Behavioral and Brain Functions

RESEARCH Open Access

Unraveling genetic risk contributions to nonverbal status in autism spectrum disorder probands

Huan Liu[3,4†], Shenghan Wang[1†], Binbin Cao[5†], Jijun Zhu[1], Zhifang Huang[3,4], Pan Li[1], Shunjie Zhang[1], Xian Liu[3,4], Jing Yu[3,4], Zhongting Huang[1], Linzhuo Lv[3,4], Fuqiang Cai[2], Weixin Liu[2], Zhijian Song[2], Yuxin Liu[1], Tao Pang[7], Suhua Chang[7], Ying Chen[3,4], Junfang Chen[1,6*] and Wen-Xiong Chen[3,4,5*]

Abstract
Autism spectrum disorder (ASD) presents a wide range of cognitive and language impairments. In this study, we investigated the genetic basis of non-verbal status in ASD using a comprehensive genomic approach. We identified a novel common variant, rs1944180 in *CNTN5*, significantly associated with non-verbal status through family-based Transmission Disequilibrium Testing. Polygenic risk score (PRS) analysis further showed that higher ASD PRS was significantly linked to non-verbal status ($p=0.034$), specific to ASD and not related to other conditions such as bipolar disorder, schizophrenia and three language-related traits. Using structural equation modeling (SEM), we found two causal SNPs, rs1247761 located in *KCNMA1* and rs2524290 in *RAB3IL1*, linking ASD with language traits. The model indicated a unidirectional effect, with ASD driving language impairments. Additionally, *de novo* mutations (DNMs) were found to be related with ASD and interaction between common variants and DNMs significantly impacted non-verbal status ($p=0.038$). Our findings also identified 5 high-risk ASD genes, and DNMs were enriched in glycosylation-related pathways. These results offer new insights into the genetic mechanisms underlying language deficits in ASD.

Keywords Autism spectrum disorder, Common variants, *De Novo* mutations, Whole-exome sequencing, Non-verbal status

[†]Huan Liu, Shenghan Wang and Binbin Cao contributed equally to this work.

*Correspondence:
Junfang Chen
junfang_chen@fudan.edu.cn
Wen-Xiong Chen
gzchcwx@126.com
[1]Center for Intelligent Medicine, Greater Bay Area Institute of Precision Medicine (Guangzhou), School of Life Sciences, Fudan University, Shanghai, China
[2]School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China
[3]Department of Behavioral Development, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 510623, Guangdong, China
[4]The Assessment and Intervention Center for Autistic Children, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[5]Department of Neurology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[6]Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, China
[7]NHC Key Laboratory of Mental Health (Peking University), Peking University Sixth Hospital, Peking University Institute of Mental Health, National Clinical Research Center for Mental Disorders (Peking University Sixth Hospital), Beijing, China

Background

Autism Spectrum Disorder (ASD) is a serious neurodevelopment disorder with core symptoms of deficit in social relatedness and repetitive behaviors, and patients have reduced interests, delays in language development, and sensory abnormalities [1, 2, 3]. To date, the global prevalence of ASD continues to rise, reaching 1 – 2%, and it typically occurs in infancy and early childhood [4, 5].

One of the most significant challenges associated with ASD is the pervasive impairment in communication abilities, both verbal and non-verbal. Language delays are among the hallmark features of most ASD, often complicating social interactions and persisting into adulthood [3]. These language deficits vary significantly across the spectrum. While some individuals with ASD are non-verbal, others acquire complex language abilities yet continue to struggle with pragmatic use in the social contexts. The heterogeneity of these language impairments is further complicated by co-occurring physical and mental health conditions, such as intellectual disability, epilepsy, anxiety, and sensory processing disorders [6, 7, 8]. This variability makes ASD a highly heterogeneous disorder, with affected individuals displaying a wide range of cognitive, communicative, and behavioral symptoms [6].

Basically, ASD is highly heritable and therefore is expected to have a substantial contribution from common and rare variation transmitted from parents to their autistic offspring [9, 10, 11, 12]. Studies on twins have shown that the genetic liability of ASD is 64 – 91% [13]. Common inherited variants are estimated to account for approximately 49% of the genetic liability [9], and 21% of diagnoses of ASD are contributed by different *de novo* mutations (DNMs) [14]. Furthermore, large-scale whole-genome sequencing studies have identified hundreds of genes harboring potentially pathogenic variants, including single nucleotide variants (SNVs), small insertions/deletions (indels), structural variants, tandem repeat expansions, and DNMs [10, 12, 15, 16, 17], and they revealed the polygenic nature and high genetic heterogeneity of ASD. These studies highlight the complex interplay between inherited and *de novo* genetic variants in shaping the diverse clinical manifestations of ASD.

Despite significant progress in identifying the genetic underpinnings of ASD, much remains to be uncovered regarding how genetic variants contribute to language impairments. Language delay, a hallmark of ASD, is commonly observed as slower development of language skills compared to neurotypical peers. Genetic studies have increasingly focused on exploring the relationship between polygenic risk scores (PRS) and language delays in ASD [18]. For example, research has established a significant association between polygenic risk for ASD and language delays in multiplex families, emphasizing the genetic contributions to delayed verbal abilities [3]. However, despite these insights, a critical gap remains in understanding the genetic contributions to non-verbal status, defined as the inability to actively produce at least five spontaneous meaningful words [8, 19].

To address the significant gap in understanding the genetic contributions to non-verbal status in ASD, this study employs a comprehensive and innovative approach integrating multiple advanced genomic methodologies, as outlined in Fig. 1. We conducted whole exome sequencing (WES) on 54 ASD families (53 trios and 1 quad), incorporating both *de novo* mutation and common variant analysis to explore their potential contribution on non-verbal status. Strict quality control measures were applied to ensure high accuracy in variant detection and genotype quality. We utilized a family-based Transmission Disequilibrium Test (TDT) to investigate whether specific ASD risk genes were over-transmitted to probands. Additionally, both polygenic risk score (PRS) analysis and polygenic TDT were employed to assess the cumulative genetic burden of common variants and their relationship to clinical traits, particularly non-verbal ability. Genomic structural equation modeling (genomicSEM) was applied to investigate causal relationships between genetic variants and non-verbal status, adding depth to our analysis of the genetic factors underlying language impairments in ASD. Furthermore, the Transmission and *De Novo* Association (TADA) model was used to identify high-confidence ASD risk genes. Finally, gene set enrichment analysis was conducted to uncover biologically relevant pathways, linking genetic variants to functional outcomes.

This multi-layered approach, combining cutting-edge genomic techniques with statistical models, allows for a comprehensive investigation of both inherited and *de novo* mutations. By integrating these methodologies, our study offers novel insights into the genetic architecture of non-verbal status in ASD, while also providing a framework for future research into the genetic mechanisms underlying complex traits in ASD.

Methods

Study sample description

We recruited 54 families (53 trios and 1 quad), comprising 55 children diagnosed with ASD and their unaffected parents, between June 2018 and August 2020, from Guangzhou Women and Children's Medical Center, Guangzhou Medical University. The eligible children, met the inclusion criteria consisting of Diagnosis and statistical Manual of Mental Diseases version-5 (DSM-5), Autism Diagnostic Interview-Revised (ADI-R) [20], and Autism Diagnostic Observation Schedule (ADOS) [21]. For patients who were initially diagnosed before two years old should be followed up to make the definitive diagnosis at least at age of two years. We made

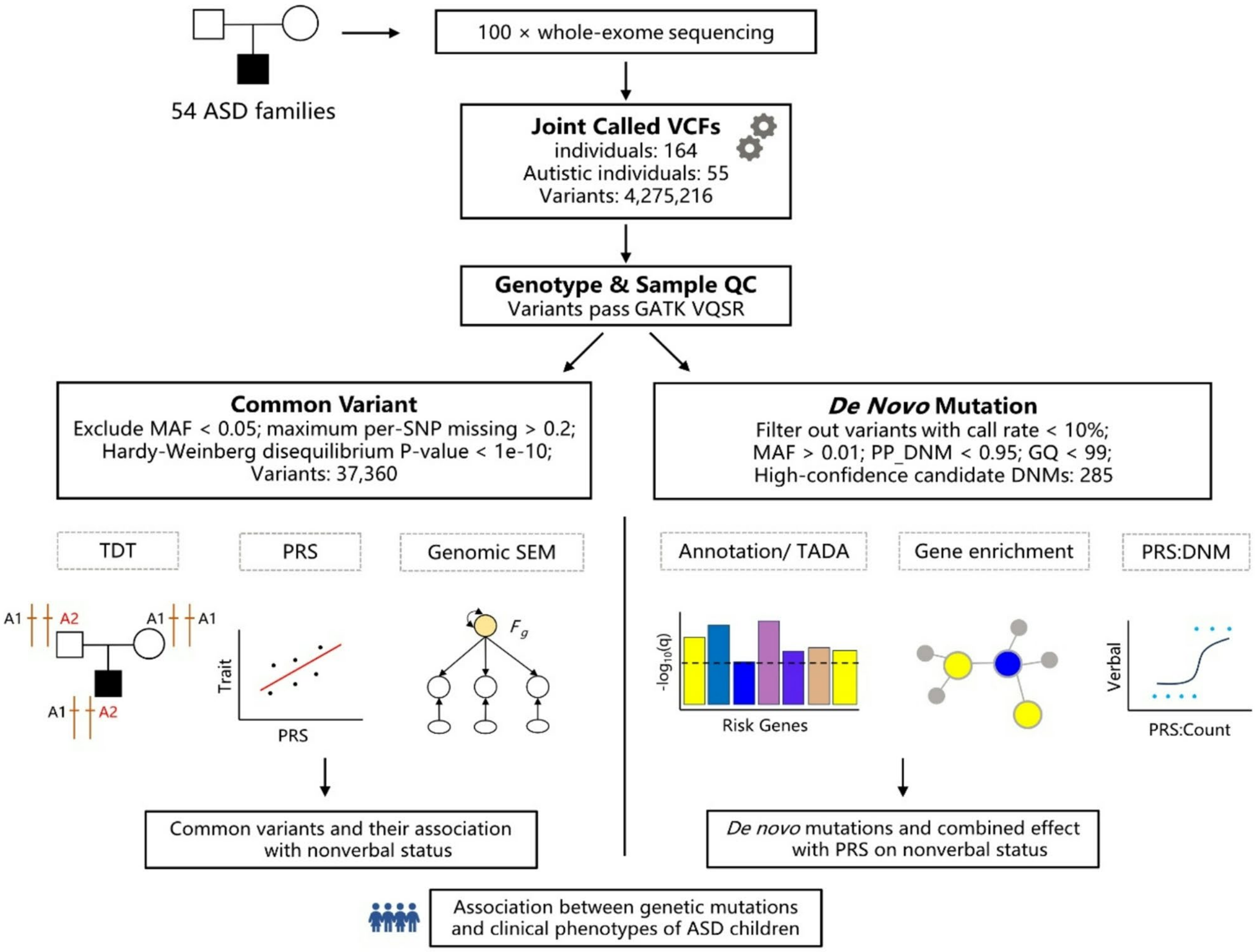
54 ASD families
100 × whole-exome sequencing
Joint Called VCFs
individuals: 164
Autistic individuals: 55
Variants: 4,275,216
Genotype & Sample QC
Variants pass GATK VQSR
Common Variant
Exclude MAF < 0.05; maximum per-SNP missing > 0.2;
Hardy-Weinberg disequilibrium P-value < 1e-10;
Variants: 37,360
De Novo Mutation
Filter out variants with call rate < 10%;
MAF > 0.01; PP_DNM < 0.95; GQ < 99;
High-confidence candidate DNMs: 285
TDT
PRS
Genomic SEM
Annotation/ TADA
Gene enrichment
PRS:DNM
A1
A2
A1
A1
A1
A2
Trait
PRS
F_g
-log10(q)
Risk Genes
Verbal
PRS:Count
Common variants and their association with nonverbal status
De novo mutations and combined effect with PRS on nonverbal status
Association between genetic mutations and clinical phenotypes of ASD children

Fig. 1 Overview diagram of study analyses. VCF = Variant Call Format, QC = Quality Control, VQSR = Variant Quality Score Recalibration, PRS = polygenic risk score, SEM = structural equation model

extensive clinical evaluations including relevant demographic data, neurological assessments, developmental quotient (DQ) assessment by Gesell Development Diagnosis Scale (GDDS) [22, 23]/ intelligence quotient (IQ) assessment by Chinese Wechsler Intelligence Scale for children- IV Version (CWISC-IV) [24, 25] or by Chinese Wechsler Young Children Scale of Intelligence-IV Version (CWYCSI-IV) [26, 27, 28]. The Childhood Autism Rating Scale (CARS) was applied to evaluate the severity of autism [29]. The patients with less than five spontaneous functional words were defined as non-verbal autistic patients [8, 19]. An autistic patient who had a language degeneration, was defined when he/she had a normal development for the first one to two years of life, followed by a loss of previously acquired language skills [30, 31]. Common evaluations in patients including hematological, biochemical, and metabolic tests and brain MRI, showed no abnormalities. Metabolic diseases, intoxications and any diagnosis of syndromic autism (X-Fragile, Tuberosis Sclerosis, Angelman Syndrome et al.) were ruled out. The study was approved by the Clinical Research Ethics Committee of Guangzhou Women's and Children's Medical Center, and informed consent for participation was obtained from their parents/guardians. Blood samples of the patients and their parents were obtained from who gave informed consent.

Data quality control and common variant calling

Genomic DNA was extracted from peripheral blood using QIAamp® Blood Mini Kit (Qiagen, Germany), and fragmentized by Covaris ultrasonicator followed by library preparation. High throughput sequencing was then performed on Illumina Novaseq6000 platform with 150 bp paired-end reads.

WES was conducted on all samples with a coverage depth of ≥100×reads. Initial quality filtering of raw sequencing reads was performed using Trimmomatic-0.39, after which the high-quality reads were aligned to the human reference genome (hg38) using BWA. Variant calling followed the Genome

Analysis Toolkit (GATK) best practices pipeline, including key steps like BaseRecalibrator, HaplotypeCaller, VariantRecalibrator [32]. A total of 4,275,216 variants were initially identified. For genotype quality control, we filtered genotypes with a genotype quality score below 25 and removed the variants on the X and Y chromosomes. The accuracy of variant calls was estimated using the GATK VQSR approach with GATK version 4.3.0.0. Low-complexity regions and variants that failed VQSR were removed, resulting in a high-quality set of 3,824,296 unique variants. After variant calling, ANNOVAR was applied to annotate variants' related genes and reveal the potential functions [33]. Meanwhile, we leveraged an ASD related project *ClinVar*[34] to identify variants that were discovered to be related to ASD.

To define a set of common exonic variants, we used PLINK v1.9 [35] 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}. Applying the scaling described above, we leaved 37,360 variants passed the filters. Since our samples consist entirely of trios and 1 quad, family-based Transmission Disequilibrium Test (TDT) was utilized to detect the significant loci using PLINK with the parameter: --tdt [35, 36].

We performed polygenic Transmission Disequilibrium Test (pTDT) [37] using the ASD-associated PRS in two groups: verbal and non-verbal. pTDT deviation distribution indicates the over-transmission of polygenic risk from parents to probands. P values denote the probability that the mean of the pTDT deviation distribution is 0 (two-sided, one-sample t test), indicating the significance of the over-transmission of PRS from parents to probands.

De novo detection, annotation and filtering

DNMs were called using GATK PossibleDeNovo [38, 39] based on the hg38 reference genome, as these mutations were found only in the children of our trios and 1 quad. For DNM quality control, variants with call rate < 10% or a Hardy-Weinberg equilibrium p value less than 1×10^{-12} were excluded, leaving 74,436 unique variants.

Furthermore, to minimize false positives, DeNovoGear v1.1.1 [40] was further employed following the initial filtering procedures. The GATK VCF file for each family was used as input to DeNovoGear. Variants that failed GATK FILTER, had a non-reference parental genotype, or had filtering allele frequency > 1% in the ExAC data (https://gnomad.broadinstitute.org/downloads) were removed. Putative DNMs were defined as those with PP_DNM > 0.95 and GQ 99. In the end, 285 DNMs were identified and used in the following study.

With these DNMs, we employed Ensembl Variant Effect Predictor (VEP) to define the High-impact, Moderate-impact, and Loss of Function (LoF) categories [41]. Possible-damaging missense DNMs were defined as the variants predicted to be damaging by at least two of the twenty prediction algorithms: SIFT, SIFT4G, Polyphen2 HDIV, Polyphen2 HVAR, LRT, Mutation Taster, Mutation Assessor, FATHMM, PROVEAN, MetaSVM, MetaLR, MetaRNN, PrimateAI, DEOGEN2, BayesDel addAF, BayesDel noAF, ClinPred, LIST.S2, fathmm.MKL, fathmm.XF annotated by dbNSFP4.2a.

Polygenic risk scores (PRS)

A PRS was calculated to estimate the genetic predisposition for ASD in the probands, using summary statistic from the meta-analysis of ASD by the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) and the Psychiatric Genomics Consortium (PGC) released in November 2017[10]. Considering the ethnic differences between discovery and target dataset, Polygenic Risk Score-Continuous Shrinkage (PRS-CS) was applied to eliminate this discrepancy and improve cross-population polygenic prediction [42]. Additionally, summary statistics for four language-related abilities [43], educational attainment [44], schizophrenia [45], and bipolar disorder [46] were included to rigorously validate disease specificity (Tabel S3). Associations between condition-specific PRS and phenotypes were analyzed using linear models, adjusting for covariates such as age and sex.

Gene set enrichment analysis

Gene set enrichment analysis was employed using the R packages clusterProfiler [47] to explore functional annotations based on Gene Ontology (GO). The core function *enrichGO* was utilized to provide species-specific GO annotation, relying on human genome-wide annotation package (org.Hs.eg.db) [48] released by the Bioconductor project. This approach allowed for the identification of biologically meaningful gene sets enriched in the dataset, offering insights into the functional pathways potentially involved in the observed phenotypes. In addition, we performed further pathway enrichment analysis using the Enrichr [49] web tool.

Genomic structural equation modeling

To investigate the genetic causal relationships between ASD and three language-related traits, we used Genomic Structural Equation Modeling (GenomicSEM, version 0.0.5c) [50] to model their joint genetic architecture. The common factor model was applied to explore how a common factor, defined by genetic indicators, is regressed onto a single nucleotide polymorphism (SNP). This allows for estimation of a set of summary statistics for the common factor that represent the SNP effects on the common factor. To better investigate the causality between ASD and language-related traits, we adopted a powerful extension - multivariate GWAS - of Genomic SEM to run user specified models that include SNP effects.

Transmission and *De Novo*Association Test (TADA)

The Transmission and *De Novo* Association Test (TADA) is a Bayesian model used to estimate the risk associated with genetic mutations based on WES data [15, 51]. TADA requires a mutational model which accounts for gene size and sequence composition to predict the expected number of mutations per gene, given the sample size. We utilized TADA to evaluate two categories of *de novo* variation, namely missense variants and frameshift variants. The genes with a Bayes Factor greater than 100 were considered as significant results under the TADA model.

Results

Clinical data

Amount to 55 ASD patients (53 trios and 1 quad) including two twins from 54 families were collected in this study. Of 55 patients, 44 were male and 11 were female, with a median age of 3.00 (range from 1.67 to 11.50) years at first diagnosis. Based on the DSM-5, ADIR and ADOS, all 55 patients were diagnosed with ASD. All 55 patients had CARS scores above 30, indicating clinically significant ASD. Among them, 94.7% (52/55) were classified as having mild to moderate ASD, and 5.3% (3/55) as severe. Children with DQ/IQ scores below 70 were considered to have developmental delay or intellectual disability of varying severity. Using a DQ/IQ cutoff of 70, 21.8% (12/55) were categorized as high-functioning, while 78.2% (43/55) as low-functioning. Additionally, 63.6% (35/55) were nonverbal, and 7.3% (4/55) showed signs of language or social regression (Table S1). We observed a modest but statistically significant correlation between nonverbal status and autism severity (CARS) in our cohort (correlation coefficient = 0.31, p value = 0.020) (Table S2).

Common variants are associated with language ability

To identify the significant loci associated with ASD in our study, we applied the family-based TDT. One SNP, rs1944180, surpassed exome-wide significance ($0.05/37{,}360 = 1.34 \times 10^{-6}$), a novel variant, located in the intronic region of *CNTN5* on chromosome 11 which has not been previously reported in *ClinVar* (Fig. 2A, Table S4). However, no significant correlation ($p > 0.05$) was found between *CNTN5* and language ability, suggesting a polygenic regulatory effect on language ability in ASD. Based on the PGC ASD summary statistics, we calculated PRS and conducted the pTDT analysis to explore the relationship between ASD associated PRS and non-verbal status. Although the pTDT did not show significant over-transmission of PRS across the three groups, including all ASD children, non-verbal ASD children, and verbal ASD children, it revealed a trend of increased genetic risk transmission in the non-verbal group (Fig. 2B). Importantly, higher PRS was significantly associated with non-verbal status after adjusting for age, sex, severity of autism and DQ/IQ scores (Nagelkerke's $R^2 = 0.421$, $p = 0.034$) (Fig. 2C). To further assess the specificity of this finding, we calculated PRS calculated based on summary statistics from related disorders and language traits, including bipolar disorder, schizophrenia, educational attainment and three types of language ability (non-word reading, word reading, spelling) (Fig. 2D). None of these disorders or traits, except for ASD, exhibited a significant correlation between PRS and the non-verbal status in probands. Furthermore, we conducted GO pathway enrichment analysis using the SNPs underlying the PRS associated with the non-verbal status. The result showed significant enrichment in pathways related to transporter, neuronal and synaptic functions (Fig. 2E). In addition, non-verbal status was modestly but significantly associated with severity of autism (CARS) (correlation coefficient = 0.31, p value = 0.020) and cognitive ability (DQ/IQ) (correlation coefficient = -0.28, p value = 0.036) (Table S2), suggesting that individuals with severe ASD may be more likely to exhibit language impairment.

SEM infers new insight into ASD with Language impairment

To better understand the genetic overlap and relationship between ASD and non-verbal status, we applied SEM with ASD PGC summary statistic along with three language ability summary statistics (Fig. 3). ASD PGC summary statistics were filtered by the SNPs significantly associated with non-verbal status, resulting in a total of 617 SNPs (Table S5). This method modeled the shared genetic architecture of ASD with language-related traits such as word reading, non-word reading and spelling. Exploratory factor models with 2 factors were fitted to the data. The two-factor model explained the majority of the variance and standard error in parentheses. Multivariate GWAS followed the model specification and calculated the significance of SNPs affecting the factor. In the final model, the first factor explained 66% of the variance of ASD incidence, while the second factor explained 34%, 38%, and 35% of the variance in language-related traits, respectively. All results met the exome-wide significance threshold, indicating a genetic overlap between ASD and language-related ability. On the other hand, language-related traits were found to be highly consistent. Although ASD shows genetic overlaps with these traits,

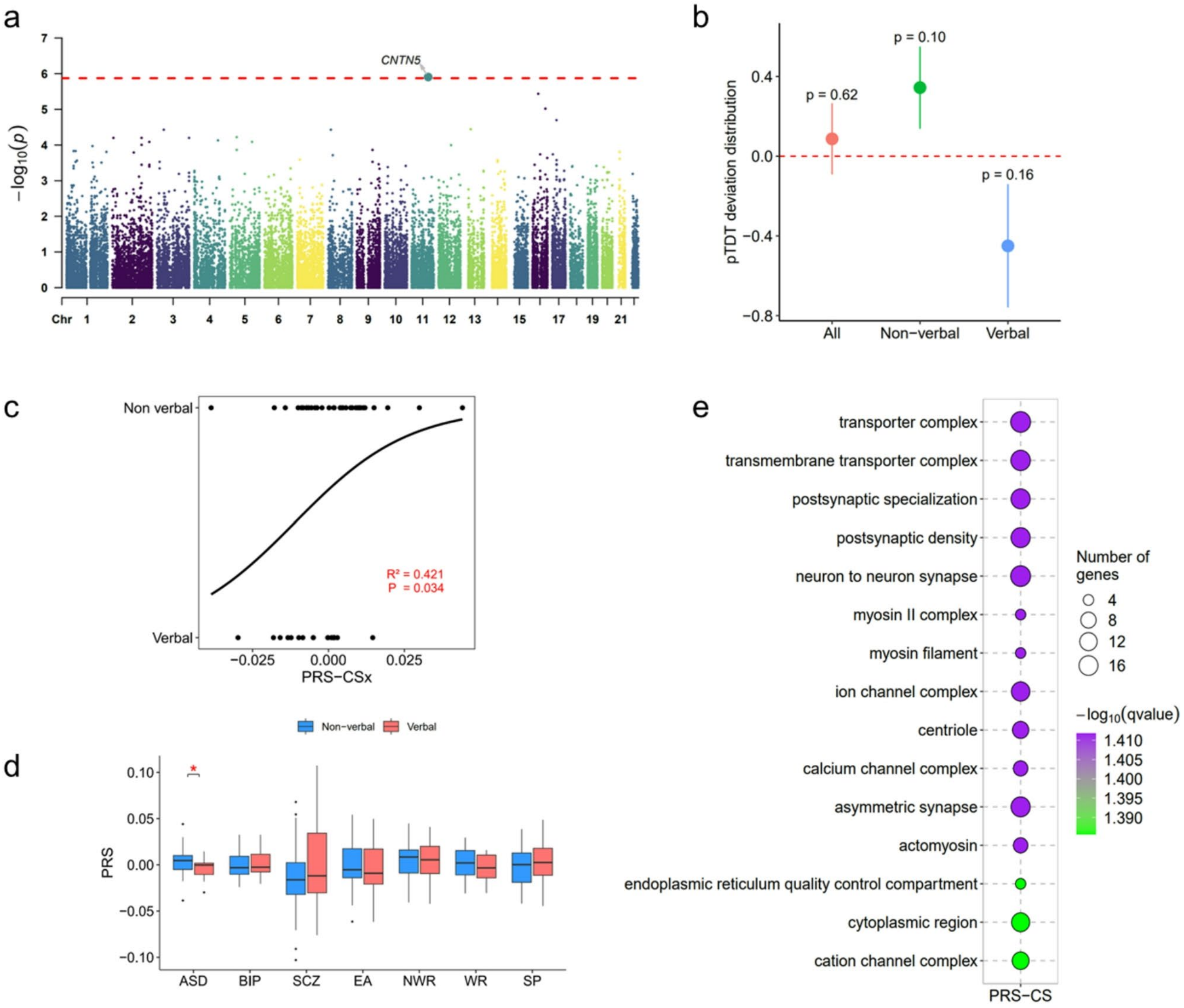

Fig. 2 The identification and influence of common variants. **(a)** Manhattan plot of TDT results. The red dashed line represents the exome-wide significance (p-value < 1.34×10^{-6}). **(b)** Polygenic transmission disequilibrium test (pTDT) in all ASD children, non-verbal ASD children and verbal ASD children. The y-axis represents the deviation distribution of the over-transmitted inheritance. P values denote the probability that the mean of the pTDT deviation distribution is 0 (two-sided, one-sample t test) and indicates the significance of the over-transmission of PRS from parents to probands. **(c)** Correlation between PRS and verbal ability with Nagelkerke's $R^2 = 0.421$ and P value = 0.034, adjusting for age, sex, severity of autism (CARS) and scores of DQ/IQ. **(d)** The significance test of PRS with and without verbal ability across discovery datasets of different diseases or traits. For these seven traits, only when ASD summary statistic was used as the discovery dataset did the comparison between those with and without verbal ability show significant differences, adjusted for age, sex, severity of autism (CARS) and scores of DQ/IQ. (Table S6). n(Verbal) = 20, n(Non-verbal) = 35. ASD = autism spectrum disorder, BIP = bipolar disorder, SCZ = schizophrenia, EA = educational attainment, NWR = non-word reading, WR = word reading, SP = spelling. **(e)** GO enrichment bubble plot of common variants significantly associated with non-verbal status. All displayed pathways have a q-value less than 0.05

the model indicates that they also have unique unshared components, aligning with genetic correlation estimates lower than one. Multivariate GWAS results identified 80 SNPs exceeding the significance threshold of 0.05, with two SNPs (rs1247761 and rs2524290) achieving exome-wide significance based on Bonferroni-corrected threshold for the tested SNPs ($0.05/617 = 8.10 \times 10^{-5}$) (Table S7). Moreover, the multivariate GWAS results highlighted a complex interplay between the two key factors, with two high significant SNPs influencing the second factor through the first. The high correlation between factor 1 (ASD) and factor 2 (language-related traits) suggested a directional effect, where the first factor played a more significant role in shaping the second. Intriguingly, this pattern was not reciprocated, with factor 2 showing a less significant impact on factor 1. This observation supports a unidirectional causal relationship, indicating ASD is more likely to be the cause of language impairments rather than be a consequence of them.

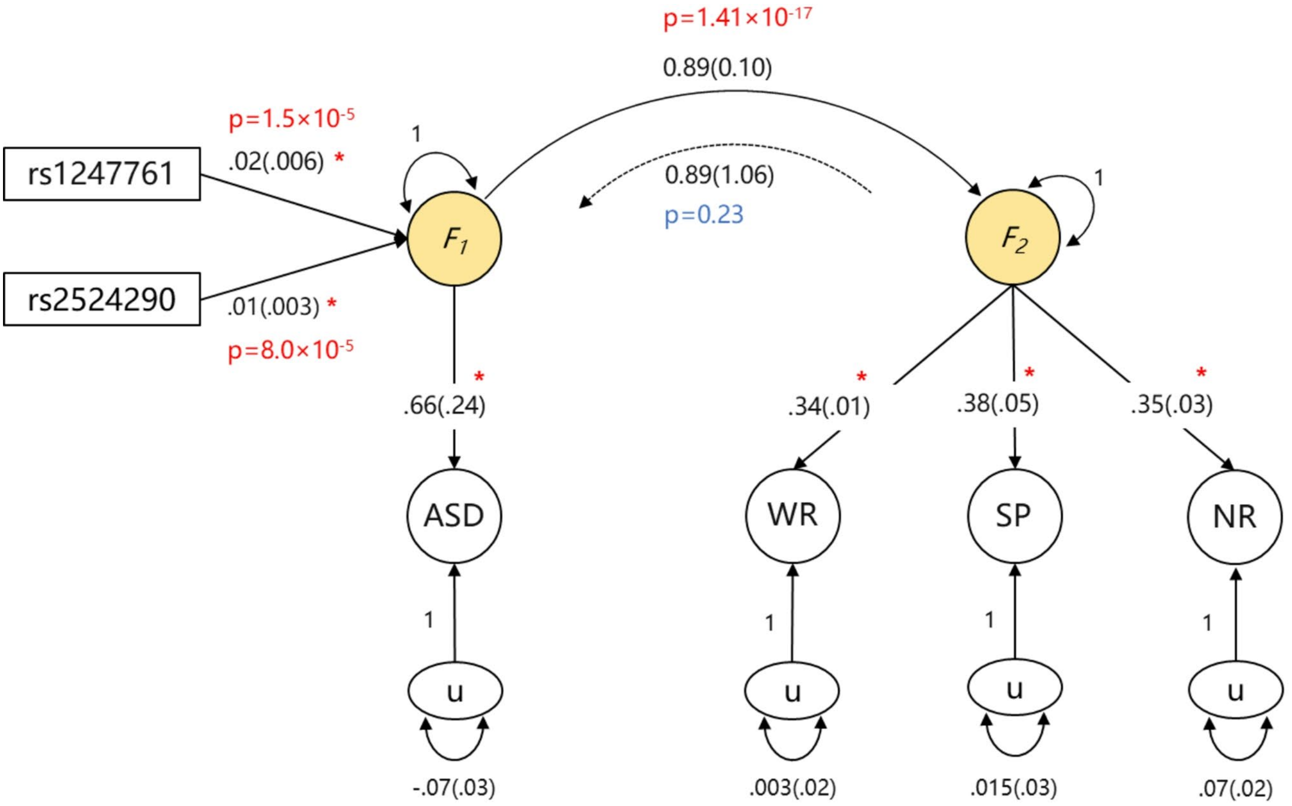

Fig. 3 Structural diagram of GenomicSEM. Two factor model fitted to ASD and language-related summary statistics. Solid and dashed paths represent factor loadings with $P < 0.001$ and $P > 0.001$, respectively. Standardized factor loadings are shown, with SE in parentheses. The capital letters F1 and F2 represent the genetic variables; the u variables represent the residual genetic variance not explained by the models. Two SNPs are significant associated with factor 1 with exome-wide significant threshold. ASD = autism spectrum disorder; WR = word reading; SP = spelling; NR = non-word reading

Identification of *de novo*mutation in ASD probands

After variants filtering described in Method, we discovered 285 DNMs which was distinct between parents and probands. We classified DNMs into three categories, High-impact, Moderate-impact, and LoF. Typically, the High-impact variants lead to the truncation of protein products, and we found 20 High-impact DNMs which were frameshift variants and 11 Moderate-impact disruptive variants (Table S8). Moderate-impact variants have changes, but didn't get truncating the protein sequence, like missense SNVs and disruptive in-frame variants. Among 20 High- and 11 Moderate-impact variants, we discovered 19 LoF variants, and 4 possible damaging variants defined by at least two of twenty prediction algorithms (Table S9).

To discover the risk effect of gene discovered via DNMs, TADA model [15, 51] was applied to integrate protein-truncating and missense variants, stratifying autosomal genes by FDR for association. 5 genes were defined by TADA model, revealing the high risk of candidate genes (Fig. 4A, Table S10). To further investigate the functional relevance of these genes, we performed enrichment analysis, which revealed that the implicated DNMs were significantly associated with biological pathways related to O-linked glycosylation (Fig. 4B-D).

Association between ASD related DNMs and non-verbal status

To evaluate the impact of *de novo* mutation on the non-verbal status, we computed the associated between the number of DNMs from TADA risk genes and non-verbal status in ASD patients. However, no significant result was found ($p = 0.12$). We then examined the interaction effects of common variants and DNMs on verbal status. We employed the PRS and count of TADA DNMs to represent this interaction using a logistic regression model. Adjusting for age and sex, the interaction effects were found to have a marginally significant impact on non-verbal status ($p = 0.124$, $z = 1.54$). When further adjusting for age, sex, CARS, and DQ/IQ, the p-value reached statistical significance at 0.038, accompanied by a z-value of 2.07. This finding suggests that the interaction between common and *de novo* polygenic mutation is significantly associated with non-verbal status.

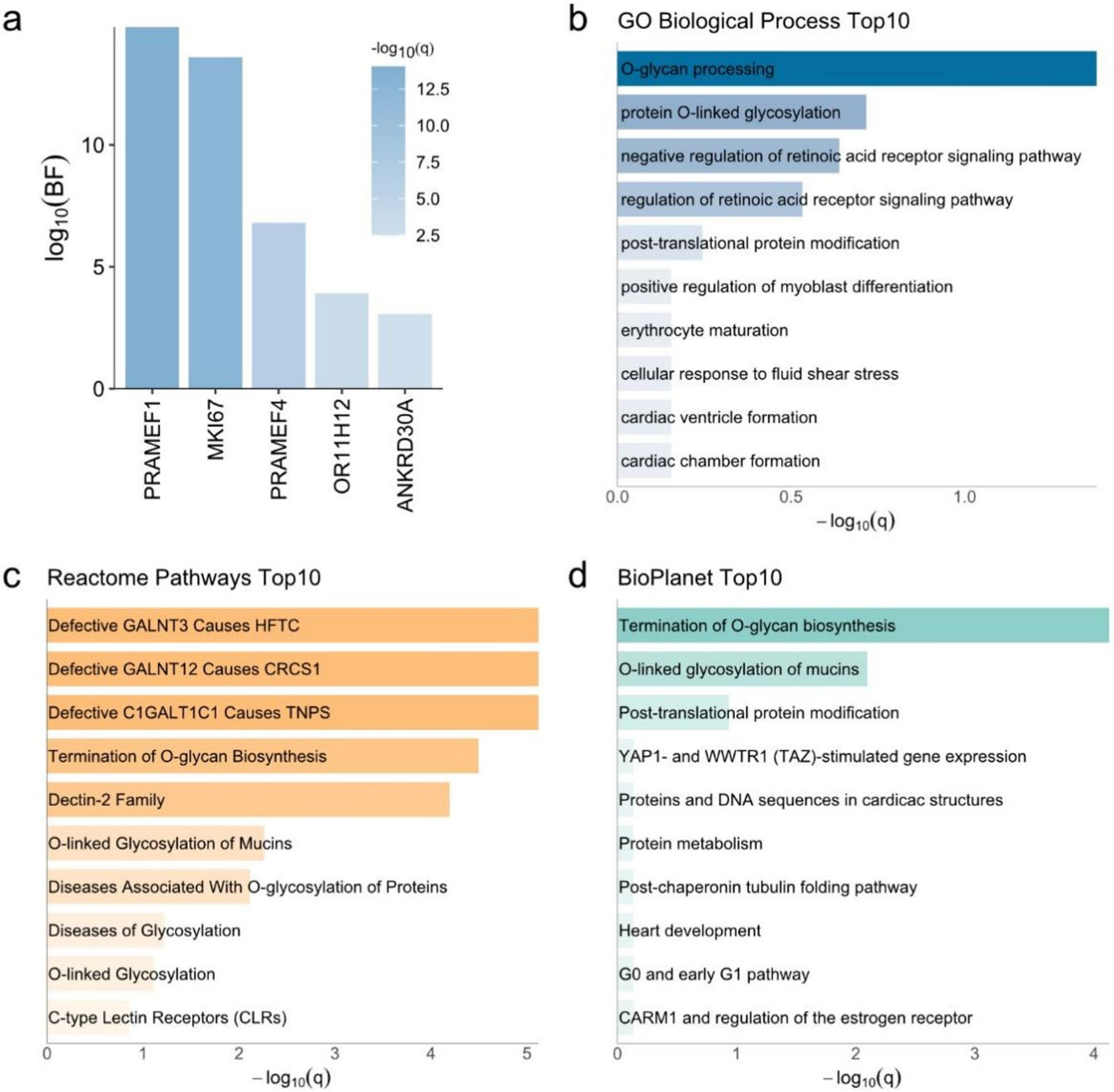

Fig. 4 Identification and enrichment analysis of DNMs. **(a)** Evidence supporting the ASD risk genes using TADA with FDR p-value < 0.05. **(b)** GO enrichment analysis of DNMs. **(c)** Reactome enrichment analysis of DNMs. **(d)** BioPlanet enrichment analysis of DNMs

Discussion

In this study, we investigated the genetic factors contributing to non-verbal status in ASD through a comprehensive genomic approach. Our findings revealed significant associations between common and *de novo* mutations and language ability in individuals with ASD. Notably, we identified a common variant in *CNTN5* significantly associated with non-verbal status through family-based Transmission Disequilibrium Testing. In addition, we found strong correlations between polygenic risk scores and language abilities, highlighting the cumulative influence of common variants on verbal ability. Using SEM, we discovered two causal SNPs, rs1247761 and rs2524290, linking both ASD and language traits. We also identified *de novo* mutations associated with glycosylation-related pathways, as well as significant interactions between these mutations and common variants impacting non-verbal status in ASD. These results provide a novel understanding of the genetic underpinnings of language deficits in ASD, particularly in relation to non-verbal status.

Our results align with previous studies indicating the polygenic nature of ASD and the involvement of common genetic variants in its clinical presentation [3]. The identification of *CNTN5* as associated with ASD extends previous research on this gene's role in neural development [52, 53]. Prior studies on polygenic risk have primarily focused on general ASD susceptibility [16]; however, both population-based PRS and polygenic TDT directly connects these common variants to language impairments, offering a more refined understanding of their contributions to specific nonverbal status within ASD.

The application of structural equation modeling in our study to uncover causal relationships between genetic variants, ASD, and language traits is novel. We identified two key SNPs, rs1247761 and rs2524290, which play significant roles in both ASD and language deficits. rs1247761, as an intron variant of *KCNMA1*, has been linked to neurodevelopment [54, 55], specifically in potassium channel activity and neuronal excitability, which is critical for both ASD and language processing.

rs2524290 is an intron variant of *RAB3IL1*, which can promote the exchange of GDP to GTP and encode guanine nucleotide exchange factor (GEF) which may activate RAB3A, a GTPase that regulates synaptic vesicle exocytosis [56]. By demonstrating the pleiotropic effects of these SNPs, our study not only deepens the understanding of ASD's genetic architecture but also opens up new possibilities for therapeutic interventions. Targeting these shared genetic factors could provide an integrated treatment strategy to address both social and communication deficits in ASD.

In terms of *de novo* mutations, our findings contribute to the growing evidence that spontaneous genetic changes play a critical role in more severe ASD phenotypes. The association between *de novo* mutations and O-linked glycosylation pathways adds a new dimension to understanding the mechanisms implicated in ASD and suggests possible links between glycosylation processes and ASD [57].

Further evidence of the effect of DNMs was observed in our correlation analysis of DNMs and non-verbal status. Though no significant result was found ($p = 0.12$) between the number of DNMs from TADA risk genes and non-verbal status in ASD patients, the interaction effects of common variants and DNMs on non-verbal status was found to be significant after adjusting ($p = 0.038$). We observed that the interplay between common and *de novo* polygenic mutations is significantly linked to non-verbal status. These results not only align with prior research conclusions but also underscores the role of *de novo* mutations in ASD, which may play an indispensable role in the pathogenesis of ASD [3, 18].

By identifying key genetic variants linked to non-verbal status, we open the door for more precise genetic screening methods that could be employed in early diagnosis, particularly for individuals at risk of severe verbal or communication impairments [3, 18]. These genetic markers may also serve as valuable targets for personalized therapeutic interventions aimed at improving language skills, especially in non-verbal or minimally verbal individuals [8]. Furthermore, the discovery of *de novo* mutations related to glycosylation pathways suggests that interventions targeting glycosylation may have meaningful downstream effects on outcomes in ASD [57], providing a new direction for clinical practice.

Despite these significant findings, there are limitations to our study that should be acknowledged. First, our sample size, though informative, is relatively small, which may limit the generalizability of our results. While the familial and *de novo* mutations identified provide valuable insights, larger cohorts with more diverse populations will be necessary to validate our findings and uncover additional relevant genetic factors. Additionally, while we identified significant associations between genetic variants and language impairments, the precise biological mechanisms by which these variants influence neural circuits and behavior remain unclear. Future functional studies will be needed to elucidate how these genetic variants disrupt language-related pathways in the brain.

Our study opens several new avenues for future research. First, expanding the sample size to include more diverse populations could help validate our findings and uncover additional genetic factors related to non-verbal status in ASD. Second, longitudinal or interventional studies that track the developmental trajectory of language abilities in individuals with known genetic variants could provide deeper insights into how these mutations influence language acquisition over time. Functional studies examining how *CNTN5* and other identified variants affect neural connectivity could also clarify their roles in ASD phenotypes. Moreover, future research could investigate gene-environment interactions to understand how epigenetic or environmental factors might exacerbate or mitigate the genetic risk for language deficits.

In conclusion, our study significantly advances the understanding of the genetic basis of non-verbal status in ASD, highlighting the roles of both common variants and *de novo* mutations. By identifying key genetic factors associated with language impairments, we provide a foundation for future diagnostic and therapeutic developments tailored to individuals with severe verbal deficits. These findings have the potential to shift the research paradigm by linking glycosylation processes and communication deficits at the genetic level, ultimately contributing to more effective, personalized treatments for individuals on the autism spectrum.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12993-025-00278-x.

Supplementary Material 1

Acknowledgements
We thank the Clinical Biological Resource Bank of Guangzhou Women and Children's Medical Center for providing the clinical samples.

Author contributions
Huan Liu, Shenghan Wang and Binbin Cao had full access to all data and are responsible for its integrity. Shenghan Wang was responsible for the accuracy of data analysis. Jijun Zhu, Pan Li, Shunjie Zhang and Zhongting Huang contributed to the pipeline development and method evaluation. Data collection and preprocessing was performed by Shenghan Wang, Huan Liu, Binbin Cao, Zhifang Huang, and Fuqiang Cai. Xian Liu, Jing Yu, Linzhuo Lv, Weixin Liu, Zhijian Song, Yuxin Liu, Ying Chen and Tao Pang contributed to acquisition, data analysis, or interpretation. Suhua Chang provided critical revisions to the manuscript. Huan Liu, Shenghan Wang and Binbin Cao contributed to drafting of the manuscript. Corresponding authors Junfang Chen and Wen-xiong Chen oversaw the study's conception, design, execution, and provided critical manuscript revisions. All authors have read and approved the final manuscript for publication.

Funding
This work was supported by a startup grant from the Greater Bay Area Institute of Precision Medicine (Guangzhou) to JC (Grant No: I0007), the National Social Science Foundation (Grant No: 32370639), Natural Science Foundation of Guangdong Province (Grant No: 2024A1515012116), the grants of Guangzhou Science and Technology plan City School (College) joint funding project (202102010232) and Guangdong Provincial Health Commission Suitable Technology Promotion Project (202303192222167450) to CWX, and partly supported by the major Scientific and Technological Projects of Brain Science and Brain-like Research of Guangzhou (202007030002), and Brain Science and Brain-like Research major project of Ministry of Science and Technology of China (2021ZD0200522-07).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This study was approved by the Clinical Research Ethics Committee of Guangzhou Women's and Children's Medical Center. All procedures performed in this study involving human participants were conducted in accordance with the ethical standards of the institutional and/or national research committee.

Competing interests
The authors declare no competing interests.

Received: 13 November 2024 / Accepted: 16 May 2025
Published online: 07 June 2025

References

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5 (Vol. 5). American psychiatric association Washington, DC vol. 5. (2013).
2. Penner M, et al. Concordance of diagnosis of autism spectrum disorder made by pediatricians vs a multidisciplinary specialist team. JAMA Netw Open. 2023;6:E2252879.
3. Matilde Cirnigliaro, Chang TS, Arteaga SA, Geschwind DH. The contributions of rare inherited and polygenic risk to ASD in multiplex families. *Proc. Natl. Acad. Sci.* 120, e2215632120 (2023).
4. Kieling C, et al. Worldwide prevalence and disability from mental disorders across childhood and adolescence: evidence from the global burden of disease study. JAMA Psychiatry. 2024;81:347–56.
5. Zeidan J, et al. Global prevalence of autism: A systematic review update. Autism Res. 2022;15:778–90.
6. Lord C, et al. The lancet commission on the future of care and clinical research in autism. Lancet. 2021;399:271–334.
7. Liu X, et al. Prevalence of epilepsy in autism spectrum disorders: A systematic review and meta-analysis. Autism. 2022;26:33–50.
8. Chen WX, et al. De Novo mutations within metabolism networks of amino acid/protein/energy in Chinese autistic children with intellectual disability. Hum Genomics. 2022;16:1–14.
9. Gaugler T, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46:881–5.
10. Grove J, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51:431–44.
11. Wilfert AB, et al. Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nat Genet. 2021;53:1125–34.
12. Fu JM, et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat Genet. 2022;54:1320–31.
13. Tick B, Bolton P, Happé F, Rutter M, Rijsdijk F. Heritability of autism spectrum disorders: A meta-analysis of twin studies. J Child Psychol Psychiatry Allied Discip. 2016;57:585–95.
14. Iossifov I, et al. The contribution of de Novo coding mutations to autism spectrum disorder. Nature. 2014;515:216–21.
15. Satterstrom FK, et al. Large-Scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020;180:568–e58423.
16. Trost B, et al. Genomic architecture of autism from comprehensive whole-genome sequence annotation. Cell. 2022;185:4409–e442718.
17. Zhou X, et al. Integrating de Novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat Genet. 2022;54:1305–19.
18. Yousaf A et al. Quantitative genome-wide association study of six phenotypic subdomains identifies novel genome-wide significant variants in autism spectrum disorder. Transl Psychiatry 10, 215 (2020).
19. Chiang HM. Expressive communication of children with autism: the use of challenging behaviour. J Intellect Disabil Res. 2008;52:966–72.
20. Lord C, Rutter M, Couteur A. Le. Autism diagnostic interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24:659–85.
21. Lord C et al. The autism diagnostic observation Schedule–Generic: A standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 30, 205–223 (2000).
22. Ping L, et al. Children's health adverse associations of both prenatal and postnatal exposure to organophosphorous pesticides with infant neurodevelopment. Environ Health Perspect. 2016;124:1637–43.
23. Zhou L, et al. Prenatal maternal stress in relation to the effects of prenatal lead exposure on toddler cognitive development. Neurotoxicology. 2017;59:71–8.
24. Yang P, et al. Wechsler intelligence scale for children 4th edition-Chinese version index scores in Taiwanese children with attention-deficit/hyperactivity disorder. Psychiatry Clin Neurosci. 2013;67:83–91.
25. Guo J, et al. Prenatal exposure to mixture of heavy metals, pesticides and phenols and IQ in children at 7 years of age: the SMBCS study. Environ Int. 2020;139:105692.
26. Bin X et al. Children's intelligence quotient following general anesthesia for dental care: a clinical observation by Chinese Wechsler young children scale of intelligence. Journal of Peking University(Health Sciences). 2016;48(2): 336–340.
27. Mikkelsen TB, Osler M, Olsen SF. Validity of protein, retinol, folic acid and n–3 fatty acid intakes estimated from the food-frequency questionnaire used in the Danish National birth cohort. Public Health Nutr. 2006;9:771–8.
28. Huang C, Martorell R, Ren A, Li Z. Cognition and behavioural development in early childhood: the role of birth weight and postnatal growth. Int J Epidemiol. 2013;42:160–71.
29. Rellini E, Tortolani D, Trillo S, Carbone S, Montecchi F. Childhood autism rating scale (CARS) and autism behavior checklist (ABC) correspondence and conflicts with DSM-IV criteria in diagnosis of autism. J Autism Dev Disord. 2004;34:703–8.

30. Pickles A, et al. Predictors of Language regression and its association with subsequent communication development in children with autism. J Child Psychol Psychiatry Allied Discip. 2022;63:1243–51.
31. Ozonoff S, Iosif AM. Changing conceptualizations of regression: what prospective studies reveal about the onset of autism spectrum disorder. Neurosci Biobehav Rev. 2019;100:296–304.
32. Ryan P et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxIV* 1–22 (2018).
33. Wang K, Li M, Hakonarson HANNOVAR. Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:1–7.
34. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:980–5.
35. Purcell S, et al. A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81. 2007;PLINK:559–75.
36. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet. 1995;57:455–64.
37. Weiner DJ, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017;49:978–85.
38. Van der Auwera G, O'Connor B. Genomics in the cloud: using docker, GATK, and WDL in Terra(1st Edition). O'Reilly Media (2020).
39. Yuan B, et al. Identification of de Novo mutations in the Chinese autism spectrum disorder cohort via Whole– Exome sequencing unveils brain regions implicated in autism. Neurosci Bull. 2023. https://doi.org/10.1007/s12264-023-01037-6.
40. Ramu A, et al. DeNovoGear: de Novo indel and point mutation discovery and phasing. Nat Methods. 2013;10:985–7.
41. McLaren W, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:1–14.
42. Ruan Y, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet. 2022;54:573–80.
43. Eising E, et al. Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. PNAS. 2022;0:1–12.
44. Jiang L, Zheng Z, Fang H, Yang J. A generalized linear mixed model association tool for biobank-scale data. Nat Genet. 2021;53:1616–21.
45. Trubetskoy V, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022;604:502–8.
46. Mullins N, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021;53:817–29.
47. Wu T, et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation. 2021;2:100141.
48. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2. (2019).
49. Xie Z, et al. Gene set knowledge discovery with enrichr. Curr Protoc. 2021;1:1–51.
50. Grotzinger AD, et al. Genomic SEM provides insights into the multivariate genetic architecture of complex traits. Nat Hum Behav. 2019;3:513–25.
51. He X et al. Integrated model of de Novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet 9(8): e1003671, (2013).
52. Lionel AC, et al. Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD. Sci Transl Med. 2011;3:1–11.
53. Schmilovich Z, et al. Copy-number variants in the contactin-5 gene are a potential risk factor for autism spectrum disorder. Res Autism Spectr Disord. 2022;99:1–12.
54. Miller JP, Moldenhauer HJ, Keros S, Meredith AL. An emerging spectrum of variants and clinical features in KCNMA1-linked channelopathy. Channels. 2021;15:447–64.
55. Moldenhauer HJ, Dinsdale RL, Alvarez S, Fernández-Jaén A, Meredith AL. Effect of an autism-associated KCNMB2 variant, G124R, on BK channel properties. Curr Res Physiol. 2022;5:404–13.
56. Quevedo MF, et al. Grab recruitment by Rab27A-Rabphilin3a triggers Rab3A activation in human sperm exocytosis. Biochim Biophys Acta - Mol Cell Res. 2019;1866:612–22.
57. Pradeep P, Kang H, Lee B. Glycosylation and behavioral symptoms in neurological disorders. Transl Psychiatry 13, 154 (2023).

Publisher's note