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

Bo Yuan
Center for Excellence in Brain Science and Intelligence Technology
Mengdi Wang
Institute of Biophysics
Xinran Wu
Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University
Peipei Cheng
Shanghai Mental Health Center
Ran Zhang
Center for Excellence in Brain Science and Intelligence Technology
Ran Zhang
Shanghai Mental Health Center
Shunying Yu
Shanghai Mental Health Center
Yasong Du
Shanghai Mental Health Center
Jie Zhang
Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University
Xiaoqun Wang
Beijing Normal University
Zilong Qiu (zqiu@ion.ac.cn)
Center for Excellence in Brain Science and Intelligence Technology

Research Article

Keywords: Autism spectrum disorders, whole-exome sequencing, single-cell sequencing, brain imaging

Posted Date: May 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-729083/v2

Abstract

Background

Autism spectrum disorder (ASD) is a highly heritable neurodevelopmental disorder characterized by deficits in social interactions and repetitive behaviors. Although hundreds of ASD risk genes, implicated in synaptic formation and transcriptional regulation, have been identified through human genetic studies, the East Asian ASD cohorts are still under-represented in the genome-wide genetic studies.

Methods

Here we performed whole-exome sequencing on 369 ASD trios including probands and unaffected parents of Chinese origin. A joint-calling analytical pipeline based on GATK toolkits was applied to examine *de novo* variants.

Results

We identified numerous *de novo* variants including 55 high-impact variants and 165 moderate-impact variants, as well as *de novo* copy number variations containing known ASD-related genes. Importantly, combining with single-cell sequencing data from the developing human brain, we found that expression of genes with *de novo* mutations were specifically enriched in pre-, post-central gyrus (PRC, PC) and banks of superior temporal (BST) regions in the human brain. By further analyzing the brain imaging data with ASD and healthy controls, we found that the gray volume of the right BST in ASD patients significantly decreased comparing to healthy controls, suggesting the potential structural deficits associated with ASD. Finally, we found that there was decreased in the seed-based functional connectivity (FC) between BST/PC/PRC and sensory areas, insula, as well as frontal lobes in ASD patients.

Limitations

: Due to the limited sample size of the current ASD genetic study, the further statistical analysis including false discover rate (FDR) cannot be performed.

Conclusion

This work indicated that the integrated analysis with genome-wide genetic screening, single-cell sequencing and brain imaging data would reveal novel insights into brain regions contributing to etiology of ASD.

Background

Current prevalence of ASD has approximately increased to 1 in 54 children aged 8 years old in the United States, and males are four times more susceptible for ASD than females(1). The epidemiology survey in China showed that the prevalence of ASD may range to 0.2–0.4%, suggesting that the difference in geography and clinical diagnosis criteria may lead to disparate impacts on studies of ASD etiology(2). Recently, tremendous efforts in ASD genetic studies using whole-exome and whole-genome sequencing have built up high-throughput assessment pipelines for protein-disrupting variants in large ASD cohorts, in which *de novo* single nucleotide variants (SNVs), insertions and deletions (INDELs) and copy number variants (CNVs), as well as rare inherited variants are major contributors for genetic risks of ASD(3–6). Thus it is critical to further classify genetic causes from accumulated ASD genetic studies in consideration of neurobiological evidences. The online SFARI gene database provided an important public resource for ASD risk genes, in which over one thousand of ASD candidate genes were prioritized with genetic and neurobiological evidence (Category S, 1, 2, 3) (7).

Although genomic information of large cohorts consisting of over 40 thousand ASD patients and unaffected parents have been collected, East Asian populations are still underrepresented groups(8). Some genetic studies on Chinese ASD cohorts using targeted multiplex sequencing technology focused on a group of genes associated with neurodevelopmental disorders, which cannot yield comprehensive genome-wide information about ASD risk genes(9, 10). Two available genetic studies on Chinese ASD cohorts using whole-genome sequencing methods included less than 40 trios, limiting the power of genomic sequencing(11, 12).

In this study, we performed whole-exome sequencing analysis in a Chinese cohort including 369 ASD probands with their parents. 150 bp paired-end sequencing short reads were mapped against human reference genome build 38 (GRCh38/hg38). SNVs and INDELs were jointly called across all samples and filtered by GATK Variant Quality Score Recalibration (VQSR) and Convolutional Neural Network (CNN) tools. Together with analysis of single-cell sequencing data from the developing human brain, we found that expression of genes with *de novo* mutations were specifically enriched in pre-, post-central gyrus (PRC, PC) and BST (banks of superior temporal) regions in the human brain.

By further analysis of the brain imaging data with ASD and healthy controls, we found that BST of the right hemisphere in ASD patients significantly decreased gray volume comparing to healthy controls, suggesting the potential structural deficits associated with BST in autistic patients. Finally, after analyzing the seed-based functional connectivity (FC) of these regions, we found the decrease in FC between BST/PC/PRC and sensory areas, insula, as well as frontal lobes in ASD. This work indicated that the in-depth combinatorial analysis of ASD risk genes from genome-wide screening, the single-cell sequencing and brain imaging data would unveil the brain regions implicated in ASD and thus provide an analytical framework illustrating the genetic basis and neurobiological mechanism for ASD.

Methods

Samples and ethics statement

We analyzed a sample set consisting of 369 ASD probands and 706 parents from 353 pedigrees recruited from Department of the Child and Adolescent Psychiatry, Shanghai Mental Health Center. Of the families 15 are multiplex that have two ASD children and 338 are trios. The fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) was used for ASD diagnoses made by trained psychiatrists. We obtained assent from the Institutional Review Board (IRB), Shanghai Mental Health Center of Shanghai Jiao Tong University (FWA number 00003065, IROG number 0002202). Dr. Yi-Feng Xu approved and signed our study with ethical review number 2016–4. Written informed consent was obtained from parents in consideration of the fact that all patients were minors. All participants were screened using the appropriate protocol approved by the IRB.

Whole-exome sequencing

Genomic DNA extracted from blood samples were sequenced at Shanghai Biotechnology Corporation (SBC) and WuXi NextCODE on Illumina HiSeq sequencers using the Agilent SureSelect Human All Exon V5 exome capture kit. Some samples were sequenced at Euler Genomics on Illumina HiSeq sequencers using the IDT xGen Exome Research Panel v1 exome capture kit. 150 bp paired-end sequencing reads were aligned to human genome build 38 (GRCh38/hg38) using the Burrows-Wheeler Aligner (BWA)(47), Picard tools MarkIlluminaAdapters, SamToFastq and MergeBamAlignment (http://broadinstitute.github.io/picard/) aggregated into a BAM file. Per-individual coverages of the target regions calculated by Qualimap 2 are shown in Figure S1A(48). Picard tools MarkDuplicates, SortSam and SetNmMdAndUqTags were used to mark duplicates, sort bam files by chromosome coordinates, and add essential tags. Single nucleotide variants (SNVs) and insertions/deletions (INDELs) were jointly called across all samples using the Genome Analysis Toolkit (GATK) HaplotypeCaller 4.2.0.0(49). Variant call accuracy was estimated using the GATK Variant Quality Score Recalibration (VQSR) approach and GATK CNN (Convolutional Neural Network) Variant Filter. The VCF file (format v4.2) was produced by the Broad sequencing and calling pipeline with GATK version 4.2.0.0.

We included variant calls with the PASS flag in the downstream analyses. Variants (SNVs and INDELs) were annotated based on the hg38 database using VEP(50). By following the definition of calculated variant consequences by VEP, we classified variants into those having HIGH, MODERATE, LOW and MODIFIER impacts.

Population stratification using genotyping information of frequent exonic SNPs

To define a set of common exonic SNPs, we Initially chosen variants that are: 1) on the InfiniumExome-24v1-1_A1 genotyping array, 2) with MAF > 0.05 in East Asian (EAS) population of ExAC(24) annotated by

VEP, and 3) biallelic in EAS. After combining the information of these SNPs in our cohort (OWN) with the data of the same SNPs in African (AFR), American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) individuals in the 1000 Genomes Project(51), we performed further filtering and linkage disequilibrium (LD)-based pruning using PLINK v1.9(52) with the following options and parameters: --maf (minor allele frequency) 0.05, --mind (maximum per-person missing) 0.2, --geno (maximum per-SNP missing) 0.2, --hwe (Hardy-Weinberg disequilibrium p-value) 1×10^{-10} and --indep (SNP window size, number of SNPs to shift and variance inflation factor threshold) 50 5 2. By using the data of 1064 SNPs that passed the filters described above, we performed multidimensional scaling with PLINK.

Identification of DNMs

We filtered out variant calls when one or more variant alleles were observed in unaffected parents of our cohort. By using the information of the remaining variant calls, we extracted candidates for DNMs using GATK PossibleDeNovo, TrioDenovo**(53)**, DeNovoGear**(54)**. Candidate DNMs called by these three tools at the same time were then stratified into SNVs and INDELs. We selected 161 DNM calls by prioritizing HIGH impact DNMs and 13 MODERATE impact DNMs. Possible-damaging missense DNMs were defined as the 14 variants predicted to be damaging by at least two of the seven prediction algorithms: SIFT**(14)**, PolyPhen-2 HumVar**(15)**, PolyPhen-2 HumDiv**(15)**, LRT**(16)**, MutationTaster**(17)**, Mutation Assessor**(18)** and PROVEAN**(19)** annotated by dbNSFP4.0a**(20, 21)**.

Results

Identification of *de novo* variant in ASD probands

We analyzed a ASD cohort consisting of 369 ASD probands and 706 parents from 353 pedigrees recruited from Department of the Child and Adolescent Psychiatry, Shanghai Mental Health Center. Among the cohort, there are 15 multiplex family containing two ASD children and 338 simplex family which have one ASD child. The fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) were used for ASD diagnoses by trained psychiatrists.

Proportion of the targeted exome regions covered with ≥ 20x or 40x of reads indicates sufficient coverage (Fig. S1A). After performing the multidimensional scaling of the genotyping data, common exonic SNPs was identified with the PLINK toolkit (a whole genome association analysis tool)(13). We found that common exonic SNPs in probands of the SMHC cohort were adjacent to the cluster of East Asian populations previously characterized, suggesting that the SMHC cohort faithfully carried genetic signatures of East Asian populations (Fig. 1A).

After performing variant filtering, we discovered a set of 220 de novo mutations (DNMs) (Table S1). We classified DNMs into three categories, including High-impact, Moderate-impact, and Possible-damaging. The High- and Moderate-impact were defined by VEP (Ensembl Variant Effect Predictor, https://asia.ensembl.org/info/docs/tools/vep/index.html). Briefly, the High-impact variants usually lead

to truncation of protein products, including gain or loss of STOP codons as well as frameshift-causing insertions and deletions (INDELs). Interestingly, among the 55 genes containing High-impact SNVs, there are 18 genes previously reported in the SFARI gene list (Category S, 1, 2, 3) such as *SCN2A, PTEN, MECP2, SRCAP, TCF4*, indicating that most genes containing high-impact SNVs in the Chinese cohorts are novel and not included in the SFARI gene database. (Fig. 1B, C).

Moderate-impact variants were defined as protein sequence changing, but not truncating, such as missense SNVs and inframe INDELs. To further categorize the severity of missense variants, we annotated missense SNVs into a new class, named Possible-damaging missense DNMs, which were defined as the variants predicted to be damaging by at least two of the seven following prediction algorithms: SIFT(14), PolyPhen-2 HumVar(15), PolyPhen-2 HumDiv(15), LRT(16), Mutation Taster(17), Mutation Assessor(18) and PROVEAN(19) annotated by dbNSFP4.0a(20, 21). Interestingly, among 165 Moderate-impact variants, there are only 23 variants are present in the SFARI gene list (Fig.1 B, C).

Over one thousand ASD risk genes in the SFARI gene list were mainly found from genetic studies in US and European studies, therefore we were wondering whether numerous genes with DNMs in the Chinese ASD cohorts which were not included in the SFARI list were really contributory to ASD or some common genetic variants may not associated with disorders. To further determine whether these genes with DNMs may be contributory to ASD, next we statistically evaluated the contributions of each *de novo* variant to ASD using the Transmission and *De Novo* Association Test-Denovo (TADA-Denovo) method. We first measure the frequency of *de novo* and missense variants in each gene with DNMR-SC-subtype data(22), then applied the TADA-Denovo method(23). We classified the DNM variants with p values obtained from TADA-Denovo test into two tiers (*, $p < 0.01$, or **, $p < 0.001$) (Table S2). We further measured the "probability of loss-of-function intolerance" (pLi) score for each variant and categorized variants with significant TADA-Denovo value into two tiers as well (>0.9 represented by $^{\#\#}$, 0.5-0.9 represented by $^{\#}$)(24). Finally, we found that 11 genes with High-impact mutations and 35 genes with Moderate-impact mutations, all of them not included in the SFARI gene list, were statistically significant with both TADA-Denovo and pLi score, further strengthening their contributions to ASD (Fig. 1C).

We would like to investigate whether genes with *de novo* variants identified in various ASD genetic studies may be overlapping. Interestingly, we found that *de novo* ASD risk genes detected in ASD probands in the SMHC cohort showed little overlapped with the list of *de novo* ASD risk genes from the Japanese cohort (Fig. S1B, C)(8). Moreover, we found that there was also little overlapping in *de novo* variants between the SMHC cohorts with other studies with 200-400 trios (Fig. S1D)(5, 25-27).

Identification of *de novo* CNVs in ASD risk genes with the WES dataset

Although the gold standard for copy number variations detection is the chromosomal microarray analysis (CMA), various toolkits has emerged to identify CNVs with the whole-exome sequencing (WES)

dataset(28). However, the current algorithms for CNV detection are not optimal for the WES dataset and incompatible with the GRCh38/hg38 reference genome.

We applied a germline CNV calling protocol based on GATK cohort mode (version 4.2.0.0) (See Supplementary Methods) and identified numerous *de novo* CNVs in the probands (Fig. 2A-N, Table S3). To exclude the false positive hits, we set 2 standards for CNV screening. First, selection of duplication or deletion signals appearing in more than 2 continuous exons. Second, CNVs should fulfill the HIGH-impact criteria, leading to protein truncation, such as deletion of START or STOP codons.

To prioritize ASD risk genes, we first examine CNVs happened in the known SFARI genes (Fig. 2A-N). We found 18 CNVs exhibiting duplication or deletions in known SFARI genes (Cat S:4 genes, Cat 1:14 genes), such as duplications of *RAI1, UBE3A* and deletions of *TBR1, SHANK3, MECP2, GIGYF1* (Fig. 2A-N). We further validated the CNV results by performing quantitative PCR, confirming the feasibility and faithfulness of our new methods (Fig. 2O).

Furthermore, among *de novo* large CNVs we found, there are 9 CNVs containing genes in the SFARI Cat 2 gene list (Table S3). There are totally 26 CNVs containing critical ASD-risk genes in the SFARI gene list (Cat S, 1, 2), suggesting that genes implicated in these *de novo* large CNVs may contribute to pathogenesis of ASD.

Expression of ASD risk genes enriched at PC, PRC and BST regions in the developing human brain

The etiology of ASD may be disruption of neural circuits associated with social behaviors, thus identification of the expression profile of gene with DNMs in the human brain would provide critical insights for which brain regions may be affected by mutations of ASD risk genes(29). To acquire the expression pattern of ASD risk genes in the single-cell resolution, we used the recent single-cell sequencing database in the developing human brain(30, 31). We grouped total 17434 transcriptomes collected from gestational week (GW) 09-26 of human fetus brains and categorized them into sub-cell types according to marker genes (Fig. 3A, Fig. S2A-C).

We first examined the expression pattern of 55 High-impact genes and 165 Moderate-impact genes in various cell types, and found that both High-impact genes and Moderate-impact genes were highly expressed in several subtypes of cells, including NPC-4, Ex-1 and In-2, as well as cajal-retzius cells (CR) (Fig. 3B, C, D). We further looked into where NPC-4, Ex-1, In-2, and cajal-retzius cells (CR) localized in the developing human brain. We found that NPC-4 was generally distributed in the four major lobes of the brain, suggesting that this specific sub-group of neural progenitor cells may be associated with ASD (Fig. 3E, F, Fig.S3A, B). However, Ex-1 and In-2 specifically enriched in some sub-regions of the brain including precentral gyrus (PRC), postcentral gyrus (PC) and banks of superior temporal sulcus (BST) regions (Fig. S3C-E).

We next investigated whether expressions of ASD risk genes may be enriched in specific brain regions of the human brain. In previous work, the single-cell sequencing were performed in 22 brain subregions in

the developing human brain (Fig. 4A)(30). Surprisingly, we found that the High- and Moderate-impact genes were significantly enriched in precentral gyrus (PRC), postcentral gyrus (PC) and banks of superior temporal sulcus (BST) regions (Fig. 4B, C, D, E). The PRC is the primary motor cortex (M1), and PC is the primary somatosensory cortex (S1). The implications of PRC and PC in ASD had been reported previously(32, 33). Interestingly, we also found the functional connectivity including right S1 (S1R) and M1 (M1R) regions were specifically decreased in *MECP2* transgenic monkeys, the non-human primate model for autism, comparing to wild-type monkeys (34-36).

Brain imaging analysis

In order to determine whether these brain regions were affected in ASD patients from different populations, we acquired imaging data from Autism Brain Imaging Data Exchange (ABIDE-I, http://fcon_1000.projects.nitrc.org/indi/abide/)(37) a publicly available database released containing 1112 subjects (539 ASDs, 573 age-matched healthy controls-HCs) from 16 international imaging sites underwent anatomical and resting-state functional MRI scans. we collected more than 200 age-matched brain imaging data from ASD or HC groups of ABIDE-I (Table S4).

To further validate whether PC/PRC and BST may have structural (gray matter) alternations in ASD patients, we first performed the voxel-based morphometry (VBM) analysis of these region in ASD and HC using T1 data. Surprisingly, we found that the gray matter volume of BST in the right hemisphere was significantly smaller in the ASD group than that in the HC group ($t = 3.61$, $p = 0.003$, t- and p-value from linear mixed model detailed in Statistic section of Methods and Materials), and this effect persisted even after controlling for medication status ($t = 3.32$, $p = 0.001$) and full-scale intelligence quotient (FIQ) ($t = 3.4$, $p = 0.0007$) (Fig. 5A, B, C).

We finally investigate the potential functional connectivity (FC) between the above regions of interests (ROIs) and the whole-brain voxels, by performing seed-based FC analysis using resting-state functional MRI scans data from ASD and HCs. Consistently, we observed a significant decrease in connectivity between BST/PC/PRC and sensory areas, insula, as well as frontal lobes in ASD compared to HC (Fig. 5D). We found a decrease of all six ROIs' functional connectivity to the occipital lobe region, which is commonly associated with vision. We also found decreased connectivity between bilateral PC/PRC to the sensorimotor region of the parietal lobe. In addition, on the right BST, we found the most widely FC decrease among all ROIs, including connections to the right insula and temporal lobes ($t = -6.05$, FWE corrected $p = 0.0002$), to the bilateral frontal lobe and to the occipital lobe (Table S5-S6 and Fig. 5D).

BST has been shown to be voice-selective areas in normal adults, which plays a role in voice recognition and social stimuli processing (38). In an fMRI study, activation of BST by speech stimulation appeared compromised in adults with ASD (39). In addition, BST also exhibited ASD-related functional connectivity alterations (40, 41), gray matter changes (such as lower surface area and greater age-related cortical thinning) (42, 43) and white matter volume reduction (44). Our study indicated that genetic predispositions in ASD patients may lead to structural and function abnormalities in brain regions

associated with processing of social information, thus providing novel candidate brain regions for intervention of ASD.

Discussion

With accumulating genomic studies on autism cohorts world-wide, the genetic architecture of ASD has emerged over the last decade. Composed of *de novo* and rare inherited mutations, genetic variants play a decisive role in determining the etiology of ASD. Although the rapid development of DNA sequencing technology, precise identification of genetic variants in the large scale genome sequencing over hundreds and thousands of ASD core trios is still very challenging.

In this work, we applied the latest GATK package (v4.2.0.0) and the GRCh38/hg38 dataset, which is compatible for ongoing update of Ensembl genome database. We focused on the identification of *de novo* variants, including SNVs, INDELs and CNVs, with the customized joint calling pipeline. Importantly, we found several critical CNVs containing ASD-risk genes, such as *SHANK3, TBR1* and *MECP2*, indicating that screening CNVs with the WES dataset would be very valuable for ASD genetic studies. Interestingly, about 30% (18/55) genes carried *de novo* High-impact variants existed in the SFARI gene list, suggesting that there are potentially novel ASD genes in the Chinese cohorts. With the in-depth analysis of TADA-Denovo and pLi evaluation, we found that there are a substantially portion of DNM discovered in the Chinese ASD cohorts appeared to be significant statically. Taken together, we suggest that although the overall genetic architecture of ASD remains similar across different populations, the frequency of individual genetic component may vary due to geographic isolations. Thus, in order to comprehensively acquire the ASD risk genes, genome-wide sequencing in large cohorts from different populations would be required.

One of intriguing hypothesis of etiology of ASD is that genes carrying genetic mutations in ASD patients may express in some specific brain regions governing social behavior-related circuits. Thus precise identification the expression pattern of ASD risk genes in human brain, with single-cell resolution, would be an ideal approach. In this work, we took advantage of single-cell sequencing database collected from various brain regions across gestational week 9–26 and found that the expression profiles of ASD risk genes in the developing human brain indeed exhibited specific patterns. ASD risk genes discovered in the Chinese ASD cohort specifically enriched in the primary somatosensory (S1-PC) and primary motor cortices (M1-PRC), as well as the BST region.

Although there are evidence suggesting that S1 and M1 cortices showed defects in functional connectivity in ASD samples (34–36), it still difficult to determine whether the defects associated with S1 or M1 are the cause or consequence of deficits in social behaviors. It is well known that ASD patients often exhibit abnormal somatosensory functions. Thus our data suggested that abnormalities in somatosensory functions may root from central control, other than peripheral functions.

The finding that right BST region is implicated in ASD pathogenesis is extremely intriguing. The brain regions in the right hemisphere are stronger associated with autism than left hemisphere are also

observed in previous work (34–36). The connections between right-hemisphere with autism are also discussed(45). Involvement of BST in social perception has been found in previous work with human subjects(46). Thus we hypothesized that dysregulation of ASD risk genes in BST likely caused abnormalities in social perception and related functions. Since BST is on the surface region of human brain, one may be able to design neuromodulation methods to activate the neural activity in the right BST region of ASD patients, through transcranial electrical or magnetic stimulations.

Taken together, this work presented numerous ASD risk genes from whole-exome sequencing of 369 Chinese ASD cohorts. Beside known ASD candidate genes present in the SFARI gene database, we further measured the probability of DNMs contributions to ASD through the TADA-Denovo test and found that quite a few ASD candidate genes appeared to be statistically significant, suggesting that whole-exome sequencing on large ASD cohorts are indeed valuable to elucidate the genetic landscape of ASD in different populations. Importantly, the combinational analysis of single-cell sequencing and brain imaging in this work presented an analytical framework in which one could address the potential etiology of ASD from genetic discoveries.

Limitations

Due to the limited sample size of the current ASD genetic study, the further statistical analysis including false discover rate (FDR) cannot be performed.

Conclusion

This work indicated that the integrated analysis with genome-wide genetic screening, single-cell sequencing and brain imaging data would reveal novel insights into brain regions contributing to etiology of ASD.

Abbreviations

ASD, Autism Spectrum Disorder; SFARI, Simons Foundation Autism Research Initiative; SNVs, single nucleotide variants; INDELs, insertions and deletions. The abbreviations for brain subregions are listed in Table S7.

Declarations

Acknowledgements

The authors thank the families for their participation in this study.

Funding: NSFC Grants (#31625013, #81941405, #32000726, #61973086)

Shanghai Brain-Intelligence Project from STCSM (16JC1420501);

Strategic Priority Research Program of the Chinese Academy of Sciences (XDBS01060200); Program of Shanghai Academic Research Leader,

The Open Large Infrastructure Research of Chinese Academy of Sciences,

Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01).

Authors' contributions

Study design: B Yuan and Z Qiu.

Experiment and data analysis: B Yuan, M Wang, X Wu.

Acquisition of Clinical information: PP Cheng, R Zhang (SMHC), Y Yu, YS Du.

Interpretation of WES data: B Yuan.

CNV validation experiment: B Yuan, PP Cheng, R Zhang (ION).

Analysis of single-cell sequencing data: M Wang, XQ Wang.

Analysis of brain imaging data, X Wu, J Zhang. Drafting of the manuscript: B Yuan, Z Qiu.

Study supervision: Z Qiu, YS Du, XQ Wang, J Zhang.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Patients were collected from outpatient Department of the Child and Adolescent Psychiatry, Shanghai Mental Health Center. Written informed consent forms were obtained from parents for all minor children.

Data and materials availability:

The datasets used and/or analyzed during the current study are available from the lead contact on reasonable request.

Ethics approval and consent to participate

Experiments were approved by the Institutional Review Board (IRB), Shanghai Mental Health Center of Shanghai Jiao Tong University (FWA number 00003065; IROG number 0002202). Ethical review number of this study is 2016–4, and committee members of IRB who approved this study was Dr. Yi-Feng Xu. All sample collections were handled according to the protocol approved by the Institutional Review Boards.

References

1. Maenner MJ, Shaw KA, Baio J, EdS, Washington A, Patrick M, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. MMWR Surveill Summ. 2020;69(4):1-12.
2. Zhou H, Xu X, Yan W, Zou X, Wu L, Luo X, et al. Prevalence of Autism Spectrum Disorder in China: A Nationwide Multi-center Population-based Study Among Children Aged 6 to 12 Years. Neurosci Bull. 2020;36(9):961-71.
3. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-15.
4. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-21.
5. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41.
6. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020;180(3):568-84 e23.
7. 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). Mol Autism. 2013;4(1):36.
8. Takata A, Miyake N, Tsurusaki Y, Fukai R, Miyatake S, Koshimizu E, et al. Integrative Analyses of De Novo Mutations Provide Deeper Biological Insights into Autism Spectrum Disorder. Cell Rep. 2018;22(3):734-47.
9. Guo H, Wang T, Wu H, Long M, Coe BP, Li H, et al. Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. Mol Autism. 2018;9:64.
10. Wang T, Guo H, Xiong B, Stessman HA, Wu H, Coe BP, et al. De novo genic mutations among a Chinese autism spectrum disorder cohort. Nat Commun. 2016;7:13316.
11. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, et al. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. Am J Hum Genet. 2013;93(2):249-63.
12. Wu J, Yu P, Jin X, Xu X, Li J, Li Z, et al. Genomic landscapes of Chinese sporadic autism spectrum disorders revealed by whole-genome sequencing. J Genet Genomics. 2018;45(10):527-38.

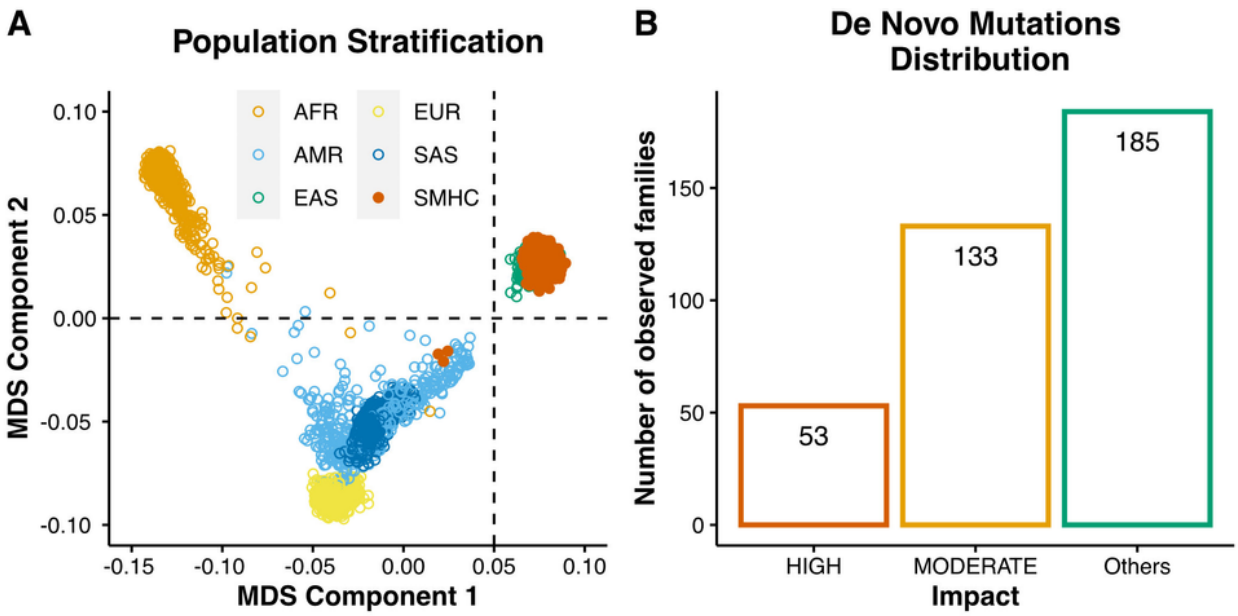
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.
14. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9.
15. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
16. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-61.
17. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-6.
18. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
19. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745-7.
20. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894-9.
21. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235-41.
22. 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.
23. 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.
24. 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):285-91.
25. Feliciano P, Zhou X, Astrovskaya I, Turner TN, Wang T, Brueggeman L, et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. NPJ Genom Med. 2019;4:19.
26. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012;485(7397):242-5.
27. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246-50.
28. Enomoto Y, Tsurusaki Y, Yokoi T, Abe-Hatano C, Ida K, Naruto T, et al. CNV analysis using whole exome sequencing identified biallelic CNVs of VPS13B in siblings with intellectual disability. Eur J Med Genet. 2020;63(1):103610.
29. Geschwind DH, State MW. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 2015;14(11):1109-20.

30. 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.
31. 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.
32. Ayub R, Sun KL, Flores RE, Lam VT, Jo B, Saggar M, et al. Thalamocortical connectivity is associated with autism symptoms in high-functioning adults with autism and typically developing adults. Transl Psychiatry. 2021;11(1):93.
33. Mizuno Y, Kagitani-Shimono K, Jung M, Makita K, Takiguchi S, Fujisawa TX, et al. Structural brain abnormalities in children and adolescents with comorbid autism spectrum disorder and attention-deficit/hyperactivity disorder. Transl Psychiatry. 2019;9(1):332.
34. Cai DC, Wang Z, Bo T, Yan S, Liu Y, Liu Z, et al. MECP2 Duplication Causes Aberrant GABA Pathways, Circuits and Behaviors in Transgenic Monkeys: Neural Mappings to Patients with Autism. J Neurosci. 2020;40(19):3799-814.
35. Liu Z, Li X, Zhang JT, Cai YJ, Cheng TL, Cheng C, et al. Autism-like behaviours and germline transmission in transgenic monkeys overexpressing MeCP2. Nature. 2016;530(7588):98-102.
36. Zhan Y, Wei J, Liang J, Xu X, He R, Robbins TW, et al. Diagnostic Classification for Human Autism and Obsessive-Compulsive Disorder Based on Machine Learning From a Primate Genetic Model. Am J Psychiatry. 2021;178(1):65-76.
37. Di Martino A, Yan CG, 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. Mol Psychiatry. 2014;19(6):659-67.
38. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000;403(6767):309-12.
39. Gervais H, Belin P, Boddaert N, Leboyer M, Coez A, Sfaello I, et al. Abnormal cortical voice processing in autism. Nature neuroscience. 2004;7(8):801-2.
40. Shih P, Keehn B, Oram JK, Leyden KM, Keown CL, Müller R-A. Functional differentiation of posterior superior temporal sulcus in autism: a functional connectivity magnetic resonance imaging study. Biological psychiatry. 2011;70(3):270-7.
41. Venkataraman A, Duncan JS, Yang DYJ, Pelphrey KA. An unbiased Bayesian approach to functional connectomics implicates social-communication networks in autism. NeuroImage: Clinical. 2015;8:356-66.
42. Mensen VT, Wierenga LM, van Dijk S, Rijks Y, Oranje B, Mandl RCW, et al. Development of cortical thickness and surface area in autism spectrum disorder. NeuroImage: Clinical. 2017;13:215-22.
43. Braden BB, Riecken C. Thinning faster? Age-related cortical thickness differences in adults with autism spectrum disorder. Research in autism spectrum disorders. 2019;64:31-8.
44. von dem Hagen EAH, Nummenmaa L, Yu R, Engell AD, Ewbank MP, Calder AJ. Autism spectrum traits in the typical population predict structure and function in the posterior superior temporal sulcus.

Cerebral cortex. 2011;21(3):493-500.
45. Ozonoff S, Miller JN. An exploration of right-hemisphere contributions to the pragmatic impairments of autism. Brain Lang. 1996;52(3):411-34.
46. Deen B, Koldewyn K, Kanwisher N, Saxe R. Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. Cereb Cortex. 2015;25(11):4596-609.
47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
48. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.
49. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018.
50. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.
51. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
52. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
53. Wei Q, Zhan X, Zhong X, Liu Y, Han Y, Chen W, et al. A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics. 2015;31(9):1375-81.
54. Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods. 2013;10(10):985-7.

Figures

Figure.1 Yuan et al.



C

HIGH	*ANK2***##*	*CACNA2D3***##*	*CTCF***##*	*MECP2***#*	*PTEN***##*	*RORB***##*
	*SCN2A***##*	*SRCAP***##*	*SYNGAP1***##*	*TCF20***##*	*TCF4***##*	*ZMYM2***##*
	*CUX1***##*	*DISC1*	*TAF6*	*NR2F1*	*INTS1*	*PIK3R2***##*
	ANKLE2	*ASTN1***##*	*CD82*	*CYP24A1*	*DHX35*	*EMILIN2*
	FBXL13	*GCC1*	*GOLGA6L6*	*GTF2A2*	*IRF2*	*ITPR2*
	*KCNV1***#*	*KLF9***#*	*KLHDC4*	*LRCH3*	*MFSD2A***#*	*NAGLU*
	*NFASC***##*	*NUPR1*	*PSPH***#*	*PXT1*	*RAB26*	*RALGAPA1***##*
	*RBMS1***##*	*REV3L***##*	*ROBO1*	*RRAD*	*SMC5***#*	*SRRM2*
	*TBX2***##*	*TFCP2*	*URB1*	*VPS37C*	*WDPCP*	*ZNF776*
	ZNF878					
MODERATE	*AHDC1*	*CASZ1*	*CHD2*	*CTNNB1**##*	*GRIN2B*	*MAGEL2**##*
	MEF2C	*NSD1*	*SCN1A*	*SCN2A***##*	*SCN8A*	*SPAST**##*
	TRAF7	*CUX1***##*	*FBN1*	*LAMB1*	*TERF2*	*CSNK1E**##*
	GTF2I	*HYDIN*	*POLA2*	*CDK13*	*HCN1*	*CACNG2**##*
	*MAP2K6**##*	*PSMC3**##*	*CTBP1**##*	*TERF2**##*	*INPP5A**##*	*NRBP1**##*
	*SIRT1**##*	*KLHL2**##*	*KLHL10**##*	*PDE1B**##*	*GLYR1**##*	*MEPCE**##*
	*TPX2**##*	*PROX1**##*	*TLK1**##*	*COG3**##*	*LRRC41**##*	*HCN1**##*
	*SLC9A1**##*	*BAHD1**##*	*NEDD4L**##*	*CLSTN3**##*	*SLC4A4**##*	*BTBD11**##*
	*CAND1**##*	*GTF2I**##*	*BLID**#*	*LYL1**#*	*RAB9B**#*	*EFNA1**#*
	*B3GNT7**#*	*METTL14**#*	*KCNC3**#*	*LRRC8D**#*	*SDK1*	*SDK1*
	YIPF2	*YIPF2*	… 103 more			

SFARI Cat1 | SFARI Cat2 | SFARI Cat3 | SFARI CatS | Recurrent Genes

**:$p_{TADA\text{-}Denovo}$<0.001, *: $p_{TADA\text{-}Denovo}$<0.01, ##: pLI>0.9, #:0.5<pLI<0.9

Figure 1

Identification of *De Novo* Mutations in ASD Probands.

(A) Multidimensional scaling plots of 369 ASD probands in our cohort (OWN) along with African (AFR), American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) individuals in the 1000 Genomes Project. Figure was generated by analyzing genotyping data of 1,064 common exonic SNPs

using PLINK. The first and second dimensions are shown. The orange, sky blue, bluish green, yellow and blue circles dots indicate AFR, AMR, EAS, EUR SAS individuals. The vermillion dots indicate ASD probands exome sequenced in this study.

(B) The number of families with various types of *de novo* mutations in our ASD cohort. Others stand for no detection of High- or Moderate-impact *de novo* mutations.

(C) The list of High-impact and Moderate-impact mutations identified in this ASD cohort. The gene also presented in the SFARI gene list were highlighted with various colors representing 4 categories (Cat S, 1-3).

Figure 2

Identification of *De Novo* CNVs

(A-N) Schematic diagrams of 18 SFARI ASD risk genes (Cat S: 4 genes, Cat 1:14 genes) with *de novo* CNVs. Blue: deleted chromosomal segment. Red: duplicated chromosomal segments.

(O) Quantitative real-time PCR verification of *de novo* CNVs. *TBR1*, t-test, $p=3.22\times10^{-7}$; *RAI1*, t-test, $p=7.40\times10^{-7}$; *MECP2*, t-test, p=0.0069; *SHANK3*, t-test, p=0.0028, *GIGYF1*, t-test, p=0.0335. All experiments were repeated more than 4 independent times. Error bars represent standard errors from four or six replicates.

Figure 3

Expression pattern of ASD risk genes among different cell types.

(A) clustering of single-cell RNA-seq data from different brain regions. Cell types were colored differently. (B) Visualization of enrichment score of the 55 High-impact genes by UMAP. (C) Visualization of enrichment score of the 165 Moderate-impact genes by UMAP. (D) Heatmap showing averaged enrichment score of ASD risk genes among different subtypes (high, red; low, blue). The top three most enriched cell types for each gene-set were shown on the sidebar. (E) Histogram showing the proportion of highly expressed genes (genes expressed in at least 25% cells from individual cell type) of the 55 High-impact genes among cell types. (F) Histogram showing the proportion of highly expressed genes (genes expressed in at least 25% cells from individual cell type) of the 165 Moderate-impact genes among cell types.

Figure 4

Expression pattern of ASD risk genes among different brain regions.

(A) The schematic diagram showing the 22 brain sub-regions used in published human single cell RNA-seq data was applied in analysis of this work. (B) UMAP displaying diverse brain regions (Abbreviations are listed in Table S7). Brain sub-regions were differently colored. (C) Heatmap showing averaged enrichment score of ASD risk genes among different brain regions (high, red; low, blue). The top three most enriched brain regions for each gene-set were shown on the sidebar. (D-E) Histogram displaying the proportion of highly expressed genes (genes expressed in at least 25% cells from individual brain region) of the 55 High-impact genes (D) and the 165 Moderate-impact genes (E) among various brain sub-regions.

Figure.5 Wu et al.

A

VBM	beta	t	p	DF
left BST / right BST	-0.25 / -1.23	-0.85 / -3.61	0.393 / 0.0003	762 / 762
left PC / right PC	-1.06 / -0.71	-1.55 / -1.14	0.121 / 0.257	762 / 762
left PRC / right PRC	1.05 / 0.80	1.47 / 1.00	0.141 / 0.317	762 / 762

B

VBM (control FIQ effect)	beta	t	p	DF
left BST / right BST	-0.35 / -1.15	-1.17 / -3.32	0.241 / 0.001	697 / 697
left PC / right PC	-0.68 / -0.77	-0.98 / -1.21	0.326 / 0.225	697 / 697
left PRC / right PRC	0.98 / 0.49	1.33 / 0.60	0.183 / 0.551	697 / 697

C

VBM (exclude subjects on medication)	beta	t	p	DF
left BST / right BST	-0.32 / -1.19	-1.04 / -3.40	0.3 / 0.0007	717 / 717
left PC / right PC	-1.17 / -0.79	-1.70 / -1.24	0.089 / 0.215	717 / 717
left PRC / right PRC	1.03 / 0.77	1.41 / 0.94	0.159 / 0.348	717 / 717

D

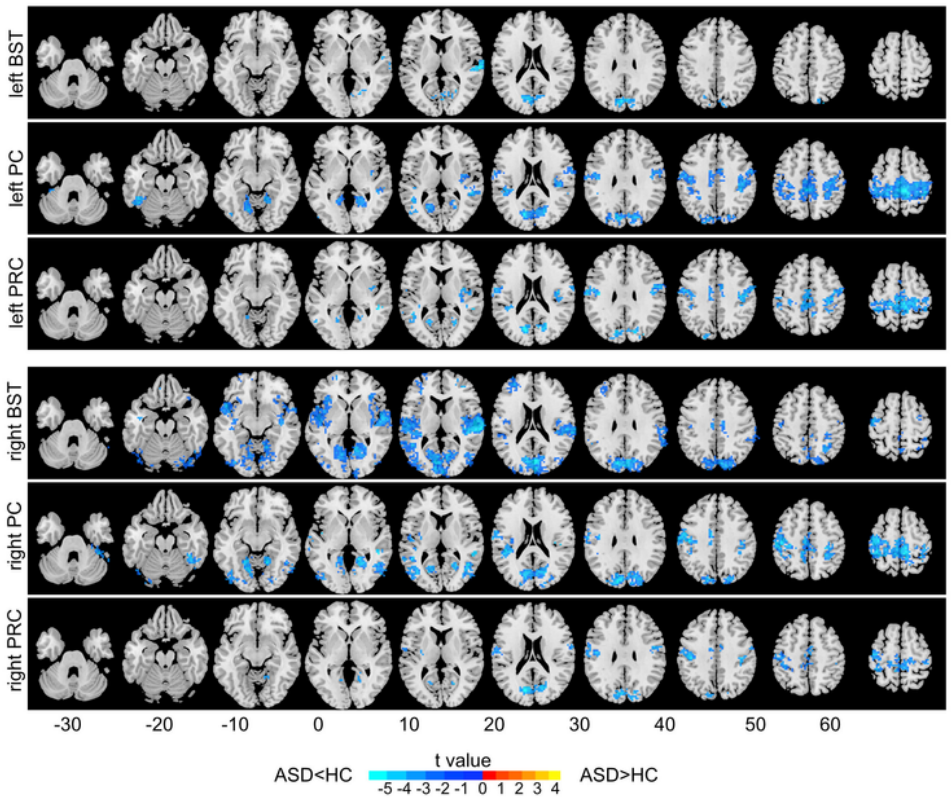


Figure 5

Structural and functional connectivity between ASD and healthy controls in PC, PRC and BST regions.

(A) VBM. Beta, t, p, and degree of freedom (DF) are obtained from linear mixed models. Beta is the regression coefficient of the fixed effect. DF is the degree of freedom of the model (equal to the sample size minus the number of parameters to be estimated). T stands for t-statistics for testing the null

hypothesis that the coefficient is equal to zero. P stands for p-value for the t-test. (B) VBM data with control of FIQ effect. (C) VMB data excludes subjects with medications.

(D) Group differences between ASD and healthy controls in seed-based resting-state Functional Connectivity（FC). 6 brain regions including bilateral BST, PC and PRC were selected as seeds. Functional connectivity between these 6 regions and the whole-brain voxels were calculated. Only the clusters of voxels with significant FC differences between the two groups (the two-sample t-test, p_{FWE}< 0.05, cluster size < 20 voxels or 180 mm^3) were displayed.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalMaterialsandMethods0427a.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.pdf
- TableS5.docx
- TableS6.docx
- TableS7.jpg