Supplementary Materials for

**Identification of *de novo* mutations in the Chinese ASD cohort via whole-exome sequencing unveils brain regions implicated in autism**

Bo Yuan2, #, Mengdi Wang4, #, Xinran Wu3, #, Peipei Cheng1, Ran Zhang2, Shunying Yu1, Jie Zhang4, \*, Yasong Du1, \*, Xiaoqun Wang3, \*, Zilong Qiu 2, \*

\*Co-corresponding authors

# These authors contribute equally to this work.

**This file includes:**

# Supplemental Figure 1-3

**Titles for Table S1-S7**

# Supplemental Materials and methods

**Supplemental references**

图表, 气泡图

描述已自动生成

**Figure S1. Quality Control Results and Enrichment Analyses of Various Functional Types of *de novo* mutations in Our and Published ASD Cohorts**

(A)Sequencing coverage performance. The proportion of the target exome region covered with the indicated numbers (≥ 20x or 40x) of reads was plotted. Data were sorted in the order of the 20x trio rank (trios with the smallest proportion covered with ≥ 20x reads on the left and the largest on the right). Orange dots indicate 20x individual coverage, and sky blue dots indicate 40x individual coverage.

(B) The overlap of ASD risk genes between the Chinese ASD cohort (SMHC) and Japanese ASD cohort (Takata). The dark yellow circle indicates genes with DNMs in the Chinese ASD cohort; the light yellow circle indicates genes with DNM mutations confirmed by Sanger sequencing in the Japanese ASD cohort; the green and magenta circles indicate genes with HIGH-impact DNMs in Chinese and Japanese ASD cohorts, respectively.

(C) The overlap of ASD risk genes between the Chinese ASD cohort (SMHC), the Japanese ASD cohort (Takata) and the SFARI gene list (including Cat S, 1, 2). The green and yellow circles indicate genes with DNMs in Chinese and Japanese ASD cohorts.

(D) The overlap of ASD risk genes between the Chinese ASD cohort (SMHC) with four other studies (samples size from 200-400 ASD trios). All genes with DNM were analyzed.

卡通人物

低可信度描述已自动生成

**Figure S2 Overview of transcriptional heterogeneity among cells from diverse brain regions.**

(A) Visualization of major cell type by UMAP. (B) Heatmap illustrating the differently expressed genes among major cell types. Well-known marker genes for each cell type were labeled on the left side of the heatmap (high, yellow; low, purple). (C) UAMP showing the expression of marker genes. *NEUROD6* and *SLC17A7* for excitatory neurons; *PAX6* and *SOX2* for NPCs; *SLC17A6* for neurons from pons; *RELN* for Cajal-Retzius cells; *GAD1* for GABAergic interneuron; *AQP4* for astrocytes; *PDGFRA* for OPC; *CX3CR1* for microglia; *HBB* for blood cells and *FN1* for endothelial.

卡通人物

低可信度描述已自动生成

**Figure S3 Regional distribution of cells from ASD risk genes enriched subtypes.**

1. Barplot displaying the regional composition of 4 ASD risk genes enriched subtypes. (B-D) Histogram showing the proportion of
2. cells for subtype NPC-4 (B), Ex-1 (C), In-2 (D) and CR (E) from each brain region.

**Table S1. All gene with *de novo* mutations found in the Chinese ASD cohorts.**

**Table S2. TADA-Denovo analysis for genes with *de novo* mutations.**

**Table S3. List of chromosomal segments with *de novo* copy number variations which including genes presented in the SFARI gene list.**

**Table S4. Demographic information of ABIDE-I participants.**

**Table S5. Seed-based functional connectivity (left hemisphere).**

**Table S6. Seed-based functional connectivity (right hemisphere).**

**Table S7. List of abbreviation for brain regions.**

# Supplemental Materials and methods

## CNV Detection

CNVs were called with GATK PreprocessIntervals, CollectReadCounts, AnnotateIntervals, FilterIntervals, DetermineGermlineContigPloidy, GermlineCNVCaller, IntervalListTools and PostprocessGermlineCNVCalls based on a cohort mode pipeline detecting germline copy number variants. All CNVs were annotated to GRCh38/hg38 by VEP and AnnotSV(1).

## Real-time Quantitative PCR Validation

To confirm *de novo* CNVs detected by WES, we performed quantitative PCR (qPCR) using DNA from probands, their parents and controls. We used the comparative CT method (delta-delta CT method) for relative quantification, with data normalized against an endogenous control sequence (glyceraldehyde 3-phosphate dehydrogenase, GAPDH) with two normal copies. Genomic DNA was amplified using SYBR Green (Thermo Fischer Scientific) and qPCR was performed on a StepOnePlus™ Real-Time PCR System (Applied Biosystems). Data were analyzed using R version 3.6.3 and pcr package(2). The primers used for qPCR are *TBR1* Forward GGGATGACGAATCAGTCAGA, *TBR1* Reverse TGGCTGGACTGAGAGAGGAG, *RAI1* Forward TCTCCAGGCCAGAAAGAAAA, *RAI1* Reverse TGAATGCCTGGAATGAATGA, *SHANK3* Forward TGCCTCACGGAGTTTTCTCT, *SHANK3* Reverse ATGCGGGACTTTATGCAAAC, *MECP2* Forward CACGGAAGCTTAAGCAAAGG, *MECP2* Reverse TCAAGCACACCTGGTCTCAG, *GAPDH* Forward ATCAAGAAGGTGGTGAAGCA, *GAPDH* Reverse TGACAAAGTGGTCGTTGAGG.

## Statistical Analyses

We statistically evaluated the observed number of dDNMs in each gene using TADA-Denovo(3). We included IMPACT HIGH and Possible-damaging missense mutations in the TADA-Denovo analysis. Parameters for this analysis were determined by following the TADA manual. Per-gene mutation rates for LOF and Possible-damaging missense DNMs were obtained from mirDNMR based on sequence context(4).

**Single cell RNA-seq data processing**

To infer the gene expression pattern of ASD risk genes in human brain, two publicly available scRNA-seq dataset were downloaded(5, 6) and assembled. The assembled gene-by-cell count matrix was then imported into R package Seurat(7) for downstream analysis. Briefly, the assembled count matrix was first loaded into function CreateSeuratObject to create a Seurat object followed by a Log-normalization process using function NormalizeData. Variable genes were then identified with function FindVariableFeatures. Next, the principal component analysis was performed by function RunPCA. Unbiased clustering was done with function FindNeighbors and FindClusters followed by Uniform manifold approximation and projection (UMAP) dimension reduction analysis with function RunUMAP.

**Identification of differentially expressed genes**

Differentially expressed genes were computed with Seurat function FindAllMarkers by setting parameter only.pos =TRUE. Genes with adjusted P-values <0.05 were selected as differentially expressed genes.

**Enrichment analysis of ASD risk genes**

To explore the expression pattern of ASD risk genes, gene-set enrichment analysis was performed using R package AUCell(8) for the 55 High-impact genes and the 165 Moderate-impact genes on the assembled scRNA-seq dataset independently. Z-score of the gene-set enrichment score was then computed with R function scale and visualized by UMAP.

**Ranking of cells by enrichment score of ASD risk genes**

To study the expression pattern of ASD risk genes among different cell types and brain regions, we computed averaged gene-set enrichment score for each cell type and brain region and then ranked the cell types and brain regions in descending order. The averaged gene-set enrichment score and rankings were visualized with Heatmap.

**Proportion of highly expressed genes among subtypes and brain regions**

To get the proportion of highly expressed genes in each of the two gene-set (the 55 High-impact genes and the 165 Moderate-impact genes), for each gene-set, we counted the number of genes that have expression in at least 25% of the cells from each subtype or brain region. The proportion of highly expressed genes was then computed by dividing the number of highly expressed genes in each subtype or brain region by the total number of genes in each gene-set.

**Brain Imaging Data Acquisition**

Imaging data were obtained from Autism Brain Imaging Data Exchange (ABIDE-I, <http://fcon_1000.projects.nitrc.org/indi/abide/>)(9), a publicly available database released in 2012. 1112 subjects (539 ASDs, 573 age-matched healthy controls) from 16 international imaging sites underwent anatomical and resting-state functional MRI scans and a series of out-of-scanner phenotypic information was also collected. The scanning parameters of each site can be found online.

**Preprocessing of Imaging data**

All the preprocessed BOLD time series data and voxel-based morphometry (VBM) data used in this study was available on Preprocessed Connectomes Project (PCP; <http://preprocessed-connectomes-project.org/>). In this paper, resting-fMRI data and T1-weight MRI data from ABIDE were preprocessed by the Data Processing Assistant for Resting-State fMRI (DPARSF, <http://rfmri.org/dpabi>)(10) pipeline, one of the four preprocessing pipelines adopted by the ABIDE Preprocessed repository; further information can be found in the literature of Craddock et al.(11).

**Voxel-Based Morphometry**

For structural MRI, preprocessing including segment, spatial normalization, modulation and spatial smoothness. The raw T1w images of each subjects were segmented into gray matter (GM), white matter (WM), and cerebral-spinal fluid (CSF) tissue classes by calling the VBM8 toolbox (http://dbm.neuro.uni-jena.de/vbm/)(12) and transformed from individual native space into the Montreal Neurological Institute (MNI) space using the DARTEL normalization method(13). Then, normalized gray matter data were modulated to obtain the grey matter volume (GMV) measure. Finally, GMV data were smoothed with an 8 mm Full-width-Half-Maximum (FWHM) Gaussian kernel. Total brain volume (TBV = GM +WM) was calculated, which would be used as covariable in further analysis. According to the quality control information given in the phenotype file provided by the PCP, all scans rated as low quality (qc\_anat\_rater\_1="fail" or qc\_anat\_rater\_2="fail") were excluded for analysis.

We abstracted GMV of 6 regions of interest (ROI) for the statistical analysis: bilateral BST, PC and PRC. All ROIs were defined by the Desikan parcellation(14), and GMV of each ROI was obtained by taking average of all voxels located in the ROI.

*Seed-based Functional Connectivity*

For resting-state fMRI, preprocessing including removing the first 4 volumes, slice timing correction and head motion correction, regression of covariables, low-frequency filtering, spatial registration, spatial normalization and smoothness(10). We used a rigid body transformation to realign the time series of images for each subject. To reduce the respiratory and cardiac effects, signals from the WM and CSF are regressed out form the BOLD signals as covariables. Since the global signal is considered to reflect respiratory, cardiac and motion information which may led to spuriously correlation(15, 16), we performed the global signal regression to remove these potential confounding. Then, linear and quadratic trends were also regressed out, and a band-pass filtering (0.01–0.1 Hz) was performed. The transformation information acquired from T1w images normalization was used to perform spatial normalization for resting-state images. Finally, we performed a smoothness using a 6 mm FWHM kernel. All the subjects with low quality (qc\_func\_rater\_1="fail" or qc\_func\_rater\_2=="fail") or large head-motion (mean framewise displacement (FD) > 0.2 mm) were excluded.

Seed-based functional connectivity (FC) of the bilateral BST, PC and PRC was calculated. For each participant, we extracted averaged timing series within all ROIs and calculated their temporal coupling with all other brain voxels using Pearson correlation, resulting in 6 FC maps.

**Statistic**

For ROI-level GMV, we examined the differences between the ASD group and the health control group using a linear mixed model (LMM, using MATLAB built-in function *fitlme*), which has been shown to effectively control for site effects in previous large-scale neuroimaging studies. For the GMV of each ROI, we set the model as follows:

*GMV~Diagnosis+Sex+Age+TBV+(1|Site)*

Where the site effect was put into model as a random intercept, while diagnosis, age and sex were used as fixed effects. T-value and p-value of regression coefficient of *Diagnosis* were obtained from LMM to reflect the GMV difference between ASD and HC. For all statistical results, we applied a significant threshold of P < 0.05 (two-tails) and corrected for multiple comparisons using Family-Wise Error (FWE) method (number of comparisons=6).

For seed-based FC, we used a two-sample t-test to detect inter-group difference. Age, sex, site (coded as dummy variables), and head motion (mean FD) were regressed out as covariates before inter-group comparison. The t statistical maps were enhanced using the threshold-free cluster enhancement (TFCE)(17) with nonparametric permutation testing, which has been proven to be able to effectively increase reproducibility and control of false positive rate in voxel-level neuroimaging analysis(18, 19). The TFCE was performed based on the statistical module of the DPARSF toolbox. The number of permutations was set to 5000 and cluster-extent threshold of Z = 2.3. We applied a threshold of P < 0.05 (two-tails) and corrected for multiple comparisons (FWE) based on the results of a permutation test using the TFCE value.

**Additional Analysis**

To confirm that our results were not due to medication use, we excluded all subjects on medication and repeated our analysis. Besides, we repeated our analyses using full-scale intelligence quotient (FIQ) as a covariate to control the potential confounding effects from subject’s intelligence. The Intelligence in the Abide-I data set is mainly measured by Wechsler Instruments (Wechsler Adult Intelligence Scale (WAIS)(20) and Wechsler Intelligence Scale for Children (WISC)(21)), while there were also some other series scales used, refer to [http://fcon\_1000.projects.nitrc.org/indi/abide/](http://fcon_1000.projects.nitrc.org/indi/abide/ABIDE_LEGEND_V1.02.pdf) for more information.

**Supplemental reference**

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

2. Ahmed M, Kim DR. pcr: an R package for quality assessment, analysis and testing of qPCR data. PeerJ. 2018;6:e4473.

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

4. Jiang Y, Li Z, Liu Z, Chen D, Wu W, Du Y, et al. mirDNMR: a gene-centered database of background de novo mutation rates in human. Nucleic Acids Res. 2017;45(D1):D796-D803.

5. Fan X, Fu Y, Zhou X, Sun L, Yang M, Wang M, et al. Single-cell transcriptome analysis reveals cell lineage specification in temporal-spatial patterns in human cortical development. Sci Adv. 2020;6(34):eaaz2978.

6. Fan X, Dong J, Zhong S, Wei Y, Wu Q, Yan L, et al. Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. Cell Res. 2018;28(7):730-45.

7. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902 e21.

8. Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083-6.

9. Di Martino A, Yan C-G, Li Q, Denio E, Castellanos FX, Alaerts K, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry. 2014;19(6):659-67.

10. Yan C, Zang Y. DPARSF: a MATLAB toolbox for" pipeline" data analysis of resting-state fMRI. Frontiers in systems neuroscience. 2010;4:13.

11. Craddock C, Benhajali Y, Chu C, Chouinard F, Evans A, Jakab A, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics. 2013;7.

12. Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005;26(3):839-51.

13. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage. 2007;38(1):95-113.

14. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 2006;31(3):968-80.

15. Chang C, Glover GH. Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. Neuroimage. 2009;47(4):1448-59.

16. Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage. 2014;84:320-41.

17. Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage. 2009;44(1):83-98.

18. Chen X, Lu B, Yan C-G. Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. Human Brain Mapping. 2018;39(1):300-18.

19. Noble S, Scheinost D, Constable RT. Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. Neuroimage. 2020;209:116468.

20. Wechsler D. Wechsler adult intelligence scale. 1955.

21. Wechsler D, Kodama H. Wechsler intelligence scale for children: Psychological corporation New York; 1949.