OXFORD

RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization

Guo-Hua Yuan (iD)†, Ying Wang (iD)†, Guang-Zhong Wang (iD) and Li Yang (iD)

Corresponding author. Li Yang, Center for Molecular Medicine, Children's Hospital, Fudan University and Shanghai Key Laboratory of Medical Epigenetics, International Laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Institutes of Biomedical Sciences, Fudan University, Dong-An Road, 131, Shanghai, China. Tel: +86-021-54237325; E-mail: liyang_fudan@fudan.edu.cn.

†Guo-Hua Yuan and Ying Wang contributed equally to this work

Abstract

Different RNAs have distinct subcellular localizations. However, nucleotide features that determine these distinct distributions of lncRNAs and mRNAs have yet to be fully addressed. Here, we develop RNAlight, a machine learning model based on LightGBM, to identify nucleotide *k*-mers contributing to the subcellular localizations of mRNAs and lncRNAs. With the Tree SHAP algorithm, RNAlight extracts nucleotide features for cytoplasmic or nuclear localization of RNAs, indicating the sequence basis for distinct RNA subcellular localizations. By assembling *k*-mers to sequence features and subsequently mapping to known RBP-associated motifs, different types of sequence features and their associated RBPs were additionally uncovered for lncRNAs and mRNAs with distinct subcellular localizations. Finally, we extended RNAlight to precisely predict the subcellular localizations of other types of RNAs, including snRNAs, snoRNAs and different circular RNA transcripts, suggesting the generality of using RNAlight for RNA subcellular localization prediction.

Keywords: RNA localization, machine learning, nucleotide feature, motif, RNA binding protein, circular RNA

Introduction

RNA localization is closely related to its biogenesis, processing and function, which also determines cell fate and polarity [1–3]. In general, most messenger RNAs (mRNAs) transcribed from protein-coding gene loci are usually processed with a series of co- and/or post- transcriptional regulation, including but not limited to 5′-cap, splicing, editing/modification and 3′-adenylation, and transported from nucleus to cytoplasm for protein translation [4, 5]. Instead, many well-studied long non-coding RNAs (lncRNAs) with the length of more than 200 nucleotides tend to be located in nucleus to regulate gene expression by associating with chromatin [6]. Nevertheless, a set of mRNA transcripts can be temporarily retained in nucleus, possibly due to the existence of inverted repeated elements in their 3′ untranslated regions (3′ UTR), by which the translation of specific proteins is retarded [7–10]. Interestingly, some lncRNAs can be exported to the cytoplasm to regulate protein translation by associating with miRNA or ribosome [11, 12]. Thus, understanding RNA molecules' subcellular localizations is important to their functional study.

A variety of approaches has been applied to study RNA subcellular localization. RNA fluorescence *in situ* hybridization (FISH) can accurately examine RNA subcellular localization in a single-RNA resolution and in living cells [13–15]. In addition, cytoplasmic and nuclear RNAs can be biochemically separated into different proportions and further examined by RT-PCR for individual RNAs or by high-throughput methods for various RNA species on a genome-wide scale. For example, CeFra-seq [16] extracted cell fractions of cytosol, insolubles, membrane and nucleus for high-throughput sequencing to identify RNA localization in these diverse cell fractions. Moreover, APEX-Seq [17], which is also an RNA-sequencing based method to examine direct proximity labeling of RNA using the peroxidase enzyme APEX2, revealed extensive patterns of localization for diverse RNA classes in distinct subcellular locales. Furthermore, by collecting the subcellular localization information of thousands of RNAs across different cell lines and species, multiple databases, such as LncATLAS [18] and RNALocate [19], have been constructed to summarize RNA subcellular localization on a genome-wide scale. These datasets not only provide information of individual RNA subcellular localization, but also render a foundation for the prediction of RNA subcellular location *in silico*. In this case, several machine learning and deep learning methods, including mRNALoc [20] and DeepLncRNA [21], have been established to predict RNA subcellular localization. Specifically, the mRNALoc

model examined cytoplasmic and nuclear localizations of mRNAs by support vector machine (SVM) based on sequence information [20], while the DeepLncRNA pipeline [21] trained a deep learning framework to predict nucleus to cytoplasm ratios of lncRNAs. Although these models have been successfully used to predict the localization of RNA based on sequence information, it remained unclear what kinds of key sequence features may contribute to the distinct localization of different types of RNAs. Meanwhile, since these reported methods were applicable to the prediction for just one specific type of RNA, mRNA or lncRNA [20–24], a universal model to perform subcellular localization prediction for different types of RNAs has been lacking.

In this study, we developed a machine learning model, RNAlight, which is based on Light Gradient Boosting Machine (LightGBM) [25], to simultaneously predict the subcellular localizations of both mRNA and lncRNA by using k-mer frequencies as input. With the integration of SHapley Additive exPlanations (Tree SHAP) [26] and k-mer assembly in RNAlight, different sequence features and their associated RNA binding proteins (RBPs) that contribute to distinct subcellular localizations of mRNA or lncRNA were further revealed. With RNAlight, subcellular localizations of various types of noncoding RNAs, including snRNAs, snoRNAs and circular RNAs, were also predicted in line with their reported functions, suggesting its general application in the study of RNA localization and function.

Results
Training RNAlight model to predict RNA subcellular localization

To train models for the prediction of RNA subcellular localizations, three published datasets, including CeFra-seq [16], APEX-Seq [17] and LncATLAS [18] (Figure 1A, Supplementary Figure S1A and B), were collected to construct a combined RNA subcellular localization library for both lncRNAs ($n = 4623$) and mRNAs ($n = 6245$) with GENCODE v30 annotation file (http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_30/gencode.v30.annotation.gtf.gz). Of note, only two major subcellular localizations, nucleus and cytoplasm, were used for model training and prediction in this study. Among four fractions in the CeFra-seq dataset, RNAs in the cytosol, insoluble and membrane fractions from the general cytoplasmic extract were all considered as cytoplasmic localization, and ones in the nuclear fraction were considered as nuclear localization. Among eight classifications in the APEX-Seq dataset, cytoplasm, ER membrane, ER lumen and outer mitochondrial membrane classifications were considered as the cytoplasm classification, while nucleus, nucleolus, lamina and nuclear pore classifications were considered as the nucleus classification. LncATLAS dataset classified lncRNAs with nuclear or cytoplasmic subcellular localization. Since experimentally examined in different conditions, such as various cell lines/methods, some RNAs were shown inconsistent with multiple localizations across these publicly available datasets, and were removed from further processing. After this filter step, about 3792 lncRNAs (1986 nuclear and 1806 cytoplasmic lncRNAs, Supplementary Table S1) and 5180 mRNAs (2256 nuclear and cytoplasmic 2924 mRNAs, Supplementary Table S2) with consistent and single localization across different datasets were combined, and randomly split into training sets (Training-lncRNA, $n = 3412$; Training-mRNA, $n = 4662$) and test sets (Test-lncRNA, $n = 380$; Test-mRNA, $n = 518$) with a 9:1 ratio for model construction and evaluation, respectively (Supplementary Figure S1A and B) ('MATERIALS AND METHODS' section).

Next, we constructed a series of machine learning and deep learning models for the prediction of RNA subcellular localization (Figure 1A). These include three machine learning models, such as canonical support vector machine (SVM), logistic regression and an RNAlight model based on the LightGBM framework that uses tree-based learning algorithms by Microsoft [25], and three deep learning models, such as convolutional neural network (CNN), recurrent neural network (RNN) and a hybrid of CNN and RNN (CNN + RNN) (Figure 1A and Supplementary Figure S2A).

Specifically, we adopted corresponding featurization methods for machine learning [27] and deep learning [28] models, respectively. For machine learning models, RNA sequences were converted to the k-mer (k equals to 3, 4 or 5) frequency matrix as input features (Figure 1A, 'MATERIALS AND METHODS' section). For deep learning models, given that the classic CNN model only accepts the fixed-length input that is connected to the fully connected layer for classification or regression tasks [29–31], each RNA sequence was processed to a fixed length (lncRNA, 4000 nt; mRNA, 9000 nt) by padding or truncating, and then converted to the tensor by one-hot encoding as input ('MATERIALS AND METHODS' section).

After training with the same sets ('MATERIALS AND METHODS' section), we compared their performances and found that RNAlight showed the best performance in the prediction of RNA (both lncRNA and mRNA) localization with cross-validation, indicated by area under the receiver operating characteristic curve (AUROC) (Figure 1B) and other performance metrics (Supplementary Table S3). Consistently, when evaluated with test sets, RNAlight also outperformed other models with the AUROC values as 0.78 and 0.80 for predicting lncRNA or mRNA subcellular localization, respectively (Figure 1C, Tables 1 and 2). Of note, all the deep learning models in our study have relatively poor performances comparing to machine learning models (Figure 1B and C). To test whether the strategy of featurization caused relatively poor performances, we next evaluated these deep learning models with padding RNA sequences from the 5′ end (Supplementary Figure S2B and C), truncating 5′ end of RNA sequences (Supplementary Figure S2D and E) or using the word2vec method (Supplementary Figure S3A). However, these alternative strategies of featurization didn't improve performances significantly (Supplementary Figure S2B and C, Supplementary Figure S3C and D), while increased time and memory consumption (Supplementary Figure S3B).

To further evaluate the performance of RNAlight, we used test sets (Test-lncRNA, $n = 380$; Test-mRNA, $n = 518$) to compare RNAlight with four previously reported prediction models, including iLoc-lncRNA [23] and lncLocator [22] for lncRNA localization or iLoc-mRNA [24] and mRNALoc [20] for mRNA localization. As illustrated by the confusion matrix in Figure 2A, RNAlight achieved a more accurate prediction for lncRNA localization than iLoc-lncRNA and lncLocator did, with the highest accuracy, F1 score (Figure 2B) and AUROC (Figure 2C). Similarly, RNAlight also outperformed other compared models in mRNA localization prediction (Figure 2D-F). In addition, when evaluating with a totally independent dataset of lncRNAs ($n = 116$, Supplementary Figure S4A) and mRNAs ($n = 809$, Supplementary Figure S4E) from Halo-seq in Hela cell line (Supplementary Table S4, 'MATERIALS AND METHODS' section) [32], RNAlight also showed generally better performance than other models on both lncRNA (Supplementary Figure S4B-D) and mRNA (Supplementary Figure S4F-H) subcellular localization prediction.

These results together suggested that the LightGBM-based RNAlight model could accurately predict the different subcellular localizations (nucleus and cytoplasm) for both lncRNAs and mRNAs, which was consistent with the superior performance of

Table 1. Evaluation of prediction models for lncRNA subcellular localization by using the lncRNA test set (Test-lncRNA, $n = 380$)

Model	Accuracy	Sensitivity	Specificity	MCC	F1 score	AUROC
RNAlight	0.72	0.76	0.68	0.45	0.74	0.78
SVM	0.69	0.77	0.60	0.37	0.72	0.75
LR	0.70	0.78	0.61	0.40	0.73	0.76
CNN	0.64	0.64	0.64	0.28	0.65	0.72
RNN	0.59	0.48	0.71	0.19	0.55	0.63
CNN + RNN	0.65	0.65	0.65	0.31	0.66	0.71

Table 2. Evaluation of prediction models for mRNA subcellular localization by using the mRNA test set (Test-mRNA, $n = 518$)

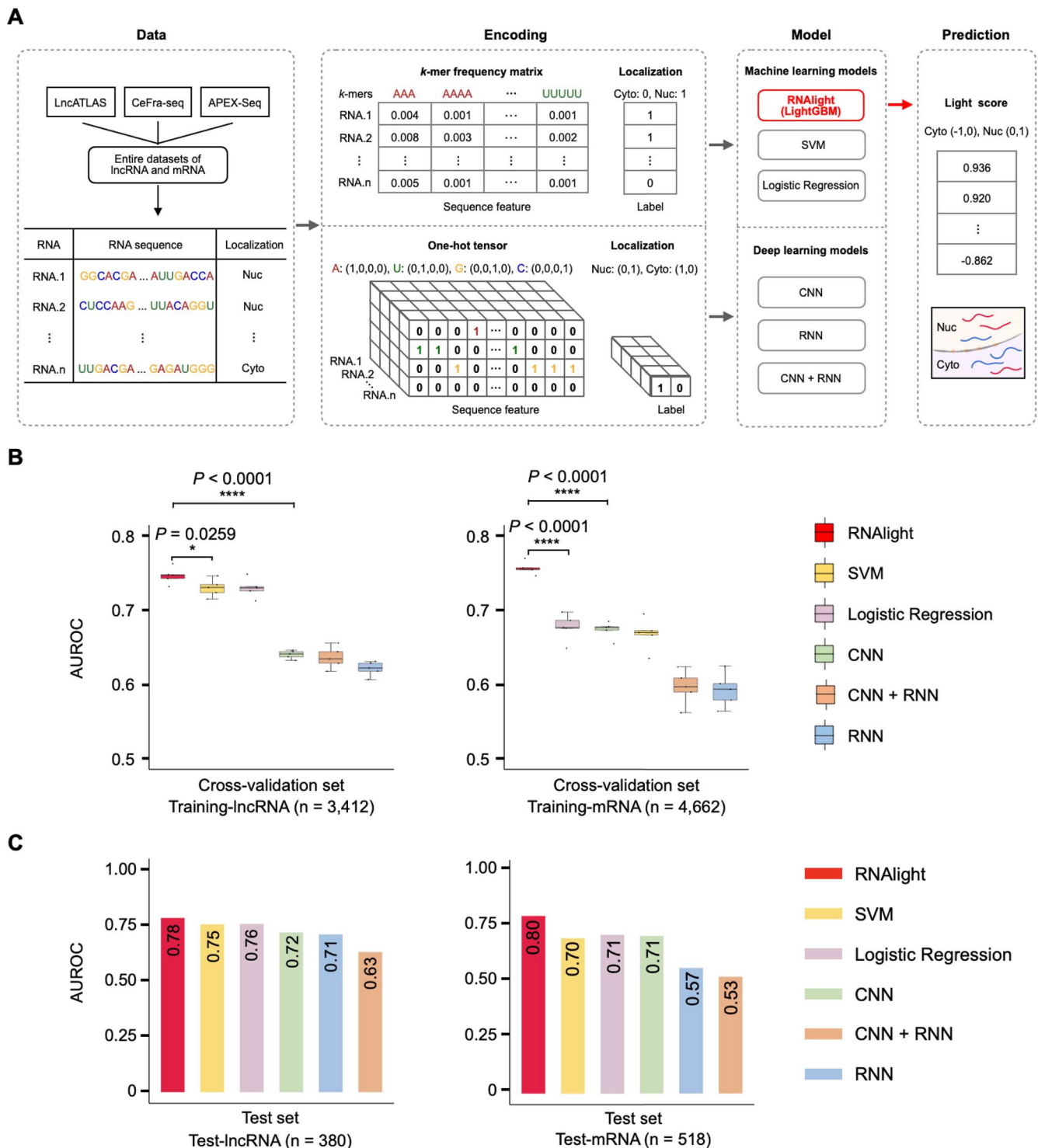| Model	Accuracy	Sensitivity	Specificity	MCC	F1 score	AUROC
RNAlight	0.73	0.59	0.84	0.45	0.66	0.80
SVM	0.64	0.37	0.86	0.26	0.48	0.70
LR	0.66	0.44	0.83	0.30	0.53	0.71
CNN	0.66	0.51	0.78	0.30	0.57	0.71
RNN	0.55	0.02	0.98	−0.02	0.03	0.53
CNN + RNN	0.58	0.07	0.98	0.13	0.12	0.57

LightGBM across a series of benchmark tests in previous studies [25, 33, 34].

Identifying distinct sequence features for lncRNA or mRNA subcellular localization with Tree SHAP algorithm

As k-mers were used as inputs for RNAlight analysis, we next attempted to identify which k-mers (sequence features) might play important roles in different (nucleus or cytoplasm) subcellular localizations of lncRNAs and/or mRNAs. In addition to the k-mer frequencies (Figure 3A, left), we also used Tree SHapley Additive exPlanations (SHAP) [26] in the LightGBM algorithm for the contribution analysis of k-mers in determining RNA nuclear or cytoplasmic localization. In theory, a positive or negative SHAP value suggests a potential role of a given k-mer on the nuclear or cytoplasmic localization of a given examined RNA, respectively ('MATERIALS AND METHODS' section). For each k-mer, different SHAP values could be quantified to show its distinct contributions on subcellular localization of different examined RNAs (Figure 3A, right). To better access the general effect of each k-mer on nuclear or cytoplasmic localization of all examined RNAs, Pearson correlation coefficients (PCCs) between k-mer frequencies and SHAP values were calculated (Figure 3A, bottom). In general, a positive PCC represents an overall nuclear localization effect of a given k-mer on analyzed lncRNAs or mRNAs, and a negative PCC represents an overall cytoplasmic localization effect of another given k-mer on analyzed lncRNAs or mRNAs ('MATERIALS AND METHODS' section). With PCC > 0.5 as cutoff, 399 k-mers were determined for nuclear localization of lncRNAs; meanwhile, 501 cytoplasm-related k-mers for cytoplasmic localization of lncRNAs were selected by PCC < −0.5 (Figure 3B and Supplementary Table S5). Importantly, several nuclear localization-related sequence elements of lncRNAs that have been previously identified by a high-throughput screening of short RNA fragments [10, 35] were successfully predicted by RNAlight (Figure 3B). Specifically, a cluster of five-mers, such as CUCCC, CCUCC and ACCUC, were identified with positive PCCs (0.726, 0.727 and 0.537, respectively) by RNAlight (Figure 3B). These five-mers can be tiled across the RCCUCCC motif (where R denotes A/G), which has been previously confirmed associated with lncRNA nuclear localization [10], suggesting the reliable prediction of RNA subcellular localization by RNAlight.

Given the fact that a spectrum of variable SHAP values could be determined for each particular k-mer, we then calculated a Z-transformed mean absolute SHAP value for each k-mer and used Z-score > 1.96 as an additional cutoff to identify most important k-mers for RNA subcellular localization among all examined lncRNAs or mRNAs. As shown in Figure 3C and Supplementary Figure S5A and S5B, 20 out of 399 nucleus-related and 43 out of 501 cytoplasm-related k-mers were individually identified to play key roles in determining lncRNA nuclear or cytoplasmic localization with Z-score > 1.96 (Supplementary Table S5). Of note, Z-scores of mean absolute SHAP values of aforementioned five-mers (CUCCC, CCUCC and ACCUC) were < 1.96, possibly due to their limited distribution among a small cluster of lncRNAs [10] but not in thousands of lncRNAs examined in the current study.

It is well known that the interaction of RNA sequences and their associated RBPs is of importance for RNA subcellular localization [10, 36]. We thus aimed to find what types of RBPs could individually bind to these different k-mers for distinct RNA subcellular localization. To achieve this goal, we assembled nucleus-related or cytoplasm-related k-mers individually to different sequence features groups [27]. Briefly, important localization-related k-mers were first mapped back to the RNA sequences. Neighboring k-mers were then ligated as candidate sequence features. Consensus sequence features were further obtained by the multiple sequence alignment based on these candidate sequence features (Supplementary Figure S6). From 20 nucleus-related k-mers identified by PCC > 0.5 and Z-score > 1.96 (Figure 3B), 190 sequence features were obtained by k-mer assembling (Supplementary Table S6); from 34 cytoplasm-related k-mers identified by PCC < −0.5 and Z-score > 1.96 (Figure 3B), eight sequence features were obtained by k-mer assembling (Supplementary Table S6). After that, we used Tomtom [37] ('MATERIALS AND METHODS' section) to map these assembled sequence features to known RBP-associated motifs reported in the CISBP-RNA (Catalog of Inferred Sequence Binding Preferences of RNA binding proteins) database [38] (Supplementary Figure S6). As a result, 27 out of 190 nucleus-related sequence features were identified to be associated with 18 RBPs for lncRNA nuclear localization (Supplementary Table S7). For example, NONO (non-POU domain containing octamer binding), a well-studied RBP that preferentially binds RNAs with AGGGA/U elements [39]

Figure 1. Computational models to predict RNA subcellular localization. (**A**) Schematic diagram of computational models for RNA subcellular localization prediction. (**B**) Cross-validation of different models for lncRNA (left) and mRNA (right) subcellular localization prediction. Each dot represents the area under the receiver operating characteristic curve (AUROC) from five-fold cross-validation (total AUROC, $n = 5$). Statistical testing was performed with one-sided Welch's t-test. In the box plots, the 25th, 50th and 75th percentiles are indicated as the top, middle and bottom lines, respectively; whiskers represent the 10th and 90th percentiles, respectively. (**C**) Evaluation of prediction models for lncRNA (left) and mRNA (right) subcellular localization by using the test sets.

and participates in paraspeckle formation through binding with nuclear *NEAT1* lncRNA [40], was identified to be associated with lncRNA nuclear localization in this study (Figure 3D). Instead, only two out of eight cytoplasm-related sequence features were identified to be individually associated with two different RBPs

for lncRNA cytoplasmic localization (Supplementary Table S7). Between these two RBPs, PCBP2 (poly(rC) binding protein 2), one major cellular poly(rC)-binding protein that is able to bind with C-rich sequences [38] and cooperates with LINC02535 to enhance the stability of mRNA in the cytoplasm [41], was predicted
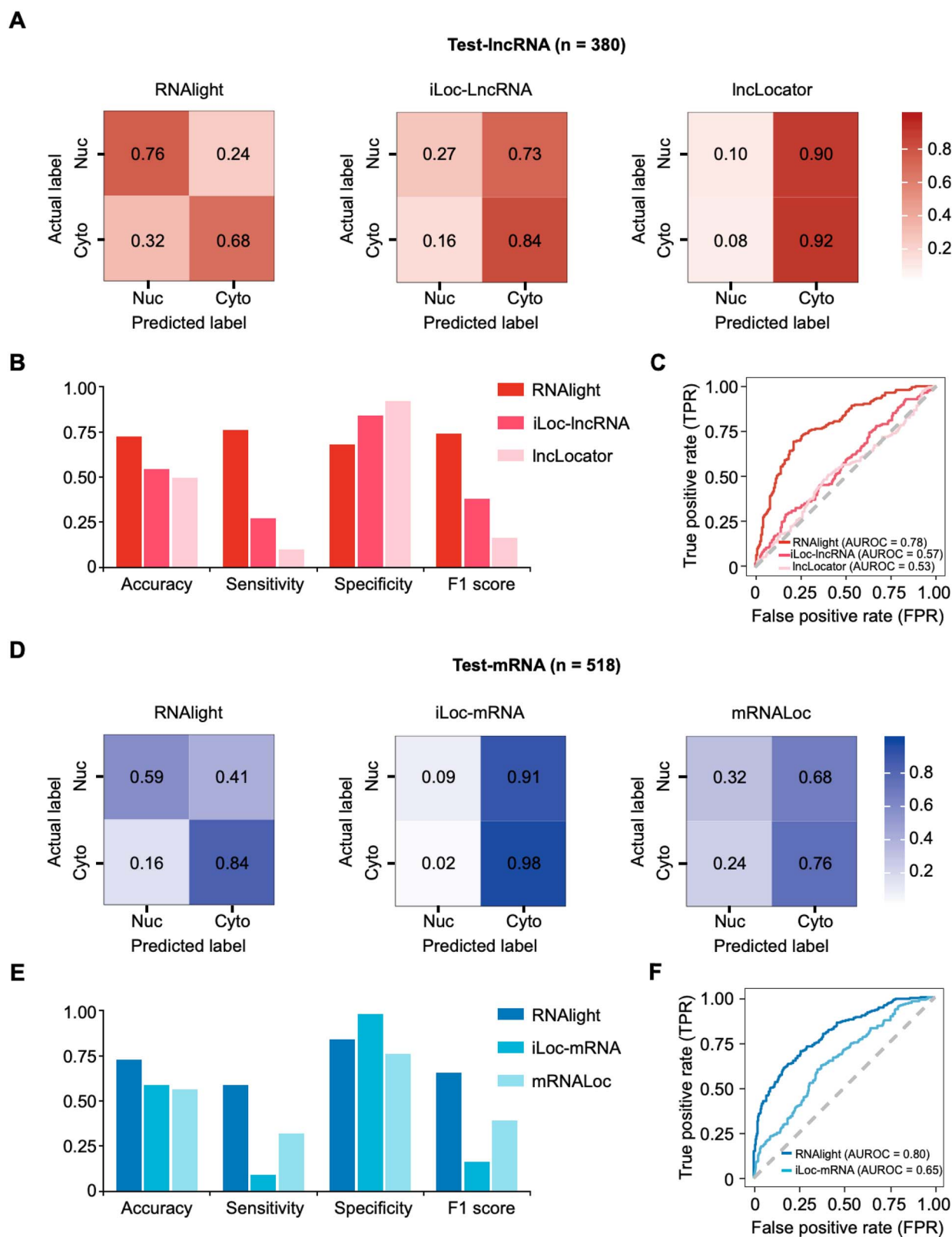
Figure 2. Comparative evaluation of RNAlight and other published models for RNA subcellular localization prediction. The confusion matrix (**A**), accuracy, sensitivity, specificity, F1 score (**B**) and AUROC curve (**C**) of RNAlight, iLoc-LncRNA and lncLocator on predicting lncRNA localization with the lncRNA test set (Test-lncRNA, $n = 380$). The confusion matrix (**D**), accuracy, sensitivity, specificity and F1 score (**E**) of RNAlight, iLoc-mRNA and mRNALoc on predicting mRNA localization with the mRNA test set (Test-mRNA, $n = 518$). (**F**) AUROC curves of RNAlight and iLoc-mRNA on predicting mRNA localization with the mRNA test set (Test-mRNA, $n = 518$).

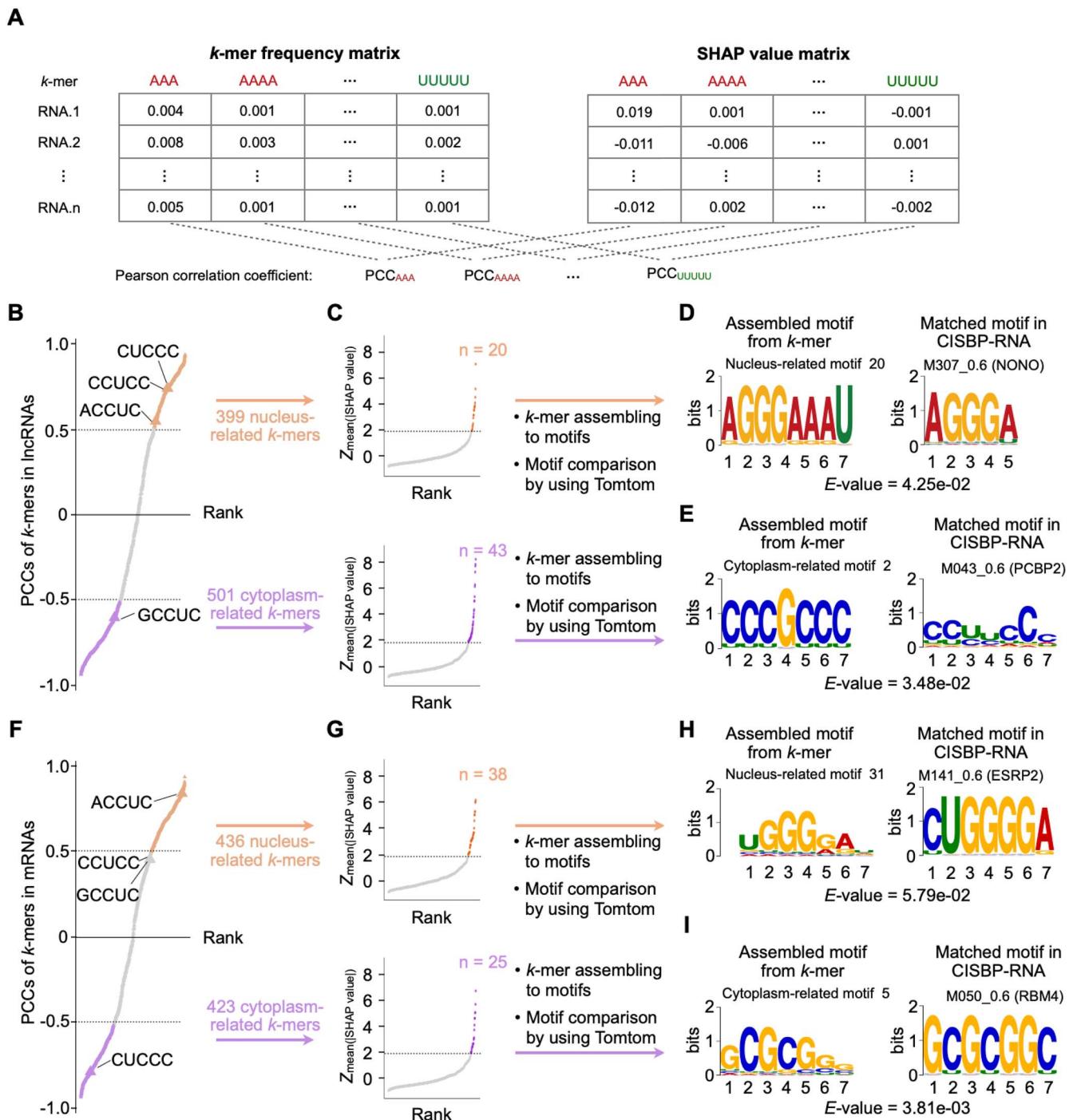**Figure 3.** Identification of localization-associated RNA binding proteins for lncRNAs and mRNAs. (**A**) Schematic drawing of PCC (Pearson correlation coefficient) calculation for each given *k*-mer. (**B**) Identification of nucleus- and cytoplasm-related *k*-mers for lncRNAs. The scatter plot shows PCCs of all 1344 *k*-mers in lncRNAs. (**C**) The scatter plots show Z-transformed mean absolute SHAP values of the kmers with PCC > 0.5 (top) or PCC < −0.5 (bottom). (**D**-**E**) Examples of *k*-mer assembled motifs (left) comparing with known RBP-associated motifs (right) for lncRNAs. (**F**) Identification of nucleus- and cytoplasm-related *k*-mers for mRNAs. The scatter plot shows PCCs of all 1344 *k*-mers in mRNAs. (**G**) The scatter plots show *Z*-transformed mean absolute SHAP values of the *k*-mers with PCC > 0.5 (top) or PCC < −0.5 (bottom). (**H**-**I**) Examples of *k*-mer assembled motifs (left) comparing with known RBP-associated motifs (right) for mRNAs.

to be involved in regulating lncRNA cytoplasmic localization (Figure 3E).

Similar analyses were parallelly performed for mRNA subcellular localization (Figure 3F-I). With PCC > 0.5 as cutoff, 436 *k*-mers were determined for nuclear localization of mRNAs, and 423 cytoplasm-related *k*-mers for cytoplasmic localization of mRNAs were selected by PCC < −0.5 (Figure 3F and

Supplementary Table S8). In addition, 38 out of 436 nucleus-related and 25 out of 423 cytoplasm-related *k*-mers were individually identified to play key roles in determining mRNA nuclear or cytoplasmic localization with Z-score > 1.96 (Figure 3G, Supplementary Figure S5C and D, Supplementary Table S8). Of note, no overlap was observed between 20 of important nucleus-related *k*-mers for lncRNAs and 38 of those for mRNAs

(Supplementary Figure S5E, left), and only one cytoplasm-related *k*-mer was observed between 43 of important cytoplasm-related *k*-mers for lncRNAs and 25 of those for mRNAs (Supplementary Figure S5E, right). This result (Supplementary Figure S5E) thus suggested distinct *cis*-element features contributing to lncRNA and mRNA subcellular localizations.

In the analysis of identifying what types of RBP could individually bind to different *k*-mers for distinct mRNA subcellular localization, 1223 sequence features were obtained by *k*-mer assembling (Supplementary Table S9) from 38 nucleus-related *k*-mers identified by PCC > 0.5 and Z-score > 1.96 (Figure 3G), and 235 sequence features were obtained by *k*-mer assembling (Supplementary Table S9) from 25 cytoplasm-related *k*-mers identified by PCC < −0.5 and Z-score > 1.96 (Figure 3G). After mapping to known RBP-associated motifs by Tomtom, 262 out of 1223 nucleus-related sequence features were identified to be associated with 54 RBPs for mRNA nuclear localization (Supplementary Table S10). For example, ESRP2 (epithelial splicing regulatory protein 2), an epithelial cell-type-specific splicing regulator which was mainly located in nucleus [42] and preferentially binds to RNA with UGGGRAD motif [38], was identified to be linked with the regulation of mRNA nuclear localization (Figure 3H). For the cytoplasmic localization of mRNAs, 67 out of 235 cytoplasm-related sequence features were identified to be associated with 16 RBPs (Supplementary Table S10), including RBM4 (RNA binding motif protein 4) (Figure 3I), an RNA-binding factor involved in mRNA splicing and translation regulation [43] with a tendency to bind with GC-rich sequences [38].

Applying RNAlight to accurately predict subcellular localizations of various types of RNAs

In contrast to previous models, RNAlight was designed to examine subcellular localization of both mRNAs and lncRNAs. As expected, RNAlight showed a preference of cytoplasmic localization for mRNA transcripts (*n* = 18 607, GENCODE v30) (the median of Light score = −0.253, Figure 4A), as most mature mRNAs are preferentially transported to the cytoplasm for protein translation [44]. However, a few mRNAs, such as *MLXIPL* and *NLRP6*, were predicted to be the nuclear localization with Light scores of 0.827 and 0.847, respectively (Supplementary Table S11), which are in line with their subcellular localizations in nuclear speckles [9]. Differently, a bimodal distribution with a slightly nuclear preference (the median of Light score = 0.037) of annotated lncRNA transcripts (*n* = 16 153) was predicted by RNAlight (Figure 4A), suggesting their regulatory roles at multiple components of cells [13], despite a greater proportion of lncRNAs were shown with a nucleus tendency [15, 45]. Accordingly, a set of known nuclear-localized lncRNAs, such as *MALAT1*, *NEAT1* and *XIST* [40, 46, 47], were predicted by RNAlight with the Light scores of 0.671, 0.899 and 0.854, respectively (Supplementary Table S11). Meanwhile, some known cytoplasmic lncRNAs, such as *ZFAS1* and *SNHG6* [48, 49], were successfully predicted as cytoplasmic localization with Light scores of −0.949 and − 0.913 (Supplementary Table S11).

To further evaluate the generality of RNAlight for the subcellular localization prediction on different types of RNAs, we extended the analysis to other RNA species that were not used for the model training, including small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and circular RNAs. It is well known that snoRNAs are distributed in eukaryotic nucleolus for rRNA/tRNA methylation or other RNA modification [51], and snRNAs are localized in the nucleoli and nucleolus as main components of spliceosome [52]. Correspondingly, RNAlight

accurately predicted their nuclear localization of snoRNAs (*n* = 553, the median of Light score = 0.277) and snRNAs (*n* = 1825, the median of Light score = 0.247) (Figure 4A). In addition, two major types of spliceosome-dependent circular RNAs, circRNAs from back-spliced exons and circular intronic RNAs (ciRNAs) from spliced intron lariats, were recently rediscovered at a genome-wide level in eukaryotes but with different subcellular localization [53,54]. It has been reported that circRNAs were generally localized in cytoplasm [55], involving in innate immunity [56,57], cell proliferation [58] and neuronal function [59], while ciRNAs were preferentially retained in nucleus to regulate Pol II transcription [60,61]. By using highly expressed circRNAs (FPB > 0.5, *n* = 800) and ciRNAs (FPB > 0.2, *n* = 104) identified by the CLEAR pipeline [62] in PA1 cells from published ribo–RNA-seq (GEO: GSE73325) [50] ('MATERIALS AND METHODS' section) as inputs, we set up to examine whether RNAlight could be further extended for circular RNA subcellular localization prediction. As shown in Figure 4B, RNAlight successfully predicted the cytoplasmic localization of circRNAs (the median of Light score = −0.274) and the nuclear localization of ciRNA (the median of Light score = 0.644). In contrast, other RNA subcellular localization prediction tools, such as iLoc-lncRNA, lncLocator, iLoc-mRNA and mRNALoc, failed to accurately reveal the subcellular localization of cytoplasmic circRNAs and nuclear ciRNAs (Supplementary Figure S7). These results together indicated RNAlight as a reliable and universal model for the subcellular localization prediction of distinct RNA species.

Discussion

Here, we report RNAlight, a machine learning model based on LightGBM for precise RNA localization prediction (Figures 1 and 2). In our hands, all examined machine learning models (RNAlight, SVM and logistic regression) outperformed deep learning models (CNN, RNN and CNN + RNN) in the prediction of RNA subcellular localization. Applying different strategies of featurization for deep learning models did not significantly improve their performances (Supplementary Figures S2 and S3). The poor performance by deep learning models is possibly due to the relatively small scale of the dataset for model development. In the future, we assume that better deep learning models can be developed with fast-accumulated large-scale datasets. Nevertheless, by integrating the Tree SHAP algorithm and *k*-mer assembly into the RNAlight model, determinant sequence features and their associated RBPs were identified to contribute to distinct mRNA or lncRNA subcellular localizations (Figure 3). Importantly, RNAlight was also shown to be a reliable model for the subcellular localization prediction of various RNAs, including mRNAs, lncRNAs, snoRNAs, snRNAs, circRNAs and ciRNAs (Figure 4), suggesting its broad application for the localization prediction of various types of RNAs.

Compared to some other methods [20–24], RNAlight has two characteristics for the prediction of RNA subcellular localization. On the one hand, by integrating the Tree SHAP algorithm and *k*-mer assembly, RNAlight not only predicts RNA subcellular localization, but also effectively identifies distinct sequence features for their different subcellular (nuclear or cytoplasmic) localization. Similar to previous studies showing that certain *cis*-elements in the 5′ end or 3′ end of RNAs contribute to their subcellular localization [8, 63–65], many *k*-mers identified by RNAlight in this study were also shown to be enriched in 5′ or 3′ ends of their host RNAs (Supplementary Figures S8–S11), indicating that RNAlight could indeed learn important sequence
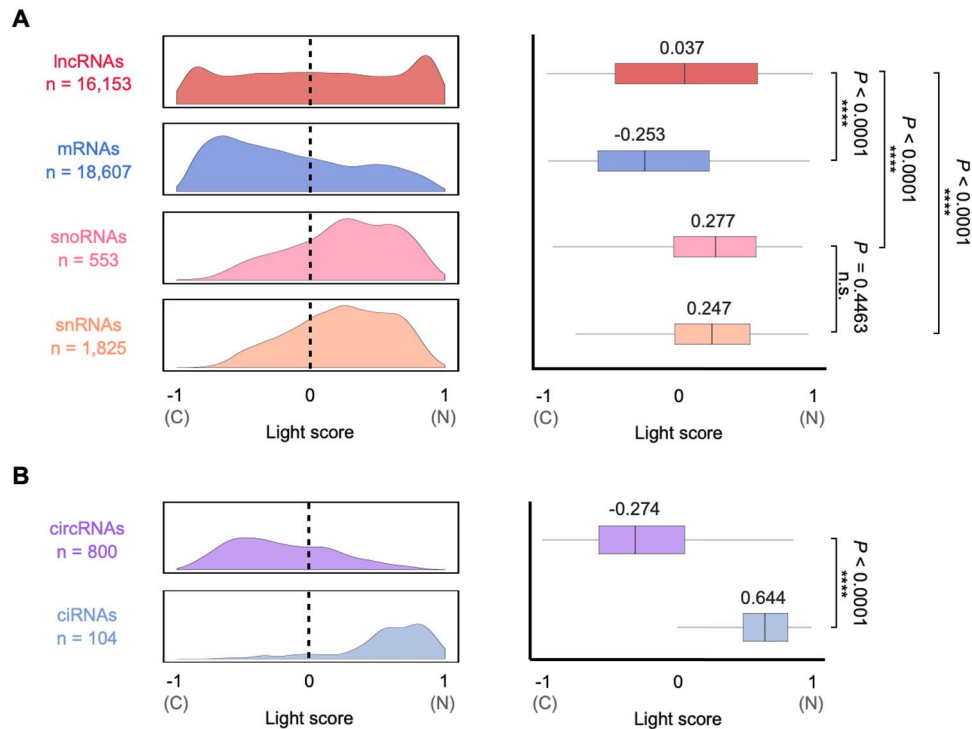
Figure 4. Application of RNAlight to predict subcellular localizations of various types of RNAs. Density (left) and box (right) charts show the distributions of Light scores reported by RNAlight across various types of RNAs, including major transcripts of lncRNA, mRNA, snoRNA and snRNA from GENCODE v30 annotation (**A**), and circRNAs and ciRNAs identified in PA1 cells from Zhang et al [50] (**B**). The range of Light score is from −1 to 1, wherein the interval of −1 to 0 or 0 to 1 indicates cytoplasmic or nuclear localization of RNA, respectively. Statistical testing was performed with two-sided Welch's *t*-test; n.s., not significant.

features that determine RNA subcellular localization. Further analyses showed that these sequence features could be associated with distinct RBPs to affect RNA subcellular localization (Figure 3). On the other hand, RNAlight has been also extended to predict subcellular localization of various RNA species, with reliable results that are consistent with previous observations from experimental examination (Figure 4), suggesting the robustness of RNAlight for the prediction of RNA subcellular localization. Interestingly, RNAlight could successfully differentiate RNA circles with exon or intron origins in their distinct subcellular localizations. Different to cytoplasmic localization of most circRNAs from back-spliced exons, ciRNAs that are produced from spliced intron lariats were predicted to be preferentially located in nucleus (Figure 4B), consistent with experimental lines of evidence [60,61]. Similarly, RNAlight showed that the inclusion of intron sequences in lncRNA, mRNA and circRNA could substantially change their subcellular localization from cytoplasm to nucleus (Supplementary Figure S12), in line with the previous study that RNA transcripts with retained introns are considered as incompletely spliced forms and generally retained in nucleus [66].

Despite of these advantages, RNAlight might also simplify the prediction of RNA subcellular localization by only inputting features from RNA primary sequences. However, when adding additional features of predicted RNA secondary structure and the compositional information, the performances of both machine learning and deep learning models were not significantly improved (Supplementary Figure S13, 'MATERIALS AND METHODS' section). Since both RNA secondary structure and compositional information were predicted or obtained from RNA primary sequences, we speculated that sequence features

might be sufficient to achieve RNA subcellular localization prediction in this scenario. In the future, experimental lines of evidence on RNA secondary structure and modification could be further considered to improve this model.

In addition, given that our study aims to identify general sequence features associated with RNA subcellular localization, RNAlight was trained with RNAs only showing single subcellular localization as many other prediction models did [20, 22–24]. It is thus a common limitation of most existing computational tools for RNA subcellular localization prediction. Nevertheless, we did not rule out the possibility that RNAs with distinct subcellular localizations across different cell types or states are functionally important. Identification of other factors, such as associated partners (RBPs, U1 snRNA, etc.) and/or cellular contexts in various conditions, is warranted to understand how distinct subcellular localizations could be regulated for a given RNA. Finally, RNAlight only focused on predicting two major RNA subcellular localizations, nucleus and cytoplasm, due to the limited datasets. We expected that detailed RNA subcellular localization could be dissected with more datasets containing high-resolution RNA localization information, such as ER, mitochondria or nucleolus.

Taken together, we reported the RNAlight model based on LightGBM to precisely predict nuclear or cytoplasmic localization of mRNAs and lncRNAs, and further identified important *k*-mer features and RBPs possibly involved in their subcellular localizations. In the future, additional datasets with extra features, including but not limited to RNA secondary structure and/or chemical modification, can be included to train a better model for characterizing complex and dynamic RNA subcellular localization.

Materials and methods
A combined library of lncRNAs and mRNAs labeled with distinct nuclear or cytoplasmic localizations

To generate a universal model for both lncRNA and mRNA subcellular localization prediction, we first collected reported datasets with lncRNA and/or mRNA subcellular localization information, and further combined them together for a combined library of lncRNA and mRNA subcellular localization.

On the one hand, several lncRNA subcellular localization datasets, including LncATLAS [18], CeFra-seq [16] and APEX-Seq [17], were collected for this study (Supplementary Figure S1A). Due to distinct methodologies in these datasets for subcellular localization analysis, different filtering strategies were then implemented in this study to select nuclear or cytoplasmic lncRNAs: (i) LncATLAS database records localization information of 6768 lncRNAs across 14 cell lines [18]. Here, the mean value of cytoplasmic/nuclear concentration index (CN–RCI_{mean}) across 13 cell lines (excluding H1 cell due to its low correlation to other cell lines, data not shown) was used for filtering: lncRNAs with (CN–RCI_{mean}) < −2 ($n = 1857$) were considered as nuclear localization, while those with (CN–RCI_{mean}) > 0 ($n = 1440$) were considered as cytoplasmic localization. (ii) CeFra-seq extracts cell fractions of cytosol, insoluble, membrane and nucleus for high-throughput sequencing and RNAs localized in these diverse cell fractions could be identified, and 14 746 lncRNAs were detected at these four cell fractions in HepG2 cells [16]. From the CeFra-seq dataset, 1621 highly expressed lncRNAs with fragments per kilobase per mapped fragments (FPKM) ≥ 1 in at least one cellular fraction were obtained for subsequent analysis. Accordingly, cytoplasmic ratio (CR) was used to distinguish nuclear and cytoplasmic lncRNAs in the CeFra-seq dataset, computed as below:

$$CR = \frac{Cyto_{FPKM}}{Cyto_{FPKM} + Nuc_{FPKM}}$$

Here, Nuc_{FPKM} is the FPKM value of an lncRNA in the nuclear fraction and $Cyto_{FPKM}$ is the maximum FPKM value of the lncRNA in the cytosol, insolubles or membrane fraction. With these criteria, 435 lncRNAs with CR < 0.4 were considered with the preference of nuclear localization, and 844 lncRNAs with CR > 0.6 were considered with the preference of cytoplasmic localization. (iii) APEX-Seq provides a practical methodology to identify RNA in distinct subcellular locales, based on APEX2-mediated proximity biotinylation of endogenous RNAs in the presence of biotin-phenol (BP) and H_2O_2 and following poly(A) + RNA sequencing [17]. RNA localized in one subcellular locale could be identified by calculating the fold change of the H_2O_2-treated sample to the untreated control sample in HEK293T cells. Here, among all lncRNAs identified in the APEX-Seq dataset ($n = 61$), those with log2(fold change) ≥ 0.75 in at least one component of the nucleus, nucleolus, lamina and nuclear pore were considered as nuclear lncRNAs ($n = 42$) and those with log2(fold change) ≥ 0.75 in at least one component of the cytoplasm, ER (endoplasmic reticulum) membrane, ER lumen and outer mitochondrial membrane were considered as cytoplasmic lncRNAs ($n = 5$).

After the aforementioned filtering, lncRNAs from these three resources were combined and lncRNAs with inconsistent but multiple localizations were removed to generate a combined library with 1986 nuclear and 1806 cytoplasmic (totally 3792) lncRNAs.

On the other hand, mRNAs with different subcellular localization were collected from CeFra-seq and APEX-Seq datasets (Supplementary Figure S1B). Similar processing parameters for CeFra-seq and APEX-Seq data were applied to select mRNAs with different subcellular localization. From the CeFra-seq dataset, 1789 and 2040 mRNAs were selected with the preference of nuclear or cytoplasmic localization, respectively. From the APEX-Seq dataset, 1145 and 1261 mRNAs were selected with the preference of nuclear or cytoplasmic localization. Finally, mRNAs with different subcellular localization from these two resources were combined and those with inconsistent but multiple localizations were filtered out, leading to a total of 5180 mRNAs with nuclear (2256) or cytoplasmic (2924) labels.

Collectively, by stringent filtering, we constructed a combined library of lncRNAs ($n = 3792$) and mRNAs ($n = 5180$) labeled with distinct nuclear or cytoplasmic localizations. These RNAs were randomly split into training sets (Training-lncRNA, $n = 3412$; Training-mRNA, $n = 4662$) and test sets (Test-lncRNA, $n = 380$; Test-mRNA, $n = 518$) with a 9:1 ratio for model training and evaluation.

An independent dataset of lncRNAs and mRNAs labeled with distinct nuclear or cytoplasmic localizations from Halo-seq

A totally independent dataset containing lncRNA and mRNA subcellular localization information from Halo-seq in HeLa cell line [32] was collected to further evaluate RNAlight with other published models. With adjusted P-value <0.05 and absolute log_2(fold change) ≥ 0.5 as cutoff, H2B-Halo enriched- and Halo-p65 depleted-RNAs were considered as nuclear localization; while H2B-Halo depleted- and Halo-p65 enriched-RNAs were considered as cytoplasmic localization. After removing redundant and bi-localized ones, 116 lncRNAs and 809 mRNAs with the nuclear or cytoplasmic label were individually obtained (Supplementary Table S4) for model evaluation.

Transcript selection

The files recording the information of RNA localization and Ensembl gene IDs of their parental genes were directly downloaded from public resources [16–18, 32]. For each RNA, we selected the major splice annotation (the −001 isoform) with the GENCODE v30 annotation file as the previously reported method [67] to obtain its primary sequence for subsequent analyses.

Featurization of RNA primary sequences

The strategy of featurization for machine-learning based models was similar to that in the previous publication [27]. Briefly, frequencies of 1344 k-mers (k equals to 3, 4, 5) that permute four nucleotides (A, T/U, G, C) were firstly computed to show their presence in each specific RNA. These frequencies were further normalized by the RNA length and integrated as a k-mer frequency matrix, which can characterize each RNA with 1344 distinct features.

For featurization in deep-learning based models, each RNA sequence was processed to a fixed length (lncRNA, 4000 nt; mRNA, 9000 nt). Specifically, lncRNAs shorter than 4000 nt or mRNAs shorter than 9000 nt in length were padded with 'N' at their 3′ ends, which covered 95% of lncRNAs and mRNAs according to the 95 percentile of the RNA length (lncRNA, 3825 nt; mRNA, 8469 nt). The rest of lncRNAs and mRNAs with longer lengths were truncated to the same length (lncRNA, 4000 nt; mRNA, 9000 nt). After that, all RNAs with the fixed length were converted to tensor by one-hot encoding as input, where each nucleotide was

transformed to a binary vector: A (1, 0, 0, 0), T/U (0,1,0,0), G(0,0,1,0), C(0,0,0,1), N(0,0,0,0).

Featurization of RNA predicted secondary structure and compositional information

Secondary structures of mRNAs and lncRNAs were predicted by Vienna RNAfold [68] with default parameters using RNA primary sequences. The bpRNA algorithm [69] was then used to annotate predicted structures into different types, including stem (S), hairpin loop (H), multi-loop (M), internal loop (I), bulge (B), external loop (X) and end (E). The secondary structure matrix was constructed to contain the lengths and ratios of different structure types and the minimum free energy (MFE) of examined RNAs (Supplementary Figure S13A, middle panel).

For compositional information of RNA, the GC content, AUGC ratio, GC skew and Z-curve of RNA sequence were calculated by following mathematical formulas:

$$GC = \frac{F_G + F_C}{F_A + F_U + F_G + F_C}$$

$$AU/GC = \frac{F_A + F_U}{F_G + F_C}$$

$$GC\ skew = \frac{F_G - F_C}{F_G + F_C}$$

$$Z - curve = \begin{cases} X = (F_A + F_G) - (F_C + F_U) \\ Y = (F_A + F_C) - (F_G + F_U) \\ Z = (F_A + F_U) - (F_G + F_C) \end{cases}$$

where F_x represents the frequency of each nucleotide (A, U, G, C).

The k-mer frequency matrix (representing sequence feature, Supplementary Figure S13A, left panel), the secondary structure matrix (representing structure feature, Supplementary Figure S13A, middle panel) and the composition matrix (representing composition feature, Supplementary Figure S13A, right panel), have been combined as the new input to train the RNAlight model (Supplementary Figure S13A). For comparison, the structure and composition features were also added to train the CNN model with a shared dense layer (Supplementary Figure S13B).

Construction of RNAlight with LightGBM framework

LightGBM is a new machine-learning implementation of gradient enhanced decision tree (GBDT) with gradient-based one-side sampling and exclusive feature building, which has several advantages, including faster training speed, higher efficiency and lower memory usage. Here, we used the LightGBM Python package (version 3.1.1.99) to train the RNAlight model by inputting the k-mer frequency matrix and labeled subcellular localizations from training sets (Training-lncRNA, $n = 3412$; Training-mRNA, $n = 4662$).

Five-fold cross-validation based on RandomizedSearchCV was performed to select the optimal hyperparameters for LightGBM. We searched 1000 combinations chosen from following hyperparameter configurations: learning rate (chosen from [0.1, 0.05, 0.02, 0.01]), the number of estimators (chosen from 24 points that were evenly spaced between 100 and 2400), the maximum depth of the individual estimators (chosen from [2–5, 10, 20, 40, 51]), the minimum number of data in one leaf (chosen from 22 points that were evenly spaced between 1 and 44), the fraction of subset on each estimator (chosen from [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]), the frequency of bagging (chosen from [0, 1, 2]), the penalty of L1 regularization (chosen from [0, 0.001, 0.005, 0.01, 0.1]) and the penalty of L2 regularization (chosen from [0, 0.001, 0.005, 0.01, 0.1]).

Construction of support vector machine model and logistic regression model

Support vector machine (SVM) and logistic regression models were both performed by Python scikit-learn package (version 0.20.3). For SVM, we searched 60 combinations through the following hyperparameter configurations: the kernel type (chosen from ['linear', 'rbf']), the penalty parameter C (chosen from [0.01, 0.1, 1, 10, 100]) and the kernel coefficient γ (chosen from [0.001, 0.005, 0.1, 0.5, 1, 2]). For logistic regression, the penalty of L2 regularization was chosen from [1e-3, 5e-3, 1e-2, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000] to optimize hyperparameters. Five-fold cross-validation based on RandomizedSearchCV was also utilized to select the appropriate set of hyperparameters.

Construction of deep learning-based models

Convolutional neural network (CNN), recurrent neural network (RNN) and a combinatorial model of CNN and RNN (CNN + RNN) were applied to train deep learning-based models under the Tensorflow (version 2.0.0) backend in Python (version 3.6.12).

CNN was developed using two convolutional layers and one dense layer with the following hyperparameters: the number of filters [32, 64, 128] in the first convolutional layer and the number of units [256, 512, 1024] in the dense layer. RNN was developed using one Bidirectional LSTM layer and one dense layer with flowing hyperparameters: the number of hidden units [16, 32, 64] in Bidirectional LSTM layer and the number of units [256, 512, 1024] in the dense layer; the combinatorial model of CNN and RNN (CNN + RNN) was developed using two convolutional layers, one Bidirectional LSTM layer and one dense layer with following hyperparameters: the number of filters [32, 64, 128] in first convolutional layer and the number of hidden units [16, 32, 64] in Bidirectional LSTM layer. We chose the model showed the highest mean AUROC (area under the receiver operating characteristic curve) in the five-fold cross-validation.

Evaluation of model performance

We used test sets, including Test-lncRNA ($n = 380$) and Test-mRNA ($n = 518$), which were excluded during model training, to evaluate model performances on predicting subcellular localizations of lncRNA and mRNA, respectively. Based on the confusion matrix from actual and predicted labels, models were assessed with the following indicators:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN \times -FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The AUROC was also calculated among the models except for mRNALoc because of its unsuitable possibilities from the output.

In our hands, the LightGBM-based RNAlight model developed in this study showed the best performance for both lncRNA and mRNA subcellular localization prediction.

Identification of nucleus- and cytoplasm-related k-mers

Tree SHAP [26] was used to measure k-mer contribution to RNA subcellular localization. In the Tree SHAP method, the SHAP value,

which is calculated on the basis of a game theoretic Shapley value for optimal credit allocations, is assigned to each identified *k*-mer in a specific RNA from RNAlight.

Nucleus- and cytoplasm-related *k*-mers were identified with two criteria: (i) For each given *k*-mer, Pearson correlation coefficients (PCCs) between the *k*-mer frequencies and SHAP values of each given *k*-mer in all examined RNAs are obtained to reflect its impact on the model output. (ii) The Z-score of mean absolute SHAP value is used to provide a general overview of its importance on the subcellular localization of all examined RNAs. Taken together, *k*-mers with PCC > 0.5 and Z-score > 1.96 suggest to be associated with nuclear localization, and ones with PCC < −0.5 and Z-score > 1.96 suggest to be associated with cytoplasmic localization.

Identification of known RBP-associated motifs related to RNA subcellular localization

Nucleus- and cytoplasm-related *k*-mers were separately assembled to consensus sequence features by the previously reported method [27]. Briefly, we tiled these *k*-mers back to each transcript (e.g. nucleus-related *k*-mers were tiled to the nucleus-localized transcripts) and joined consecutive *k*-mers together to form longer sequences as candidate sequence features. These candidate sequence features were merged to identify consensus sequences by multiple sequence alignment. Consensus sequence features were then mapped to known human RBP position weight matrices (PWMs) in CISBP-RNA database [38], which consists of RNA motifs and specificities to RBPs, by Tomtom [37] (version 5.3.0, https://meme-suite.org/meme/meme_5.3.0/tools/tomtom) to identify known RBP-associated motifs related to RNA subcellular localization.

Calculation of Light score

To use RNAlight for the prediction of a given RNA, we scaled the output of the RNAlight model as Light score:

$$\text{Light score} = 2 \times \text{probability} - 1$$

Here, the probability is the original output from RNAlight ranging from 0 to 1, representing the probability of nuclear localization about the input RNA. Scaled Light scores range from −1 to 1, of which a given Light score in the interval (−1, 0) indicates cytoplasmic localization or in the interval (0, 1) indicates nuclear localization.

Selection of circular RNAs

Highly expressed circRNAs without retained introns (FPB > 0.5, *n* = 800) and ciRNAs (FPB > 0.2, *n* = 104) in PA1 cells from published ribo–RNA-seq (GEO: GSE73325) [50] were identified by CLEAR pipeline [62] in this study, and these circular RNAs were further used to evaluate the performance of RNAlight. Circular RNA sequences were stretched by Bedtools (version 2.28.0).

Key Points
- A machine learning model, RNAlight, is developed to efficiently and sensitively predict subcellular localizations of mRNAs and lncRNAs.

- With embedded Tree SHAP algorithm, RNAlight further reveals distinct key sequence features and their associated RBPs for subcellular localizations of mRNAs or lncRNAs.
- RNAlight is successfully extended for the subcellular localization prediction of additional types of noncoding RNAs that were not used for model development, such as circular RNAs, suggesting its generality in RNA subcellular localization prediction.
- RNAlight is freely available at https://github.com/YangLab/RNAlight.

Data availability

All scripts used in this project are currently available at https://github.com/YangLab/RNAlight, including RNAlight model and related codes.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

We thank Yang laboratory for discussion and previous lab members, Zheng Luo, He-Na Zhang and Meng-Ran Wang, for their early tests on this project.

Funding

References

1. Mili S, Macara IG. RNA localization and polarity: from A(PC) to Z(BP). *Trends Cell Biol* 2009;**19**:156–64.
2. Chen LL. Linking long noncoding RNA localization and function. *Trends Biochem Sci* 2016;**41**:761–72.
3. Chen LL. Towards higher-resolution and in vivo understanding of lncRNA biogenesis and function. *Nat Methods* 2022;**19**:1152–5.
4. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 2015;**16**:651–64.
5. de Klerk E, Hoen PA. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet* 2015;**31**:128–39.
6. Yin Y, Lu JY, Zhang X, *et al.* U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* 2020;**580**:147–50.
7. Chen LL, DeCerbo JN, Carmichael GG. Alu element-mediated gene silencing. *EMBO J* 2008;**27**:1694–705.
8. Chen LL, Carmichael GG. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol Cell* 2009;**35**:467–78.

9. Bahar Halpern K, Caspi I, Lemze D, *et al.* Nuclear retention of mRNA in mammalian tissues. *Cell Rep* 2015;**13**:2653–62.

10. Lubelsky Y, Ulitsky I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* 2018;**555**:107–11.

11. Wang Y, Xu Z, Jiang J, *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell* 2013;**25**:69–80.

12. Zeng C, Fukunaga T, Hamada M. Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. *BMC Genomics* 2018;**19**:414.

13. Cabili MN, Dunagin MC, McClanahan PD, *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol* 2015;**16**:20.

14. Yao RW, Xu G, Wang Y, *et al.* Nascent Pre-rRNA sorting via phase separation drives the assembly of dense fibrillar components in the human nucleolus. *Mol Cell* 2019;**76**:767–783 e711.

15. Guo CJ, Ma XK, Xing YH, *et al.* Distinct processing of lncRNAs contributes to non-conserved functions in stem cells. *Cell* 2020;**181**:621–636 e622.

16. Benoit Bouvrette LP, Cody NAL, Bergalet J, *et al.* CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells. *RNA* 2018;**24**:98–113.

17. Fazal FM, Han S, Parker KR, *et al.* Atlas of subcellular RNA localization revealed by APEX-Seq. *Cell* 2019;**178**:473–490 e426.

18. Mas-Ponte D, Carlevaro-Fita J, Palumbo E, *et al.* LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* 2017;**23**:1080–7.

19. Zhang T, Tan P, Wang L, *et al.* RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res* 2017;**45**:D135–8.

20. Garg A, Singhal N, Kumar R, *et al.* mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res* 2020;**48**:W239–43.

21. Gudenas BL, Wang L. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Sci Rep* 2018;**8**:16385.

22. Cao Z, Pan X, Yang Y, *et al.* The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics* 2018;**34**:2185–94.

23. Su ZD, Huang Y, Zhang ZY, *et al.* iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 2018;**34**:4196–204.

24. Zhang ZY, Yang YH, Ding H, *et al.* Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief Bioinform* 2021;**22**:526–35.

25. Ke GL, Meng Q, Finley T, *et al.* LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 2017;**30**(Nips 2017):30.

26. Lundberg SM, Erion G, Chen H, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;**2**:56–67.

27. Wu KE, Parker KR, Fazal FM, *et al.* RNA-GPS predicts high-resolution RNA subcellular localization and highlights the role of splicing. *RNA* 2020;**26**:851–65.

28. Wang J, Wang L. Deep learning of the back-splicing code for circular RNA formation. *Bioinformatics* 2019;**35**:5235–42.

29. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;**26**:990–9.

30. Pan X, Shen HB. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* 2018;**34**:3427–36.

31. Kim HK, Kim Y, Lee S, *et al.* SpCas9 activity prediction by Deep-SpCas9, a deep learning-based model with high generalization performance. *Sci Adv* 2019;**5**:eaax9249.

32. Engel KL, Lo HG, Goering R, *et al.* Analysis of subcellular transcriptomes by RNA proximity labeling with Halo-seq. *Nucleic Acids Res* 2022;**50**:e24.

33. Bar N, Korem T, Weissbrod O, *et al.* A reference map of potential determinants for the human serum metabolome. *Nature* 2020;**588**:135–40.

34. Yan J, Xu Y, Cheng Q, *et al.* LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biol* 2021;**22**:271.

35. Shukla CJ, McCorkindale AL, Gerhardinger C, *et al.* High-throughput identification of RNA nuclear enrichment sequences. *EMBO J* 2018;**37**:e98452.

36. Azam S, Hou S, Zhu B, *et al.* Nuclear retention element recruits U1 snRNP components to restrain spliced lncRNAs in the nucleus. *RNA Biol* 2019;**16**:1001–9.

37. Gupta S, Stamatoyannopoulos JA, Bailey TL, *et al.* Quantifying similarity between motifs. *Genome Biol* 2007;**8**:R24.

38. Ray D, Kazan H, Cook KB, *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 2013;**499**:172–7.

39. Basu A, Dong B, Krainer AR, *et al.* The intracisternal A-particle proximal enhancer-binding protein activates transcription and is identical to the RNA- and DNA-binding protein p54nrb/NonO. *Mol Cell Biol* 1997;**17**:677–86.

40. Yamazaki T, Souquere S, Chujo T, *et al.* Functional domains of NEAT1 architectural lncRNA induce paraspeckle assembly through phase separation. *Mol Cell* 2018;**70**:1038–1053 e1037.

41. Wen D, Huang Z, Li Z, *et al.* LINC02535 co-functions with PCBP2 to regulate DNA damage repair in cervical cancer by stabilizing RRM1 mRNA. *J Cell Physiol* 2020;**235**:7592–603.

42. Warzecha CC, Sato TK, Nabet B, *et al.* ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol Cell* 2009;**33**:591–601.

43. Lin JC, Hsu M, Tarn WY. Cell stress modulates the function of splicing regulatory protein RBM4 in translation control. *Proc Natl Acad Sci U S A* 2007;**104**:2235–40.

44. Kloc M, Zearfoss NR, Etkin LD. Mechanisms of subcellular mRNA localization. *Cell* 2002;**108**:533–44.

45. Derrien T, Johnson R, Bussotti G, *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012;**22**:1775–89.

46. Hutchinson JN, Ensminger AW, Clemson CM, *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 2007;**8**:39.

47. Pandya-Jones A, Markaki Y, Serizay J, *et al.* A protein assembly mediates Xist localization and gene silencing. *Nature* 2020;**587**:145–51.

48. Xu M, Chen X, Lin K, *et al.* lncRNA SNHG6 regulates EZH2 expression by sponging miR-26a/b and miR-214 in colorectal cancer. *J Hematol Oncol* 2019;**12**:3.

49. Zhang J, Quan LN, Meng Q, *et al.* miR-548e sponged by ZFAS1 regulates metastasis and cisplatin resistance of OC by targeting CXCR4 and let-7a/BCL-XL/S signaling axis. *Mol Ther Nucleic Acids* 2020;**20**:621–38.

50. Zhang Y, Xue W, Li X, *et al.* The biogenesis of nascent circular RNAs. *Cell Rep* 2016;**15**:611–24.

51. Watkins NJ, Bohnsack MT. The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *Wiley Interdiscip Rev RNA* 2012;**3**:397–414.

52. Kastner B, Will CL, Stark H, *et al.* Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harb Perspect Biol* 2019;**11**:a032417.

53. Li X, Yang L, Chen LL. The biogenesis, functions, and challenges of circular RNAs. *Mol Cell* 2018;**71**:428–42.

54. Chen LL. The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat Rev Mol Cell Biol* 2020;**21**:475–90.

55. Salzman J, Gawad C, Wang PL, *et al.* Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 2012;**7**:e30733.

56. Li X, Liu CX, Xue W, *et al.* Coordinated circRNA biogenesis and function with NF90/NF110 in viral infection. *Mol Cell* 2017;**67**:214–227 e217.

57. Liu CX, Li X, Nan F, *et al.* Structure and degradation of circular RNAs regulate PKR activation in innate immunity. *Cell* 2019;**177**:865–880 e821.

58. Li S, Li X, Xue W, *et al.* Screening for functional circular RNAs using the CRISPR-Cas13 system. *Nat Methods* 2021;**18**:51–9.

59. You X, Vlatkovic I, Babic A, *et al.* Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat Neurosci* 2015;**18**:603–10.

60. Zhang Y, Zhang XO, Chen T, *et al.* Circular intronic long noncoding RNAs. *Mol Cell* 2013;**51**:792–806.

61. Li X, Zhang JL, Lei YN, *et al.* Linking circular intronic RNA degradation and function in transcription by RNase H1. *Sci China Life Sci* 2021;**64**:1795–809.

62. Ma XK, Wang MR, Liu CX, *et al.* CIRCexplorer3: A CLEAR pipeline for direct comparison of circular and linear RNA expression. *Genomics Proteomics Bioinformatics* 2019;**17**:511–21.

63. Meer EJ, Wang DO, Kim S, *et al.* Identification of a cis-acting element that localizes mRNA to synapses. *Proc Natl Acad Sci U S A* 2012;**109**:4639–44.

64. Wilusz JE, JnBaptiste CK, Lu LY, *et al.* A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. *Genes Dev* 2012;**26**:2392–407.

65. Carlevaro-Fita J, Rahim A, Guigo R, *et al.* Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* 2016;**22**:867–82.

66. Yoshimoto R, Kaida D, Furuno M, *et al.* Global analysis of pre-mRNA subcellular localization following splicing inhibition by spliceostatin A. *RNA* 2017;**23**:47–57.

67. Kirk JM, Kim SO, Inoue K, *et al.* Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* 2018;**50**:1474–82.

68. Lorenz R, Bernhart SH, Siederdissen C HZ, *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011;**6**:26.

69. Danaee P, Rouches M, Wiley M, *et al.* bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res* 2018;**46**:5381–94.