Supplementary Methods

Human Subjects

Some members within the pedigrees were also diagnosed with OCD and/or ADHD (including ADHD Combined type, ADHD Predominantly Inattentive type, or ADHD Predominantly Hyperactive-Impulsive type), based on the DSM-IV-TR, as part of the TIC Genetics study [1]. The OCD and ADHD diagnosis were only used in the phenotype analysis and were not included in the genetic analysis.

Whole Exome Sequencing, Variant Calling and Annotation

Variant calling was performed using the Genome Analysis Toolkit (GATK) following the best practice pipeline [2]. Briefly, paired-end sequencing reads from each individual were aligned to UCSC hg19 genome assembly using bwa (version 0.7.12) [3]. The alignment files (in BAM format) were sorted and indexed using samtools (version 0.1.19) [4]. The BAM files were then subjected to indel realignment (GATK IndelRealigner version 3.2 or 3.3) [2], mark duplicate reads (samtools markdup), and base quality score recalibration (GATK baseRecalibrator). Picard Tools CollectAlignmentSummaryMetrics and CollectWgsMetrics were used to extract sequencing summary statistics.

For variant calling, GATK haplotypecaller (GATK version 3.3) was performed on each processed BAM file. The output gVCF files were combined into one single gVCF file using GATK combineGVCF. GATK GenotypeGVCF was used for joint genotype calling, followed by VariantRecalibrator and ApplyRecalibration for variant recalibration. After recalibration, sites with a “PASS” flag were selected for downstream analyses. To ensure consistency across samples, the exome target regions were defined based on the SeqCap EZ Exome V2 kit in all samples, which covers approximately 36 megabases of the human genome.

ANNOVAR [5] was used to annotate variants to obtain information, including allele frequency (AF) in the 1000 Genomes project (1KGP) [6] and Exome Aggregation Consortium (ExAC) [7], PolyPhen-2 [8] and SIFT [9] damaging prediction scores, protein domain information, etc. ANNOVAR was run using the following command:

```
table_annovar.pl input.vcf annovar/humandb -out out.vcf -
buildver hg19 -otherinfo -protocol
refGene,cytoBand,exac03nonpsych,dbnsfp31a_interpro,gnomad211_exo
me -operation g -nastring "." -vcfinput
```

Candidate Gene Prioritization

pVAAST (pedigree Variant Annotation, Analysis and Search Tool) was used to identify candidate genes in each pedigree [10, 11]. pVAAST scores each gene with a likelihood model considering several types of variant information in each gene, including the segregation pattern, the predicted functional impact, and the AF in general populations. Before the pVAAST run, variants were filtered on their prevalence in the general population based on ANNOVAR annotation. Variant sites with the 1KGP AF < 10% and ExAC_all AF < 5% were selected for the pVAAST analysis [11, 12]. pVAAST was run under dominant mode of inheritance for all pedigrees and recessive mode of inheritance for some pedigrees if the recessive mode of inheritance cannot be excluded (Table 1). pVAAST was run following user guidelines for multiplex families in the following steps:

Step1: VAAST converter

```
perl vaast_converter --build hg19 TS_multiplex.vcf --path
vaast_converter_output/
```

Step 2: VAT (Variant Annotation Tool, one individual is shown here as an example, this step was performed for all individuals within a family)

```
VAT -f RefSeq_hg19.p10_VAAST.gff3 -a vaast_hsap_chrs_hg19.fa 4001.gvf > 4001.vat.gvf --sex male
```

Step 3: VST (Variant Selection Tool, individuals within each family were merged into one file)

```
VST -o 'U(0..$index)' -b hg19 *.vat.gvf > family1.cdr
```

Step 4: pVAAST

```
VAAST -m lrt -p 32 --indel --enable_splice_sites y -gw 1e6 -r 0.05 -pv_control family1_dominant.ctl -o family1_dominant RefSeq_hg19.p10_VAAST.gff3 control.cdr
```

Parameters used within the control file `family1.ctl` were as follows:

```
unknown_representatives: yes
inheritance_model: [dominant|recessive]
informative_site_selection: 3
simulate_genotyping_error: yes
genotyping_error_rate: 1.00E-04
penetrance_lower_bound: 0.6
penetrance_upper_bound: 1
max_prevalence_filter: 0.01
lod_score_filter: yes
clrt_score_filter: yes
nocall_filter: yes
nocall_filter_cutoff: 2
inheritance_error_filter: no
```

The output of pVAAST were parsed to csv files by a custom Python script for downstream analyses. Candidate genes were removed if all variants scored by pVAAST had AF > 0.05 in the gnomAD 2.1.1 [13]. AF for variants inside repetitive sequences were curated manually because the variant call is subject to high error rates and the variant position reporting is often not consistent among different databases. The pLI (probability of being loss-of-function intolerant) score and the missense Z score were extracted from gnomAD for each gene [14].

Candidate Gene Annotation and Filtering

Gene expression data were downloaded from three resources: the Gene Tissue Expression project (GTEx) version 8 [15, 16], the BrainSpan Atlas of the Developing Human Brain [17], and the Human Developmental Biology Resource (HDBR) expression resource of prenatal human brain development [18]. To reduce the variation caused by mitochondrial and non-coding genes, TPM (Transcript Per Million) values of coding genes were re-calculated after removing mitochondrial and non-coding genes. A gene is defined as coding if there is a protein sequence in corresponding GENCODE gene models [19] and the gene is not from mitochondria. To reduce the impact of variants in non-coding genes, for GTEx data, we excluded samples where less than 40% of sequenced mRNAs were from coding genes (TPM < 400,000). For coding genes in each sample, $normalized\ TPM = \frac{TPM}{sum\ of\ TPMs} \times 10^6$. For each gene, the median of normalized TPM values in samples of the same tissues were selected and normalized to represent the expression level of the genes in different tissues. TPM values for TD candidate genes were extracted from the three resources and a gene was removed if the max TPM values in brain tissues was less than 5.

Variant Segregation Within Pedigrees

To select genes with segregating variants, the number of affected/unaffected individuals with a candidate variant is counted within each gene. A true positive event is defined as an affected individual with the mutation and a true negative event as an unaffected family member without the mutation, respectively. A false positive event is defined as an unaffected family member with the mutation and a false negative event as an affected individual without the mutation. Unknown individuals are those whose genotype cannot be determined. For each variant, the false rate is calculated as (false positive + false negative) / (total individuals - unknown). Candidate genes that include at least one variant with true positive events ≥ 2 and false rate < 0.3 were kept.

Gene Lists from Previous NDD Studies

Risk genes for several NDDs were collected from previous studies (Table S1). The lists of genes are: TD_multiplex, genes reported in this study; TD_simplex, genes published in the previous TD literature [20-25] and genes with *de novo* mutations from TD simplex families [26]; TD_CNV, genes from a copy number variant (CNV) study of TD [26]; OCD, genes from two GWAS studies [27, 28] and one WES study of OCD [29]; ADHD, "Published Gene" from the ADHDgene database (http://adhd.psych.ac.cn/) [30]; ASD_high, ASD candidate genes annotated as syndromic or with score ≤ 2 in the Simons Foundation Autism Research Initiative (SFARI) database (12-05-2019 release) [31]; ASD_low, other SFARI genes not labelled as ASD_high [31]; OtherNeuro, genes associated with ID, EE, NDD, and SCZ summarized in [32].

Protein–Protein Interaction Network Identification

Three databases were selected to investigate Protein–Protein Interaction (PPI) networks among candidate genes, including STRING [33], ConsensusPathDB [34], and GIANT_v2 [35] (an

updated version of GIANT [36]). These three databases display the best performance for PPI network construction based on a recent benchmark study [37]. For ConsensusPathDB (CPDB), "induced network modules" analysis was performed for genes from the eight gene lists (Table S1) using only high-confidence interactions and no intermediate nodes. For TD_multiplex genes, custom Python scripts were used to obtain all high-confidence interaction genes with them. For STRING, the v11 full human PPI network was downloaded from the STRING website (https://stringdb-static.org/download/protein.links.full.v11.0/9606.protein.links.full.v11.0.txt.gz). Self-interactions of genes were removed, and a cutoff was set for the "combined score" so that the number of interactions among the genes was the same as identified in ConsensusPathDB. For GIANT_v2, the full PPI network with global evidence was downloaded from the website (http://giant-v2.princeton.edu/static//networks/global.dab) and converted to text format with a Python script (https://github.com/FunctionLab/flib/blob/master/dat.py). A cutoff for interaction score was set in the same way as in the STRING analysis. For each TD_multiplex gene, the number of all interacting genes was counted based on interactions in any of the three databases. The number of interacting genes with other multiplex families and the other 7 gene lists were also counted for each TD_multiplex gene. Fisher's exact test were performed to determine whether the number of interactions were enriched for candidate genes in each gene list. The numbers of interacting genes and enrichment p-values were used to prioritize putative causal genes.

Gene Ontology, Pathway, and Protein Complex Enrichment

Enrichment analyses were performed with over-representation analysis provided by ConsensusPathDB [34]. All pathway, Gene Ontology (GO), and protein complex-based enrichment analyses were enabled, with the minimum two genes from the input and p-value cutoff set to 1.

Enrichment p-values was determined by ConsensusPathDB using a hypergeometric test. For each enriched term (pathway, GO, or protein complex), the total number of genes of that group, input genes identified, gene counts in each of the eight gene lists (no overlap, Table 2), and multiplex family counts were determined. Terms were combined if the overlapped input genes were identical and the total count of genes of the smallest term was used for later analysis. A term was selected for further inspection if the total count of genes ≤ 200, TD_multiplex gene ≥ 2, TD_multiplex + TD_simplex + TD_CNV gene ≥ 3, and multiplex family count ≥ 2. Enrichment p-values of the eight gene lists were calculated with Fisher's exact test and the significance level ($3x10^{-5}$) is determined by the Bonferroni correction of the 1,669 selected terms (0.05/1669). The enriched terms were used as evidence to prioritizing causal genes in multiplex families.

Copy Number Variant (CNV) Analysis

Genotypes were called using Illumina GenomeStudio software (V2010.1). There were no significant differences in call rates and heterozygosity between genotyping facilities. All samples included in the analysis passed strict quality control, including expected genotypic identity within the family using PLINK [38], expected genotypic sex based on chromosome X heterozygosity and sex chromosome LRR, genotyping rate >= 97%, and all samples passing quality metrics within the CNV calling algorithms. The CNV detection was performed using the program CNVision (https://sourceforge.net/projects/cnvision) as previously described [39]. CNVision merges CNV calls from three algorithms: PennCNV [40], QuantiSNP [41], and GNOSIS [42].

References

1. Dietrich A, Fernandez TV, King RA, State MW, Tischfield JA, Hoekstra PJ *et al.* The Tourette International Collaborative Genetics (TIC Genetics) study, finding the genes causing Tourette syndrome: objectives and methods. *European child & adolescent psychiatry* 2015; **24**(2)**:** 141-151.

2. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 2011; **43**(5)**:** 491-498.

3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**(14)**:** 1754-1760.

4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 2009; **25**(16)**:** 2078-2079.

5. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols* 2015; **10**(10)**:** 1556-1566.

6. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM *et al.* A global reference for human genetic variation. *Nature* 2015; **526**(7571)**:** 68-74.

7. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536**(7616)**:** 285-291.

8. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* 2013; **Chapter 7:** Unit7 20.

9. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* 2003; **31**(13)**:** 3812-3814.

10. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J *et al.* A probabilistic disease-gene finder for personal genomes. *Genome research* 2011; **21**(9)**:** 1529-1542.

11. Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL *et al.* A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nature biotechnology* 2014; **32**(7)**:** 663-669.

12. Kennedy B, Kronenberg Z, Hu H, Moore B, Flygare S, Reese MG *et al.* Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Current protocols in human genetics* 2014; **81:** 6 14 11-25.

13. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* 2019**:** 531210.

14. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581**(7809)**:** 434-443.

15. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B *et al.* Genetic effects on gene expression across human tissues. *Nature* 2017; **550**(7675)**:** 204-213.

16. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020; **369**(6509)**:** 1318-1330.

17. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A *et al.* Transcriptional landscape of the prenatal human brain. *Nature* 2014; **508**(7495)**:** 199-+.

18. Lindsay SJ, Xu YB, Lisgo SN, Harkin LF, Copp AJ, Gerrelli D *et al.* HDBR Expression: A Unique Resource for Global and Individual Gene Expression Studies during Early Human Brain Development. *Front Neuroanat* 2016; **10**.

19. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome research* 2012; **22**(9)**:** 1760-1774.

20. Pauls DL, Fernandez TV, Mathews CA, State MW, Scharf JM. The Inheritance of Tourette Disorder: A review. *J Obsessive Compuls Relat Disord* 2014; **3**(4)**:** 380-385.

21. Huang AY, Yu D, Davis LK, Sul JH, Tsetsos F, Ramensky V *et al.* Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. *Neuron* 2017; **94**(6)**:** 1101-1111 e1107.

22. Willsey AJ, Fernandez TV, Yu D, King RA, Dietrich A, Xing J *et al.* De Novo Coding Variants Are Strongly Associated with Tourette Disorder. *Neuron* 2017; **94**(3)**:** 486-499 e489.

23. Sun N, Tischfield JA, King RA, Heiman GA. Functional Evaluations of Genes Disrupted in Patients with Tourette's Disorder. *Front Psychiatry* 2016; **7:** 11.

24. Scharf JM, Yu D, Mathews CA, Neale BM, Stewart SE, Fagerness JA *et al.* Genome-wide association study of Tourette's syndrome. *Mol Psychiatry* 2013; **18**(6)**:** 721-728.

25. Sun N, Nasello C, Deng L, Wang N, Zhang Y, Xu Z *et al.* The PNKD gene is associated with Tourette Disorder or Tic disorder in a multiplex family. *Mol Psychiatry* 2018; **23**(6)**:** 1487-1495.

26. Wang S, Mandell JD, Kumar Y, Sun N, Morris MT, Arbelaez J *et al.* De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. *Cell reports* 2018; **24**(13)**:** 3441-3454 e3412.

27. Arnold PD, Askland KD, Barlassina C, Bellodi L, Bienvenu OJ, Black D *et al.* Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol Psychiatr* 2018; **23**(5)**:** 1181-1188.

28. Mattheisen M, Samuels JF, Wang Y, Greenberg BD, Fyer AJ, McCracken JT *et al.* Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Mol Psychiatr* 2015; **20**(3)**:** 337-344.

29. Cappi C, Brentani H, Lima L, Sanders SJ, Zai G, Diniz BJ *et al.* Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways. *Translational psychiatry* 2016; **6**.

30. Zhang LY, Chang SH, Li Z, Zhang KL, Du Y, Ott J *et al.* ADHDgene: a genetic database for attention deficit hyperactivity disorder. *Nucleic acids research* 2012; **40**(D1)**:** D1003-D1009.

31. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular autism* 2013; **4**.

32. Willsey AJ, Morris MT, Wang S, Willsey HR, Sun NW, Teerikorpi N *et al.* The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. *Cell* 2018; **174**(3)**:** 505-520.

33. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* 2017; **45**(D1)**:** D362-D368.

34. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature protocols* 2016; **11**(10)**:** 1889-1907.

35. Wong AK, Krishnan A, Troyanskaya OG. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic acids research* 2018; **46**(W1)**:** W65-W70.

36. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 2015; **47**(6)**:** 569-576.

37. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P *et al.* Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Syst* 2018; **6**(4)**:** 484-495 e485.

38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007; **81**(3)**:** 559-575.

39. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 2015; **87**(6)**:** 1215-1233.

40. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* 2007; **17**(11)**:** 1665-1674.

41. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research* 2007; **35**(6)**:** 2013-2025.

42. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011; **70**(5)**:** 863-885.