Diagnostic Accuracy Study

Medicine®

OPEN

Development and validation of a novel and robust blood small nuclear RNA signature in diagnosing autism spectrum disorder

Jinxia Zhou, MD, Qian Hu, MD, Xijia Wang, MD, Wei Cheng, MD, Chunlian Pan, MD*, Xiaobin Xing, MD*

Abstract

Reliable molecular signatures are needed to improve the early and accurate diagnosis of autism spectrum disorder (ASD), and indicate physicians to provide timely intervention. This study aimed to identify a robust blood small nuclear RNA (snRNA) signature in diagnosing ASD. 186 blood samples in the microarray dataset were randomly divided into the training set (n = 112) and validation set (n = 72). Then, the microarray probe expression profiles were re-annotated into the expression profiles of 1253 snRNAs though probe sequence mapping. In the training set, least absolute shrinkage and selection operator (LASSO) penalized generalized linear model was adopted to identify the 9-snRNA signature (RNU1-16P, RNU6-1031P, RNU6-258P, RNU6-335P, RNU6-485P, RNU6-549P, RNU6-98P, RNU6ATAC26P, and RNVU1-15), and a diagnostic score was calculated for each sample according to the snRNA expression levels and the model coefficients. The score demonstrated a good diagnostic ability for ASD in the training set (area under receiver operating characteristic curve (AUC) = 0.90), validation set (AUC = 0.87), and the overall (AUC = 0.88). Moreover, the blood samples of 23 ASD patients and 23 age- and gender-matched controls were collected as the external validation set, in which the signature also showed a good diagnostic ability for ASD (AUC = 0.88). In subgroup analysis, the signature was robust when considering the confounders of gender, age, and disease subtypes, and displayed a significantly better performance among the female and younger cases ($P = .039$; $P = .002$). In comparison with a 55-gene signature deriving from the same dataset, the snRNA signature showed a better diagnostic ability (AUC: 0.88 vs 0.80, $P = .049$). In conclusion, this study identified a novel and robust blood snRNA signature in diagnosing ASD, which might help improve the diagnostic accuracy for ASD in clinical practice. Nevertheless, a large-scale prospective study was needed to validate our results.

Abbreviations: ASD = autism spectrum disorder, AUC = area under ROC curve, dNTPs = deoxy-ribonucleoside triphosphate, GWAS = genome-wide association study, LASSO = least absolute shrinkage and selection operator, PDD-NOS = pervasive developmental disorder-not otherwise specified, ROC = receiver operating characteristic, snRNAs = small nuclear RNAs.

Keywords: autism spectrum disorder, diagnosis, signature, small nuclear RNA

1. Introduction

Autism spectrum disorder (ASD) is a heterogeneous set of neurodevelopmental diseases, characterized by deficits in social communication and verbal/nonverbal interaction, as well as restricted and repetitive patterns of interests and behaviors. It has a high prevalence of approximately 0.3% to 1.2%, with 3 main subtypes of autistic disorder, Asperger's disorder and pervasive developmental disorder-not otherwise specified (PDD-NOS).[1] Despite of the disease onsite before 3 years old, most children are diagnosed with ASD after 4 years old.[2] Early intensive behavioral interventions could improve the outcomes (eg, language skills, cognitive performance, and adaptive behavior skills) in some young children with ASD.[3] Thus, it has a critical need in clinical practice to increase the diagnostic accuracy for ASD. Reliable molecular signatures could help improve the early and accurate diagnosis of ASD, and indicate physicians to provide timely intervention. Small noncoding RNAs have been indicated as new class of biomarkers in ASD, but few studies focused on the subtype of small nuclear RNAs (snRNAs).[4]

As the components of spliceosome, snRNAs were fairly conserved with a uridine-rich noncoding sequence of less than 200 nt, and involved in the splicing of precursor messenger RNA (mRNA).[4] SnRNAs could be divided into 2 main categories according to the common sequences and interactive proteins. Sm-class of snRNAs (U1, U2, U4, and U5), characterized by a 5′ trimethylguanosine cap, were synthesized by RNA polymerase II and bound several Sm proteins. Lsm snRNAs (U6 and other snRNAs), characterized by a monomethylphosphate 5′ cap, were transcribed by polymerase III and acted as a binding site for Lsm proteins. Spliceosomes have been reported in the pathogenesis of nervous system diseases, demonstrating a potential involvement of snRNA.[5]

With great advances in microarray technologies, gene expression profiles were more available, and we could further adopt the method of probe re-annotation to extract the snRNA expression profiles. In this study, we aimed to develop and validate a novel and robust snRNA signature in diagnosing ASD, which might help improve the diagnostic accuracy for ASD in clinical practice.

Editor: Massimo Tusconi.

JZ and QH contributed equally to this study.

The authors have no funding and conflicts of interest to disclose.

Department of Neurology, Puren Hospital Affiliated to Wuhan University of Science and Technology, Wuhan, China.

** Correspondence: Xiaobin Xing, Department of Neurology, Puren Hospital Affiliated to Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: xingxiaobin2019@163.com).*

How to cite this article: Zhou J, Hu Q, Wang X, Cheng W, Pan C, Xing X. Development and validation of a novel and robust blood small nuclear RNA signature in diagnosing autism spectrum disorder. Medicine 2019;98:45(e17858).

Received: 24 August 2019 / Received in final form: 4 October 2019 / Accepted: 9 October 2019

http://dx.doi.org/10.1097/MD.0000000000017858

2. Methods

2.1. Data preparation

The database of gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/) was reviewed to search proper datasets, which met the following criteria:

(1) detected blood gene expression profiles of both ASD patients and controls;
(2) availability of clinical data and probe sequences;
(3) the sample size was large enough.

Then, the tab-delimited expression value-matrix table was downloaded and $\log_2$-transformed. The study was approved by the ethnic committee of Puren Hospital Affiliated to Wuhan University of Science and Technology.

2.2. Probe re-annotation

First, we obtained the microarray probe sequences from the Affymetrix product website (http://www.affymetrix.com), as well as the human genome sequences (GRCh38.p12) and comprehensive gene annotation from the GENCODE database (https://www.gencodegenes.org).[6] Then, the HISAT software (hierarchical indexing for spliced alignment of transcripts) was applied to identify probe-matched sequences. Transcripts were included according to the following criteria:

(1) mapped by at least 1 probe sequence and without any mismatch;
(2) each probe was matched to only 1 transcript in probe-transcript pairs.

When multiple probes matched to an identical gene, the average expression value across these probes was calculated to represent the corresponded gene. SnRNA expression profiles were extracted according to the RNA types.

2.3. SnRNA signature construction, evaluation, and validation

The samples were randomly divided into the training set and validation set according to the ratio of 6:4. In the training set, a least absolute shrinkage and selection operator (LASSO) penalized generalized linear model was adopted to identify significant snRNAs. The penalty parameter was estimated by 10-fold cross-validation at 1 standard error beyond the minimum partial likelihood deviance. Then, the coefficients of significant snRNAs in the model were extracted to calculate a diagnostic score for each sample in the training set, validation set, and the overall. In receiver operating characteristic (ROC) curve analysis, area under ROC curve (AUC) was calculated to evaluate the diagnostic ability of the signature. Moreover, subgroup analysis was conducted on gender, age, and disease subtypes to assess the diagnostic stability of the signature, and DeLong's test for 2 ROC curves was performed to investigate the difference between subgroups. Finally, we also compared the snRNA signature with a 55-gene signature which derived from the same dataset.[7]

2.4. External validation

The blood samples of 23 ASD and 23 age- and gender-matched controls were obtained in Puren Hospital from September 2015 to July 2017. Written informed consent was provided before sample collection, and the present study protocol was approved by the ethnic committee of Puren Hospital. Then, total RNA was extracted using TRIzol reagent (Invitrogen, Waltham, MA), and stored at -80°C. The RNA concentration and purity were measured by the NanoDrop spectrophotometer (Thermo Fisher, Waltham, MA). Total RNA was synthesized into first-strand complementary DNA using fluorescent-labeled deoxy-ribonucleoside triphosphate (dNTPs) (Thermo Fisher, Waltham, MA), before hybridization with a customized microarray which tailed and fixed 9 snRNA probes (CapitalBio, China). The probe sequences were as follows:

RNU1-16P: GGGACTATGTTCGTGTTCTCTCCTG
RNU6-1031P: AAAATTGGAGTGATACAGAGAACAT
RNU6-258P: AAGTCGTGAAATAGTCCATATGTTA
RNU6-335P: TGCAAATTTGTGAAGAGGCACATTT
RNU6-485P: CCCTGTGCAAGGATGATATGCAAAT
RNU6-549P: GCTCACTTCAGTGGTACATATACTA
RNU6-98P: CAAATTCGAGAAATGTAGGAATTTT
RNU6ATAC26P: GAGAAGGTTAGCACTTCCCTTGGCA
RNVU1-15: GAACTCGACTGCATAACTTGTGATA

Finally, the snRNA expression levels were detected by the GenePix microarray scanner (Axon Instrument, Union City, CA), and a diagnostic score was calculated for each sample according to the signature formula.

2.5. Statistical analysis

All statistical analyses were conducted using R 3.6.0 software. The generalized linear model was constructed with glmnet 2.0-18 package, and ROC curve analysis was performed with ROCR 1.0-7 package. A 2-sided P-value $< .05$ was considered statistically significant.

3. Results

3.1. Characteristic of included dataset

The included microarray dataset of GSE18123 was based on the platform of GPL6244 (Affymetrix Human Gene 1.0 ST Array [HuGene-1_0-st]), with a total of 104 ASD (80 males [76.9%] and average age 8.1 years [2–21]) and 82 controls (48 males [58.3%] and average age 8.0 years [2–22]). The subtypes of autistic disorder, Asperger disorder, and PDD-NOS accounted for 39.4% (n = 41), 14.4% (n = 15), and 46.2% (n = 48), respectively. Then, 186 samples were randomly divided into the training set (n = 112) and validation set (n = 74).

3.2. Data preprocessing and sample clustering

The GPL6244 platform contained 861493 sequences (25 bp) aligning to 33297 probes. After probe re-annotation, a total of 8077 RNAs (32 types) were identified with 18329 specific probes, among which there were 1253 snRNAs (U1, U2, U4, U5, U6, U7, U11, and U12) mapping to 2561 probes.

Then, samples were clustered according to the distance in Pearson correlation matrices. When adopted the expression profiles of probes and genes, no outliers were detected (height <0.2) (Fig. 1). However, it was more discrete when based on the snRNA expression profiles, indicating a potential differential expression of snRNAs in ASD.

3.3. Signature construction, evaluation, and validation

In the training set, the LASSO penalized generalized linear model identified 9 significant snRNAs, which demonstrated an obvious discrepancy between ASD and control samples (Table 1). A

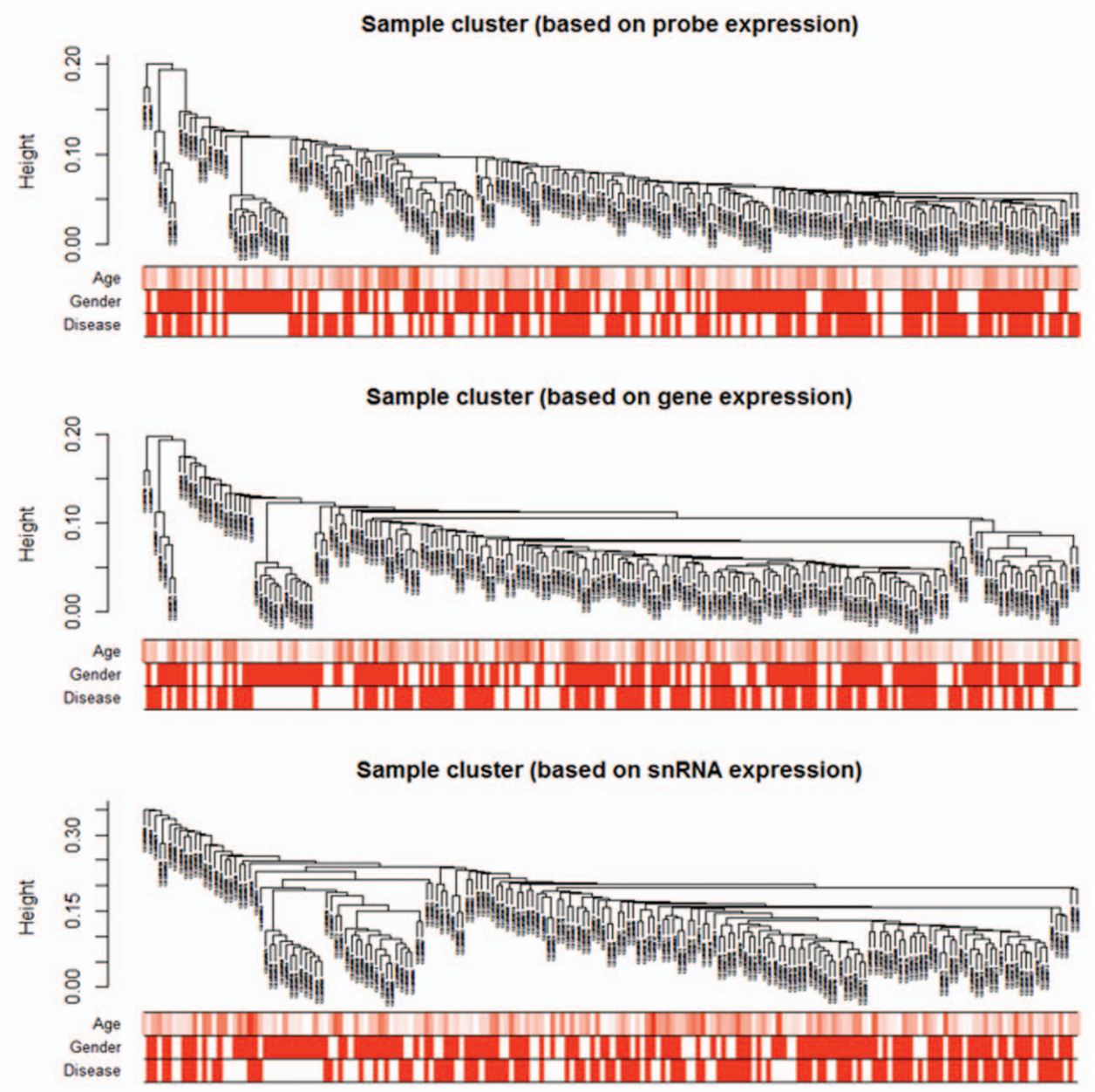

Figure 1. Sample clustering based on the expression profiles of probes, genes, and snRNAs. snRNAs = small nuclear RNAs.

Table 1

Nine snRNAs in the blood diagnostic signature of autism spectrum disorder.

snRNA symbol	snRNA name	Genomic location	Size (bases)	Probe ID	Ensembl ID
RNU1-16P	U1 Small Nuclear 16	Chr13: 113,478,915-113,479,078	164	7972921	ENSG00000202347
RNU6-1031P	U6 Small Nuclear 1031	Chr1: 67,541,127-67,541,233	107	7902225	ENSG00000207504
RNU6-258P	U6 Small Nuclear 258	Chr17: 20,126,388-20,126,488	101	8013448	ENSG00000212186
RNU6-335P	U6 Small Nuclear 335	Chr4: 183,675,603-183,675,715	113	8103892	ENSG00000201433
RNU6-485P	U6 Small Nuclear 485	Chr12: 7,118,785-7,118,891	107	7953620	ENSG00000200345
RNU6-549P	U6 Small Nuclear 549	Chr15: 64,671,263-64,671,369	107	7984215	ENSG00000207162
RNU6-98P	U6 Small Nuclear 98	ChrX: 132,580,117-132,580,223	107	8175193	ENSG00000206900
RNU6ATAC26P	U6atac Small Nuclear 26	Chr3: 57,619,447-57,619,572	126	8080683	ENSG00000210841
RNVU1-15	Variant U1 Small Nuclear 15	Chr1: 144,412,576-144,412,740	165	7919556; 7919560	ENSG00000207205

snRNAs = small nuclear RNAs.

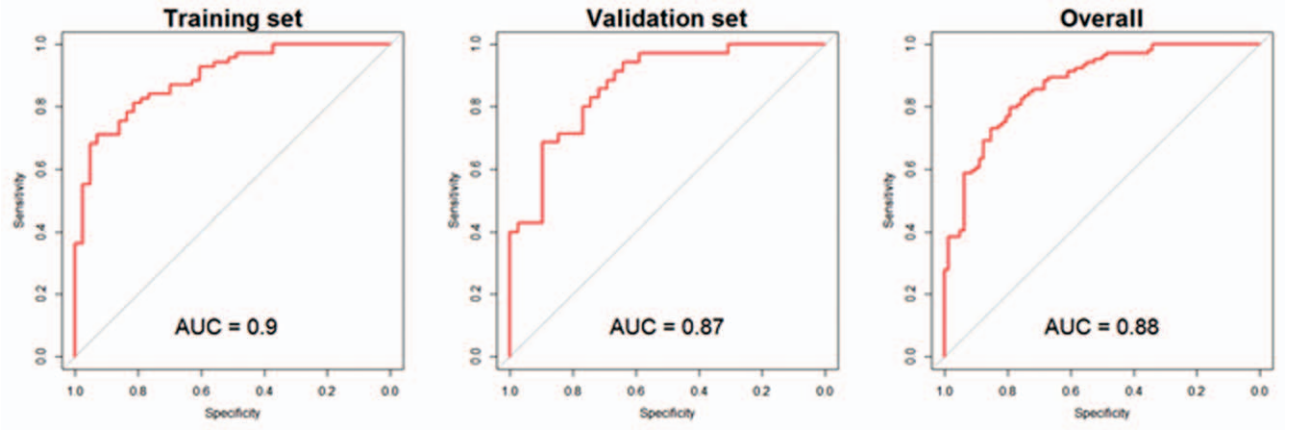

Figure 2. ROC curve analysis of the diagnostic signature. AUC = area under receiver operating characteristic curve, ROC = receiver operating characteristic.

diagnostic score was calculated for each sample according to the snRNA expression levels weighted by their coefficients in the LASSO model.

Diagnostic score = RNVU1-15*0.186 + RNU1-16P*0.141 + RNU6-258P*0.119 + RNU6-485P*(−0.244) + RNU6-549P*0.125 + RNU6ATAC26P*(−0.147) + RNU6-1031P*0.100 + RNU6-335P*(−0.124) + RNU6-98P*0.014.

The score displayed a good diagnostic ability for ASD in the training set (AUC = 0.90), validation set (AUC = 0.87), and the overall (AUC = 0.88) (Fig. 2). The score had a better performance among the females than males in the training set (AUC: 0.98 vs 0.86, $P < .001$) and the overall (AUC: 0.93 vs 0.85, $P = .039$), but it was not significant in the validation set (AUC: 0.89 vs 0.85, $P = .661$) (Fig. 3). It was also more accurate among the younger cases (<6 years) in the training set (AUC: 0.98 vs 0.83, $P < .001$) and the overall (AUC: 0.93 vs 0.84, $P = .002$), but it was not significant in the validation set (AUC: 0.90 vs 0.86, $P = .619$) (Fig. 4).

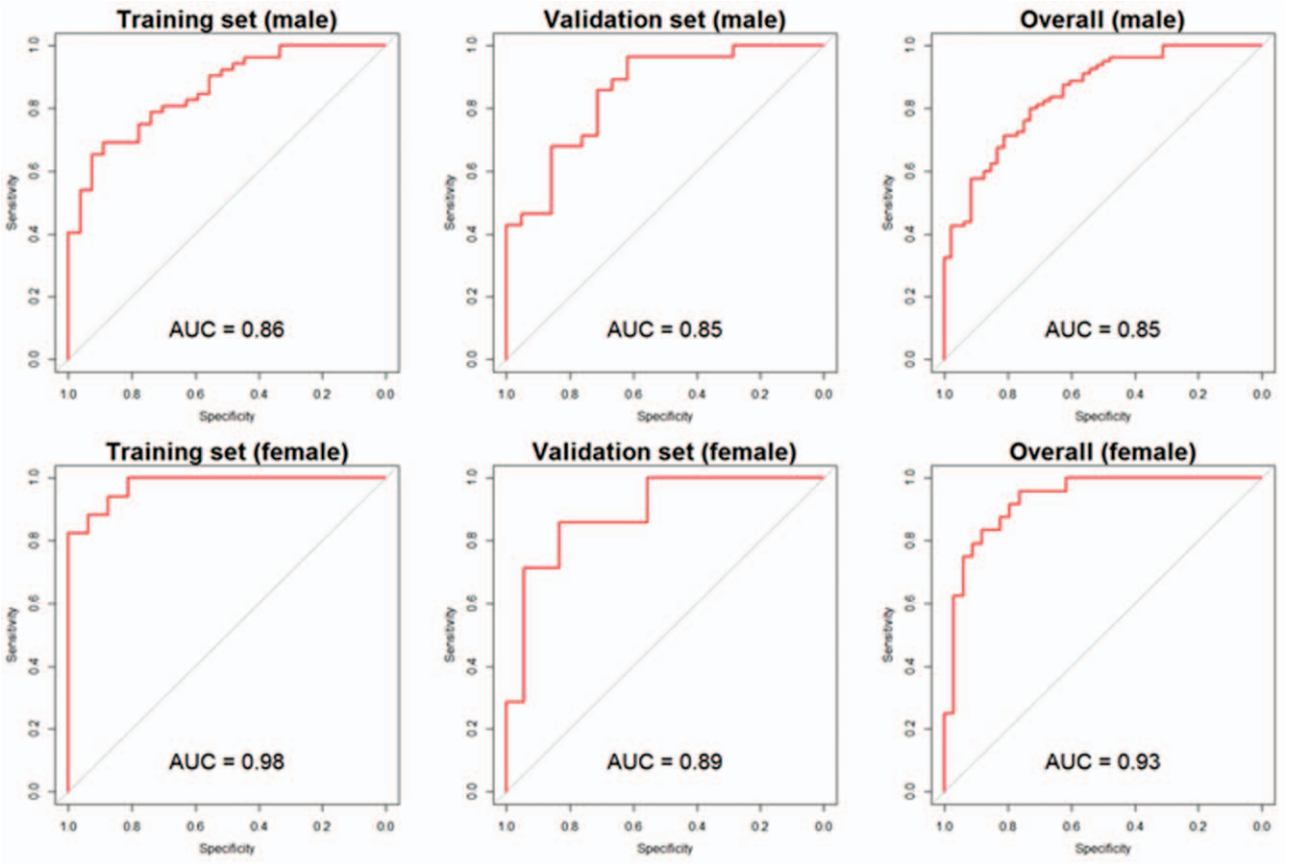

Figure 3. ROC curve analysis of the diagnostic signature in the subgroups of difference genders. AUC = area under receiver operating characteristic curve, ROC = receiver operating characteristic.

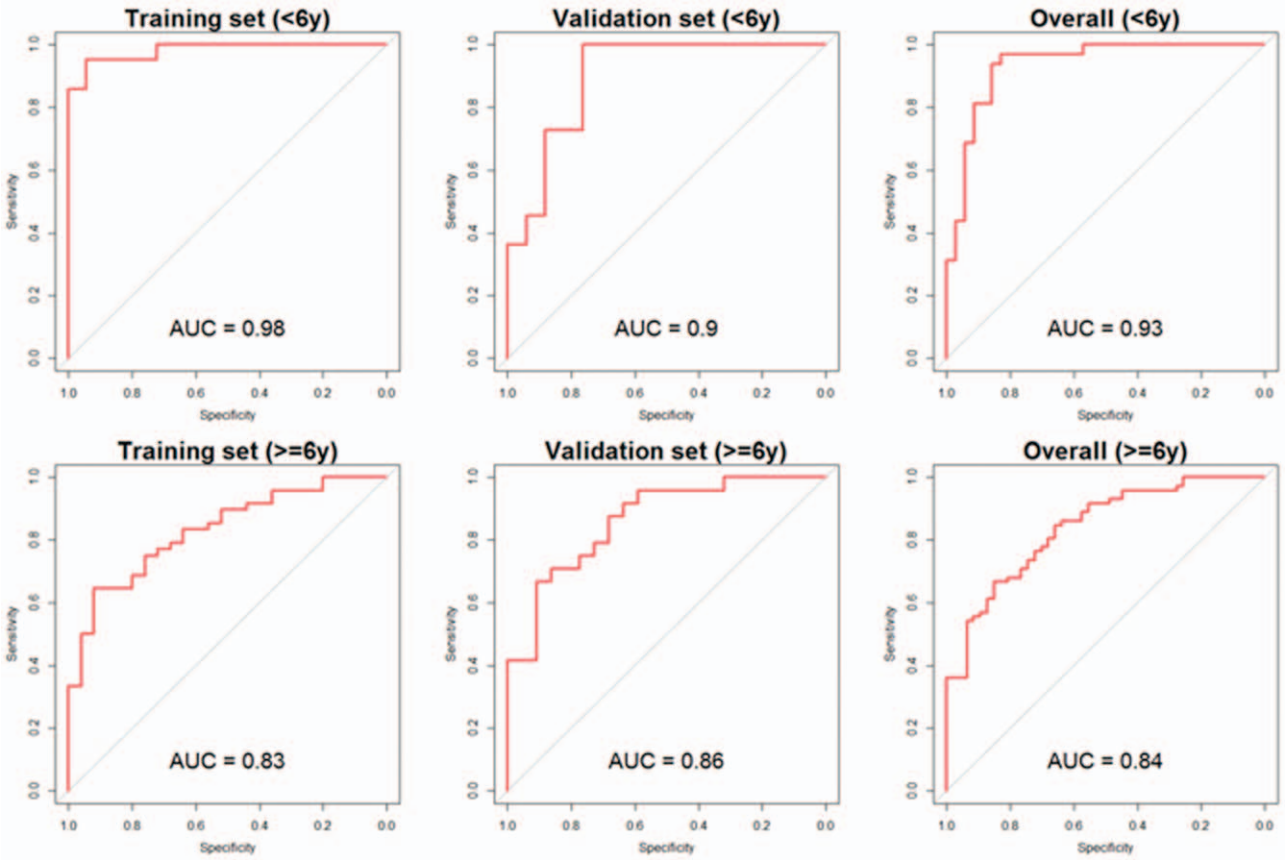

Figure 4. ROC curve analysis of the diagnostic signature in the subgroups of difference ages. AUC = area under receiver operating characteristic curve. ROC = receiver operating characteristic.

As for disease subtypes, the score had a good power in diagnosing autistic disorder (AUC: 0.85 in training set, 0.86 in validation set, and 0.85 in the overall), PDD-NOS (AUC: 0.92 in training set, 0.89 in validation set, and 0.89 in the overall), and Asperger disorder (AUC: 0.91 in training set, 0.89 in validation set, and 0.88 in the overall) (Fig. 5).

3.4. External validation

The blood samples were collected from 23 ASD patients (14 males [60.9%], and average age 8.0 years [2–16]), and 23 age- and gender-matched controls. According to the signature formula, the score displayed a good diagnostic ability for ASD (AUC = 0.88) (Fig. 6).

3.5. Comparison with a 55-gene signature

The 55-gene signature derived from the same dataset. In the training set, the 55-gene signature showed a better performance than the snRNA signature (AUC: 0.99 vs 0.90, $P < .001$), which might contribute to the co-linearity and thus make the regression coefficients unreliable (Fig. 7). In the validation set, the 55-gene signature displayed a poorer performance than the snRNA signature (AUC: 0.87 vs 0.58, $P < .001$). In general, the snRNA signature showed a better diagnostic ability than the 55-gene signature (AUC: 0.88 vs 0.80, $P = .049$).

4. Discussion

In this study, we adopted the methods of probe re-annotation and penalized generalized linear model to identify a novel and robust blood snRNA signature in diagnosing ASD. The signature showed a good stability in the subgroup analyses of age, gender, and disease subtypes. To further validate the robustness of the signature, the blood samples of 23 ASD patients and 23 age- and gender-matched controls were collected, and a customized microarray was used to detect the snRNA expression and calculate the score. The signature also demonstrated a good diagnostic ability for ASD. The 55-gene signature was the first systematic blood transcriptome signature for ASD diagnosis, which had the largest sample size and the cross-validation between 2 large-scale datasets based on different microarray platforms (GPL6244 and GPL570). Compared with the 55-gene signature, the 9-snRNA signature also showed a higher diagnostic efficiency.

The snRNA signature consisted of 9 snRNAs (RNU1-16P, RNU6-1031P, RNU6-258P, RNU6-335P, RNU6-485P, RNU6-549P, RNU6-98P, RNU6ATAC26P, and RNVU1-15). RNU1-16P had a moderate to high expression in nervous system (brain, cortex, and cerebellum) and whole blood (based on the GTEx database). In genome-wide association study (GWAS) Catalog, RNU6-1031P was associated with the human phenotypes of mean corpuscular hemoglobin, red blood cell distribution width, and vital capacity, while RNU6-258P was related with intelligence.[8] Compared with normally developed children, the

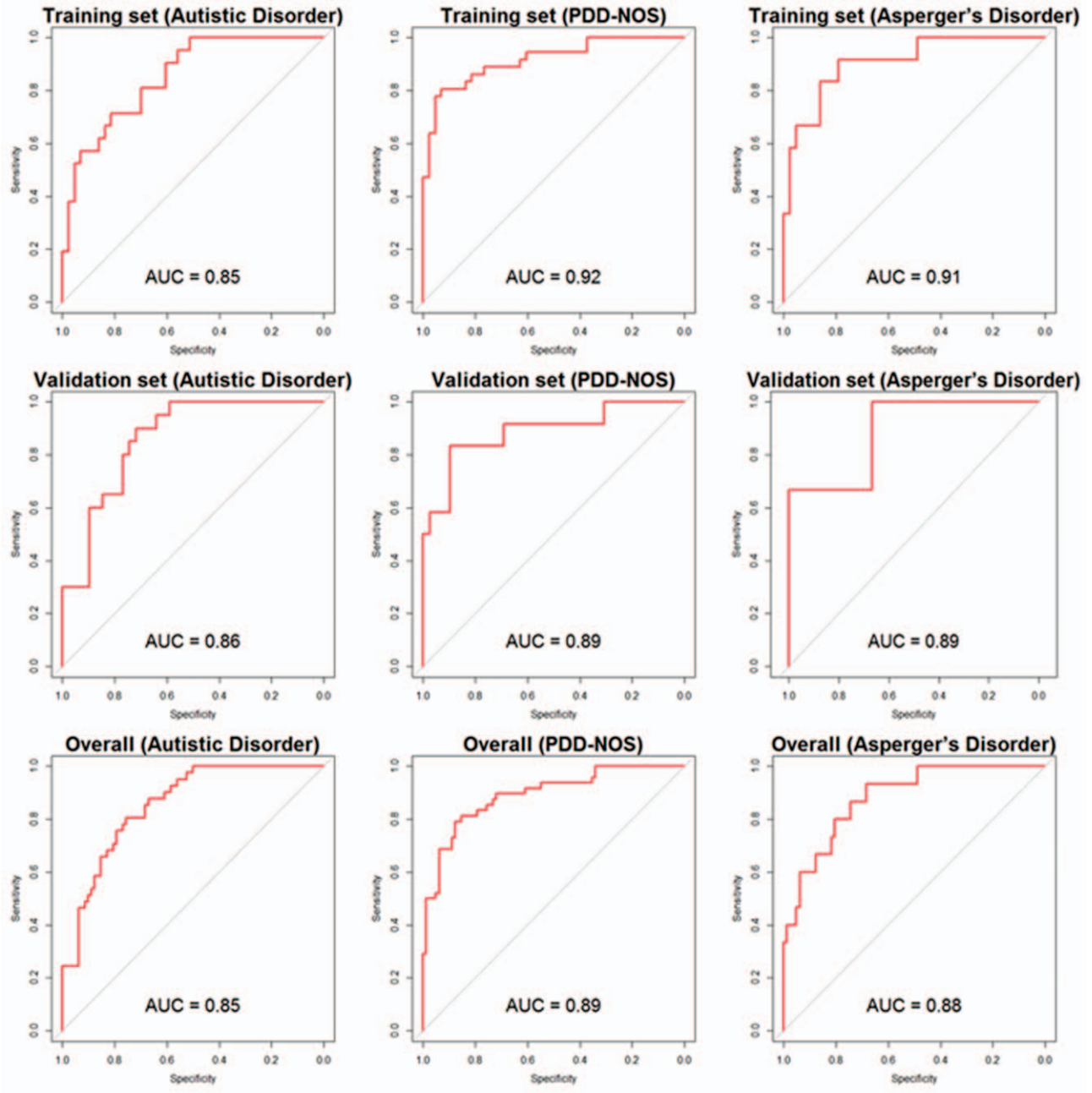

Figure 5. ROC curve analysis of the diagnostic signature in the subgroups of difference disease subtypes. AUC = area under receiver operating characteristic curve. ROC = receiver operating characteristic.

intelligence development in ASD children was significantly delayed.[9,10] There were positive correlations between age and mean corpuscular volume, and red cell distribution width in ASD children, and 24.1% cases had iron deficiency and 15.5% had anemia.[11] RNU6-549P was associated with the GWAS phenotypes of mathematical ability, self-reported educational attainment, coronary artery disease, and factor VII activating protease measurement. Previous studies suggested impaired metacognitive monitoring, mathematics under-achievement, and educational needs in ASD.[12–14] RNU6-98P was associated with the GWAS phenotype of self-reported educational attainment. RNU6ATAC26P had a moderate to high expression in nervous system and whole blood. RNVU1-15 had a moderate to high expression in nervous system and whole blood. It was associated with U1 snRNP (GO:0005685), mRNA 5′-splice site recognition (GO:0000395), pre-mRNA 5′-splice site binding (GO:0030627). A growing number of alternative splicing regulators have been reported in relation with ASD.[15,16]

The limitations in this study should be also acknowledged. First, the sample size is not as large as we expected. Second, the method of probe re-annotation could not cover all snRNAs. In the future, a large-scale prospective designed study was needed to validate this snRNA signature.

In conclusion, through probe re-annotation and penalized generalized linear model, we identified a novel and robust blood snRNA signature in diagnosing ASD, which might help improve

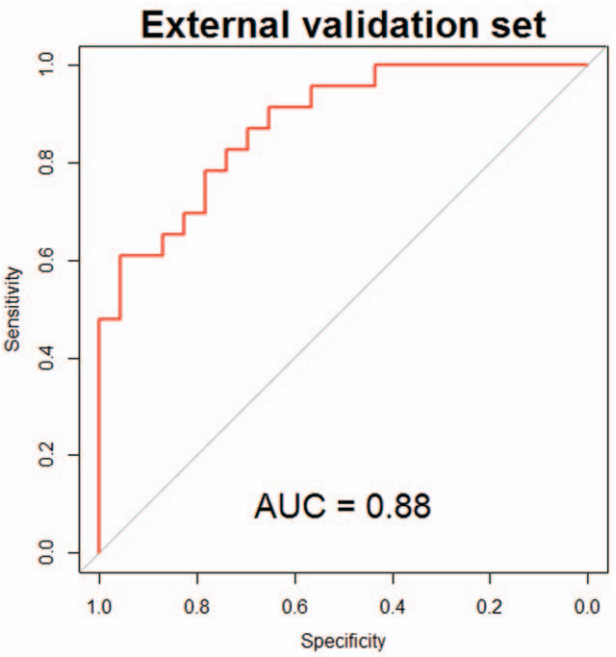

Figure 6. ROC curve analysis of the diagnostic signature in the external validation set. AUC = area under receiver operating characteristic curve. ROC = receiver operating characteristic.

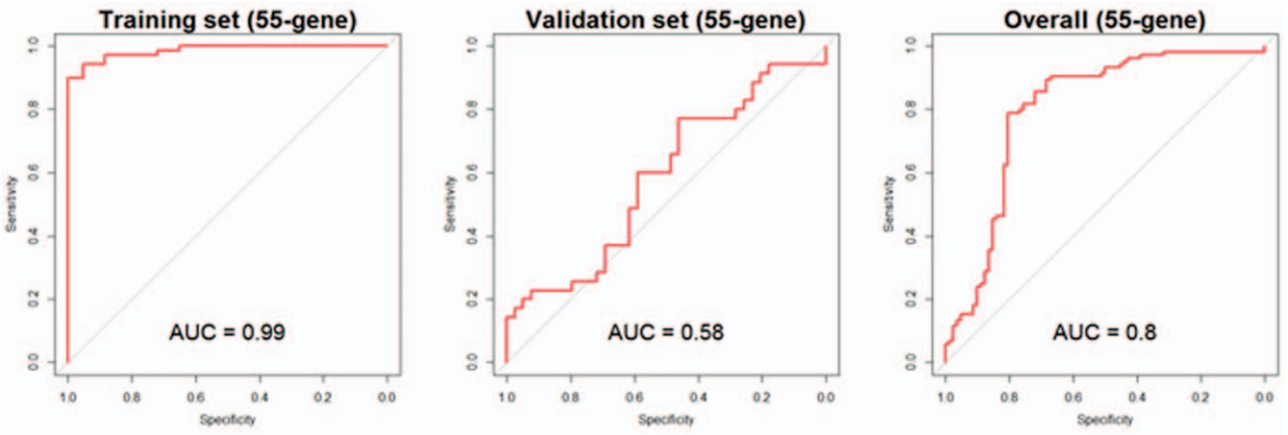

Figure 7. ROC curve analysis of the 55-gene diagnostic signature. AUC = area under receiver operating characteristic curve. ROC = receiver operating characteristic.

the diagnostic accuracy for ASD in clinical practice. Nevertheless, a large-scale prospective study was needed to validate our results.

Author contributions

Data curation: Jinxia Zhou, Qian Hu, Chunlian Pan.
Formal analysis: Jinxia Zhou, Qian Hu.
Investigation: Wei Cheng.
Methodology: Jinxia Zhou, Qian Hu, Xijia Wang, Wei Cheng, Chunlian Pan.
Software: Qian Hu, Wei Cheng.
Supervision: Jinxia Zhou.
Validation: Jinxia Zhou, Xijia Wang.
Visualization: Jinxia Zhou, Xijia Wang, Chunlian Pan.
Writing – original draft: Jinxia Zhou.

References

[1] Simashkova NV, Boksha IS, Klyushnik TP, et al. Diagnosis and management of autism spectrum disorders in russia: clinical-biological approaches. J Autism Dev Disord 2019;49:3906–14.
[2] Christensen DL, Baio J, Van Naarden Braun K, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years–autism and developmental disabilities monitoring network, 11 sites, United States, 2012. MMWR Surveill Summ 2016;65:1–23.
[3] Warren Z, McPheeters ML, Sathe N, et al. A systematic review of early intensive intervention for autism spectrum disorders. Pediatrics 2011;127:e1303–11.
[4] Watson CN, Belli A, Di Pietro V. Small non-coding RNAs: new class of biomarkers and potential therapeutic targets in neurodegenerative disease. Front Genet 2019;10:364.
[5] Bai B, Hales CM, Chen PC, et al. U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. Proc Natl Acad Sci U S A 2013;110:16562–7.

[6] Han LO, Li XY, Cao MM, et al. Development and validation of an individualized diagnostic signature in thyroid cancer. Cancer Med 2018;7:1135–40.
[7] Kong SW, Collins CD, Shimizu-Motohashi Y, et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. PLoS One 2012;7:e49475.
[8] MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res 2017;45:D896–901.
[9] Bedford SA, Park MTM, Devenyi GA, et al. Large-scale analyses of the relationship between sex, age and intelligence quotient heterogeneity and cortical morphometry in autism spectrum disorder. Mol Psychiatry 2019; doi: 10.1038/s41380-019-0420-6.
[10] Dryburgh E, McKenna S, Rekik I. Predicting full-scale and verbal intelligence scores from functional connectomic data in individuals with autism spectrum disorder. Brain Imaging Behav 2019; doi: 10.1007/s11682-019-00111-w.
[11] Hergüner S, Keleşoğlu FM, Tanıdır C, et al. Ferritin and iron levels in children with autistic disorder. Eur J Pediatr 2012;171:143–6.
[12] Maras K, Gamble T, Brosnan M. Supporting metacognitive monitoring in mathematics learning for young people with autism spectrum disorder: a classroom-based study. Autism 2017;23: 60–70.
[13] Hetzroni O, Agada H, Leikin M. Creativity in autism: an examination of general and mathematical creative thinking among children with autism spectrum disorder and children with typical development. J Autism Dev Disord 2019;49:3833–44.
[14] Saggers B, Tones M, Dunne J, et al. Promoting a collective voice from parents, educators and allied health professionals on the educational needs of students on the autism spectrum. J Autism Dev Disord 2019;49:3845–65.
[15] Singh RK, Cooper TA. Pre-mRNA splicing in disease and therapeutics. Trends Mol Med 2012;18:472–82.
[16] Gonatopoulos-Pournatzis T, Wu M, Braunschweig U, et al. Genome-wide CRISPR-Cas9 interrogation of splicing networks reveals a mechanism for recognition of autism-misregulated neuronal microexons. Mol Cell 2018;72:510–24.e12.