De novo damaging coding mutations are strongly associated with obsessive-compulsive disorder and overlap with autism

Carolina Cappi[a,1], Melody E. Oliphant[b,1], Zsanett Péter[b], Gwyneth Zai[c,d], Catherine A. W. Sullivan[e], Abha R. Gupta[b,e], Ellen J. Hoffman[b], Manmeet Virdee[b], A. Jeremy Willsey[f], Roseli G. Shavitt[a], Euripedes C. Miguel[a], James L. Kennedy[c,d], Margaret A. Richter[d,g], and Thomas V. Fernandez[b,h,2]

[a]Department of Psychiatry, School of Medicine, University of São Paulo, R. Dr. Ovídio Pires de Campos, 785, 3º andar, sala 9, São Paulo, SP, 05403-010, Brazil.

[b]Yale Child Study Center, Yale University School of Medicine, New Haven, CT, 06519, USA.

[c]Neurogenetics Section, Molecular Brain Science Department, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON M5T 1R8, Canada.

[d]Department of Psychiatry, University of Toronto, Toronto, ON M5T 1R8, Canada.

[e]Department of Pediatrics, Yale University School of Medicine, New Haven, CT, 06519, USA.

[f]Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA, 94143, USA.

[g]Frederick W. Thompson Anxiety Disorders Centre, Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, Toronto, ON M4N 3M5, Canada.

[h]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, 06519, USA.

[1]These authors contributed equally to this work.

[2]Lead contact. Please address correspondence to: thomas.fernandez@yale.edu (T. V. F.)

ABSTRACT

Obsessive-compulsive disorder (OCD) is a debilitating developmental neuropsychiatric disorder with a genetic risk component, yet identification of high-confidence risk genes has been challenging. We performed whole-exome sequencing in 222 OCD parent-child trios (184 trios after quality control), finding strong evidence that de novo likely gene disrupting and predicted damaging missense variants contribute to OCD risk. Together, these de novo damaging variants are enriched in OCD probands (RR 1.52, p=0.0005). We identified two high-confidence risk genes, each containing two de novo damaging variants in unrelated probands: *CHD8 (Chromodomain Helicase DNA Binding Protein 8)* and *SCUBE1 (Signal Peptide, CUB Domain And EGF Like Domain Containing 1)*. Based on our data, we estimate that 34% of de novo damaging variants seen in OCD contribute to risk, and that de novo damaging variants in approximately 335 genes contribute to risk in 22% of OCD cases. Furthermore, genes harboring de novo damaging variants in OCD are enriched for those reported in neurodevelopmental disorders, particularly autism spectrum disorders. An exploratory network analysis reveals significant functional connectivity and enrichment in canonical pathways related to immune response.

Key words: obsessive-compulsive disorder, de novo variants, whole exome sequencing, gene discovery, *CHD8*, *SCUBE1,* autism, developmental disorders

SIGNIFICANCE STATEMENT

Decades of genetic studies in obsessive-compulsive disorder (OCD) have yet to provide reproducible, statistically significant findings. Following an approach that has led to tremendous success in gene discovery for several neuropsychiatric disorders, here we report findings from DNA whole-exome sequencing of patients with OCD and their parents. We find strong evidence for the contribution of spontaneous, or de novo, predicted-damaging genetic variants to OCD risk, identify two high-confidence risk genes, and detect significant overlap with genes previously identified in autism. These results change the status quo of OCD genetics by identifying novel OCD risk genes, clarifying the genetic landscape of OCD with respect to de novo variation, and suggesting underlying biological pathways that will improve our understanding of OCD biology.

\body

INTRODUCTION

Obsessive-compulsive disorder (OCD) is an often-disabling developmental neuropsychiatric disorder with onset typically during adolescence or young adulthood and a lifetime prevalence of 1.5-2.5% (1-5). Obsessions are intrusive thoughts, images, or urges experienced as irrational, excessive, and accompanied by anxiety or discomfort. Compulsions are behaviors undertaken to mitigate obsessions or subjective feelings (i.e., the need to relieve a tactile sensation or achieve a “just right” feeling); they are usually repetitive, stereotyped, and excessive (6, 7). The anxiety or distress associated with obsessions and compulsions and the time spent on them are sources of lifelong morbidity in OCD, having profound negative effects on both patients’ and families’ quality of life. Symptoms can be so disabling that the World Health Organization has ranked OCD among the 10 most debilitating disorders of any kind, in terms of lost earnings and diminished quality of life (8, 9). Furthermore, OCD has been linked to significantly increased mortality, even after controlling for comorbid psychiatric conditions, which can occur in up to 75% of cases (10, 11). Treatment-refractory disease is common, with about 40% of patients resistant to current pharmacological and psychotherapeutic treatments, and untreated OCD generally persists and becomes chronic (12, 13). The causes and underlying pathophysiology of OCD are not well understood, which has limited the development of new treatments and interventions. For these reasons, there is an urgent need for more research to elucidate OCD risk factors and disease mechanisms.

Twin and family studies provide strong evidence for a substantial genetic contribution to OCD risk, with modern estimates of heritability around 40-50% (14-17), yet progress in identifying risk genes has been slow. Decades of linkage, common-variant candidate gene association studies, and more recent genome-wide association studies in OCD (18-20) have yielded few reproducible associations and therefore have provided limited insights into disease

biology. Further efforts are clearly needed to identify specific OCD risk variants and to confirm vulnerability pathways by modern genome-wide and comprehensive variant discovery approaches.

In contrast, genetic research into several other neuropsychiatric disorders has seen great progress in recent years, partly attributable to an increasing effort to evaluate the contribution of rare genetic sequence variation, especially de novo variants (arising spontaneously in the parental germ cells or in a zygote shortly after conception). This approach has shown great success for systematic risk gene discovery in other genetically complex neuropsychiatric disorders (21-23), particularly autism spectrum disorders (24-27). While an individual rare variant is unlikely to explain a sizeable fraction of disease risk in the context of a heterogeneous genetic architecture, concurrent investigations of multiple genes implicated by rare sequence and structural variation highlight convergence toward a limited number of important underlying biological mechanisms (28). Therefore, there is a proven avenue for risk gene discovery in complex neuropsychiatric disorders that has yet to be fully leveraged in OCD.

Following these previous studies in other disorders and our pilot study suggesting a role for de novo single nucleotide variants (SNVs) in OCD risk (29), we performed whole-exome sequencing (WES) in 222 OCD parent-child trios to identify de novo SNVs and insertion-deletion variants (indels). In 184 OCD trios passing quality control, we find strong evidence for the contribution of de novo likely gene disrupting (LGD; disruption of a stop codon, canonical splice site, or a frameshift indel) as well as predicted damaging missense (Mis-D) variants to OCD. Furthermore, we identified two high-confidence candidate risk genes based on observing gene-level recurrence of de novo damaging (LGD + Mis-D) variants in unrelated probands: *CHD8 (Chromodomain Helicase DNA Binding Protein 8)* and *SCUBE1 (Signal Peptide, CUB Domain And EGF Like Domain Containing 1)*. We estimated that 22% of OCD cases will harbor a de novo damaging SNV or indel mediating OCD risk, and that there are approximately 335

genes affected by such variants contributing to the risk. Finally, we detected significant overlap between genes with damaging de novo variants in OCD and autism.

RESULTS

Damaging de novo SNVs and indels are associated with OCD risk

Exome sequencing was performed on 222 OCD parent-child trios. WES data from 855 unaffected trios, already sequenced from the Simons Simplex Collection, were pooled with our OCD trios for joint variant calling. After quality control methods, our sample size for a burden analysis was 184 OCD and 777 unaffected trios (Figure 1, Table 1, Table S1). To compare the de novo mutation rates between cases and controls, we limited our analysis to loci with at least 20x coverage in all members of a trio, as this was our pre-defined threshold for calling a de novo variant (see Methods). Based on our OCD pilot study (29) and work in other neurodevelopmental disorders (22, 24, 26, 30), we expected to find an enrichment of de novo LGD variants (stop codon, frameshift, or canonical splice-site variants) in OCD probands versus controls. We found a statistically significant increased rate of de novo LGD variants in OCD cases, confirming our hypothesis (rate ratio [RR] 1.93, 95% Confidence Interval [CI] 1.19-3.09, p=0.01). Furthermore, de novo missense variants predicted to be damaging by PolyPhen2 (Mis-D; Polyphen2 HDIV score ≥0.957) were also over-represented in OCD probands (RR 1.43, CI 1.13-1.80, p=0.006). Taken together, damaging de novo coding variants (LGD and Mis-D) occur more often in OCD probands versus controls (RR 1.52, CI 1.23-1.86, p=0.0005). We did not detect a difference in mutation rates for de novo synonymous variants (RR 0.99, CI 0.75-1.31, p=0.5). (See Table 1, Figure 2A)

Damaging de novo SNVs and indels contribute to OCD risk in 22% of cases

Next, we estimated the fraction of observed de novo mutations that contribute to OCD risk, based on our dataset. By dividing the de novo mutation rate difference between cases and controls by the rate in cases, we estimate that 49.2% (CI 3.4-95.0%) of de novo LGD and 29.5% (CI 6.0-53.0%) of de novo Mis-D mutations contribute to OCD risk. As a group, we estimate that 33.9% (CI 13.3-54.6%) of damaging (LGD + Mis-D) de novo mutations contribute to OCD risk (Figure 2B).

We also used our data to estimate the proportion of cases harboring a de novo mutation contributing to OCD risk. By subtracting the percentage of controls from the percentage of OCD probands with at least one de novo mutation, we estimate that 15.0% (CI 3.1-26.9%) have a de novo Mis-D mutation and 7.3% (CI 0.50-14.0%) have a de novo LGD mutation mediating OCD risk. As a group, we estimate that 22.2% (CI 8.7-35.8%) of cases have a damaging de novo mutation contributing to OCD risk (Figure 2B).

Recurrent damaging de novo variants identify two candidate risk genes

Having established that de novo damaging variants occur more frequently in OCD probands, we next asked whether these variants cluster within specific genes. We identified three genes with multiple (≥2) de novo LGD or Mis-D variants in unrelated probands. Using TADA-Denovo (31) and previously established false discovery rate (FDR) thresholds, two of these genes met criteria for high-confidence risk genes (q<0.1): *SCUBE1* (*Signal Peptide, CUB Domain And EGF Like Domain Containing 1*; q=0.040) and *CHD8* (*Chromodomain Helicase DNA Binding Protein 8*; q=0.043) (Table 2). A third gene, *TTN* (*Titin*), did not meet this threshold (q=0.47) (Table 2).

Approximately 335 genes contribute to OCD risk

Based on OCD proband vulnerability to de novo damaging variants in our dataset, we used two methods to estimate the number of genes contributing to OCD risk. Using a maximum

likelihood estimation (MLE) method (32), we determined the most likely number of genes to be 335 (Figure S1). This agrees with an alternate method based on the "unseen species problem" (33); the estimated number of OCD risk genes using this alternate method is 317 (95% CI 190-454).

Next, we used the estimated number of OCD risk genes (n=335) to predict the likely future gene discovery yield as additional OCD trios are investigated by WES. Based upon 10,000 simulations at each cohort size, we predict discovery of the following numbers of risk genes as we sequence more OCD parent-child trios: 24 probable risk genes, including 11 high-confidence risk genes (24 / 11 genes) at 500 trios; 77 / 40 genes at 1,000 trios; 202 / 113 genes at 2,000 trios; 323 / 189 genes at 3,000 trios (Figure S2).

Overlap with ASD and CHD8 target genes

Using DNENRICH (34), we found significant overlap between genes harboring de novo damaging variants in OCD (n=89, excludes occurrences in controls) and several gene sets from the literature (Table 3, Table S4). Our OCD genes were significantly enriched for genes harboring de novo nonsynonymous (LGD, missense) variants in autism (ASD), genes achieving TADA q<0.1 in ASD, genes with genome-wide significant statistical evidence for association with developmental disorders, and genes that are targets of CHD8 in the developing human brain. There was no significant enrichment for genes harboring de novo variants in intellectual disability (ID) or schizophrenia (SCZ), no enrichment for de novo synonymous variants in any disorder, and no enrichment for any class of de novo variation in unaffected siblings in the SSC (Table 3, Table S4).

Exploratory pathway and network analyses

Using our list of genes harboring de novo damaging variants in OCD (n=89), we performed exploratory analyses to determine shared underlying canonical pathways and

functional connectivity. Using the GeNets algorithm, OCD genes mapping onto a meta-network displayed significantly more connectivity than expected by chance (p=0.026), and 68 additional "tier 1" candidate genes were predicted, based on their high connectivity to our input genes (Figure S6, Table S5). Our input gene list is significantly enriched for canonical pathways related to immune response, particularly the complement system, which was a top result from two pathway analysis tools, MetaCore and IPA. Other significant canonical pathways include granulocyte-macrophage colony stimulating factor (GM-CSF) signaling, neurotrophin/tyrosine kinase signaling, B cell receptor signaling, and focal adhesion kinase signaling (Table S6).

DISCUSSION

By whole-exome sequencing of OCD parent-child trios, we have demonstrated a strong association between de novo damaging (LGD and Mis-D) coding variants and OCD cases (Table 1, Figure 2). As seen in studies of other neurodevelopmental disorders, these results can be leveraged to systematically identify OCD risk genes. In the current study, two genes, *CHD8* and *SCUBE1*, have a FDR q<0.1, meeting criteria for high-confidence association with OCD (Table 2).

SCUBE1 has not been extensively studied. While it is expressed in the developing brain and nervous system (35, 36), functional studies to date have focused mostly on its potential role in platelet activation and adhesion (37, 38). A study in mice has shown downregulated *SCUBE1* expression in response to inflammatory stimuli (36), but this gene has not yet been implicated in disorders of the brain or nervous system. Interestingly, increased levels of pro-inflammatory markers have been reported in several studies of children and adolescents with OCD (39-42).

On the other hand, there are several recent and ongoing studies of *CHD8*, a gene that has emerged as having the strongest association with autism spectrum disorder via the

identification of multiple de novo LGD variants in unrelated parent-child trios (Figure S5) (33, 43-45). *CHD8* is highly expressed in the developing brain (46). It encodes an ATP-dependent chromatin remodeler that binds to tri-methylated histone H3 lysine 4, a post-translational histone modification present at active promoters (47-49). Loss of *CHD8* function appears to contribute to autism pathology by disrupting the expression of its target genes, which are themselves enriched for high confidence autism risk genes (46). While OCD subjects with de novo damaging *CHD8* variants in our study do not meet any diagnostic criteria for autism, this finding suggests there may be overlapping biological mechanisms between the two disorders, and leads us to hypothesize that genes regulated by *CHD8* may similarly be enriched for OCD risk genes. Indeed, we see significant overlap between our OCD genes and ASD genes, as well as *CHD8* gene targets mapped in the developing human brain (46) (Table 3).

Based on data from this study, we estimate that 34% of de novo damaging mutations seen in OCD carry risk (Figure 2B) and that 335 genes confer risk in 22% of patients (Figure 2B). Given our OCD sample size, the 95% confidence intervals around these contribution estimates are wide, and need refinement by continued sequencing of OCD trios.

Mindful of the fact that more than half of genes harboring de novo damaging variants in our study may not be true risk genes, we consider our pathway and network analyses as exploratory at this stage. Nevertheless, we see preliminary evidence that genes identified by de novo damaging variants in OCD are functionally connected to a greater degree than expected by chance (Figure S6, Table S5). Furthermore, these genes may be enriched in immunological and cell signaling canonical pathways (Table S6), consistent with our pilot study of exome sequencing in 20 OCD trios (29). These analyses should be repeated as more high-confidence OCD risk genes are identified.

While not rising to the level of a high-confidence risk gene in this study, it is notable that we identified an OCD de novo damaging (Mis-D) variant in *DLGAP1* (*discs, large homolog-*

associated protein 1). In a genome-wide association study by the International OCD Foundation Genetic Collaborative (IOCDF-GC), the lowest p-values for their case-control analysis were found for two SNPs located within *DLGAP1* ($2.49x10^{-6}$, $3.44x10^{-6}$) (20). A subsequent GWAS by the OCD Collaborative Genetics Study (OCGAS) identified a SNP nearby this gene with a prominent signal ($p=2.67x10^{-4}$) (18). Furthermore, a rare paternally-inherited duplication in *DLGAP1* was recently reported in a child with OCD, Tourette syndrome, and anxiety (50). *DLGAP1* is a member of the neuronal postsynaptic density complex and is in the same family as *DLGAP3* (*SAPAP3*), a gene associated with OCD-like behaviors in a knockout mouse model (51). Therefore, evidence is beginning to converge on this gene as one of great interest in OCD genetics.

Successful gene discovery by leveraging gene-level recurrent de novo variation in autism, where over 65 genes have now been identified (24, 26, 27), and the results presented here for OCD, strongly reinforce the value of continuing WES in larger cohorts of OCD parent-child trios. Our models predict that by increasing the sample size of this study to 500 trios, we will gain 9 additional high-confidence risk genes and 22 probable risk genes. Further increasing to 1,000 trios will yield a total of 40 high-confidence and 77 probable risk genes. Discovering risk genes will change the status quo in OCD genetics by allowing new studies in model systems (e.g. animal models, induced pluripotent stem cells) and network analyses. Such studies will provide insights into OCD pathophysiology that are critical prerequisites for the discovery of novel therapeutic targets to alleviate the suffering of those with OCD.

MATERIALS AND METHODS

Subjects

This study was approved by the local institutional review boards of all participating institutions, and appropriate informed consent was obtained from participating subjects. 222 parent-child trios (139 male, 83 female), consisting of offspring meeting criteria for the diagnosis of obsessive-compulsive disorder, as defined by the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV-TR or DSM-5) (52, 53), and their unaffected parents, were recruited for DNA sequencing. Trios were recruited at three sites: the University of São Paulo School of Medicine Obsessive-Compulsive Spectrum Disorders Program (42 trios), Centre for Addiction and Mental Health and the Frederick W. Thompson Anxiety Disorders Centre at the Sunnybrook Health Sciences Centre in Toronto (77 trios), and Yale University School of Medicine (61 trios). Additionally, we included 42 trios with OCD and chronic tics that were recruited for a separate study by TIC Genetics (54, 55). Other diagnostic criteria included: onset of symptoms prior to age 18 years; no previously diagnosed neurological disorder, intellectual disability, psychotic disorder, or OCD occurring exclusively in the context of depression; no known history of OCD in first degree relatives. Final diagnostic status was assigned based on the consensus of an experienced interviewer and a psychiatrist or psychologist after independent review and administration of a structured diagnostic interview. We prioritized the study of simplex OCD trios to increase the likelihood of detecting de novo sequence and structural variants. Available phenotype information, including gender and parental age, is included in Table S1.

Whole-exome sequencing (WES)

Exome capture and sequencing of blood-derived DNA from 222 affected children and their parents (666 samples total) were performed at the Yale Center for Genomic Analysis (YCGA), using the NimbleGen SeqCap EZExomeV2 (109 trios) or MedExome (113 trios)

capture libraries (Roche NimbleGen, Madison, WI, USA) and the Illumina HiSeq 2000 platform (74 bp paired-end reads; Illumina, San Diego, CA). We multiplexed six samples during each capture reaction and sequencing lane, pooling parents and probands when possible. WES data from 855 unaffected parent-child trios (2565 samples total) were obtained from the Simons Simplex Collection via the NIH Data Archive (https://ndar.nih.gov/edit_collection.html?id=2042). These children and their parents have no evidence of autism spectrum or other neurodevelopmental disorders (56). Like our OCD samples, control WES was from blood-derived DNA and sequenced on the Illumina HiSeq 2000 sequencing platform after capture with the NimbleGen SeqCap EZExomeV2 library.

Sequence alignment and variant calling

Alignment and variant calling of the sequencing reads followed the latest Genome Analysis Toolkit (GATK) (57) Best Practices guidelines. Reads were aligned using BWA-mem (58) to the b37 human reference sequence with decoy sequences. Picard's MarkDuplicates tool was used to mark PCR duplicates (https://broadinstitute.github.io/picard/). A target bed file was created by taking the intersection of the EZExomeV2 and MedExome target regions. GATK was used to realign indels, recalibrate quality scores, and generate GVCF files for each sample using the HaplotypeCaller tool. All samples were called jointly using GATK's GenotypeGVCFs tool, variant score recalibration was applied to the called variants, and all variant call data was written to a VCF file. This pipeline uses GATK's Best Practices parameters and the default parameters for BWA and Picard. Only passing variants were used in downstream analyses. Variants were annotated against the RefSeq hg19 gene definitions and multiple external databases of variant population frequency, conservation scores, variation intolerance, mutation severity, and predicted functional effects using ANNOVAR (59).

Quality control and de novo variant calling

Relatedness statistics were calculated based on the method of Manichaikul et al. (60), implemented in VCFtools v0.1.14.10 (61). Trios were omitted if expected family relationships were not confirmed or if there were unexpected relationships within or between families. Trios were omitted if > 5 de novo variants were observed. PLINK/SEQ (34) (i-stats; https://psychgen.u.hpc.mssm.edu/plinkseq/stats.shtml), PicardTools, and GATK DepthOfCoverage tools were used to generate quality metrics (Table S1). To identify outliers that might confound our case-control analysis, we performed principal components analysis (PCA) using this data. A scree plot determined the number of principal components accounting for the greatest proportion of variance, and we removed trios with family members falling more than three standard deviations from the mean in any of these principal components. See Figure S3, Table S1, and Supplementary Methods.

We used stringent thresholds for identifying de novo mutations because DNA from control subjects (recruited by the Simons Foundation Autism Research Initiative) was not available for confirmation by Sanger sequencing. De novo variants were called using an in-house script that required the following: child is heterozygous for a variant with alternate allele frequency between 0.3 and 0.7 in the child and < 0.05 in the parents, sequencing depth (DP) ≥ 20 in all family members at the variant position, alternate allele depth (AD) ≥ 5, observed allele frequency (AC) < 0.01 (1%) among all cases and controls, mapping quality (MQ) ≥ 30. False positive calls were removed by in silico visualization. We performed Sanger sequencing on the probands and parents for 383 putative de novo variants from this and another exome sequencing project using the same calling thresholds; 370 were confirmed, resulting in 96.6% specificity.

Mutation rate analysis

Within each cohort, we calculated the rate of de novo mutations per base pair. For accurate rate calculation, we first determined the number of “callable” base pairs per family

using the GATK DepthOfCoverage tool. We considered only bases covered at ≥ 20x in all family members, with base quality ≥ 20, and map quality ≥ 30; these thresholds match those required for GATK and de novo variant calling. For each cohort, we summed the "callable" base pairs in every family and used this number as the denominator for de novo rate calculations. The resulting rate was divided by two to give haploid rates. Confidence intervals were calculated using the *pois.conf.int (pois.exact)* function from the epitools v0.5-9 package in R. We compared de novo mutation rates in cases versus controls (burden analysis) using a one-tailed rate ratio test in R (https://cran.r-project.org/package=rateratio.test), considering only those variants present with a frequency of <0.01 in the ExAC v0.3.1 database (62). See Table 1, Table S1, and Supplementary Methods.

TADA analysis

Prior exome analyses demonstrated that the observation of even a small number of rare de novo mutations in the same gene among unrelated individuals can provide considerable statistical power to establish association (33). We used the Transmitted And De novo Association (TADA-Denovo) test as a statistical method for risk gene discovery based on gene-level recurrence of de novo mutations within the classes of variants that we found enriched in OCD (27, 31). Parameter calculations and a detailed description of the method are given in Supplementary Methods. In summary, TADA-Denovo generates random mutational data based on each gene's specified mutation rate to determine null distributions, then calculates a p-value and a false discovery rate (FDR) q-value for each gene using a Bayesian "direct posterior approach." A low q-value represents strong evidence for OCD association. See Table 2, Table S3, and Supplementary Methods.

Estimation of number of risk genes

We first used a maximum likelihood estimation (MLE) method to estimate the number of genes contributing risk to OCD, based on vulnerability to de novo damaging variants (32). For every number of risk genes from 1 to 2,500, we simulated 95 variants (the number of damaging de novo variants observed in probands in our case-control burden analysis). Variant simulations were performed 100,000 times at each number of risk genes. Following each simulation, a percentage of variants was randomly assigned to the risk genes. The percentage of variants assigned to risk genes was determined by the fraction of de novo damaging variants estimated to carry OCD risk, and variant simulations were weighted by gene size and GC content (31). We then counted the number of risk and non-risk genes containing two variants and the number containing three or more variants. The frequency of concordance between our simulated and observed data was calculated. A curve was plotted to show the concordance frequency (y-axis) at each assumed number of risk genes (x-axis), and the peak was taken as the estimate of the most likely number of risk genes (32). See Figure S1 and Supplementary Methods.

Next, we used an alternate method for estimating the number of risk genes, using a statistical method based on the "unseen species" problem (33). This method uses the frequency and number of observed variant types (or species) to infer how many species are present in the population. See Supplementary Methods for details of these calculations.

Estimation of future risk gene discovery

We used the estimated number of OCD risk genes to predict the likely future gene discovery yield as additional OCD trios are investigated by WES. Fixing the gene number at 335 (from MLE estimate above), we varied the cohort size (from 25 to 3000, in increments of 25). At each cohort size, we simulated a number of variants matching the observed mutation rate in OCD probands. Simulated variants were randomly assigned to the risk genes. The percentage of variants assigned to risk genes was determined by the fraction of de novo damaging variants estimated to carry OCD risk, and variant simulations were weighted by gene size and GC

content (31). At each cohort size, 10,000 simulations were performed. LGD and Mis-D variants were generated separately. Simulated variants were then combined and given as input to the TADA-Denovo algorithm, using the same parameters described above for the observed data. The number of high confidence ($q<0.1$) and probable ($q<0.3$) risk genes were recorded and plotted using polynomial regression fitting; this regression model allows prediction of the number of genes identified at a specified cohort size. See Figure S2 and Supplementary Methods.

Gene set overlap

We used DNENRICH (34) (https://psychgen.u.hpc.mssm.edu/dnenrich/) to test whether OCD genes harboring de novo damaging mutations (89 genes; excluding two genes, *TTN* and *CACNA1E*, found to harbor de novo damaging variants in control subjects) were significantly enriched among genes identified in autism (ASD), schizophrenia (SCZ), developmental disorders (DD), Tourette's disorder (TD), and intellectual disability (ID). Gene lists were obtained from a recent cross-disorder study (63) that included de novo single nucleotide and indel variants from multiple exome sequencing studies in ASD (24, 26, 43, 64, 65), SCZ (34, 66-69), and ID (23, 70-72). Three of these studies also included de novo variants present in unaffected siblings (26, 64, 67). DD (21) and TD (54) genes were obtained from recently published WES studies. DNENRICH simulates random mutations while accounting for gene size, trinucleotide context, and mutational effect. We performed 100,000 permutations, comparing the observed and expected overlap with each gene set. Empirical p-values were generated, based on a one-sided enrichment analysis under a binomial model of greater than expected hits per gene set. We tested for overlap between our OCD genes and those in ASD, SCZ, DD, TD, ID, and unaffected siblings that harbored de novo (LGD, nonsynonymous, synonymous) mutations. We also tested for overlap with ASD genes achieving $q<0.01$ in a recent meta-analysis (27) and DD genes achieving genome-wide statistical significance (21). Finally, given that we identified

CHD8 as an OCD risk gene in our study, we tested for overlap with lists of genes that are targets of CHD8 in human brain (46), See Table 3, Table S4, and Supplementary Methods.

Exploratory pathway and network analyses

To determine whether all genes harboring de novo damaging variants in OCD are enriched for specific biological pathways, we used the same gene list from our gene set overlap analysis (n=89) to identify the most significant canonical pathways suggested by MetaCore (Thomson Reuters, version 6.30, build 68780, https://portal.genego.com/) and Ingenuity Pathway Analysis (IPA, build version 430520M, content version 31813282, release date 2016-12-05; Ingenuity Systems, http://www.ingenuity.com/). The following default settings were used for MetaCore: Analyze Single Experiment Tool, species Homo sapiens, threshold 0, p-value 1, signals both. The following default settings were used for IPA: Reference set: Ingenuity Knowledge Base (Genes Only); direct and indirect relationships; does not include endogenous chemicals; consider only relationships where species = human and confidence = experimentally observed. See Table S6.

Using the GeNets algorithm (https://apps.broadinstitute.org/genets), we mapped all 89 genes harboring de novo damaging mutations in OCD onto the GeNets Metanetwork v1.0 to determine whether they are functionally connected. The GeNets Metanetwork contains integrated protein-protein interactions from InWeb3 (73), ConcensusPathDB (http://consensuspathdb.org) (74), and 5,057 drug-target interactions (75); the total network size is 530,532 interactions. The GeNets algorithm determines the density of the mapped network (density = number of edges / number of possible edges) and compares this to computed densities for randomly sampled gene sets. An empirically determined p-value is generated. The network is determined to be significantly more connected than random if the density is greater than 95% of the randomly sampled gene sets. Additionally, in the process of mapping our genes onto the Metanetwork, additional candidate genes are predicted, based on their connectivity to

our input genes. An overall network connectivity p-value is generated, both with and without these additional predicted candidates. Also, as part of the GeNets analysis, gene "communities" were determined, defined as genes that are more connected to one another than they are to other groups of genes. See Figure S6 and Table S5. All GeNets results for this analysis are also available in interactive form here:

https://www.broadinstitute.org/genets#/visualize/58d9425ea4e00291af652379

ACKNOWLEDGEMENTS

We wish to thank the families who have participated in and contributed to this study. Control subject data were obtained from the NIH-supported National Database for Autism Research (NDAR). NDAR is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in autism. Dataset identifier: 2042. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDAR. Additionally, we wish to thank the Tourette International Collaborative Genetics Study for contributing published genetic data. This study was supported by grants from the Allison Family Foundation [TVF]; FAPESP (process number: 2014/01585-5) and CNPq (process number: 460928/2014-7) [CC]; the Ontario Mental Health Foundation (OMHF) and private donations from the Frederick W. Thompson family [MAR, JLK]. JLK is a Scientific Advisory Board member of AssureRx. JLK has received speaker honoraria and expenses from Eli Lilly and Novartis, consultant honoraria and expenses from Roche, and expenses from AssureRx. MAR has received research support through grants from Roche and speaker honoraria from Lundbeck. TVF has received research support from Shire, the Simons Foundation, and the National Institute of Mental Health; neither funded this project. The authors declare no potential conflict of interest.

REFERENCES

1. Fontenelle LF, Mendlowicz MV, & Versiani M (2006) The descriptive epidemiology of obsessive-compulsive disorder. *Progress in neuro-psychopharmacology & biological psychiatry* 30(3):327-337.

2. Karno M, Golding JM, Sorenson SB, & Burnam MA (1988) The epidemiology of obsessive-compulsive disorder in five US communities. *Arch Gen Psychiatry* 45(12):1094-1099.

3. Ruscio AM, Stein DJ, Chiu WT, & Kessler RC (2010) The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Mol Psychiatry* 15(1):53-63.

4. Weissman MM*, et al.* (1994) The cross national epidemiology of obsessive compulsive disorder. The Cross National Collaborative Group. *J Clin Psychiatry* 55 Suppl:5-10.

5. Torres AR*, et al.* (2006) Obsessive-compulsive disorder: prevalence, comorbidity, impact, and help-seeking in the British National Psychiatric Morbidity Survey of 2000. *Am J Psychiatry* 163(11):1978-1985.

6. Miguel EC*, et al.* (2000) Sensory phenomena in obsessive-compulsive disorder and Tourette's disorder. *J Clin Psychiatry* 61(2):150-156; quiz 157.

7. Shavitt RG*, et al.* (2014) Phenomenology of OCD: lessons from a large multicenter study and implications for ICD-11. *J Psychiatr Res* 57:141-148.

8. Veale D & Roberts A (2014) Obsessive-compulsive disorder. *BMJ (Clinical research ed.)* 348:g2183.

9. Bobes J*, et al.* (2001) Quality of life and disability in patients with obsessive-compulsive disorder. *Eur Psychiatry* 16(4):239-245.

10. Meier SM*, et al.* (2016) Mortality Among Persons With Obsessive-Compulsive Disorder in Denmark. *JAMA psychiatry* 73(3):268-274.

11. Lochner C*, et al.* (2014) Comorbidity in obsessive-compulsive disorder (OCD): a report from the International College of Obsessive-Compulsive Spectrum Disorders (ICOCS). *Comprehensive psychiatry* 55(7):1513-1519.

12. Skoog G & Skoog I (1999) A 40-year follow-up of patients with obsessive-compulsive disorder [see commetns]. *Arch Gen Psychiatry* 56(2):121-127.

13. Eisen JL*, et al.* (2013) Five-year course of obsessive-compulsive disorder: predictors of remission and relapse. *J Clin Psychiatry* 74(3):233-239.

14. Monzani B, Rijsdijk F, Harris J, & Mataix-Cols D (2014) The structure of genetic and environmental risk factors for dimensional representations of DSM-5 obsessive-compulsive spectrum disorders. *JAMA psychiatry* 71(2):182-189.

15. Pauls DL, Alsobrook JP, Goodman W, Rasmussen S, & Leckman JF (1995) A family study of obsessive-compulsive disorder. *Am J Psychiatry* 152(1):76-84.

16. Taylor S (2011) Etiology of obsessions and compulsions: a meta-analysis and narrative review of twin studies. *Clinical psychology review* 31(8):1361-1372.

17. van Grootheest DS, Cath DC, Beekman AT, & Boomsma DI (2005) Twin studies on obsessive-compulsive disorder: a review. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 8(5):450-458.

18. Mattheisen M*, et al.* (2014) Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Mol Psychiatry*.

19. Costas J*, et al.* (2016) Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. *Transl Psychiatry* 6:e768.

20. Stewart SE*, et al.* (2013) Genome-wide association study of obsessive-compulsive disorder. *Mol Psychiatry* 18(7):788-798.

21. Deciphering Developmental Disorders Study (2017) Prevalence and architecture of de novo mutations in developmental disorders. *Nature*.

22. Allen AS*, et al.* (2013) De novo mutations in epileptic encephalopathies. *Nature* 501(7466):217-221.

23. Rauch A*, et al.* (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380(9854):1674-1682.

24. De Rubeis S*, et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515(7526):209-215.

25. Dong S*, et al.* (2014) De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell reports* 9(1):16-23.

26. Iossifov I*, et al.* (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515(7526):216-221.

27. Sanders SJ*, et al.* (2015) Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 87(6):1215-1233.

28. Willsey AJ*, et al.* (2013) Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155(5):997-1007.

29. Cappi C*, et al.* (2016) Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways. *Transl Psychiatry* 6:e764.

30. Deciphering Developmental Disorders Study (2015) Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519(7542):223-228.

31. He X*, et al.* (2013) Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* 9(8):e1003671.

32. Homsy J, *et al.* (2015) De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* 350(6265):1262-1266.

33. Sanders SJ, *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485(7397):237-241.

34. Fromer M, *et al.* (2014) De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506(7487):179-184.

35. Grimmond S, *et al.* (2000) Cloning, mapping, and expression analysis of a gene encoding a novel mammalian EGF-related protein (SCUBE1). *Genomics* 70(1):74-81.

36. Yang RB, *et al.* (2002) Identification of a novel family of cell-surface proteins expressed in human vascular endothelium. *J Biol Chem* 277(48):46364-46373.

37. Tu CF, *et al.* (2006) Localization and characterization of a novel secreted protein SCUBE1 in human platelets. *Cardiovascular research* 71(3):486-495.

38. Wu MY, *et al.* (2014) Inhibition of the plasma SCUBE1, a novel platelet adhesive protein, protects mice against thrombosis. *Arteriosclerosis, thrombosis, and vascular biology* 34(7):1390-1398.

39. Gabbay V, *et al.* (2009) A cytokine study in children and adolescents with Tourette's disorder. *Progress in neuro-psychopharmacology & biological psychiatry* 33(6):967-971.

40. Mittleman BB, *et al.* (1997) Cerebrospinal fluid cytokines in pediatric neuropsychiatric disease. *Journal of immunology (Baltimore, Md. : 1950)* 159(6):2994-2999.

41. Leckman JF, *et al.* (2005) Increased serum levels of interleukin-12 and tumor necrosis factor-alpha in Tourette's syndrome. *Biol Psychiatry* 57(6):667-673.

42. Mitchell RH & Goldstein BI (2014) Inflammation in children and adolescents with neuropsychiatric disorders: a systematic review. *J Am Acad Child Adolesc Psychiatry* 53(3):274-296.

43. Neale BM, *et al.* (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485(7397):242-245.

44. Bernier R, *et al.* (2014) Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* 158(2):263-276.

45. O'Roak BJ, *et al.* (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* 338(6114):1619-1622.

46. Cotney J, *et al.* (2015) The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nature communications* 6:6404.

47. Thompson BA, Tremblay V, Lin G, & Bochar DA (2008) CHD8 is an ATP-dependent chromatin remodeling factor that regulates beta-catenin target genes. *Mol Cell Biol* 28(12):3894-3904.

48. Yuan CC, *et al.* (2007) CHD8 associates with human Staf and contributes to efficient U6 RNA polymerase III transcription. *Mol Cell Biol* 27(24):8729-8738.

49. Barski A, *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823-837.

50. Gazzellone MJ, *et al.* (2016) Uncovering obsessive-compulsive disorder risk genes in a pediatric cohort by high-resolution analysis of copy number variation. *Journal of neurodevelopmental disorders* 8:36.

51. Welch JM, *et al.* (2007) Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice. *Nature* 448(7156):894-900.

52. American Psychiatric Association (2000) *Diagnostic and statistical manual of mental disorders : DSM-IV-TR* (American Psychiatric Association, Washington, DC).

53. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders : DSM-5.

54. Willsey AJ, *et al.* (2017) De Novo Coding Variants Are Strongly Associated with Tourette Disorder. *Neuron* 94(3):486-499 e489.

55. Dietrich A, *et al.* (2015) The Tourette International Collaborative Genetics (TIC Genetics) study, finding the genes causing Tourette syndrome: objectives and methods. *European Child & Adolescent Psychiatry* 24(2):141-151.

56. Fischbach GD & Lord C (2010) The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* 68(2):192-195.

57. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.

58. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.

59. Wang K, Li M, & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.

60. Manichaikul A, *et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* 26(22):2867-2873.

61. Danecek P, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics (Oxford, England)* 27(15):2156-2158.

62. Lek M, *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.

63. Shohat S, Ben-David E, & Shifman S (2017) Varying Intolerance of Gene Pathways to Mutational Classes Explain Genetic Convergence across Neuropsychiatric Disorders. *Cell reports* 18(9):2217-2227.

64. Iossifov I, *et al.* (2012) De novo gene disruptions in children on the autistic spectrum. *Neuron* 74(2):285-299.

65. O'Roak BJ, *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485(7397):246-250.

66. Girard SL, *et al.* (2011) Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 43(9):860-863.

67. Gulsuner S, *et al.* (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 154(3):518-529.

68. McCarthy SE, *et al.* (2014) De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol Psychiatry* 19(6):652-658.

69. Xu B, *et al.* (2011) Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 43(9):864-868.

70. de Ligt J, *et al.* (2012) Diagnostic exome sequencing in persons with severe intellectual disability. *The New England journal of medicine* 367(20):1921-1929.

71. Gilissen C, *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511(7509):344-347.

72. Hamdan FF, *et al.* (2014) De novo mutations in moderate or severe intellectual disability. *PLoS Genet* 10(10):e1004772.

73. Lage K, *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 25(3):309-316.

74. Kamburov A, Stelzl U, Lehrach H, & Herwig R (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 41(Database issue):D793-800.

75. Rask-Andersen M, Masuram S, & Schioth HB (2014) The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual review of pharmacology and toxicology* 54:9-26.

FIGURE LEGENDS

Figure 1 – Study summary. We performed whole exome sequencing on 222 OCD and 855 control parent-child trios. After quality control, 184 OCD and 777 control trios remained for subsequent analyses. A burden analysis compared the rates of de novo single nucleotide (SNVs) and insertion-deletion (indel) variants between cases and controls. Next, we used the TADA-Denovo algorithm to assess the significance of gene-level recurrence of de novo damaging variants in our OCD group, identifying two high-confidence risk genes. Exploratory network, pathway, and cross-disorder analyses were then performed using genes harboring de novo damaging variants in our OCD subjects. Finally, based on the number of de novo damaging variants in OCD versus controls, we estimated the number of genes contributing to OCD risk, and used this estimate to predict future risk gene discovery as additional OCD parent-child trios are studied by exome sequencing.

Figure 2 – De novo damaging variants are associated with OCD risk. (A) Bar chart comparing the rates of de novo mutation types between OCD cases (red) and controls (blue). Comparisons are between per base pair (bp) mutation rates, considering only those "callable" loci in each family and cohort that meet required sequencing depth and quality scores to support high confidence de novo variant calling. Mutation rates were compared using a one-tailed rate ratio test. Statistically significant comparisons ($p<0.05$) are marked with asterisks. Error bars show 95% confidence intervals. **(B)** For the enriched classes of de novo variants, we quantified their contribution to OCD risk in two ways. First, we estimated the percentage of observed variants carrying risk by dividing the difference in rates (estimated coding variants per individual, see Table 1 and Methods) by the rate in OCD. Second, we estimated the percentage of cases with a mutation mediating risk by subtracting the proportion of controls carrying a mutation from the proportion in OCD probands carrying a mutation.

Figure 1 – Study summary

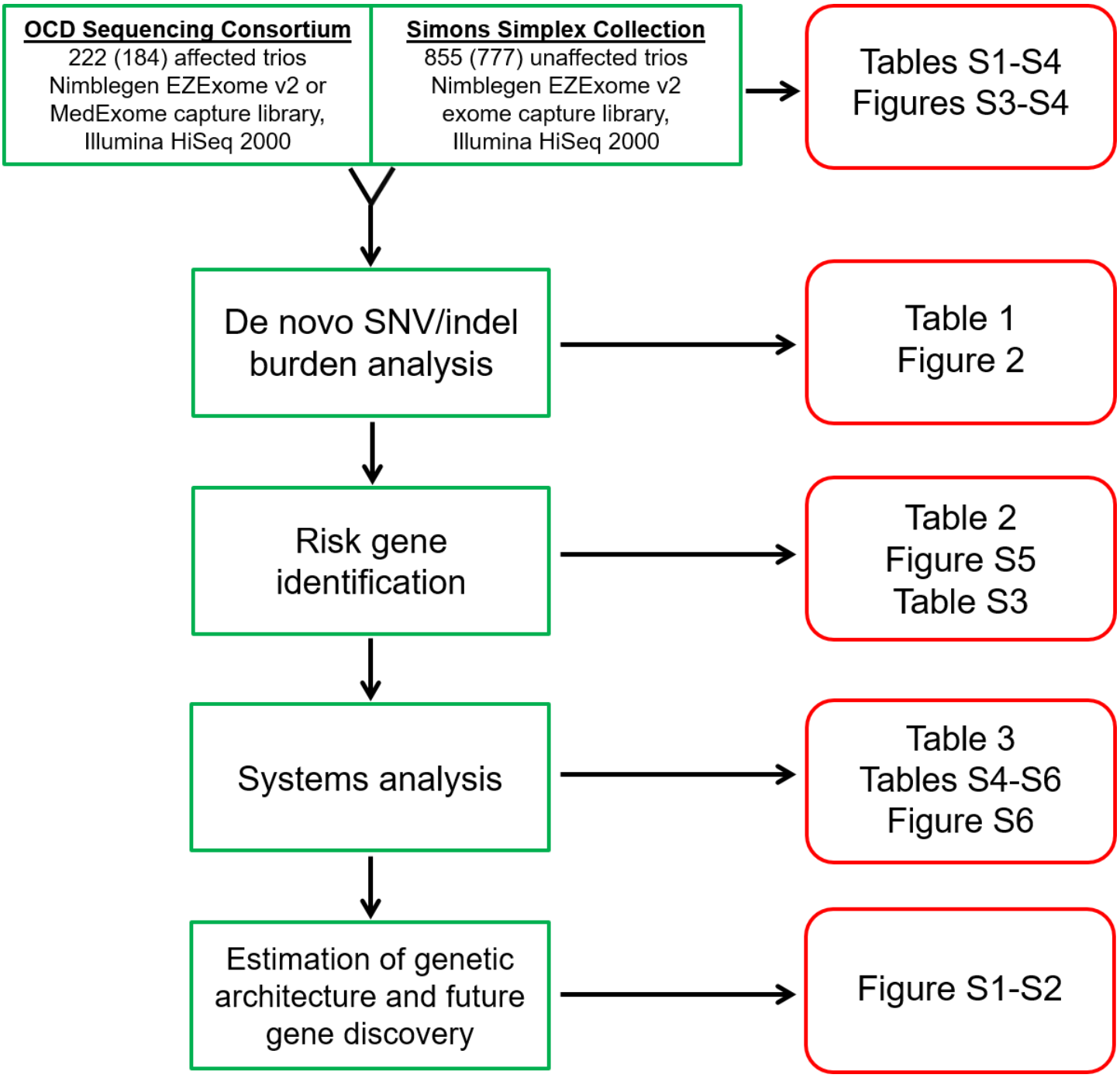
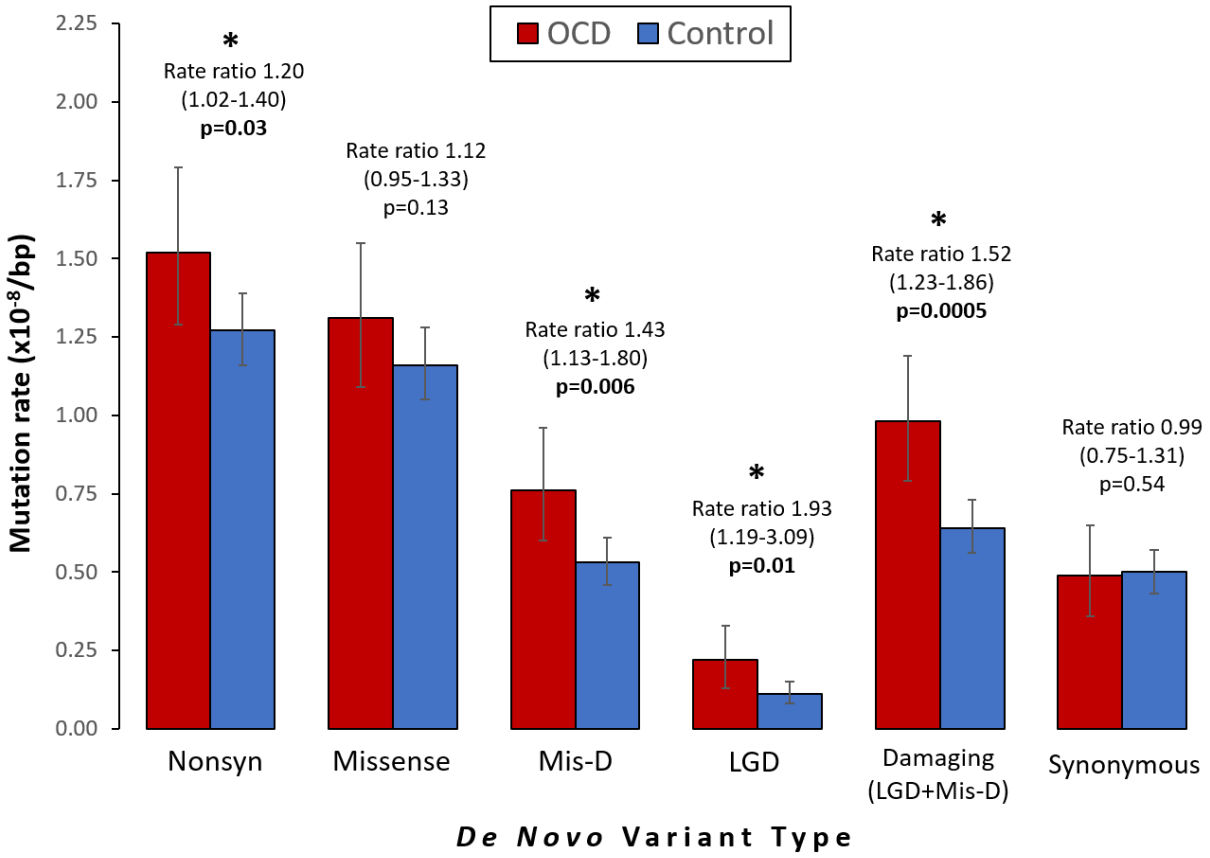


Figure 2 – De novo damaging .variants are associated with OCD risk

A.



B.

De novo variant type	Mutations per individual (95% CI)		% of mutations carrying risk (95% CI)	% of cases with mutation mediating risk (95% CI)
	OCD (N=184)	Control (N=777)		
Predicted damaging missense (Mis-D)	0.51 (0.41-0.65)	0.36 (0.31-0.41)	29.5 (6.0-53.0)	15.0 (3.1-26.9)
Likely Gene Disrupting (LGD)	0.15 (0.088-0.22)	0.075 (0.054-0.10)	49.2 (3.4-95.0)	7.3 (0.50-14.0)
Damaging (LGD + Mis-D)	0.66 (0.53-0.81)	0.43 (0.38-0.49)	33.9 (13.3-54.6)	22.2 (8.7-35.8)

Table 1 – Distribution of de novo variants in OCD cases and controls

De novo variant type[a]	Variant counts		Mutation rate (x10^{-8}) per bp (95% CI)[j]		Estimated coding variants per individual (95% CI)[k]		Rate ratio (95% CI)	p-value[l]
	OCD (N=184)	Control (N=777)	OCD (N=184)	Control (N=777)	OCD (N=184)	Control (N=777)		
All[b]	207	701	2.02 (1.75-2.31)	1.80 (1.67-1.94)	1.37 (1.18-1.56)	1.22 (1.13-1.31)	1.12 (0.95-1.31)	0.11
Coding[c]	200	662	2.06 (1.78-2.36)	1.80 (1.67-1.95)	1.39 (1.20-1.60)	1.22 (1.13-1.32)	1.14 (0.99-1.30)	0.06
Synonymous SNV	48	182	0.49 (0.36-0.65)	0.50 (0.43-0.57)	0.33 (0.24-0.44)	0.34 (0.29-0.39)	0.99 (0.75-1.31)	0.54
Nonsynonymous[d]	148	467	1.52 (1.29-1.79)	1.27 (1.16-1.39)	1.03 (0.87-1.21)	0.86 (0.78-0.94)	**1.20 (1.02-1.40)**	**0.03**
All Missense (Mis)	127	426	1.31 (1.09-1.55)	1.16 (1.05-1.28)	0.89 (0.74-0.20)	0.78 (0.71-0.87)	1.12 (0.95-1.33)	0.13
Mis-D[e]	74	195	0.76 (0.60-0.96)	0.53 (0.46-0.61)	0.51 (0.41-0.65)	0.36 (0.31-0.41)	**1.43 (1.13-1.80)**	**0.006**
MIs-P[f]	18	79	0.19 (0.11-0.29)	0.22 (0.17-0.27)	0.13 (0.074-0.20)	0.15 (0.12-0.18)	0.86 (0.53-1.34)	0.76
Mis-B[g]	33	147	0.34 (0.23-0.48)	0.40 (0.34-0.47)	0.23 (0.16-0.32)	0.27 (0.23-0.32)	0.85 (0.60-1.17)	0.83
Likely Gene Disrupting (LGD)[h]	21	41	0.22 (0.13-0.33)	0.11 (0.080-0.15)	0.15 (0.088-0.22)	0.074 (0.054-0.10)	**1.93 (1.19-3.09)**	**0.01**
Damaging (LGD + Mis-D)	95	236	0.98 (0.79-1.19)	0.64 (0.56-0.73)	0.66 (0.53-0.81)	0.43 (0.38-0.49)	**1.52 (1.23-1.86)**	**0.0005**
LGD SNV	14	20	0.14 (0.079-0.24)	0.055 (0.033-0.084)	0.095 (0.053-0.16)	0.037 (0.022-0.057)	**2.64 (1.39-4.93)**	**0.006**
LGD frameshift indel	7	21	0.072 (0.029-0.15)	0.057 (0.035-0.088)	0.049 (0.020-0.10)	0.039 (0.024-0.060)	1.28 (0.53-2.72)	0.37

Nonframeshift indel	2	5	0.021 (0.0025-0.074)	0.014 (0.0044-0.032)	0.014 (0.00017-0.050)	0.0095 (0.0030-0.022)	1.51 (0.21-7.28)	0.44
Unknown[i]	2	8	0.021 (0.0025-0.074)	0.022 (0.0094-0.043)	0.014 (0.00017-0.050)	0.015 (0.0064-0.029)	0.94 (0.14-3.88)	0.65

[a]Variants were annotated with Annovar, using RefSeq hg19 gene definitions. [b]"All" includes coding and non-coding variants. [c]"Coding" variants include synonymous, nonsynonymous, nonframeshift, and those annotated as "unknown" by Annovar. [d]"Nonsynonymous" variants include all missense and LGD variants. [e]"Mis-D" are "probably damaging" missense variants with a Polyphen2 (HDIV) score ≥0.957. [f]Mis-P are "possibly damaging" missense variants with a Polyphen2 (HDIV) score <0.957 and ≥0.453. [g]Mis-B are "benign" missense variants with a Polyphen2 (HDIV) score <0.453. Two OCD missense variants and five control missense variants had no prediction by Polyphen2, but were included in the "All Missense (Mis)" variant type. [h]LGD variants are those altering a stop codon, canonical splice site, and frameshift indels. [i]"Unknown" variants are not included in the synonymous or nonsynonymous counts. [j]De novo mutation rates were calculated as the number of variants divided by the number of haploid "callable" bases (see Methods). [k]The estimated number of de novo mutations per individual was calculated by multiplying the mutation rate by the size of the RefSeq hg19 coding exome (33,828,798 bp). [l]Rates were compared using a one-sided rate ratio test. Rate ratios, 95% CI, and p-values that are statistically significant ($p<0.05$) are underlined and in bold. See also Figure 2A.

Table 2 – Risk gene discovery in OCD

Gene	# LGD	# Mis-D	p-value	q-value (FDR)
SCUBE1	**1**	**1**	**2.1×10^{-6}**	**0.040**
CHD8	**1**	**1**	**2.5×10^{-6}**	**0.043**
TTN	0	3	0.0005	0.47

We used the TADA-Denovo algorithm (He et al., 2013) to estimate the likelihood of observing gene-level recurrence of de novo damaging variants in three genes in unrelated individuals. Two of these genes, *SCUBE1* and *CHD8*, exceeded the false discovery rate (FDR) threshold for high-confidence risk genes ($q<0.1$). Despite observing three de novo damaging variants in *TTN*, this gene did not meet criteria for a high-confidence or even a probable ($q<0.3$) risk gene, owing to its large size and high expected de novo mutation rate. See also Figure S5 and Tables S2-S3.

Table 3 – Overlap between OCD de novo damaging mutations and gene sets

Comparison gene set[a]	LGD Obs[b]	LGD Exp[c]	LGD O/E[d]	LGD P[e]	Missense Obs	Missense Exp	Missense O/E	Missense P	Synonymous Obs	Synonymous Exp	Synonymous O/E	Synonymous P
ASD	9	4.55	1.98	**0.038**	25	17.94	1.39	**0.043**	9	8.06	1.12	0.41
SCZ	1	1.00	1.00	0.64	7	5.40	1.30	0.29	0	1.84	0	1
DD	2	2.77	0.72	0.77	5	6.50	0.77	0.79	4	2.27	1.76	0.19
TD	0	0.26	0	1	2	1.90	1.05	0.57	2	0.77	2.61	0.18
ID	0	0.49	0	1	1	1.28	0.78	0.72	0	0.43	0	1
Unaffected	2	1.35	1.48	0.39	5	8.98	0.56	0.95	2	4.20	0.48	0.93

	Obs	Exp	O/E	P
DD significant[f]	4	0.84	4.76	**0.010**
ASD – q<0.1[g]	3	0.65	4.62	**0.028**
CHD8 brain[h]	20	13.02	1.54	**0.030**

[a]Comparisons for autism (ASD), schizophrenia (SCZ), developmental disorders (DD), Tourette's disorder (TD), intellectual disability (ID), and unaffected siblings are between genes harboring de novo LGD, missense, or synonymous variants in these phenotypes and those harboring damaging (LGD or Mis-D) variants in OCD (n=89). Gene lists and their references are in Table S4. [b]Observed number of genes overlapping between sets. [c]Expected number of genes overlapping between sets, determined by 100,000 random mutation set simulations using DNENRICH (Fromer et al., 2014). [d]O/E, observed divided by expected number of genes overlapping between sets. [e]P-value is one-sided under a binomial model of greater than expected hits per gene set, calculated by DNENRICH using 100,000 permutations. [f]93 genes with genome-wide significant statistical evidence for association with developmental disorders (Deciphering Developmental Disorders Study, 2017). [g]65 ASD genes with False Discovery Rate (q) < 0.1 by TADA, considering data across exome sequencing studies (Sanders et al., 2015). [h]Genes in human brain with promoters targeted by CHD8 (Cotney et al., 2015). See also Tables S2 and S4.

SUPPLEMENTARY METHODS

Principal Component Analysis (PCA)

PCA was performed on all sequencing quality metrics (Table S1) in R using the following code:

```
library(xlsx)
library("FactoMineR")
library("factoextra")
library("corrplot")

## Load Data from Table S1, first tab
data1 <- read.delim("Table_S1.xlsx")
# select only certain columns
data1.temp <- data1[c(2,14:45)]
# make first column the row names
data1.active <- data.frame(data1.temp[,-1], row.names=data1.temp[,1])

## Load additional data for later use (non-numeric labels/groups)
# select only certain columns
data2.temp <- data1[c(4,3,1)]
# make first column the row names
data2.active <- data.frame(data2.temp[,-1], row.names=data2.temp[,1])

## Principal component analysis
pdf("PCA_factor_maps.pdf")
res.pca <- PCA(data1.active, scale.unit = TRUE, ncp = 10, graph = TRUE, axes
= c(1,2))
dev.off()

print(res.pca)

## Export PCA coordinates to determine outliers (Table S1)
indcoord<-res.pca$ind$coord
write.xlsx(indcoord, "Table_S1.xlsx")

## Estimate the number of components in Principal Component Analysis
(Factominer)
sink("EstmateNumberPCs.txt")
estim_ncp(data1.active, ncp.min=0, ncp.max=NULL, scale=TRUE, method="Smooth")
sink()

## Variances of the principal components
eigenvalues <- res.pca$eig

## Make scree plot using base graphics : A scree plot is a graph of the
eigenvalues/variances associated with components (Figure S3.A).
pdf("ScreePlot.pdf")
barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
       main = "Variances",
```

```
       xlab = "Principal Components",
       ylab = "Percentage of Variance",
       col ="steelblue")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
      type="b", pch=19, col = "red")
dev.off()

## Make cumulative variance graph (Figure S3.B)
pdf("ScreePlot_cumulative.pdf")
barplot(eigenvalues[, 3], names.arg=1:nrow(eigenvalues),
        main = "Variances",
        xlab = "Principal Components",
        ylab = "Cumulative Percentage of Variance",
        col ="steelblue")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 3],
      type="b", pch=19, col = "red")
dev.off()

## GRAPHS OF VARIABLES
pdf("PCA_factor_maps_variables.pdf")
fviz_pca_var(res.pca, col.var="contrib") + scale_color_gradient2(low="white",
mid="blue", high="red", midpoint=55)+theme_bw()
dev.off()

## GRAPHS OF INDIVIDUALS (Figure S3.C)
pca = prcomp(data1.active, scale = TRUE)
pdf("PCA_prcomp_factor_map_indiv.pdf")
plot(pca$x, pch = 20, col = c(rep("red", 366), rep("blue", 1200)))
dev.off()
```

Calculating “callable” base pairs

Within each cohort, we calculated the rate of de novo mutations per base pair. For accurate rate calculation, we first determined the number of “callable” base pairs per family using the GATK DepthOfCoverage tool. We considered only bases covered at ≥ 20x in all family members, with base quality ≥ 20, and map quality ≥ 30; these thresholds match those required for GATK and de novo variant calling. The following command was used to calculate the callable base pairs in each trio:

```
java -jar GenomeAnalysisTK.jar -T DepthOfCoverage -R human_g1k_v37.fasta -o
FamilyID -I FamilyID.list -L target_intersection.bed --minMappingQuality 30 -
-minBaseQuality 20 --summaryCoverageThreshold 20
```

`FamilyID.list` contains names and locations of the three trio bam files. The .bed file contains the genomic intervals over which to calculate the callable base pairs. To calculate the

coding callable base pairs (used for coding mutation rates, e.g. synonymous, nonsynonymous, missense, etc., see Table 1, Table S2), we used a bed file with intervals spanning the intersection of both capture array target intervals and the RefSeq coding intervals (32,027,823 bp total). To calculate all callable base pairs (used for the total coding + noncoding mutation rate, see "All" in Table 1), we used a bed file with intervals spanning the intersection of both capture array target intervals (33,973,867 bp total). The number of coding and total callable base pairs for every family passing is listed in Table S1.

Contribution of de novo mutations to OCD risk

For every proband and control subject passing QC (obtained from Table S1), we made a file containing the number of Mis-D, LGD, and damaging (Mis-D + LGD) variants (obtained from Table S2), then calculated the haploid mutation rate per subject for each variant type. Haploid mutation rates were calculated by dividing the number of mutations by the twice the number of callable coding bases ("CallableExomeCoding" in Table S1).

The following R code was then used to calculate the percentage of cases with a mutation mediating risk and the percentage of mutations carrying risk, along with 95% confidence intervals for each. The number of de novo mutations per individual was calculated by multiplying the mutation rate by the size of the RefSeq hg19 coding exome (33,828,798 bp).

```
library(magrittr)
library(dplyr)

#### Load files containing haploid mutation rates
caseVarFile="OCD_rates.txt"
caseVarData <- read.table(caseVarFile, sep="\t", header=T)

ctrlVarFile="SSC_rates.txt"
ctrlVarData <- read.table(ctrlVarFile, sep="\t", header=T)

#### Calculations for Mis-D variants
MisD_contrib <- t.test(caseVarData$haploidRatesMisD,
ctrlVarData$haploidRatesMisD, paired = F)

df_MisD_contrib <- data.frame(case.rate = MisD_contrib$estimate[1] * 33828798
* 2,
```

```
                    ctrl.rate = MisD_contrib$estimate[2] * 33828798 * 2,
                    diff.lower.ci = MisD_contrib$conf.int[1] * 33828798 * 2,
                    diff.upper.ci = MisD_contrib$conf.int[2] * 33828798 * 2)
df_MisD_contrib <- df_MisD_contrib %>% mutate(diff = case.rate - ctrl.rate,
percent = diff / case.rate,
                    percent.lower.ci = diff.lower.ci / case.rate,
                    percent.upper.ci = diff.upper.ci / case.rate)

#### Calculations for LGD variants
LGD_contrib <- t.test(caseVarData$haploidRatesLGD,
ctrlVarData$haploidRatesLGD, paired = F)
df_LGD_contrib <- data.frame(case.rate = LGD_contrib$estimate[1] * 33828798 *
2,
                    ctrl.rate = LGD_contrib$estimate[2] * 33828798 * 2,
                    diff.lower.ci = LGD_contrib$conf.int[1] * 33828798 * 2,
                    diff.upper.ci = LGD_contrib$conf.int[2] * 33828798 * 2)
df_LGD_contrib <- df_LGD_contrib %>% mutate(diff = case.rate - ctrl.rate,
percent = diff / case.rate,
                percent.lower.ci = diff.lower.ci / case.rate,
                percent.upper.ci = diff.upper.ci / case.rate)

#### Calculations for Damaging variants
Damaging_contrib <- t.test(caseVarData$haploidRatesDamaging,
ctrlVarData$haploidRatesDamaging, paired = F)
df_Damaging_contrib <- data.frame(case.rate = Damaging_contrib$estimate[1] *
33828798 * 2,
                ctrl.rate = Damaging_contrib$estimate[2] * 33828798 * 2,
                diff.lower.ci = Damaging_contrib$conf.int[1] * 33828798 * 2,
                diff.upper.ci = Damaging_contrib$conf.int[2] * 33828798 * 2)
df_Damaging_contrib <- df_Damaging_contrib %>% mutate(diff = case.rate -
ctrl.rate, percent = diff / case.rate,
                percent.lower.ci = diff.lower.ci / case.rate,
                percent.upper.ci = diff.upper.ci / case.rate)
```

Mutation rates for variant simulations

To perform subsequent maximum likelihood estimation and TADA analyses, we used published per gene de novo mutation rates from unaffected parent-child trios (1). For the control samples in our dataset, we calculated the proportion of the overall coding mutation rate that comprised LGD and Mis-D mutations, and then used these proportions to calculate the expected LGD and Mis-D mutation rate per gene.

The following R code was used to generate the mutation rate tables:

```
library(denovolyzeR)
library(plyr)
```

```
########################
#### Mutation type fractions from controls in OCD project
########################

# fractions of overall coding mutation rate for each variant type in SSC
controls (see Table 1)
fracLGD <- 0.0611
fracMisD <- 0.2944

########################
#### Get published de novo mutation rtaes
########################

denovolyzer <- viewProbabilityTable()
mutationProbs <- denovolyzer[ , c("geneName", "all")]
mutationProbs <- rename(mutationProbs, c("geneName"="gene.name"))  #rename
column
mutationProbs <- rename(mutationProbs, c("all"="mut.rate"))  #rename column
#save(mutationProbs, file = "denovolyzer_rates_all_unadjusted.RData")
#write.table(mutationProbs, "denovolyzer_rates_all_unadjusted.txt", sep="\t")

########################
#### Add LGD and Mis-D de novo mutation rates based on fractions seen in our
study
########################

mutationProbs$lgd <- mutationProbs$mut.rate * fracLGD
mutationProbs$misD <- mutationProbs$mut.rate * fracMisD

save(mutationProbs, file = "de_novo_mutation_rates.RData")
write.table(mutationProbs, "de_novo_mutation_rates.RData", sep="\t")
```

TADA analysis

The enrichment of de novo LGD and Mis-D mutations in OCD raises the possibility that these classes of mutations target a set of genes that mediates OCD risk. To test this hypothesis, we used the transmitted and de novo association (TADA) test, a Bayesian model that can effectively combine data from de novo mutations, inherited variants in families, and standing variants in the population (via case-control cohorts) to significantly increase the power of gene discovery. In this study, we elected not to include the rare inherited exome variants because their confirmation rate is not known and their contribution to the TADA score is minimal given their lower relative risks (2, 3). Instead, we used a specialized version of TADA that analyzes only the de novo mutations from exome sequencing data, called TADA-Denovo (2).

The code and documentation for this tool can be found here (TADA.v1.1.R; http://wpicr.wpic.pitt.edu/WPICCompGen/TADA/TADA_homepage.htm).

The TADA-Denovo test considers two types of variants, de novo LGD and de novo severe missense (those predicted by PolyPhen2-HDIV to be "probably damaging" to protein function, abbreviated as "Mis-D"). The main input is the number of de novo LGD and number of de novo Mis-D variants per gene. Additionally, mutation rates (mu) for all human genes are provided, based on Sanders et al. (4). The test analyzes each of these event types separately, then combines the evidence in a Bayesian fashion, weighing each type of mutation differently.

To compute the Bayes factors and p-values, TADA-Denovo requires the following parameters:

- n.family: the number of OCD parent-child trios passing QC (184)
- The fold-enrichment (λ) for each mutational class is calculated as follows, where X_{case} and X_{ctrl} are the number of mutations in cases and controls, respectively; S_{case} and S_{ctrl} are the number of synonymous mutations in case and controls, respectively: $\lambda = (X_{case} / (X_{ctrl} \cdot (S_{case}/S_{ctrl})))$. As shown in the code below, we used the estimated coding variants per individual (Table 1) and multiplied by the number of cases or controls. Fold-enrichment for LGD was calculated as 2.09, and fold-enrichment for Mis-D was 1.46.
- The fraction of causal genes (π) is the estimated number of OCD risk genes (335, calculated by MLE as detailed below) divided by the number of RefSeq genes with mutational rates included in the TADA-Denovo algorithm. $\pi = 335 / 19618 = 0.0171$
- gamma.mean.dn: The average relative risk (γ) is related to the fold-enrichment (λ) and the fraction of causal genes (π) by the following equation: $\pi \cdot (\gamma-1) = \lambda-1$. Solving for γ gives LGD $\gamma = 64.74$, Mis-D $\gamma = 27.91$).

Using these parameters, TADA-Denovo calculates the Bayes factors of all input genes. Next, it computes the p-value for each gene by generating random mutational data, based on each gene's specified mutation rate, to obtain a null distribution of Bayes factors. We used 1,000 samplings of de novo mutations in each gene to determine null distributions. Finally, TADA-Denovo calculates an FDR q-value for each gene using a Bayesian "direct posterior approach." A low q value represents strong evidence for OCD association. Genes with FDR < 0.3 are considered probable risk genes, and those with FDR < 0.1 are high-confidence risk genes.

The following R code performed the TADA-Denovo analysis:

```
source("TADA.v1.1.R")

# set.seed(100)

### read mutation rates and counts (see Table S3, second tab)
tada.file="Table_S3.txt"
tada.data=read.table(tada.file,header=T)

### Number of mutations and TADA parameters

numLgdMutations <- (0.15*184)
lgdRate <- numLgdMutations/184
numControlLgdMutations <- (0.074*777)
controlLgdRate <- numControlLgdMutations/777
lgdRiskFraction <- (lgdRate - controlLgdRate) / lgdRate

numMis3Mutations <- (0.51*184)
mis3Rate <- numMis3Mutations/184
numControlMis3Mutations <- (0.36*777)
controlMis3Rate <- numControlMis3Mutations/777
mis3RiskFraction <- (mis3Rate - controlMis3Rate) / mis3Rate

numGenes <- 335  # from MLE analysis below

nPerms <- 1000

pi <- numGenes / nrow(tada.data)
pi0 <- 1-pi

numSilentMutations <- (0.33*184)
numControlSilentMutations <- (0.34*777)

dn.lof.lambda <- (numLgdMutations) / (numControlLgdMutations *
(numSilentMutations/numControlSilentMutations))
dn.lof.relativeRisk <- 1 + ((dn.lof.lambda-1) / pi)
```

```
dn.mis3.lambda <- (numMis3Mutations) / (numControlMis3Mutations *
(numSilentMutations/numControlSilentMutations)) # num Mis3 in ctrls, num
silent in cases, num silent in ctrls
dn.mis3.relativeRisk <- 1 + ((dn.mis3.lambda-1) / pi)

n.family = 184
n = data.frame(dn=n.family, ca=NA, cn=NA)
sample.counts <- list(cls1=n, cls2=n)

### create the mutational data used by TADA-Denovo
cls1.counts=data.frame(dn=tada.data$dn.cls1, ca=NA, cn=NA)
rownames(cls1.counts)=tada.data$gene.id
cls2.counts=data.frame(dn=tada.data$dn.cls2, ca=NA, cn=NA)
rownames(cls2.counts)=tada.data$gene.id
tada.counts=list(cls1=cls1.counts,cls2=cls2.counts)

### set up mutation rates
mu=data.frame(cls1=tada.data$mut.cls1,cls2=tada.data$mut.cls2)

### specify de novo only analyses
denovo.only=data.frame(cls1=TRUE,cls2=TRUE)

### set up parameters -
cls1=
data.frame(gamma.mean.dn.=dn.lof.relativeRisk,beta.dn=1,gamma.mean.CC=NA,beta
.CC=NA ,rho1=NA,nu1=NA,rho0=NA,nu0=NA)
cls2=
data.frame(gamma.mean.dn=dn.mis3.relativeRisk,beta.dn=1,gamma.mean.CC=NA,beta
.CC=NA,rho1=NA,nu1=NA,rho0=NA,nu0=NA)
hyperpar=list(cls1=cls1,cls2=cls2)

### running TADA-Denovo
re.TADA <- do.call(cbind.data.frame, TADA(tada.counts=tada.counts,
sample.counts=sample.counts, mu=mu, hyperpar=hyperpar,
denovo.only=denovo.only))

### Bayesian FDR control
re.TADA$qval=Bayesian.FDR(re.TADA$BF.total, pi0 = pi0)

### run permutation to get the null distributions to use for calculating p-
values for TADA
re.TADA.null=do.call(cbind.data.frame, TADAnull(tada.counts=tada.counts,
sample.counts=sample.counts, mu=mu, hyperpar=hyperpar,
denovo.only=denovo.only, nrep=nPerms))
re.TADA$pval=bayesFactor.pvalue(re.TADA$BF.total,re.TADA.null$BFnull.total)

### display top 10 genes based on BF.total
re.TADA[order(-re.TADA$BF.total)[1:10],]

### write all table to file - See Table S3
write.table(re.TADA, "TADA_denovo_results_.txt", sep="\t")
save.image(file="TADA_OCD_Workspace_tada.RData")
```

Applying this model to 184 OCD parent-child trios passing QC identifies two high confidence OCD risk genes (q<0.1) (Table 2, Table S3).

Maximum Likelihood Estimation (MLE) method for estimating the number of OCD risk genes

The following R code was used to perform these calculations:

```
library(ggplot2)
library(parallel)
library(data.table)

plotDir <- getwd()
load("de_novo_mutation_rates.RData") # see Table S3, first tab
mutationProbs <- as.data.table(mutationProbs)

K <- 95  # total OCD de novo damaging mutations (Mis-D+LGD)
R2 <- 2 # number of above mutations hitting same gene twice
R3 <- 1 # number of above mutations hitting sane gene three times
M1 <- 95/184  # observed rate of de novo damaging mutations in OCD
M2 <- 236/777 # observed rate of de novo damaging mutations in controls
E <- (M1-M2)/M1 # estimating fraction of de novo damaging variants carrying
risk
nPerms <- 100000 # number of permutations to perform at each assumed number
of risk genes
maxGenes <- 2500 # perform permutations from 1 to this number

# get number of cores available
numCores <- max(1, detectCores() - 1)

################################################################################
# FUNCTIONS
################################################################################
getRecurrence <- function(G, mutationProbs, K, E, R2, R3, nPerms){
    permutationVector <- lapply(1:nPerms, function(x) {
        riskGeneIndex <- sample(1:nrow(mutationProbs), G, replace = F)
        riskGenes <- mutationProbs[riskGeneIndex,]
        nonRiskGenes <- mutationProbs[-riskGeneIndex,]

        C1 <- rbinom(1,K,E)
        C2 <- K-C1

        C1geneMutations <- sample(riskGenes$gene.name, C1, replace = T, prob
= riskGenes$damaging)
        C2geneMutations <- sample(nonRiskGenes$gene.name, C2, replace = T,
prob = nonRiskGenes$damaging)

        allGeneMutations <- c(C1geneMutations, C2geneMutations)
        length(which(table(allGeneMutations)==2))==R2 &
length(which(table(allGeneMutations)>=3))==R3
    })
```

```
    proportionMatchingObserved <-
length(which(unlist(permutationVector)))/nPerms

}
################################################################################
# RUN
################################################################################
RNGkind("L'Ecuyer-CMRG")
set.seed(1)
mc.reset.stream()

permutationTest <- mclapply(1:maxGenes,
                    function(x) getRecurrence(x, mutationProbs, K, E, R2, R3,
nPerms),
                    mc.cores = numCores, mc.set.seed = T )

save(permutationTest, file = paste("UpTo", maxGenes, "genes", nPerms,
"perms", "MaxLikelihoodPermutation.RData", sep="_"))

toPlot <- data.frame(likelihood = unlist(permutationTest), nGenes =
1:length(unlist(permutationTest)))

save(toPlot, file = paste("UpTo", maxGenes, "genes", nPerms, "perms",
"MaxLikelihoodPermutationToPlot.RData", sep="_"))

p <- ggplot(toPlot, aes(x=nGenes, y=likelihood))
p <- p + geom_line() + geom_smooth()
ggsave(file.path(plotDir, paste("UpTo", maxGenes, "genes", nPerms, "perms",
"MaxLikelihood.pdf", sep="_") ), p)

save.image(file = paste("Workspace", nPerms, "perms.RData", sep="_"))
```

"Unseen species" method for estimating the number of OCD risk genes

Following a method used previously in estimating the number of risk genes for autism spectrum disorder (4), we used the following R code to obtain a second estimate of the number of risk genes (C) in OCD and 95% confidence intervals.

```
library(epitools)
# total OCD de novo damaging mutations (Mis-D+LGD)
damaging <- 95
# scale the observed number of damaging mutations in controls to get expected
number in cases
expectedDamaging <- ceiling( (236) * 184/777 )
# calculate the number of risk associated mutations
d <- damaging - expectedDamaging
nRecurrent <-  3
expectedNumRecurrent <- 2 # there were 8 in 777 controls, so (8*184/777)
# total number of observed risk genes
c <- d - (nRecurrent + 2)
# number of genes mutated once
```

```
c1 <- c - nRecurrent
# probability that newly added mutation hits a previously mutated gene
u = 1 - c1/d
#Estimate the number of risk genes
( C <-  c/u + 1*d*(1-u)/u )
# 316.875

## calculating 95% confidence interval for number of risk genes
# import tab delimited file with number of damaging mutations per control
sample
ctrlDamagingFile="SSC_damaging.txt"
ctrlDamagingData <- read.table(ctrlDamagingFile, sep="\t", header=T)
# calculate upper and lower number of damaging mutations per control sample
ci_pois <- pois.exact(sum(ctrlDamagingData$NumDamaging), pt =
nrow(ctrlDamagingData), conf.level = 0.95)

# repeat above risk gene calculations using lower value - gives upper
estimate
expectedDamaging_low <- ceiling((ci_pois$lower*777) * 184/777 )
d_low <- damaging - expectedDamaging_low
c_low <- d_low - (nRecurrent + 2)
c1_low <- c_low - nRecurrent
u_low = 1 - c1_low/d_low
( C_low <-  c_low/u_low + 1*d_low*(1-u_low)/u_low )
# 454.25

# repeat above risk gene calculations using upper value - gives lower
estimate
expectedDamaging_hi <- ceiling((ci_pois$upper*777) * 184/777 )
d_hi <- damaging - expectedDamaging_hi
c_hi <- d_hi - (nRecurrent + 2)
c1_hi <- c_hi - nRecurrent
u_hi = 1 - c1_hi/d_hi
( C_hi <-  c_hi/u_hi + 1*d_hi*(1-u_hi)/u_hi )
# 189.875
```

Predicting the number of risk genes identified by cohort size

The following R code was used to perform these calculations:

```
library(ggplot2)
library(parallel)
library(reshape2)

plotDir <- getwd()
source(file = "TADA.v1.1.R")

load("de_novo_mutation_rates.RData")

numLgdMutations <- (0.15*184)
lgdRate <- numLgdMutations/184
numControlLgdMutations <- (0.074*777)
controlLgdRate <- numControlLgdMutations/777
```

```
lgdRiskFraction <- (lgdRate - controlLgdRate) / lgdRate

numMis3Mutations <- (0.51*184)
mis3Rate <- numMis3Mutations/184
numControlMis3Mutations <- (0.36*777)
controlMis3Rate <- numControlMis3Mutations/777
mis3RiskFraction <- (mis3Rate - controlMis3Rate) / mis3Rate

numGenes <- 335
nPerms <- 10000

# get number of cores available
numCores <- max(1, detectCores() - 1)

pi <- 0.01707615
pi0 <- 1-pi

numSilentMutations <- (0.33*184)
numControlSilentMutations <- (0.34*777)

dn.lof.lambda <- (numLgdMutations) / (numControlLgdMutations *
(numSilentMutations/numControlSilentMutations))
dn.lof.relativeRisk <- 1 + ((dn.lof.lambda-1) / pi)
dn.mis3.lambda <- (numMis3Mutations) / (numControlMis3Mutations *
(numSilentMutations/numControlSilentMutations)) # num Mis3 in ctrls, num
silent in cases, num silent in ctrls
dn.mis3.relativeRisk <- 1 + ((dn.mis3.lambda-1) / pi)


################################################################################
# FUNCTIONS
################################################################################

getGenes <- function(numGenes_f=numGenes, mutationProbs_f, cohortSize_f,
mutationRate_f, riskFraction_f, probability_f=c("lgd", "mis3")[1]){
        numMutations <- ceiling(cohortSize_f * mutationRate_f)
        riskGeneIndex <- sample(1:nrow(mutationProbs_f), numGenes_f, replace
= F)
        riskGenes <- mutationProbs_f[riskGeneIndex,]
        nonRiskGenes <- mutationProbs_f[-riskGeneIndex,]

        C1 <- rbinom(1, numMutations, riskFraction_f)
        C2 <- numMutations - C1

        C1geneMutations <- sample(riskGenes$gene.name, C1, replace = T, prob
= riskGenes$probability)
        C2geneMutations <- sample(nonRiskGenes$gene.name, C2, replace = T,
prob = nonRiskGenes$probability)

        allGeneMutations <- c(C1geneMutations, C2geneMutations)
}

runIteration <- function(numGenes_f=numGenes, mutationProbs_f, cohortSize_f,
lgdMutationRate_f, mis3MutationRate_f, lgdRiskFraction_f, mis3RiskFraction_f,
nTadaRep_f = 100){
    lgdMutations <- getGenes(numGenes, mutationProbs_f, cohortSize_f,
lgdMutationRate_f, lgdRiskFraction_f, "lgd")
```

```
    lgdMutations_df <- data.frame(gene=lgdMutations, lof=1, mis3=0,
stringsAsFactors = F)
    mis3Mutations <- getGenes(numGenes, mutationProbs_f, cohortSize_f,
mis3MutationRate_f, mis3RiskFraction_f, "lgd")
    mis3Mutations_df <- data.frame(gene=mis3Mutations, lof=0, mis3=1,
stringsAsFactors = F)
    combinedMutations <- rbind(lgdMutations_df, mis3Mutations_df)
    combinedMutations <- aggregate(combinedMutations[,c("lof", "mis3")],
by=list(combinedMutations$gene), sum)
    colnames(combinedMutations) <- c("gene.id", "dn.lof", "dn.mis3")
    tadaResults <- runTada(cohortSize_f = cohortSize_f, mutationTable_f =
combinedMutations, mutationProbs_f = mutationProbs_f, nTadaRep_f =
nTadaRep_f)
    return(tadaResults)
}

runTada <- function(cohortSize_f, mutationTable_f, mutationProbs_f,
nTadaRep_f = 100){
    tada.data <- merge(mutationTable_f, mutationProbs_f[,c("gene.name",
"lgd", "mis3")], by.x="gene.id", by.y="gene.name")
    names(tada.data)[which(names(tada.data)=="lgd")] <- "mut.lof"
    names(tada.data)[which(names(tada.data)=="mis3")] <- "mut.mis3"

    n.family = cohortSize_f
    n = data.frame(dn=n.family, ca=NA, cn=NA)
    sample.counts <- list(cls1=n, cls2=n)

    cls1.counts=data.frame(dn=tada.data$dn.lof, ca=NA, cn=NA)
    rownames(cls1.counts)=tada.data$gene.id
    cls2.counts=data.frame(dn=tada.data$dn.mis3, ca=NA, cn=NA)
    rownames(cls2.counts)=tada.data$gene.id
    tada.counts=list(cls1=cls1.counts,cls2=cls2.counts)

    mu=data.frame(cls1=tada.data$mut.lof,cls2=tada.data$mut.mis3)

    denovo.only=data.frame(cls1=TRUE,cls2=TRUE)

    cls1=
data.frame(gamma.mean.dn=dn.lof.relativeRisk,beta.dn=1,gamma.mean.CC=NA,beta.
CC=NA ,rho1=NA,nu1=NA,rho0=NA,nu0=NA)
    cls2= data.frame(gamma.mean.dn=
dn.mis3.relativeRisk,beta.dn=1,gamma.mean.CC=NA,beta.CC=NA,rho1=NA,nu1=NA,rho
0=NA,nu0=NA)
    hyperpar=list(cls1=cls1,cls2=cls2)


    re.TADA <- do.call(cbind.data.frame, TADA(tada.counts=tada.counts,
sample.counts=sample.counts, mu=mu, hyperpar=hyperpar,
denovo.only=denovo.only))
    re.TADA$qval=Bayesian.FDR(re.TADA$BF.total, pi0 = pi0)

    tadaResults <- re.TADA[order(re.TADA$qval, decreasing = F), ]

    probableGenes <- length(which(tadaResults$qval<0.3))
    highConfidenceGenes <- length(which(tadaResults$qval<0.1))
    return(data.frame(probable = probableGenes, highConfidence =
highConfidenceGenes))
```

```
}

################################################################################
# RUN
################################################################################

RNGkind("L'Ecuyer-CMRG")
set.seed(1)
mc.reset.stream()

tadaSimulations <- mclapply(seq(from=25, to=3000, by=25), function(x)
    lapply(1:nPerms, function(y) runIteration(numGenes_f = numGenes,
mutationProbs_f = mutationProbs, cohortSize_f = x, lgdMutationRate_f =
lgdRate, mis3MutationRate_f = mis3Rate, lgdRiskFraction_f = lgdRiskFraction,
mis3RiskFraction_f = mis3RiskFraction)),
    mc.cores = numCores, mc.set.seed = T)

save(tadaSimulations, file = paste("tadaSimulations", nPerms, "perms",
"forGeneDiscoveryEstimate_noPval.RData", sep="_"))

resultsByCohortSize <- lapply(tadaSimulations, function(x) do.call(rbind, x))

averageGeneDiscoveryByCohortSize <- lapply(resultsByCohortSize, function(x)
apply(x, 2, mean))

averageGeneDiscoveryByCohortSize_DF <- as.data.frame(do.call("rbind",
averageGeneDiscoveryByCohortSize))
averageGeneDiscoveryByCohortSize_DF$cohortSize <- seq(from=25, to=3000,
by=25)

save(averageGeneDiscoveryByCohortSize_DF, file =
paste("averageGeneDiscoveryByCohortSize", nPerms, "perms", ".RData",
sep="_"))

toPlot <- melt(averageGeneDiscoveryByCohortSize_DF,
measure.vars=c("probable", "highConfidence"),
               variable.name = "confidenceThreshold", value.name =
"numGenes")

save(toPlot, file = paste("averageGeneDiscoveryByCohortSizetoPlot", nPerms,
"perms", ".RData", sep="_"))

p <- ggplot(toPlot, aes(x=cohortSize, y=numGenes, col=confidenceThreshold))
p <- p + geom_line()

ggsave(p, file=file.path(plotDir, paste("averageGeneDiscoveryByCohortSize",
nPerms, "perms.pdf", sep="_")))
```

Gene set overlap using DNENRICH

The following Linux commands were used to run the DNENRICH analysis:

```
dnenrich . 100000 alias.txt refseq_gene_sizes.txt gene_lists.set
mutations_ocd_damaging.mut > results;

csh extractDnenrichResults.csh results > results.txt
```

References

1. Ware JS, Samocha KE, Homsy J, & Daly MJ (2015) Interpreting de novo Variation in Human Disease Using denovolyzeR. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* 87:7.25.21-15.

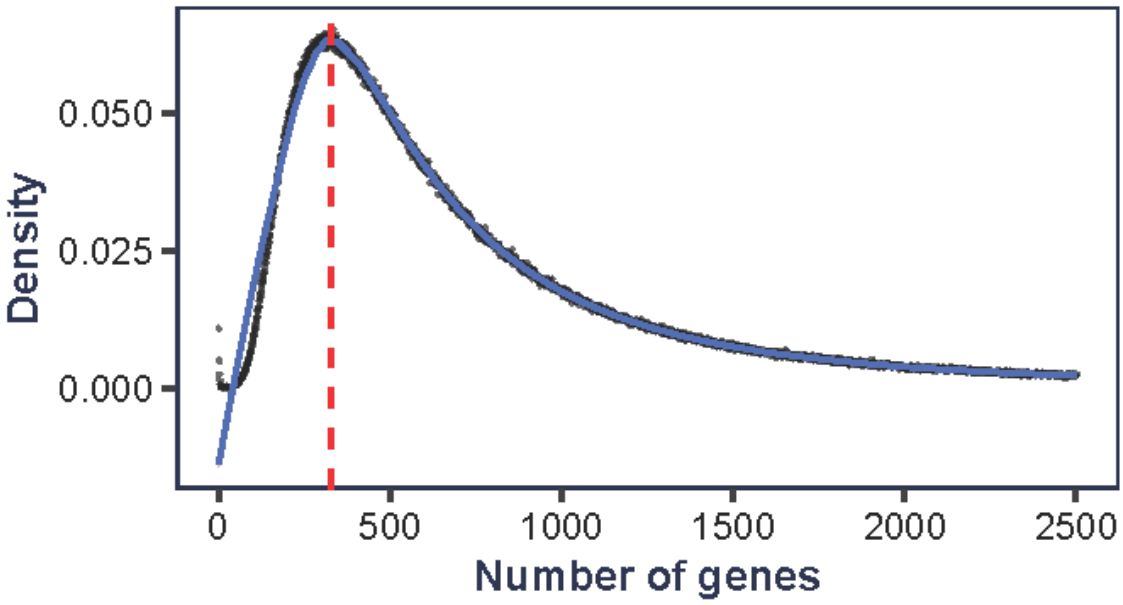
2. He X*, et al.* (2013) Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* 9(8):e1003671.

3. Sanders SJ*, et al.* (2015) Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 87(6):1215-1233.

4. Sanders SJ*, et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485(7397):237-241.

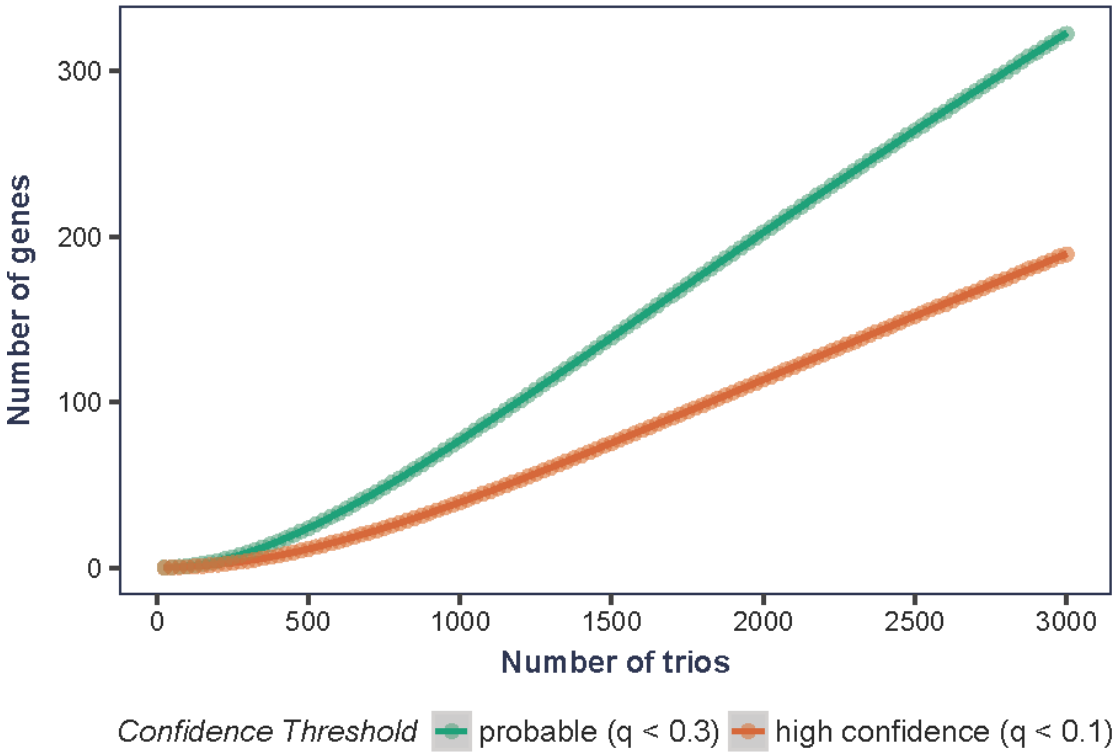
Supplementary Figures

Figure S1 – Maximum Likelihood Estimate (MLE) of number of OCD risk genes.



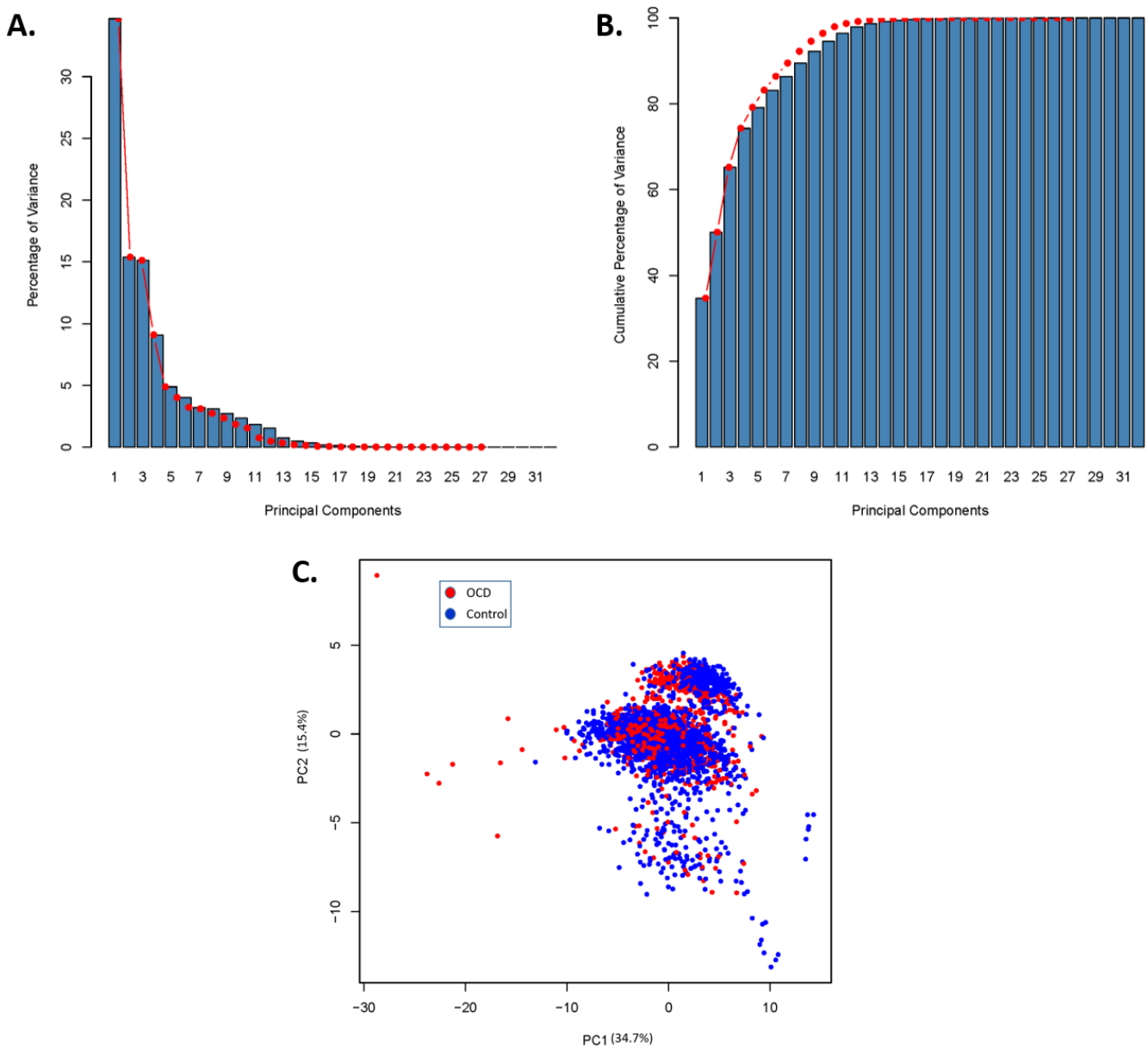
Assuming each number of possible risk genes between 1-2,500, 100,000 simulations were conducted to determine the number of risk genes that yielded the closest agreement between our observed and simulated data. In each simulation, we generated 95 variants (the number of de novo damaging variants observed in our OCD sample), then randomly assigned a percentage of variants (determined by the fraction of de novo damaging variants estimated to carry OCD risk) to the risk genes, recording the frequency at which the number of genes with two and three recurrent variants matched the number observed in our study (2 and 1, respectively). This MLE method yields an estimate of 335 OCD risk genes (red vertical line), a number that is in close agreement with that from an alternative "unseen species" method (317 genes, 95% CI: 190-454).

Figure S2 – Gene discovery by number of trios sequenced.



Using the MLE estimate of 335 risk genes, we estimated the number of probable (q<0.3) and high-confidence (q<0.1) risk genes that will be discovered as more OCD trios are sequenced. We performed 10,000 simulations at each cohort size from 25-3,000 trios, randomly generating variants and assigning to risk genes in agreement with the proportions seen in our data, then applying the TADA-Denovo algorithm.

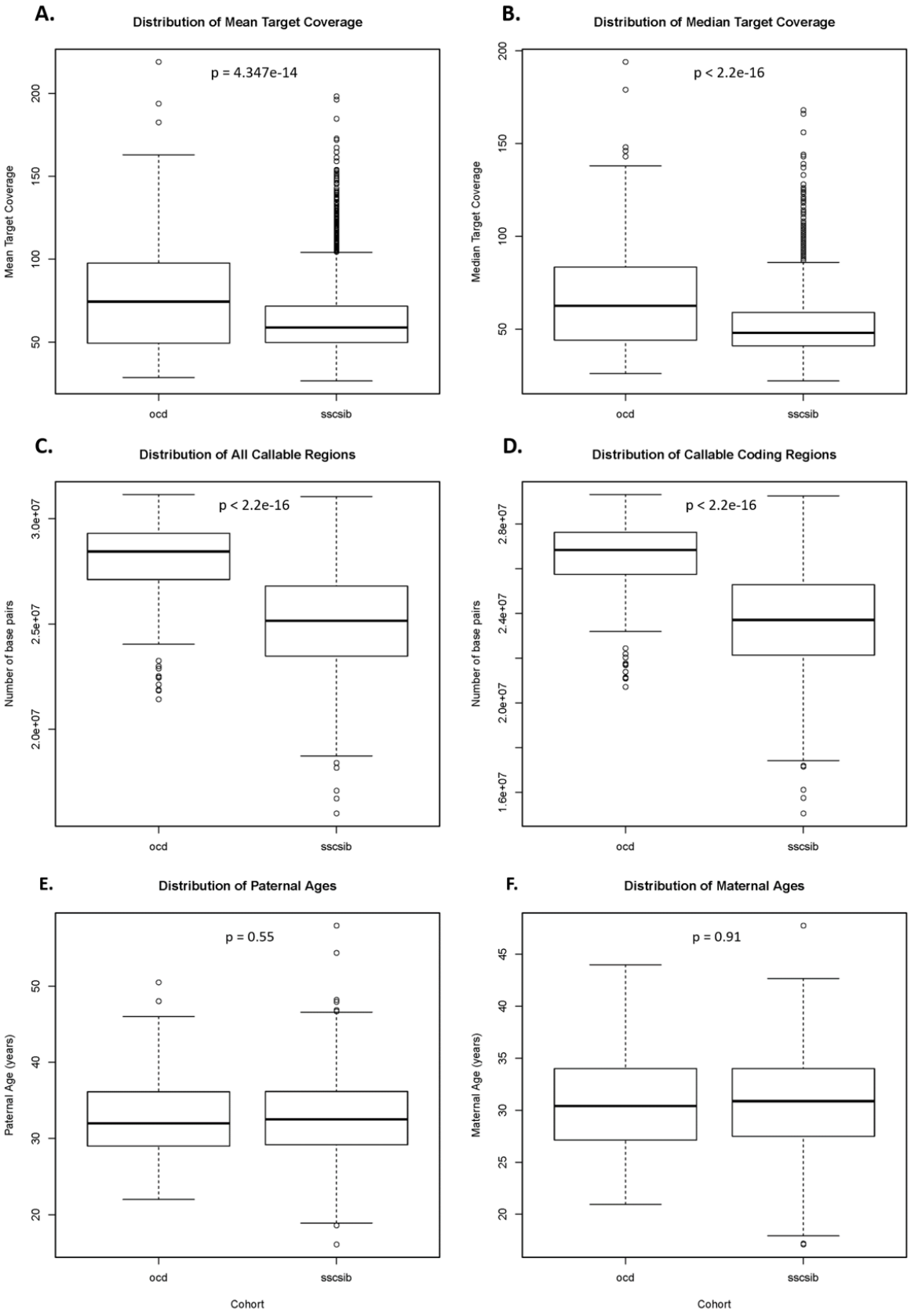
Figure S3 – PCA scree and individual plots.



Scree plots following Principal Components Analysis (PCA), showing (A) the percentage of variance captured by each of the first 32 principal components, and (B) the cumulative percentage of variance captured by these same components in the exome metrics data from cases and controls. The "elbow" of the scree plot is visualized to be around the 5th principal component. This was confirmed by the Factominer R code function "estim_ncp()". The first 5 PCs capture almost 80% of the variance, and this number of PCs was used to determine PCA outliers during quality control (see Table S1 and Supplementary Methods). (C) Individual plots for the first two principal components, based on PCA of exome sequencing quality metrics. OCD cases are plotted in red, and controls in blue. The first two PCs together capture 50.1% of the variance. R code to

generate this data and figure are in Supplementary Methods, and individual PC factor values are in Table S1. This figure includes PCA outliers (>3 standard deviations from the mean in PCs 1-5), which were removed during quality control, prior to further analysis of case-control data.

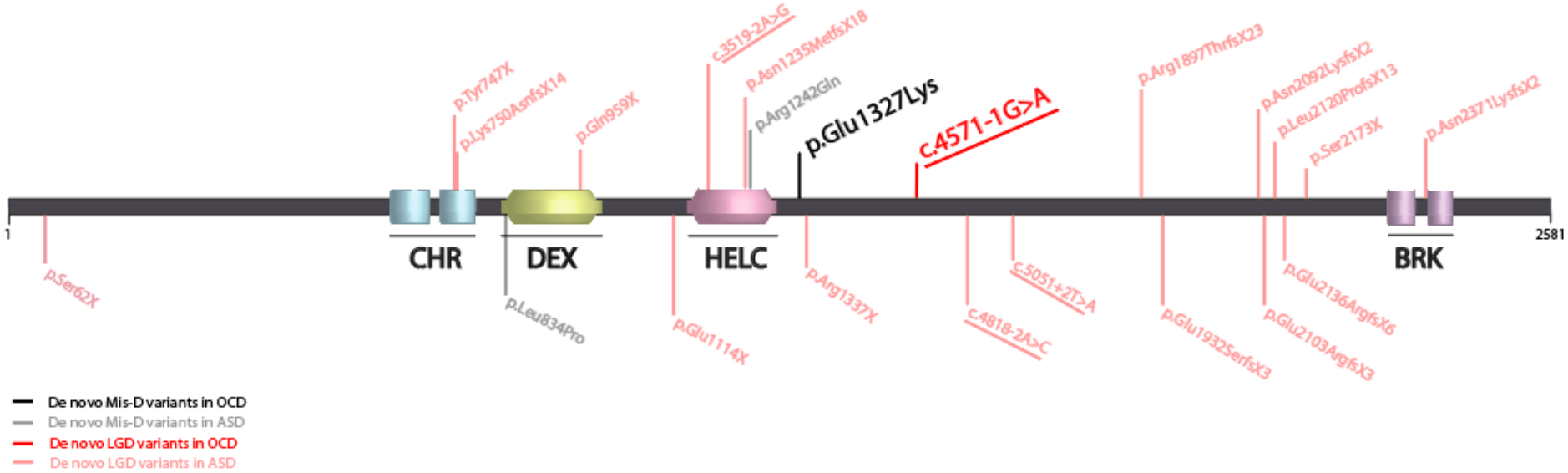
Figure S4 – Sequencing coverage and parental age distributions.



Distribution boxplots of values for (A) mean target coverage, (B) median target coverage, (C) number of base pairs in all “callable” regions, (D) number of base pairs in coding “callable” regions, (E) paternal age, and (F)

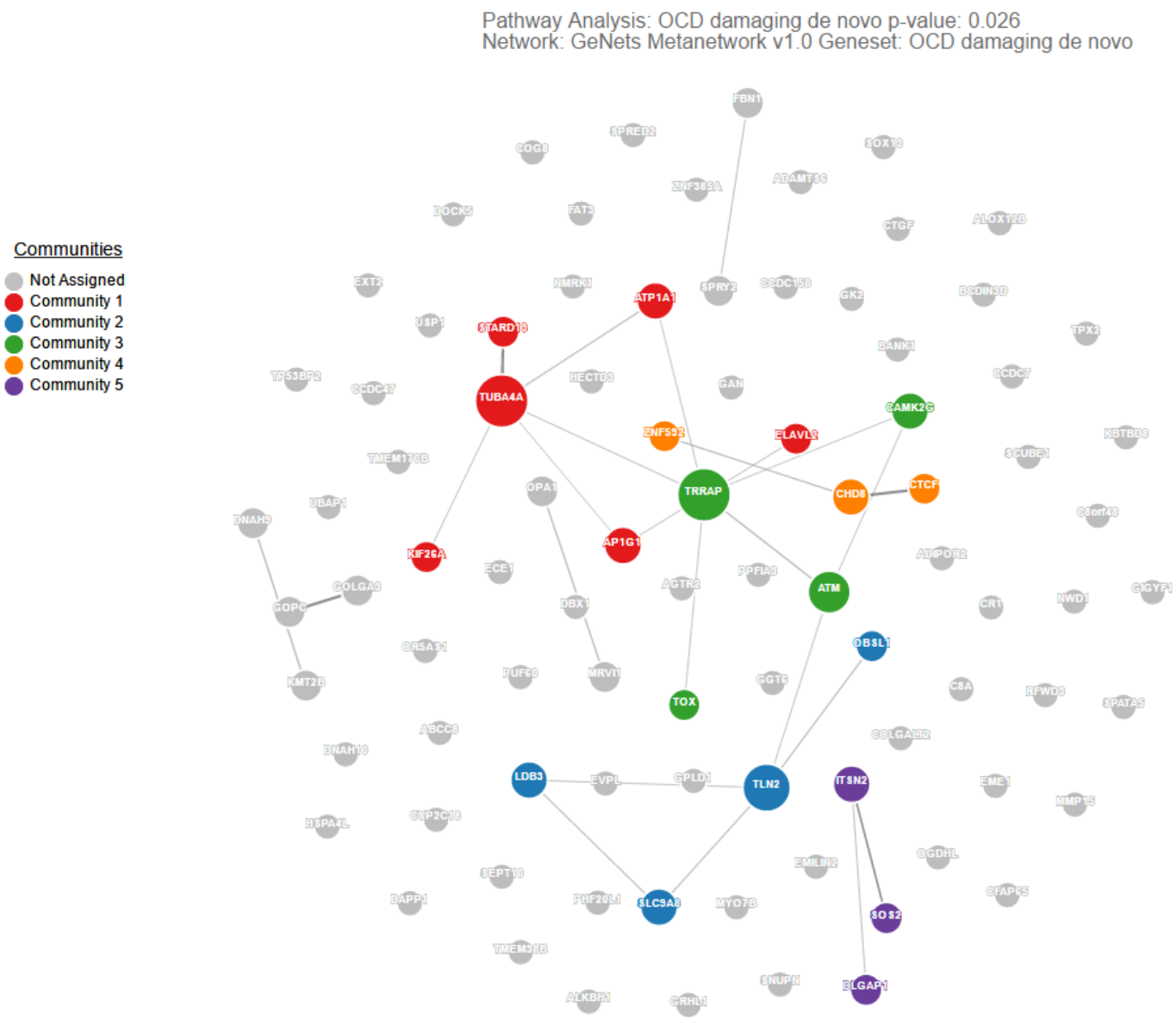
maternal age for both OCD and control cohorts. For each cohort, the box extends from the first through third quartiles, and the horizontal line is at the second quartile (median) of the data. Whiskers extend to the largest non-outliers, and outlier data points are plotted individually. For each comparison, a p-value was calculated using a two-sided Wilcoxon rank sum test with continuity correction. Panels A-D show increased opportunity for variant calling in the OCD cohort, necessitating the use of de novo mutation rate comparisons within the callable exome, as explained in the main text and methods. Panels E-F show no significant difference in parental ages between case and control cohorts. Also see Table S1.

Figure S5 – CHD8 variants in OCD and ASD.



Two de novo likely gene disrupting (LGD, red) and damaging missense (Mis-D, black) variants identified in CHD8 among OCD probands are indicated. ASD-associated de novo LGD and Mis-D mutations reported in the Simons Foundation Autism Research Initiative (SFARI) database (accessed April 12, 2017) are also shown in muted colors. Only variants with identifiable allele or residue sequence positions in the SFARI database were included in the above protein diagram, and splice site variants across cohorts are indicated by the respective allele change and underlined. Annotated protein domains predicted with confidence by the Simple Modular Architecture Research Tool (SMART) are shown as follows: CHR, chromatin organization modifier domain (blue), DEX, DEAD-like helicases superfamily (yellow), HELC, helicase superfamily c-terminal domain (pink), and BRK, domain of unknown function associated with CHROMO domain helicases (purple).

Figure S6 – GeNets network analysis without candidates.



Using the GeNets algorithm (https://apps.broadinstitute.org/genets), we mapped all 89 genes harboring de novo damaging mutations in OCD (excluding two genes, *TTN* and *CACNA1E*, found to harbor de novo damaging variants in control subjects) onto the GeNets Metanetwork v1.0 to determine whether they are functionally connected. The density of the mapped network (density = number of edges / number of possible edges) was greater than 95% of randomly sampled gene sets, indicating that the network is significantly more connected than random (p=0.026). In the network, node (gene) size is proportional to the number of connections. The color is assigned by community, defined as a gene set that is more connected to one another

than to another group of genes. Results from this analysis are available in interactive form here:

https://www.broadinstitute.org/genets#/visualize/58d9425ea4e00291af652379

Supplemental Tables

Table S1 – Phenotype, exome sequencing metrics, and principal components analysis.

(see "TableS1.xlsx")

First tab contains individual-level sample information (columns A-K), including family ID, individual ID, phenotype, cohort, collection site, gender, capture platform, size of "callable exome", and parental age (years) at birth, where available. Column L lists reasons for any sample exclusions by quality control methods; "0" indicates that the sample was not excluded, and was included in subsequent analyses. Columns M-AH list individual sample sequencing metrics generated using PicardTools, and GATK DepthOfCoverage tools. Columns AI-AS list individual sample sequencing metrics generated using PLINK/SEQ (i-stats; https://psychgen.u.hpc.mssm.edu/plinkseq/stats.shtml). Columns B, M-AS were included in Principal Components Analysis (PCA). Third tab contains cohort-level metrics calculated using samples passing quality control. ±95% confidence intervals are given, when applicable. Fourth tab contains coordinates generated for each sample for the top 10 principal components following PCA. The code used to generate this data is included in Supplementary Methods. Using these coordinates, we removed trios with family members falling more than three standard deviations from the mean in any of the first five principal components; this information is contained in the fifth tab.

Table S2 – Annotated de novo variants in OCD and controls.

(see "TableS2.xlsx")

Detailed information on all high confidence de novo variants in cases and controls. These variants were annotated using Annovar, based on RefSeq hg19 gene definitions. Column

descriptions are provided in a separate tab of this file. A third tab provides the number of each de novo variant type per sample.

Table S3 – Gene-level de novo mutation rates, variant counts, and TADA-Denovo results.

(see "TableS3.xlsx")

First tab contains de novo mutation rates used to perform subsequent maximum likelihood estimation (MLE) and TADA-Denovo analyses. The following mutation rates are listed for each gene: overall, likely gene disrupting (lgd), predicted damaging missense (misD), and all damaging (lgd + misD). These mutation rates were previously published (Ware et al., 2015) from unaffected parent-child trios. The code used to generate the mutation rate table is provided in Supplementary Methods. Second tab contains the input file for the TADA-Denovo algorithm. Gene-level expected mutation rates for LGD ("mut.cls1" column) and Mis-D variants ("mut.cls2" column) are listed, along with their respective observed mutation counts in our OCD data ("dn.cls1" and "dn.cls2", respectively). Code for running TADA-Denovo is given in Supplementary Methods. Third tab contains the final output results from TADA-Denovo code provided in Supplementary Methods. Genes harboring more than one damaging de novo (LGD or Mis-D) variant in OCD probands are highlighted in yellow (*SCUBE1, CHD8, TTN*). Two of these genes (*SCUBE1* and *CHD8*) exceeded thresholds for being considered a probable (qval < 0.3) or high confidence (qval < 0.1) risk gene.

Table S4 – DNENRICH gene lists and results.

(see "TableS4.xlsx")

See Supplementary Methods for details of DNENRICH analysis. First tab contains input for DNENRICH analysis. Each row represents a de novo damaging mutation in an OCD proband. Second tab contains the input gene lists to determine enrichment for our OCD damaging de novo mutations. Third tab contains final results output from DNENRICH. Significantly enriched gene sets are highlighted.

Table S5 – GeNets network connectivity analysis results.

(see "TableS5.xlsx")

Complete results from GeNets network analysis of de novo damaging variants found in OCD probands. First tab contains summary statistics of the resulting network, considered both with and without nearby predicted candidate genes. Second tab contains the input gene list and the candidate genes predicted by the network analysis. Third tab groups genes (without predicted candidates) into nearby "communities" that are more connected with each other than their neighbors. Fourth tab contains network edges without the predicted candidates. Fifth tab contains gene community groupings, including predicted candidates. Sixth tab contains network edges including predicted candidates. See Methods for further details of this analysis.

TableS6 – MetaCore and Ingenuity Pathway Analysis (IPA) gene enrichment analysis results.

(see "TableS6.xlsx")

Complete results from Metacore (first tab) and IPA (second tab) gene enrichment analyses, with p-values calculated by each analysis algorithm. See Methods for details of these analyses.