PSI-MOUSE: Predicting Mouse Pseudouridine Sites From Sequence and Genome-Derived Features

Evolutionary Bioinformatics
Volume 16: 1–9
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934320925752
SAGE

Bowen Song[1], Kunqi Chen[1], Yujiao Tang[1], Jialin Ma[2], Jia Meng[1] and Zhen Wei[1]

[1]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China. [2]Cancer Genome Computational Analysis, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

ABSTRACT: Pseudouridine (Ψ) is the first discovered and the most prevalent posttranscriptional modification, which has been widely studied during the past decades. Pseudouridine was observed in almost all kinds of RNAs and shown to have important biological functions. Currently, the time-consuming and high-cost procedures of experimental approaches limit its uses in real-life Ψ site detection. Alternatively, by taking advantage of the explosive growth of Ψ sequencing data, the computational methods may provide a more cost-effective avenue. To date, the existing mouse Ψ site predictors were all developed based on sequence-derived features, and their performance can be further improved by adding the domain knowledge derived feature. Therefore, it is highly desirable to propose a genomic feature-based computational method to increase the accuracy and efficiency of the identification of Ψ RNA modification in the mouse transcriptome. In our study, a predictive framework PSI-MOUSE was built. Besides the conventional sequence-based features, PSI-MOUSE first introduced 38 additional genomic features derived from the mouse genome, which achieved a satisfactory improvement in the prediction performance, compared with other existing models. Moreover, PSI-MOUSE also features in automatically annotating the putative Ψ sites with diverse types of posttranscriptional regulations (RNA-binding protein [RBP]-binding regions, miRNA-RNA interactions, and splicing sites), which can serve as a useful research tool for the study of Ψ RNA modification in the mouse genome. Finally, 3282 experimentally validated mouse Ψ sites were also collected in a database with customized query functions. For the convenience of academic users, a website was built to provide a user-friendly interface for the query and analysis on the database. The website is freely accessible at www.xjtlu.edu.cn/biologicalsciences/psimouse and http://psimouse.rnamd.com. We introduced the genome-derived features to mouse for the first time, and we achieved a good performance in mouse Ψ site prediction. Compared with the existing state-of-art methods, our newly developed approach PSI-MOUSE obtained a substantial improvement in prediction accuracy, marking the reliable contributions of genomic features for the prediction of RNA modifications in a species other than human.

KEYWORDS: Pseudouridine sites, genomic feature, web-server

RECEIVED: January 22, 2020. **ACCEPTED:** March 30, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been supported by the National Natural Science Foundation of China (grant no. 31671373); XJTLU Key Program Special Fund (grant no. KSF-T-01).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Zhen Wei, Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China. Email: zhen.wei@xjtlu.edu.cn

Introduction

Pseudouridine (Ψ) is the first discovered[1] and the most abundant RNA modification[2] that has been widely studied in the past decades. To date, 13 pseudouridine synthases (PUSs) have been identified, which are responsible for the catalytic reactions of Ψ sites in various RNA types,[3-7] as well as its differential functions in nuclear miRNA processing.[8] It has been shown that pseudouridylation in the rRNA and tRNA can maintain the ribosome entry sites[9] and stabilize its structure.[10] Pseudouridylation of tRNA-derived fragments were also found to impact stem cell commitment during early embryogenesis through translation control.[11] In mRNA, Ψ is contributed to the regulation of transcript stability,[12] translation efficiency, and RNA immune response.[13] The mRNA structure was proved to be recognized as the target of PUS 1, which implicates that modulation of RNA structure may be a possible mechanism to regulate mRNA pseudouridylation.[14] Pseudouridine was also found to be responded to environmental changes like serum starvation and been regulated by those signals accordingly.[15] In the mouse model, the associations between diseases and abnormal pseudouridylation have been reported.[16-18] For example, the sequence-specific pseudouridylation of ribosomal RNAs is linked to the growth and metastasis of xenograft tumors.[16] The mutation of DKC1, a Ψ synthase (PUS) that affects the accumulation of telomerase RNA,[17] can lead to dyskeratosis congenita (DC) and cancer.[18]

Currently, the transcriptome-wide distribution of Ψ RNA modification has been profiled by several types of high-throughput sequencing techniques, which contains Pseudo-seq,[15] Ψ-seq,[12] PSI-seq,[19] CeU-seq,[20] and ribosomal binding site (RBS)-seq.[21] The first 4 approaches were developed with similar principles, in which RNA is treated with the N-cyclohexyl-N′-(2-morpholinoethyl)-carbodiimide-metho-*p*-toluenesulfonate (CMC), and the reverse transcription is stopped on Ψ position by the presence of a bulky group. As a recently developed novel Ψ site-detection approach, the RBS-seq[21] is realized based on a modified version of the RNA bisulfite sequencing, which could result in better sensitivity. Although thousands of experimentally validated Ψ sites were revealed by the 5 sequencing techniques mentioned above, only 2 mouse Ψ data sets are publicly available, while both of them are identified by CeU-seq technique. Several computational models have been developed[22-27] to support the prediction of mouse Ψ sites in a given RNA transcript. However, all those

Table 1. Base-resolution data set used for Ψ site prediction.

SPECIES	DATA SET ID	CELL LINE	SITE #	OVERLAPPED #	TECHNIQUE	SOURCE
Mouse	M1	Brain	1566	46	CeU-Seq	Li et al[20]
	M2	Liver	1484			

All the mouse Ψ data sets from CeU-Seq can be freely downloaded from PSI-MOUSE website. Only the Ψ sites not previously used as training data were considered during performance evaluation, so the training sites and testing sites did not overlap.

methods are based on sequence-derived features. The published prediction models were reported to have reasonable performances, but their accuracy still has room for improvement. Moreover, it is worth noting that, in a more recent study, Dou et al applied a feature extraction approach bi-profile Bayes (BPB)[28] to obtain sequence information from both positive and negative RNA sequences.[29] Interestingly, their results concluded that the overall performances for Ψ identification using BPB features as well as the combined features (a combination of sequence and BPB features) were not obviously enhanced, with the general accuracies ranging from 60% to 70%. Therefore, it is highly desirable to develop a high-accuracy approach for the identification of Ψ modification in mouse transcriptome, not only for taking the best use of experimentally detected data but also to further improve the real-life biological usages of Ψ predictors.

In this study, we first introduced genome-derived information to the predictive framework of Ψ RNA modification in the mouse transcriptome. The genomic feature was first introduced and implemented in research for human m^6A site prediction,[30] which achieved a marked improvement in performance compared with previously sequence-based only predictors. In a more recent study, the combination of genomic and sequence features was applied to site prediction work of human Ψ RNA modification,[31] which validated the consistent performance of genome-derived information in a different type of human RNA modification. However, both these 2 studies focused on human transcriptome only, no evidence was found if genome-derived features work equally effective in a species other than human. In our study PSI-MOUSE, we extracted 38 genomic features from mouse transcripts. Combining the hand-crafted features with conventional sequence-based information, a high-accuracy mouse Ψ site predictor was established. We further trained and validated our predictive framework in the yeast genome and compared the performance with other competing methods. PSI-MOUSE also systematically checks whether the putative Ψ sites localize in regions with various post-transcriptional regulation factors, that is, RNA-binding protein (RBP)-binding regions, miRNA targets, and splicing sites, which can serve as a useful computational approach to facilitate the study of mouse Ψ RNA modification. The web-based server of PSI-MOUSE is freely accessed at www.xjtlu.edu.cn/biologicalsciences/psimouse and http://psimouse.rnamd.com.

Materials and Methods

Training and testing data

The known base-resolution mouse Ψ sites used for training and benchmarking were obtained from CeU-seq (Table 1) downloaded from Gene Expression Omnibus (GEO), we then filtered the experimentally validated Ψ sites to remain only those located on known transcripts for the extraction of genomic features. For the extraction of negative data, the unmodified U sites located on the same transcripts of positive Ψ sites were randomly selected. To fully utilize the underrepresented positive data, we match 10 negative sites for each of the positive site, from which 10 separate predictors were established with 1:1 positive-to-negative ratio, and their prediction performances were averaged during the evaluation. The known mouse Ψ sites from CeU-seq were detected under 2 independent conditions (brain and liver, see Table 1). In order to perform data set-level leave-one-out cross-validation, data set M1 was used for training first, while its performance was validated by data set M2 generated under another condition. Subsequently, data set M2 was used for training, with data set M1 for independent testing.

In order to faithfully compare the performance of our approach with competing methods, we further used the mouse and yeast Ψ data sets collected from Chen et al's[23] study (Supplementary Table S1), which were also applied as benchmark data sets in many previous published studies.[25-27]

Feature extraction

Genome-derived features. The previous Ψ predictors were all developed by extracting sequence-based information from RNA segments. Inspired by the successful application of genomic features used in human m^6A prediction, a total of 38 mouse and 22 yeast genome-derived features were generated to capture the attributes of the transcriptomic topologies for Ψ sites. The Genomic Features R/Bioconductor package with transcript annotations mm10 TxDb package and sacCer3 TxDb package[32] were used to generate Genomic Features 1-16, which are dummy variables indicating the overlapping state with the corresponding transcriptomic regions. We checked whether the uridine sites localize in a variety of types of topological features, that is, 5′UTR, 3′UTR, and intron. In order to overcome the isoform ambiguity from the annotation, only the transcript sub-regions on the primary (longest) transcripts of each gene were extracted. The second category of features (number 17-20) calculates the relative position of uridine sites in the transcript regions. The features are real numbered values between 0 and 1, where the value close to 0 indicates the relative approximation with the 5′ start of the region. The values are set to 0 if the sites do not belong to this region. Genomic features 21 to 25 calculate the length of the corresponding transcript regions where uridine sites localize. The values are also set to 0 for the sites not overlapping with the transcript region. The distances between uridine to the splicing junctions and the nearest Ψ site are captured in features 26 to 29. The clustering status of uridine sites was presented in features 30 and 31, which were defined by the number of neighboring uridine sites within the 100 bp and 1000 bp flanking regions of the target uridine sites. Besides, the RNA secondary structures around the uridine site are predicted using RNAfold with the Vienna RNA package[33] and presented in features 32 and 33. Finally, the various functional properties of the transcripts containing the Ψ RNA modification were represented by features 34 to 38, such as the number of isoforms and the transcript types of the coding. The detailed information of all genomic features we used in the Ψ site prediction was listed in Table 2.

Sequence-derived feature. To try to achieve the best performance in accuracy, 2 kinds of sequence-based features were also extracted and combined with genomic features mentioned above. The chemical properties of nucleotides and nucleotide density were first created and used for splice site prediction[34] and then being widely used in the predictive framework of RNA modification.[35-37] For chemical properties, the 4 types of nucleotides were classified into 3 categories. First, the adenosine and guanosine have 2 rings in their structure, while the cytidine and uridine only have one ring. Second, the guanosine and cytidine have stronger hydrogen bonding than that of adenosine and uridine. And finally, adenosine and cytidine contain the amino group, while guanosine and uridine contain the keto group. Based on these 3 principles, the *i*th nucleotide from sequence *S* may be encoded by a vector $S_i = (x_i, y_i, z_i)$

$$x_i = \begin{cases} 1 \text{ if } s_i \in \{A,G\} \\ 0 \text{ if } s_i \in \{C,U\} \end{cases}, \; y_i = \begin{cases} 1 \text{ if } s_i \in \{A,C\} \\ 0 \text{ if } s_i \in \{G,U\} \end{cases}, \\ z_i = \begin{cases} 1 \text{ if } s_i \in \{A,U\} \\ 0 \text{ if } s_i \in \{C,G\} \end{cases} \tag{1}$$

Thus, the A, C, G, U can be encoded as a vector (1,1,1), (0,1,0), (1,0,0), and (0,0,1), respectively. For the nucleotide density, the cumulative nucleotide frequency of nucleotide in *i*th position is calculated. The density of nucleotide in *i*th position is defined as the occurrences of the nucleotide A_i before $i+1$ position divided by its position *i*: $d_i = A_i / i$. If we take the RNA fragment "GAUACGCUA" as an example, the cumulative frequency of adenosine at the second, fourth, and ninth position is encoded as .50 (1/2), .50 (2/4), and .33 (3/9), respectively; and for cytidine is .20 (1/5) and .29 (2/7) at the fifth and seventh position of the example sequence.

Machine-learning approach used for Ψ site prediction

Support vector machine (SVM) has been widely applied for the prediction of genomic and proteomic data in computational biology, such as the microRNA target prediction,[38] and protein phosphorylation prediction.[39] In terms of RNA modification, the SVM was also used in previous m6A,[40] m5C,[41] and m1A[42] prediction and was shown to be a robust and effective machine-learning algorithm. In this project, the *R* interface of LIBSVM was used to build the predictor using the kernel of radial basis function.[43] In the data set-level leave-one-out experiment, only the Ψ sites not previously used as training data were considered in independent testing, which avoided the leakage between the training and testing data. Consequently, the performance evaluated on the testing data can justly reflect the ability of the purposed predictor to identify unknown Ψ sites.

Model performance evaluation

To evaluate the performance, we calculated the receiver operating characteristic (ROC) curve (sensitivity against 1-specificity) and the area under ROC curve (AUROC) as the main performance evaluation metric. The sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and overall accuracy (ACC) were calculated as other indicators to evaluate the reliability of the model. All the prediction processing is based on *R* language. To compare the performance of the newly proposed model with the existing Ψ site predictors, we reproduced the machine learning scaffold of PPUS, iRNA-PseU, and PseUI by realizing each of their sequence-based encoding methods, and the predictors are learned with the training data used in our study

$$Sn = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP) \times(TN+FN)}} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where *TP* represents the true positive, while *TN* represents the true negative; *FP* is the number of false positives, and *FN* the number of false negatives.

Table 2. Genome-derived features used for mouse and yeast Ψ site prediction.

ID	NAME	DESCRIPTION	NOTE	SPECIES
1	UTR5	5′ UTR	Dummy variables indicating whether the site is overlapped to the topological region on the major RNA transcript.	Mouse only
2	UTR3	3′ UTR		
3	**cds**	**Coding sequence**		Mouse and yeast
4	**Stop_codons**	**Stop codons flanked by 100bp**		
5	**Start_codons**	**Start codons flanked by 100bp**		
6	**TSS**	**Downstream 100bp of TSS**		
7	**TSS_A**	**Downstream 100bp of TSS on A**		
8	**exon_stop**	**Exons containing stop codons**		
9	alternative_exon	Alternative exons		Mouse only
10	**constitutive_exon**	**Constitutive exons**		Mouse and yeast
11	internal_exon	Internal exons		Mouse only
12	**long_exon**	**Long exons (exon length ⩾ 400bp)**		Mouse and yeast
13	**last_exon**	**5′ last_exon**		
14	**last_exon_400bp**	**5′ 400bp of the last exons**		
15	last_exon_sc400	5′ 400bp of the last exons containing stop codons		Mouse only
16	intron	intron		
17	**pos_cds**	**Relative position on coding sequence**	Relative position on the region	Mouse and yeast
18	pos_UTR5	Relative position on 5′ UTR		Mouse only
19	pos_UTR3	Relative position on 3′ UTR		
20	**pos_exons**	**Relative position on exon**		Mouse and yeast
21	length_UTR5	5′ UTR length	The region length in bp.	Mouse only
22	length_UTR3	3′ UTR length		
23	**length_gene_ex**	**Mature transcript length**		Mouse and yeast
24	**length_cds**	**Coding sequence length**		
25	**length_gene_full**	**Full transcript length**		
26	**dist_sj_5_p2000**	**Distance to the 5′ splicing junction**	Nucleotide distances toward the splicing junctions or the nearest neighboring sites.	
27	**dist_sj_3_p2000**	**Distance to the 3′ splicing junction**		
28	dist_nearest_p200	Distance to the closest neighbor truncated at 200bp		Mouse only
29	dist_nearest_p2000	Distance to the closest neighbor truncated at 2000bp		
30	**clust_f100**	**Clustering information at 100pb**	Clustering information of modification sites	Mouse and yeast
31	**clust_f1000**	**Clustering information at 1000pb**		
32	struc_hybridize	Predicted RNA hybridized region	RNA secondary structures	Mouse only
33	struc_loop	Predicted RNA loop region		
34	sncRNA	sncRNA	Genomic properties	
35	lncRNA	lncRNA		
36	**isoform_num**	**Number of isoforms**		Mouse and yeast
37	**exon_num**	**Number of exons**		
38	**GC_cont_genes**	**GC composition of genes**		

Genomic features generated from both mouse and yeast genome are in bold.

Estimate the probability of mouse Ψ RNA modification

In our study, the likelihood ratio (LR) is used to indicate how likely the computationally predicted result to be a true Ψ RNA modification. A LR of "100" suggests that it is an experimentally validated ψ site. The calculation of LR is showed as follows

$$LR = \frac{P(observation \mid \Psi)}{P(observation \mid \mathrm{U})} \tag{6}$$

The statistical significance of LR is assessed by an upper bound of the *p* value. It indicates how extreme the observed LR is among all the mouse transcriptome *U* sites. It is calculated from the relative ranking of the putative Ψ sites, that is, if only 0.1% of *U* sites have an LR score larger than a specific *U* site, then the upper bound of the *p* value of this site is .001. A transcriptome-wide prediction of mouse Ψ site was applied using PSI-MOUSE, a putative Ψ site is considered as high confidence if its LR within top 0.5% of LR (corresponding to p value < .005) of all the transcriptome *U* sites, followed by medium confidence (.005 < p value < .05) and low confidence (p value > .05).

Gene annotation and posttranscriptional regulation association analysis

To help academic users gain a more comprehensive understanding of the predicted mouse Ψ sites, various types of functional genomic information were automatically annotated in PSI-MOUSE for each putative Ψ site using ANNOVAR package,[44] that is, gene symbol, Ensembl gene ID, gene region, and gene type. PSI-MOUSE also checks whether the putative Ψ sites localized in regions with RBPs, miRNA-RNA targets, and splice sites, which help users to explore the potential effect of Ψ RNA modification on posttranscriptional regulation. The RBPs regions were obtained from POSTAR2,[45] the miRNA-RNA targets were downloaded from miRanda[46] and starBase2,[47] and the Canonical splice sites (GT-AG) from UCSC[32] annotations. For splicing sites analysis, specifically, we expanded 100 bp upstream region from 5′ splicing sites and 100 bp downstream region from 3′ splicing sites, the putative Ψ sites were then mapped with these regions. The Supplementary Table S2 shows detailed information related to posttranscriptional regulation analysis.

Results

As the process of RNA library preparation for existing data sets was based on polyA selection, to avoid the under-representation of the intronic Ψ sites in the data, we designed 2 modes for mouse Ψ site prediction. For the mature mRNA mode, the positive and negative Ψ sites mapped to intron regions were both filtered to only remain the sites localized on the mature mRNA transcripts. While for the full transcript mode, all Ψ sites were considered in the prediction. Consequently, the performance of the predictor was evaluated under these 2 modes, respectively. In addition, we applied feature selection to identify the most significant subset of features, which was implemented using Perturb method.[48] The functional insights of genomic features in our tool can be reflected by the feature importance plot. The feature importance scores can indicate the scientific significance of genomic markers in terms of its predictability for Ψ. According to the rank of the importance, the *N* most significant features were reserved in the prediction and evaluated with 5-fold cross-validation. In the process of feature selection, the Ψ sites from mouse brain data set were used as training data, while the mouse liver Ψ sites were used for testing purposes. Finally, the top 26 and 28 genomic features contribute to the best prediction performance under mouse full transcript and mature mRNA model, respectively (see Figure 1).

We observed that our approach PSI-MOUSE achieved remarkable improvement in prediction accuracy compared with existing encoding methods for Ψ prediction when tested on 5-fold cross-validation and independent data sets (see Table 3). Besides, the SVM algorithm applied in our PSI-MOUSE model, we further re-built predictors using 4 other machine-learning approaches, which are Random Forest, Naïve Bayes, decision tree, and generalized linear model. When tested on independent data sets, our model, PSI-MOUSE, achieved a better performance than the competitive predictors in all conditions (Table 4). This result suggested that the features used by the prediction model contribute to most of the predictive power, and the impact of machine-learning algorithm is limited. The detailed evaluation by the sensitivity, specificity, ACC, and MCC were summarized in Supplementary Table S3 and S4 for the full-transcript model and mRNA model, respectively.

Besides, the study of iPseU-CNN,[27] iPseU-NCP,[25] and XG-PseU[26] all used data sets from Chen et al's study to develop and validate their predictive pipelines. To faithfully compare the performance of our approach with these competing methods, we further trained and tested our predictive framework using data sets from Chen et al's study. For mouse Ψ site prediction, data set M_944 was used for training purpose and tested on 5-fold cross-validation. Consistent with previous results, PSI-MOUSE achieved a major improvement in all conditions compared with other state-of-the-art predictors (Table 5), suggesting the reliability of the approach. For Ψ site prediction in yeast genome, data set S_628 was used to train the predictor, while its performance was tested on independent testing data set S_200. When tested on 5-fold cross-validation and independent data set, we observed that our approach, which integrated additional genomic features besides the conventional sequence features, achieved reasonable improvement in prediction performance compared with other competing predictors, respectively (Supplementary Table S5). To sum up, for the first time, we generated genome-derived features from mouse and yeast transcriptome, respectively. The predictive framework achieved satisfactory improvement compared with

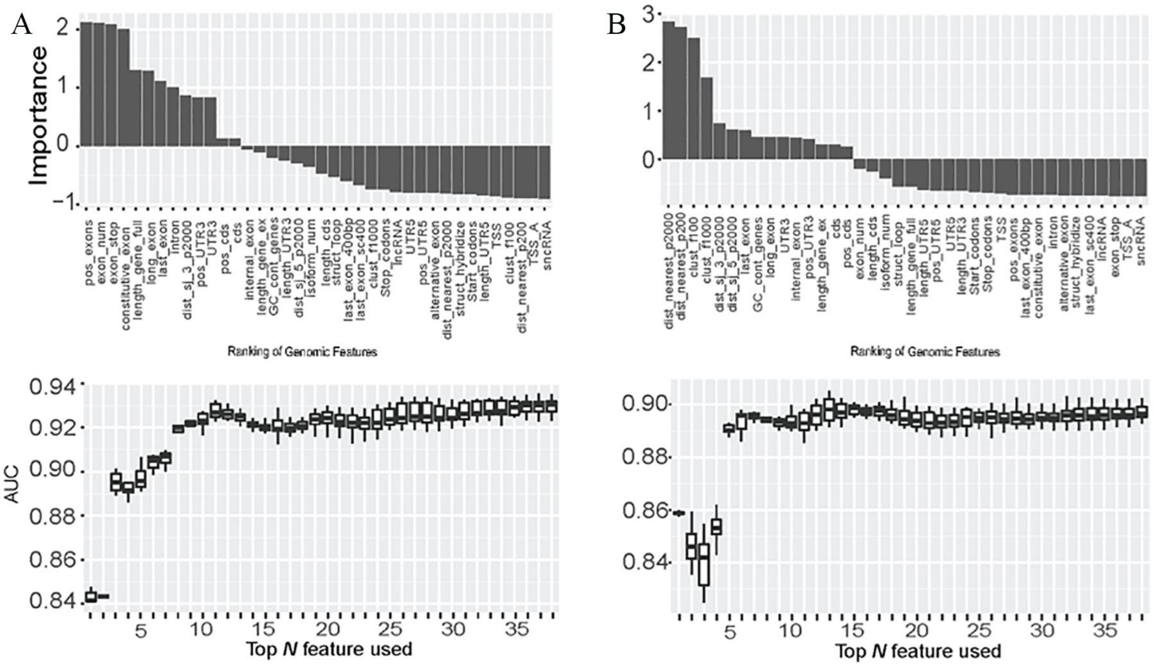


Figure 1. Feature selection of the genome-derived features for mouse Ψ site prediction. Top 26 and 28 genomic features were used in further prediction under mouse full transcript model (A) and mouse mature mRNA model (B), respectively. mRNA indicates messenger RNA.

Table 3. Performance evaluation of mouse Ψ site predictors (AUROC).

MODE	TESTING METHOD	PREDICTOR	BASE-RESOLUTION TECHNIQUE AND DATA SET ID		
			CeU-Seq		
			BRAIN	LIVER	AVERAGE
Full transcript	Cross-validation	PSI-MOUSE	0.982	0.982	0.982
		iRNA-PseU	0.798	0.791	0.795
		PPUS	0.795	0.789	0.792
		PseUI	0.715	0.701	0.708
	Independent data set	PSI-MOUSE	0.981	0.981	0.981
		iRNA-PseU	0.779	0.778	0.778
		PPUS	0.791	0.789	0.790
		PseUI	0.698	0.703	0.701
Mature mRNA	Cross-validation	PSI-MOUSE	0.943	0.948	0.945
		iRNA-PseU	0.799	0.790	0.794
		PPUS	0.794	0.789	0.792
		PseUI	0.676	0.664	0.670
	Independent data set	PSI-MOUSE	0.932	0.927	0.929
		iRNA-PseU	0.784	0.786	0.785
		PPUS	0.775	0.777	0.776
		PseUI	0.653	0.657	0.655

Abbreviation: AUROC, area under ROC curve.
Using the sequence-based encoding methods described in each existing predictor, we faithfully reproduced the iRNA-PseU, PPUS, and PseUI using the same training data of PSI-MOUSE. The 5-folds cross-validation and independent data set testing were performed to evaluate prediction accuracy under full-transcript and mature mRNA modes, respectively. When the independent data set testing was performed, the observations used for testing were excluded from the training data.

Table 4. Performance on an independent data set using different machine-learning algorithms (AUROC).

		INDEPENDENT TESTING (AVERAGE)			
		PSI-MOUSE	iRNA-PseU	PPUS	PseUI
Full transcript	SVM	0.981	0.778	0.790	0.701
	RF	0.984	0.787	0.789	0.802
	GLM	0.977	0.777	0.775	0.724
	NB	0.976	0.684	0.774	0.684
	DT	0.967	0.710	0.687	0.658
Mature mRNA	SVM	0.929	0.785	0.776	0.655
	RF	0.945	0.782	0.788	0.796
	GLM	0.942	0.777	0.772	0.719
	NB	0.898	0.684	0.772	0.706
	DT	0.901	0.719	0.704	0.670

Abbreviations: AUROC, area under ROC curve; DT: decision tree; GLM: generalized linear model; NB: Naïve Bayes; RF: random forest; SVM: support vector machine. The iRNA-PseU, PPUS, and PseUI were rebuilt using the same training data of PSI-MOUSE with their own sequence-based encoding methods.

Table 5. Performance evaluation of mouse Ψ predictors using data set M_944.

MODEL	TRAINING DATA SET	5-FOLD CROSS-VALIDATION			
		ACC (%)	MCC	Sn (%)	Sp (%)
PSI-MOUSE	M_944	91.97	0.84	86.62	97.31
iRNA-PseU		69.07	0.38	73.31	64.83
PseUI		70.44	0.41	74.58	66.31
iPseU-CNN		71.81	0.44	74.49	69.11
iPseU-NCP		71.82	0.44	67.37	76.27
XG-PseU		72.03	0.45	76.48	67.57

Abbreviations: ACC, accuracy; MCC, Matthews correlation coefficient; Sn, sensitivity; Sp, specificity.

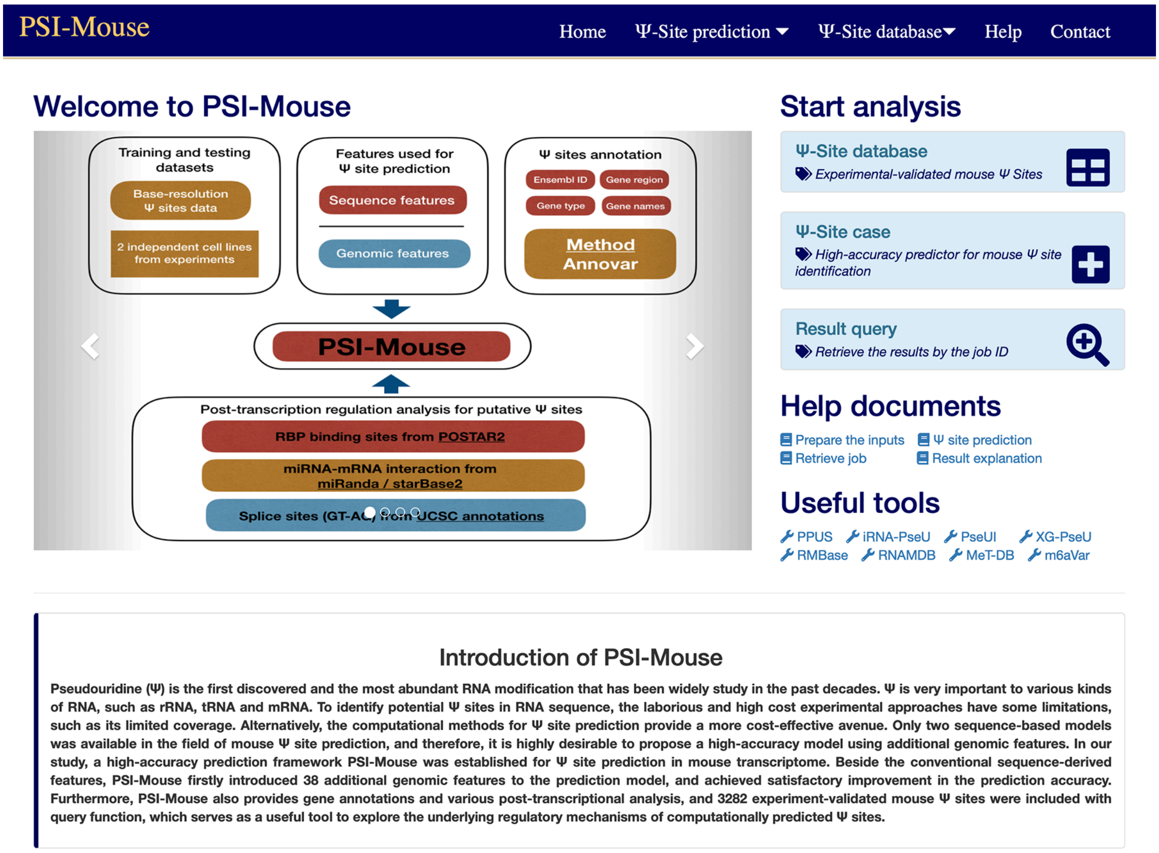
previous published predictors, marking the reliable contributions of genomic features for the prediction of RNA modifications in species other than human.

Web server interface

A web server for our newly purposed model was built using Hyper Text Markup Language (HTML), cascading style sheets (CSSs), and hypertext preprocessor (PHP). All the components of PSI-MOUSE can be easily accessed through the homepage of PSI-MOUSE (see Figure 2). The webserver of PSI-MOUSE provides different prediction modes, which allows users to customize their prediction under different pipelines, that is, 2 feature sets, various combinations of annotation methods. Users can upload either genome coordinate in txt format, or the FASTA file containing RNA sequences as the input of PSI-MOUSE. The database of PSI-MOUSE collected 3282 experimentally validated mouse Ψ sites, users can also query the database by Gene and Chromosome region. All the materials presented in the database and web server can be freely downloaded. PSI-MOUSE is freely accessed at www.xjtlu.edu.cn/biologicalsciences/psimouse and http://psimouse.rnamd.com.

Discussion

With the massive amount of data generated from various types of high-throughput sequencing techniques, many computational methods have been developed to facilitate the research of RNA modification, such as site prediction and data collection works,[49-53] RNA modification-associated genetic variants analysis tools,[54,55] as well as functional annotation tools.[56-59] In this study, we innovatively represented 38 RNA topological features on the mouse genome, and a prediction framework PSI-MOUSE was built upon them. Compared with existing works that based on sequence-derived features only, PSI-MOUSE achieved a significant improvement in performance accuracy, indicating the successful application of genome-derived features in the mouse



Welcome to PSI-Mouse

Start analysis

Ψ-Site database: *Experimental-validated mouse Ψ Sites*

Ψ-Site case: *High-accuracy predictor for mouse Ψ site identification*

Result query: *Retrieve the results by the job ID*

Help documents

Useful tools

Introduction of PSI-Mouse

Pseudouridine (Ψ) is the first discovered and the most abundant RNA modification that has been widely study in the past decades. Ψ is very important to various kinds of RNA, such as rRNA, tRNA and mRNA. To identify potential Ψ sites in RNA sequence, the laborious and high cost experimental approaches have some limitations, such as its limited coverage. Alternatively, the computational methods for Ψ site prediction provide a more cost-effective avenue. Only two sequence-based models was available in the field of mouse Ψ site prediction, and therefore, it is highly desirable to propose a high-accuracy model using additional genomic features. In our study, a high-accuracy prediction framework PSI-Mouse was established for Ψ site prediction in mouse transcriptome. Beside the conventional sequence-derived features, PSI-Mouse firstly introduced 38 additional genomic features to the prediction model, and achieved satisfactory improvement in the prediction accuracy. Furthermore, PSI-Mouse also provides gene annotations and various post-transcriptional analysis, and 3282 experiment-validated mouse Ψ sites were included with query function, which serves as a useful tool to explore the underlying regulatory mechanisms of computationally predicted Ψ sites.

Figure 2. Screenshot of the homepage of PSI-MOUSE. The web server and database of PSI-MOUSE can be easily accessed from the homepage. The webserver of PSI-MOUSE currently does not allow us to submit prediction jobs using Safari iOS system.

transcriptome. In addition, 3282 experimentally validated mouse Ψ sites with functional annotations were also collected in our work.

However, the mouse experimentally validated Ψ sites are only available in 2 conditions detected from the same sequencing technique, and thus, the prediction performance may be over-estimated. Besides, we further validated our approach using mouse and yeast Ψ data sets collected from Chen et al's study, which were also applied in many state-of-the-art Ψ site prediction studies. We observed a remarkable improvement of prediction accuracy in the mouse genome, as well as a relatively minor bonus in the yeast genome. The genomic features derived from annotation file of yeast transcript were obviously less than that of from mouse transcript; therefore, the application of genomic features in yeast can be further studied, and the performance of yeast Ψ site predictor can still be improved by deriving more informative genomic patterns from yeast genome. In future research, PSI-MOUSE will be further updated with more experimental identified Ψ sites detected under multiple technical contexts. In addition, the PSI-MOUSE scheme can be further expanded to the prediction of other RNA modification such as m^1A[60] and m^7G.[61]

Author Contributions

ZW initialized the project. ZW and BS designed the research plan. ZW and KC constructed the genomic features considered in mouse pseudouridine site prediction. BS performed the development of the pseudouridine site web server. YT and BS built the website. BS and ZW drafted the manuscript. All authors read, critically revised, and approved the final manuscript.

ORCID iDs

Bowen Song https://orcid.org/0000-0002-8586-0573
Kunqi Chen https://orcid.org/0000-0002-6025-8957

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Cohn WE, Volkin E. Nucleoside-5′-phosphates from ribonucleic acid. *Nature*. 1951;167:483-484.
2. Jacob R, Zander S, Gutschner T. The dark side of the epitranscriptome: chemical modifications in long non-coding RNAs. *Int J Mol Sci*. 2017;18:2387.
3. Hamma T, Ferre-D'Amare AR. Pseudouridine synthases. *Chem Biol*. 2006;13:1125-1135.
4. McCleverty CJ, Hornsby M, Spraggon G, Kreusch A. Crystal structure of human Pus10, a novel pseudouridine synthase. *J Mol Biol*. 2007;373:1243-1254.
5. Shaheen R, Han L, Faqeih E, et al. A homozygous truncating mutation in PUS3 expands the role of tRNA modification in normal cognition. *Hum Genet*. 2016;135:707-713.
6. Chen J, Patton JR. Pseudouridine synthase 3 from mouse modifies the anticodon loop of tRNA. *Biochemistry*. 2000;39:12723-12730.
7. Zhao X, Patton JR, Davis SL, Florence B, Ames SJ, Spanjaard RA. Regulation of nuclear receptor activity by a pseudouridine synthase through posttranscriptional modification of steroid receptor RNA activator. *Mol Cell*. 2004;15:549-558.

8. Song J, Zhuang Y, Zhu C, et al. Differential roles of human PUS10 in miRNA processing and tRNA pseudouridylation. *Nat Chem Biol*. 2020;16:160-169.
9. Jack K, Bellodi C, Landry DM, et al. rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells. *Mol Cell*. 2011;44:660-666.
10. Kierzek E, Malgowska M, Lisowiec J, Turner DH, Gdaniec Z, Kierzek R. The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res*. 2014;42:3492-3501.
11. Guzzi N, Ciesla M, Ngoc PCT, et al. Pseudouridylation of tRNA-derived fragments steers translational control in stem cells. *Cell*. 2018;173:1204.e26-1216.e26.
12. Schwartz S, Bernstein DA, Mumbach MR, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*. 2014;159:148-162.
13. Karijolich J, Yu YT. The new era of RNA modification. *RNA*. 2015;21: 659-660.
14. Carlile TM, Martinez NM, Schaening C, et al. mRNA structure determines modification by pseudouridine synthase 1. *Nat Chem Biol*. 2019;15:966-974.
15. Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*. 2014;515:143-146.
16. Cui L, Nakano K, Obchoei S, et al. Small nucleolar noncoding RNA SNORA23, up-regulated in human pancreatic ductal adenocarcinoma, regulates expression of spectrin repeat-containing nuclear envelope 2 to promote growth and metastasis of xenograft tumors in mice. *Gastroenterology*. 2017;153:292-306.
17. Mochizuki Y, He J, Kulkarni S, Bessler M, Mason PJ. Mouse dyskerin mutations affect accumulation of telomerase RNA and small nucleolar RNA, telomerase activity, and ribosomal RNA processing. *Proc Natl Acad Sci USA*. 2004;101:10756-10761.
18. Ruggero D, Grisendi S, Piazza F, et al. Dyskeratosis congenita and cancer in mice deficient in ribosomal RNA modification. *Science*. 2003;299:259-262.
19. Lovejoy AF, Riordan DP, Brown PO. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS ONE*. 2014;9:e110799.
20. Li X, Zhu P, Ma S, et al. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat Chem Biol*. 2015;11:592-597.
21. Khoddami V, Yerra A, Mosbruger TL, Fleming AM, Burrows CJ, Cairns BR. Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc Natl Acad Sci USA*. 2019;116:6784-6789.
22. He J, Fang T, Zhang Z, Huang B, Zhu X, Xiong Y. PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics*. 2018;19:306.
23. Chen W, Tang H, Ye J, Lin H, Chou KC. iRNA-PseU: identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids*. 2016;5:e332.
24. Li YH, Zhang G, Cui Q. PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics*. 2015;31:3362-3364.
25. Nguyen-Vo TH, Nguyen QH, Do TTT, Nguyen TN, Rahardja S, Nguyen BP. iPseU-NCP: identifying RNA pseudouridine sites using random forest and NCP-encoded features. *BMC Genomics*. 2019;20:971.
26. Liu K, Chen W, Lin H. XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. *Mol Genet Genomics*. 2019;295:13-21.
27. Tahir M, Tayara H, Chong KT. iPseU-CNN: identifying RNA pseudouridine sites using convolutional neural networks. *Mol Ther Nucleic Acids*. 2019;16:463-470.
28. Shao J, Xu D, Tsai SN, Wang Y, Ngai SM. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS ONE*. 2009;4:e4920.
29. Dou L, Li X, Ding H, Xu L, Xiang H. Is there any sequence feature in the RNA pseudouridine modification prediction problem? *Mol Ther Nucleic Acids*. 2019;19:293-303.
30. Chen K, Wei Z, Zhang Q, et al. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res*. 2019;47:e41.
31. Song B, Tang Y, Wei Z, et al. PIANO: a web server for pseudouridine-site (Ψ) identification and functional annotation. *Front Genet*. 2020;11:88.
32. Lawrence M, Huber W, Pages H, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;9:e1003118.
33. Lorenz R, Bernhart SH, Honer Zu, Siederdissen C, et al. ViennaRNA package 2.0. *Algorithms Mol Biol*. 2011;6:26.
34. Bari ATMG, Reaz MR, Choi HJ, Jeong BS. *DNA Encoding for Splice Site Prediction in Large DNA Sequence*. Berlin; Heidelberg: Springer; 2013:46-58.
35. Yang H, Lv H, Ding H, Chen W, Lin H. iRNA-2OM: a sequence-based predictor for identifying 2′-O-methylation sites in Homo sapiens. *J Comput Biol*. 2018;25:1266-1277.
36. Chen W, Feng P, Tang H, Ding H, Lin H. RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes. *Sci Rep*. 2016;6:31080.
37. Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N6-methyladenosine sites. *J Biomol Struct Dyn*. 2017;35:683-687.
38. Liu H, Yue D, Chen Y, Gao SJ, Huang Y. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*. 2010;11:476.
39. Wong Y-H, Lee TY, Liang HK, et al. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res*. 2007;35:W588-W594.
40. Wei L, Chen H, Su R. M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol Ther Nucleic Acids*. 2018;12:635-644.
41. Feng P, Ding H, Chen W, Lin H. Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. *Mol Biosyst*. 2016;12:3307-3311.
42. Feng P, Ding H, Yang H, Chen W, Lin H, Chou KC. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol Ther Nucleic Acids*. 2017;7:155-163.
43. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2:1-27.
44. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164-e164.
45. Zhu Y, Xu G, Yang YT, et al. POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res*. 2018;47:D203-D211.
46. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015;4:e05005.
47. Li J-H, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-seq data. *Nucleic Acids Res*. 2013;42:D92-D97.
48. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Model*. 2003;160:249-264.
49. Liu H, Wang H, Wei Z, et al. MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res*. 2018;46:D281-D287.
50. Liu L, Lei X, Meng J, Wei Z. WITMSG: large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features. *Curr Genomics*. 2020;21:67-76.
51. Han Y, Feng J, Xia L, et al. CVm6A: a visualization and exploration database for m(6)As in cell lines. *Cells*. 2019;8:168.
52. Xuan J-J, Sun WJ, Lin PH, et al. RMBase v2. 0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res*. 2017;46:D327-D334.
53. Liu L, Lei X, Fang Z, Tang Y, Meng J, Wei Z. LITHOPHONE: improving lncRNA methylation site prediction using an ensemble predictor [published online ahead of print May 06 2020]. *Front Genet*. doi:10.3389/fgene.2020.00545.
54. Song B, Tang Y, Chen K, et al. m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human [published online ahead of print March 12, 2020]. *Bioinformatics*. doi:10.1093/bioinformatics/btaa178.
55. Zheng Y, Nie P, Peng D, et al. m6AVar: a database of functional variants involved in m6A modification. *Nucleic Acids Res*. 2018;46:D139-D145.
56. Tang Y, Chen K, Wu X, et al. DRUM: inference of disease-associated m6A RNA methylation sites from a multi-layer heterogeneous network. *Front Genet*. 2019;10:266.
57. Zhang S-Y, Zhang SW, Fan XN, et al. Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. *PLoS Comput Biol*. 2019;15:e1006663.
58. Xue H, Wei Z, Chen K, Tang Y, Wu X, Su J. Prediction of RNA methylation status from gene expression data using classification and regression methods. *Evol Bioinf*. 2020;16:1-5.
59. Wu X, Wei Z, Chen K, et al. m6Acomet: large-scale functional prediction of individual m(6)A RNA methylation sites from an RNA co-methylation network. *BMC Bioinformatics*. 2019;20:223.
60. Safra M, Sas-Chen A, Nir R, et al. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature*. 2017;551:251-255.
61. Zhang LS, Liu C, Ma H, et al. Transcriptome-wide mapping of internal N(7)-methylguanosine methylome in mammalian mRNA. *Mol Cell*. 2019;74:1304-1316.