noncoding regions, we would expect the majority of the most clinically significant CNVs to occur in coding regions so that they will severely impact gene dosage. We aim to develop and optimize a computational approach to detect rare inherited and *de novo* CNVs from our WES of OCD and control trios. Based on previous findings in the literature, we expect to find an increased burden of deletions in probands compared to controls.

Aim 3: Identify New OCD Risk Genes and Biological Pathways

We will use the variants detected in the first two aims to identify putative OCD risk genes. Genes containing multiple germline or mosaic *de novo* variants or overlapping novel *de novo* CNVs will be deemed to possibly contribute OCD risk. We will construct networks of genes co-expressed across space and time in brain development and look for networks enriched for OCD risk genes, which could point to specific brain regions and developmental timepoints underlying OCD pathogenesis. Presuming correlated expression levels across space and time suggest similar function or regulation for a set of genes, we can associate other genes within these networks with OCD as well (32). We also will use gene ontology and pathway analysis tools to associate specific biological pathways with the set of risk genes.

MATERIALS AND METHODS

Data collection and processing

Participant recruitment, sample collection, and whole-exome sequencing (WES) were performed as described in Cappi et al., 2019 (44). In brief, we generated WES data from peripheral blood DNA of 222 parent-child OCD trios collected from sites in Toronto, Canada; São Paulo, Brazil; and New Haven, USA; and from a separate Tourette International Collaborative Genetics study that included patients with both OCD and chronic tics (55, 56). All samples were sequenced at the Yale Center for Genome Analysis (YCGA) using the NimbleGen SeqCap EZExomeV2 (109 trios) or MedExome (113 trios) capture libraries (Roche NimbleGen, Madison, WI) and the Illumina HiSeq 2000 platform (74-bp paired-end reads) (Illumina, San Diego, CA). These data were compared to WES from peripheral blood DNA in 855 control trios without OCD from the Simons Simplex Collection (57), sequenced at YCGA using the NimbleGen SeqCap EZExomeV2 and the Illumina HiSeq 2000 platform. These WES data were aligned using our lab's well-validated analysis pipeline following the latest Genome Analysis Toolkit (GATK) Best Practices guidelines (58). From this sample set, we retained 184 OCD trios (117 male probands; 67 female) and 777 control trios (356 male children; 421 female) that passed strict quality control measures, including removal of outlier trios based on principal component analysis of sequencing quality metrics (44).

Following sample collection and data processing, I performed all elements of the work described below, including the development and implementation of variant (PZV and CNV) calling approaches, mutation rate analyses, and risk gene and pathway analyses.

Variant Calling

In-house computational pipelines built from pre-existing tools were developed to detect PZVs and CNVs from WES data (Figure 2).

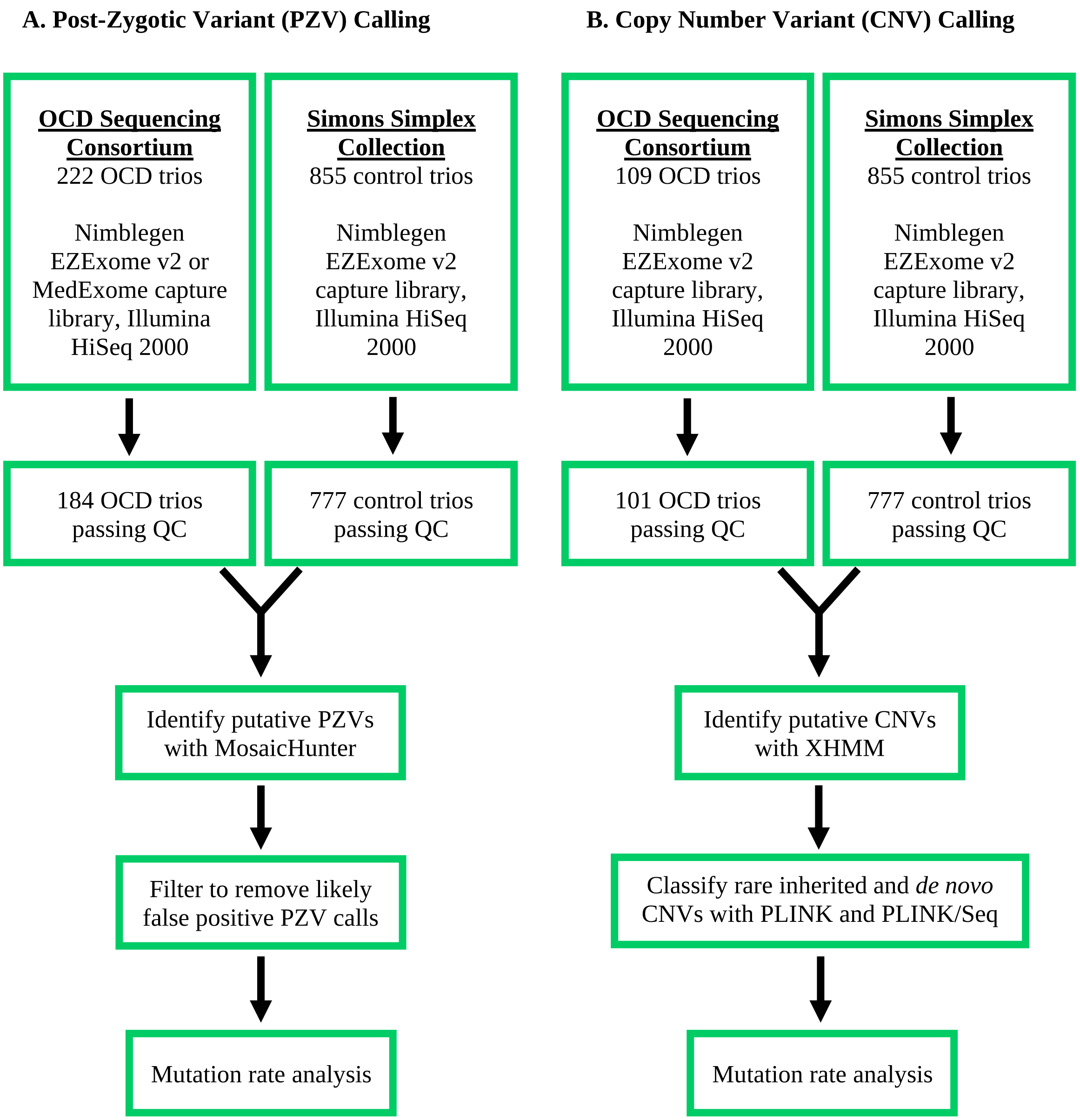
A. Post-Zygotic Variant (PZV) Calling
B. Copy Number Variant (CNV) Calling
OCD Sequencing Consortium
222 OCD trios
Nimblegen EZExome v2 or MedExome capture library, Illumina HiSeq 2000
Simons Simplex Collection
855 control trios
Nimblegen EZExome v2 capture library, Illumina HiSeq 2000
OCD Sequencing Consortium
109 OCD trios
Nimblegen EZExome v2 capture library, Illumina HiSeq 2000
Simons Simplex Collection
855 control trios
Nimblegen EZExome v2 capture library, Illumina HiSeq 2000
184 OCD trios passing QC
777 control trios passing QC
101 OCD trios passing QC
777 control trios passing QC
Identify putative PZVs with MosaicHunter
Identify putative CNVs with XHMM
Filter to remove likely false positive PZV calls
Classify rare inherited and *de novo* CNVs with PLINK and PLINK/Seq
Mutation rate analysis
Mutation rate analysis

Figure 3. Variant calling pipelines for samples from the OCD Sequencing Consortium (44) and Simons Simplex Collection (57). **(A)** 184 OCD trios and 777 control trios passed quality control (QC) metrics for exome sequencing and all were included in the PZV analysis. PZVs were detected with MosaicHunter (59) and subsequently filtered to remove likely false positive variant calls. **(B)** 101 OCD trios and 777 control trios sequenced with the same capture library were used to call CNVs, which were detected

with XHMM and classified as transmitted (inherited) or *de novo* in the children using PLINK and PLINK/Seq tools (60, 61).

PZV Calling with MosaicHunter

We called putative PZVs from our aligned and indexed WES for 184 OCD trios and 777 control trios passing QC with MosaicHunter, a Bayesian-based genotyping tool (Figure 3A). MosaicHunter was developed to call single-nucleotide mosaic variants in non-cancer contexts, i.e. when a known normal control from the same individual is not available to compared to the tissue of interest (59). We used the trio mode of the tool, which incorporates WES from the parents into the calling algorithm, and the exome mode, which employs a beta-binomial model that accounts for capture bias and over-dispersion in WES to better fit the data. We applied these settings to our WES to identify low–allele frequency, potentially mosaic SNVs in probands and in control children. MosaicHunter was set to discard variants with a frequency of more than 0.05 in the Single Nucleotide Polymorphism Database (62), variants with ≥10 sequencing reads in the parents and ≥25 reads in the child, and variants falling in regions with indels or CNVs in the child. All other parameters were left as their default settings, and reference genome b37d5 was used (b37 human reference genome with decoy sequences). For each trio, MosaicHunter generated an output file containing all calls found to violate Mendelian inheritance, i.e. both *de novo* germline variants and PZVs. We discarded the output for one outlier OCD trio with an excess of variants.

In addition to the filtering steps built into MosaicHunter, we applied inclusion criteria to the output data to reduce the number of false positive PZVs in our final dataset.

These criteria include: ≥0.7 posterior probability of being mosaic in the child, ≥1 child likelihood ratio of mosaic vs. heterozygous, ≥0.5 posterior probability each parent does not carry the alternate allele (reference homozygous genotype), no more than two reads with the alternate allele in either parent, no duplicates of the variant across families, and ≤0.001 (≤0.1%) frequency in non-Finnish European populations according to the Exome Aggregation Consortium (ExAC) database (63). We removed all G>T variants with fewer than 8 T alleles, as these are highly likely to be false positive calls caused by oxidative damage to samples after collection (64).

CNV Calling with XHMM

We called putative CNVs from the same WES data, using 101 of the OCD trios that were sequenced with the same capture library (Nimblegen EZ Exome V2) as the 777 control trios (Figure 3B). Sequencing read depths were calculated using GATK's DepthOfCoverage tool. Calls were generated using eXome-Hidden Markov Model (XHMM), a statistical package designed specifically to detect CNVs from normalized read-depth data from targeted sequencing (61). Members of one OCD trio and four control trios were filtered by the XHMM default quality control methods and consequently were not included in analyses. We then used an in-house pipeline following a protocol (61) combining PLINK, Plink/Seq, and ANNOVAR software to annotate rare CNVs (frequency <1% among all individuals in the sample set) in the children as inherited or *de novo*. Plink/Seq quality thresholds for *de novo* calls were set at SQ ≥ 70 (high probability of a CNV in the child) and NQ ≥ 70 (high probability of no CNV in the parents). Following annotation, we discarded maternal and paternal CNVs not transmitted to the child. We discarded one additional outlier OCD trio with an excess of CNV calls (>20) in the child. After obtaining a set of *de novo* CNV calls, we used the AnnotSV webtool (65) to identify *de novo* CNVs that were not present in the Database of Genomic Variants (DGV; not previously detected in the human population) (66).

Burden Analysis

Mutation Rates of PZVs

Within cases and controls, we calculated the rates of single-nucleotide PZVs per base pair. To account for differences in coverage between the two cohorts, we calculated the number of callable base pairs per trio using the GATK DepthOfCoverage tool (58). Callable bases were defined as those with a sequencing depth of at least 20 reads in all three family members at that genomic position. To perform the burden analysis (comparing mutation rates in cases vs. controls), we used the rateratio.test R package to calculate mutation rate ratios with a two-sided p-value (67). We used the wANNOVAR webtool using RefSeq hg19 gene definitions (analogous to b37d5, our reference genome) to classify PZVs as LGD (adding/removing a stop codon or altering a canonical splice site), nonsynonymous (predicted to alter a gene-encoded protein sequence), synonymous (within the coding sequence but not affecting the protein product), or noncoding (68, 69). For nonsynonymous variants, we used PolyPhen-2 to computationally predict the effects of detected PZVs on protein function (70).

Rates of CNVs

We calculated CNV rates as the number of CNVs per individual and as the proportion of individuals in each cohort with at least one CNV. For both measures, we performed the burden analysis with the rateratio.test R package as described above using a two-sided p-value. Rate measurements were calculated together and separately for deletions and duplications, and by size bin (<10 kb, 10-30 kb, >30 kb). We did not perform a comparison of CNV lengths between cases and controls as the start and end points (breakpoints) of CNVs may fall outside the exomic intervals targeted by WES, rendering length measurements inaccurate.

Exploratory Risk Gene Pathway, and Expression Analyses

We used the wANNOVAR webtool to identify genes containing our putative PZVs and the AnnotSV webtool to identify genes overlapping *de novo* CNVs. Genes overlapping novel (not present in DGV) *de novo* CNVs were labeled as putative OCD risk genes and used as the input gene list for our pathway analysis. Metascape was used to perform pathway analyses using ontology terms pulled from KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways and CORUM knowledgebases (71). All known genes in the human genome were used in the enrichment background to calculate an enrichment factor (the ratio between the observed counts and the counts expected by chance) and an associated p-value. These analyses were inputted into Cytoscape to generate and visualize an interactive enrichment network of ontology terms for the gene list (72). Spatio-temporal expression analyses were conducted using the Cell-type Specific Expression Analysis (CSEA) tool (73).

RESULTS

Mutation Rates and Burden Analysis

PZV Rates

The rate of all single-nucleotide PZVs per base pair was not significantly different between OCD probands (4.90 x 10^{-9}) and controls (4.93 x 10^{-9}), rate ratio = 0.994 (95% confidence interval = 0.613-1.56), two-sided p = 1. Of the putative PZVs identified in OCD probands, none are likely gene disrupting (LGD; alteration of a splice site or stop codon) and 28% are missense mutations predicted by PolyPhen-2 to be probably damaging (Mis-D). The rate of putative damaging PZVs (LGD and Mis-D) per base pair also is not significantly different in OCD probands (1.45 x 10^{-9}) than in controls (1.09 x 10^{-9}), rate ratio = 1.33 (95% confidence interval = 0.475-3.27), two-sided p = 0.653 (Table 1). We observe no recurrence of PZVs in the same gene in unrelated probands (Table 2).

Variant class	**Variant count**		**Mutation rate (×10^{-9}) per bp**		**Estimated variants per individual**		**Rate ratio (95% CI)**	**p-value**
	OCD n=183	Control n=777	OCD n=183	Control n=777	OCD n=183	Control n=777		
All	25	96	4.90	4.93	0.166	0.167	0.994 (0. 613-1.56)	1
Coding	18	76	3.72	4.14	0.126	0.140	0.899 (0.506-1.52)	0.797
Synonymous	6	16	1.24	0.872	0.0420	0.0295	1.42 (0.456-3.83)	0.605
Nonsynonymous	12	58	2.48	3.16	0.0840	0.107	0.785 (0.384-1.48)	0.594

All Missense	12	55	2.48	3.00	0.0840	0.101	0.828 (0.404-1.56)	0.676
Mis-D	7	17	1.45	0.927	0.0490	0.0314	1.56 (0.548-3.96)	0.440
Mis-P	2	10	0.414	0.545	0.0140	0.0184	0.759 (0.0809-3.56)	1
Mis-B	3	28	0.621	1.53	0.0210	0.0516	0.407 (0.0791-1.32)	0.176
LGD	0	3	0	0.164	0	0.00553	--[a]	--
Damaging (LGD + Mis-D)	7	20	1.45	1.09	0.0490	0.0369	1.33 (0.475-3.27)	0.653
Unknown	0	2	0	0.109	0	0.00369	--[a]	--

Table 1. PZVs are not enriched in OCD probands compared to unaffected controls. Variants were annotated with Annovar, using RefSeq hg19 gene definitions. “Nonsynonymous” variants include all missense and LGD variants. Mis-D are “probably damaging” missense variants, Mis-P are “possibly damaging” missense variants, and Mis-B are “benign” missense variants based on PolyPhen-2 scoring. LGD variants are those adding/removing a stop codon or affecting a canonical splice site. “Unknown” variants are included as coding variants but are not included in the synonymous or nonsynonymous counts. Mutation rates were calculated as the number of variants divided by the number of “callable” bases (see supplementary methods). Estimates of inherited mutations per individual were calculated by multiplying the mutation rate by the size of the RefSeq hg19 coding exome (33,828,798 bp). Mutation rates were compared with a two-sided rate ratio test.

[a] Rate ratio is not calculable.

The mutant allele fractions of PZVs was not significantly different between the OCD trios (mean = 0.236, SD = 0.126) and controls (mean = 0.213, SD = 0.107), t = 0.922, df = 119, two-sided p = 0.358. Additionally, when considering only PZVs with a mutant allele fraction of at least 0.2, as in Dou et. al (49), the mutation rate for damaging PZVs in OCD (6.21 x 10^{-10}) was not significantly different from that in controls (3.82 x 10^{-10}), RR= 1.63 (95% CI = 0.271-7.12), two-sided p = 0.696. No children in the OCD or control trios harbored multiple likely damaging PZVs, so we did not perform a separate calculation comparing the proportion of children with at least one PZV in each group. Table 2 shows all putative PZVs detected in the OCD trios.

Position	**Base pair change**	**Variant type**	**PolyPhen2 prediction**	**Gene(s)**	**Family ID**
1:10166307	G>A	missense	B	*UBE4B*	8186
2:64113007	G>T	missense	D	*UGP2*	8167
2:102018919	C>T	synonymous		*RFX8*	8097
2:118578877	G>A	intronic		*DDX18*	8144
2:179476243	G>A	missense	D	*TTN*	8214
4:887739	G>C	missense	D	*GAK*	8141
4:144620034	A>G	synonymous		*FREM3*	8197
5:172123965	C>G	intergenic		*NEURL1B; LOC101928093*	8065
6:109312016	C>T	missense	D	*SESN1*	8074
7:48065402	G>A	intronic		*SUN3*	8206
7:99996911	C>T	intronic		*PILRA*	8167
8:22012961	G>T	synonymous		*LGI3*	8138
10:33140834	T>G	missense	P	*CCDC7*	8211
12:95604617	T>C	missense	B	*FGD6*	8183

16:69482126	A>G	intronic		*CYB5B*	8139
17:8381651	C>A	intronic		*MYH10*	8040
17:18167126	G>A	missense	B	*MIEF2*	8038
17:26488199	G>A	missense	D	*NLK*	8149
19:1083061	C>T	missense	P	*ARHGAP45*	8018
19:19765409	C>T	missense	D	*ATP13A1*	8094
20:33855171	C>T	synonymous		*MMP24*	8140
21:28793094	A>G	intergenic		*ADAMTS5;* *LINC00113*	8168
22:51133459	C>T	synonymous		*SHANK3*	8042
X:110970140	C>T	synonymous		*ALG13*	8002
X:115303777	T>C	missense	D	*AGTR2*	8172

Table 2. Putative PZVs detected in OCD samples. For PolyPhen2 missense variant predictions, D represents Mis-D or "probably damaging," P represents Mis-P or "possibly damaging," and B represents Mis-B or "benign." Two PZVs are found in the same child from family 8167.

CNV Rates

The rate of all rare (*de novo* and inherited) CNVs per child is 2.42 in OCD cases and 1.72 in controls, rate ratio = 1.41 (95% confidence interval = 1.23-1.62), two-sided p = 2.65x10^{-6}. The rate of rare *de novo* CNVs per child is 0.305 in OCD cases and 0.0854 in controls, rate ratio = 3.55 (95% confidence interval = 2.22-5.54), two-sided p = 2.97x10^{-7}. For novel *de novo* CNVs not present in DGV, the rate per child is 0.192 in OCD cases and 0.0492 in controls, rate ratio = 3.90 (95% confidence interval = 2.13-6.94), two-sided p = 1.91x10^{-5}. These findings are shown in Table 3.

Variant class	**Size (kb)**	**CNV type**	**CNV count (CNVs per person)**		**Rate ratio (95% CI)**	**p-value**
			OCD 99 trios	**Control 773 trios**		
All rare	**<10**	**DEL**	26 (0.263)	222 (0.287)	0.914 (0.584-1.38)	0.758
		DUP	59 (0.596)	328 (0.424)	1.40 (1.05-1.86)	0.0242
		DEL+DUP	85 (0.859)	550 (0.712)	1.21 (0.949-1.52)	0.126
	10-30	**DEL**	22 (0.901)	115 (2.37)	1.49 (0.222-0.149)	0.120
		DUP	39 (0.394)	179 (0.232)	1.70 (1.17-2.42)	0.00565
		DEL+DUP	61 (0.616)	294 (0.380)	1.62 (1.21-2.14)	0.00140
	>30	**DEL**	35 (0.354)	156 (0.202)	1.75 (1.18-2.54)	0.00600
		DUP	59 (0.596)	327 (0.423)	1.41 (1.05-1.86)	0.0230
		DEL+DUP	94 (0.949)	483 (0.625)	1.52 (1.21-1.90)	0.000473
	Total DEL		83 (0.838)	493 (0.638)	1.31 (1.03-1.66)	0.0288
	Total DUP		157 (1.59)	834 (1.08)	1.47 (1.23-1.75)	2.65x10^{-5}
	Total DEL+DUP		240 (2.42)	1327 (1.72)	1.41 (1.23-1.62)	2.65x10^{-6}
Rare *de novo*	**<10**	**DEL**	3 (0.0303)	8 (0.0103)	2.93 (0.500-12.2)	0.241
		DUP	12 (0.121)	27 (0.0349)	3.47 (1.60-7.08)	0.00187
		DEL+DUP	15 (0.152)	35 (0.0453)	3.35 (1.70-6.29)	0.000611
	10-30	**DEL**	3	9	2.60	0.295

			(0.0303)	(0.0116)	(0.453-10.4)	
		DUP	2 (0.0202)	8 (0.0103)	1.95 (0.202-9.78)	0.633
		DEL+DUP	5 (0.0505)	17 (0.0220)	2.30 (0.662-6.48)	0.192
	>30	**DEL**	6 (0.0606)	6 (0.00776)	7.81 (2.09-29.2)	0.00215
		DUP	4 (0.0404)	8 (0.0103)	3.90 (0.860-14.6)	0.0777
		DEL+DUP	10 (0.101)	14 (0.0181)	5.58 (2.22-13.5)	0.000307
	Total DEL		12 (0.121)	23 (0.0298)	4.07 (1.85-8.53)	0.000614
	Total DUP		18 (0.182)	43 (0.0556)	3.27 (1.77-5.79)	0.000202
	Total DEL+DUP		30 (0.305)	66 (0.0854)	3.55 (2.22-5.54)	2.97×10^{-7}
Novel *de novo*	**<10**	**DEL**	2 (0.0202)	5 (0.00647)	3.12 (0.297-19.1)	0.368
		DUP	8 (0.0808)	19 (0.0246)	3.287 (1.25-7.86)	0.0167
		DEL+DUP	10 (0.101)	24 (0.0310)	3.25 (1.39-7.06)	0.00707
	10-30	**DEL**	2 (0.0202)	5 (0.00647)	3.12 (0.297-19.1)	0.368
		DUP	2 (0.0202)	5 (0.00647)	3.12 (0.297-19.1)	0.368
		DEL+DUP	4 (0.0404)	10 (0.0129)	3.12 (0.715-10.8)	0.131
	>30	**DEL**	4 (0.0404)	0 (0.000)	Inf (5.15-Inf)	0.000332
		DUP	1 (0.0101)	4 (0.00517)	1.95 (0.0396-19.7)	0.905
		DEL+DUP	5 (0.0505)	4 (0.00517)	9.76 (2.10-49.2)	0.00320
	Total DEL		8	10	6.25	0.000840

	(0.0808)	(0.0129)	(2.14-17.6)	
Total DUP	11 (0.111)	28 (0.0362)	3.07 (1.38-6.36)	0.00650
Total DEL+DUP	19 (0.192)	38 (0.0492)	3.90 (2.13-6.94)	1.91x10^{-5}

Table 3. Rates of rare CNVs are higher in OCD probands compared to unaffected controls. DEL = deletion; DUP = duplication. Significance threshold with Bonferroni correction for multiple testing is 0.00139.

Considering deletions and duplications together, the proportion of children with at least one CNV is not statistically significantly different between OCD cases and controls for all rare and for novel *de novo* CNVs following multiple testing correction; however, significantly more OCD cases have at least one rare *de novo* CNV compared to controls (see Table 4). Looking only at deletions, the proportion of children with at least one rare *de novo* or inherited deletion is higher in OCD trios (0.606) than in controls (0.448), chi-square = 8.86, p=0.00292. The proportion of children with at least one rare *de novo* deletion is higher in OCD trios (0.111) than in controls (0.0298), chi-square=15.5, p=0.000082. The proportion of children with at least one novel *de novo* deletion is higher in OCD trios (0.0707) than in controls (0.0130), chi-square=15.3, p=0.000091. In contrast, the proportion of children with at least one CNV is not significantly different between cases and controls for all rare *de novo* or inherited duplications (OCD=0.747, control=0.658, chi-square=3.14, p=0.0765), for rare *de novo* duplications (OCD=0.0909, control=0.0543, chi-square=2.13, p=0.144), and for novel *de novo* duplications

(OCD=0.0404, control=0.0349, chi-square=0.0767, p=0.782). These data are shown in Figure 4 and Table 4.

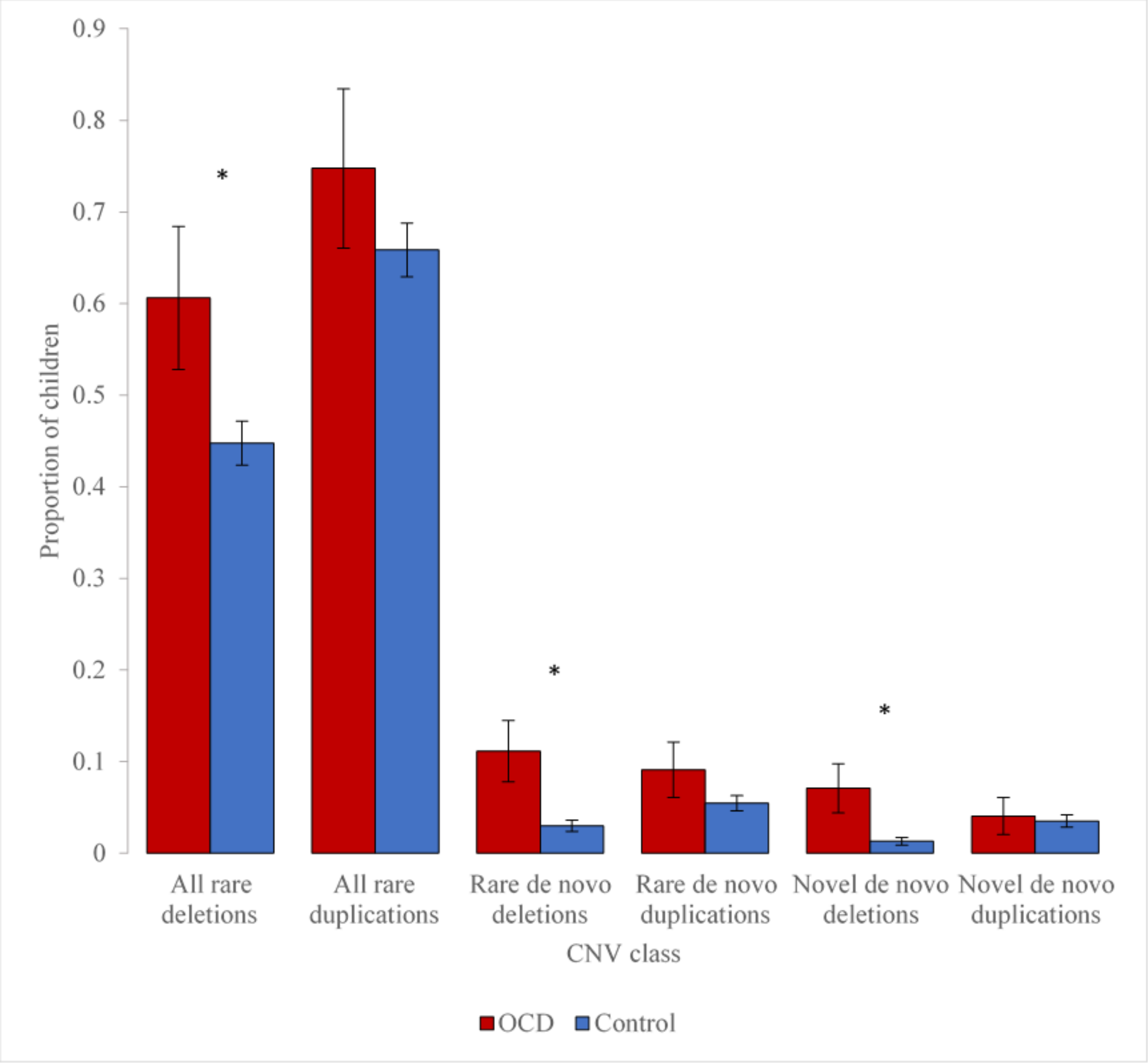

Figure 4. Rare, rare *de novo*, and novel *de novo* deletion rates are increased in OCD cases compared to controls. Within each CNV category, rates are calculated as proportion of children with at least one CNV.

Variant class	CNV type	Proportion of children with at least one CNV		Chi-square statistic	p-value
		OCD (99 trios)	Control (773 trios)		
All rare	**DEL**	0.606	0.448	8.86	0.00292
	DUP	0.747	0.658	3.14	0.0765
	DEL+DUP	0.869	0.796	2.97	0.0846
Rare *de novo*	**DEL**	0.111	0.0298	15.5	0.000082
	DUP	0.0909	0.0543	2.13	0.144
	DEL+DUP	0.182	0.084	10.1	0.00148
Novel *de novo*	**DEL**	0.0707	0.0130	15.3	0.000091
	DUP	0.0404	0.0349	0.0767	0.782
	DEL+DUP	0.0909	0.0479	3.25	0.0713

Table 4. Proportion of children with at least one CNV. Deletions are enriched in OCD cases vs. controls across all variant classes. Significance threshold with Bonferroni correction for multiple testing is 0.00556.

The 19 putative novel *de novo* CNVs detected are present in nine OCD cases, with one case containing eight of these CNVs and another containing four. In total, these 19 CNVs overlap 31 genes (Table 5).

Chromosome:start-end	CNV type	CNV size	Gene(s)
2:26587170-26611971	DUP	24801	*SELENOI*
2:100217897-100218083	DUP	186	*AFF3*
2:112536252-112545897	DEL	9645	*ANAPC1*
2:167055182-167085482	DEL	30300	*SCN1A-AS1; SCN9A*

2:170657471-170681107	DUP	23636	*METTL5; SNORD3K; SSB*
2:230456296-230456604	DUP	308	*DNER*
3:129546646-129547221	DUP	575	*TMCC1*
3:130305350-130318654	DEL	13304	*COL6A6*
5:80911292-80946158	DEL	34866	*SSBP2*
6:42897309-42897459	DUP	150	*CNPY3; CNPY3-GNMT*
7:26245988-26251794	DEL	5806	*CBX3*
8:101718922-101730116	DEL	11194	*PABPC1*
11:129780371-129780551	DUP	180	*PRDM10*
12:69124890-69279669	DEL	154779	*CPM; LOC100130075; MDM2; NUP107; SLC35E3*
15:29367124-30092905	DUP	725781	*APBA2; FAM189A1; LOC100130111; NSMCE3; TJP1*
15:45777361-45783079	DUP	5718	*SLC30A4*
17:71205668-71205907	DUP	239	*FAM104A*
18:45368198-45423127	DEL	54929	*SMAD2*
19:49926469-49926596	DUP	127	*PTH2*

Table 5. Putative novel *de novo* CNVs detected in OCD samples.

Pathway Analysis

For the list of 31 genes overlapping novel *de novo* OCD CNVs (the subclass of CNVs with the highest rate ratio in the burden analysis), Metascape found eight ontology terms that had a p-value < 0.01, had an enrichment factor > 1.5, and were associated with at least three genes from the input list. Pathway analysis showed the highest enrichment for the ontology terms cell cycle (p=0.00035), associated with SMAD2, MDM2, and ANAPC1 genes; protein export from the nucleus (p=0.0010), associated with MDM2, SSB, and NUP107; and SUMO E3 ligases (p=0.0011), associated with MDM2, NSMCE3, and NUP107. Figure 5 shows all eight terms mapped in a network based on relatedness of terms, with terms with a similarity > 0.3 connected by edges (lines). Table 6 shows the genes contributing to each term and the corresponding p-value.

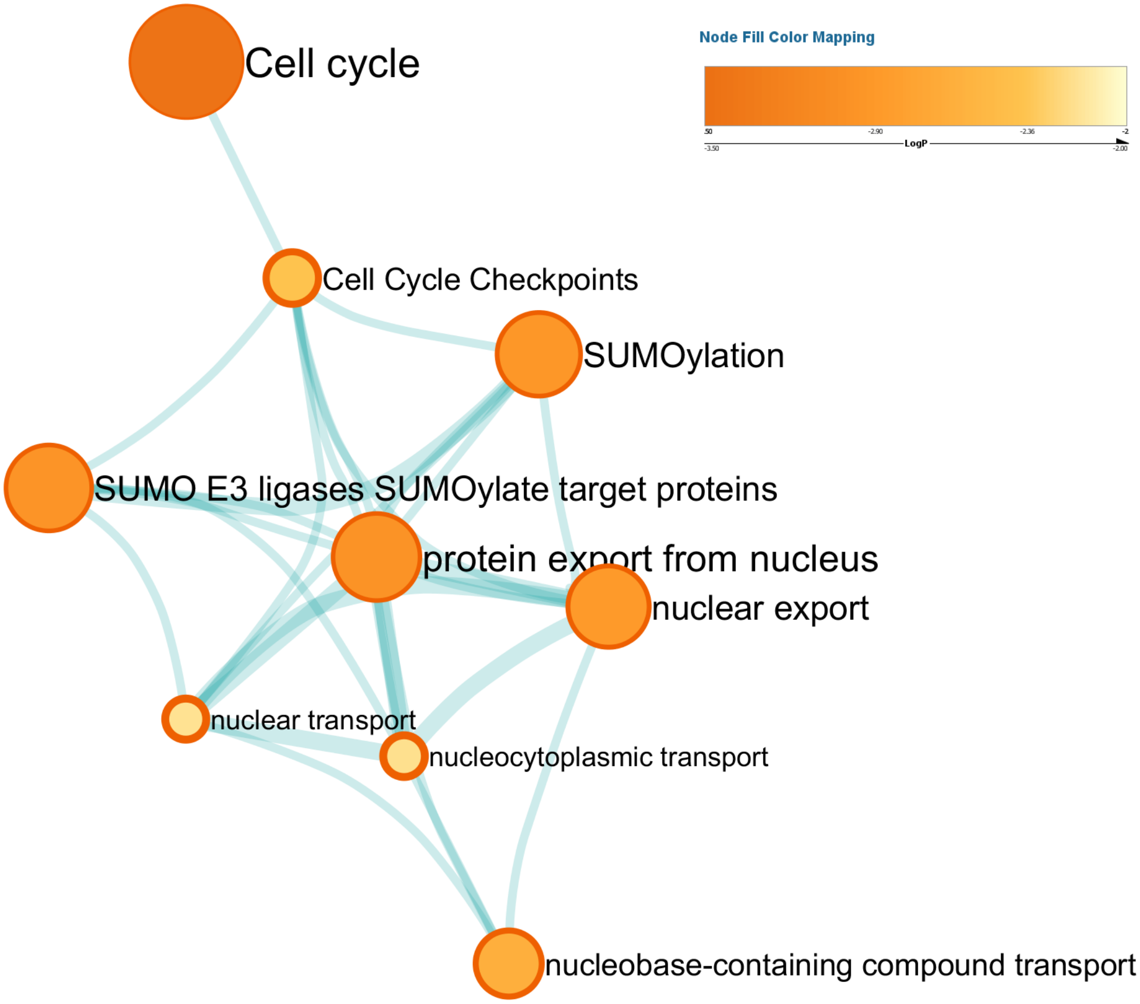
Cell cycle
Node Fill Color Mapping
.50
-2.90
-2.36
-2
LogP
-3.50
-2.00
Cell Cycle Checkpoints
SUMOylation
SUMO E3 ligases SUMOylate target proteins
protein export from nucleus
nuclear export
nuclear transport
nucleocytoplasmic transport
nucleobase-containing compound transport

Figure 5. Pathway analysis of 31 genes overlapping novel *de novo* OCD CNVs. Nodes are labeled by their corresponding gene ontology description, colored according to their p-value (see legend), and connected by edges to show relatedness.

Ontology term	**Gene hits**	**p-Value**
Cell cycle	*SMAD2; MDM2; ANAPC1*	0.00035
Protein export from nucleus	*MDM2; SSB; NUP107*	0.0010
SUMO E3 ligases SUMOylate target proteins	*MDM2; NSMCE3; NUP107*	0.0011
SUMOylation	*MDM2; NSMCE3; NUP107*	0.0012
Nuclear export	*MDM2; SSB; NUP107*	0.0013
Nucleobase-containing compound transport	*SSB; SLC35E3; NUP107*	0.0024
Cell cycle checkpoints	*MDM2; NUP107; ANAPC1*	0.0042
Nucleocytoplasmic transport	*MDM2; SSB; NUP107*	0.0064

Table 6. Significant gene ontology terms from pathway analysis the set of 31 genes overlapping novel *de novo* OCD CNVs. Genes ("hits") contributing to each term and the p-value for enrichment of the term are shown.

Clinical Features of Notable Cases

Five CNVs, each from a separate OCD patient, contribute to the eight significant ontology terms from the pathway analysis (Table 7). All five patients have other psychiatric comorbidities, notably Tourette Syndrome (TS) in all cases. The child from family 8134 previously was found to have a *de novo*, germline, predicted-damaging missense mutation in the *CHD8* gene (44), though there is no history of ASD or intellectual disability in the patient or in any close family members.

Family ID	**CNV type**	**Gene(s)**	**Sex**	**Age of onset[a]**	**Other clinical features**
8134	DUP	*METTL5*; *SNORD3K*; *SSB*	Male	--[b]	TS, ADHD, SAD, irritability
8168	DEL	*ANAPC1*	Male	11	TS
8171	DEL	*CPM*; *LOC100130075*; *MDM2*; *NUP107*; *SLC35E3*	Male	5	TS, ADHD
8205	DEL	*SMAD2*	Male	6	TS, current ADHD symptoms, ASD/PDD, congenital anomalies
8221	DUP	APBA2; FAM189A1; LOC100130111; NSMCE3; TJP1	Male	8	TS, ADHD, ASD/PDD

Table 7. Clinical features of OCD cases with putative novel *de novo* CNVs contributing to significant ontology terms. TS = Tourette Syndrome, ADHD = attention deficit hyperactivity disorder, SAD = separation anxiety disorder, ASD/PDD = flagged for autism spectrum disorders/pervasive developmental disorders. Congenital anomalies are flagged in the phenotypic data but not further specified.

[a] Age of OCD onset in years.

[b] Data not available.

Expression Analysis

We performed an expression analysis of 31 genes overlapping novel *de novo* OCD CNVs across brain regions and developmental time periods. No expression networks within the CSEA brain region and development expression dataset were enriched for these putative OCD risk genes. Of note, the Cerebellum Early Fetal network contained 4 OCD risk genes (*DNER, PRDM10, SCN9A*, *TMCC1*), though it did not reach statistical significance after multiple testing correction (uncorrected p=0.021, corrected p=0.936).

DISCUSSION

We successfully applied variant calling approaches to detect PZVs and CNVs in WES. These variants still require experimental validation before we can use them in more rigorous risk gene and other downstream analyses. However, we can begin to draw inferences from our burden analyses about the potential role of these types of variation in OCD pathogenesis and from our pathway analysis about potential biological mechanisms significant in OCD.

In this study, we first aimed to characterize the contribution of PZVs to OCD. Our burden analysis showed that, counter to our expectations, likely damaging PZVs were not enriched in OCD cases compared to controls. This finding does not support our hypothesis that damaging PZVs contribute to OCD pathogenesis. However, we may have had difficulty detecting PZVs due to tissue- or cell-specificity. Somatic variants may be present only in a specific tissue, and even within this tissue may affect only a subset of cell types at certain developmental stages. A challenge to our approach is that by studying whole blood, we may not detect tissue-specific mosaicism (for example, mosaic variants present only in the brain). Furthermore, studying bulk tissue (blood) instead of single cells may decrease our ability to detect mosaicism due to the presence of normal cells in the tissue, which may overwhelm the signal from a minority of cells harboring damaging mutations.

Additionally, our burden analysis may be hampered by insufficient power to detect differences between the cohorts. Our power to detect differences in the rate of all PZVs for cases and controls is estimated to be 0.828 based on rate ratios previously reported in the literature (see Power Calculations in Supplementary Methods). This power should be sufficient (above 0.8) to detect differences, suggesting our failure to find a difference in the rate of all PZVs reflects a true lack of statistical significance. This mirrors our previous finding of no statistical difference in the rate of all germline *de novo* variants in the same samples (44). However, we had hypothesized that we would see a significantly higher rate of likely damaging PZVs in cases vs. controls. For this subset of PZVs, our power to detect differences in mutation rate between the two groups is 0.423, which is significantly below what we would like our power to be when using a significance cutoff of 0.05. As we continue collecting WES for more OCD trios and our sample size increases, we may have more power to detect significant differences in damaging PZV burden.

The proportion of children with putative rare deletions of all classes was greater for OCD patients than controls, while there was no difference in the proportion of OCD and control children with rare duplications. This result is consistent with previous microarray studies of CNVs in OCD (11, 51, 52). We might expect that deletions are more likely to have a deleterious effect compared to duplications by inducing a haploinsufficiency-like effect. However, duplications also could have a highly damaging effect if their endpoint falls within a gene and disrupts the protein-coding region. Like the PZVs, the detected putative CNVs should be validated experimentally to remove false positives. If the enrichment of OCD patients with deletions holds after validation, this finding would provide further evidence that rare deletions play a role in OCD pathogenesis. Additionally, as we continue to sequence more OCD trios, we may detect additional risk genes and new biological pathways or expression networks enriched for these genes.

Genes overlapping novel *de novo* CNVs in OCD patients are associated with ontology terms related to cell cycle and nuclear transport. The associate with cell cycle terms is consistent with findings that many genes related to neurodevelopmental disorders play a role in neural stem cell proliferation and differentiation and that particular genes are associated both with neurodevelopmental disorders and with cancers (74). Similarly, many genes related to nuclear transport or nuclear localization, namely those encoding transcription factors and chromatin modifiers, have been associated with neurodevelopmental disorders (75-79). These findings are consistent with our previous study of germline SNVs and indels in OCD WES. In the previous study we identified *CHD8*, which encodes a DNA helicase that regulates gene expression through chromatin remodeling, as a high-confidence OCD risk gene.

Like many patients with OCD, most of the OCD cases with CNVs contributing to the significant ontology terms had multiple comorbid psychiatric disorders. All cases had TS and four of the five had ADHD diagnoses or symptoms present at the time of evaluation. Notably, two cases were flagged for a diagnosis of autism, one of which also was flagged for congenital anomalies (unspecified in our available clinical data). These cases highlight the challenges in teasing apart the contribution of genetic variants to OCD and to comorbid features. Given recent evidence that OCD and TS have overlapping genetic etiologies, future risk gene analyses in OCD should examine the overlap with genes implicated in TS (44, 80). Future efforts to collect patients with OCD and without comorbid disorders may help isolate potential OCD-specific genetic etiologies. Further, deep phenotyping of enrolled patients would allow us to interrogate genetic factors that could contribute to the clinical heterogeneity of OCD. By attempting to sort patients based on clinical phenotype, we could parse out any different genetic features between OCD subtypes.

Future Directions

We intend to validate our detected likely damaging PZVs and *de novo* CNVs in OCD cases with digital droplet PCR (ddPCR), a technique capable of validating very low-frequency mosaic variants (81). In ddPCR, the DNA sample is diluted into droplets, each containing one molecule of the target allele. A PCR reaction with fluorescent tags marking the target region is run in each droplet, and quantification of the tag signal allows calculation of allele frequency. Based on these validation results, we will optimize the pipeline parameters to obtain high-confidence sets of variants. Following validations, we will compare our set of detected CNVs in the WES data generated from our control samples to microarray data previously generated in the same samples (57). Additionally, we will compare the rates of CNVs detected in our WES for OCD patients to rates of CNVs found in previous OCD studies that used microarray data. These comparisons will test our hypothesis that detecting CNVs from WES data with our pipeline allows us to detect more CNVs, particularly smaller ones, than can be detected using microarray data.

If our finding of CNV enrichment in OCD cases holds after validation, we will use information about these variants to assess the level of significance of putative OCD risk genes using the Transmission And *De novo* Association (TADA) statistical method. TADA uses information about inherited and *de novo* variants to predict a gene's likelihood of association with a disease and strongly implicates a gene in disease if damaging *de novo* mutations are seen in the same gene in more than one unrelated proband (82). This statistical method has been used to identify 99 high-confidence risk genes in autism (83, 84), and we have used it to identify risk genes based on germline *de novo* variants in OCD (44) and Tourette syndrome (56). By incorporating CNVs into this model, we are likely to identify more OCD risk genes that will rise to the level of statistical significance.

Long term, this research will help lay the groundwork for further research into the molecular basis of OCD. Specific genes and biologic pathways implicated by our analyses will provide jumping-off points to guide later studies examining molecular mechanisms (e.g. animal and cell culture models). Ultimately, these mechanistic studies will point to potential drug targets and will allow for development and testing of crucial new therapeutic options for patients with OCD.

SUPPLEMENTARY METHODS

Sequence Alignment

Sequence reads obtained from WES were aligned to the b37d5 human reference genome using the Burrows-Wheeler Aligner tool, PCR duplicates were marked using Picard's MarkDuplicates tool, and tab-delimited text file (BAM file) containing the aligned exome data was generated (85). The BAM file for each individual's exome was used as the input for MosaicHunter and for XHMM.

Power Calculations

To estimate our power to detect differences in mutation frequency between our OCD and control samples (86), we defined the following variables:

Group	Control children	Children with OCD
Mean callable base pairs	t_1	t_2
Sample size	N_1	N_2
Number of PZVs	X_1	X_2
Rate of PZVs per individual	λ_1	λ_2

The rate ratio of PZVs is $RR = \frac{\lambda_2}{\lambda_1}$. $RR_0 = 1$, representing the null hypothesis that the mutation rates of the two groups are not statistically different. $RR_a > 1$, representing the alternative hypothesis that the mutation rate in children with OCD is significantly greater than that in control children.

Assuming nonequal mutation rates for group 1 (control children) and group 2 (children with OCD) with unconstrained maximum likelihood estimates, we can calculate the test statistic for testing the ratio of two Poisson rates as

$$W_1 = \frac{X_2 - X_1\left(\frac{\sqrt{RR_0}}{d}\right)}{\sqrt{X_2 + X_1\left(\frac{RR_0}{d}\right)^2}}$$

where

$$d = \frac{t_1 N_1}{t_2 N_2}$$

Based on our samples, we can calculate

$$d = \frac{(2.506 * 10^7)777}{(2.787 * 10^7)183} = \ 3.818$$

To calculate power, we use

$$Power(W_1) = 1 - \ \phi\left(\frac{z_{1-\alpha}\sigma_1 - \mu_1}{\sigma_1}\right)$$

where

$$\mu_1 = \left(\frac{RR_a}{d} - \frac{RR_0}{d}\right) t_1 N_1 \lambda_1$$

$$\sigma_1 = \sqrt{\left(\frac{dRR_a + {RR_o}^2}{d^2}\right) t_1 N_1 \lambda_1}$$

$$\phi(z) = \int_{-\infty}^{z} Normal(0{,}1)$$

Our significance threshold is α = 0.05, so our critical value $z_{1-\alpha} = 1.645$ using the standard normal distribution and assuming infinite degrees of freedom.

We estimated RR_a using the rate ratio for all mosaic variants found by Freed and Pevsner in children with ASD compared to their unaffected siblings, which was 1.73 (46). This allows us to calculate

$$\mu_1 = \left(\frac{1.73}{3.818} - \frac{1}{3.818}\right)(2.506 * 10^7)(777)(4.93 * \ 10^{-9}) = 18.35$$

$$\sigma_1 = \sqrt{\left(\frac{3.818 * 1.73 + 1^2}{3.818^2}\right)(2.506 * 10^7)(777)(4.93 * \ 10^{-9})} = 7.077$$

$$\phi\left(\frac{1.645 * 7.077 - 18.35}{7.077}\right) = \phi(-0.9479)$$

We calculate $\phi(-0.9479)$ by integrating from -∞ to -0.9479 over a normal distribution with mean = 0 and standard deviation = 1, giving

$$\phi(-0.9479) = \int_{-\infty}^{-0.9479} Normal(0,1) = 0.172$$

This gives us

$$Power(W_1) = 1 - 0.172 = \ 0.828$$

for detecting differences in the rate of all PZVs. Repeating these calculations for our ability to detect differences in rates of likely damaging PZVs using the rate ratio from Freed and Pevsner for Mis-D and LGD mosaic mutations, which is 1.58, gives us a power of 0.423.

Callable Bases

The number of “callable” bases within each trio was calculated as previously described (44) and used to calculate PZV rates to minimize bias in calling variants between case and control cohorts. Using the GATK DepthOfCoverage tool, we calculated the number of bases covered at ≥ 20x in all family members, with base quality ≥20, and map quality ≥ 30 (the same thresholds required for GATK and *de novo* variant calling). For each cohort, we summed the callable base pairs in every family. The sum of coding and noncoding callable bases was used as the denominator for calculating rates of all PZVs (5100562503 bases for 183 OCD trios and 19474297328 bases for 777 control trios). The sum of only coding callable bases was used as the denominator for all other rate calculations (4833549696 bases for 183 OCD trios and 18342070930 bases for 777 control trios).

REFERENCES

1. Pauls DL. The genetics of obsessive-compulsive disorder: a review. Dialogues Clin Neurosci. 2010;12(2):149-63.
2. Meier SM, Mattheisen M, Mors O, Schendel DE, Mortensen PB, Plessen KJ. Mortality Among Persons With Obsessive-Compulsive Disorder in Denmark. JAMA Psychiatry. 2016;73(3):268-74.
3. O'Connor K, Todorov C, Robillard S, Borgeat F, Brault M. Cognitive-behaviour therapy and medication in the treatment of obsessive-compulsive disorder: a controlled study. Can J Psychiatry. 1999;44(1):64-71.
4. Bloch MH, Landeros-Weisenberger A, Rosario MC, Pittenger C, Leckman JF. Meta-analysis of the symptom structure of obsessive-compulsive disorder. Am J Psychiatry. 2008;165(12):1532-42.
5. Mataix-Cols D, Rosario-Campos MC, Leckman JF. A multidimensional model of obsessive-compulsive disorder. Am J Psychiatry. 2005;162(2):228-38.
6. McKay D, Abramowitz JS, Calamari JE, Kyrios M, Radomsky A, Sookman D, et al. A critical evaluation of obsessive-compulsive disorder subtypes: symptoms versus mechanisms. Clin Psychol Rev. 2004;24(3):283-313.
7. Purty A, Nestadt G, Samuels JF, Viswanath B. Genetics of obsessive-compulsive disorder. Indian J Psychiatry. 2019;61(Suppl 1):S37-S42.
8. Pauls DL, Abramovitch A, Rauch SL, Geller DA. Obsessive–compulsive disorder: an integrative genetic and neurobiological perspective. Nature Reviews Neuroscience. 2014;15(6):410-24.
9. van Grootheest DS, Cath DC, Beekman AT, Boomsma DI. Twin studies on obsessive-compulsive disorder: a review. Twin Res Hum Genet. 2005;8(5):450-8.
10. Mataix-Cols D, Boman M, Monzani B, Rück C, Serlachius E, Långström N, et al. Population-based, multigenerational family clustering study of obsessive-compulsive disorder. JAMA psychiatry. 2013;70(7):709-17.
11. Grunblatt E, Oneda B, Ekici AB, Ball J, Geissler J, Uebe S, et al. High resolution chromosomal microarray analysis in paediatric obsessive-compulsive disorder. BMC Med Genomics. 2017;10(1):68.
12. Hudziak JJ, Van Beijsterveldt C, Althoff RR, Stanger C, Rettew DC, Nelson EC, et al. Genetic and Environmental Contributions to the Child Behavior ChecklistObsessive-Compulsive Scale: A Cross-cultural Twin Study. Archives of General Psychiatry. 2004;61(6):608-16.
13. Monzani B, Rijsdijk F, Harris J, Mataix-Cols D. The structure of genetic and environmental risk factors for dimensional representations of DSM-5 obsessive-compulsive spectrum disorders. JAMA psychiatry. 2014;71(2):182-9.

14. Eley TC, Bolton D, O'connor TG, Perrin S, Smith P, Plomin R. A twin study of anxiety-related behaviours in pre-school children. Journal of Child Psychology and Psychiatry. 2003;44(7):945-60.

15. Fernandez T, Leckman J, Pittenger C. Neurogenetics: Handbook of Clinical Neurology. 2015.

16. Geschwind DH, Flint J. Genetics and genomics of psychiatric disease. Science. 2015;349(6255):1489-94.

17. Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Magi R, Reschen ME, et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat Genet. 2015;47(12):1415-25.

18. Stewart SE, Yu D, Scharf JM, Neale BM, Fagerness JA, Mathews CA, et al. Genome-wide association study of obsessive-compulsive disorder. Mol Psychiatry. 2013;18(7):788-98.

19. Mattheisen M, Samuels JF, Wang Y, Greenberg BD, Fyer AJ, McCracken JT, et al. Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. Mol Psychiatry. 2015;20(3):337-44.

20. International Obsessive Compulsive Disorder Foundation Genetics C, Studies OCDCGA. Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. Mol Psychiatry. 2018;23(5):1181-8.

21. Costas J, Carrera N, Alonso P, Gurriaran X, Segalas C, Real E, et al. Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. Transl Psychiatry. 2016;6:e768.

22. Bosker FJ, Hartman CA, Nolte IM, Prins BP, Terpstra P, Posthuma D, et al. Poor replication of candidate genes for major depressive disorder using genome-wide association data. Mol Psychiatry. 2011;16(5):516-32.

23. Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet. 2009;5(2):e1000373.

24. Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry. 2008;13(6):570-84.

25. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry. 2011;168(10):1041-9.

26. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet. 2003;361(9360):865-72.

27. Koenen KC, Duncan LE, Liberzon I, Ressler KJ. From candidate genes to genome-wide association: the challenges and promise of posttraumatic stress disorder genetic studies. Biol Psychiatry. 2013;74(9):634-6.

28. Taylor S. Molecular genetics of obsessive-compulsive disorder: a comprehensive meta-analysis of genetic association studies. Mol Psychiatry. 2013;18(7):799-805.

29. Gomes CKF, Vieira-Fonseca T, Melo-Felippe FB, de Salles Andrade JB, Fontenelle LF, Kohlrausch FB. Association analysis of SLC6A4 and HTR2A genes with obsessive-compulsive disorder: Influence of the STin2 polymorphism. Compr Psychiatry. 2018;82:1-6.

30. Sampaio AS, Hounie AG, Petribu K, Cappi C, Morais I, Vallada H, et al. COMT and MAO-A polymorphisms and obsessive-compulsive disorder: a family-based association study. PLoS One. 2015;10(3):e0119592.

31. Walitza S, Marinova Z, Grunblatt E, Lazic SE, Remschmidt H, Vloet TD, et al. Trio study and meta-analysis support the association of genetic variation at the serotonin transporter with early-onset obsessive-compulsive disorder. Neurosci Lett. 2014;580:100-3.

32. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013;155(5):997-1007.

33. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. Genome Biol. 2016;17(1):241.

34. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415-25.

35. Hanna GL, Veenstra-VanderWeele J, Cox NJ, Boehnke M, Himle JA, Curtis GC, et al. Genome-wide linkage analysis of families with obsessive-compulsive disorder ascertained through pediatric probands. Am J Med Genet. 2002;114(5):541-52.

36. Hanna GL, Veenstra-Vanderweele J, Cox NJ, Van Etten M, Fischer DJ, Himle JA, et al. Evidence for a susceptibility locus on chromosome 10p15 in early-onset obsessive-compulsive disorder. Biol Psychiatry. 2007;62(8):856-62.

37. Mathews CA, Badner JA, Andresen JM, Sheppard B, Himle JA, Grant JE, et al. Genome-wide linkage analysis of obsessive-compulsive disorder implicates chromosome 1p36. Biol Psychiatry. 2012;72(8):629-36.

38. Ross J, Badner J, Garrido H, Sheppard B, Chavira DA, Grados M, et al. Genomewide linkage analysis in Costa Rican families implicates chromosome 15q14 as a candidate region for OCD. Hum Genet. 2011;130(6):795-805.

39. Shugart YY, Samuels J, Willour VL, Grados MA, Greenberg BD, Knowles JA, et al. Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q. Mol Psychiatry. 2006;11(8):763-70.

40. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, et al. Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet. 2011;43(9):864-8.

41. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat Genet. 2011;43(9):860-3.

42. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41.

43. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585-9.

44. Cappi C, Oliphant ME, Péter Z, Zai G, do Rosário MC, Sullivan CA, et al. De Novo Damaging DNA Coding Mutations Are Associated With Obsessive-Compulsive Disorder and Overlap With Tourette’s Disorder and Autism. Biological psychiatry. 2019.

45. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843-51.

46. Freed D, Pevsner J. The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLoS Genet. 2016;12(9):e1006245.

47. Krupp DR, Barnard RA, Duffourd Y, Evans SA, Mulqueen RM, Bernier R, et al. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. Am J Hum Genet. 2017;101(3):369-90.

48. Lim ET, Uddin M, De Rubeis S, Chan Y, Kamumbu AS, Zhang X, et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. Nat Neurosci. 2017;20(9):1217-24.

49. Dou Y, Yang X, Li Z, Wang S, Zhang Z, Ye AY, et al. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. Hum Mutat. 2017;38(8):1002-13.

50. Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, Veltman JA, et al. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. Am J Hum Genet. 2015;97(1):67-74.

51. McGrath LM, Yu D, Marshall C, Davis LK, Thiruvahindrapuram B, Li B, et al. Copy number variation in obsessive-compulsive disorder and tourette syndrome: a cross-disorder study. J Am Acad Child Adolesc Psychiatry. 2014;53(8):910-9.

52. Gazzellone MJ, Zarrei M, Burton CL, Walker S, Uddin M, Shaheen SM, et al. Uncovering obsessive-compulsive disorder risk genes in a pediatric cohort by high-resolution analysis of copy number variation. J Neurodev Disord. 2016;8:36.

53. Poultney CS, Goldberg AP, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwara Y, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am J Hum Genet. 2013;93(4):607-19.

54. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.

55. Dietrich A, Fernandez TV, King RA, State MW, Tischfield JA, Hoekstra PJ, et al. The Tourette International Collaborative Genetics (TIC Genetics) study, finding the genes causing Tourette syndrome: objectives and methods. Eur Child Adolesc Psychiatry. 2015;24(2):141-51.

56. Willsey AJ, Fernandez TV, Yu D, King RA, Dietrich A, Xing J, et al. De Novo Coding Variants Are Strongly Associated with Tourette Disorder. Neuron. 2017;94(3):486-99 e9.

57. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68(2):192-5.

58. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

59. Huang AY, Zhang Z, Ye AY, Dou Y, Yan L, Yang X, et al. MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. Nucleic Acids Res. 2017;45(10):e76.

60. 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;91(4):597-607.

61. Fromer M, Purcell SM. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr Protoc Hum Genet. 2014;81:7 23 1-1.

62. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000;28(1):352-5.

63. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840-D5.

64. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41(6):e67.

65. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics. 2018;34(20):3572-4.

66. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986-92.

67. Fay M, Fay MM. Package ‘rateratio. test’. 2014.

68. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet. 2012;49(7):433-6.

69. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10(10):1556-66.

70. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.

71. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.

72. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2010;27(3):431-2.

73. Xu X, Wells AB, O'Brien DR, Nehorai A, Dougherty JD. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. J Neurosci. 2014;34(4):1420-31.

74. Ernst C. Proliferation and Differentiation Deficits are a Major Convergence Point for Neurodevelopmental Disorders. Trends Neurosci. 2016;39(5):290-9.

75. Bain JM, Cho MT, Telegrafi A, Wilson A, Brooks S, Botti C, et al. Variants in HNRNPH2 on the X Chromosome Are Associated with a Neurodevelopmental Disorder in Females. Am J Hum Genet. 2016;99(3):728-34.

76. Estruch SB, Graham SA, Quevedo M, Vino A, Dekkers DHW, Deriziotis P, et al. Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Hum Mol Genet. 2018.

77. den Hoed J, Sollis E, Venselaar H, Estruch SB, Deriziotis P, Fisher SE. Functional characterization of TBR1 variants in neurodevelopmental disorder. Sci Rep. 2018;8(1):14279.

78. Jones KA, Luo Y, Dukes-Rimsky L, Srivastava DP, Koul-Tewari R, Russell TA, et al. Neurodevelopmental disorder-associated ZBTB20 gene variants affect dendritic and synaptic structure. PLoS One. 2018;13(10):e0203760.

79. Cappuyns E, Huyghebaert J, Vandeweyer G, Kooy RF. Mutations in ADNP affect expression and subcellular localization of the protein. Cell Cycle. 2018;17(9):1068-75.

80. Wang S, Mandell JD, Kumar Y, Sun N, Morris MT, Arbelaez J, et al. De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. Cell Rep. 2018;24(13):3441-54 e12.

81. Zhou B, Haney MS, Zhu X, Pattni R, Abyzov A, Urban AE. Detection and Quantification of Mosaic Genomic DNA Variation in Primary Somatic Tissues Using ddPCR: Analysis of Mosaic Transposable-Element Insertions, Copy-Number Variants, and Single-Nucleotide Variants. Methods Mol Biol. 2018;1768:173-90.

82. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013;9(8):e1003671.

83. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-33.

84. Feliciano P, Zhou X, Astrovskaya I, Turner TN, Wang T, Brueggeman L, et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. NPJ Genom Med. 2019;4:19.

85. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.

86. Gu K, Ng HK, Tang ML, Schucany WR. Testing the ratio of two poisson rates. Biom J. 2008;50(2):283-98.