Science Advances
AAAS

Supplementary Materials for

De novo mutations identified by whole-genome sequencing implicate chromatin modifications in obsessive-compulsive disorder

Guan Ning Lin*, Weichen Song, Weidi Wang, Pei Wang, Huan Yu, Wenxiang Cai, Xue Jiang, Wu Huang, Wei Qian, Yucan Chen, Miao Chen, Shunying Yu, Tingting Xu, Yumei Jiao, Qiang Liu, Chen Zhang, Zhenghui Yi, Qing Fan, Jue Chen, Zhen Wang*

*Corresponding author. Email: nickgnlin@sjtu.edu.cn (G.N.L.); wangzhen@smhc.org.cn (Z.W.)

Published 12 January 2022, *Sci. Adv.* **8**, eabi6180 (2022)
DOI: 10.1126/sciadv.abi6180

The PDF file includes:

Supplementary Materials and Methods
Figs. S1 to S3
Table S1
Legends for tables S2 to S5

Other Supplementary Material for this manuscript includes the following:

Tables S2 to S5

SUPPLEMENTARY MATERIALS AND METHODS

Studied subjects

All participants' informed consent was obtained, as approved by Institutional Review Boards of Shanghai Mental Health Center. We collected 53 unrelated parent-offspring families (34 male, 19 female), based on the availability of genomic DNA from whole blood and completeness of phenotype information. The parent-offspring families are composed of 52 trios (each family with one OCD-affected offspring, 156 samples in total), and one family with two OCD-affected offspring (four samples in total). Families were all recruited at Shanghai Mental Health Center in Shanghai. Detailed information for each participant was provided in Supplementary Table 1. One of the trios was excluded due to DNA sample contamination. All the OCD affected patients met the following conditions: (a) the patients were diagnosed as having OCD according to Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria; (b) between the age of 18–65 years; (c) Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) total score cut-off of 16. The patients were excluded if they (a) included DSM-IV criteria for other disorders other than OCD; (b) had moderate to severe suicidal ideation; (c) were pregnant or lactating females. And all parents with any DSM-IV Axis I psychiatric diagnosis were excluded. The above final diagnostic status was approved by the experienced psychiatrists.

Whole-genome sequencing (WGS) and sample filtering

Genomic DNA extracted from whole blood- or lymphoblast-derived cell lines (LCLs) were assessed for quality by PicoGreen and gel electrophoresis and then sequenced by Novogene (NovoGene Biosciences Inc. Beijing, China). DNA quantity was measured by Qubit 3.0, and at least one µg of non-degraded genomic DNA was used for genomic library preparation and WGS. Novogene performed additional quality controls, including DNA quality assessment, sex check, and comparison of samples with results from 96-SNP genotyping assay to avoid sample mix-up. Next, we sequenced all trio samples, which have never been previously sequenced on Illumina HiSeq 4000 sequencing platform (150 bp paired-end reads). The average depth of coverage of our WGS data was 29.9x, with an average 99.87% median alignment rate. After one trio was excluded for quality control, the remaining 53 trios were submitted for subsequent analyses.

Estimation of relative mutation rates

The genome-wide per base pair mutation rate for each patient was calculated as a total number of *de novo* mutations divided by the total number of base pairs with more than 20x coverage in the genome. We estimated average mutation rates in a 95% confidence interval for all 53 OCD patients by "*t.test*" function in R. Final estimation was further divided by two to account for bi-allelic mutation occurrence.

As for OCD relative mutation rate of DNMs to controls, we first downloaded ranges of all exonic regions and UTR regions of hg19 from UCSC Genome Browser and calculated the total numbers of base pairs on coding regions with more than 20x coverage across the whole genome. DNMs from SSC siblings were used as controls. The following calculation of the relative mutation rate among different clinical phenotypes was similar to the above. We collected DNMs affected ASD, SCZ, DD, and ID patients and these corresponding phenotypical information from PsyMuKB. Then ORs were applied to measure the relative rate of DNMs in these other psychiatric disorders compared to our OCD patients.

Definition of subtypes of coding mutations

According to Annovar annotations, we first partitioned all coding mutations: we defined frameshift indel, start site/stop site/splice site mutations as Loss of Function (LoF) mutations, non-synonymous SNV as missense mutations. We further defined a subset of missense as severe mutations (mis3): missense mutations with CADD-phred score ≥ 20, PolyPhen prediction is “D”, and SIFT prediction is “D”. All LoF and mis3 mutations were combined as damaging mutations.

Comparison of mutation severity

We utilized the Combined Annotation Dependent Depletion online tool to calculate the CADD-phred scores representing the predicted severity of input mutations. We compared CADD-phred scores between our OCD patients and SSC sibling from Ann et al. for both exonic mutations and genomic mutations. Since CADD-phred corresponds to the percentile rank of mutation severity and does not follow a normal distribution, we applied the cumulative distribution curve to visualize the distribution of CADD-phred score and Wilcoxon rank-sum test to compare the predicted severity.

Comparison of GC content and gene length

To evaluate the influence of GC bias in different genome regions, we compared the GC content and gene length between DNMs in our OCD data and SSC controls. Data for GC content and gene length were collected from GenBank. *P*-values were calculated by a two-sided Wilcoxon rank-sum test.

Weighted gene co-expression network analysis (WGCNA)

BrainSpan RNA-sequencing data was applied to the following co-expression network analysis. We split the expression profiles into two different period sets by prenatal and postnatal samples, including all brain regions. For each subset, only genes with RPKM > 0 in at least half of the samples and coefficient of variance > 0.3 were retained for WGCNA analysis. The remaining data were log10-transformed. The soft threshold was set at 12 for the prenatal subset and 16 for a postnatal subset. The minimum module size was set at 20. We constructed the signed networks by blockwiseModules function in the WGCNA R package based on Pearson correlation coefficient, and partitioned genes into modules with a tree cut height of 0.3.

After the detection of co-expression modules in both subsets, we tested whether genes carrying OCD coding mutations were enriched in some of the modules by Fisher's exact test. The enriched modules were then used for Gene Ontology enrichment and network analysis. For both enriched modules, we calculated the Topology Overlap Measure (TOM) matrix and gene-module correlation (kME) for all genes to build a co-expression network in Cytoscape.

Analysis of co-expression patterns of *SETD5*, *KDM3B*, and *ASXL3* in the prefrontal cortex of OCD patients and healthy control

To explore the potential dysregulation between three key genes (SETD5, KDM3B, and ASXL3) and three neurotransmitter systems (glutamate, serotonin, and dopamine), we first manually collected three gene lists of these neurotransmitter systems (Supplementary Table 5). Then, we calculated Pearson Correlation Coefficients (PCC) between the key genes and all other genes in prefrontal cortex expression data from Jaffe et al.. The calculation was applied separately for OCD patients and controls. The difference of PCC between OCD patients and controls, $|\Delta\text{Co-exp}|=|\text{PCC}_{\text{OCD}}\text{-PCC}_{\text{control}}|$, was then calculated to measure dysregulation of target genes by corresponding key genes. We used the Wilcoxon rank-sum test to check if overall co-expression between the chromatin modifiers and neurotransmitters was significant between OCD patients and controls. To see whether any single neurotransmitter gene was significantly dysregulated, we compared its |ΔCo-exp| to the whole distribution of |ΔCo-exp| of corresponding key genes and generated a nominal *P*-value.

Plasmids

The SETD5-FLAG and ASXL3-FLAG expression plasmids were purchased from Youze Biotechnology Co, Ltd. (Hunan, China), and the SETD5 R77C and ASXL3 F1460Lfs*5 Mutated plasmids were introduced using the Q5® Site-Directed

Mutagenesis Kit (New England BioLabs). All plasmids were confirmed via Sanger sequencing.

Western blot analysis and antibody

HEK293T cells were lysed in 1× lysis buffer on ice for 30 min and centrifuged at 12,000 rpm for 15 min. Cell lysates were separated by electrophoresis on 4–20% SDS-PAGE gels, then transferred onto nitrocellulose membranes. The membranes were blocked with 5% milk in TBST for one hour and incubated overnight at 4 °C with the corresponding primary antibody: rabbit anti-FLAG and anti-β-Tubulin (Sigma); rabbit anti-H3K9me1(Abcam); rabbit anti-H2AK119ub(Cell Signaling). The process was followed by incubation with the HRP-labeled Goat Anti-Rabbit IgG (Beyotime) for one hour at room temperature 20–25 °C. Quantitation of immunoblots was performed via densitometric analysis using the ImageJ software.

Comparison of OCD and TD *de novo* mutation genes

We downloaded publicly available DNM data of OCD and TD from PsyMuKB database. A total of two OCD studies and three TD studies were included. All genes with at least one missense or LoF mutation were recorded. We combined our non-synonymous mutation genes with published OCD DNM genes and compared them with all previous TD DNM genes.

To analyze the overlap of TD and OCD genes, we performed hypergeometric tests with a background gene list as all protein-coding genes. *P*-values were calculated for two-tailed tests. To analyze the functional characteristics of TD and OCD genes, we first obtained a list of 15 gene sets, mainly about the central nervous system, from psyMuKB. We tested enrichment TD/OCD genes in all these gene sets by Fisher's exact test. Then, we performed a Pearson correlation analysis on -log10 (OR) of genes from two diseases. To analyze the spatiotemporal- and cell-type-specific expression patterns of TD and OCD genes, We applied EWCE R package, a bootstrap enrichment tool based on gene-cell type specificity matrix, to conduct enrichment analysis on two datasets: (1) BrainSpan for period and region analysis; (2) Dronc data for adult brain cell-type analysis. Specificity was defined in the corresponding paper. We first calculated the specificity matrix of two expression datasets using the "generate.celltype.data" function. Next, the enrichment of all TD/OCD genes was tested by "bootstrap.enrichment.test" function. We defined the background gene list as all genes annotated in the tested expression dataset.

SUPPLEMENTARY FIGURES

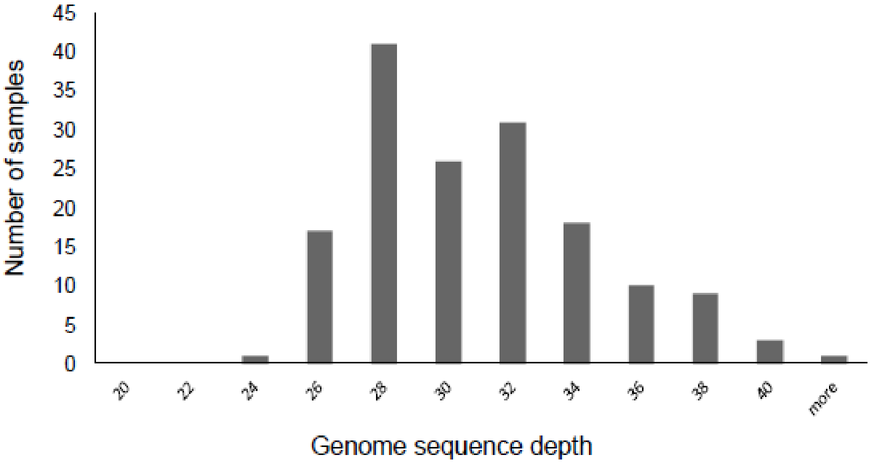

Figure S1. Sequence depth across the whole genome in the samples

The distribution shows the whole genome sequence depth of the 157 samples.

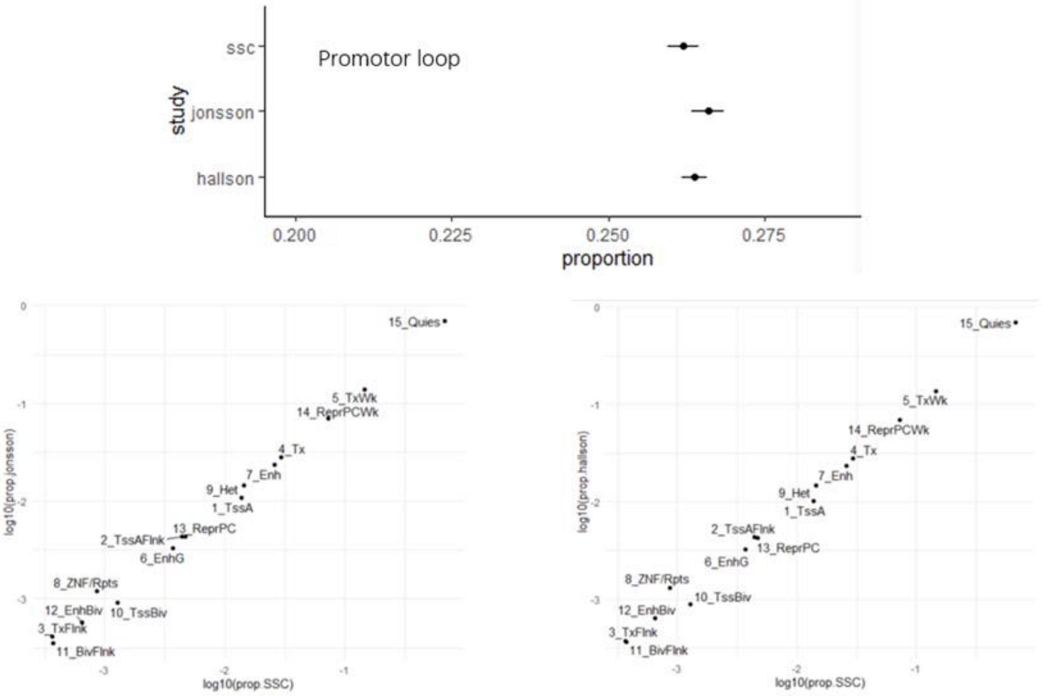

Figure S2. Comparison of mutation distribution among three control datasets.

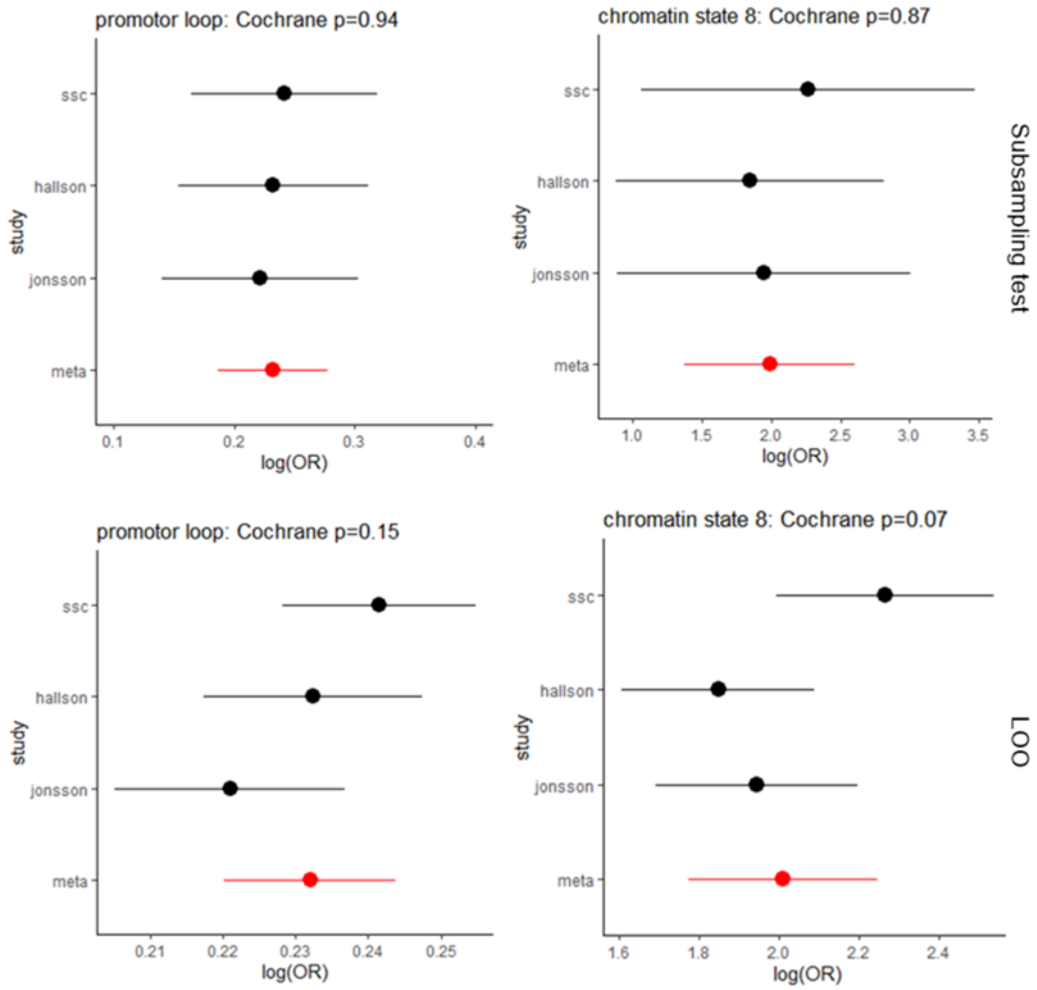

Figure S3. Sensitivity test for cross-study enrichment analysis.

SUPPLEMENTARY TABLES

Table S1. Demographic and clinical information of participants

		Proband	Parents
N		53	104
Gender	Male	34	52
	Female	19	52
Age (years, Mean ± SD)		26.13 ± 6.98	53.66 ± 8.18
Education (years, Mean ± SD)		13.75 ± 2.74	9.58 ± 4.37
Total duration (years, Mean ± SD)		7.54 ± 6.50	n.a.
Y-BOCS (Mean ± SD)		23.38 ± 8.18[a]	n.a.
BDI (Mean ± SD)		16.04 ± 10.93[a]	n.a.
BAI (Mean ± SD)		11.11 ± 9.40[a]	n.a.

Table S2. All OCD DNMs identified in this study.

This table is provided as a separate spreadsheet.

Table S3. Summary of high confidence structural variants discovered by SV2

This table is provided as a separate spreadsheet.

Table S4. GO function enrichment of module M16 of prenatal period co-expression network and M11 of postnatal period co-expression network.

This table is provided as a separate spreadsheet.

Table S5. Correlation results between expressions of key chromatin modifiers and neurotransmitter system genes in OCD cases and controls data.

This table is provided as a separate spreadsheet.