frontiers | Frontiers in Neurology

TYPE Original Research
PUBLISHED 17 July 2023
DOI 10.3389/fneur.2023.1203375

OPEN ACCESS

EDITED BY
Madelyn Gillentine,
Seattle Children's Hospital, United States

REVIEWED BY
Yingxue Ding,
Capital Medical University, China
Priyadarsan Parida,
Gandhi Institute of Engineering and Technology University (GIET), India

*CORRESPONDENCE
Huitao Tang
✉ tanghuitao2021@126.com
Shufang Chen
✉ chenshufang189@126.com
Daojiang Shen
✉ zjyysdj@126.com

†These authors have contributed equally to this work

RECEIVED 10 April 2023
ACCEPTED 26 June 2023
PUBLISHED 17 July 2023

CITATION
Tang H, Liang J, Chai K, Gu H, Ye W, Cao P, Chen S and Shen D (2023) Artificial intelligence and bioinformatics analyze markers of children's transcriptional genome to predict autism spectrum disorder.
Front. Neurol. 14:1203375.
doi: 10.3389/fneur.2023.1203375

Artificial intelligence and bioinformatics analyze markers of children's transcriptional genome to predict autism spectrum disorder

Huitao Tang[1*†], Jiawei Liang[2†], Keping Chai[1], Huaqian Gu[1], Weiping Ye[1], Panlong Cao[1], Shufang Chen[1*] and Daojiang Shen[1*]

[1]Department of Pediatrics, Zhejiang Hospital, Hangzhou, China, [2]College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

Introduction: Autism spectrum disorder (ASD), characterized by difficulties in social interaction and communication as well as restricted interests and repetitive behaviors, is extremely challenging to diagnose in toddlers. Early diagnosis and intervention are crucial however.

Methods: In this study, we developed a machine learning classification model based on mRNA expression data from the peripheral blood of 128 toddlers with ASD and 126 controls. Differentially expressed genes (DEGs) between ASD and controls were identified.

Results: We identified genes such as UBE4B, SPATA2 and RBM3 as DEGs, mainly involved in immune-related pathways. 21 genes were screened as key biomarkers using LASSO regression, yielding an accuracy of 86%. A neural network model based on these 21 genes achieved an AUC of 0.88.

Discussion: Our findings suggest that the identified neurotransmitters and 21 immune-related biomarkers may facilitate the early diagnosis of ASD. The mRNA expression profile sheds light on the biological underpinnings of ASD in toddlers and potential biomarkers for early identification. Nevertheless, larger samples are needed to validate these biomarkers.

KEYWORDS

autistic spectrum disorder, biomarkers, RNA-Seq, neural network, LASSO

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by the impairment of social and communication skills and repetitive movements in early childhood (1). Genetic factors are associated with the susceptibility to and development of ASD, with an estimated heritability of 50–83% (2). To date, the pathological mechanisms and curative treatment of ASD have not been clarified (3, 4). Early diagnosis and interventions could significantly improve the life of ASD toddlers (5). Reportedly, ASD is difficult to diagnose in toddlers under 53 months of age (6).

The current methods are either expensive or subjective, and their application in the diagnosis of ASD is limited. Research on the use of electroencephalography (EEG) data from children with ASD to train neural networks for predicting ASD has yielded promising results. However, EEG signals are susceptible to ambient noise, and there may be significant individual differences, which pose practical limitations in their application (7). Similarly, studies utilizing infant functional magnetic resonance imaging (fMRI) data and machine

learning methods have shown potential for the classification of ASD. Nevertheless, fMRI scanning can be challenging for infants who require a calm and still environment, leading to poor image quality and motion artifacts (8). Additionally, genomic data analysis has identified genetic risk variants associated with ASD. However, single gene mutations alone are not sufficient for fully explaining the complexity of ASD (9). Therefore, further research is necessary to explore the potential of these approaches in accurately predicting ASD in children.

Identifying ASD biomarkers may help with the early diagnosis of ASD. However, the well-known neural system biomarkers of ASD are rarely applied due to the difficulty of sample collection. Compared to neural system tissue, peripheral blood is more readily available to screen for biomarkers, but identifying blood biomarkers for ASD and using them to diagnose ASD are two core issues that need to be resolved. The impetus herein was interrogating mRNA expression profiles of peripheral blood from ASD subjects and controls to obtain biomarkers that are amenable to ASD diagnostics. Expression microarray data from 128 ASD and 126 control toddlers were analyzed. Differential expression analysis revealed that 1,027 genes (adjusted $P < 0.05$) were dysregulated in ASD; the ingenuity pathway analysis of the top 200 genes identified immune response, neurotransmission, and cell proliferation pathways as enriched. The least absolute shrinkage and selection operator (LASSO) regression identified 21 candidate biomarkers, including *GDI1*, *HYAL3*, and *ANAPC7*. Binary logistic regression and neural network models were developed utilizing these 21 biomarkers, achieving a satisfactory accuracy of 86 and 88%, respectively. These models demonstrated the potential of the identified biomarkers for the early detection of ASD.

In aggregate, we adduced 21 candidate peripheral blood biomarkers related to immune functions, growth factors, and neurotransmitter functions in ASD. Our methodology and results adumbrate the value of biomolecular approaches coupled with machine learning for illuminating pathological mechanisms underlying ASD and developing diagnostic modalities. Although the findings are promising, further validation in larger, heterogeneous populations and comparison with prevailing diagnostics are necessary to determine their clinical utility.

2. Methods

2.1. Data acquisition and preprocessing

The data used in this study were obtained from the Gene Expression Omnibus (GEO) database in NCBI (Gene Expression Omnibus, http://www.ncbi.nlm.gov/geo), and the access number is GSE111175 (10, 11), GSE42133 (12, 13). The platform is Illumina HumanHT-12 V4.0 Expression BeadChip. Gene expression data of 128 ASD and 126 CON samples were identified. In addition, 38 developmental delay (27 language delay samples, 9 pervasive developmental disorder not otherwise specified samples, 1 socially and emotionally delayed sample, 1 global developmental delay sample) samples in GSE111175 were identified for specificity detection of the machine learning model. Principal component analysis (PCA) was performed to visualize the batch effect between the two datasets. The batch effect was eliminated using the SVA package based on the R language (14). Expression matrix probes without corresponding annotation were removed. Finally, we obtained normalized and batch effect-removed RNA expression data, which contained 254 samples and 24,698 genes.

2.2. Diagnostic criteria and procedures for ASD

The diagnosis of ASD was determined using the datasets GSE111175 and GSE42133, which also provide the age, the Autism Diagnostic Observation Schedule (Module T, 1, or 2) (ADOS) scores (15), the Mullen Scales of Early Learning (MSEL) (16) scores of the ASD and CON samples, and the Vineland Adaptive Behavior Scales (17) scores of the ASD and CON samples. A *t*-test was performed to calculate the *p*-values of scores between the two groups.

2.3. Clustering methods

A cohort of 128 toddlers diagnosed with ASD was assessed using a dataset consisting of 12 clinical features, including scores from the ADOS, MSEL, and Vineland. In particular, ADOS scores encompassed communication, sociability, circumscribed, and repetitious behaviors, while MSEL scores encompassed precocious learning aggregate and fine motor skills. The Vineland scores encompassed activities of daily living and motor adaptive functioning.

To analyze the data, both hierarchical and K-means clustering methods were employed, with the number of clusters set to two for each technique. The K-means algorithm utilizes a random state of 42. By calculating the spatial distances among the 12 indices across the entire cohort, the toddlers were effectively stratified into two distinct clusters, providing valuable insights into the heterogeneity of ASD profiles.

2.4. Identification of differential expression genes

The Wilcox test in the R-package "limma" was performed to screen for differentially expressed genes (DEGs) (18). The Benjamini–Hochberg method, which is a multiple testing method, was performed to calculate the adjusted *p*-values and reduce false-positive DEGs. The heatmap and scatter plot of DEGs were plotted using the R language packages "pheatmap" and "ggpubr." A total of

Abbreviations: ASD, Autism spectrum disorder; GEO, Gene Expression Omnibus; CON, control; DEGs, differentially expressed genes; AUC, accuracy; ROC, receiver-operating characteristic; LASSO, least absolute shrinkage and selection operator; IPA, ingenuity pathway analysis; GSEA, gene set enrichment analysis; PCA, principal component analysis; ADOS, Autism Diagnostic Observations Schedule; MSEL, Mullen Scales of Early Learning; fMRI, functional magnetic resonance imaging; EEG, electroencephalography.

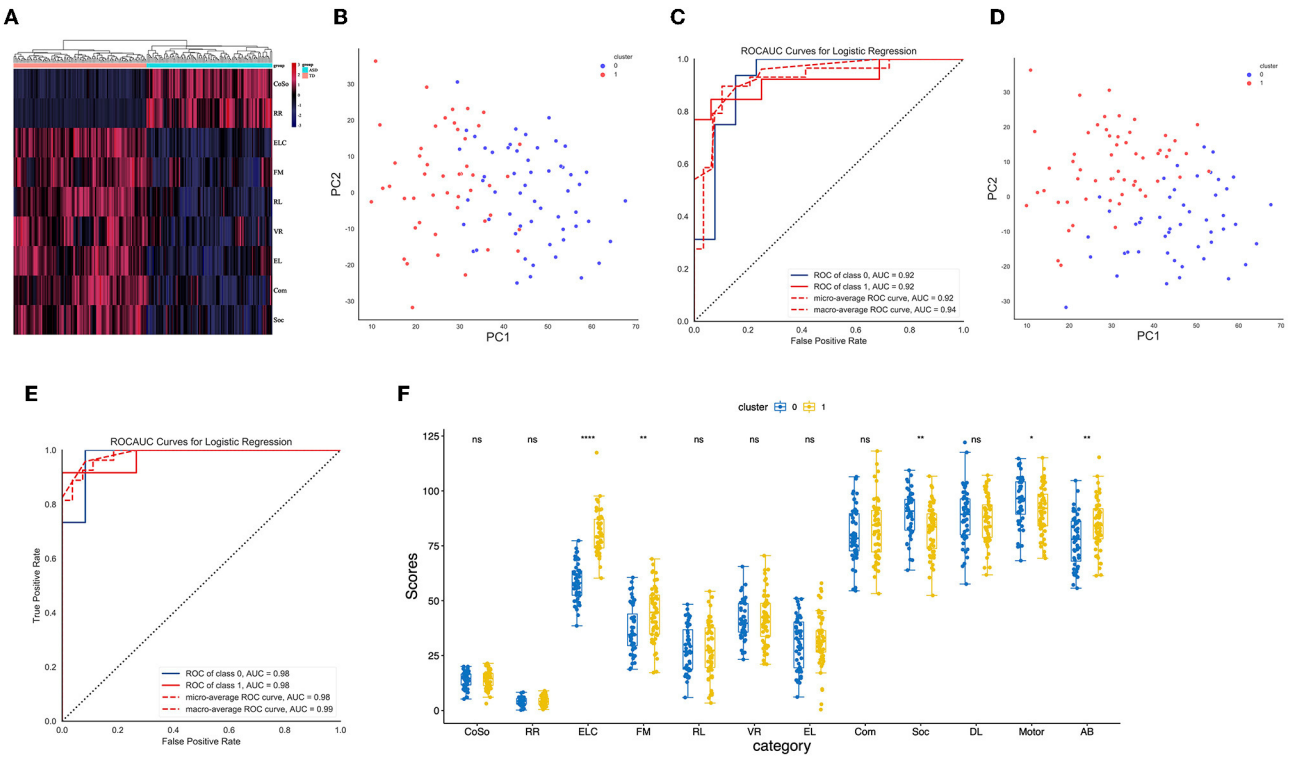

FIGURE 1
(A) A heatmap of the clinical feature score for CON and ASD samples. **(B)** Distribution of hierarchical clusters 0 and 1 of ASD in PC1 and PC2 spaces. **(C)** Logistic regression identifies the receiver-operating characteristic (ROC) curves of hierarchical ASD-0 and ASD-1 **(D)** Distribution of ASD-0 and ASD-1 clusters in PC1 and PC2 spaces identified by k-means. **(E)** Logistic regression identifies the ROC curves of k-means ASD-0 and ASD-1. **(F)** The boxplot shows the difference in clinical feature scores between ASD-0 and ASD-1.

200 DEGs with the largest absolute value of LogFC were used as the top 200 candidates for subsequent analysis.

2.5. Functional enrichment analysis of DEGs

The Gene Ontology enrichment analysis of DEGs was based on R language packages "org.Hs.eg.db" and "enrichplot." The gene set enrichment analysis (GSEA) function enrichment analysis was performed using the R language package "ReactomePA" (19, 20). The number of permutations was set to 100, and a p-value of <0.05 and an FDR of <0.25 were considered statistically significant. Ingenuity pathway analysis (IPA) was conducted using IPA software with input data for the gene symbol and the logFC value of DEGs with an adjusted p-value of <0.05.

2.6. The construction of machine learning models

The LASSO regression model, which is suitable for dimensionality reduction of high-dimensional data, was used to screen genes with non-zero coefficients. The expression data of the screened genes were used to train the logistic regression model and construct the nomograph. The glmnet (21), rms, and regplot (22) packages based on the R language were used to implement the above method. The expression data of the screened genes were used to train and optimize the neural network model using the sklearn package in Python.

2.7. The acquisition and pre-processing of the test dataset

The test dataset was GSE26415 (23). The platform was Agilent-014850 Whole Human Genome Microarray 4x44K G4112F. The original datasets contained four groups of 21 toddlers each: toddlers with ASDs, healthy age- and sex-matched subjects (ASD and CON, ages ranging from 18 to 35 months), healthy mothers who had toddlers with ASD (ASDmos), and CONmos (ages ranging from 33 to 55 months). Gene expression data of the 21 ASD and 21 CON participants were retained. The normalized data were downloaded, and the expression matrix was obtained.

2.8. Validation of ASD early diagnosis nomogram

A calibration curve method was adopted to evaluate the coordination between the predicted and actual ASD. A decision

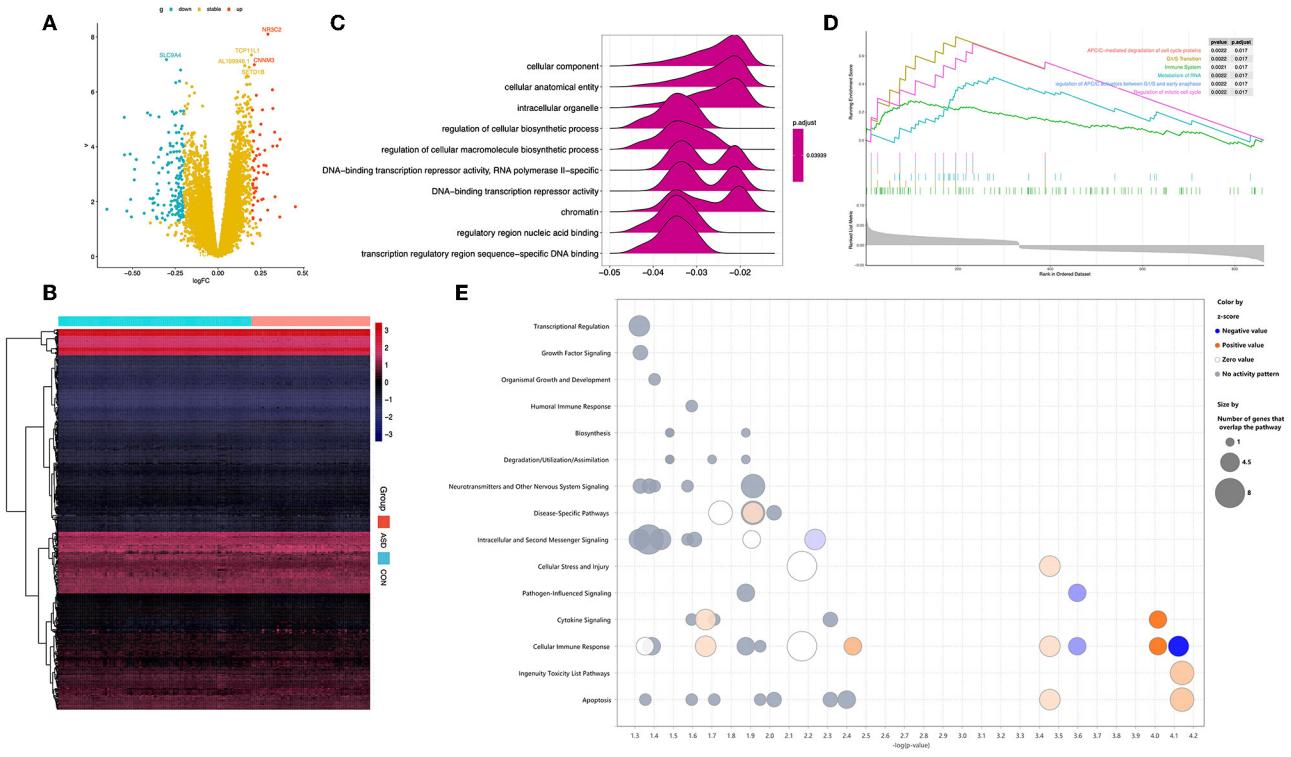

FIGURE 2
Identification of DEGs between ASD and CON and DEGs functional enrichment analysis. **(A)** Volcano diagram: Each point represents a protein: downregulated (blue), upregulated (red), and not significant (yellow). **(B)** A heatmap of the expression of DEGs in ASD and CON. **(C)** TOP-6 GSEA analysis of the KEGG pathway in DEGs. **(D)** Statistically significant Gene Ontology terms with a false discovery rate of <0.05. **(E)** The canonical pathway of IPA shows the signaling pathways in which DEGs are involved.

curve was adopted to quantify the net benefits of different threshold probabilities in the ASD cohort and evaluate the clinical utility of the strong line chart. The net benefits were calculated by subtracting the proportion of all false positive patients from the proportion of true positive patients. The relative harm caused by the intervention was balanced, and the negative consequences of unnecessary interventions were eliminated. The visualization of the decision curve was based on the RMDA package of the R language. The generalization ability and accuracy of the trained model were evaluated using the test datasets.

3. Results

3.1. Identification and removal of batch effects and normalization of datasets

We identified an obvious batch effect between the two datasets (GSE111175 and GSE42133) via PCA (Supplementary Figure S1A), which could be removed via the SVA package (Supplementary Figure S1B). To eliminate the within-group inconsistency and normalize the dataset, limma packages were used (Supplementary Figures S1C, D).

3.2. The analysis of participant characteristics and clinical information

First, differences in the clinical scores between the ASD and CON participants were identified (Supplementary Table S1, Figure 1A). Specifically, we compared the ADOS scores of the ASD and CON participants and observed that the ASD participants had significantly higher ADOS scores than the CON participants. Similarly, MSEL score differences between the two groups were identified. Furthermore, we also observed that the Vineland scores of the ASD participants were lower than those of the CON participants.

To assess whether there was a within-group bias within the ASD cohort, such as the presence of a good or poor ASD group, hierarchical cluster methods were employed. In particular, 61 ASD toddlers in cluster 0 and 67 ASD toddlers in cluster 1 were identified (Figure 1B). The confidence of the clusters was verified by logistic regression (AUC = 0.92, Figure 1C). To improve the accuracy of clustering, the k-means clustering method was employed. Ultimately, 66 ASD toddlers in cluster 0 and 62 ASD toddlers in cluster 1 were identified (Figure 1D). The confidence of the clusters was verified by logistic regression (AUC = 0.98, Figure 1E). Furthermore, we identified that early learning composite (ELC) and adaptive behavior (AB) scores in ASD cluster 1 were higher than those in ASD cluster 0 (Figure 1F).

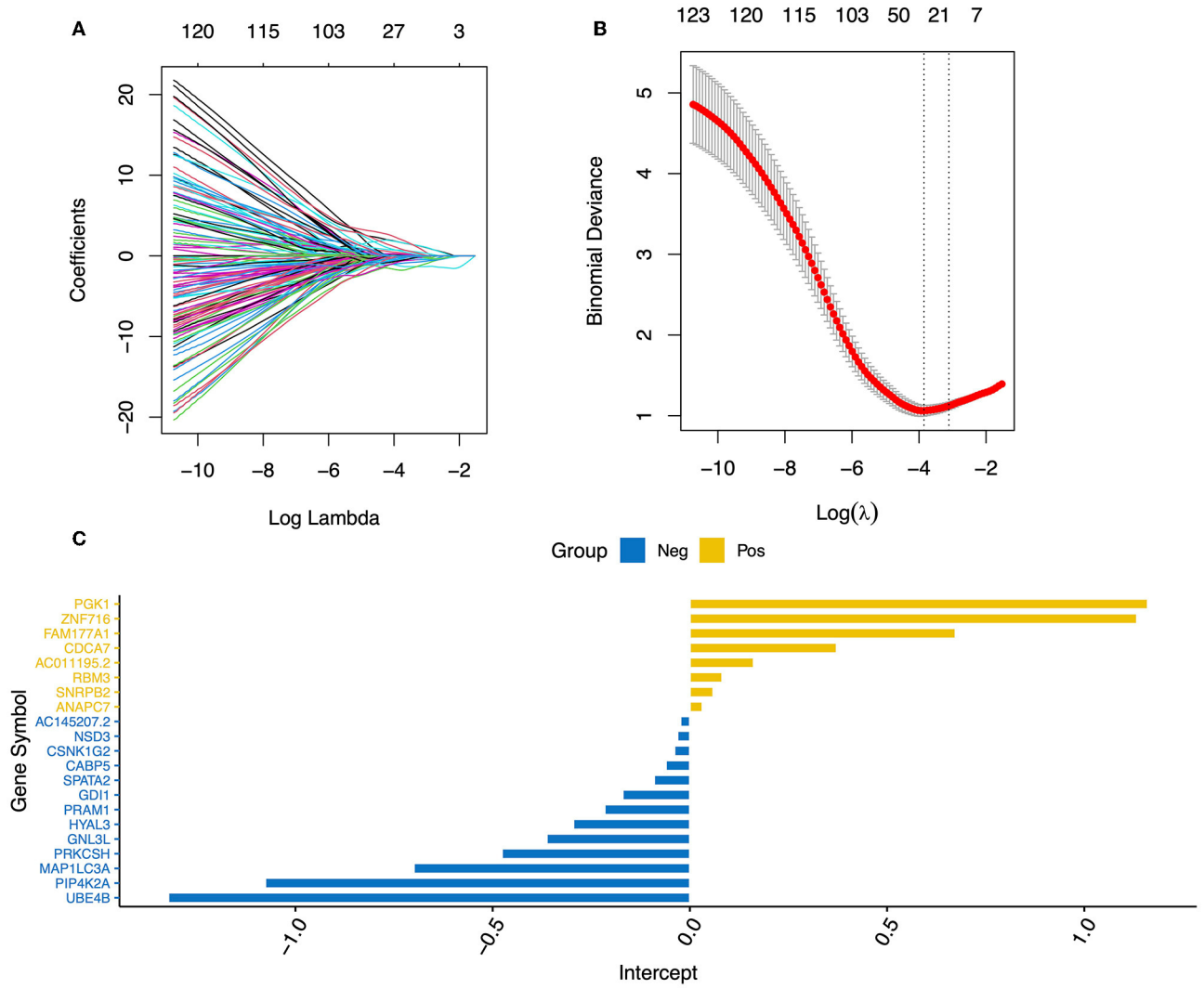

FIGURE 3
Blood feature gene selection using the LASSO binary logistic regression model. **(A)** LASSO coefficient profiles of the 200 DEGs. **(B)** A coefficient profile plot was produced against the log (lambda) sequence in the LASSO model. The optimal parameter (lambda) was selected as the first black dotted line indicated. **(C)** A bar graph of the coefficients of the 21 genes selected.

3.3. Identification of DEGs between ASD and CON, and the functions of DEGs

We identified 1,027 DEGs between 128 ASD and 126 CON participants based on the pre-processing expression data (Figures 2A, B). GO analysis was performed to identify the function of these DEGs; regulatory region nucleic acid binding and transcription regulatory region sequence-specific DNA binding were identified as the main molecular functions. Furthermore, the results of GSEA indicated that the main function of DEGs was APC-mediated degradation of cell cycle proteins, the immune system, and so on (Figures 2C, D). We also analyzed the canonical pathways of DEGs through IPA and found that the functions of the DEGs are enriched in immune-related signaling pathways such as cellular immune response and cytokine signaling. In addition, growth-related signaling pathways, neurotransmitter pathways, and other nervous signaling pathways were identified (Figure 2E).

3.4. Construction of nomogram for the early diagnosis of ASD

To limit the overfitting of machine learning models and screen the DEGs, LASSO regression was performed. Specifically, we selected 200 genes with large logFC absolute values from 1,027 DEGs for LASSO regression, and 21 DEGs were retained (Figures 3A–C). These genes included *ZNF716, SNRPB2, AC145207.2, UBE4B, SPATA2, RBM3, PRKCSH, PRAM1, PIP4K2A, PGK1, NSD3, MAP1LC3A, HYAL3, GNL3,* and so on. Logistic regression analysis and nomography were performed to construct a diagnosis model of ASD based on the expression data of the 21 DEGs (Figures 4A, B). In addition, the calibration curve demonstrated the consistency between the predictions of the nomograph and the cohort (Figure 4C).

To evaluate the generalization ability of the logistic regression model and the benefits of the nomogram in clinical use, we assessed

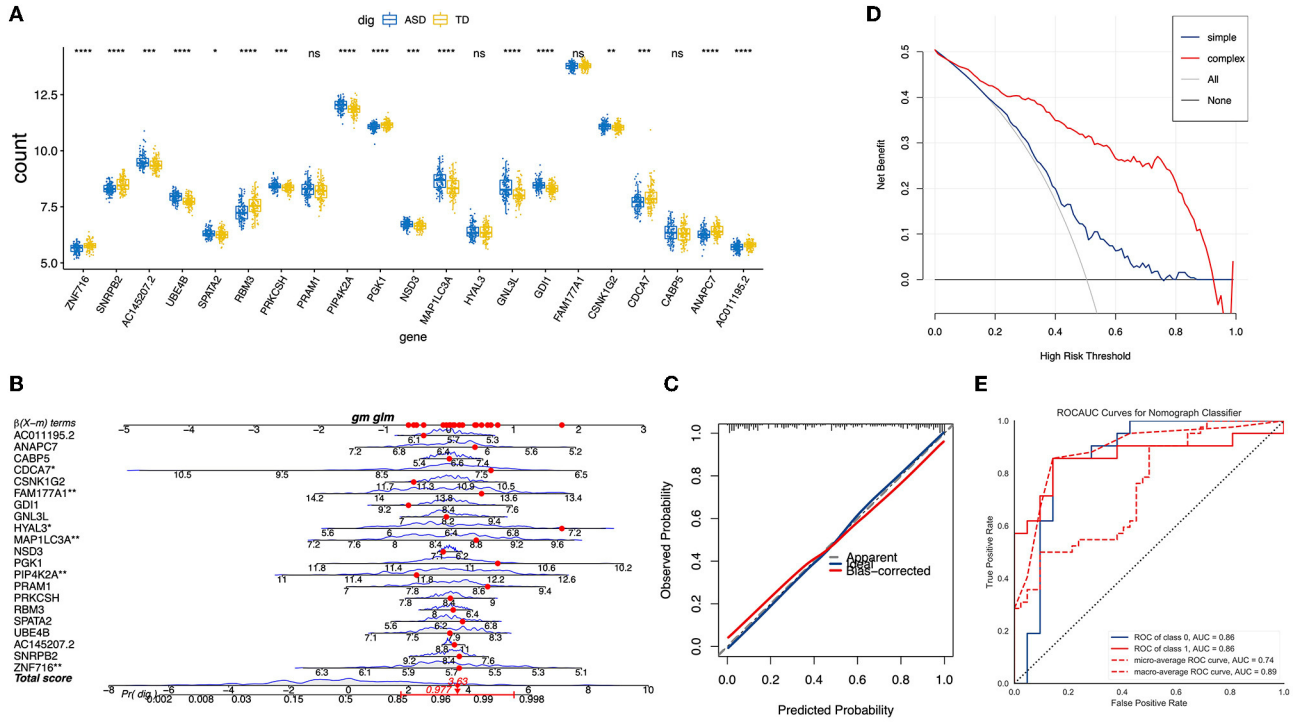

FIGURE 4
Construction and validation of ASD diagnosis nomogram. **(A)** A boxplot of the expression values of the 21 genes selected in different samples. **(B)** The nomogram for predicting the risk of ASD based on 21 genes. **(C)** The calibration curves of the ASD prediction nomogram. **(D)** The ROC curve of the nomogram in the validation cohort. **(E)** The decision curve analysis for the nomogram.

the accuracy of prediction of the trained logistic regression model in the test dataset, as shown in Figure 4D, AUC = 0.86. Furthermore, the decision curve of the nomogram indicated a higher net benefit of using the complex model than using the single gene model (Figure 4E).

3.5. Construction of the artificial neural network model

To further improve the accuracy of ASD diagnosis, we used the Python-based sklearn package to construct a neural network model that could predict ASD. Specifically, the above 21 gene expression matrix was used to train the neural network model, which had five hidden layers, and the number of neurons included in each hidden layer was 10, 10, 8, 10, and 10, as shown in Figure 5A. The number of iterations was set to 200, ReLU was selected as the activation function, the weight optimizer was the optimizer of the quasi-Newton method, and the regularization parameter was 0.00001. The average AUC of the model in the training dataset was 0.88, as shown in Figure 5B. Next, we evaluated the generalization ability of the trained neural network model in the test dataset (n ASD = 21, n CON = 21), as shown in Figure 5C, AUC = 0.88. The precision-recall curve (Figure 5D, average precision = 0.91) indicated that the trained neural network had improved generalization and accuracy.

In addition, to prevent the differences within the ASD dataset from affecting the prediction results of the model, the data of ASD good and ASD poor were used to construct the neural network model. Firstly, 66 ASD good and 126 CON samples were used to construct the neural network model of the above structure, and the accuracy of the model in predicting the ASD good and CON samples was 0.87 through 5-fold cross-validation (Supplementary Figure S2A). Similarly, we also built the neural network model of ASD poor and CON samples and found that the accuracy of the model was 0.85 (Supplementary Figure S2B). To explore the 21 genes' ability to distinguish between ASD good and poor samples, 66 ASD good and 62 ASD poor samples were used to construct the neural network, and we found that the accuracy of this model was 0.68 (Supplementary Figure S2C). Finally, to explore the specificity of the neural network model in predicting ASDs, 38 toddlers with developmental delays were considered in the analysis, and the accuracy of the model in predicting 38 developmental delays was 0.52 (Supplementary Figure S2D).

4. Discussion

Early diagnosis and targeted interventions for ASD in toddlers are imperative to optimize neurodevelopmental outcomes and quality of life (5); yet, current diagnostic methods are subjective, cost-prohibitive, and constrained by developmental maturation. The current diagnosis of ASD includes multiple dimensions,

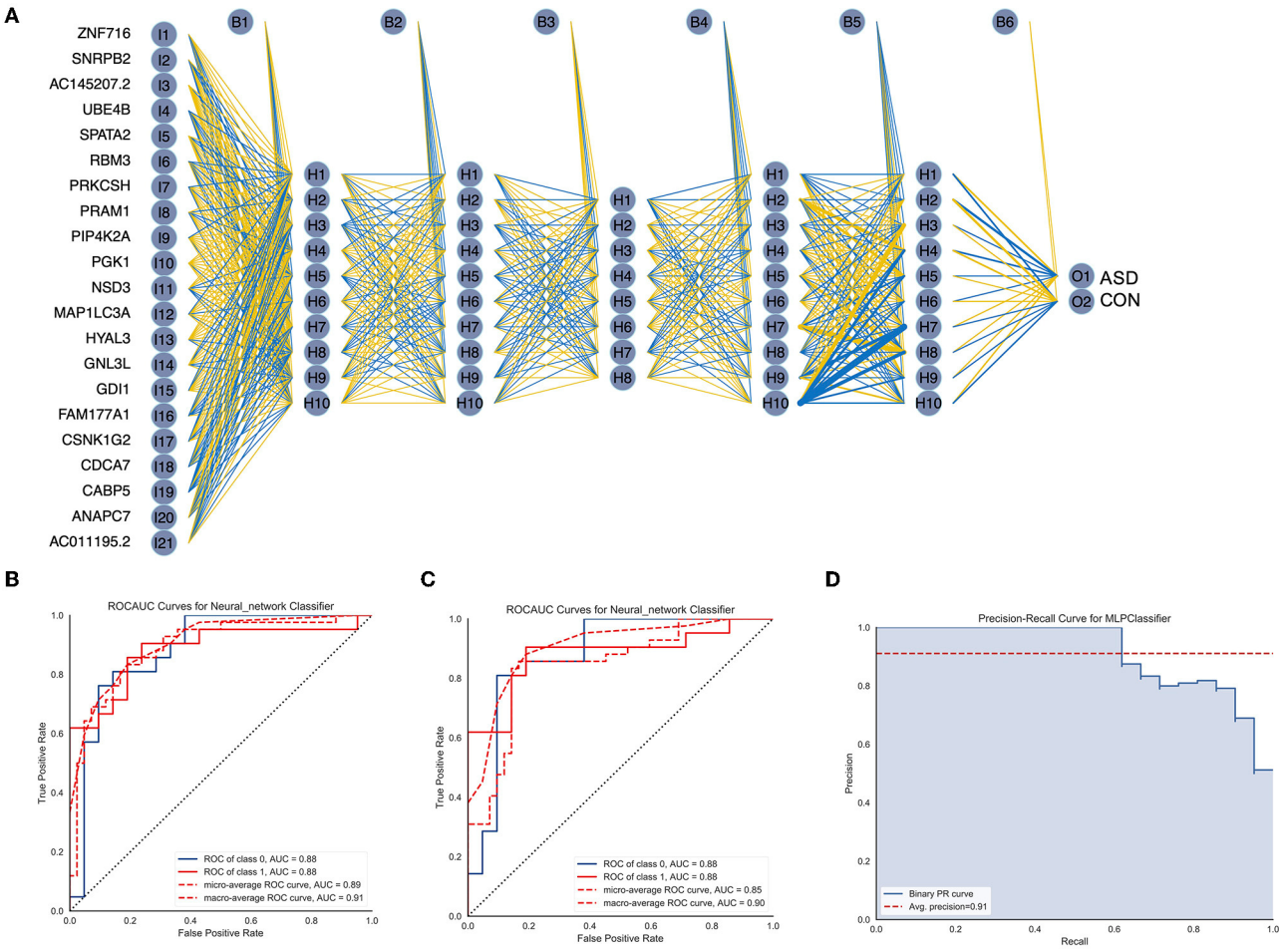

FIGURE 5
Construction and validation of the ASD diagnosis neural network. **(A)** Results of neural network visualization. **(B)** The ROC curve of the neural network in the training cohort. **(C)** The ROC curve of the neural network in the validation cohort. **(D)** The precision-recall curve for the neural network classifier.

such as vision and hearing examination to exclude sensory impairment, genetic and neurological testing, interviews with parents, neurological or psychiatrist observation of children, and development and behavioral tests of children (24–26). These special evaluation or inspection methods are either subjective or expensive. In addition, the cognitive assessment for toddlers may vary depending on their age. Thus, there is an urgent need to develop a safe, convenient, and accurate method for ASD diagnosis.

Biomarker-based machine learning models could enable scalable, accurate, and objective early detection of ASD, which is critical to transforming prognosis. Compared with other methods, biomarkers are more objective and convenient (27, 28). With the development of machine learning algorithms, the accuracy of diagnosis can be improved by combining machine learning methods with biomarker data. However, two issues remain and must be resolved. First, machine learning usually requires a large amount of data (29, 30). To expand the dataset, datasets GSE111175 and GSE42133 from the GEO database, which included 128 ASD participants and 126 CON participants in the datasets, were obtained. Second, limiting the features of the data to avoid overfitting the model and improving the generalization ability of the model are necessary. In this study, several methods were employed to reduce the dataset's dimensions, such as selecting the top 200 genes from 1,027 DEGs (Figures 2A, B) and LASSO regression (Figure 3). In addition, 21 key DEGs were identified as the biomarkers to diagnose ASD (Figures 3, 4A). To evaluate the generalization ability of the model, we compared the predicted results of the trained dataset with the predicted results of the test dataset (GSE26415, 21 CON, 21 ASD) and found that the difference in accuracy between the two was negligible, indicating that there was no serious overfitting in the trained model (Figure 4D). We also found that the net benefit of using the complex model was higher than the benefit of using a single gene model, indicating the strong robustness of the trained model (Figure 4E). To obtain a more accurate model, the neural network model was also trained in this study (Figure 5A). Similarly, we found that the accuracy of the neural network in the test data was higher than that of the nomograph (AUC = 0.88, Figures 5B, C). Although the improvement in the accuracy of the model was not as obvious

as in other studies (13, 23), we established the robustness of the model across different datasets. To prevent the interference of intra-group differences in ASDs with the prediction results, we redefined neural network models as ASD good and ASD poor, respectively. Compared with the original neural network model, the accuracy of the ASD good and ASD poor models decreased, which may have been caused by the reduction of the dataset size (Supplementary Figures S2A, B). In addition, we found that the neural network model used to distinguish ASD good from ASD poor did not show high accuracy, which may have been due to the insignificant difference in biomarkers between the two groups (Supplementary Figure S2C).

Many previous studies have identified biomarkers related to the immune system in blood samples from individuals with ASD (13, 31–33), which is consistent with our results that the DEG pathways are mainly distributed in the immune system (Figures 2C–E). Furthermore, the growth factor, organismal growth and development, neurotransmitters, and other nervous system signaling were established through IPA analysis, which may indicate the novel and specific pathways related to ASD biomarkers (Figure 2E). Among these 21 biomarkers, we identified that *GDI1* expression in ASD was significantly increased (Figure 4A), which can be explained by the recurrent duplications of *GDI1* in ASD (34). In a previous study, *GDI1*-related pathways were reported to be a vesicle-mediated transport (35), which may be related to the neurotransmitter pathway identified in the IPA (Figure 2E). A genome-wide association study analysis showed that *HYAL3* is one of the pathogenic genes of attention-deficit hyperactivity disorder (36). In our study, the expression of *HYAL3* in ASD was significantly increased (Figure 4A), indicating that *HYAL3* may also cause ASD. Similarly, a decrease in *ANAPC7* expression in ASD was also identified (Figure 4A). Studies have shown that the loss of *ANAPC7* is associated with intellectual disability syndrome (37), which may also explain the intellectual disability in toddlers with ASD.

The final limitation of our study is the generalization ability of the neural network model, which can be better assessed with the help of multiple test datasets. We encountered difficulty in finding the expression profile data of blood leukocyte samples for ASD in open databases, which prevented us from verifying the generalizability of the model in this study. Thus, in a future study, we will explore whether specific data are available for this purpose.

5. Conclusion

In this study, we obtained blood RNA-seq data of CON and ASD toddlers from public GEO datasets in NCBI and identified the potential biomarkers of ASD, including ANAPC7 and HYAL3, through a series of analyses. Machine learning can be used on the expression data of the biomarkers to obtain models with high prediction accuracy. By predicting different datasets, we established a certain level of generalizability of our model. We also predicted the developmental delay dataset and established that our model has a certain specificity. Further improvements to the model may shed some light on its clinical application.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111175; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42133.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

HT contributed to the study design, performed the experiments, and contributed to the writing of the manuscript. JL contributed to the data collection and writing of the manuscript. KC, HG, WY, PC, and SC contributed to the study design. DS contributed to the study design, performed the experiments, and contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We acknowledge the GEO database for providing their platforms and contributors for uploading meaningful datasets. We also acknowledge Chenxi Li, Ph.D., of West Lake University for helping with the IPA analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2023.1203375/full#supplementary-material

References

1. Lai M-C, Lombardo MV, Baron-Cohen S. Autism. *Lancet.* (2014) 383:896–910. doi: 10.1016/S0140-6736(13)61539-1

2. Sandin S, Lichtenstein P, Kuja-Halkola R, Hultman C, Larsson H, Reichenberg A. The heritability of autism spectrum disorder. *JAMA.* (2017) 318:1182–4. doi: 10.1001/jama.2017.12141

3. Chang J, Gilman SR, Chiang AH, Sanders SJ, Vitkup D. Genotype to phenotype relationships in autism spectrum disorders. *Nat Neurosci.* (2015) 18:191–8. doi: 10.1038/nn.3907

4. Bhandari R, Paliwal JK, Kuhad A. Neuropsychopathology of autism spectrum disorder: complex interplay of genetic, epigenetic, and environmental factors. *Adv Neurobiol.* (2020) 24:97–141. doi: 10.1007/978-3-030-30402-7_4

5. Famitafreshi H, Karimian M. Overview of the recent advances in pathophysiology and treatment for autism. *CNS Neurol Disord Drug Targets.* (2018) 17:590–4. doi: 10.2174/1871527317666180706141654

6. Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, Centers for Disease Control and Prevention (CDC). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveill Summ.* (2014) 63:1–21.

7. Bosl WJ, Tager-Flusberg H, Nelson CA. EEG analytics for early detection of autism spectrum disorder: a data-driven approach. *Sci Rep.* (2018) 8:6828. doi: 10.1038/s41598-018-24318-x

8. Santana CP, de Carvalho EA, Rodrigues ID, Bastos GS, de Souza AD, de Brito LL. rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. *Sci Rep.* (2022) 12:6030. doi: 10.1038/s41598-022-09821-6

9. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet.* (2019) 51:431–44. doi: 10.1038/s41588-019-0344-8

10. Gazestani VH, Pramparo T, Nalabolu S, Kellman BP, Murray S, Lopez L, et al. Perturbed gene network containing PI3K-AKT, RAS-ERK and WNT-β-catenin pathways in leukocytes is linked to ASD genetics and symptom severity. *Nat Neurosci.* (2019) 22:1624–34. doi: 10.1038/s41593-019-0489-x

11. Lombardo MV, Eyler L, Pramparo T, Gazestani VH, Hagler DJ, Chen C-H, et al. Atypical genomic cortical patterning in autism with poor early language outcome. *Sci Adv.* (2021) 7:eabh1663. doi: 10.1126/sciadv.abh1663

12. Pramparo T, Lombardo MV, Campbell K, Barnes CC, Marinero S, Solso S, et al. Cell cycle networks link gene expression dysregulation, mutation, and brain maldevelopment in autistic toddlers. *Mol Syst Biol.* (2015) 11:841. doi: 10.15252/msb.20156108

13. Pramparo T, Pierce K, Lombardo MV, Carter Barnes C, Marinero S, Ahrens-Barbeau C, et al. Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry.* (2015) 72:386–94. doi: 10.1001/jamapsychiatry.2014.3008

14. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* (2012) 28:882–3. doi: 10.1093/bioinformatics/bts034

15. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, 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.* (2000) 30:205–23. doi: 10.1037/t17256-000

16. Hutchins T, Vivanti G, Mateljevic N, Jou RJ, Shic F, Cornew L, et al. Mullen Scales of Early Learning. In: Volkmar FR, editor. *Encyclopedia of Autism Spectrum Disorders*. New York, NY: Springer New York (2013). p. 1941–1946.

17. Sparrow SS. Vineland Adaptive Behavior Scales. In: Kreutzer JS, DeLuca J, Caplan B, editors. *Encyclopedia of Clinical Neuropsychology*. New York, NY: Springer New York (2011). p. 2618–2621. doi: 10.1007/978-0-387-79948-3_1602

18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007

19. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst.* (2016) 12:477–9. doi: 10.1039/C5MB00663E

20. Powers RK, Goodspeed A, Pielke-Lombardo H, Tan A-C, Costello JC. GSEA-InContext: identifying novel and common patterns in expression experiments. *Bioinformatics.* (2018) 34:i555–64. doi: 10.1093/bioinformatics/bty271

21. Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Epigenetics.* (2019) 11:123. doi: 10.1186/s13148-019-0730-1

22. Zhang Z, Cortese G, Combescure C, Marshall R, Lee M, Lim HJ, et al. Overview of model validation for survival regression model with competing risks using melanoma study data. *Ann Transl Med.* (2018) 6:325. doi: 10.21037/atm.2018.07.38

23. Kuwano Y, Kamio Y, Kawai T, Katsuura S, Inada N, Takaki A, et al. Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children. *PLoS ONE.* (2011) 6:e24723. doi: 10.1371/journal.pone.0024723

24. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, et al. Using whole-exome sequencing to identify inherited causes of autism. *Neuron.* (2013) 77:259–73. doi: 10.1016/j.neuron.2012.11.002

25. Robertson CE, Baron-Cohen S. Sensory perception in autism. *Nat Rev Neurosci.* (2017) 18:671–84. doi: 10.1038/nrn.2017.112

26. Brosnan M. An exploratory study of a dimensional assessment of the diagnostic criteria for autism. *J Autism Dev Disord.* (2020) 50:4158–64. doi: 10.1007/s10803-020-04474-8

27. Farah R, Haraty H, Salame Z, Fares Y, Ojcius DM, Said Sadier N. Salivary biomarkers for the diagnosis and monitoring of neurological diseases. *Biomed J.* (2018) 41:63–87. doi: 10.1016/j.bj.2018.03.004

28. Galiana-Simal A, Muñoz-Martinez V, Calero-Bueno P, Vela-Romero M, Beato-Fernandez L. Towards a future molecular diagnosis of autism: Recent advances in biomarkers research from saliva samples. *Int J Dev Neurosci.* (2018) 67:1–5. doi: 10.1016/j.ijdevneu.2018.03.004

29. Blum AL, Langley P. Selection of relevant features and examples in machine learning. *Artif Intell.* (1997) 97:245–71. doi: 10.1016/S0004-3702(97)00063-5

30. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* (2014) 15:1929–58. doi: 10.5555/2627435.2670313

31. Ashwood P, Corbett BA, Kantor A, Schulman H, Van de Water J, Amaral DG. In search of cellular immunophenotypes in the blood of children with autism. *PLoS ONE.* (2011) 6:e19299. doi: 10.1371/journal.pone.0019299

32. Bjorklund G, Saad K, Chirumbolo S, Kern JK, Geier DA, Geier MR, et al. Immune dysfunction and neuroinflammation in autism spectrum disorder. *Acta Neurobiol Exp.* (2016) 76:257–68. doi: 10.21307/ane-2017-025

33. McCarthy MM, Wright CL. Convergence of sex differences and the neuroimmune system in autism spectrum disorder. *Biol Psychiatry.* (2017) 81:402–10. doi: 10.1016/j.biopsych.2016.10.004

34. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am J Hum Genet.* (2014) 94:677–94. doi: 10.1016/j.ajhg.2014.03.018

35. John A, Ng-Cordell E, Hanna N, Brkic D, Baker K. The neurodevelopmental spectrum of synaptic vesicle cycling disorders. *J Neurochem.* (2021) 157:208–28. doi: 10.1111/jnc.15135

36. Fahira A, Li Z, Liu N, Shi Y. Prediction of causal genes and gene expression analysis of attention-deficit hyperactivity disorder in the different brain region, a comprehensive integrative analysis of ADHD. *Behav Brain Res.* (2019) 364:183–92. doi: 10.1016/j.bbr.2019.02.010

37. Ferguson CJ, Urso O, Bodrug T, Gassaway BM, Watson ER, Prabu JR, et al. APC7 mediates ubiquitin signaling in constitutive heterochromatin in the developing mammalian brain. *Mol Cell.* (2022) 82:90–105.e13. doi: 10.1016/j.molcel.2021.11.031