Research Paper

An artificial neural network model to diagnose non-obstructive azoospermia based on RNA-binding protein-related genes

Fan Peng[1,*], Bahaerguli Muhuitijiang[2,3,*], Jiawei Zhou[2,3,*], Haoyu Liang[4], Yu Zhang[1], Ranran Zhou[1,3]

[1]Department of Urology, Baoan Central Hospital of Shen Zhen, Shenzhen 518102, China
[2]Department of Urology, Nanfang Hospital, Southern Medical University, Guangzhou 510000, China
[3]The First School of Clinical Medicine, Southern Medical University, Guangzhou 510000, China
[4]Department of Urology, The Third Affiliated Hospital, Southern Medical University, Guangzhou 510000, China
[*]Equal contribution

Correspondence to: Ranran Zhou; **email:** 2966075781@qq.com, https://orcid.org/0000-0003-2464-0646
Keywords: machine learning, artificial neural network, diagnosis, non-obstructive azoospermia, RNA-binding protein
Received: November 16, 2022 **Accepted:** April 15, 2023 **Published:** April 24, 2023

ABSTRACT

Non-obstructive azoospermia (NOA) is a severe form of male infertility, but its pathological mechanisms and diagnostic biomarkers remain obscure. Since the dysregulation of RNA-binding proteins (RBPs) had nonnegligible effects on spermatogenesis, we aimed to investigate the functions and diagnosis values of RBPs in NOA. 58 testicular samples (control = 11, NOA = 47) from Gene Expression Omnibus (GEO) were set as the training cohort. Three public datasets, containing GSE45885 (control = 4, NOA = 27), GSE45887 (control = 4, NOA = 16), and GSE145467 (control = 10, NOA = 10), and 44 clinical samples from the local hospital (control = 27, NOA = 17) were used for validation. Through a series of bioinformatical analyses and machine learning algorithms, including genomic difference detection, protein-protein interaction network analysis, LASSO, SVM-RFE, and Boruta, DDX20 and NCBP2 were determined as significant predictors of NOA. Single-cell RNA sequencing of 432 testicular cell samples from NOA patients indicated that DDX20 and NCBP2 were associated with spermatogenesis (false discovery rate < 0.05). Based on the transcriptome expressions of DDX20 and NCBP2, we constructed multiple diagnosis models using logistic regression, random forest, and artificial neural network (ANN). The ANN model exhibited the most reliable predictive performance in the training cohort (AUC = 0.840), GSE45885 (AUC = 0.731), GSE45887 (AUC = 0.781), GSE145467 (AUC = 0.850), and local cohort (AUC = 0.623). Totally, an ANN diagnosis model based on RBP DDX20 and NCBP2 was developed and externally validated in NOA, functioning as a promising tool in clinical practice.

INTRODUCTION

As the most grievous situation of male infertility, azoospermia refers to the complete absence of spermatozoa in the ejaculate. The prevalence of azoospermia is fairly high, accounting for 1% of males and over 10% of infertile males [1]. Around 30% of azoospermia patients exhibit obstructive azoospermia (OA) due to the physical blockage of the sperm outflow tract, and these cases usually show normal spermatogenesis. Non-obstructive azoospermia (NOA), which accounts for 70% of the azoospermia cases, is characterized by spermatogenetic failure and testicular dysfunction [2]. NOA is mainly caused by various genetic diseases, adverse drug effects, and malignant tumors [2, 3]. Therefore, the surgical removal of the blockage and the testicular puncture are valid in OA, while the current treatment strategies for

NOA often fail [4]. Seeking the underlying biological mechanisms of NOA has attracted increasing attention in the past decades, and some critical genes were uncovered, such as VASA [5], CHD5 [6], and MEIOB [7], providing possible therapeutic targets. Nevertheless, our understandings of the genetic alternation of NOA remain limited. A prominent manifestation is that the examination of any NOA-specific gene has not been recommended in current clinical guidelines [8–10].

RNA-binding proteins (RBPs) are a group of proteins capable of regulating a plethora of cellular post-transcriptional processes, the perturbation of which leads to impaired spermatogenesis. Previous studies have demonstrated that RBPs tremendously influence the mammalian reproductive system. For instance, RBP Rbm46 knockout mice had reduced testes size and spermatogenetic defect and thus were infertile [11]; Boule was able to bind to a tremendous amount of spermatogenesis-related mRNAs and was involved in the spermatogenetic process in mice testes through forming amyloid-like aggregation both *in vivo* and *in vitro* [12]; The loss of RBP Tulp2 led to infertility in male mice by reducing the quantity and quality of sperms [13]. Given that a large number of proteins need to be properly expressed during spermato-genesis, it is inevitable that RBPs exert nonnegligible functions in the spermatogenetic process and, naturally, in the initiation and progression of NOA. However, the number of studies focusing on the relationship between RBPs and NOA is currently inadequate.

The present study collected the RBPs from previous reports and compared their expression levels between control and NOA testicular samples. Multiple bioinformatical and machine learning algorithms, including protein-protein interaction (PPI) network analysis, least absolute shrinkage and selection operator (LASSO), support vector machine-recursive feature elimination (SVM-RFE), and Boruta, were adopted for feature selection. Logistic regression (LR), random forest (RF), and artificial neural network (ANN) were harnessed to construct the diagnosis models. The GSE9210 dataset obtained from Gene Expression Omnibus (GEO) was set as the training cohort, while GSE45885, GSE45887, and GSE145467 from GEO were used for external validation. Importantly, we collected the seminal plasma and testicular tissue of 27 control and 17 NOA samples from the local hospital to re-confirm the reliability of the models. Single-cell RNA sequencing (scRNA-seq) data of 432 testicular cell samples from NOA patients were used to investigate the association of the unearthed genes and spermatogenesis.

MATERIALS AND METHODS

Data retrieval and processing

A sum of 1542 RBPs was gleaned from the report of Gerstberger et al. [14], as listed in Supplementary Table 1. The transcriptome sequencing data of the testicular tissue from 11 control and 47 NOA patients in the GSE9210 dataset [15] was directly downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/) as the training dataset. At the same time, GSE45885 [16] (control = 4, NOA = 27), GSE45887 [17] (control = 4, NOA = 16), and GSE145467 (control = 10, NOA = 10) were obtained from GEO as the external validation datasets. The control cases were defined as the subjects with normal spermatogenesis, including healthy donors and OA patients. The chip probe IDs were converted into gene symbols using R software (version 4.1.0) following the annotation files. The average expression value would be adopted if multiple probe IDs corresponded to the same gene symbol. The RNA expression values in these cohorts were all normalized with log2 (x + 1) transformation, and the sva package in R software was used to reduce the batch effects of these experiments as possible.

The scRNA-seq of 432 testicular cell samples from NOA patients was obtained from GEO (GSE157421) [18] to investigate the association of the screened genes with spermatogenesis. The processing of analyses of the scRNA-seq was described in the previous study in detail [19]. More information on these public datasets mentioned above is shown in Table 1.

Clinical sample collection

A total of 44 participants, including 27 cases with normal spermatogenesis (OA) and 17 NOA subjects, were enrolled in this project between January 2021 and May 2022 at the Bao'an Central Hospital of Shenzhen (China). The study protocol was reviewed and approved by the Ethics Committee of Bao'an Central Hospital of Shenzhen, and all the patients signed the informed consent. The paraffin-embedded testicular biopsy specimens of these patients were provided by the Department of Pathology in Bao'an Central Hospital of Shenzhen. The semen samples were collected by masturbation following 3–5 days of sexual abstinence. The seminal supernatant plasma was obtained after centrifuging semen at 3,000 g for 20 min and then immediately stored at −80°C for RNA extraction. Other critical clinicopathological parameters, inclusive of age, Johnsen's Score, follicle-stimulating hormone (FSH) levels, luteinizing hormone (LH) levels, and testosterone (T) levels, were retrospectively documented as well.

Table 1. The detailed information of the public datasets from GEO.

ID	Platform	Experimental type	Tissue	Control	NOA	Region
GSE9210	GPL887	Microarray	Testes	11	47	Japan
GSE45885	GPL6244	Microarray	Testes	4	27	Norway
GSE45887	GPL6244	Microarray	Testes	4	16	Norway
GSE145467	GPL4133	Microarray	Testes	10	10	Unknown
GSE157421	GPL20301	Single-cell RNA sequencing	Testes	–	432	China

Abbreviations: GEO: gene expression omnibus; NOA: non-obstructive azoospermia.

Table 2. The primer sequence used in this study.

Gene	Sequence (5′–3′)
NCBP2	F: AAAACGCCATGCGGTACATAA
	R: GCCTGCCCTCCTTAAAGCC
DDX20	F: GCTGCGGGCTCGATTTAATTG
	R: GTCCAAAGCTATGGTGGAGAAC
GAPDH	F: GGAGCGAGATCCCTCCAAAAT
	R: GGCTGTTGTCATACTTCTCATGG

Real-time quantitative PCR experiments

We conducted real-time quantitative PCR (RT-qPCR) experiments to quantify the mRNA expression levels of the screened genes in the seminal plasma of the local cohort. The total RNA of the seminal plasma samples was isolated using TRIzol Reagent (Invitrogen, USA) following the manufacturer's protocols. The PrimeScript RT Reagent Kit (Takara, China) was used to perform the reverse transcription. The PCR experiments were then carried out based on ABI 7600 system (Applied Biosystems, USA) with the SYBR Premix ExTaq kit (Takara, China). GAPDH was chosen as the internal reference gene, and all the detected values were normalized with the $2^{-\Delta\Delta Ct}$ method. The primer sequence of GAPDH, DDX20, and NCBP2 was designed and synthesized by the TSINGKE Company (Guangzhou, China), which is shown in Table 2.

Immunohistochemical staining

The immunohistochemical (IHC) staining of the paraffin-embedded testicular samples of the local cohort was implemented to investigate the protein expression levels and distribution of DDX20. The testicular slides were de-paraffinized in xylene and then added to the ethanol following the below concentration: 100% ethanol (4 min), 90% ethanol (4 min), 80% ethanol (4 min), and 70% ethanol (4 min). Subsequently, the slides were blocked in phosphate-buffered saline (PBS) supplemented with 5% bovine serum albumin (BSA) for 1 hour at room temperature and incubated with the primary antibody (rabbit anti-DDX20, 1:100, Proteintech, China) at 4°C overnight. After washing the slides 3 times with PBS, we incubated the slices with anti-rabbit secondary antibodies (Proteintech, China). Nikon Eclipse 90i system (Nikon, Japan) was used to get the images, and the Image-Pro Plus (version 6.0, Media Cybernetics, USA) software was used to measure the integral optical density (IOD), which represented the protein expression levels. For each sample, 3 slices were randomly chosen to conduct the IHC staining, and 5 different microscopic fields of a slide were randomly selected to evaluate the levels of DDX20.

Gene expression difference analysis and functional annotation

The mRNA expression divergence of the 1542 RBPs between the control and NOA testicular samples was detected with the limma package in R. The filtering criteria were set as follows: $|logFC| > 1$ and false discovery rate (FDR) < 0.05. The biological gene function enrichment was performed using the Metascape online tool (https://metascape.org/) to identify the associated Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways with $P < 0.05$ filtering threshold.

PPI network construction and analysis

We uploaded the differentially-expressed RBPs to the STRING database (https://string-db.org/) to construct the PPI network to investigate the interaction of these genes. The confidence level was set to 0.4, and the genes without association with other nodes were

excluded. The Cytoscape software (version 3.8.0) was utilized to visualize the PPI network. The cytoHubba plug-in in the Cytoscape software was used to measure the importance of the genes in the network. We chose the Top 20 genes showing the highest degree for further analyses.

Feature selection via machine learning algorithms

We employed multiple feature selection algorithms to investigate the significant diagnosis biomarkers in NOA, including LASSO, SVM-RFE, and Boruta. LASSO Binomial regression with nested cross-validation to select the optimal predictor was built using the glmnet R package. SVM-RFE algorithm, which was based on the backward feature elimination that recursively removes the least ranking feature, was conducted by the caret package. The Boruta algorithm, which was built around the random forest, was also used to remove the irrelevant and redundant features through the Boruta package in R, where the variables labelled with "Confirmed" were identified. Following these results, we included the RBPs con-determined by LASSO, SVM-RFE, and Boruta in the diagnosis model development.

Construction and validation of diagnosis models using LR, RF, and ANN

We implemented 3 commonly used machine learning-based methods to construct the diagnosis models, including LR, RF, and ANN, to improve the predictive power and robustness. The LR model was developed based on the "glm" function with default settings in R software. The RF model was established using the randomForest package with the following parameters: ntree = 500, mtry = 3, importance = T, and proximity = T. The ANN model was constructed according to the neuralnet R package, which contained one input layer, one hidden layer, and one output layer. In the hidden layer, we applied 5 nodes, and rectified linear unit was utilized as an activation function. Two nodes (control and NOA) were set in the output layer, where a softmax function was employed. According to the ANN model, the classification score of each subject was calculated.

To ensure the comparability of the LR, RF, and ANN models, we regarded that the sample would be classified as a control case if its probability predicted by the LR and ANN models was less than 0.5; otherwise, it would be considered an NOA sample. The receiver operating characteristic (ROC) analyses were performed to measure the predictive performance of the models in different cohorts through the pROC package. The confusion matrices, and other statistical indexes, such as accuracy, precision, recall, F-measure, sensitivity, specificity, positive predictive value, and negative predictive value, were also applied in this study.

The functionally-related genes

The Top 20 genes interacting with the screened genes were obtained from the GeneMANIA database (http://genemania.org/) with default settings. The interaction types included physical interactions, co-expression, prediction, co-localization, genetic interactions, pathways, and shared protein domains.

Gene set enrichment analysis

Here, we adopted the single-gene gene set enrichment analysis (GSEA) strategy to investigate the association between the screened genes and spermatogenesis. According to the median expression value of the particular gene, 432 testicular cell samples were divided into high- and low-gene expression groups, followed by the GSEA analysis. The GSEA was conducted via the GSEA software (version 4.1.0) with default settings, and the reference gene sets (Hallmark version 7.2) were downloaded from the Molecular Signature Database (https://www.gsea-msigdb.org/gsea/msigdb/). The term with Nominal $P < 0.05$ and FDR < 0.05 was considered to be statistically significant.

Statistical analyses

The statistical analyses of the whole study were based on the R software (version 4.1.0) and GraphPad Prism 8 (version 8.4.3). The data in this study are presented as n (%) or mean ± standard deviation (SD). The two-tailed Student's *t*-test was used to compare the difference in the RT-qPCR experiments, and The Welch-corrected *t*-test was used for IODs. Unless otherwise specified, $P < 0.05$ was significant. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

RESULTS

51 RBPs were differentially expressed between control and NOA samples

The schematic workflow chart, which graphically describes the methodology of this study, is presented in Figure 1. First, we compared the mRNA expression levels of 1542 RBPs collected from previous studies in the testicular samples of control and NOA cases in the GSE9210 cohort. The results indicated that 51 of 1542 RBPs were differentially expressed (Supplementary Table 2), as shown in the volcano plot (Figure 2A) and heat map (Figure 2B). The functional annotation displayed that the 51 differentially-expressed genes (DEGs) were mainly involved in translation regulation,

RNA metabolism, RNA stability regulation, and spermatogenetic process, implying the tremendous effect of RBPs on the pathogenesis of NOA (Figure 2C).

PPI network construction

We constructed the PPI network to further explore the internal contact and interactions among the 51 DEGs at the protein level. Figure 3A illustrates the established PPI network, where the size of the nodes represented the absolute value of their corresponding logFC in the GSE9210 cohort. The importance and influence of the genes in the network were quantified as degrees, and the Top 20 genes with the highest degree were identified and selected for the next step of research (Figure 3B, Supplementary Table 3).

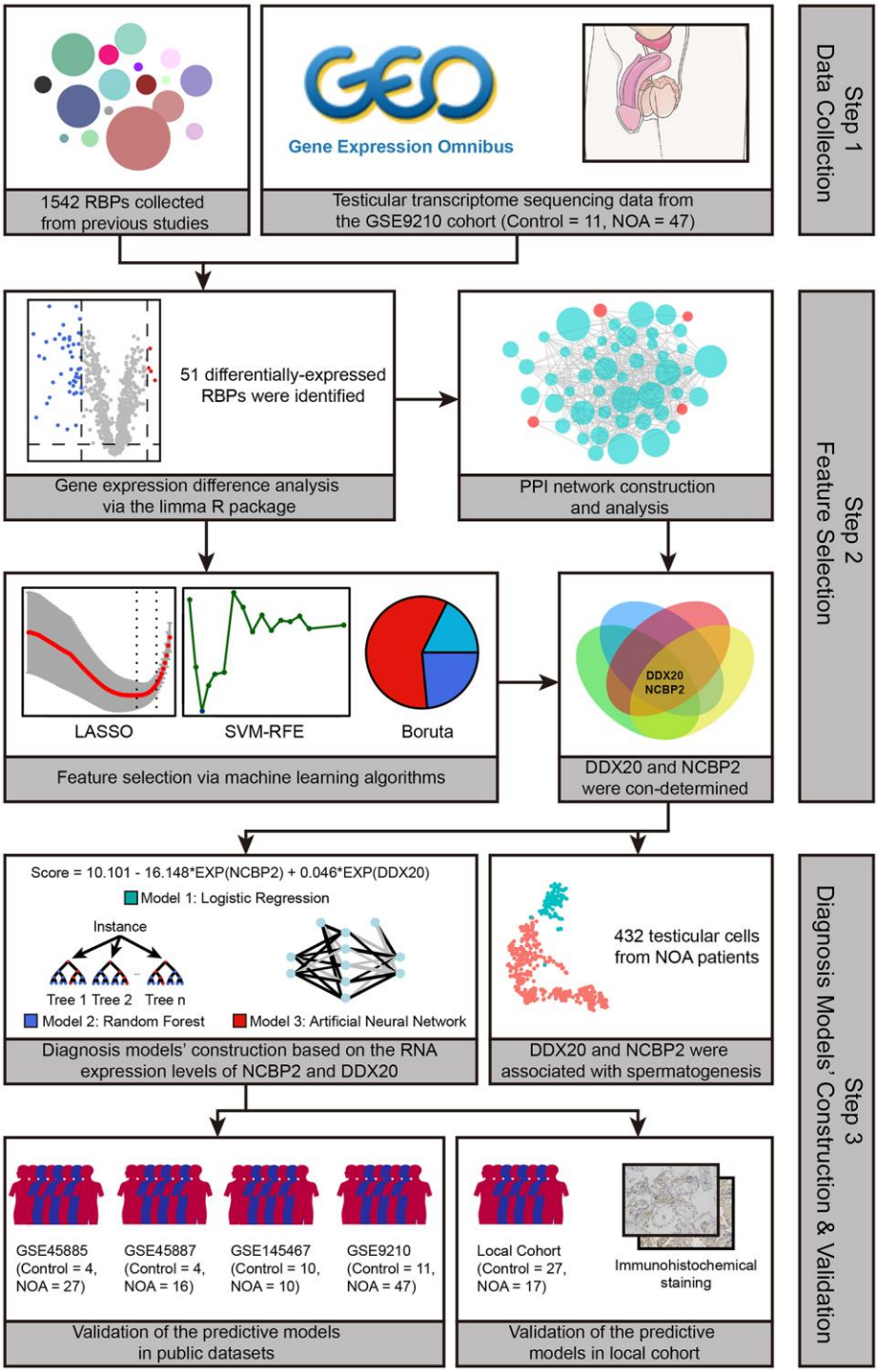

Figure 1. The workflow of the present study.

DDX20 and NCBP2 were identified via feature selection algorithms and PPI network analysis

5 genes were identified as important features of NOA through LASSO regression (Figure 4A), including NCBP2, DDX20, TSN, SRPK2, and CARHSP1. The coefficients of these genes in the LASSO regression model were −0.121, −0.703, −0.770, −0.921, and −0.178, respectively (Figure 4B). Simultaneously, 30 of 51 DEGs were selected via the Boruta algorithm (Figure 4C, Supplementary Table 4), and 6 genes, including NCBP2, DDX20, CCDC86, TSN, CARHSP1, and TDRD7, were screened by SVM-RFE (Figure 4D). Ultimately, NCBP2 and DDX20 were identified by integrating the Top 20 genes with the highest degree in the PPI network and these feature selection results (Figure 4E) and then included in the diagnosis model construction.

External validation of DDX20 and NCBP2

The scRNA-seq data of 432 testicular cell samples indicated that DDX20 (Nominal $P < 0.001$, FDR < 0.001) and NCBP2 (Nominal $P < 0.01$, FDR < 0.05) were both positively associated with the spermatogenetic process (Figure 5A), re-confirming that

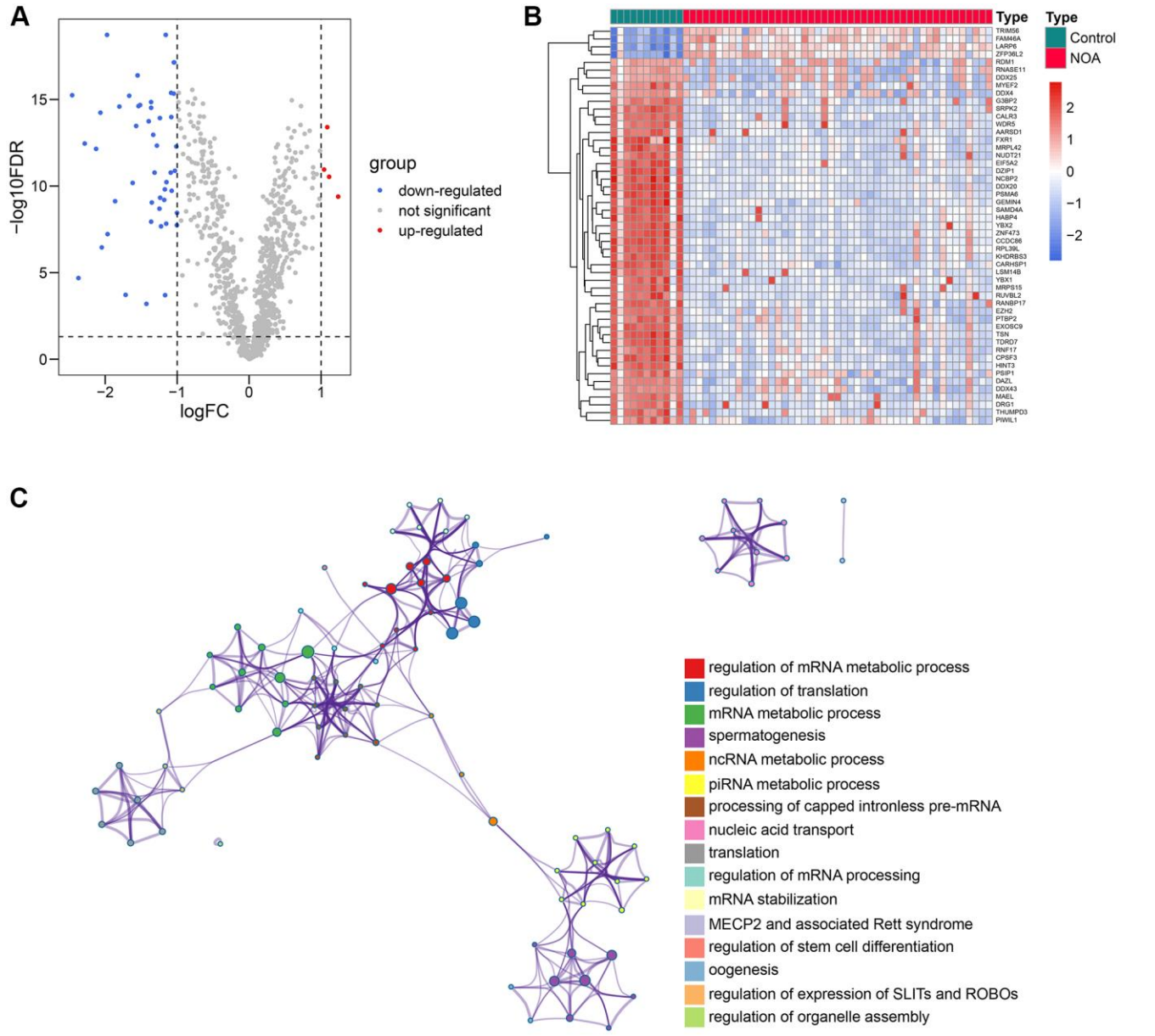

Figure 2. 51 differentially-expressed RBPs and their functional enrichment. (**A**, **B**) The volcano plot (**A**) and the heat map (**B**) indicated that 51 of 1542 RBPs were differentially expressed between the control and NOA testicular samples. (**C**) The functional annotation of the 51 RBPs. Abbreviations: RBP: RNA-binding protein; NOA: non-obstructive azoospermia.

DDX20 and NCBP2 were significant biomarkers to NOA. Next, we gleaned the seminal plasma and testicular biopsy of 27 control and 17 NOA patients from the local hospital for validation. Compared with the control samples, the NOA samples exhibited lower mRNA levels of DDX20 ($P < 0.01$, Figure 5B) and NCBP2 ($P < 0.05$, Figure 5C) in seminal plasma, suggesting that the levels of DDX20 and NCBP2 in seminal plasma were also promising diagnostic biomarkers for NOA. ROC analysis showed that DDX20 in seminal plasma was a powerful classifier for NOA (area under the curve [AUC] = 0.826, 95% confidence interval [CI] = 0.706–0.946, Figure 5D), while the predictive performance of NCBP2 was relatively low (AUC = 0.693, 95% CI = 0.534–0.852, Figure 5D), which might be caused by the heterogeneity across different cohorts. Hence, we then investigated the protein levels of DDX20 in the local cohort using IHC staining, and the results supported the conclusion drawn before that DDX20 was significantly down-regulated in NOA testicular samples ($P < 0.05$, Figure 5E).

The performance of LR, RF, and ANN diagnosis models

The present study utilized multiple datasets, including GSE9210, GSE45885, GSE45887, and GSE145467, and local clinical samples to verify the predictive ability of the established model. The detailed clinicopathological parameters of the training and external validation cohorts are displayed in Table 3. It should be stated that we used the mRNA expression values in the seminal plasma, other than in the testicular samples, to validate the models in the local cohort because the fresh testicular samples were unavailable given the policy formulated by the ethics committee of our hospital. Since we have detected the expressions of DDX20 and NCBP2 in the seminal plasma and found that both genes were down-regulated in NOA samples, which corresponded to the results in the training dataset, we thought that the validation in the seminal plasma samples from the local cohort was still acceptable.

First, an LR diagnosis model was constructed as follows: Score = 10.101–16.148 × *EXP*(*NCBP*2) + 0.046 × *EXP*(*DDX*20), where the EXP meant the mRNA expression value of the gene. The predictive ability of the LR model in the training cohort was quite high (AUC = 0.955, 95% CI = 0.865–1.000, Figure 6A). However, its performance in the GSE45885 cohort (AUC = 0.514, 95% CI = 0.256–0.772, Figure 6B) and the GSE45887 cohort (AUC = 0.531, 95% CI = 0.267–0.795, Figure 6C) was non-ideal. The AUCs of the LR model in the GSE145467 and local cohorts are 0.700

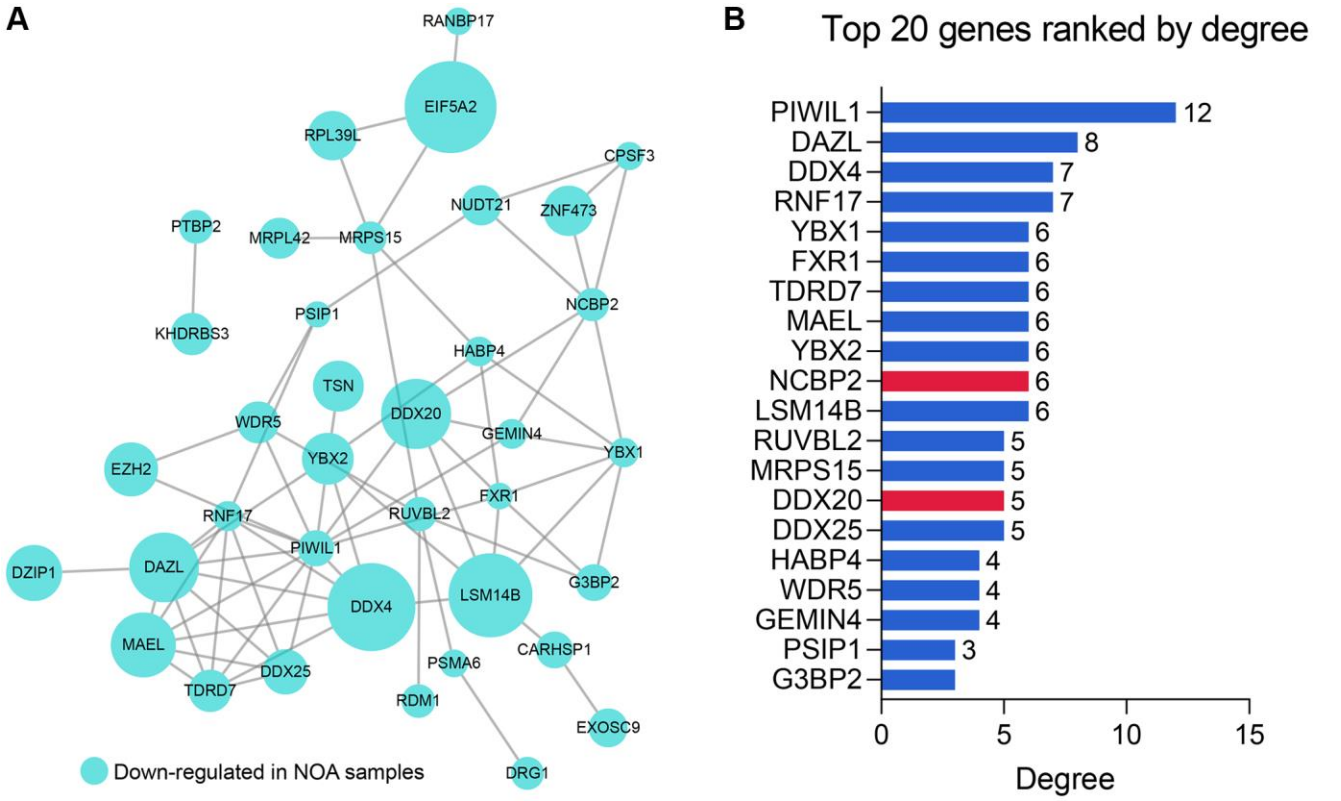

Figure 3. The PPI network analysis of the 51 RBPs. (**A**) The PPI network of the 51 RBPs, where the genes unconnected with other genes were excluded. (**B**) The Top 20 genes with the highest degree in the PPI network. Abbreviation: PPI: protein-protein interaction network.

(95% CI = 0.493–0.907, Figure 6D) and 0.597 (95% CI = 0.465–0.729, Figure 6E), respectively. The confusion matrices of the LR model in these cohorts are shown in Figure 6F–6J, respectively. Generally, the predictive ability of the LR model was far from satisfactory, especially in the external validation cohorts, enlightening us to utilize more tools to construct the diagnosis models.

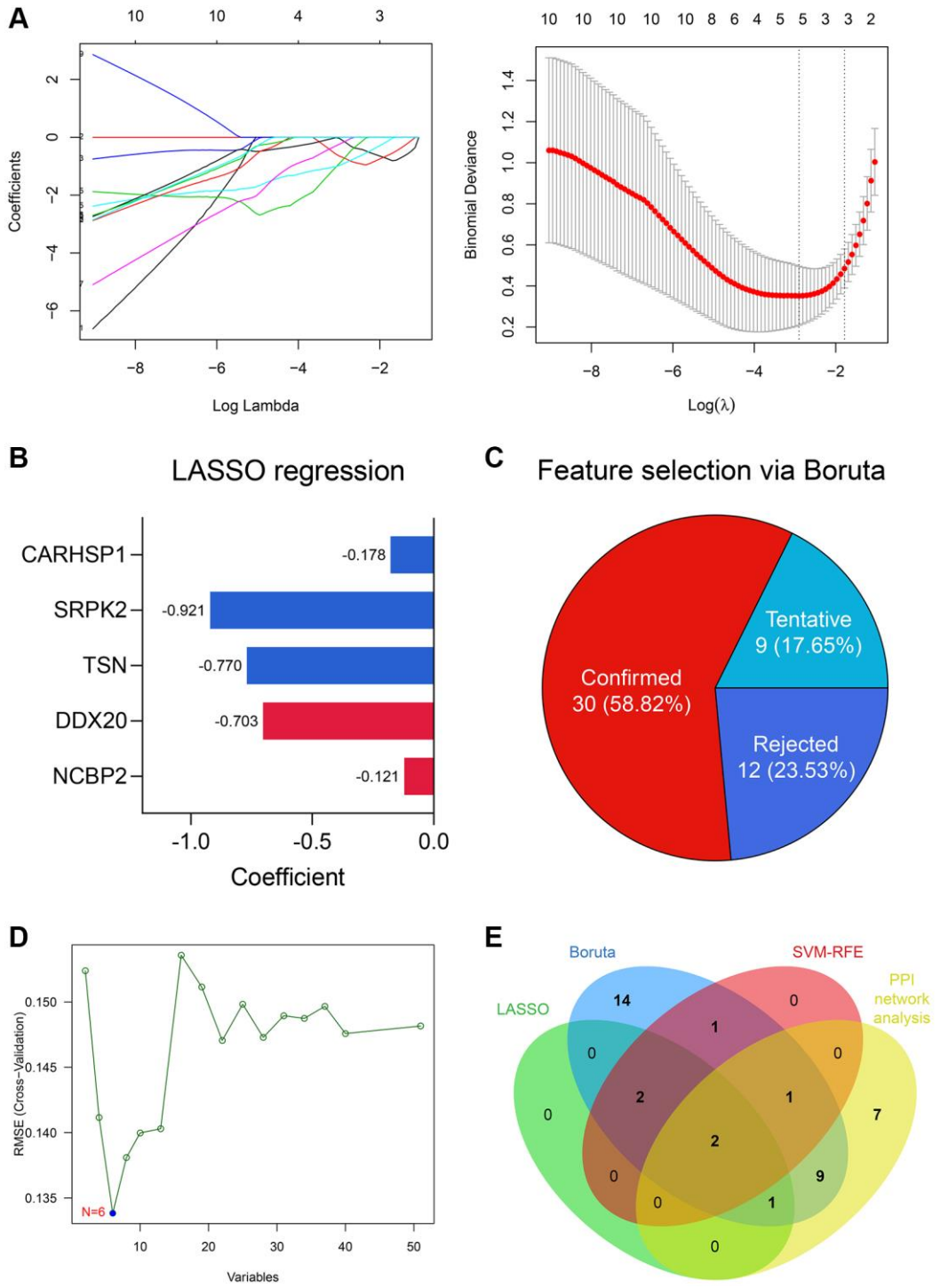

Figure 4. DDX20 and NCBP2 were con-determined via feature selection methods and PPI network analysis. (**A**) 5 genes, including NCBP2, DDX20, TSN, SRPK2, and CARHSP1, were identified as significant features to NOA via LASSO regression. (**B**) The coefficients of the 5 selected genes in the LASSO regression model. (**C**) 30 genes were determined as important features via the Boruta algorithm. (**D**) 6 genes, including NCBP2, DDX20, CCDC86, TSN, CARHSP1, and TDRD7, were selected by the SVM-RFE. (**E**) DDX20 and NCBP2 were con-determined by the machine learning algorithms and PPI network analysis.

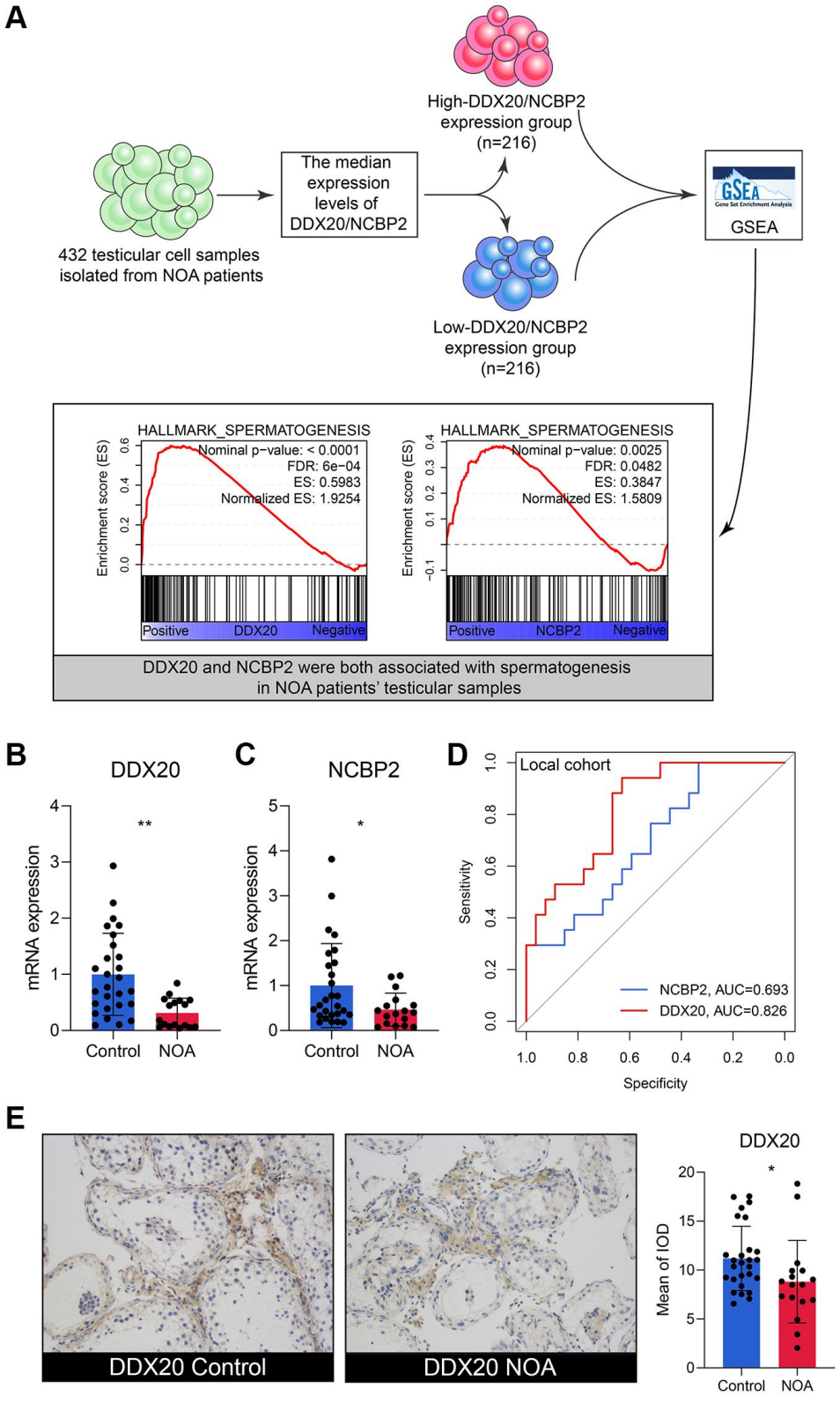

Figure 5. The external validation of DDX20 and NCBP2. (**A**) The single-cell RNA-sequencing analysis of 432 testicular cell samples isolated from NOA patients displayed that DDX20 and NCBP2 were both positively associated with spermatogenesis. (**B**, **C**) DDX20 (**B**) and NCBP2 (**C**) were down-regulated in the seminal plasma samples of NOA patients from the local hospital. (**D**) The ROC curve exhibited the diagnosis ability of seminal plasma DDX20 and NCBP2 to NOA in the local cohort. (**E**) The protein expression levels of DDX20 were obviously down-regulated in the testicular samples from NOA patients in the local cohort, which were detected by the immunohistochemical staining. Abbreviation: ROC: receiver operating characteristic.

Table 3. The clinicopathological features of the cohorts enrolled in this study.

Characteristics	GSE9210		GSE45885		GSE45887		GSE145467		Local cohort	
	Control (*n* = 11)	NOA (*n* = 47)	Control (*n* = 4)	NOA (*n* = 27)	Control (*n* = 4)	NOA (*n* = 16)	Control (*n* = 10)	NOA (*n* = 10)	Control (*n* = 27)	NOA (*n* = 17)
Age (years)	33.3 ± 8.5	35.0 ± 5.7	–	32.1 ± 4.05	–	31.3 ± 1.8	–	–	31.8 ± 9.2	33.4 ± 7.4
Johnsen's score	7.9 ± 1.2	2.4 ± 1.3	–	4.9 ± 2.5	–	–	–	–	7.3 ± 1.7	2.7 ± 2.5
FSH (mIU/ml)	10.1 ± 9.3	29.2 ± 9.1	–	–	–	–	–	–	11.1 ± 7.6	21.6 ± 8.8
LH (mIU/ml)	4.5 ± 2.3	8.8 ± 4.8	–	–	–	–	–	–	4.3 ± 0.9	8.1 ± 3.5
T (ng/ml)	4.8 ± 1.7	3.5 ± 1.6	–	–	–	–	–	–	5.1 ± 0.7	3.7 ± 1.2

Abbreviations: FSH: follicle-stimulating hormone; LH: luteinizing hormone; T: testosterone; NOA: non-obstructive azoospermia.

Subsequently, we established an RF model to classify the NOA samples. The RF model showed superiority to the routine LR model in the training dataset (AUC = 1.000, 95% CI = 1.000–1.000, Figure 7A), GSE45885 dataset (AUC = 0.676, 95% CI = 0.385–0.967, Figure 7B), GSE45887 dataset (AUC = 0.656, 0.381–0.932, Figure 7C), GSE145467 dataset (AUC = 0.750, 95% CI = 0.562–0.938, Figure 7D), and local cohort (AUC = 0.656, 95% CI = 0.547–0.765, Figure 7E). Figure 7F–7J represent the confusion matrices of the RF model in each cohort.

ANN was also a widely-used method for diagnosis model establishment, and many ANN diagnosis models have been proposed and exhibited high reliability and precision in multiple diseases [20–22]. Thus, we then developed an ANN diagnosis model based on the expressions of DDX20 and NCBP2 in NOA, as displayed in Figure 8A. Similar to those previous contributions, the established ANN model showed high predictive performance across the training cohort (AUC = 0.840, 95% CI = 0.773–0.908, Figure 8B), GSE45885 cohort (AUC = 0.731, 95% CI = 0.446–1.000, Figure 8C), GSE45887 cohort (AUC = 0.781, 95% CI = 0.517–1.000, Figure 8D), GSE145467 cohort (AUC = 0.850, 95% CI = 0.700–1.000, Figure 8E), and local cohort (AUC = 0.623, 95% CI = 0.482–0.765, Figure 8F). Figure 8G–8K displayed the confusion matrices of these cohorts. The performance of the ANN model in the local cohort was not satisfying (AUC < 0.7), but we held that the result was still acceptable considering the different sample types and gene expression detection methods. As a whole, the ANN model was a promising tool to classify the NOA samples on the background of the high heterogeneity among different cohorts.

Here, we measure the predictive performance of these models mainly from the aspect of AUC. However, the other assessment indexes, containing accuracy, precision, recall, F-measure, sensitivity, specificity, positive predictive value, and negative predictive value, were also provided for reference, as listed in Table 4.

The functions of the DDX20- and NCBP2-asociated genes

The Top 20 genes showing the highest connection with DDX20 and NCBP2 are displayed in Figure 9A and 9B, respectively, along with their interaction patterns. The DDX20-associated genes mainly involved cellular transcription, RNA modification, RNA splicing, RNA localization, and RNA stability maintenance (Figure 9C). The NCBP2-associated genes were mainly enriched in mRNA and miRNA processing, RNA stability regulation, and DNA repair (Figure 9D). These data revealed clues further to elucidate the biological functions of DDX20 and NCBP2.

DISCUSSION

The rapid development of gene sequencing technologies and the tremendous advancement of computational biology and machine learning algorithms help to improve our understanding of the genetic biomarkers, associated pathogenesis, and latent therapeutic targets in multiple reproductive diseases covering a broad spectrum of prostate cancer [23], spontaneous miscarriage [24], testicular cancer [25], and NOA [26]. Investigating novel biomarkers from a particular aspect, such as transcriptional factor [27] and macrophage polarization [28], has become a popular and effective maneuver. However, the studies about the expression profiles and predictive values of RBPs are rarely seen in NOA for the moment despite the nonnegligible effects of RBPs on spermatogenesis, as discussed above. Hence, seeking more genetic biomarkers based on RBP-related genes is urgently demanded on the background of our poor knowledge of the mechanisms of NOA.

Herein, we utilized the 58 testicular samples inclusive of 11 control and 47 NOA cases from the GEO as the training cohort. 51 of 1542 RBPs reported by previous studies were differentially expressed between the control and NOA subjects. DDX20 and NCBP2 were ultimately determined as

the significant features through the PPI network analysis, LASSO regression, SVM-RFE, and Boruta. Subsequently, we collected the clinical samples from 27 control and 17 NOA patients in the local hospital to verify the expression divergence of DDX20 and NCBP2. Intriguingly, we found that DDX20 and NCBP2 were significantly down-regulated in the seminal plasma samples, in addition to the testicular

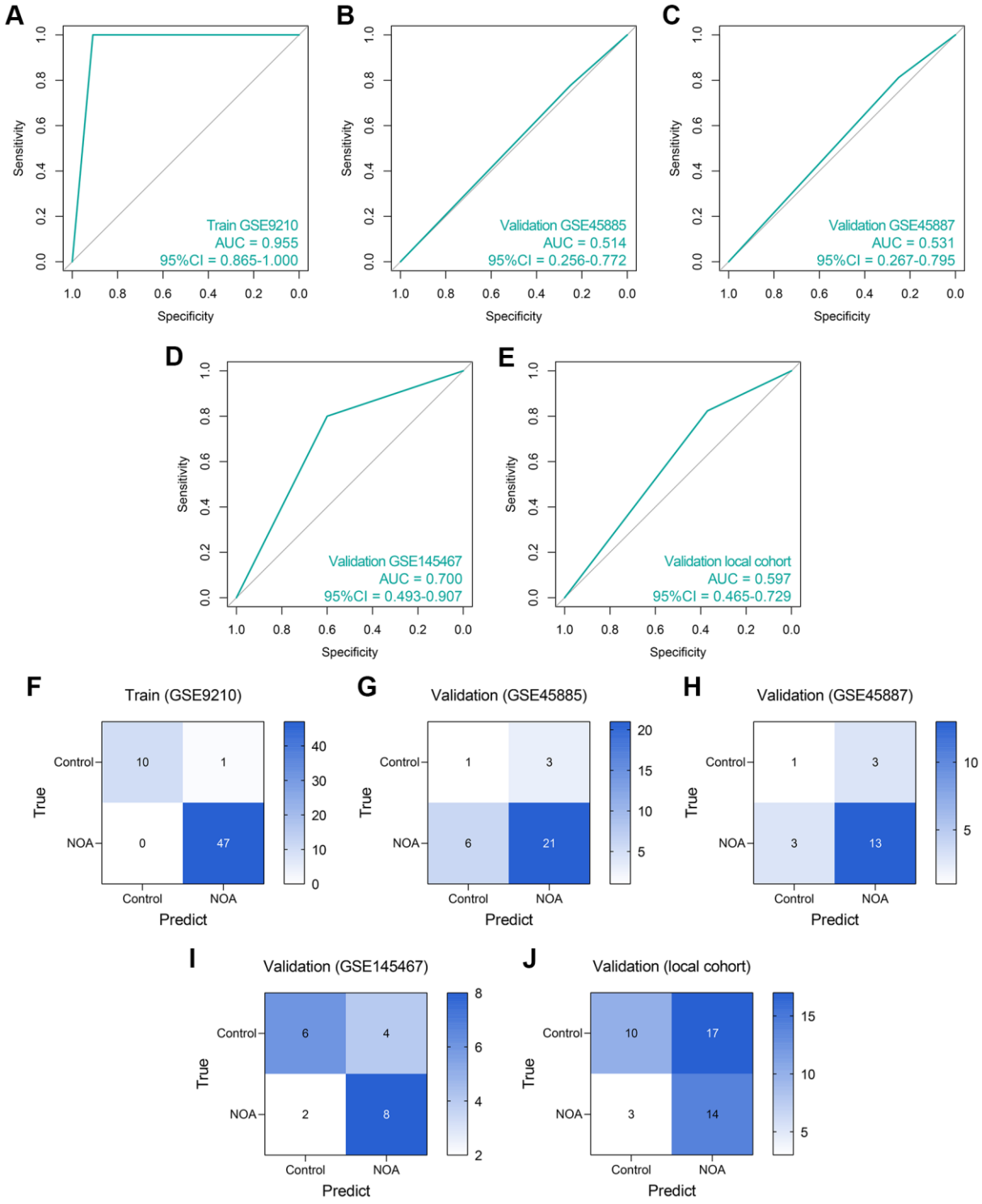

Figure 6. The predictive performance of an LR diagnosis model in each cohort. (**A**–**E**) The ROC analyses of the LR model in the training cohort (**A**), GSE45885 cohort (**B**), GSE45887 cohort (**C**), GSE145467 cohort (**D**), and the local cohort (**E**). (**F**–**J**) The confusion matrices of the LR model in the training cohort (**F**), GSE45885 cohort (**G**), GSE45887 cohort (**H**), GSE145467 cohort (**I**), and the local cohort (**J**). Abbreviation: LR: logistic regression.

samples, of NOA subjects, suggesting that DDX20 and NCBP2 could serve as potential non-invasive diagnostic biomarkers. The scRNA-seq analysis of 432 testicular cell samples isolated from NOA patients indicated that DDX20 and NCBP2 were both associated with the spermatogenetic process, re-confirming the pivotal roles DDX20 and NCBP2 played in NOA.

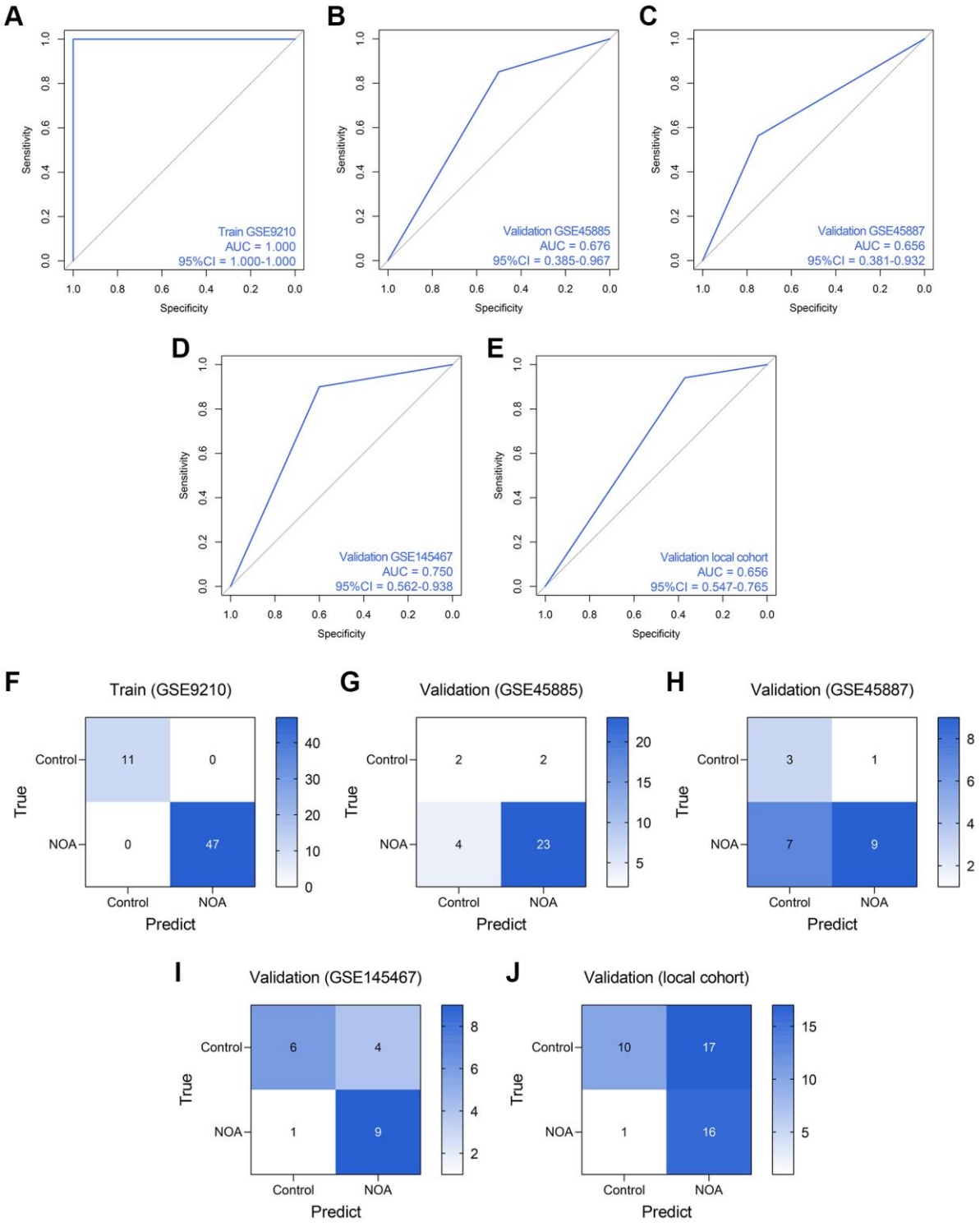

Figure 7. The predictive performance of an RF diagnosis model in each cohort. (**A**–**E**) The ROC analyses of the RF model in the training cohort (**A**), GSE45885 cohort (**B**), GSE45887 cohort (**C**), GSE145467 cohort (**D**), and the local cohort (**E**). (**F**–**J**) The confusion matrices of the RF model in the training cohort (**F**), GSE45885 cohort (**G**), GSE45887 cohort (**H**), GSE145467 cohort (**I**), and the local cohort (**J**). Abbreviation: RF: random forest.

DDX20 encoded a DEAD box protein and was first reported as an RBP interacting with miR-140-3p by Takata and his colleagues [29]. DEAD box proteins, represented by VASA, a widely-accepted germ-line specific marker, were considered critical regulatory factors in spermatogenesis via modulating multiple RNA metabolism processes [30]. The other DEAD box proteins involved in spermatogenesis include

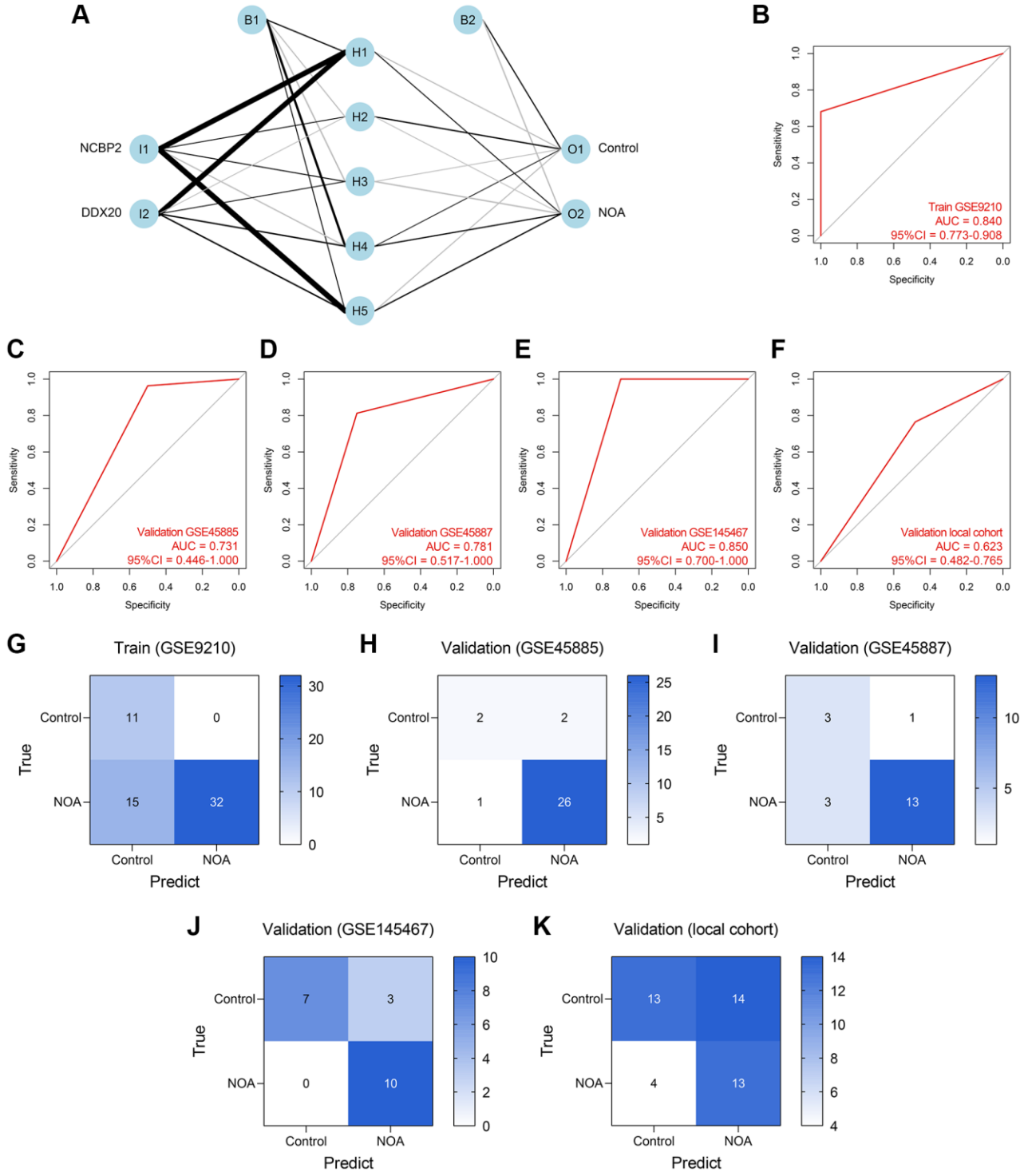

Figure 8. Establishment and validation of an ANN diagnosis model. (**A**) An ANN model containing one input layer, one hidden layer, and one output layer was constructed to diagnose NOA. (**B–F**) The ROC analyses of the ANN model in the training cohort (**B**), GSE45885 cohort (**C**), GSE45887 cohort (**D**), GSE145467 cohort (**E**), and the local cohort (**F**). (**G–K**) The confusion matrices of the RF model in the training cohort (**G**), GSE45885 cohort (**H**), GSE45887 cohort (**I**), GSE145467 cohort (**J**), and the local cohort (**K**). Abbreviation: ANN: artificial neural network.

Table 4. The predictive performance of the established models in each cohort.

Cohort	Accuracy	Precision	Recall	F-measure	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Logistic regression model								
GSE9210	0.983	0.979	1.000	0.989	1.000	0.909	0.979	1.000
GSE45885	0.710	0.875	0.778	0.824	0.778	0.250	0.875	0.143
GSE45887	0.700	0.813	0.813	0.813	0.813	0.250	0.813	0.250
GSE145467	0.700	0.667	0.800	0.727	0.800	0.600	0.667	0.750
Local Cohort	0.545	0.452	0.824	0.583	0.824	0.370	0.452	0.769
Random forest model								
GSE9210	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GSE45885	0.806	0.920	0.852	0.885	0.852	0.500	0.920	0.333
GSE45887	0.600	0.900	0.563	0.692	0.563	0.750	0.900	0.300
GSE145467	0.750	0.692	0.900	0.783	0.900	0.600	0.692	0.857
Local Cohort	0.591	0.485	0.941	0.640	0.941	0.370	0.485	0.909
Artificial neural network model								
GSE9210	0.741	1.000	0.681	0.810	0.681	1.000	1.000	0.423
GSE45885	0.903	0.929	0.963	0.945	0.963	0.500	0.929	0.667
GSE45887	0.800	0.929	0.813	0.867	0.813	0.750	0.929	0.500
GSE145467	0.850	0.769	1.000	0.870	1.000	0.700	0.769	1.000
Local Cohort	0.591	0.481	0.765	0.591	0.765	0.481	0.481	0.765

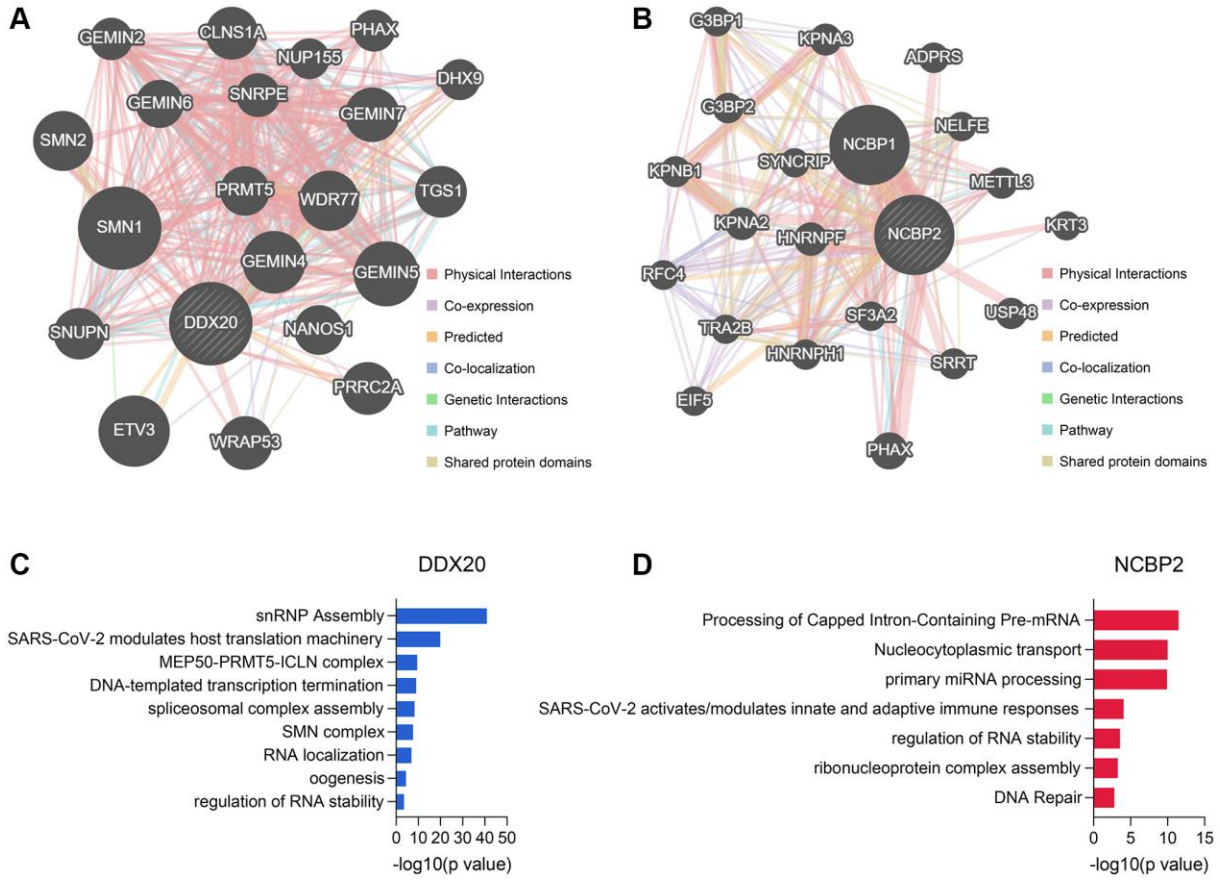

Figure 9. The genes associated with DDX20 and NCBP2 and their functional enrichment. (**A**, **B**) The Top 20 genes showing the closest connection with DDX20 (**A**) and NCBP2 (**B**). (**C**, **D**) The functional annotation of the DDX20- (**C**) and NCBP2-associated (**D**) genes, which was obtained from the Metascape database.

DDX3 [31], DDX25 [32], DDX23 [33], and MEL-46 [34]. Here, we first found that the disturbance of DDX20 was correlated with the spermatogenetic process, and its expressions in the seminal plasma and testes could act as a diagnostic biomarker in NOA, broadening our knowledge of the DEAD box protein family in spermatogenesis. The protein encoded by NCBP2 has an RNP domain commonly found in RBPs and was regarded as a regulator in DNA damage and repair, cell cycle, and cellular apoptosis [35]. However, the association between NCBP2 and spermatogenesis remains unclear. In all, we first found that DDX20 and NCBP2 could serve as biomarkers in NOA, shedding novel insights into the pathogenesis from an angle of RBP.

The reduced cost of gene sequencing renders the genetic diagnosis of NOA to attract increasing attention, and many great efforts have been paid. For example, Kherraf et al. employed whole-exome sequencing to construct a 7-gene panel to improve the classification of NOA, helping the patients receive a clearer diagnosis [36]. Given the satisfying performance of machine learning algorithms, especially ANN, in various diseases [37, 38], we then used LR, RF, and ANN to construct the diagnosis models based on the mRNA expression levels of DDX20 and NCBP2 and validated and compared their predictive ability in the training cohort, 3 external public validation datasets, and the local cohort. Similar to those previous works, the ANN model was observed to exhibit the highest predictive ability on average. It is worth mentioning that to the best of our knowledge, no ANN model based on genetic biomarkers has been constructed in NOA up to now. Our study reveals that ANN modelling has great potency in NOA diagnosis and deserves more attention.

The flaws of the present study should also be acknowledged. First, only the protein levels of DDX20 were detected in the clinical samples, and the detection of the protein levels of NCBP2 was lacking due to the limited financial support. Second, although the models established in this study have been verified in 4 public datasets and clinical samples, a prospective, multi-center, and large-scale clinical trial would be more beneficial to clarify the usefulness of the models. Third, our study analyzed the expression profiles, diagnosis values, and spermatogenetic association of DDX20 and NCBP2 in NOA, but their concrete biological functions in spermatogenesis remain unclear. A deeper experimental exploration is required to better elucidate the associated mechanisms in the near future.

Collectively, an ANN diagnosis model to NOA based on RBP DDX20 and NCBP2 was presented, which was externally validated in multiple public datasets and clinical samples, providing the possible cut-in points to clarify the pathogenesis and a promising tool in clinical practice.

AUTHOR CONTRIBUTIONS

RRZ and YZ designed the whole study, retrieved the public datasets, and collected the clinical samples. RRZ and FP provided the financial support to conduct the experiments. FP, BM, and JWZ developed the algorithms and performed the experiments. FP and HYL wrote the original draft. RRZ and YZ helped to edit the original manuscript. All authors reviewed the manuscript and approved the submission version.

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interests.

ETHICAL STATEMENT AND CONSENT

The protocol of this study has been reviewed and approved by the Ethics Committee of Bao'an Central Hospital of Shenzhen based on the Declaration of Helsinki. The informed consent of each subject was obtained before the testicular biopsy.

FUNDING

The work received no funding support from any agency.

REFERENCES

1. Willott GM. Frequency of azoospermia. Forensic Sci Int. 1982; 20:9–10. https://doi.org/10.1016/0379-0738(82)90099-8 PMID:7095683

2. Kasak L, Laan M. Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives. Hum Genet. 2021; 140:135–54. https://doi.org/10.1007/s00439-020-02112-y PMID:31955275

3. Tournaye H, Krausz C, Oates RD. Novel concepts in the aetiology of male reproductive impairment. Lancet Diabetes Endocrinol. 2017; 5:544–53. https://doi.org/10.1016/S2213-8587(16)30040-7 PMID:27395771

4. Wu X, Lin D, Sun F, Cheng CY. Male Infertility in Humans: An Update on Non-obstructive Azoospermia (NOA) and Obstructive Azoospermia (OA). Adv Exp Med Biol. 2021; 1288:161–73. https://doi.org/10.1007/978-3-030-77779-1_8 PMID:34453736

5. Amirian M, Azizi H, Hashemi Karoii D, Skutella T. VASA protein and gene expression analysis of human non-obstructive azoospermia and normal by immuno-histochemistry, immunocytochemistry, and bio-informatics analysis. Sci Rep. 2022; 12:17259. https://doi.org/10.1038/s41598-022-22137-9 PMID:36241908

6. Naeimi N, Mohseni Kouchesfehani H, Heidari Z, Mahmoudzadeh-Sagheb H, Movahed S. CHD5 gene (rs9434741) might be a genetic risk factor for infertility in non-obstructive azoospermia and severe oligozoospermia. Andrologia. 2022; 54:e14590. https://doi.org/10.1111/and.14590 PMID:36102082

7. Wang Y, Liu L, Tan C, Meng G, Meng L, Nie H, Du J, Lu GX, Lin G, He WB, Tan YQ. Novel MEIOB variants cause primary ovarian insufficiency and non-obstructive azoospermia. Front Genet. 2022; 13:936264. https://doi.org/10.3389/fgene.2022.936264 PMID:35991565

8. Jarvi K, Lo K, Grober E, Mak V, Fischer A, Grantmyre J, Zini A, Chan P, Patry G, Chow V, Domes T. The workup and management of azoospermic males. Can Urol Assoc J. 2015; 9:229–35. https://doi.org/10.5489/cuaj.3209 PMID:26316904

9. Jungwirth A, Giwercman A, Tournaye H, Diemer T, Kopa Z, Dohle G, Krausz C, and European Association of Urology Working Group on Male Infertility. European Association of Urology guidelines on Male Infertility: the 2012 update. Eur Urol. 2012; 62:324–32. https://doi.org/10.1016/j.eururo.2012.04.048 PMID:22591628

10. Practice Committee of the American Society for Reproductive Medicine in collaboration with the Society for Male Reproduction and Urology. Evaluation of the azoospermic male: a committee opinion. Fertil Steril. 2018; 109:777–82. https://doi.org/10.1016/j.fertnstert.2018.01.043 PMID:29778371

11. Peart NJ, Johnson TA, Lee S, Sears MJ, Yang F, Quesnel-Vallières M, Feng H, Recinos Y, Barash Y, Zhang C, Hermann BP, Wang PJ, Geyer CB, Carstens RP. The germ cell-specific RNA binding protein RBM46 is essential for spermatogonial differentiation in mice. PLoS Genet. 2022; 18:e1010416. https://doi.org/10.1371/journal.pgen.1010416 PMID:36129965

12. Su Y, Guo X, Zang M, Xie Z, Zhao T, Xu EY. RNA binding protein BOULE forms aggregates in mammalian testis. J Biomed Res. 2022; 36:255–68. https://doi.org/10.7555/JBR.36.20220072 PMID:35965435

13. Zheng M, Chen X, Cui Y, Li W, Dai H, Yue Q, Zhang H, Zheng Y, Guo X, Zhu H. TULP2, a New RNA-Binding Protein, Is Required for Mouse Spermatid Differentiation and Male Fertility. Front Cell Dev Biol. 2021; 9:623738. https://doi.org/10.3389/fcell.2021.623738 PMID:33763418

14. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014; 15:829–45. https://doi.org/10.1038/nrg3813 PMID:25365966

15. Okada H, Tajima A, Shichiri K, Tanaka A, Tanaka K, Inoue I. Genome-wide expression of azoospermia testes demonstrates a specific profile and implicates ART3 in genetic susceptibility. PLoS Genet. 2008; 4:e26. https://doi.org/10.1371/journal.pgen.0040026 PMID:18266473

16. Malcher A, Rozwadowska N, Stokowy T, Kolanowski T, Jedrzejczak P, Zietkowiak W, Kurpisz M. Potential biomarkers of nonobstructive azoospermia identified in microarray gene expression analysis. Fertil Steril. 2013; 100:1686–94.e1-7. https://doi.org/10.1016/j.fertnstert.2013.07.1999 PMID:24012201

17. Malcher A, Rozwadowska N, Stokowy T, Jedrzejczak P, Zietkowiak W, Kurpisz M. The gene expression analysis of paracrine/autocrine factors in patients with spermatogenetic failure compared with normal spermatogenesis. Am J Reprod Immunol. 2013; 70:522–8. https://doi.org/10.1111/aji.12149 PMID:23869807

18. Wang M, Xu Y, Zhang Y, Chen Y, Chang G, An G, Yang X, Zheng C, Zhao J, Liu Z, Wang D, Miao K, Rao S, et al. Deciphering the autophagy regulatory network via single-cell transcriptome analysis reveals a requirement for autophagy homeostasis in spermatogenesis. Theranostics. 2021; 11:5010–27. https://doi.org/10.7150/thno.55645 PMID:33754041

19. Zhou R, Liang J, Chen Q, Tian H, Yang C, Liu C. A 3-Gene Random Forest Model to Diagnose Non-obstructive Azoospermia Based on Transcription Factor-Related Henes. Reprod Sci. 2023; 30:233–46. https://doi.org/10.1007/s43032-022-01008-8 PMID:35715550

20. Sun D, Peng H, Wu Z. Establishment and Analysis of a Combined Diagnostic Model of Alzheimer's Disease With Random Forest and Artificial Neural Network. Front Aging Neurosci. 2022; 14:921906.

https://doi.org/10.3389/fnagi.2022.921906
PMID:35847663

21. Tian Y, Yang J, Lan M, Zou T. Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure. Aging (Albany NY). 2020; 12:26221–35. https://doi.org/10.18632/aging.202405 PMID:33401250

22. Xie NN, Wang FF, Zhou J, Liu C, Qu F. Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network. Biomed Res Int. 2020; 2020:2613091. https://doi.org/10.1155/2020/2613091 PMID:32884937

23. Li Q, Wu B, Daba M, Gao X, Chen B, Song G, Zeng K, Miao J, Yuan X, Liu J, Wang Z, Liu B. Identification of Calcium Channel-Related Gene P2RX2 for Prognosis and Immune Infiltration in Prostate Cancer. Dis Markers. 2022; 2022:8058160. https://doi.org/10.1155/2022/8058160 PMID:36246559

24. Luan CX, Xie WD, Liu D, Li W, Yuan ZW. Candidate Circulating Biomarkers of Spontaneous Miscarriage After IVF-ET Identified via Coupling Machine Learning and Serum Lipidomics Profiling. Reprod Sci. 2022; 29:750–60. https://doi.org/10.1007/s43032-021-00830-w PMID:35075613

25. Fu Y, Sun S, Bi J, Kong C, Shi D. An RNA-binding protein-related risk signature can predict the prognosis and tumor immunity of patients with testicular germ cell tumors. Am J Transl Res. 2022; 14:2825–43. PMID:35702133

26. Zhang Y, Tang Y, Huang J, Liu H, Liu X, Zhou Y, Ma C, Wang Q, Yang J, Sun F, Zhang X. Circulating microRNAs in seminal plasma as predictors of sperm retrieval in microdissection testicular sperm extraction. Ann Transl Med. 2022; 10:392. https://doi.org/10.21037/atm-21-5100 PMID:35530943

27. He H, Yu F, Shen W, Chen K, Zhang L, Lou S, Zhang Q, Chen S, Yuan X, Jia X, Zhou Y. The Novel Key Genes of Non-obstructive Azoospermia Affect Spermatogenesis: Transcriptomic Analysis Based on RNA-Seq and scRNA-Seq Data. Front Genet. 2021; 12:608629. https://doi.org/10.3389/fgene.2021.608629 PMID:33732283

28. Zheng W, Zhang S, Jiang S, Huang Z, Chen X, Guo H, Li M, Zheng S. Evaluation of immune status in testis and macrophage polarization associated with testicular damage in patients with nonobstructive azoospermia. Am J Reprod Immunol. 2021; 86:e13481. https://doi.org/10.1111/aji.13481 PMID:34192390

29. Takata A, Otsuka M, Yoshikawa T, Kishikawa T, Kudo Y, Goto T, Yoshida H, Koike K. A miRNA machinery component DDX20 controls NF-κB via microRNA-140 function. Biochem Biophys Res Commun. 2012; 420:564–9. https://doi.org/10.1016/j.bbrc.2012.03.034 PMID:22445758

30. Chen W, Brown JS, He T, Wu WS, Tu S, Weng Z, Zhang D, Lee HC. GLH/VASA helicases promote germ granule formation to ensure the fidelity of piRNA-mediated transcriptome surveillance. Nat Commun. 2022; 13:5306. https://doi.org/10.1038/s41467-022-32880-2 PMID:36085149

31. Liu WS, Wang A, Yang Y, Chang TC, Landrito E, Yasue H. Molecular characterization of the DDX3Y gene and its homologs in cattle. Cytogenet Genome Res. 2009; 126:318–28. https://doi.org/10.1159/000266168 PMID:20016128

32. Tsai-Morris CH, Sheng Y, Gutti RK, Tang PZ, Dufau ML. Gonadotropin-regulated testicular RNA helicase (GRTH/DDX25): a multifunctional protein essential for spermatogenesis. J Androl. 2010; 31:45–52. https://doi.org/10.2164/jandrol.109.008219 PMID:19875492

33. Konishi T, Uodome N, Sugimoto A. The Caenorhabditis elegans DDX-23, a homolog of yeast splicing factor PRP28, is required for the sperm-oocyte switch and differentiation of various cell types. Dev Dyn. 2008; 237:2367–77. https://doi.org/10.1002/dvdy.21649 PMID:18729217

34. Minasaki R, Puoti A, Streit A. The DEAD-box protein MEL-46 is required in the germ line of the nematode Caenorhabditis elegans. BMC Dev Biol. 2009; 9:35. https://doi.org/10.1186/1471-213X-9-35 PMID:19534797

35. Singh MD, Jensen M, Lasser M, Huber E, Yusuff T, Pizzo L, Lifschutz B, Desai I, Kubina A, Yennawar S, Kim S, Iyer J, Rincon-Limas DE, et al. NCBP2 modulates neurodevelopmental defects of the 3q29 deletion in Drosophila and Xenopus laevis models. PLoS Genet. 2020; 16:e1008590. https://doi.org/10.1371/journal.pgen.1008590 PMID:32053595

36. Kherraf ZE, Cazin C, Bouker A, Fourati Ben Mustapha S, Hennebicq S, Septier A, Coutton C, Raymond L, Nouchy M, Thierry-Mieg N, Zouari R, Arnoult C, Ray PF.

Whole-exome sequencing improves the diagnosis and care of men with non-obstructive azoospermia. Am J Hum Genet. 2022; 109:508–17.
https://doi.org/10.1016/j.ajhg.2022.01.011
PMID:35172124

37. Chen M, Wu GB, Xie ZW, Shi DL, Luo M. A novel diagnostic four-gene signature for hepatocellular carcinoma based on artificial neural network: Development, validation, and drug screening. Front Genet. 2022; 13:942166.
https://doi.org/10.3389/fgene.2022.942166
PMID:36246599

38. Lugtu EJ, Ramos DB, Agpalza AJ, Cabral EA, Carandang RP, Dee JE, Martinez A, Jose JE, Santillan A, Bangaoil R, Albano PM, Tomas RC. Artificial neural network in the discrimination of lung cancer based on infrared spectroscopy. PLoS One. 2022; 17:e0268329.
https://doi.org/10.1371/journal.pone.0268329
PMID:35551276

SUPPLEMENTARY MATERIALS

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Table 1.

Supplementary Table 1. 1542 RBPs collected from previous reports.

Supplementary Table 2. 51 RBPs showing the RNA expression difference between control and NOA samples.

RBPs	logFC	AveExpr	*t*	*P*-value	FDR	B
NCBP2	–1.164687726	–1.01510361	–15.65914545	1.83E-22	1.90E-19	40.84652815
DDX20	–1.981351445	–1.256276019	–15.4569828	3.36E-22	1.90E-19	40.24705778
PSMA6	–1.04580801	–0.593954932	–14.11702322	2.16E-20	7.40E-18	36.14151813
CCDC86	–1.052820552	–1.579661768	–14.05676552	2.62E-20	7.40E-18	35.95141099
TSN	–1.556463958	–1.458954503	–13.4536697	1.84E-19	4.17E-17	34.02233281
GEMIN4	–1.089842819	–1.39742396	–12.61506455	3.01E-18	4.25E-16	31.25990374
CPSF3	–1.053481046	–0.269150663	–12.55089869	3.74E-18	4.70E-16	31.0447018
EIF5A2	–2.465618315	–2.041042187	–12.45715696	5.14E-18	5.81E-16	30.7293308
DZIP1	–1.672745206	–0.779102049	–12.40795237	6.07E-18	6.25E-16	30.56333073
TDRD7	–1.368686437	–1.133101657	–12.09505473	1.78E-17	1.44E-15	29.50028074
RPL39L	–1.521431228	–1.676756631	–11.89177303	3.59E-17	2.14E-15	28.80279562
ZNF473	–1.54377798	–1.342913734	–11.82184714	4.58E-17	2.47E-15	28.56163306
SAMD4A	–1.80946998	–1.834060238	–11.79041026	5.11E-17	2.63E-15	28.45300701
KHDRBS3	–1.368138893	–1.349001538	–11.73163021	6.28E-17	3.09E-15	28.24955889
SRPK2	–2.073289186	–1.444866371	–11.52317684	1.30E-16	5.90E-15	27.52449873
HABP4	–1.087948044	–1.004264542	–11.31664897	2.71E-16	1.06E-14	26.80069981
CARHSP1	–1.246263907	–1.237840966	–11.26739822	3.22E-16	1.21E-14	26.62730444
HINT3	–1.401923087	–1.583747782	–11.13008743	5.26E-16	1.86E-14	26.14228678
YBX2	–1.576988412	–1.725600589	–10.92210224	1.11E-15	3.48E-14	25.40321439
LARP6	1.079538654	0.630573824	10.86291927	1.37E-15	4.08E-14	25.19194891
MRPL42	–1.339480172	–1.005859006	–10.53158899	4.55E-15	1.12E-13	24.00151447
LSM14B	–2.290982508	–2.204647215	–10.15495172	1.81E-14	3.58E-13	22.63298213
EXOSC9	–1.288382785	–0.439625159	–10.06058765	2.56E-14	4.74E-13	22.28765618
FXR1	–1.018466865	–0.390312658	–10.02650068	2.90E-14	5.29E-13	22.16268071
CALR3	–2.133446923	–1.657728028	–9.922753033	4.26E-14	7.30E-13	21.78155103
TRIM56	1.035837998	0.947452277	9.08252601	9.94E-13	1.14E-11	18.65637084
RANBP17	–1.038119382	–0.834252701	–9.034483444	1.19E-12	1.34E-11	18.4758085
RNF17	–1.096010711	–0.419352204	–8.950511996	1.64E-12	1.72E-11	18.15978426
NUDT21	–1.320680742	–1.348681463	–8.947807388	1.66E-12	1.72E-11	18.1495967
FAM46A	1.105819374	1.036928608	8.783675358	3.09E-12	2.98E-11	17.5303634
MRPS15	–1.154592926	–1.112376746	–8.571113319	6.96E-12	6.05E-11	16.72572174
EZH2	–1.625135047	–1.422316674	–8.530601646	8.13E-12	6.76E-11	16.57205215
PTBP2	–1.177276056	–0.725024783	–8.290871874	2.04E-11	1.59E-10	15.66089028
YBX1	–1.082648888	–1.272797356	–8.23570308	2.52E-11	1.91E-10	15.45080955

ZFP36L2	1.2309675	0.937835892	8.008135343	6.04E-11	4.17E-10	14.5829598
G3BP2	–1.243105015	–1.079792805	–7.962686977	7.20E-11	4.82E-10	14.40942825
RUVBL2	–1.186127446	–1.587202473	–7.878862924	9.95E-11	6.46E-10	14.08922302
MAEL	–1.868901115	–1.771197607	–7.831033391	1.20E-10	7.65E-10	13.90644116
WDR5	–1.359031222	–1.727862331	–7.777332285	1.47E-10	9.15E-10	13.70116652
THUMPD3	–1.252777472	–0.923589423	–7.544794643	3.62E-10	2.08E-09	12.81185111
AARSD1	–1.008879832	–1.189050641	–7.384985781	6.71E-10	3.63E-09	12.20058137
MYEF2	–1.368124041	–0.701933592	–7.053598831	2.41E-09	1.17E-08	10.93417168
DRG1	–1.158392434	–1.473746051	–6.97794412	3.23E-09	1.52E-08	10.64549489
PSIP1	–1.010477262	–0.394447535	–6.918768578	4.06E-09	1.86E-08	10.41986273
PIWIL1	–1.230109401	–0.92726326	–6.874058726	4.83E-09	2.16E-08	10.24949506
DAZL	–1.973456426	–1.623255357	–6.589229654	1.45E-08	6.10E-08	9.166869481
DDX43	–2.054685325	–1.066758398	–6.086147677	9.91E-08	3.57E-07	7.271458078
DDX4	–2.37842902	–1.352155656	–4.899463425	8.11E-06	2.10E-05	2.960449039
RNASE11	–1.722923321	–1.272160398	–4.208908711	9.06E-05	0.000197766	0.625916397
RDM1	–1.173722882	–1.147959994	–4.196040207	9.46E-05	0.000205521	0.584080555
DDX25	–1.434031937	–1.312842028	–3.823569584	0.000324367	0.000647018	–0.595832854

Supplementary Table 3. The importance of the genes in the PPI network.

RBPs	MCC	DMNC	MNC	Degree	EPC	Bottle Neck	Ec Centricity	Closeness	Radiality	Betweenness	Stress	Clustering Coefficient
PIWIL1	252	0.5854	7	12	15.044	16	0.18919	20.28333	5.67568	295.34444	646	0.28788
DAZL	247	0.5854	7	8	14.183	2	0.18919	17.23333	5.31399	85.76111	176	0.57143
DDX4	128	0.4756	7	7	14.187	3	0.18919	17.15	5.36963	53.49444	152	0.61905
RNF17	241	0.6657	6	7	14.002	2	0.15766	16.38333	5.14706	43.96667	92	0.66667
YBX1	8	0.2842	4	6	12.307	6	0.23649	16.75	5.48092	146.27222	260	0.26667
FXR1	10	0.2378	6	6	13.011	6	0.23649	17.5	5.62003	93.7	234	0.33333
TDRD7	240	0.6657	6	6	13.822	1	0.15766	15.63333	5.06359	1.06667	8	0.93333
MAEL	240	0.6657	6	6	13.78	1	0.15766	15.63333	5.06359	1.06667	8	0.93333
YBX2	10	0.3789	4	6	12.914	8	0.18919	16.86667	5.42528	162.26667	428	0.26667
NCBP2	8	0.309	3	6	10.874	4	0.18919	15.4	5.11924	138.90556	286	0.26667
LSM14B	7	0.309	3	6	12.595	4	0.18919	16.48333	5.34181	152.85556	382	0.2
RUVBL2	5	0	1	5	8.256	8	0.18919	15.78333	5.31399	254.75	366	0
MRPS15	5	0.3078	2	5	7.4	6	0.18919	15.11667	5.14706	262.91667	474	0.1
DDX20	8	0.2593	5	5	12.672	1	0.18919	16.65	5.42528	55.15556	160	0.4
DDX25	120	0.6483	5	5	13.306	1	0.15766	14.83333	4.95231	0	0	1
HABP4	4	0.3078	2	4	11.075	9	0.23649	15.91667	5.4531	208.41111	442	0.16667
WDR5	4	0.3078	2	4	10.185	9	0.18919	16.11667	5.4531	159.86111	260	0.16667
GEMIN4	6	0.2842	4	4	11.468	3	0.18919	15.9	5.34181	33.04444	94	0.5
PSIP1	3	0	1	3	8.829	4	0.18919	14.15	5.09141	56.17778	114	0
G3BP2	3	0.3078	2	3	9.416	1	0.23649	14.66667	5.28617	68.66667	106	0.33333
EIF5A2	3	0.3078	2	3	4.23	2	0.15766	11.55	4.3124	66	116	0.33333
NUDT21	3	0.3078	2	3	7.723	2	0.15766	12.86667	4.70191	25.31667	54	0.33333

CPSF3	4	0.309	3	3	6.957	1	0.15766	11.76667	4.34022	3	8	0.66667
EZH2	2	0.3078	2	2	8.015	1	0.18919	13.81667	5.09141	0	0	1
RPL39L	2	0.3078	2	2	4.006	1	0.15766	11.05	4.28458	0	0	1
PSMA6	2	0	1	2	3.843	2	0.15766	11.41667	4.45151	66	84	0
ZNF473	2	0.3078	2	2	5.922	1	0.15766	10.93333	4.22893	0	0	1
CARHSP1	2	0	1	2	5.772	2	0.15766	11.66667	4.47933	66	148	0
TSN	1	0	1	1	5.426	1	0.15766	11.2	4.50715	0	0	0
RDM1	1	0	1	1	3.667	1	0.15766	10.75	4.39587	0	0	0
MRPL42	1	0	1	1	3.718	1	0.15766	10.38333	4.22893	0	0	0
PTBP2	1	0	1	1	1.357	1	0.05405	1	0.16216	0	0	0
KHDRBS3	1	0	1	1	1.357	1	0.05405	1	0.16216	0	0	0
RANBP17	1	0	1	1	2.34	1	0.13514	8.59286	3.39428	0	0	0
DRG1	1	0	1	1	2.128	1	0.13514	8.65952	3.53339	0	0	0
DZIP1	1	0	1	1	5.838	1	0.15766	11.18333	4.39587	0	0	0
EXOSC9	1	0	1	1	2.569	1	0.13514	8.78571	3.56121	0	0	0

Supplementary Table 4. The genes identified by feature selection algorithms.

Algorithm	Genes
LASSO	NCBP2, DDX20, TSN, SRPK2, CARHSP1
SVM-RFE	NCBP2, DDX20, CCDC86, TSN, CARHSP1, TDRD7
Boruta	NCBP2, DDX20, PSMA6, CCDC86, TSN, GEMIN4, CPSF3, EIF5A2, DZIP1, TDRD7, RPL39L, ZNF473, SAMD4A, KHDRBS3, SRPK2, HABP4, CARHSP1, HINT3, YBX2, LARP6, LSM14B, EXOSC9, CALR3, TRIM56, RANBP17, NUDT21, FAM46A, G3BP2, MYEF2, DDX4