Computational and Structural Biotechnology Journal 27 (2025) 1305–1317

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

TetraRNA, a tetra-class machine learning model for deciphering the coding potential derivation of RNA world

Hanrui Bai [a,b,1], Jie Wang [c,1], Xiaoke Jiang [b,1], Zhen Guo [d], Wenjing Yang [b], Zitian Yang [b], Jing Li [b,*], Changning Liu [a,b,*]

[a] *College of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230026, China*
[b] *CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Yunnan Key Laboratory of Crop Wild Relatives Omics, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming 650223, China*
[c] *Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, Cologne 50829, Germany*
[d] *College of Science and Engineering, Saint Louis University, St. Louis, MO 63103, USA*

ARTICLE INFO

Keywords:
Translated ncRNA
Untranslated mRNA
CncRNA
Tetra-class classification model
TetraRNADB database

ABSTRACT

CncRNAs (coding and noncoding RNAs) are a class of bifunctional RNAs that that has both coding and noncoding biological activity. An increasing number of cncRNAs are being identified, prompting reassessment of our knowledge of RNA. However, most existing RNA classification tools are based on binary classification models which are not effective in distinguishing cncRNAs from mRNAs or long noncoding RNAs (lncRNAs). Our statistical analysis demonstrated that mRNA-derived cncRNAs (untranslated mRNAs, untr-mRNAs) and lncRNA-derived cncRNAs (translated ncRNAs, tr-ncRNAs) do not fall in the same cluster. Therefore, in this study, we devised a novel tetra-class RNA classification model that is systematically optimized for RNA feature extraction. According to our model, all human RNAs can be reclassified into one of four categories — mRNA, untr-mRNA, lncRNA, and tr-ncRNA — representing a novel RNA classification system and allowing the discovery of more potential cncRNAs. Further analysis revealed significant differences among the four types of RNAs in tissue-specific expression, functional annotation, sequence composition, and other factors, providing insights into their divergent evolution trajectories. Moreover, investigation of the small tr-ncRNA peptides demonstrated that their evolution is coordinated with that of the the conserved functional small RNAs associated with them. All analysis results have been integrated into a database — TetraRNADB accessible online (http://tetrarnadb.liu-lab.com/).

1. Introduction

RNAs are traditionally divided into protein-coding RNAs and noncoding RNAs based on their protein-coding potential, and are thought to function either as a protein or as RNA, respectively. However, recent research suggests that the function of RNA may not always be uniform. For example, the ENOD40 gene was originally thought to function solely in the form of RNA [1]. However, subsequent research found that ENOD40 can also encode 12aa/24aa protein, and binds to sucrose synthase to regulate sucrose utilization in root nodules [2]. In *Drosophila*, the Osk gene codes a scaffolding protein involved in building the cytoplasmic structures required for embryonic development [3]. However, there are *Bruno* binding sites (BREs) in the 3′UTR region of Osk mRNA. *Bruno* is a negative regulator of translation that inhibits early oogenesis, and the 3′UTR region of Osk mRNA can bind *Bruno* and isolate it, thereby regulating the oogenesis process [4].

The RNAs with both protein-coding and noncoding roles are referred to as cncRNA (coding and noncoding RNA) [5]. If a molecule is initially identified as a noncoding RNA (ncRNA) but later found to possess coding functionality, such as ENOD40 RNA, it is classified as a translational ncRNA (tr-ncRNA). Conversely, if a molecule is initially characterized as an mRNA but subsequently discovered to function in a noncoding capacity, such as Osk RNA, it is termed an untranslated mRNA (untr-mRNA).

The emergence of cncRNAs is probably not a random occurrence. To date, cncRNAs have been detected in numerous species with selective

[*] Correspondence to: No. 21 Qingsong Road, Panlong District, Kunming, Yunnan 650223, China.
E-mail addresses: lijing3@xtbg.ac.cn (J. Li), liuchangning@xtbg.ac.cn (C. Liu).
[1] Hanrui Bai, Jie Wang and Xiaoke Jiang contributed equally to this work and these three authors should be regarded as Joint First Authors.

https://doi.org/10.1016/j.csbj.2025.03.039
Received 25 November 2024; Received in revised form 20 March 2025; Accepted 24 March 2025
Available online 26 March 2025

conservation and some important biological functions [6]. With advances in mass spectrometry and ribosome analysis techniques, many short peptides originating from cncRNAs have been discovered and have been demonstrated to play important roles in various processes [7,8]. On the other hand, traditional RNA classification tools are mostly based on the idea of dichotomy, without considering the existence of cncRNAs, and to date there have been only preliminary attempts to develop the systematic analysis tools necessary for the identification and classification of bifunctional RNAs. Ulveling et al. proposed a method to identify bifunctional RNAs by analyzing mRNAs whose ORFs are disrupted by alternative splicing events [9], and Liu et al. developed the LncReader tool to distinguish tr-ncRNAs from lncRNAs [10]. Postic et al. modified the existing binary classification model IRSOM to identify cncRNAs as those likely rejecting both coding and noncoding classes [11]. However, none of these methods has systematically examined whether and how the features of cncRNAs differ from those of coding mRNAs or noncoding RNAs. Therefore, the current dichotomous framework of RNA classification must in principle be altered to have multiple classification scope, and be able to demonstrate the distinctive characteristics of different types of RNA precisely and in detail.

As more cncRNAs have been discovered experimentally, research attention has increasingly focused on this issue. The construction of small-protein-related databases (such as SmProt [12] database and MtSSPdb [13]) and cncRNA-related databases (including cncRNADB [14] and Translnc database [15]) can systematically consolidate data resources on cncRNAs, thereby advancing ongoing research into cncRNAs. The data in these databases are mostly derived from existing experimental and mass spectrometry data. Due to the lack of RNA multi-classification tools, many cncRNAs remain unidentified and have not been included in these databases. Furthermore, by distinguishing RNA categories, we hope to gain a deeper understanding of the origin of RNA bifunctionality and even the evolution of other biological macromolecules.

In this study, we compared the existing cncRNA data with data from mRNA and lncRNA using various binary RNA classification tools, and then used Decision Tree (DT) and Random Forest (RF) algorithms to identify the best-performing Tetra-class model. The model was named RF-based TetraRNA, and included fully optimized extraction of RNA features. Using this model, we further divided all human RNAs into four classes: mRNA, untr-mRNA, lncRNA and tr-ncRNA, and obtained the first landscape of human tetra-classification RNAs by effectively separating the cncRNAs. We then analyzed the characteristics and functions of the four RNA classes and performed a corresponding evolutionary analysis, which revealed the significant differences among the four types

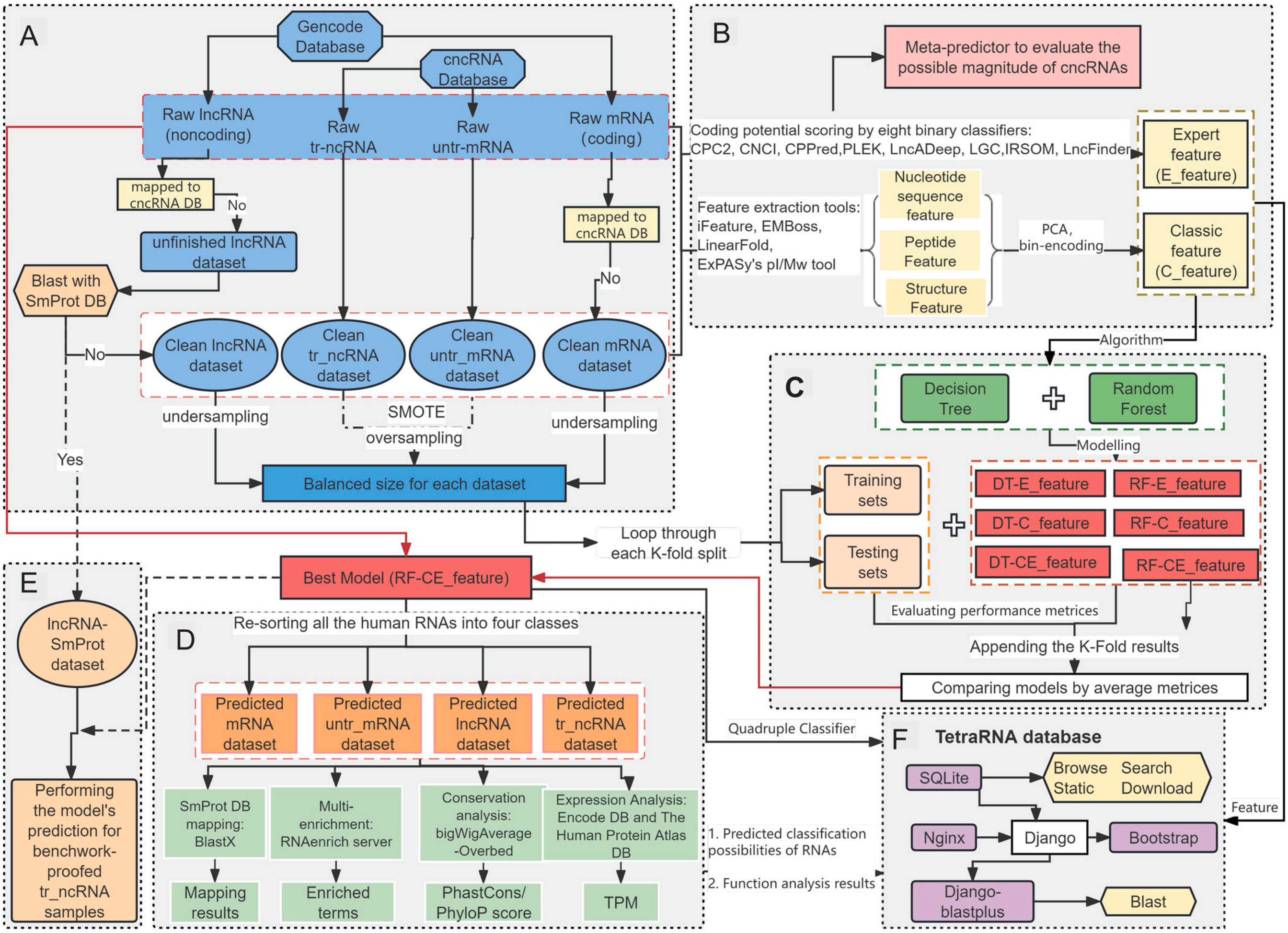
A
Gencode Database
cncRNA Database
Raw lncRNA (noncoding)
Raw tr-ncRNA
Raw untr-mRNA
Raw mRNA (coding)
mapped to cncRNA DB
No
unfinished lncRNA dataset
mapped to cncRNA DB
No
Blast with SmProt DB
No
Yes
Clean lncRNA dataset
Clean tr_ncRNA dataset
Clean untr_mRNA dataset
Clean mRNA dataset
undersampling
SMOTE
oversampling
undersampling
Balanced size for each dataset
B
Meta-predictor to evaluate the possible magnitude of cncRNAs
Coding potential scoring by eight binary classifiers: CPC2, CNCI, CPPred,PLEK, LncADeep, LGC,IRSOM, LncFinder
Expert feature (E_feature)
Feature extraction tools: iFeature, EMBoss, LinearFold, ExPASy's pI/Mw tool
Nucleotide sequence feature
Peptide Feature
Structure Feature
PCA, bin-encoding
Classic feature (C_feature)
Algorithm
C
Decision Tree
Random Forest
Modelling
Loop through each K-fold split
Training sets
Testing sets
DT-E_feature
RF-E_feature
DT-C_feature
RF-C_feature
DT-CE_feature
RF-CE_feature
Evaluating performance metrices
Appending the K-Fold results
Comparing models by average metrices
Best Model (RF-CE_feature)
E
lncRNA-SmProt dataset
Performing the model's prediction for benchwork-proofed tr_ncRNA samples
D
Re-sorting all the human RNAs into four classes
Predicted mRNA dataset
Predicted untr_mRNA dataset
Predicted lncRNA dataset
Predicted tr_ncRNA dataset
SmProt DB mapping: BlastX
Multi-enrichment: RNAenrich server
Conservation analysis: bigWigAverage -Overbed
Expression Analysis: Encode DB and The Human Protein Atlas DB
Mapping results
Enriched terms
PhastCons/ PhyloP score
TPM
Quadruple Classifier
1. Predicted classification possibilities of RNAs
2. Function analysis results
F
TetraRNA database
SQLite
Browse Search Static Download
Feature
Nginx
Django
Bootstrap
Django-blastplus
Blast

Fig. 1. Workflow of this study. A, Constructing the clean datasets for the four types of RNAs. The datasets we used to establish our models were downloaded from GENCODE and the cncRNA database. All datasets would be resampled to ensure that the various datasets were balanced in their qualities. B, Feature extraction. In addition to sequence-intrinsic composition, features (called C features in this study) were extracted from multi-scale secondary structure and EIIP-based physicochemical properties, and the coding potential scores from eight binary classifiers were collected as E features. C, Construction of classification models. The optimal feature combination and machine learning algorithm were used to develop a new method for multi-class RNA identification. D, Re-classification of RNA types and primary exploration of the functions of four types of RNAs. E, The lncRNA-Smprot dataset was used as proofed tr-ncRNAs to evaluate the multi-classifier. F, Database construction.

of RNAs and gave hints into the evolutionary path of the different RNAs. Finally, to facilitate the use of this tool for researchers, we created a database called TetraRNADB (http://tetrarnadb.liu-lab.com/), which supports the download of the TetraRNA model and allows the tetra-classification of human RNAs.

2. Materials and methods

2.1. Construction a tetra-classification machine-learning model for the efficient prediction of categories of RNA

2.1.1. Dataset construction

We collected four human RNA datasets (mRNA, lncRNA, untr-mRNA and tr-ncRNA) from the cncRNA database [14] and GENCODE [16] with annotation file version GRCh38.p14 (Fig. 1A). A total of 85727 mRNAs, 48915 lncRNAs, 5833 tr-ncRNA and 865 untr-mRNAs were obtained as raw RNA datasets. In the raw lncRNA dataset, some lncRNAs (∼26,000 RNAs) mapped to small proteins (and thus potential tr-ncRNA candidates) in the SmProt v2.0 database [12] (BLAST e-value < 1e-5), and were therefore assigned to the lncRNA-SmProt dataset and not included in the model construction. (Fig. 1A and Fig. 1E; details in Supplementary File S1 and Supplementary File S2). Finally, the lncRNAs without those in lncRNA-SmProt dataset and the other three types of RNAs together consisting of the clean datasets to build multi-classification model for RNAs. (Fig. 1A).

2.1.2. Silhouette Coefficient, R $_{(i,j)}$ score and Davies-Bouldin Index

The four types of RNA in the raw data sets formed four known RNA clusters. We use the Scikit-learn.metrics package to calculate the Silhouette Coefficient Score (S_i) of each cluster according to Eq. 1, and the R $_{(i,j)}$ score and Davies-Bouldin Index (DBI) coefficient score between the clusters according to the Eqs. 2 and 3. The level of aggregation of each cluster was assessed using the S_i. When the S_i value approaches 1, the cluster is more aggregated; when it is near −1, the cluster is more dispersed. The proximity of the clusters was gauged using DBI and R $_{(i,j)}$. The distance between two clusters increases with decreasing R $_{(i,j)}$ and increasing DBI.

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{1}$$

In Eq. 1, a_i is the average Euclidean distance between sample i and other samples in the same cluster, and b_i is the min average Euclidean distance between sample i and other samples in all other clusters.

$$R_{(i,j)} = \frac{avg(C_i) + avg(C_j)}{d_{cen}(C_i, C_j)} \tag{2}$$

In Eq. 2, $avg(C_i)$ is the average Euclidean distance from all samples in cluster i to the center of the cluster, and $d_{cen}(C_i, C_j)$ is the Euclidean distance between the centers of clusters i and j.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{3}$$

In Eq. 3, k is the cluster number, and R_{ij} was calculated according to Eq. 2.

2.1.3. Construction of a RNA four-class classification model with a RF and DT framework

We extracted 351 classic features (C-features) of RNA sequences including 126 nucleotide sequence features, 222 peptide features, 3 structural features using the iFeatureOmega program [17] and the "cai" and "chip" functions in EMBOSS [18] following dimensionality reduction. Eight coding potential scores were also calculated from binary RNA classification tools and were named expert features (E-features, Supplementary Table S1). We also combined C-features and E-features as CE-features (Fig. 1B and Supplementary File S3). Finally, six models, including the DT-C model, DT-E model, DT-CE model, RF-C model, RF-E model and RF-CE model were built using these three different kinds of feature and DT/RF machine learning algorithms (Fig. 1C). Details pertaining to feature extraction and model construction are given in Supplementary File S2.

2.1.4. Evaluation of model performance

To evaluate the performance of each model when predicting RNA classification, the following indicators were used: accuracy, precision, recall, and f1-score (Fig. 1C), and ROC-AUC curves were also calculated among the models (Fig. 1C). Due to the particularity of the multi-classification model, we used the Scikit-learn library and adopted two strategies, macro and micro, when calculating the ROC-AUC curves [19]. Based on the results predicted by the best model, we divided all the human RNAs in the raw dataset into four groups: new mRNA, lncRNA, untr-mRNA, and tr-ncRNA, which would be used for further analysis (Fig. 1D). The best model was named the TetraRNA model.

2.2. Comparison of TetraRNA and other RNA classification tools

To compare TetraRNA with basic binary classification tools, we first converted the four-class classification into a two-class classification and divided the RNAs into coding and noncoding RNAs. For this analysis, we regarded lncRNA and tr-ncRNA as noncoding RNA, and mRNA and untr-mRNA as coding RNA. We randomly sampled ten groups of RNAs from the clean datasets to form the Test-basic dataset to compare the performance of TetraRNA, CNCI (version 2) [20], CPC2 (version 1.0.1) [21], CPPred [22], IRSOM [23], LGC (version 1.0) [24], LncADeep (version 1.0) [25], LncFinder (version 1.1.6) [26] and PLEK (version 1.2) [27]. Each test group contained 500 mRNAs, 500 untr-mRNAs, 500 tr-ncRNAs and 500 lncRNAs. In addition, we integrated the prediction results from the eight binary tools by majority voting, and compared them with the predictions of TetraRNA to measure the performance of the tools (details in Supplementary File S2). All the tools ran with default parameters.

In order to assess the performance of TetraRNA in distinguishing lncRNA from tr-ncRNA, we randomly sampled ten groups of RNAs from the clean datasets to form a Test-binary dataset to compare the performance of IRSOM2, LncReader and TetraRNA. Each group contained 1000 tr-ncRNAs and 1000 lncRNAs. IRSOM2, LncReader and TetraRNA were run with default parameters.

2.3. The TetraRNA model provides the first landscape of tetra-classified human RNAs and cncRNAs

To further confirm the reliability of the TetraRNA model, we used six models to predict the types of RNA from the lncRNA-SmProt dataset. All models were run with default parameters (details in Supplementary File S2).

After model evaluation and comparison with other models, the TetraRNA model was used to re-predict the types of RNA in the raw datasets. The parameters were set to default. The results predicted by the TetraRNA model were used for subsequent analysis.

In this study, small protein sequences from Ribo-seq, the literature, related databases and mass spectrometry data were downloaded from the SmProt v2.0 database. Then, BLAST was used to map the sequences of the tr-ncRNA predicted by the TetraRNA model to these small protein sequences. The e-value was set to 1e-5. The number of tr-ncRNAs in the comparison and the number of tr-ncRNAs supported with data from different sources were counted.

2.4. Trait comparison among four types of RNA

2.4.1. Function analysis of four classes of RNA

We used the R package "biomaRt" to convert all gene IDs into entrz ID types for further analysis [28]. To obtain more information about the functions of these human cncRNAs, we used the R package "clusterProfiler" [29] to perform GO and KEGG enrichment analysis. The reactome pathway of the four types of genes was obtained from the Reactome Pathway Knowledgebase [30]. Next, the RNAenrich (http://idrblab.cn/rnaenrich/) web server was used to visualize the results of the GO and signal pathway enrichment [31]. The Benjamini-Hochberg (FDR) adjustment was set as the P-value adjustment method, while other parameters were set to default.

2.4.2. Conservation analysis of the four classes of RNA

The bigWigAverageOverBed tool and bigwig file downloaded from the UCSC Genome Browser were used to obtain the phastCons score and phyloP score for each exon [32–34]. We then calculated the average phastCons score and average phyloP score of each RNA, according to Eq. 4.

$$Transcript\_mean\_score = \frac{\sum_{i=1}^{n}(s_i \times L_{ei})}{L_t} \quad (4)$$

In Eq. 4, s_i is the score of exon i, the L_t is the transcript length and L_{ei} is the length of exon i.

2.4.3. Expression and DNA methylation analyses of the four classes of RNA

The TPM values of single-cell sequencing of human genes in different tissues were taken from The Human Protein Atlas database [35] and the TPM values of RNA-seq of human and mouse were downloaded from ENCODE database [36]. The average methylation scores of genes were obtained from the MethBank database with project IDs GSE127282, GSE127284, GSE127285, GSE127289, GSE127291, GSE127292, GSE127295, GSE127299, GSE127320, GSE127321, GSE16256, GSE46401, and GSE52578, and were divided into several groups according to tissues and sequence regions [37].

2.5. Characterization of peptides derived from tr-ncRNA for exploration of their potential evolutionary process

2.5.1. Statistics pertaining to peptides derived from tr-ncRNAs

BLAST was used to compare the predicted human tr-ncRNAs with the transcripts of *Macaca mulatta* (rhesus macaque), *Gorilla gorilla* (gorilla) and *Pan troglodytes* (chimpanzee) (e-value < 1e-5). The taxon tree was created in TimeTree [38]. Human tr-ncRNAs with homologous sequences in the species *P. troglodytes*, *G. gorilla* and *M. mulatta* were classified into the “Macaca” group; human tr-ncRNAs with homologous sequences in both *P. troglodytes* and *G. gorilla* species were classified into the “Gorilla” group; and human tr-ncRNAs with homologous sequences only in the species *P. troglodytes* were classified into the “Pan” group. We used the ORF Finder [39] to predict the ORF region of each tr-ncRNA and took the longest ORF region as the final prediction result. The ExPASy’s ProtParam [40] tool was used to calculate the instability coefficient and Gravy index, and the proportions of 20 amino acids were assessed in the predicted protein sequences.

2.5.2. Analysis and comparison of mouse and human tr-ncRNAs and lncRNAs

The snoRNAs and haripin microRNAs obtained from the Rfam database [41] and miRBase [42] were mapped to human tr-ncRNAs and lncRNAs, respectively, using BLAST (e-value < 1e-5). The intergenic sequences were obtained according to the annotation file and then were divided into sequences of approximate length based on the average length of the tr-ncRNAs and lncRNAs. Finally, BLAST was used to map hairpin microRNAs and snoRNAs to the intergenic sequences with the same parameters as the random background (details in Supplementary File S2). Mouse lncRNAs were download from GENCODE database [16]. BLAST was then used to find the homologous lncRNA sequences in human and mouse, (e-value 1e-5). The homologous transcript pairs used in further analyses were aligned in MAFFT V7.520 [43] and visualized using GeneDoc [44]. The putative ORFs in the sequences were marked according to the Translnc database [15].

2.6. Statistical analysis

T tests were used to assess the result significance in the Silhouette Coefficient, Phastcons and Phylop scores, as well as the TPM between mRNA and untr-mRNA, and between lncRNA and tr-ncRNA. Moreover, t tests were also used to assess the result significance in the Instability indices and Gravy scores of small proteins coded by tr-ncRNAs among different groups, as well as the TPM between homologous genes in human and mouse.

2.7. Construction of the tetra-RNA database

The online cncRNA database TetraRNADB (http://tetrarnadb.liu-lab.com/) was constructed using the Django framework, and the web front-end was developed using Bootstrap. All data is stored in the SQLite database (Fig. 1F). All the results described in this article and all the feature tables of human tetra-classified RNAs can be browsed and downloaded from TetraRNADB (http://tetrarnadb.liu-lab.com/download/).

3. Results

3.1. Construction of a tetra-classification machine-learning model for efficient prediction of RNA categories

We analyzed a total of about 85,727 mRNAs, 48,915 lncRNAs, 865 untr-mRNAs and 5833 tr-ncRNAs. Of these, mRNA transcripts had the highest average number of exons; cncRNA genes, including untr-mRNA and tr-ncRNA, had the highest average number of transcripts (compared to mRNA and lncRNA genes); and lncRNA genes had the lowest numbers of transcripts and exons (Supplementary Table S2). Through analysis of the distribution of scores of different classes between the eight known lncRNA prediction tools, we found that cncRNAs tended to drift out of their original categories. For instance, the score of untr-mRNAs, which was similar to that of mRNAs, deviated from mRNAs and trended toward noncoding RNAs, while tr-ncRNAs deviated from lncRNAs and trended toward coding RNAs, despite having similar scores to lncRNAs (Supplementary Figure S1). This suggests that coding potential scores in different tools may actually contain some underlying combination of features that could be valuable for the identification of cncRNAs.

We assessed the silhouette coefficient of the C-features of each RNA dataset (Fig. 2A). The results showed that there were significant differences in the degree of intraclass aggregation for different datasets. In comparison, mRNAs showed higher homogeneity (average silhouette coefficient 0.05), while the score of lncRNAs was the lowest (-0.06), indicating there may be a large heterogeneity within this group. Next, we examined the similarities between different datasets as measured by the $R_{(i,j)}$ score (Fig. 2B). Our results showed that mRNAs and lncRNAs had the lowest similarity ($R_{mRNA\text{-}lncRNA} = 0.97$), while the relationships between mRNAs and untr-mRNAs, as well as between lncRNA and tr-ncRNA, were more interconnected (Fig. 2B). We then used the k-means method to pre-estimate the potential number of cncRNAs (more details in Supplementary File S2 and Supplementary Table S3). The estimated results suggest that the real number of cncRNAs may be much larger than the number currently known.

By combining C-features, E-features, the RF algorithm and the DT algorithm, we compared the performance of six different models (Fig. 2C ~ H). The RF-CE model had the highest metrics of all models in all the categories using 10-fold cross-validation. The random forest algorithm was found to be superior to the decision tree; while in feature selection, the combination of E features with C features outperformed either E features or C features alone, suggesting these two features are complementary for model construction. The best performing RF-CE model was finally defined as the TetraRNA model.

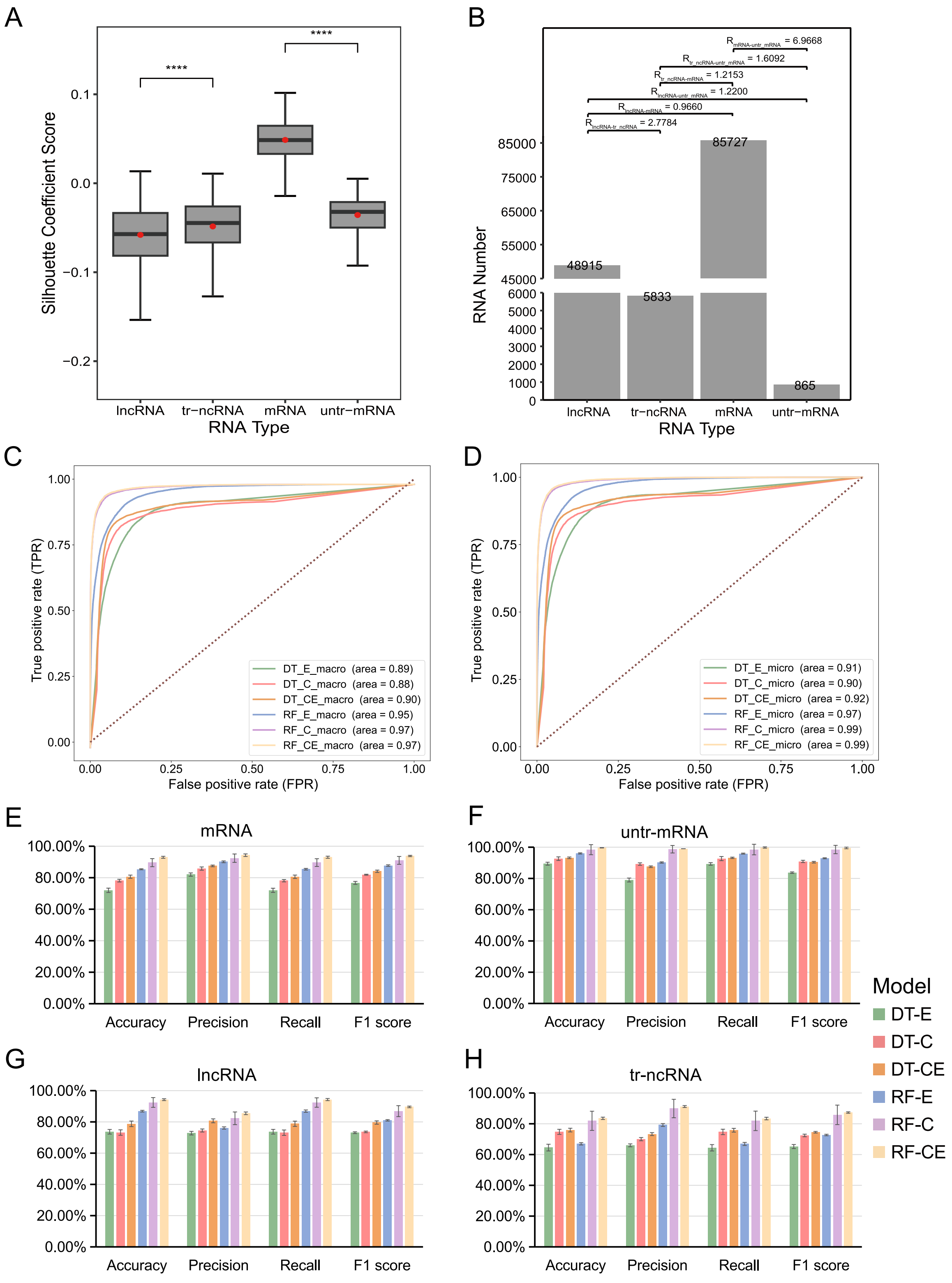

Fig. 2. Characteristics of clean datasets and model evaluation. LncRNAs, mRNAs, tr-ncRNAs and untr-mRNAs were regarded as four different clusters. The silhouette coefficient scores, number of RNAs and the $R_{(i,j)}$ are shown in parts A and B separately. The number of each type of RNA is given on the top of the bar and $R_{(i,j)}$ between each two types of RNA is marked above the bar. T tests were used to assess the result significance among groups. * ** * represents the $0 < \text{p-value} \leq 0.0001$. C, Macro average ROC curves of the six models. D, Micro average ROC curves of the six models. E~H, Accuracy, precision, recall and f1 score of DT-E model, RF-E model, DT-C model, RF-C model, DT-CE model and RF-CE model in the four types of RNAs.

We next evaluated the contribution of each feature to the models using feature importance (Fig. 3 and Supplementary File S4). In all the E-features, the LncADeep score demonstrated the highest importance (0.35), followed by the CNCI score (SVM algorithm) and the LncFiner score (SVM algorithm), with the LGCscore (maximum likelihood estimation) and the IRSOM score (deep learning algorithm) showing relatively lower importance (Fig. 3A). The algorithms seem to vary in power, therefore, it is likely that feature selection in different tools may be responsible for the importance of the different E features. For instance, LncADeep, CNCI and LncFiner all extracted the hexamer score, which naturally reflects base usage frequency and codon bias, as a classification feature. LncADeep and LncFiner further explored the intrinsic composition of the hexamer score. For example, LncADeep extracted the entropy density profile (EDP) of the hexamer score and LncFiner calculated the hexmer's Euclidean-distance and Logarithm-distance, which are likely to facilitate further distinction of RNA categories. According to the feature importance ranking of all the C-features, protein features accounted for the largest proportion (80 %), while structural features were not found in the top 15 features (Fig. 3B). Three ORF-related features (ORFl, ORFc and ORFi) were all listed in the top 15 features, further demonstrating that the ability of RNA to encode proteins plays an important role in distinguishing the four RNA types. In the top 24 features in both the DT-CE model and the RF-CE model (Fig. 3C), peptide/protein features accounted for 50.00 %, far exceeding sequence features (16.67 %) and expert features (33.33 %), indicating that the subgroup of peptide/protein features played an important role in RNA multiple classification. Most of these components represent the peptide/protein differentiation in amino acid composition, physicochemical properties and even function. There was a good agreement with previous results that all the eight components in the E-features showed remarkable importance in CE-feature-involved models, with a ranking above 20. Simultaneously, the components in the C-features were also actively involved in improving models. For example, peptide length, amphiphilic pseudo-amino-acid composition (APAAC), pseudo K-tuple reduced amino acids composition (PseKRAAC), MW and Pseudo-Amino Acid Composition (PAAC) were all listed in the top 10 (Fig. 4C). The subgroups of nucleic-acid sequence features also had different degrees of contribution, including PseDNC_CG (rank 15) and PCPseDNC-CG (rank 17).

3.2. Comparison of TetraRNA with other RNA classification tools

We compared the ability of the new classification model TetraRNA with that of eight classic bianary lncRNA prediction models to distinguish between coding and noncoding RNAs. All models performed well in classifying coding RNA and noncoding RNA (Fig. 4A). However, the

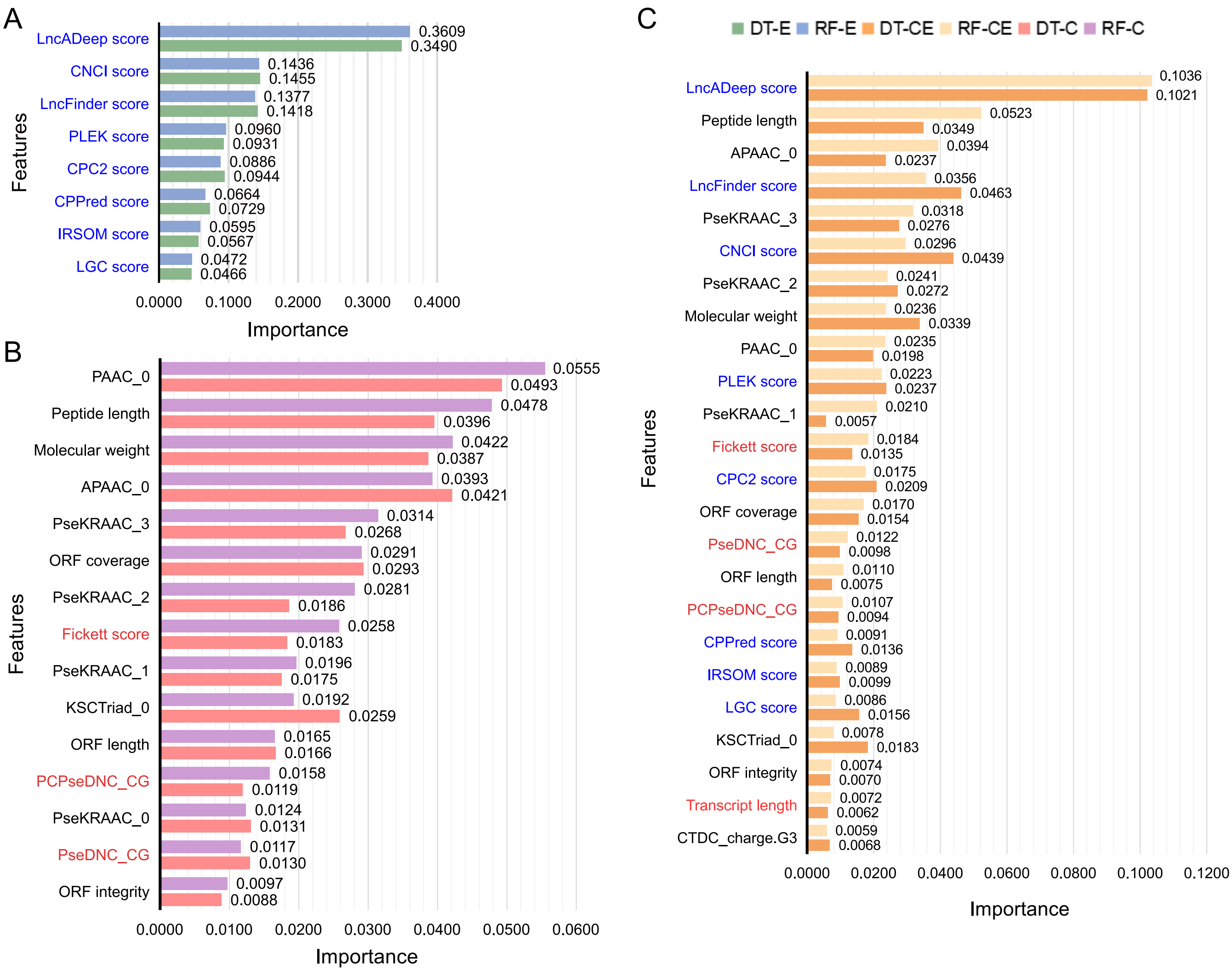

Fig. 3. Feature importance of six models. A, Top 8 features in DT-E model and RF-E model. B, Top 15 features in DT-C model and RF-C model. C, Top 24 features in DT-CE model and RF-CE model. The expert features were marked as blue text; the peptide features were marked as black text and the nucleotide sequence features were marked as red text.

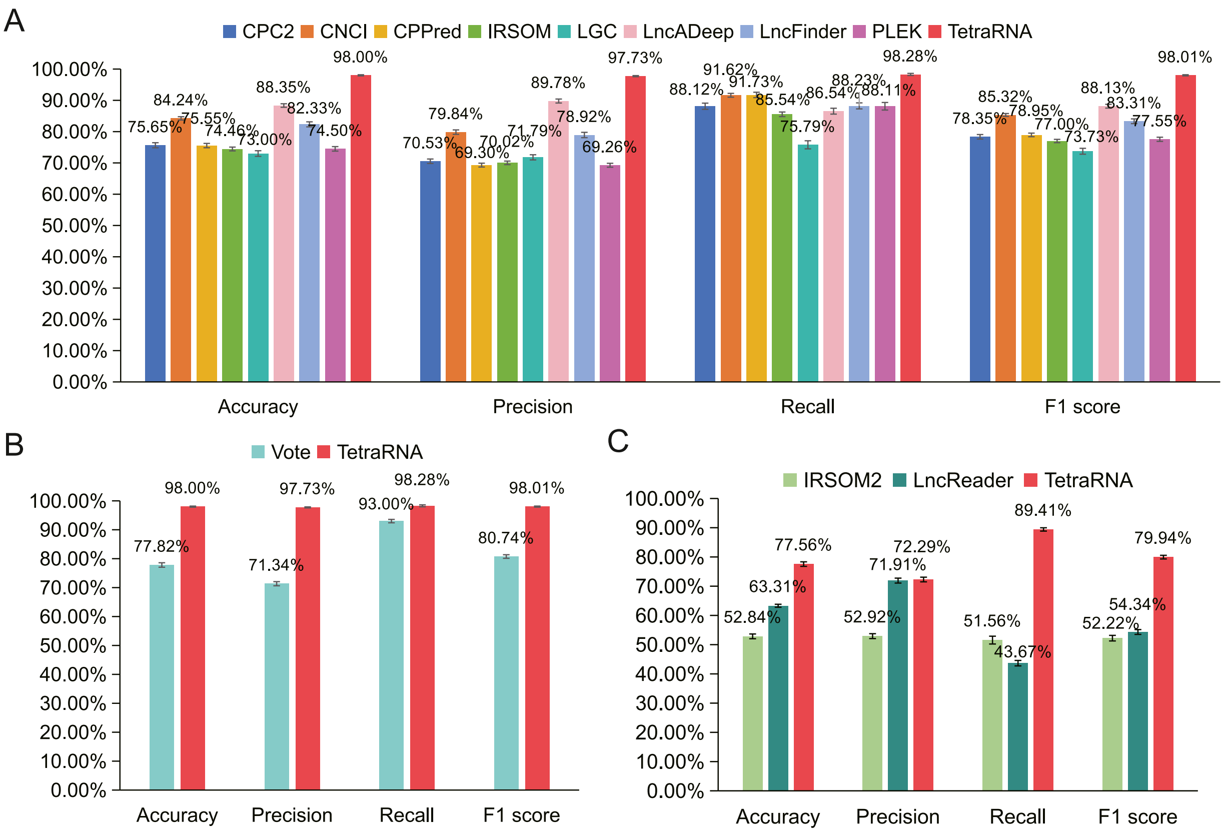

Fig. 4. Comparison of performance in RNA classification tools. A, Accuracy, precision, recall and f1 score of TetraRNA and basic binary-classification tools using the Test-basic dataset. B, Accuracy, precision, recall and f1 score of TetraRNA and basic binary-classification tools through majority voting using the Test-basic dataset. C, Accuracy, precision, recall and f1 score of TetraRNA and bi-functional RNA identifiers using the Test-binary dataset.

TetraRNA model had the highest accuracy, precision, recall and f1-score. We also used the majority voting method (Supplementary File S2) to integrate the results of the eight tools and further compared the classification of TetraRNA with that of common binary classification tools. The results of this analysis also show that TetraRNA had a better ability to distinguish between coding RNA and non-coding RNA than the other models (Fig. 4B). The TetraRNA model is therefore currently the most effective model for distinguishing between coding RNA and non-coding RNAs.

Since both IRSOM2 and LncReader are classification tools that can identify cncRNAs, we further compared the ability of TetraRNA to identify tr-ncRNAs with that of these two models. From Fig. 4C, it can be seen that the TetraRNA model had the highest accuracy, precision, recall and f1-score in tr-ncRNA identification.

3.3. *The TetraRNA model provides the first tetra-classification system of human cncRNAs*

Since changes in class may be associated with altered knowledge of RNA function, we applied the best-performing TetraRNA model to re-predict the categories of all human RNAs. The prediction results revealed a refreshing landscape of human tetra-classified RNAs, including 83,729 mRNAs, 33,425 lncRNAs, 23,235 tr-ncRNAs and 951 untr-mRNAs.

As shown in Fig. 5A, the sizes of all the original RNA categories detected changed significantly. Fewer mRNAs and lncRNAs were identified, with lncRNA numbers decreasing by 38.30 % (from 48,915 to 33,425), which was consistent with our previous clustering results suggesting the huge internal heterogeneity of lncRNAs. In contrast, higher numbers of cncRNAs were identified, and numbers of tr-ncRNAs increased dramatically by almost fourfold (from 5833 to 23,235) (Fig. 5A). Furthermore, each RNA category exhibited varying degrees of divergence. For example, although most of the original mRNAs (85,727 RNAs) were still predicted to be mRNAs (82,101 RNAs), 969 were categorized as tr-ncRNAs and 2635 as lncRNAs (Fig. 5B, Supplementary Figure S2). On the other hand, the original lncRNA category showed the greatest divergence, with over a third of the original lncRNAs newly predicted as tr-ncRNA (17,049 RNAs) (Fig. 5B, Supplementary Figure S2). Intriguingly, most of the original cncRNA categories were predicted into their own categories (89.44 % for tr-ncRNA, 99.88 % for untr-mRNA). These results further demonstrate the vitality of our model for the discrimination and categorization of cncRNAs.

As the category with the largest increase in predicted numbers, tr-ncRNA was further analyzed using the existing benchmark data sources. We compared our predicted tr-ncRNAs with the data in the SmProt database. This database lists peptides and small protein identified using various lines of experimental evidence, including Ribo-seq, mass spectrum (MS), literature and known databases (details in Supplementary File S2). Notably, 95.77 % of predicted tr-ncRNAs (22,250 out of 23,235) matched those in the SmProt database and were mostly supported by Ribo-seq analysis (Fig. 5C-5E). Overall, 1703 tr-ncRNAs were backed by four sources, 5843 tr-ncRNAs by three sources, 7393 tr-ncRNAs by two sources and 7312 tr-ncRNAs by a single source (Fig. 5E). This evidence from many sources further confirmed the reliability of our prediction of tr-ncRNAs.

We next applied all the models to the lncRNA-SmProt dataset, which represented potential tr-ncRNAs. The lncRNA-SmProt dataset was also divided into four categories, but here the predominant category was tr-ncRNA, and lncRNA was the second most important (Fig. 5F). In good agreement with previous results, the TetraRNA (RF-CE) model showed the best performance in identification of tr-ncRNAs, that is, about 65.52 % of the RNAs in the lncRNA-SmProt dataset were identified as tr-

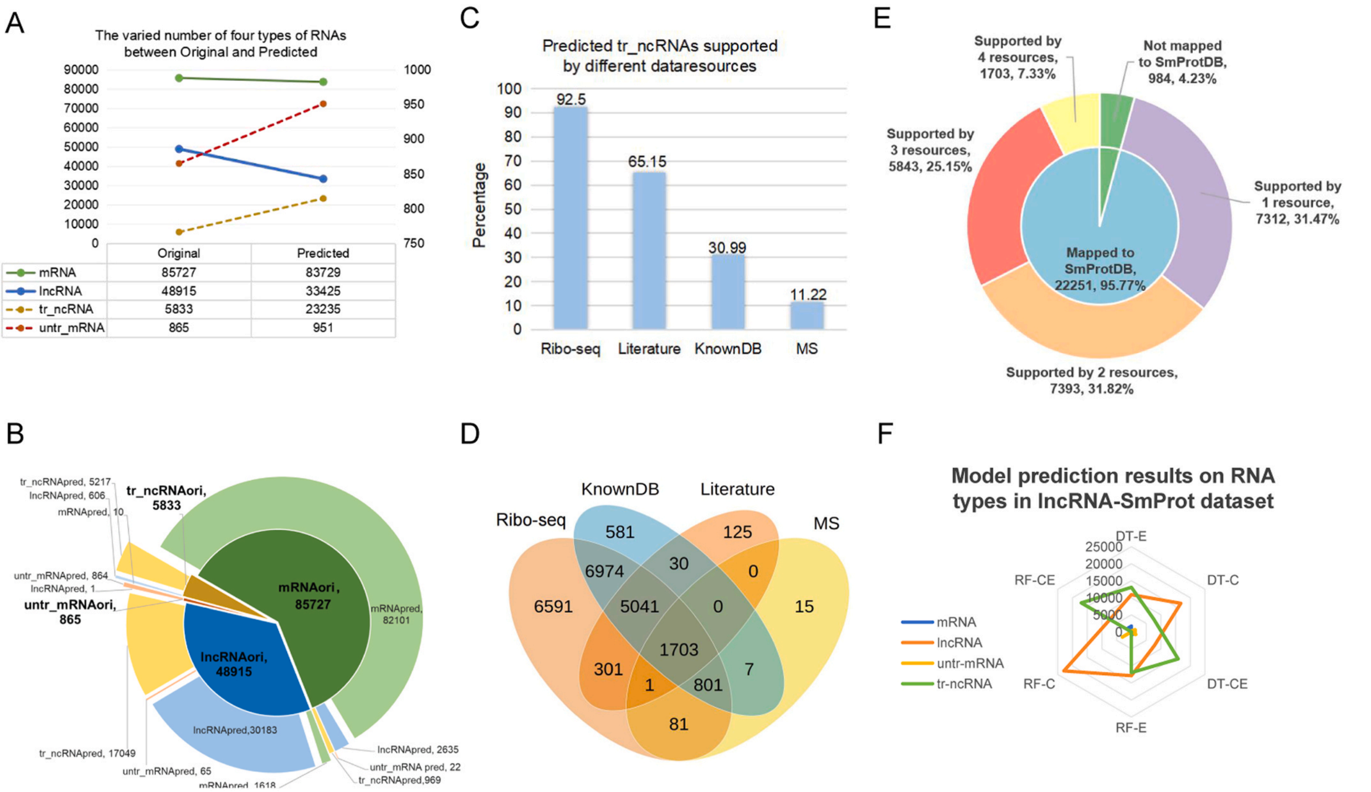

Fig. 5. Evaluation of model prediction results. A, Variation between original and predicted numbers of the four types of RNAs. B, Prediction of RNA types in raw datasets from the RF-CE model. C, The proportion of predicted tr-ncRNA supported by small proteins from ribo-seq profiles, known database (knowndb), literature mining results and mass spectrometry (MS) separately in the SmProt database are shown in histogram. D, Venn diagram showing the number of predicted tr-ncRNA supported by small proteins from ribo-seq profiles, knowndb, literature mining results and MS in SmProt database. E, Small pie chart: the number and proportion of predicted tr-ncRNAs mapped / not mapped to the SmProt database. Large pie chart: the number and proportion of predicted tr-ncRNA supported by small proteins from one to four resources (ribo-seq profiles, knowndb, literature mining results and MS). In each chart, the first number represents the number of RNAs and the second number represents the proportion of the total. **F**, Results of model prediction of RNA type in lncRNA-SmProt dataset. Radar chart showing the number of each type of RNA as predicted by different models using lncRNA-SmPort dataset.

ncRNA, which was a much higher number than predicted by other models (Fig. 5F). These results suggested that the combination of experimental and *in silico* methods has the potential to greatly improve the screening of cncRNAs.

3.4. Trait differentiation among four types of RNA

To determine whether the four RNA classes differ in their attributions, we first investigated two conservation assessment methods, PhastCons (Fig. 6A) and PhyloP (Fig. 6B), to evaluate the levels of conservation of different RNA classes. The PhastCons score calculates the probability that each nucleotide belongs to a conserved element, whereas phyloP is more appropriate for evaluating signatures of selection at particular nucleotides, with positive values implying conservation. We found that the results of these two evaluation systems were perfectly consistent. The conservation of mRNA and untr-mRNA was significantly higher than that of lncRNA and tr-ncRNA, and with tr-ncRNA being more highly conserved than lncRNA. Intriguingly, the conservation of untr-mRNA was also higher than that of mRNA, suggesting that that bifunctional RNAs could be more conserved than the monofunctional RNAs they are similar to. We also extracted eight common features of various types of RNA after prediction. (Suplementary Figure S3). Compared with the other two types of RNA, mRNA and untr-mRNA tended to have longer transcripts, longer ORF regions, higher GC content, and lower free energy. This may be because these two RNAs can encode proteins. Moreover, compared with mRNAs, untr-mRNAs had shorter transcript and ORF lengths, lower GC content in the CDS region, a smaller proportion of the ORF region, and higher free energy. This may be because untr-mRNA can also function directly as an RNA, so the various features of untr-mRNA will tend to be close to those of the other two types of RNA. Similar phenomena were also seen in lncRNAs and tr-ncRNAs (Suplementary Figure S3).

We next examined the expression levels of genes in the four RNA classes in various tissues and organs (Fig. 6C). Here, there were obvious differences among all the RNA categories. The expression levels of untr-mRNA were the highest while those of lncRNA were the lowest. In general, the expression of untr-mRNA was higher than mRNA in all tissues except the testis, while the expression of tr-ncRNA was higher than that of lncRNA in all tissues except the brain. In addition, untr-mRNA showed explosive expression (fold change >4) compared with mRNA in vascular, skeletal muscles, brain and adipose tissue, which would exert strong regulatory activity on the metabolism and cell differentiation and development. Compared with lncRNA, the expression of tr-ncRNA was significantly increased in the skin, ovary, pancreas, kidney and colon (fold change > 2), with the highest expression occurring in the ovary, skin, heart and immune system.

We next took advantage of single-cell sequencing data to examine the expression of four classes of RNAs in different types of brain and testis cells (Fig. 6D and E). We found that the expression levels of both tr-ncRNAs and lncRNAs were higher than those of mRNAs in all kinds of brain cells. Moreover, the expression of tr-ncRNA in all kinds of cells in the brain was lower than that of lncRNA. That is, the expression of tr-ncRNA in the brain was restricted or finely regulated compared to other tissues. We found the 3’UTR region of tr-ncRNAs had a greater CpG length and higher average level of methylation than did this region of lncRNA in the brain, which may explain the lower expression levels

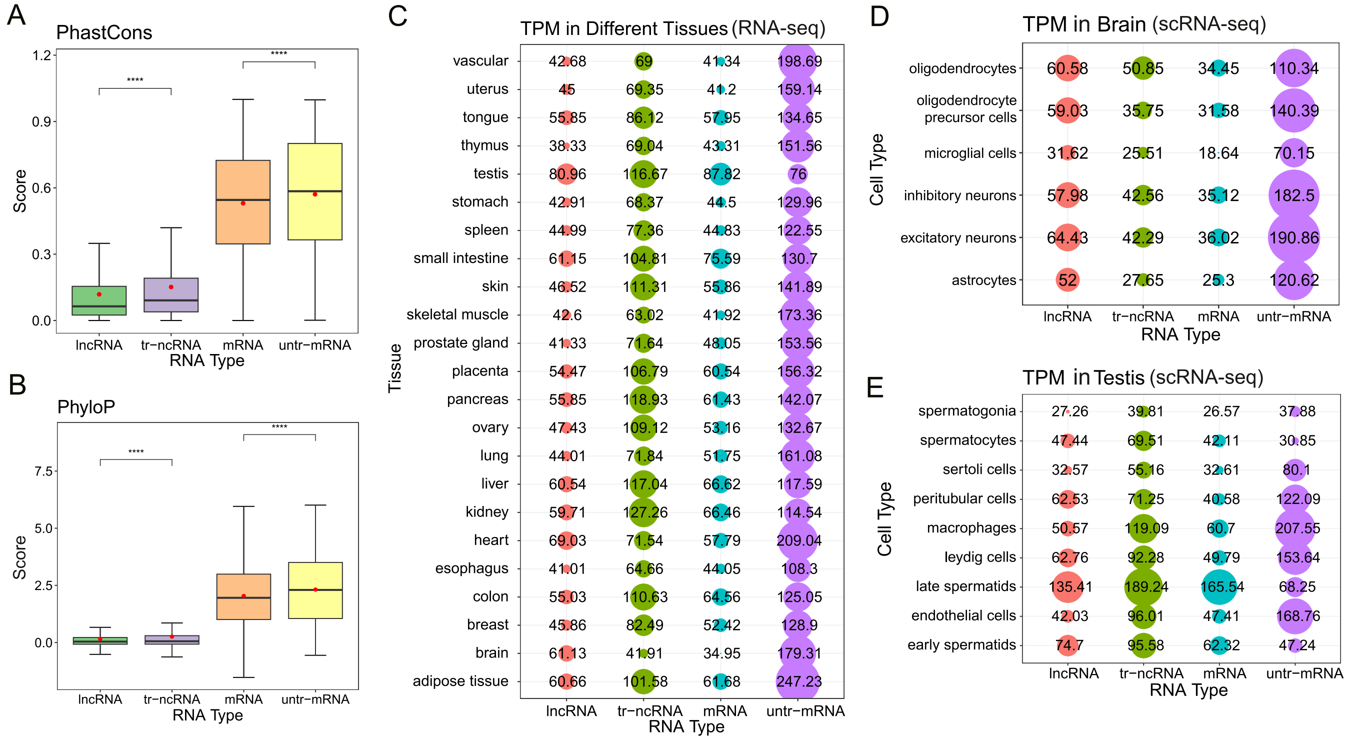

Fig. 6. Functional feature analysis of different human RNA classes. A, The scores of PhastCons and B, PhyloP of human lncRNA, tr-ncRNA, mRNA, and untr-mRNA. Higher scores represent higher degrees of conservation. C, Expression levels of human lncRNA, tr-ncRNA, mRNA, and untr-mRNA in 23 tissues calculated using RNA-seq. D, Expression levels of human lncRNA, tr-ncRNA, mRNA, and untr-mRNA in the brain calculated using scRNA-seq. E, Expression levels of human lncRNA, tr-ncRNA, mRNA, and untr-mRNA in the testis as calculated using scRNA-seq. The dots and the numbers represent the TPM values of the RNAs. The bigger the dot, the higher the TPM. T tests were used to assess the result significance among groups. * ** * represents 0 < p-value ≤ 0.0001.

seen in tr-ncRNAs. Our enrichment results showed that tr-ncRNA was involved in Alzheimer's disease (Supplementary Figure S4). We also found the average methylation levels of the untr-mRNA promoter region were higher than those of the mRNA promoter region in the testis, indicating that the transcription of untr-mRNA is inhibited in the testis compared with that of mRNA, and resulting in lower expression levels in testis tissue (Fig. 6C and Supplementary Figure S5). However, the expression of untr-mRNA was found to be higher than that of mRNA in most cell types in testis with the exception of spermatocytes, and late and early spermatids, suggesting that untr-mRNA might have differential functions in the testis during spermatogenesis.

3.5. *Characterization of peptides derived from tr-ncRNA and exploration of their potential evolutionary processes*

To further investigate peptides derived from tr-ncRNAs, we constructed homologous peptide families in primate species, including *M. mulatta*, *G. gorilla* and *P. troglodytes*. In total, there were 21 homologs in the "Macaca" group, 69 in the "Gorilla" group and 155 in the "Pan" group (Fig. 7A). Apparently, most of the tr-ncRNA- derived peptides are not conserved, and the number of homologs increased as the evolutionary time (time between species divergence) shortened. Next, we evaluated the intrinsic characteristics of these conserved peptides in terms of their instability coefficient and hydrophobicity (Gravy index), which are closely related to their functional importance and structural stability [45]. As shown in Fig. 7B, the stability of the peptides in the "Pan" group, which have a relatively short evolutionary period, is poor (instability coefficient > 50), while those peptides in the "Macaca" and "Gorilla" groups were significantly more stable (instability coefficient < 40). More interestingly, we found that, with the elongation of evolutionary period, the hydrophobicity of the peptides gradually decreased, and there were significant differences in hydrophobicity between different evolutionary stages (p-value < 0.05) (Fig. 7C).

We also evaluated the proportion change of all amino acids among different groups (Fig. 7D). The peptides in the newly evolved "Pan" group tended to have an increased proportion of amino acids containing functionally-complex residues, such as "Leucine (L)" zippers, which are involved in the formation of transmembrane helices, "Serine (S)", which can be used as a prosthetic group to bind metal ions [46–48], and "Tryptophan (W)", which iscapable of UV absorption and fluorescence emission [49]. In addition, the more conserved the peptide, the more stable the residue was in the structure. For example, Alanine (A), which does not have long side chains to form a special conformation and can therefore occur anywhere in a protein to increase structural flexibility [50], was significantly elevated in the relatively conserved peptides in the "Macaca" group. Likewise, Asparagine (N), which often appears at the corner of protein folding structures and can form flexible protein scaffolds with other amino acids [51], and Glutamate (E) whose peptide bond is stable without any active sites for acid proteases [52], were also markedly increased in the conserved peptides.

To further investigate why tr-ncRNAs were more highly conserved than lncRNAs, we mapped lncRNAs and tr-ncRNAs to known microRNAs and snoRNAs, and compared the distribution of these small RNAs in different regions, including tr-ncRNA, lncRNA, and randomly selected intergenic regions. As shown in Fig. 8A, the distribution ratios of small RNAs in both tr-ncRNA and lncRNA were significantly higher than in the intergenic regions, while the ratio of those mapping to tr-ncRNAs was significantly higher than that in lncRNA (Chi-Squared Test, P-value<0.01). These data suggest that the emergence of tr-ncRNAs may be correlated with the emergence of functional small RNAs in these regions, and these conserved functional small RNAs may in turn confer high sequence conservation to the tr-ncRNAs. Previous studies have also shown that some ncRNAs can serve as host genes for small RNAs, which have higher expression level and unique characteristics compared to other ordinary lncRNAs, and some may even evolve functional ORFs [53,54].

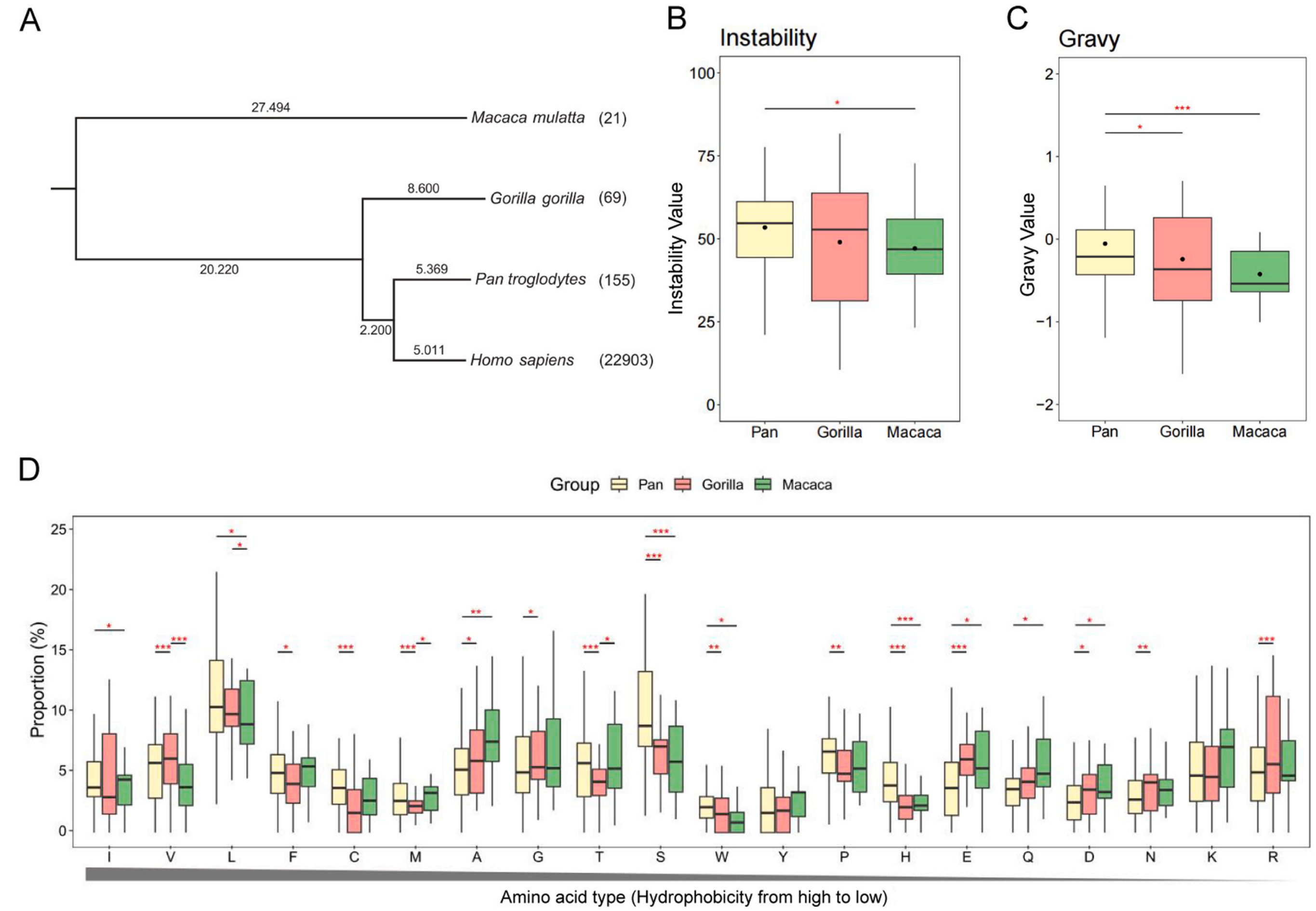

Fig. 7. The characters of peptide sequences of human tr-ncRNAs conserved in *Macaca mulatta*, *Gorilla gorilla* and *Pan troglodytes*. A, Taxon tree of four primate species. The evolution time (MYA) is marked on the branch. The number after the species name is the number of conserved tr-ncRNAs. B, The instability indices of small proteins coded by tr-ncRNAs. C, The Gravy values of small proteins coded by tr-ncRNAs. D, The proportion of 20 types of amino acids sorted by hydrophobility. Human tr-ncRNAs divided in different groups are represented by different colors. The t test was used to assess the result significance among groups. * ** represents $0 < \text{p-value} \leq 0.0001$, * * represents p-values from $0.0001 < \text{p-value} \leq 0.01$, and * represents p-values from $0.01 < \text{p-value} \leq 0.05$.

To further investigate this, we selected the tr-ncRNA gene *GAS5* (ENSG00000234741) and its mouse homolog (ENSMUSG00000053332) for further analysis. The *GAS5* gene is known as a tumor suppressor gene that regulates cell proliferation, invasion, migration, and apoptosis [55]. The *GAS5* genes in both human and mouse are the host genes of snoRNAs, and human *GAS5* gene has been reported to contain 10 snoRNAs in their sequences [56]. The annotation file suggested that the human *GAS5* gene has more than 100 transcripts, much higher than the average, indicating a very high transcription frequency. Significant differences in multiple tissue expression between human and mouse *GAS5* genes were also found, suggesting possible functional differentiation between them (Fig. 8B). We next aligned the sequences of the human and mouse *GAS5* genes (Supplementary File S5 and S6) and assessed the alignment identity of different regions (Figs. 8C and 8D). The identity of the entire *GAS5* gene sequence was only 44.96 %, indicating that this gene may have undergone rapid evolution. The snoRNA region was the most conserved (80.72 %), while the ORF region was the least conserved (41.96 %), which may be related to the main feature of this gene as a snoRNA host rather than in protein coding. Finally, we draw a schematic diagram illustrating the alignment results (Fig. 8d), and marked the ORF regions according to the Translnc database. Consistent with the annotation file records, most of the regions in the human GAS5 gene sequence are transcribed into RNA (marked as exon regions in Fig. 8d). The human *GAS5* gene is highly transcribed, and most regions of its gene sequence are covered by transcripts, on which small ORFs were detected. The overall conservation between human and mouse *GAS5* was very low, but we found that the sequences were highly conserved in their special functional snoRNA regions. Some transcripts of the human *GAS5* gene were identified as tr-ncRNAs, but mouse *GAS5* transcripts were currently considered to be lncRNAs. Based on these analyses of *GAS5* gene pairs, we speculate that the RNAs encoded by the human *GAS5* gene became tr-ncRNA rather than lncRNA probably because of a special structure on the gene sequence (for example the snoRNA coding region), which allows more ribosomes to bind and results in a higher transcription frequency and higher expression levels. Peptides encoded by this gene are therefore not completely degraded and can be eventually detected. Although this speculation still needs verification, it may give us a deeper understanding of tr-ncRNAs and lncRNAs.

4. Discussion

Current RNA identification methods primarily concentrate on determining whether a transcript has the potential to encode a protein, stemming from historical cognitive limitations concerning RNA functions. However, recent research into RNAs suggests that RNA categories should not be only limited to two. Indeed, Kang et al. suggested that the coding capacity of transcripts should not be simply summarized as coding or non-coding, but rather as a continuous quantitative index [57]. In this study, we designed for the first time an RNA multiple

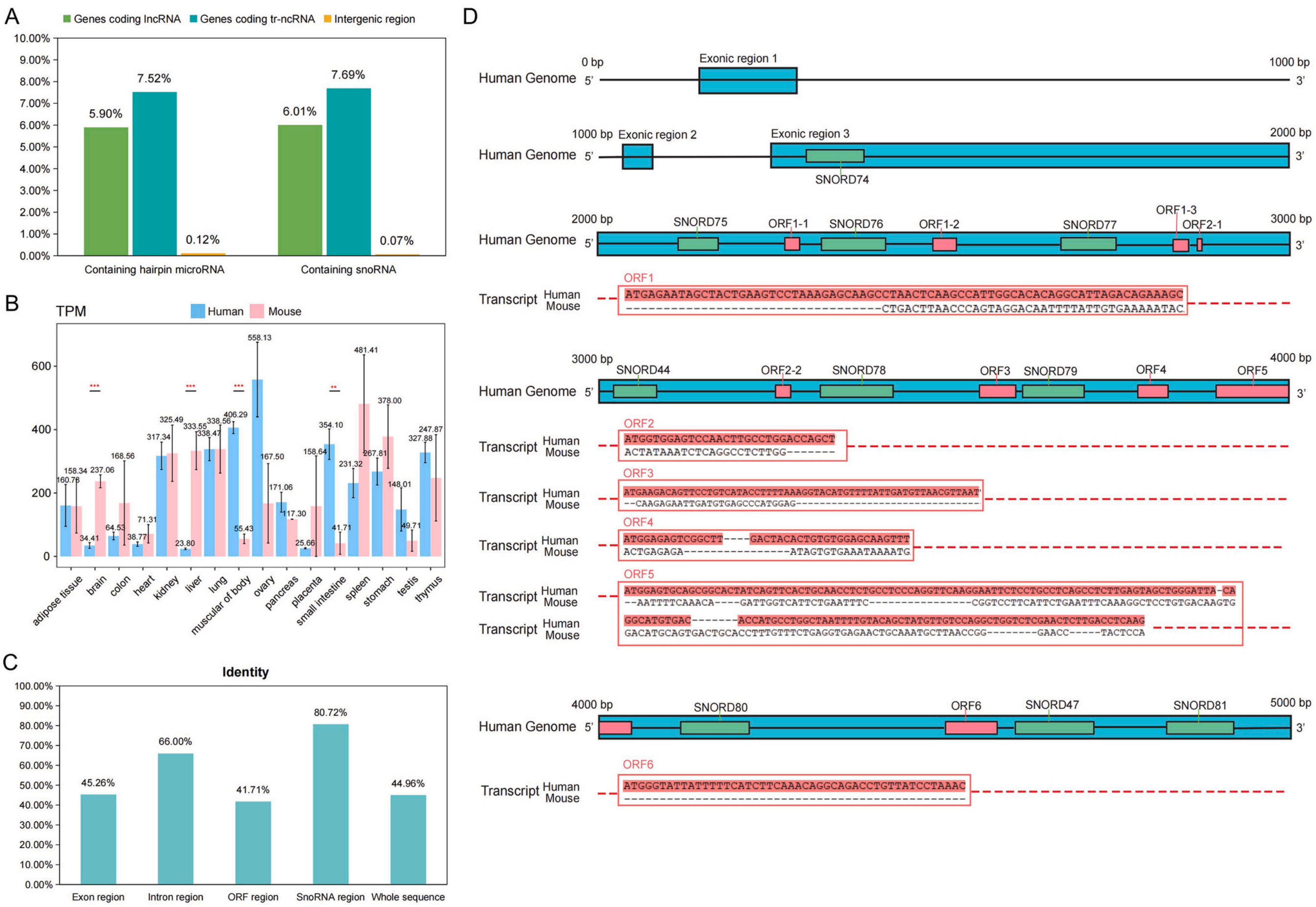

Fig. 8. Analysis of homologous sequences in human and mouse. A, The proportion of tr-ncRNA, lncRNA and intergenic regions mapped to microRNA and snoRNA. B, Expression levels of homologous genes in different tissues in human and mouse. T tests were used to assess the result significance among groups. * ** represents the 0 < p-value ≤ 0.0001, and * * represents the 0.0001 < p-value ≤ 0.01. C, Identity of different regions in the alignment of human and mouse homologous gene sequences. D, Schematic diagram of human and mouse homologous gene sequence alignment. The blue boxes represent exons, the red boxes represent ORF regions, and the green boxes represent snoRNAs. The alignment results of the ORF regions of transcripts are shown in the figure.

classification model based on machine learning methods, and succeeded in effectively identifying the bifunctional RNAs from the previously determined categories of mRNA or lncRNA. Previously, tr-ncRNAs and untr-mRNAs were categorized with mRNA and lncRNA, making them hard to examine in detail. However, our model found a high number of tr-ncRNAs and untr-mRNAs, which will form the foundation for further research into these two forms of RNA. We then used our model to repredict the categories of all human RNAs. Our results demonstrated that the abundance of cncRNAs was significantly underestimated, particularly in the case of tr-ncRNAs. This suggests that the translation of small proteins in non-coding regions is likely widespread, which is consistent with recent large-scale mass spectrometry and ribosome analysis [58, 59].

To further clarify the significance of the production of these peptides, we extracted the encoded products of human tr-ncRNAs and compared their characteristics with the homologous tr-ncRNA products in other species and divided into different groups. Our results indicated that the more conserved the peptides are, the lower their hydrophobicity and the higher their stability. Our findings are consistent with those of Kesner et al., who found that the polypeptide produced by the untranslated region of RNA has a hydrophobic C-terminal tail, and the less evolved the polypeptide, the stronger the hydrophobicity [60]. Meanwhile, the amino acid components of peptides have been partially adapted to the requirements of structural stability and functional evolution. These analyses are likely to afford us profound insights into the de novo genesis of proteins. Through careful distinction, we can differentiate between random occurrences and those that signify the pinnacle of evolutionary processes. In the long run, such recent investigations, including the development of RNA Affinity Purification Sequencing (RAP-seq), have endowed us with the capacity to establish unequivocal linkages between RNAs and proteins. This advancement, coupled with the identification of common motifs in RNA-binding proteins and the understanding of RNA's role in methylation processes, enhances our understanding of the intricate molecular mechanisms underlying biological systems.

We performed a global comparison between the different categories of RNAs, and found that they were significantly different from each other in terms of sequence characteristics, evolutionary degree, and other aspects. It is worth noting that although the coding potential scores of untr-mRNAs were lower than those of mRNAs, their PhastCons and PhyloP scores were higher. The factors driving the shift from coding to noncoding is therefore interesting and worthy of further exploration. On the other hand, tr-ncRNAs have relatively higher PhastCons and PhyloP scores than lncRNAs, indicating that tr-ncRNAs have been subjected to greater selection pressure compared to lncRNAs. Analysis of the expression levels of genes of four RNA classes also showed higher expression levels and specific tissue and organ expression profiles in untr-mRNAs and tr-ncRNAs. All these results suggest that untr-mRNAs and tr-ncRNAs have evolved away from mRNAs and lncRNAs, and have different roles.

Our analysis of tr-ncRNAs has revealed two questions. Firstly, we found that the higher conservation of tr-ncRNAs than of lncRNAs is not due to the peptides they encode. Is it the small RNAs they encode that improve the conservation? Secondly, if the peptides of tr-ncRNAs arose randomly, why are the ORFs in some ncRNAs translated at higher frequency than others? Perhaps the small RNAs they encode could also shed light on this problem. It is possible that these functional small RNAs need to be transcribed in large quantities and are therefore present in higher copy numbers in order to function under selection pressure. Transcription levels in this case would be much higher than those of ordinary lncRNAs, ultimately leading to better capture of ribosome units and subsequent translation of the ORFs located on them into proteins. These randomly translated small peptides may further evolve into new functional proteins.

Our model still requires improvement. Feature selection and model complexity present another area for consideration. Currently, tree-based models are chosen for their high interpretability, aiding in the comprehension of the model's internal mechanisms. Nevertheless, given the potential availability of larger datasets in the future, it is advisable to delve into deep learning techniques. Deep learning models might be able to capture more complex relationships in the data, leading to more accurate and powerful predictions. After sorting feature importance, we found that protein features have the most significant impact on the classification of RNAs. The strong influence of protein features implies that the widespread application of mass spectrometry could potentially be a powerful tool in the identification of tr-ncRNAs. Mass spectrometry can provide detailed information about the associated proteins, which in turn can help to accurately classify RNAs. Furthermore, the lack of more reliable structural features is likely to be the reason why few untrmRNAs have been discovered to date. The prediction of an untrmRNA suggests that the functions of its potential secondary structure should be thoroughly analyzed. This study was limited by computational efficiency and consistency across RNA categories, and we therefore extracted structural features using a sliding window. Obtaining more representative structural features of full-length RNA could potentially enhance the performance of the model. Proper weighting of structure- and translation-related features could be a potential improvement in the future. In addition, the TetraRNA model could be improved by incorporating known RBP interaction motifs, ribosome occupancy scores, or structural motifs relevant to RNA function in the future.

Although our model's predictions show good alignment with existing databases, the absence of functional experiments such as ribosome profiling and mass spectrometry means that the biological validation is incomplete. Experimental validations of the model results remain necessary. These functional experiments can provide more in-depth insights into the biological mechanisms underlying the model's predictions, strengthening the reliability and scientific value of the model. By focusing on these limitations and future directions, we hope to enhance the model's future performance. Moreover, model generalizability can be improved in the future. Our model, trained on human RNA data, has yet to prove its effectiveness in non-human species. This lack of cross-species validation restricts its broader application in biology. To address this, future studies are planned to incorporate cross-species datasets into the training process, which could potentially expand the model's scope and applicability. In conclusion, while our model has shown promising results, there are clear areas for improvement.

This study examined the landscape of human RNAs from the standpoint of four different classification groups. It is our hope that this will prompt researchers to a fresh viewpoint on RNA function. We have summarized the results of our work in the database TetraRNADB (http://tetrarnadb.liu-lab.com/), including the landscape of human tetra-classification RNAs, the functional differences among four types of RNA methylation levels, function enrichment, conservation and expression levels, the normalized features of human RNAs and the TetraRNA model. Our research is poised to contribute significantly to the RNA world hypothesis, shedding new light on the fundamental roles of RNAs and their potential in catalysis and regulation.

CRediT authorship contribution statement

Bai Hanrui: Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Data curation. **Wang Jie:** Software, Methodology, Formal analysis. **Yang Wenjing:** Methodology, Investigation. **Yang Zitian:** Formal analysis. **Jiang Xiaoke:** Visualization, Software, Methodology, Data curation. **Guo Zhen:** Software, Methodology. **Li Jing:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Conceptualization. **Liu Changning:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

We are very grateful to the members in the laboratory for their helpful discussions and technical assistance. This work was supported by the National Natural Science Foundation of China (Grant No. 31970609), 'Top Talents Program in Science and Technology' from Yunnan Province and Startup Fund from Xishuangbanna Tropical Botanical Garden. The funders had no role in the study design, data collection, analysis and interpretation, or preparation of the manuscript. Publication costs are funded by the National Natural Science Foundation of China (Grant No. 31970609).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.03.039.

Data Availability

The database TetraRNADB is freely available online without login requirement at: http://tetrarnadb.liu-lab.com/. The landscape of human tetra-classification RNAs and the TetraRNA model are involved in the TetraRNADB database.

References

[1] Yang WC, Katinakis P, Hendriks P, Smolders A, de Vries F, Spee J, et al. Characterization of GmENOD40, a gene showing novel patterns of cell-specific expression during soybean nodule development. Plant J 1993;3:573–85. https://doi.org/10.1046/j.1365-313x.1993.03040573.x.

[2] Rohrig H, Schmidt J, Miklashevichs E, Schell J, John M. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. Proc Natl Acad Sci 2002;99:1915–20. https://doi.org/10.1073/pnas.022664799.

[3] Ephrussi A, Lehmann R. Induction of germ cell formation by oskar. Nature 1992; 358:387–92. https://doi.org/10.1038/358387a0.

[4] Kanke M, Jambor H, Reich J, Marches B, Gstir R, Ryu YH, et al. oskar RNA plays multiple noncoding roles to support oogenesis and maintain integrity of the germline/soma distinction. RNA 2015;21:1096–109. https://doi.org/10.1261/rna.048298.114.

[5] Kumari P, Sampath K. cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. Semin Cell Dev Biol 2015;47–48:40–51. https://doi.org/10.1016/j.semcdb.2015.10.024.

[6] Li J, Liu C. Coding or noncoding, the converging concepts of RNAs. Front Genet 2019;10:496. https://doi.org/10.3389/fgene.2019.00496.

[7] Bonilauri B, Holetz FB, Dallagiovanna B. Long non-coding RNAs associated with ribosomes in human adipose-derived stem cells: from RNAs to microproteins. Biomolecules 2021;11. https://doi.org/10.3390/biom11111673.

[8] Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science 2013;341:1116–20. https://doi.org/10.1126/science.1238802.

[9] Ulveling D, Francastel C, Hubé F. Identification of potentially new bifunctional RNA based on genome-wide data-mining of alternative splicing events. Biochimie 2011;93:2024–7. https://doi.org/10.1016/j.biochi.2011.06.019.

[10] Liu T, Zou B, He M, Hu Y, Dou Y, Cui T, et al. LncReader: identification of dual functional long noncoding RNAs using a multi-head self-attention mechanism. Brief Bioinform 2023;24 (https://doi.org/Avast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants).

[11] Postic G, Tav C, Platon L, Zehraoui F, Tahi F. IRSOM2: a web server for predicting bifunctional RNAs. Nucleic Acids Res 2023;51:W281–8. https://doi.org/10.1093/nar/gkad381.

[12] Li Y, Zhou H, Chen X, Zheng Y, Kang Q, Hao D, et al. SmProt: a reliable repository with comprehensive annotation of small proteins identified from ribosome profiling. Genom Proteom Bioinforma 2021;19:602–10. https://doi.org/10.1016/j.gpb.2021.09.002.

[13] Boschiero C, Dai X, Lundquist PK, Roy S, Christian de Bang T, Zhang S, et al. MtSSPdb: the medicago truncatula small secreted peptide database. Plant Physiol 2020;183:399–413. https://doi.org/10.1104/pp.19.01088.

[14] Huang Y, Wang J, Zhao Y, Wang H, Liu T, Li Y, et al. cncRNAdb: a manually curated resource of experimentally supported RNAs with both protein-coding and noncoding function. Nucleic Acids Res 2021;49:D65–70. https://doi.org/10.1093/nar/gkaa791.

[15] Lv D, Chang Z, Cai Y, Li J, Wang L, Jiang Q, et al. TransLnc: a comprehensive resource for translatable lncRNAs extends immunopeptidome. Nucleic Acids Res 2022;50:D413–20. https://doi.org/10.1093/nar/gkab847.

[16] Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. Gencode 2021. Nucleic Acids Res 2021;49:D916–23. https://doi.org/10.1093/nar/gkaa1087.

[17] Chen Z, Liu X, Zhao P, Li C, Wang Y, Li F, et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. Nucleic Acids Res 2022;50:W434–47. https://doi.org/10.1093/nar/gkac351.

[18] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet 2000;16:276–7. https://doi.org/10.1016/s0168-9525(00)02024-2.

[19] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn. Mach Learn Python 2011;12:2825–30.

[20] Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res 2013;41:e166. https://doi.org/10.1093/nar/gkt646.

[21] Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res 2017;45:W12–6. https://doi.org/10.1093/nar/gkx428.

[22] Tong X, Liu S. CPPred: coding potential prediction based on the global description of RNA sequence. Nucleic Acids Res 2019;47:e43. https://doi.org/10.1093/nar/gkz087.

[23] Platon L, Zehraoui F, Bendahmane A, Tahi F. IRSOM, a reliable identifier of ncRNAs based on supervised self-organizing maps with rejection. Bioinformatics 2018;34:i620–8. https://doi.org/10.1093/bioinformatics/bty572.

[24] Wang G, Yin H, Li B, Yu C, Wang F, Xu X, et al. Characterization and identification of long non-coding RNAs based on feature relationship. Bioinformatics 2019;35:2949–56. https://doi.org/10.1093/bioinformatics/btz008.

[25] Yang C, Yang L, Zhou M, Xie H, Zhang C, Wang MD, et al. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. Bioinformatics 2018;34:3825–34. https://doi.org/10.1093/bioinformatics/bty428.

[26] Han S, Liang Y, Ma Q, Xu Y, Zhang Y, Du W, et al. LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. Brief Bioinform 2019;20:2009–27. https://doi.org/10.1093/bib/bby065.

[27] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinforma 2014;15:311. https://doi.org/10.1186/1471-2105-15-311.

[28] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 2009;4:1184–91. https://doi.org/10.1038/nprot.2009.97.

[29] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.

[30] Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res 2022;50:D687–92. https://doi.org/10.1093/nar/gkab1028.

[31] Zhang S, Amahong K, Zhang Y, Hu X, Huang S, Lu M, et al. RNAenrich: a web server for non-coding RNA enrichment. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad421.

[32] Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 2010;26:2204–7. https://doi.org/10.1093/bioinformatics/btq351.

[33] Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC genome browser database: 2021 update. Nucleic Acids Res 2021;49:D1046–57. https://doi.org/10.1093/nar/gkaa1070.

[34] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110–21.

[35] Pontén F, Jirström K, Uhlen M. The human protein atlas–a tool for pathology. J Pathol 2008;216:387–93. https://doi.org/10.1002/path.2440.

[36] Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the encyclopedia of DNA elements (ENCODE) data portal. Nucleic Acids Res 2020;48:D882–9. https://doi.org/10.1093/nar/gkz1062.

[37] Zhang M, Zong W, Zou D, Wang G, Zhao W, Yang F, et al. MethBank 4.0: an updated database of DNA methylation across a variety of species. Nucleic Acids Res 2023;51:D208–16. https://doi.org/10.1093/nar/gkac969.

[38] Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, et al. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol 2022;39. https://doi.org/10.1093/molbev/msac174.

[39] Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF-FINDER: a vector for high-throughput gene identification. Gene 2002;282:33–41. https://doi.org/10.1016/s0378-1119(01)00819-8.

[40] Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, et al. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 2012;40:W597–603. https://doi.org/10.1093/nar/gks400.

[41] Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 2021;49:D192–200. https://doi.org/10.1093/nar/gkaa1047.

[42] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47:D155–62. https://doi.org/10.1093/nar/gky1141.

[43] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.

[44] Nicholas KB, Nicholas HB. GeneDoc: a tool for editing and annotating multiple sequence alignments. EMBnet N 1997:1–4.

[45] Chen H, Kim J, Kendall DA. Competition between functional signal peptides demonstrates variation in affinity for the secretion pathway. J Bacteriol 1996;178:6658–64.

[46] Gurezka R, Laage R, Brosig B, Langosch D. A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. J Biol Chem 1999;274:9265–70. https://doi.org/10.1074/jbc.274.14.9265.

[47] Gurezka R, Langosch D. In vitro selection of membrane-spanning leucine zipper protein-protein interaction motifs using POSSYCCAT. J Biol Chem 2001;276:45580–7. https://doi.org/10.1074/jbc.M105362200.

[48] Walther TH, Ulrich AS. Transmembrane helix assembly and the role of salt bridges. Curr Opin Struct Biol 2014;27:63–8. https://doi.org/10.1016/j.sbi.2014.05.003.

[49] Liu H, Zhang H, Jin B. Fluorescence of tryptophan in aqueous solution. Spectrochim Acta Part A: Mol Biomol Spectrosc 2013;106:54–9. https://doi.org/10.1016/j.saa.2012.12.065.

[50] Kemp DS, Boyd JG, Muendel CC. The helical s constant for alanine in water derived from template-nucleated helices. Nature 1991;352:451–4. https://doi.org/10.1038/352451a0.

[51] Cardoso MH, Chan LY, Candido ES, Buccini DF, Rezende SB, Torres MDT, et al. An N-capping asparagine-lysine-proline (NKP) motif contributes to a hybrid flexible/stable multifunctional peptide scaffold. Chem Sci 2022;13:9410–24. https://doi.org/10.1039/d1sc06998e.

[52] Godoy-Ruiz R, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM. Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. J Mol Biol 2004;336:313–8. https://doi.org/10.1016/j.jmb.2003.12.048.

[53] Monziani A, Ulitsky I. Noncoding snoRNA host genes are a distinct subclass of long noncoding RNAs. Trends Genet 2023;39:908–23. https://doi.org/10.1016/j.tig.2023.09.001.

[54] Sun Q, Song YJ, Prasanth KV. One locus with two roles: microRNA-independent functions of microRNA-host-gene locus-encoded long noncoding RNAs. Wiley Inter Rev RNA 2021;12:e1625. https://doi.org/10.1002/wrna.1625.

[55] Kaur J, Salehen N, Norazit A, Rahman AA, Murad NAA, Rahman NMANA, et al. Tumor suppressive effects of GAS5 in cancer cells. Noncoding RNA 2022;8:39. https://doi.org/10.3390/ncrna8030039.

[56] Smith CM, Steitz JA. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. Mol Cell Biol 1998;18:6897–909. https://doi.org/10.1128/MCB.18.12.6897.

[57] Kang Y-J, Li J-Y, Ke L, Jiang S, Yang D-C, Hou M, et al. Quantitative model suggests both intrinsic and contextual features contribute to the transcript coding ability determination in cells. Brief Bioinforma 2022;23:bbab483. https://doi.org/10.1093/bib/bbab483.

[58] Tharakan R, Sawa A. Minireview: novel micropeptide discovery by proteomics and deep sequencing methods. Front Genet 2021;12:651485. https://doi.org/10.3389/fgene.2021.651485.

[59] Bazin J, Baerenfaller K, Gosai SJ, Gregory BD, Crespi M, Bailey-Serres J. Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. Proc Natl Acad Sci USA 2017;114:E10018–27. https://doi.org/10.1073/pnas.1708433114.

[60] Kesner JS, Chen Z, Shi P, Aparicio AO, Murphy MR, Guo Y, et al. Noncoding translation mitigation. Nature 2023;617:395–402. https://doi.org/10.1038/s41586-023-05946-4.