



## OPEN Transcriptomic profiling and machine learning reveal novel RNA signatures for enhanced molecular characterization of Hashimoto's thyroiditis

Zefeng Li<sup>1,2</sup>, Qiuyu Xu<sup>2</sup>, Fengxu Xiao<sup>1</sup>, Yipeng Cui<sup>1</sup>, Jue jiang<sup>1</sup>, Qi Zhou<sup>1</sup>, Jiangwei Yan<sup>3</sup>✉, Yu Sun<sup>4</sup>✉ & Miao Li<sup>1</sup>

While ultrasonography effectively diagnoses Hashimoto's thyroiditis (HT), exploring its transcriptomic landscape could reveal valuable insights into disease mechanisms. This study aimed to identify HT-associated RNA signatures and investigate their potential for enhanced molecular characterization. Samples comprising 31 HT patients and 30 healthy controls underwent RNA sequencing of peripheral blood. Differential expression analysis identified transcriptomic features, which were integrated using multi-omics factor analysis. Pathway enrichment, co-expression, and regulatory network analyses were performed. A novel machine-learning model was developed for HT molecular characterization using stacking techniques. HT patients exhibited increased thyroid volume, elevated tissue hardness, and higher antibody levels despite being in the early subclinical stage. Analysis identified 79 HT-associated transcriptomic features (3 mRNA, 6 miRNA, 64 lncRNA, 6 circRNA). Co-expression (77 nodes, 266 edges) and regulatory (18 nodes, 45 edges) networks revealed significant hub genes and modules associated with HT. Enrichment analysis highlighted dysregulation in immune system, cell adhesion and migration, and RNA/protein regulation pathways. The novel stacking-model achieved 95% accuracy and 97% AUC for HT molecular characterization. This study demonstrates the value of transcriptome analysis in uncovering HT-associated signatures, providing insights into molecular changes and potentially guiding future research on disease mechanisms and therapeutic strategies.

**Keywords** Hashimoto's thyroiditis, Transcriptome analysis, Signatures, Machine learning, Molecular characterization

### Abbreviations

ACC	Accuracy
AI	Artificial intelligence
AUC	Area under the curve
CatBoost	Categorical boosting
circRNA	Circular RNA
CNN	Convolutional neural network
DT	Decision tree
ET	Extra tree
HT	Hashimoto's thyroiditis
ENCI	Experimental non-coding RNA interactome
EPC	Edge Percolated Component

<sup>1</sup>Department of Medical Ultrasound, The Second Affiliated Hospital, Xi'an Jiaotong University, 157 Xiwu Road, Xi'an 710004, China. <sup>2</sup>Key Laboratory of National Health Commission for Forensic Sciences, Xi'an Jiaotong University Health Science Center, 76 Yanta West Road, Xi'an 710061, China. <sup>3</sup>Department of Genetics, School of Medicine & Forensics, Shanxi Medical University, 56 Xinjian South Road, Taiyuan 030001, China. <sup>4</sup>Department of Endocrinology and Metabolism, Qilu Hospital of Shandong University, 107 Wenhua West Road, Ji'nan 250012, China. ✉email: yanjw@sxmu.edu.cn; sunyujn@aliyun.com

FPKM	Fragments per kilobase of exon model per million reads mapped
FT3	Free triiodothyronine
FT4	Free thyroxine
GBDT	Gradient boosting decision tree
GO	Gene ontology
HC	Healthy control
HGNC	HUGO gene nomenclature committee
KEGG	Kyoto encyclopedia of genes and genomes
KNN	K-nearest neighbor
LR	Logistic regression
lncRNA	Long non-coding RNA
ML	Machine learning
miRNA	Micro RNA
MOFA	Multi-omics factor analysis
MCC	Maximal clique centrality
MLP	Multilayer perceptron
ncRNA	Non-coding RNAs
rRNA	Ribosomal RNA
RNA-seq	RNA-sequencing
RF	Random forest
SVM	Support vector machine
SWE	Shear wave elastography
TPE	Tree-structured Parzen Estimator approach
TGAb	Thyroglobulin antibodies
TPOAb	Thyroid peroxidase antibodies
TSH	Thyrotropin
XGBoost	EXtreme gradient boosting

Hashimoto's thyroiditis (HT) is an autoimmune disease characterized by specific thyroid pathology<sup>1</sup>. Its prevalence has increased in recent decades, affecting approximately 0.3–1.5 cases per 1000 individuals<sup>2,3</sup>. Pathological features include lymphocyte infiltration, thyroid enlargement, cell destruction, atrophy, and fibrosis<sup>4</sup>. Thyroglobulin antibodies (TGAb) and thyroid peroxidase antibodies (TPOAb)<sup>5</sup> are serological markers indicating autoimmune attack on the thyroid. HT often presents with thyroid morphological changes and elevated antibody levels, sometimes without obvious symptoms or significant thyroid dysfunction<sup>6</sup>, posing challenges in molecular characterization.

Non-coding RNAs (ncRNAs) are RNA transcripts that don't encode proteins but play vital roles in regulating gene transcription and protein translation. Altered expression patterns and regulatory networks of ncRNAs have shown potential for enhancing our understanding of chronic diseases like cancer, neurodegenerative disorders, and cardiovascular diseases<sup>7–9</sup>. Specific ncRNA expression studies offer promise as molecular signatures<sup>10</sup>. In HT research, Zhao et al. identified 6 miRNAs as potential plasma markers<sup>11</sup>, Peng et al. found that upregulated lncRNA-IFNG-AS1 contributes to HT's pathogenesis<sup>12</sup>, and Xiong et al. highlighted the role of has\_cir\_0089172 in HT pathogenesis by sequestering miR-125-3p<sup>13</sup>. However, identifying expression patterns and regulatory network changes in HT remains challenging, with implications for understanding disease mechanisms and potential therapeutic strategies.

Traditional methods of analyzing differential expression focus on individual RNA types, providing detailed biological information but neglecting interrelationships and joint regulation between different RNA types<sup>14</sup>. Integrative analysis methods, such as multi-omics factor analysis (MOFA), are needed to capture complex relationships by combining multiple RNA sequencing data types<sup>15,16</sup>. However, careful feature selection, appropriate data preprocessing, and efficient computing methods are necessary to enhance MOFA's performance with high-dimensional RNA-seq data. Artificial intelligence models, including machine learning, have shown promise in disease characterization and personalized medicine<sup>17,18</sup>. In HT research, Zhao et al. used a deep learning model based on CNN to analyze ultrasound images<sup>19</sup>, while Peng et al. developed a machine learning model using clinical and laboratory data to characterize HT and identify key factors<sup>20</sup>. Integrating differential expression analysis, MOFA, and machine learning to identify RNA signatures and transcriptional regulatory networks in HT requires further research.

This study aims to analyze whole-transcriptome sequencing data from HT patients. By integrating differential expression analysis and MOFA models, we seek to identify transcriptomic signatures for characterizing HT and potential regulatory mechanisms. Co-expression and regulatory networks will be constructed to reveal changes in gene regulation. Additionally, a novel machine learning stacking model will be developed to assess the potential of these signatures for enhanced molecular characterization of HT. This comprehensive approach aims to provide valuable insights into HT molecular mechanisms and identify potential targets for future research and therapeutic strategies.

## Methods

### Study participants

This study involved two distinct cohorts from different medical centers. Between June 2022 and May 2023, a total of 31 early HT patients and 30 healthy controls (HC) were recruited from Qilu Hospital of Shandong University and the Second Affiliated Hospital of Xi'an Jiaotong University. The diagnosis of HT was based on specific criteria such as thyroid enlargement, characteristic ultrasound images, and elevated thyroid antibodies,

with no clinical or biochemical evidence of thyroid dysfunction. To further identify early HT, a longitudinal assessment was conducted over a period of six months, with evaluations performed every two months, to observe the development of autoantibodies and ultrasound changes. If subjects experience thyroid dysfunction during the assessment period, they are considered to be in a stage other than early HT and excluded from this study. Participants with prior history of thyroid diseases, thyroid cancer, or thyroid surgery were excluded. The healthy control group exhibited normal thyroid function, no ultrasound abnormalities, and tested negative for thyroid autoantibodies. Exclusion criteria included the presence of cardiovascular, autoimmune, infectious, musculoskeletal, or malignant diseases, as well as recent surgery or trauma. Thyroid function tests were conducted using the Roche Cobas 6000 E601 module immunoassay analyzer. Demographic information was collected through a questionnaire. The study was ethically approved by the Medical Ethics Committee of Xi'an Jiaotong University and conducted following the Helsinki Declaration guidelines (NO. 2020844). Written informed consent was obtained from all participants.

### RNA sequencing and data pre-processing

Peripheral blood samples were collected during the confirmed early HT assessment for transcriptome sequencing. Total RNA was extracted from peripheral blood samples of each individual using the PAXgene Blood RNA Kit (BD Biosciences, USA). Ribosomal RNA (rRNA) was depleted, and libraries were prepared using the TruSeq Small RNA Library Preparation Kit. The libraries were sequenced on the NovaSeq 6000 platform, generating 150 bp paired-end reads. Raw data quality control was conducted using FastQC (v0.11.8)<sup>21</sup>. miRNA identification was performed using Bowtie (v1.1.2)<sup>22</sup>, and miRNA expression profiles were generated with miRDeep2. The filtered and quality-controlled sequences, referred to as “clean reads”, were aligned using HISAT2 (v2.0.4)<sup>23</sup>. Transcript assembly and gene expression estimation were performed using StringTie (v1.3.1)<sup>24</sup>. After quality control and alignment, transcripts were classified through a systematic pipeline. For protein-coding genes, transcripts were mapped to the human reference genome (GRCh38) and annotated based on RefSeq database using StringTie. miRNAs were identified through alignment to miRBase (v22) using Bowtie and quantified with miRDeep2. For lncRNA identification, we first excluded known protein-coding transcripts and small RNAs, then selected transcripts longer than 200 nucleotides, and assessed their coding potential using CPC2 and CNCI tools. circRNAs were identified by detecting back-spliced junction reads using find\_circ and CIRI2 algorithms, with at least two unique back-spliced reads required for circRNA annotation. All identified transcripts were further filtered based on expression level (FPKM > 0.1) to ensure reliable quantification. The resulting gene expression matrices for mRNA, miRNA, lncRNA, and circRNA were utilized for subsequent analysis.

### Identification of HT-associated RNA signatures

#### *Differential expression analysis*

Differential expression analysis was conducted on four RNA expression matrices (mRNA, miRNA, lncRNA, and circRNA) using the limma<sup>25</sup> and edgeR<sup>26</sup> packages. The parallel implementation of both limma and edgeR packages for identifying DEGs was strategically chosen to leverage their complementary strengths. Limma excels in handling complex experimental designs through its sophisticated linear modeling framework and empirical Bayes methods, which are particularly effective for controlling false discovery rates in multi-factor analyses. Meanwhile, edgeR specifically addresses the challenges of RNA-seq count data through its negative binomial distribution-based statistical framework, making it especially robust for analyzing genes with low expression levels. This dual-package approach enhances the reliability of our differential expression analysis by combining limma's statistical power in handling experimental complexity with edgeR's specialized capabilities for RNA-seq data characteristics. The intersection of results from both methods provides a more stringent and confident set of differentially expressed genes. Differentially expressed RNAs meeting the significance criteria were identified (P-value < 0.05 & |log<sub>2</sub>FC| > 1). The intersection of differentially expressed RNA lists obtained from both packages was used for further analysis.

#### *Multi-omics factor analysis*

MOFA<sup>16</sup> was employed to integrate the expression matrices of the four RNA types identified in the differential expression analysis. Prior to generating the MOFA model, z-score normalization was applied to eliminate biases caused by library size discrepancies. The normalization was performed using the formula:

$$x_{ij}^* = \frac{x_{ij} - \mu_i}{\sigma_i}$$

Where  $x_{ij}$  represents the expression value of RNA  $j$  in sample  $i$ ,  $\mu_i$  represents the mean expression value of sample  $i$ , and  $\sigma_i$  represents the standard deviation of sample  $i$ . A 15-factor MOFA model was generated using 10,000 iterations in ‘slow’ convergence mode.

Subsequent analysis was performed using the MOFA+tool<sup>15</sup>. The sample factor matrix was extracted to examine correlations between factors and clinical variables. Factors showing significant differences between the HT and control groups were identified. High-contributing weight features meeting the criteria (weight  $\geq$  mean + 2 \* standard deviation & weight  $\leq$  mean - 2 \* standard deviation) under these factors were selected as the characteristic RNA signatures.

### Pathway enrichment analysis

For the mRNAs among the characteristic RNA signatures, we performed enrichment analysis based on Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) using the clusterProfiler package<sup>27</sup>.

Enrichment information for cellular components, biological processes, molecular functions, and KEGG pathways with  $P$ -value  $< 0.05$  was considered statistically significant. This analysis aimed to provide insights into the biological processes and pathways potentially involved in HT pathogenesis.

### Analysis of the biological basis underlying HT-associated RNA signatures

#### *Construction of co-expression networks*

To identify potential regulatory relationships, we conducted a co-expression network analysis for the characteristic RNA signatures, focusing on lncRNA-miRNA, lncRNA-mRNA, miRNA-mRNA, and circRNA-miRNA interactions. Pearson correlation coefficient was used to examine the correlation between RNA pairs. RNA pairs with a statistical significance ( $P$ -value  $< 0.05$ ) and meaningful correlation strength ( $|r| > 0.2$ ) were included to construct the co-expression network, where RNAs represented nodes and significant correlations were depicted as edges. Visualization of the co-expression network was performed using Cytoscape. For node ranking, we employed CytoHubba with five attribute-ranking methods: Betweenness, Closeness, Degree, Edge Percolated Component (EPC), and Maximal Clique Centrality (MCC). The shared top 5 ranking nodes in attributes were considered as hub RNAs in the co-expression network. To identify modules with potential similar expression patterns, we utilized ClusterONE with the following parameters: Minimum size = 3, Minimum density = Auto, Edge weight = unweighted, Node penalty = 2, Haircut threshold = 0, Merging method = Single-pass, Similarity = Match coefficient, Overlap threshold = 0.8, Seeding method = From unused nodes.

#### *Development of regulatory networks based on experimental non-coding RNA interactome*

To construct a novel human Experimental Non-coding RNA Interactome (ENCI), we integrated the latest data from ten databases, including DIANA<sup>28</sup>, lncBook<sup>29</sup>, miRTarBase<sup>30,31</sup>, miRNet<sup>32</sup>, NPinter4<sup>33</sup>, RAIN<sup>34</sup>, RISE<sup>35</sup>, Starbase<sup>36</sup>, Circbank<sup>37</sup>, RNAInter<sup>38</sup>, and data from a study by Morselli et al.<sup>39</sup>. Standardized IDs were used for different RNA types, with HUGO Gene Nomenclature Committee (HGNC)<sup>40</sup> ID for mRNA and lncRNA, ENSEMBL ID for lncRNA not in HGNC, miRBase<sup>41</sup> ID for miRNA, and circBase<sup>42</sup> ID for circRNA. Self-interactions and duplicate records were removed for data quality assurance. Based on the ENCI, regulatory networks were constructed for characteristic RNA signatures, and CytoHubba and clusterONE software were used to identify hub RNAs and modules in these networks.

### Development of a stacking machine learning model for enhanced HT molecular characterization

We extracted and integrated expression matrices corresponding to characteristic RNA signatures as the dataset for machine learning (ML) model construction. The dataset was randomly split into two independent subsets: one for model training (training set: 21 HT, 20 HC) and the other for model evaluation (testing set: 10 HT, 10 HC). Our modeling strategy involved combining 10-fold bagging and two-layer stacking. In the first layer, we trained 10 base models using 9/10 of the training set and made predictions on the remaining 1/10 of the data. The base models included logistic regression (LR)<sup>43</sup>, decision tree (DT)<sup>44</sup>, random forest (RF)<sup>45</sup>, extra tree (ET)<sup>46</sup>, gradient boosting decision tree (GBDT)<sup>47</sup>, XGBoost (eXtreme Gradient Boosting)<sup>48</sup>, support vector machine (SVM)<sup>49</sup>, multilayer perceptron (MLP)<sup>50</sup>, CatBoost<sup>51</sup>, and K-nearest neighbors (KNN)<sup>52</sup>. In the second layer, we concatenated the predictions from the first layer with the original training set as input. The second-layer models included the ten ML models from the first layer, and the final model was a logistic regression model. The selection of logistic regression as the final model was based on several key considerations. First, logistic regression excels in binary classification tasks and is particularly effective when combining predictions from multiple models, making it ideal for our stacking architecture. Second, its linear nature helps prevent overfitting when integrating diverse predictions from the double-layer models. Third, the model provides interpretable probability outputs and clear insights into the contribution of each base model, which is crucial for understanding the relative importance of different RNA signatures in HT characterization. Furthermore, in our preliminary evaluations, logistic regression demonstrated stable performance in handling the transformed feature space created by the double-layer predictions, while maintaining computational efficiency. Hyperparameter optimization and fine-tuning of the models were performed using the Tree-structured Parzen Estimator Approach (TPE)-based Bayesian optimization algorithm. Model performance and reliability were assessed using accuracy (ACC) and area under the curve (AUC) metrics on the testing set.

## Results

### Demographic and clinical features of the study participants

This study included 61 participants (31 HT, 30 HC), with similar demographic variables between the groups. Laboratory examinations revealed significant differences in thyroid morphology, function, hormone, and immunity levels (all  $p$  values  $< 0.05$ ). The HT group showed increased thyroid volume, elevated tissue hardness, and higher levels of TPOAb and TGAb antibodies. Thyroid function indicators (FT3, FT4) showed varying degrees of abnormality, reflecting different stages of HT progression. Table 1 summarizes the demographic and clinical characteristics of the participants.

### Identification and characterization of HT-associated RNA signatures

#### *Distinct RNA profiles in HT by differential expression analyses*

Differential expression analysis using the limma and edgeR packages identified significant mRNAs, miRNAs, lncRNAs, and circRNAs. The intersection analysis of both methods revealed distinct regulatory patterns across different RNA types. Specifically, we identified 1279 differentially expressed mRNAs (52 up-regulated, 1227 down-regulated), 9 miRNAs (3 up-regulated, 6 down-regulated), 652 lncRNAs (35 up-regulated, 617 down-regulated), and 18 circRNAs (3 up-regulated, 15 down-regulated) (Fig. 1). Notably, the predominant down-

Parameters	Controls	HT cases	P-value
Gender (Female)	28(93%)	31(100%)	0.458
Marital status (Married)	26(87%)	25(81%)	0.772
Age (Y)	41.8 ± 13.6	42.1 ± 12.2	0.921
Weight (kg)	59.2 ± 7.6	60.1 ± 9.6	0.705
Thyroid volume (cm <sup>3</sup> )	0.2 ± 0.0	0.3 ± 0.1	<b>3.95E-05</b>
SWE (KPa)	14.3 ± 5.2	43.3 ± 10.0	<b>1.11E-20</b>
FT3 (pmol/l)*	3.7 ± 0.6	4.3 ± 0.8	<b>1.10E-03</b>
FT4 (pmol/l)*	14.2 ± 1.7	16.6 ± 2.6	<b>9.02E-05</b>
TSH (μIU/ml)*	1.7 ± 1.1	5.0 ± 0.8	<b>3.17E-20</b>
TPO-Ab (IU/ml)*	11.9 ± 7.6	568.4 ± 294.9	<b>7.50E-15</b>
TG-Ab (IU/ml)*	31.9 ± 25.4	697.8 ± 270.2	<b>1.32E-19</b>

**Table 1.** The demographic and clinical characteristics of the subjects. Fisher's exact tests were utilized for categorical variables, such as gender and marital status. For continuous variables, t-tests were employed. If the variances are unequal, Welch's t-test was used instead. Significant values are in bold. HT: Hashimoto thyroiditis, SD: Standard Deviation. Data are shown as mean ± SD. \*The normal value ranges of the five biochemical indicators FT3, FT4, TSH, TPOAb, and TGAb in the sample recruitment hospitals are as follows: FT3 (3.0–8.1), FT4 (12–22), TSH (0.3–4.5), TPOAb (0–34), and TGAb (0–115).

regulation pattern, particularly evident in mRNAs (96% down-regulated) and lncRNAs (95% down-regulated), suggests a broad transcriptional suppression in HT pathogenesis. The detailed results can be found in the supplementary materials (Table S1).

#### Key RNA signatures in HT by multi-omics factor analyses

MOFA model was used to identify key RNA signatures for HT molecular characterization. The model explained variances of 3.7%, 39.4%, 8.8%, and 39.9% for mRNA, miRNA, lncRNA, and circRNA data, respectively (Fig. 2A, Table S2). Factor correlation analysis indicated low correlations, reflecting a good model fit (Fig. 2B, Table S2). Variance decomposition analysis revealed that Factor7 explained the main mRNA variance, while miRNA variance was primarily explained by Factor1 and Factor3. lncRNA variance was distributed across multiple factors, with Factor2, Factor4, and Factor12 being prominent. Factors 1 and 2 predominantly explained circRNA variance (Fig. 2C, Table S2). Significant differences were observed in Factor3 ( $p=0.016$ ) and Factor10 ( $p=0.006$ ) between HT and HC groups (Fig. 2D and E, Table S2). RNA signatures meeting specific criteria (weight  $\geq$  mean + 2 \* standard deviation & weight  $\leq$  mean - 2 \* standard deviation) were identified based on feature weight matrices for Factor3 and Factor10, resulting in 79 characteristic RNA signatures including 3 mRNAs, 6 miRNAs, 64 lncRNAs, and 6 circRNAs (Table S2). The expression heatmap of these RNA signatures was presented in Fig. 2F. Enrichment analysis revealed the involvement of these RNA signatures in various biological processes, including peroxisome organization, peroxisomal membrane transport, immune system processes, Golgi apparatus function, and lipid metabolism (Fig. 2G, Table S2). These findings provide insights into the molecular mechanisms potentially involved in HT pathogenesis and progression.

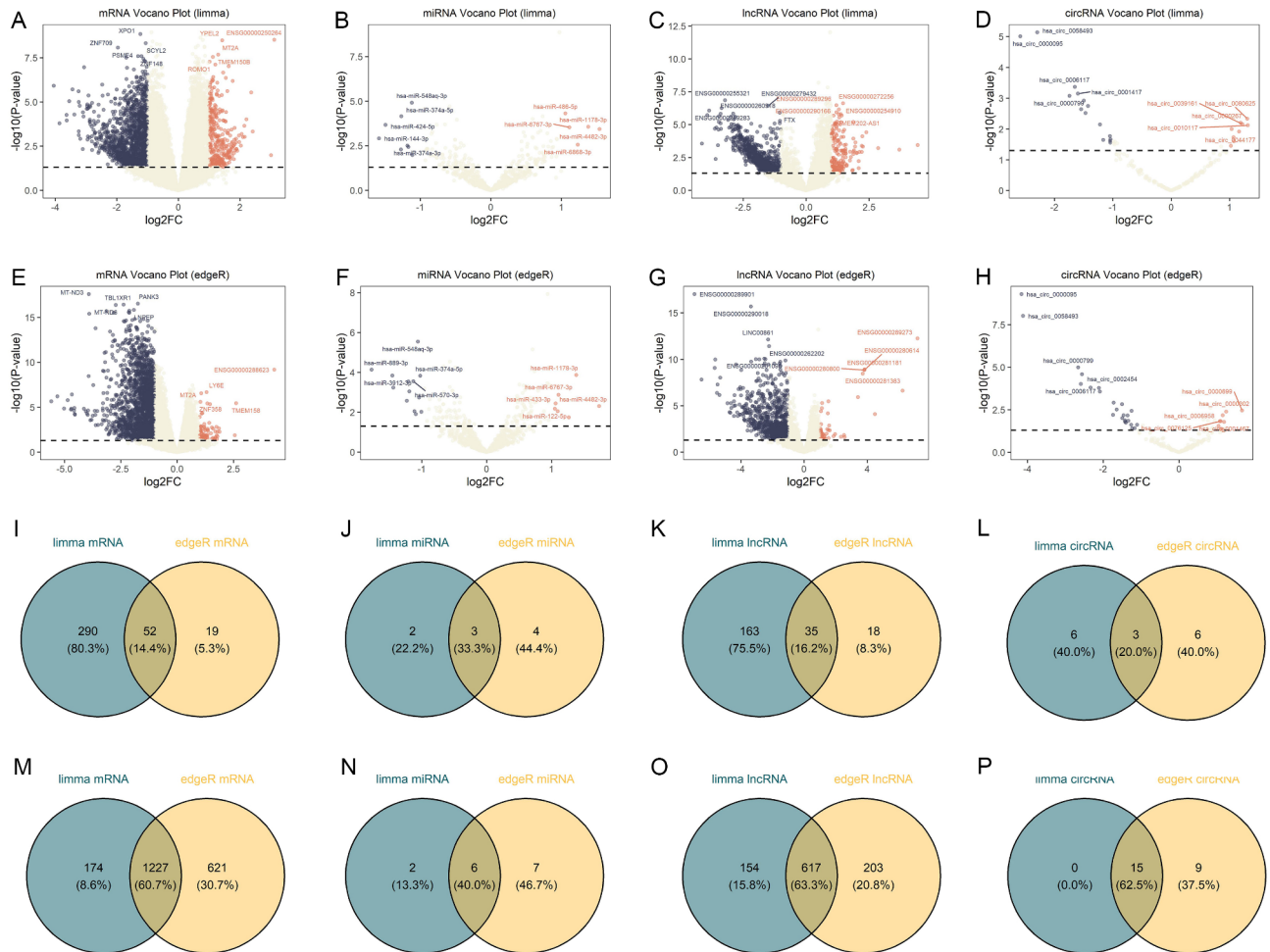
## Molecular interactions and regulatory mechanisms in HT

#### Hub genes and modules in HT by co-expression networks

We categorized RNA signatures based on RNA interactions (lncRNA-miRNA, lncRNA-mRNA, miRNA-mRNA, circRNA-miRNA) and performed correlation tests, resulting in co-expression networks. The networks comprised 77 nodes and 266 edges, including lncRNA-miRNA (65 nodes, 121 edges), lncRNA-mRNA (66 nodes, 122 edges), miRNA-mRNA (7 nodes, 8 edges), and circRNA-miRNA (9 nodes, 15 edges) (Fig. 3A, Table S3). Using CytoHubba, we identified hsa-miR-548aq-3p, PEX13, MARCHF1, and hsa-miR-374a-5p as hub RNAs from the top shared nodes in attributes (Table S3). ClusterONE software extracted two modules with similar expression patterns, containing 5 RNAs each (Fig. 3B,I). Module 1 exhibited significant patterns among its RNA pairs (all  $p$  values < 0.05, Fig. 3C–H). Module 2 showed significant patterns for all pairs except between ENSG00000214955 and hsa-miR-6767-3p ( $p=0.36$ ) (all other  $p$  values < 0.05, Fig. 3J–P).

#### Dysregulation of multiple processes in HT by regulatory networks

By integrating information from 10 databases and the study by Morselli et al., an ENCI with 210,672 nodes and 30,392,876 edges was constructed (Table S4). Analysis of the 79 characteristic RNA signatures using the ENCI identified 18 RNAs with direct interactions, forming a regulatory network with 18 nodes and 45 edges (Fig. 4A). Hub RNAs in the network, ranked by CytoHubba, included hsa-miR-144-3p, hsa-miR-424-5p, hsa-miR-374a-5p, and hsa-miR-548aq-3p (Table S4). The regulatory network was clustered into four modules using ClusterONE software (Fig. 4B–E). The first module comprised 6 RNAs (hsa-miR-548aq-3p, hsa-miR-374a-5p, PEX13, MARCHF1, hsa-miR-144-3p, PTENP1-AS). The second module consisted of 5 RNAs (BACH1-IT2, hsa-miR-1178-3p, PDC-AS1, hsa-miR-374a-5p, PTENP1-A). The third module included 5 RNAs (hsa-miR-1178-3p, NUTM2B-AS1, hsa-miR-424-5p, hsa-miR-6767-3p, ARHGAP15-AS1). The fourth module contained 5 RNAs (BACH1-IT2, hsa-miR-1178-3p, HIF1A-AS3, hsa-miR-374a-5p, PTENP1-AS). GO and KEGG enrichment

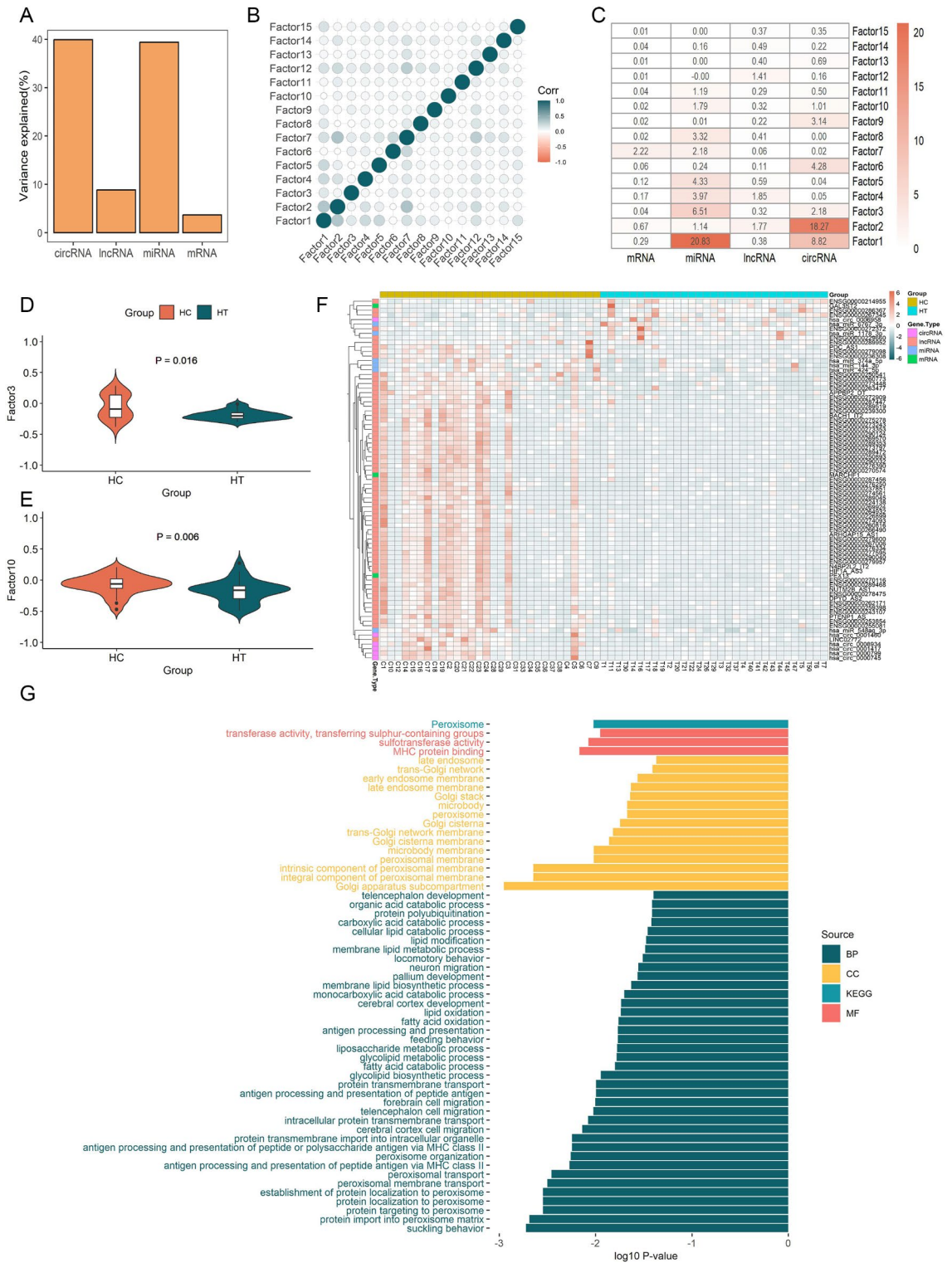


**Fig. 1.** Differential expression analysis of RNA signatures in Hashimoto's thyroiditis. **(A–D)** Volcano plots depicting the results of differential expression analysis based on limma for four RNA-seq datasets. Significantly differentially expressed RNAs were identified: 1743 mRNAs (342 Up, 1401 Down), 13 miRNAs (5 Up, 8 Down), 969 lncRNAs (198 Up, 771 Down), and 24 circRNAs (9 Up, 15 Down). **(E–H)** Volcano plots showing the results of differential expression analysis based on edgeR for the same four RNA-seq datasets. Significantly differentially expressed RNAs were identified: 1919 mRNAs (71 Up, 1848 Down), 20 miRNAs (7 Up, 13 Down), 873 lncRNAs (53 Up, 820 Down), and 33 circRNAs (9 Up, 24 Down). **(I–L)** Venn diagrams illustrating the intersection of upregulated differentially expressed RNAs identified by both limma and edgeR for the four RNA-seq datasets. **(M–P)** Venn diagrams displaying the intersection of downregulated differentially expressed RNAs identified by both limma and edgeR for the four RNA-seq datasets. Intersection analysis of the two methods resulted in a final set of preliminary RNA signature candidates: 1279 mRNAs (52 Up, 1227 Down), 9 miRNAs (3 Up, 6 Down), 652 lncRNAs (35 Up, 617 Down), and 18 circRNAs (3 Up, 15 Down). ENSEMBL IDs are shown only for transcripts without officially assigned gene symbols, primarily novel lncRNAs; all other transcripts are displayed using their HGNC gene symbols.

analysis of the 18 nodes revealed their involvement in key biological processes and pathways, such as immune system regulation, cell adhesion and migration, RNA and protein regulation, lipid metabolism, and vesicle and membrane regulation (Table S4, Fig. 4F). These findings suggest that HT is characterized by dysregulation of multiple biological processes and pathways, potentially contributing to its complex pathogenesis and progression.

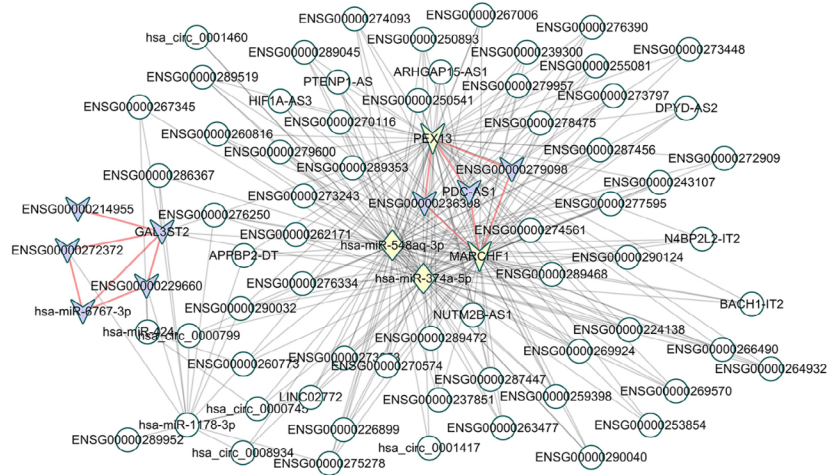
### Enhanced HT molecular characterization via ensemble machine learning

We utilized the expression matrix of 79 RNA signatures to build ML models. Twenty randomly selected samples (10 HT, 10 HC) were used for testing, while the rest were used for training. Ten ML models mentioned in the methods were employed, along with a stacking model for cross-model comparison (hyperparameter settings in Table S5). Figure 5A showed the overview of our newly constructed stacking model. Figure 5B displayed the prediction results on the testing set (showing only models with ACC > 80%). The stacking model showed the best performance, achieving an ACC of 95% and an AUC of 97%. These results demonstrate the superiority of the multi-model fusion approach over individual models in characterizing HT molecular features using a small sample dataset. Our RNA signature selection method combined with the stacking model has potential for enhancing HT molecular characterization.



**Fig. 2.** Multi-omic factor analysis of RNA signatures in Hashimoto's thyroiditis. **(A)** Bar plot demonstrating the total variance explanation of the model for the four RNA-seq datasets. **(B)** Heatmap representing the correlation between factors. **(C)** Heatmap displaying the variance explanation for each factor across the four RNA types. **(D, E)** Violin plots indicating the values of Factor 3 and Factor 10 across Hashimoto's thyroiditis and healthy control samples. **(F)** Expression heatmap visualizing the characteristic RNA signatures identified through multi-omic factor analysis. **(G)** Bar plot presenting the results of GO and KEGG enrichment analysis for the mRNAs included in the characteristic RNA signatures. ENSEMBL IDs are shown only for transcripts without officially assigned gene symbols, primarily novel lncRNAs; all other transcripts are displayed using their HGNC gene symbols.

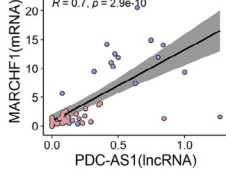
A



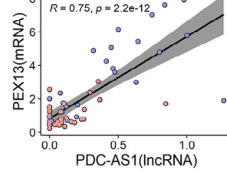
B



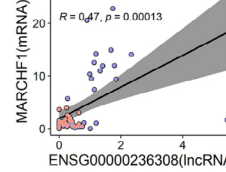
C



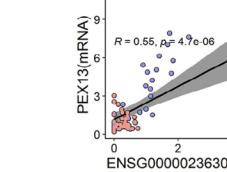
D



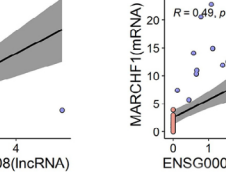
E



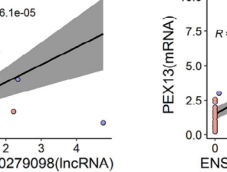
F



G



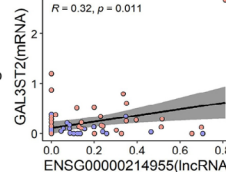
H



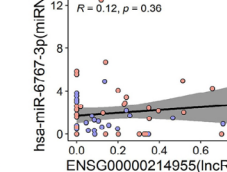
I



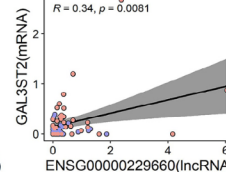
J



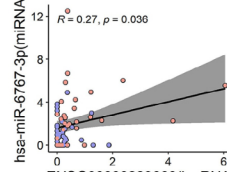
K



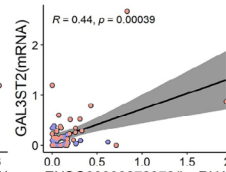
L



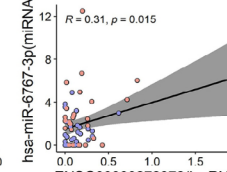
M



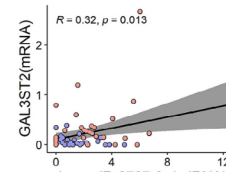
N



O



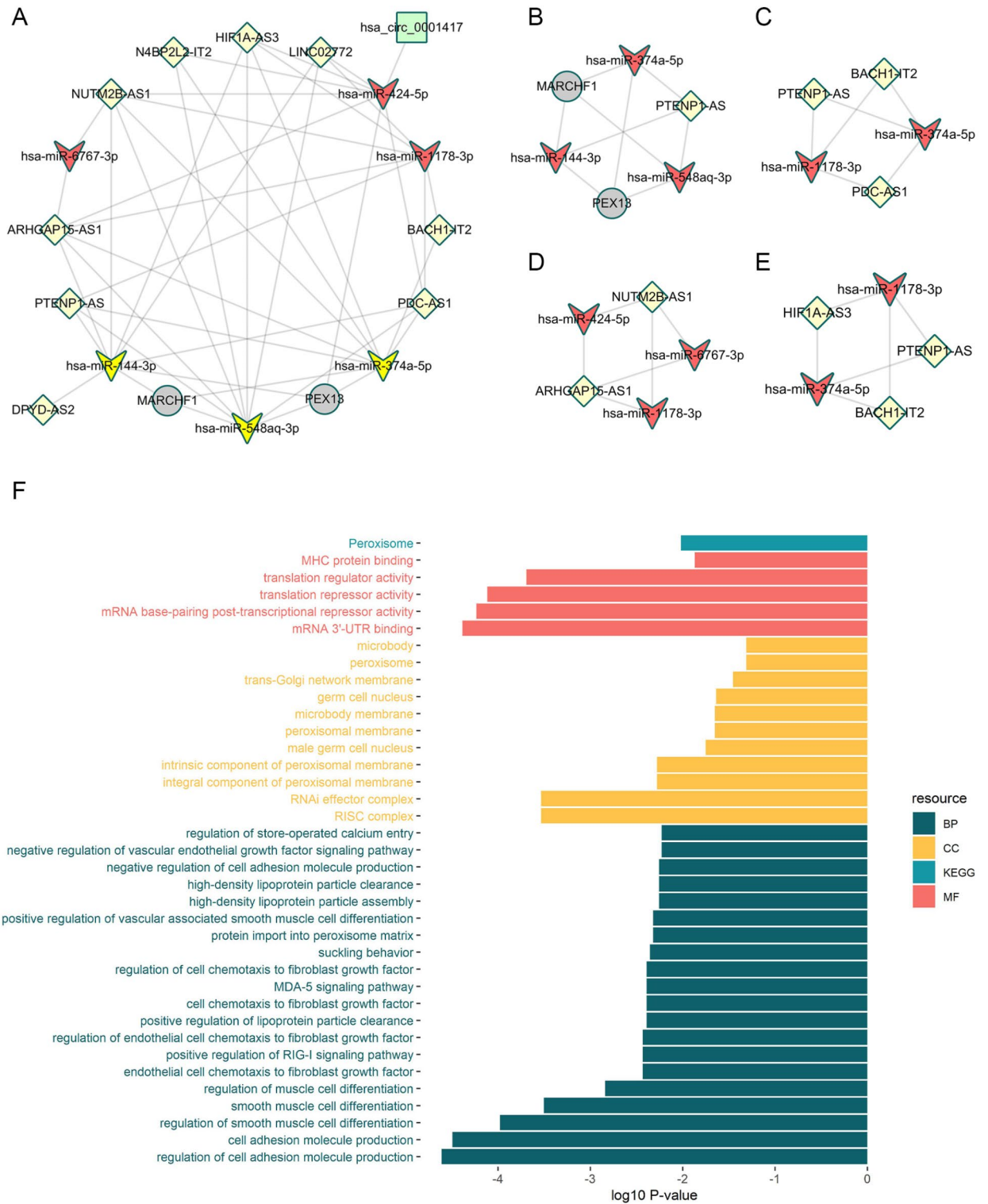
P



Group ● HC ● HT

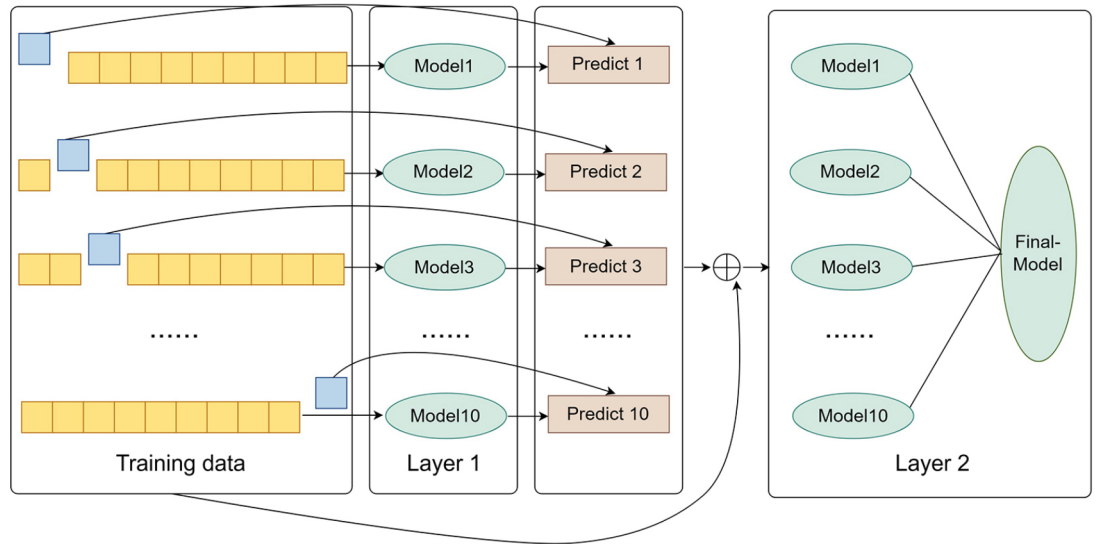
**Fig. 3.** Results of co-expression network analysis in Hashimoto's thyroiditis. **(A)** Visualization of the co-expression network. The yellow nodes represent hub genes, “V”-shaped nodes indicate modules, and lines between nodes represent interactions. **(B)** Visualization of Module 1. This module consisted of 5 RNA signatures (MARCHF1, PDC-AS1, ENSG00000236308, ENSG00000279098, PEX13). **(C–H)** Expression correlation plots for molecular pairs within Module 1. Red points represent the HT group, purple points represent the healthy control group, the black line indicates the fitted line, and the gray interval represents the confidence interval. **(I)** Visualization of Module 2. This module comprised 5 RNA signatures (ENSG00000272372, GAL3ST2, hsa-miR-6767-3p, ENSG00000214955, ENSG00000229660). **(J–P)** Expression correlation plots for molecular pairs within Module 2. Red points represent the HT group, purple points represent the healthy control group, the black line indicates the fitted line, and the gray interval represents the confidence interval. ENSEMBL IDs are shown only for transcripts without officially assigned gene symbols, primarily novel lncRNAs; all other transcripts are displayed using their HGNC gene symbols.



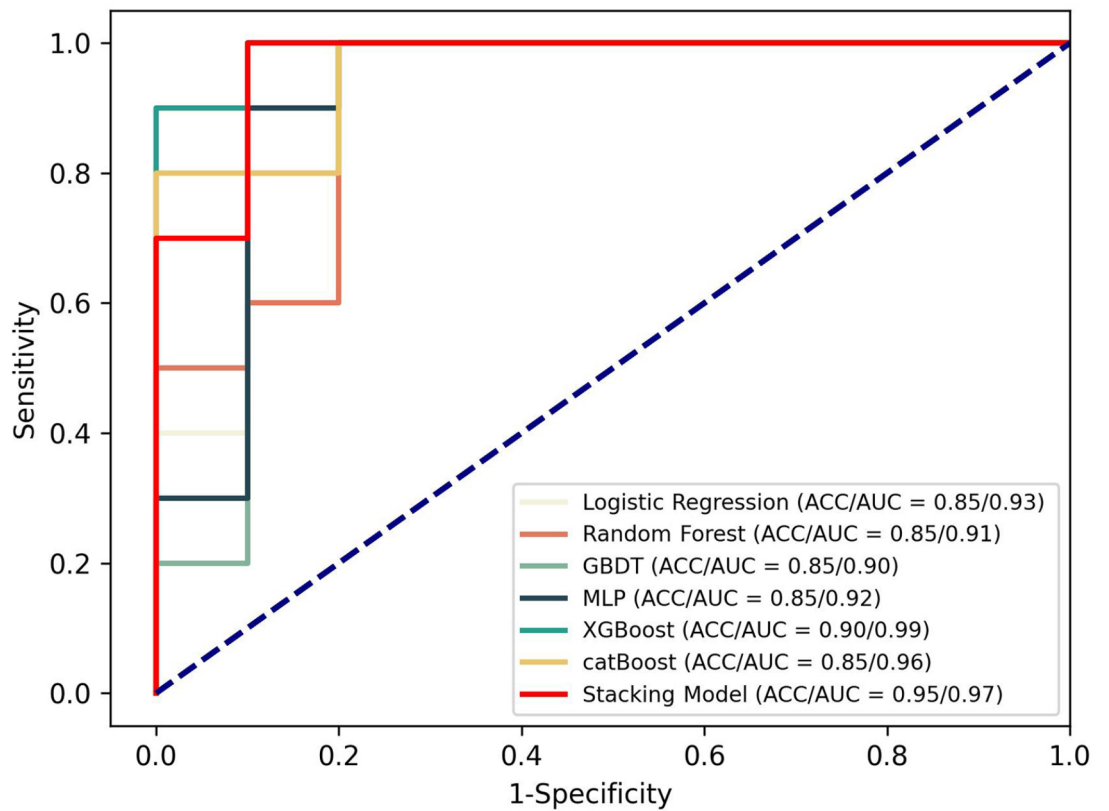


**Fig. 4.** Regulatory network analysis of RNA signatures in Hashimoto's thyroiditis. (A) Visualization of the regulatory network. Gray circular nodes represent mRNAs, red "V"-shaped nodes represent miRNAs, yellow diamond-shaped nodes represent lncRNAs, and green square nodes represent circRNAs. (B–F) Visualization of four modules within the regulatory network, highlighting potential functional RNA signature clusters. (G) Bar plot presenting the results of GO and KEGG enrichment analysis for the regulatory network, providing insights into the biological processes and pathways associated with the identified RNA signatures.

A



B



**Fig. 5.** Stacking model construction and performance evaluation for molecular characterization of Hashimoto’s thyroiditis. **(A)** Overview of the proposed stacking model integrating multiple machine learning algorithms for enhanced molecular characterization of HT. **(B)** Comparison of various machine learning models, including the stacking model, on the independent testing set, demonstrating the performance in distinguishing Hashimoto’s thyroiditis from healthy controls based on the identified RNA signatures.

## Discussion

In this study, we performed a comprehensive transcriptomic analysis of HT patients and healthy individuals, identifying 79 characteristic RNA signatures. We constructed co-expression and regulatory networks based on these signatures and developed a Stacking model for enhanced molecular characterization of HT. Our model demonstrated excellent performance on an independent testing set, highlighting its potential for improving HT molecular profiling. These findings provide valuable insights into HT's molecular mechanisms and may contribute to the development of more precise diagnostic and therapeutic strategies.

The widespread down-regulation observed across different RNA types provides important insights into HT pathogenesis. The down-regulated mRNAs were primarily enriched in pathways related to thyroid hormone synthesis and metabolism, suggesting compromised thyroid function. The dysregulated miRNAs, particularly the up-regulated hsa-miR-144-3p and hsa-miR-374a-5p, may contribute to immune dysregulation through their targeting of immune-related genes. The extensive down-regulation of lncRNAs implies altered transcriptional regulation, potentially affecting thyroid-specific gene expression programs. These regulatory patterns align with the progressive nature of HT and may represent early molecular events in disease development.

Previous transcriptomic studies often focused solely on differential expression analysis, overlooking the complex interactions between multiple RNA types. To address this limitation, we introduced MOFA, which simultaneously identifies factors explaining variation across different omics datasets. However, MOFA's computational demands and sensitivity to data quality posed challenges. As noise or outliers can impact its accuracy, it is very necessary to reduce the dimensionality of RNA-seq data and remove irrelevant variables. To overcome these, we combined differential expression analysis with MOFA for screening HT characteristic RNA signatures and employed a fusion strategy using Bagging and Stacking to characterize HT based on the identified signatures. Bagging reduces overfitting by training diverse learners on subsets<sup>53</sup>, while stacking integrates different learners to reduce variance<sup>54</sup>. This strategy significantly mitigates model fluctuations caused by data changes and enhances performance for the complex task of molecularly characterizing HT.

HT is a complex autoimmune disease with a multifaceted pathogenesis. Integrating different RNA data types provides comprehensive insights into gene regulation, potentially identifying novel factors and pathways involved in HT<sup>55</sup>. Analysis of characteristic RNA signatures and enrichment helps uncover key molecules and processes involved in HT dysregulation. For instance, PEX13 and MARCHF1 impact immune cell activity. PEX13 regulates peroxisome balance, while MARCHF1 influences MHC II expression and immune recognition<sup>56,57</sup>. MiRNAs like hsa-miR-144-3p and hsa-miR-374a-5p control inflammation and autoimmune-associated genes<sup>58</sup>. While these molecules have known roles in immune regulation and inflammation, their specific involvement in HT pathogenesis and identification as hub genes in HT-specific regulatory networks represents a novel finding. Similarly, hsa-miR-1178-3p interacts with BRCA2, potentially influencing thyroid cancer risk<sup>59</sup>. HIF1A-AS3 expression is higher in multiple sclerosis<sup>60</sup>. This combination of known immune regulators and novel HT-associated RNAs suggests both the biological relevance of our findings and their potential to reveal previously unknown aspects of HT pathogenesis. These findings suggest common expression patterns and regulatory changes among autoimmune diseases, with PEX13, MARCHF1, hsa-miR-144-3p, hsa-miR-374a-5p, hsa-miR-1178-3p, and HIF1A-AS3 potentially playing significant roles in HT pathogenesis. Our enrichment analysis supports these findings, indicating the involvement of immune system processes, cell processes, and RNA/protein regulation in HT pathogenesis. Notably, the characteristic RNA signature co-expression network resembles the regulatory network but contains more nodes and edges, suggesting additional molecules and relationships worth exploring further.

While our study provides valuable insights into HT's molecular landscape, there are limitations to consider. The sample size was relatively small, and the study population was homogeneous, which may limit the generalizability of our findings. Although we implemented several strategies to maximize reliability, including rigorous patient selection with longitudinal assessment, 10-fold bagging in our stacking model, and balanced training-testing sets, external validation using larger and more diverse cohorts remains essential. This validation should include independent cohorts from different populations, cross-center studies, and testing across various disease stages and subtypes. Further research is necessary to functionally characterize the identified RNA signatures and explore their potential as therapeutic targets. Cellular and animal models can help elucidate their mechanisms in HT pathogenesis. Integrating other omics data, such as proteomics and metabolomics, will further enhance our understanding of HT's molecular features. Addressing these limitations through additional investigations will advance our understanding of HT and potentially uncover novel therapeutic targets.

## Conclusions

Our study has yielded several novel and significant findings in HT research. First, we identified a unique set of 79 characteristic RNA signatures specific to early-stage HT, including previously unreported associations such as hsa-miR-548aq-3p. Second, we constructed the first comprehensive experimental non-coding RNA interactome (ENCI) for HT, integrating data from ten databases and revealing novel regulatory networks with 18 nodes and 45 edges. Third, we discovered two distinct co-expression modules and four regulatory modules that provide new insights into potential HT pathogenic mechanisms. Finally, our innovative stacking model achieved superior performance (95% accuracy, 97% AUC) in molecular characterization of early HT, demonstrating the potential of machine learning in disease diagnosis. These computational findings establish a foundation for future experimental validation and potential therapeutic development.

In summary, this study performed a comprehensive transcriptomic analysis of peripheral blood from HT patients, identifying 79 characteristic RNA signatures through the integration of differential expression analysis and MOFA. Our co-expression and regulatory network analyses revealed key molecular interactions, including previously unreported RNA relationships and functional modules in HT. The constructed stacking

model achieved promising performance in molecular characterization, suggesting the potential utility of RNA signatures in HT diagnosis. While experimental validation is needed, our findings provide new insights into the complex RNA regulatory networks in HT and establish a foundation for future mechanistic studies and potential therapeutic developments. This computational framework also demonstrates the value of integrative approaches in understanding autoimmune disease pathogenesis.

### Data availability

The datasets generated and/or analysed during the current study are available in the GitHub repository (<https://github.com/zefenglee/HT>). Interested parties can contact Dr. Miao Li for data access.

Received: 8 September 2024; Accepted: 21 November 2024

Published online: 03 January 2025

### References

- Rapoport, B. Pathophysiology of Hashimoto's thyroiditis and hypothyroidism. *Annu. Rev. Med.* **42**, 91–96. <https://doi.org/10.1146/annurev.me.42.020191.000515> (1991).
- Ott, J. et al. The incidence of lymphocytic thyroid infiltration and Hashimoto's thyroiditis increased in patients operated for benign goiter over a 31-year period. *Virchows Arch.* **459**(3), 277–281. <https://doi.org/10.1007/s00428-011-1130-x> (2011).
- Caturegli, P. et al. Hashimoto's thyroiditis: Celebrating the centennial through the lens of the Johns Hopkins hospital surgical pathology records. *Thyroid* **23**(2), 142–150. <https://doi.org/10.1089/thy.2012.0554> (2013).
- Ralli, M. et al. Hashimoto's thyroiditis: An update on pathogenic mechanisms, diagnostic protocols, therapeutic strategies, and potential malignant transformation. *Autoimmun. Rev.* **19**(10), 102649. <https://doi.org/10.1016/j.autrev.2020.102649> (2020).
- Biondi, B., Cappola, A. R. & Cooper, D. S. Subclinical hypothyroidism: A review. *Jama* **322**(2), 153–160. <https://doi.org/10.1001/jama.2019.9052> (2019).
- Parvathaneni, A., Fischman, D., Cheriya, P. Hashimoto's thyroiditis (2012).
- Anastasiadou, E., Jacob, L. S. & Slack, F. J. Non-coding RNA networks in cancer. *Nat. Rev. Cancer* **18**(1), 5–18. <https://doi.org/10.1038/nrc.2017.99> (2018).
- Salta, E. & De Strooper, B. Non-coding RNAs in neurodegeneration. *Nat. Rev. Neurosci.* **18**(10), 627–640. <https://doi.org/10.1038/nrn.2017.90> (2017).
- Jusic, A. et al. Non-coding RNAs in age-related cardiovascular diseases. *Ageing Res Rev* **77**, 101610. <https://doi.org/10.1016/j.arr.2022.101610> (2022).
- Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**(1), 56–68. <https://doi.org/10.1038/nrg2918> (2011).
- Zhao, L. et al. Differential expression levels of plasma microRNA in Hashimoto's disease. *Gene* **642**, 152–158. <https://doi.org/10.1016/j.gene.2017.10.053> (2018).
- Peng, H. et al. The long non-coding RNA IFNG-AS1 promotes T helper type 1 cells response in patients with Hashimoto's thyroiditis. *Sci. Rep.* **5**, 17702. <https://doi.org/10.1038/srep17702> (2015).
- Xiong, S. et al. Circular RNA expression profiling and the potential role of hsa\_circ\_0089172 in Hashimoto's thyroiditis via sponging miR125a-3p. *Mol. Ther. Nucleic Acids* **17**, 38–48. <https://doi.org/10.1016/j.omtn.2019.05.004> (2019).
- Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13. <https://doi.org/10.1186/s13059-016-0881-8> (2016).
- Argelaguet, R. et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**(1), 111. <https://doi.org/10.1186/s13059-020-02015-1> (2020).
- Argelaguet, R. et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**(6), e8124. <https://doi.org/10.15252/msb.20178124> (2018).
- Bera, K. et al. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat. Rev. Clin. Oncol.* **19**(2), 132–146. <https://doi.org/10.1038/s41571-021-00560-7> (2022).
- Loftus, T. J. et al. Artificial intelligence and surgical decision-making. *JAMA Surg.* **155**(2), 148–158. <https://doi.org/10.1001/jamasurg.2019.4917> (2020).
- Zhao, W. et al. Convolutional neural network-based computer-assisted diagnosis of Hashimoto's thyroiditis on ultrasound. *J. Clin. Endocrinol. Metab.* **107**(4), 953–963. <https://doi.org/10.1210/clinem/dgab870> (2022).
- Li, P. et al. Prediction models constructed for Hashimoto's thyroiditis risk based on clinical and laboratory factors. *Front. Endocrinol. (Lausanne)* **13**, 886953. <https://doi.org/10.3389/fendo.2022.886953> (2022).
- Andrews, S. *FastQC a Quality Control Tool for High Throughput Sequence Data* (2014).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**(4), 357–359. <https://doi.org/10.1038/nmeth.1923> (2012).
- Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**(8), 907–915. <https://doi.org/10.1038/s41587-019-0201-4> (2019).
- Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**(3), 290–295. <https://doi.org/10.1038/nbt.3122> (2015).
- Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**(7), e47. <https://doi.org/10.1093/nar/gkv007> (2015).
- Dai, Z. et al. edgeR: A versatile tool for the analysis of shRNA-seq and CRISPR-Cas genetic screens. *F1000Res* **3**, 95. <https://doi.org/10.12688/f1000research.3928.2> (2014).
- Yu, G. et al. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omic* **16**(5), 284–287. <https://doi.org/10.1089/omi.2011.0118> (2012).
- Karagkouni, D. et al. DIANA-LncBase v3: Indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic Acids Res.* **48**(D1), D101–D110. <https://doi.org/10.1093/nar/gkz1036> (2020).
- Ma, L. et al. LncBook: A curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.* **47**(D1), D128–d134. <https://doi.org/10.1093/nar/gky960> (2019).
- Hsu, S. D. et al. miRTarBase: A database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **39**, D163–D169. <https://doi.org/10.1093/nar/gkq1107> (2011).
- Huang, H. Y. et al. miRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* **48**(D1), D148–d154. <https://doi.org/10.1093/nar/gkz896> (2020).
- Chang, L. et al. miRNet 20: Network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Res.* **48**(W1), W244–w251. <https://doi.org/10.1093/nar/gkaa467> (2020).
- Teng, X. et al. NPInter v4.0: An integrated database of ncRNA interactions. *Nucleic Acids Res.* **48**(D1), D160–D165. <https://doi.org/10.1093/nar/gkz969> (2020).

34. Junge, A., Refsgaard, J. C., Garde, C. et al. RAIN: RNA-protein association and interaction networks. *Database Oxford* **2017**. <https://doi.org/10.1093/database/baw167> (2017).
35. Gong, J. et al. RISE: A database of RNA interactome from sequencing experiments. *Nucleic Acids Res.* **46**(D1), D194–D201. <https://doi.org/10.1093/nar/gkx864> (2018).
36. Li, J. H. et al. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42**, D92–D97. <https://doi.org/10.1093/nar/gkt1248> (2014).
37. Liu, M. et al. Circbank: A comprehensive database for circRNA with standard nomenclature. *RNA Biol.* **16**(7), 899–905. <https://doi.org/10.1080/15476286.2019.1600395> (2019).
38. Kang, J. et al. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Res.* **50**(D1), D326–D332. <https://doi.org/10.1093/nar/gkab997> (2022).
39. Morselli Gysi, D. & Barabási, A. L. Non-coding RNAs improve the predictive power of network medicine. *Proc. Natl. Acad. Sci. U. S. A.* **120**(45), e2301342120. <https://doi.org/10.1073/pnas.2301342120> (2023).
40. Bruford, E. A. et al. HUGO gene nomenclature committee (HGNC) recommendations for the designation of gene fusions. *Leukemia* **35**(11), 3040–3043. <https://doi.org/10.1038/s41375-021-01436-6> (2021).
41. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: From microRNA sequences to function. *Nucleic Acids Res.* **47**(D1), D155–D162. <https://doi.org/10.1093/nar/gky1141> (2019).
42. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: A database for circular RNAs. *Rna* **20**(11), 1666–1670. <https://doi.org/10.1261/rna.043687.113> (2014).
43. Berger, A. L., Pietra, V. J. D. & Pietra, S. A. D. A maximum entropy approach to natural language processing. *Comput. Linguist.* **22**(1), 39–71 (1996).
44. Krzywinski, M. & Altman, N. Classification and regression trees. *Nat. Methods* **14**(8), 757–758. <https://doi.org/10.1038/nmeth.4370> (2017).
45. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
46. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1> (2006).
47. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**(4), 367–378. [https://doi.org/10.1016/s0167-9473\(01\)00065-2](https://doi.org/10.1016/s0167-9473(01)00065-2) (2002).
48. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Association for Computing Machinery: San Francisco, California, USA* (2016).
49. Boser, B. A *Training Algorithm for Optimal Margin Classifiers* (ACM Press, 1992).
50. Hinton, G. E. Connectionist learning procedures. *Artif. Intell.* **40**, 185–234 (1989).
51. Prokhorenkova, L. et al. *CatBoost: Unbiased Boosting with Categorical Features* (Curran Associates Inc., 2018).
52. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**(1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964> (1967).
53. Ghoggh, B. & Crowley, M. *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial*. [arXiv:1905.12787](https://arxiv.org/abs/1905.12787) [stat.ML,cs.LG]
54. Breiman, L. Stacked regressions. *Mach. Learn.* **24**(1), 49–64. <https://doi.org/10.1023/A:1018046112532> (1996).
55. Salmena, L. et al. A ceRNA hypothesis: The Rosetta stone of a hidden RNA language?. *Cell* **146**(3), 353–358. <https://doi.org/10.1016/j.cell.2011.07.014> (2011).
56. Demers, N. D. et al. PEX13 prevents pexophagy by regulating ubiquitinated PEX5 and peroxisomal ROS. *Autophagy* **19**(6), 1781–1802. <https://doi.org/10.1080/15548627.2022.2160566> (2023).
57. Di Cara, F., Andreoletti, P., Trompier, D., et al. Peroxisomes in immune response and inflammation. *Int. J. Mol. Sci.* **20**(16). <https://doi.org/10.3390/ijms20163877> (2019).
58. Hu, S. et al. MicroRNA-144-3p may participate in the pathogenesis of preeclampsia by targeting Cox-2. *Mol. Med. Rep.* **19**(6), 4655–4662. <https://doi.org/10.3892/mmr.2019.10150> (2019).
59. Guo, N. et al. BRCA2 3'-UTR polymorphism rs15869 alters susceptibility to papillary thyroid carcinoma via binding hsa-mir-1178-3p. *Pharmgenomics Pers. Med.* **14**, 533–544. <https://doi.org/10.2147/pgpm.S300783> (2021).
60. Rodriguez-Lorenzo, S. et al. Altered secretory and neuroprotective function of the choroid plexus in progressive multiple sclerosis. *Acta Neuropathol. Commun.* **8**(1), 35. <https://doi.org/10.1186/s40478-020-00903-y> (2020).

## Acknowledgements

We would like to express our sincere gratitude to the patients and their families for their understanding and willingness to participate in this study. We are also grateful to the investigators and site staff for their dedicated efforts and contributions. Without their support and commitment, this research would not have been possible.

## Author contributions

M.L., Y.S. and J.Y. designed and supervised the project. Q.X., F.X., Y.C., J.J., and Q.Z. collected samples. M.L. and J.Y. participated in part of the sequencing work. Z.L. conducted data analyses. Q.X. and M.L. organized the clinical data. Z.L. and Q.X. wrote the manuscript with input from co-authors. M.L., Y.S. and J.Y. revised the manuscript. All authors reviewed and approved the manuscript before submission.

## Funding

This work was supported by the Natural Science Foundation of China (82071952, 82370806, 82170813, 82030058) and Shandong Provincial Natural Science Foundation (ZR2022ZD14). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-80728-0>.

**Correspondence** and requests for materials should be addressed to J.Y. or Y.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025