Identification of Autism Risk Genes in a Chinese Cohort via Whole-Exome Sequencing with the Joint Calling Analysis

Bo Yuan

Institute of Neuroscience, Chinese Academy of Sciences

Peipei Cheng

Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine

Ran Zhang

Institute of Neuroscience, Chinese Academy of Sciences

Yasong Du

Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine

Zilong Qiu (✉ zqiu@ion.ac.cn)

Institute of Neuroscience, Chinese Academy of Sciences <https://orcid.org/0000-0003-4286-3288>

Short report

Keywords: Autism spectrum disorder, Whole-exome sequencing, Joint calling, Copy number variations

Posted Date: November 18th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-106326/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

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, transcriptional regulation, and chromatin remodeling, have been identified, the genetic analysis on east Asian ASD cohorts in the whole-geome or whole-exome level is still limited(1-5). Here we performed whole-exome sequencing on 168 ASD probands with their unaffected parents of Chinese origin. We applied a joint calling analytical pipeline based on GATK best practices and identified numerous *de novo* variants including single nucleotide variants (SNVs) and insertion or deletions (INDELs). By querying the Simons foundation autism research initiative (SFARI) gene database, we found that there were potential novel ASD risk genes in East Asian cohorts, which did not exist in European American populations. Furthermore, our analysis pipeline identified *de novo* copy number variations (CNVs) of known ASD-related gene based on a sufficiently large sample size, validated by quantitative PCR. Our work indicated that there may be differences in potential ASD genetic components existing across different geographical populations, suggesting that genomic analysis over large cohorts are required for each population in order to precisely identify ASD risk genes.

Introduction

Current prevalence of ASD has approximately increased to 1 in 49 children in the United States, and males are four times more susceptible for ASD than females(6). Recently, tremendous efforts in ASD genetic studies using whole-exome and whole-genome sequencing have enabled high-throughput assessment of 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 of genetic risks for ASD(1, 7, 8). Although genomic information of large cohorts consisting of tens of thousands ASD patients have been collected, east Asian populations are still underrepresented groups. Whether geographical factors may contribute to genetic causes of ASD remained to be addressed(5).

In this study, we collected 168 ASD probands with their parents and performed whole-exome sequencing analysis. 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. CNVs were called using a cohort mode GATK pipeline detecting germline copy number variants.

Interestingly, we found that potential ASD risk genes identified in this study are largely distinct from the result in SFARI databases and ASD gene candidates in the Japanese population, suggesting that geographical difference may play a critical role in genetic variations leading to psychiatric diseases including ASD.

Materials And Methods

Samples and ethics statement

We analyzed a sample set consisting of 168 ASD probands and 326 parents from 163 pedigrees recruited from Department of the Child and Adolescent Psychiatry, Shanghai Mental Health Center. Of the families 5 are multiplex that have two ASD children and 158 are trios. The fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) were 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

260 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. 234 other 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)(1), 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(2). Picard tools MarkDuplicates, SortSam and SetNmMdAndUqTags was used for marking duplicates, sorting by chromosome coordinates and adding essential tags. Single nucleotide variants (SNVs) and insertions / deletions (INDELs) were jointly called across all samples using the Genome Analysis Toolkit (GATK) HaplotypeCaller 4.1.4.1(3). 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 4.1.4.1.

We included variant calls with PASS flag in the downstream analyses. Variants (SNVs and INDELs) were annotated on the basis of the hg38 database using VEP(4). 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 data of common exonic SNPs

To define a set of common exonic SNPs, we first selected variants that are; 1) on the InfiniumExome-24v1-1_A1 genotyping array, 2) with MAF > 0.05 in East Asian (EAS) population of ExAC(5) 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(6), we performed further filtering and linkage disequilibrium (LD)-based pruning using PLINK v1.9(7) 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 (N of individuals = 326). By using the information of the remaining variant calls, we extracted candidates for DNMs using GATK PossibleDeNovo, TrioDenovo(8), DeNovoGear(9). Candidate DNMs called by these three tools at the same time were then stratified into SNVs and INDELs. We selected 98 DNM calls by prioritizing HIGH impact DNMs and MODERATE impact DNMs into consensus damaging missense. (CD-missense) DNMs were defined as the variants predicted to be damaging by at least two of the seven prediction algorithms: SIFT(10), PolyPhen-2 HumVar(11), PolyPhen-2 HumDiv(11), LRT(12), MutationTaster(13), Mutation Assessor(14) and PROVEAN(15) annotated by dbNSFP4.0a(16, 17).

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(18).

Real-time Quantitative PCR Validation

To confirm *de novo* CNVs detected by WES, quantitative PCR (qPCR) was performed using DNA from probands, their parents and controls. The comparative CT method (delta-delta CT method) was used 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(19). 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(20). We included IMPACT HIGH and CD 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 CD missense DNMs were obtained from mirDNMR based on sequence context(21).

Results

Identification of *de novo* variant in ASD probands

We analyzed a sample set consisting of 168 ASD probands and 326 parents from 163 pedigrees recruited from Department of the Child and Adolescent Psychiatry, Shanghai Mental Health Center. Among the cohort, there are 5 multiplex family containing two ASD children and the rest 158 family are trios having 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 target exome region covered with $\geq 10x$ or $30x$ of reads indicates sufficient coverage (Figure. S1A). After performing the multidimensional scaling of the genotyping data of common exonic SNPs was performed by using PLINK (a whole genome association analysis toolset)(9), we found that all probands in this cohort were included in the cluster of East Asian individuals (Figure. 1A).

After performing variant filtering, we discovered a set of 442 *de novo* mutations (DNMs) (Table S1). We classified SNV and INDELs into three classes, 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 product, such as gain or loss of STOP codons as well as frameshift-causing INDELs. We identified 11 HIGH-impact SNVs and 8 HIGH-impact INDELs (Figure. 1B, C). Interestingly, among the 11 genes containing HIGH-impact SNVs, there are 5 genes previously reported in the SFARI gene list (*SCN2A*, *POGZ*, *MECP2*, *SRCAP*, *TCF4*). However, in the 7 genes containing HIGH-impact INDELs (there are 2 recurrent INDELs in *SYNGAP1* genes), only *SYNGAP1* and *CUX1* are reported in the SFARI gene list, suggesting that there are substantially non-SFARI ASD gene in the Chinese cohorts (Figure. 1B, C).

MODERATE-impact variants were defined as protein sequence changing, but not truncating, such as missense variants and inframe INDELs. We found there are 15 inframe INDELs classified as MODERATE-impact variants (Table S1). To further categorize the severity of missense variants, we annotated missense 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(10), PolyPhen-2 HumVar(11), PolyPhen-2 HumDiv(11), LRT(12), Mutation Taster(13), Mutation Assessor(14) and PROVEAN(15) annotated by dbNSFP4.0a(16, 17). We found 64 Possible damaging missense DNMs (Table S1).

Next we statistically assessed the observed number of *de novo* variants in each gene using Transmission and *De Novo* Association Test-Denovo (TADA-Denovo) and identified one gene significantly enriched for *de novo* mutations (*SYNGAP1* q val < 0.05) (Table S2). Overall, *de novo* ASD risk genes detected in ASD probands from the Chinese cohort showed little overlapped with the list of *de novo* ASD risk genes in ASD probands from the Japanese cohort(Figure. 1D)(5). Only a few SFARI genes, including *SYNGAP1*, *POGZ* and *NCOA6* were found in both East Asian cohorts.

Odds Ratio may not be a good measure of genetic risks for ASD.

We next re-annotated DNM data from 4872 ASD probands and 1943 unaffected siblings originally from db-denovo v.1.6.1 with the same pipeline as used for our dataset. By comparing the proportion of individuals carrying one or more HIGH, MODERATE, LOW and MODERATE impact mutations in the case groups with controls, We confirmed that carriers of HIGH impact DNM were significantly enriched in both our cohort and the db-denovo ASD cohort ($p = 2.792 \times 10^{-11}$, odds ratio [OR] = 3.182105 in our ASD cohort; $p = 2.843 \times 10^{-6}$, odds ratio [OR] = 1.418789 in the db-denovo ASD cohort, Figure. S1B). However, there was no enrichment of MODERATE-impact DNM carriers with lower ORs in db-denovo case cohorts ($p = 1.852 \times 10^{-6}$, odds ratio [OR] = 2.343237 in our ASD cohort; $p = 0.3342$, odds ratio [OR] = 0.9531462 in the db-denovo ASD cohort, Figure. S1B). And there was a statistically significant enrichment of LOW impact DNM carriers with a pair of contradictory ORs in both case cohorts ($p = 0.01748$, odds ratio [OR] = 1.464373 in our ASD cohort; $p = 2.321 \times 10^{-4}$, odds ratio [OR] = 0.8257816 in the db-denovo ASD cohort, Figure. S1B).

Furthermore, the odds ratio of MODIFIER impact DNM carriers in the db-denovo ASD cohort suggests that a type of mild mutations inhibit the onset of ASD($p = 0.7444$, odds ratio [OR] = 1.185185 in our ASD cohort; $p = 2.2 \times 10^{-16}$, odds ratio [OR] = 0.1182785 in the db-denovo ASD cohort, Figure. S1B). Taken together, these results indicate that odds ratio can only be partially used to determine the effect of different mutation types on the incidence of ASD.

Identification of 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(18). However, the current reported algorithms for CNV detection is 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 4.1.4.1) (See Supplementary Methods) and identified numerous *de novo* CNVs in the probands (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 criterial, 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 (Figure. 2A-G). We found 8 CNVs exhibiting duplication or deletions in known SFARI genes, such as duplications of *AMT*, *RAI1*, *TBC1D23*, and deletions of *TBR1*, *SHANK3*, *MECP2*, *GIGYF1* (Figure. 2A-G). We further validated the CNV results by performing quantitative PCR (Figure. 2H), confirming the feasibility and faithfulness of our new methods.

Importantly, we further identified *de novo* large CNVs, containing multiple genes (Figure. S2A-H, Figure. S3A-H). To investigate whether these candidate genes may be involved in brain development, we examine the expression pattern of candidate genes which exhibited either duplications or deletions in ASD patients in the GTEx Analysis Release V8 database (dbGaP Accession phs000424.v8.p2). We found that numerous candidate genes indeed were expressed in the central nervous system (Figure. S4), suggesting that genes implicated in these *de novo* large CNVs may contribute to pathogenesis 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 (v4.1.4.1) 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 40% genes carried *de novo* HIGH-impact variants (18), existed in the SFARI gene list, suggesting that there are potentially novel ASD genes in the Chinese cohorts. Consistently, among the 46 genes carried *de novo* HIGH-impact variants in the Japanese ASD cohorts, only 8 genes appeared in the SFARI gene list, suggesting that geographical factors may play a critical role in contributing the ASD. Taken together, we suggest that although the overall genetic architecture of ASD remains similar across different populations, the genetic components 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.

Abbreviations

ASD, Autism Spectrum Disorder; SFARI, Simons Foundation Autism Research Initiative; SNVs, single nucleotide variants; INDELs, insertions and deletions.

Declarations

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 our study is 2016–4, and committee members of IRB who approved this study was Dr. Yi-Feng Xu. Patients were collected from outpatient Department of the Child and Adolescent Psychiatry, Shanghai Mental Health Center. Written informed consent was obtained from parents for all minor children and those who were unable to give consent. All participants were ascertained using the protocol approved by the appropriate Institutional Review Boards.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the NSFC Grants (#31625013, #81941405, #32000726); 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, and the Shanghai Municipal Science and Technology Major Project (#2018SHZDZX05).

Authors' contributions

All authors contributed to the work and meet the criteria for authorship. Study concept and design: B Yuan and Z Qiu. Experiment and data analysis: B Yuan. Acquisition of Clinical information: PP Cheng, YS Du. Interpretation of WES data: B Yuan. Experiment: B Yuan, PP Cheng, R Zhang, Drafting of the manuscript: Study supervision: Z Qiu and YS Du.

Acknowledgements

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

References

1. 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.
2. O'Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C, et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat Commun*. 2014;5:5595.
3. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*. 2012;338(6114):1619-22.
4. 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.
5. 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.
6. Maenner MJ, Shaw KA, Baio J, Washington A, Patrick M, DiRienzo 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(No. SS-4):1-12.
7. 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.
8. 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):1215-33.
9. 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.
10. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc*. 2016;11(1):1-9.
11. 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.

12. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res.* 2009;19(9):1553-61.
13. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods.* 2010;7(8):575-6.
14. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39(17):e118.
15. 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.
16. 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.
17. 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.
18. 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.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-60.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc.* 2016;11(1):1-9.
29. 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.

30. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res.* 2009;19(9):1553-61.
31. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods.* 2010;7(8):575-6.
32. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39(17):e118.
33. 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.
34. 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.
35. 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.
36. 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.
37. Ahmed M, Kim DR. pcr: an R package for quality assessment, analysis and testing of qPCR data. *PeerJ.* 2018;6:e4473.
38. 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.
39. 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.

Figures

Figure 1

Figure 1

Identification of De Novo Mutations in ASD Probands. (A) Multidimensional scaling plots of 168 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 red, blue, green, magenta and yellow circles dots indicate AFR, AMR, EAS, EUR SAS

individuals. The orange dots indicate 168 ASD probands exome sequenced in this study. (B) The number of families with various functional types of de novo mutations in our ASD cohort. (C) The list of HIGH-impact mutations identified in our ASD cohort. (D) The overlap of mutant genes between our ASD cohort and Japanese ASD cohort. The red circle indicates genes with HIGH-impact SNVs and INDELs, MODERATE-impact INDELs and Possible damaging missense DNMs in our ASD cohort; the blue circle indicates genes with mutations confirmed by Sanger sequencing in Japanese ASD cohort; the green and magenta circle indicate genes with HIGH-impact DNMs in our and Japanese ASD cohorts, respectively.

Figure 2

Figure 2

(A-G) Schematic diagrams of seven SFARI ASD risk genes with de novo CNVs. (H) Quantitative real-time PCR verification of de novo CNVs. TBR1, n=4, t-test, $p=3.22 \times 10^{-7}$; RAI1, n=6, t-test, $p=7.40 \times 10^{-7}$; SHANK3, n=4, t-test, $p=0.0028$; MECP2, n=4, t-test, $p=0.0069$; GIGYF1, n=3, t-test, $p=0.0335$. Error bars represent standard errors from three, four or six replicates.

Supplementary Files

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

- [FigureS1.jpg](#)
- [FigureS2.jpg](#)
- [FigureS3.jpg](#)
- [FigureS4.jpg](#)
- [TableS1.xlsx](#)
- [TableS2.98candidategenescalculatedbyTADADenovo.docx](#)
- [TableS3.FilteredDeNovoCNVinformation.xlsx](#)