Molecular Therapy
Nucleic Acids
Original Article

m5UPred: A Web Server for the Prediction of RNA 5-Methyluridine Sites from Sequences

Jie Jiang,[1,5] Bowen Song,[2,5] Yujiao Tang,[1,5] Kunqi Chen,[1,4] Zhen Wei,[1,5] and Jia Meng[1,3,5]

[1]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; [2]Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; [3]AI University Research Centre, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; [4]Institute of Ageing & Chronic Disease, University of Liverpool, L7 8TX, Liverpool, UK; [5]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, L7 8TX, Liverpool, UK

As one of the widely occurring RNA modifications, 5-methyluridine (m^5U) has recently been shown to play critical roles in various biological functions and disease pathogenesis, such as under stress response and during breast cancer development. Precise identification of m^5U sites on RNA is vital for the understanding of the regulatory mechanisms of RNA life. We present here m5UPred, the first web server for *in silico* identification of m^5U sites from the primary sequences of RNA. Built upon the support vector machine (SVM) algorithm and the biochemical encoding scheme, m5UPred achieved reasonable prediction performance with the area under the receiver operating characteristic curve (AUC) greater than 0.954 by 5-fold cross-validation and independent testing datasets. To critically test and validate the performance of our newly proposed predictor, the experimentally validated m^5U sites were further separated by high-throughput sequencing techniques (miCLIP-Seq and FICC-Seq) and cell types (HEK293 and HAP1). When tested on cross-technique and cross-cell-type validation using independent datasets, m5UPred achieved an average AUC of 0.922 and 0.926 under mature mRNA mode, respectively, showing reasonable accuracy and reliability. The m5UPred web server is freely accessible now and it should make a useful tool for the researchers who are interested in m^5U RNA modification.

INTRODUCTION

The development of high-resolution transcriptome mapping and quantification technologies has made epigenetic modifications of RNA one of the fastest-growing fields in biological research in the past several years. Over 170 post-transcriptional modifications have been identified, with the majority of them occurred in tRNAs and rRNAs.[1] RNA modifications showed imperative roles in varied biological functions (e.g., embryonic stem cell development,[2] cancer cell survival, migration,[3] and response to environmental exposures[4]). Over 100 RNA modification enzyme mutations have been associated with human diseases.[5] Besides the well-known role of fine-tuning RNA structures and functions to regulate gene expression and protein synthesis by RNA modifications,[6] there are many other functions associated with RNA modifications. Some post-transcriptional RNA modifications have shown to be dynamic processes that have regulatory roles similar to post-translational protein modifications in controlling cell-type-specific functions.[7]

However, as one of the most abundant RNA modifications, the identification and functional characterization of 5-methyluridine (m^5U) remain extremely limited in the literature. As a pyrimidine modification, m^5U involves methylation at the 5-carbon position of uridine, which may be the first pyrimidine methyltransferases evolved to catalyze the pyrimidine-C5 methylation.[8] The enzymes catalyzing the modification of m^5U are TrmA in *Escherichia coli*,[9,10] Trm2p in *Saccharomyces cerevisiae*,[11] and TRMT2A and TRMT2B in mammals.[12,13] m^5U has been reported to participate in the development of breast cancer,[5] systemic lupus erythematosus,[14] and regulation of stress response and development in plants.[15] Accurate identification of m^5U sites is crucial to understanding fundamental biological processes and functions in all species. Some wet-lab experimental methods, such as miCLIP-Seq, iCLIP, and FICC-seq, have been developed to identify m^5U sites.[16] However, RNA sequencing could be a high-cost and time-consuming process, and the specificity of antibodies used for immunoprecipitation restricts the delivery of accurate sequencing results. Thus, only very limited data have been generated for m^5U sites so far. In this study, we would like to propose an *in silico* technique for the identification of m^5U sites based on sequence-derived information using a support vector machine (SVM) algorithm.

A number of computational methods have been developed to predict epigenetic modifications of RNA, including m6A site predictors WHISTLE,[17] SRAMP,[18] Gene2vec,[19] iRNA-Methyl,[20] and M6AMRFS;[21] m5C site predictors RNAm5Cfinder,[22] iRNA-m5C,[23] and M5C-HPCR;[24] pseudouridine site predictors iRNA-PseU,[25] PseUI,[26] PPUS,[27] and PIANO;[28] and one-stop platform iMRM for simultaneously identifying various RNA modifications in multiple species.[29] However, to the best of our knowledge, there is no m^5U site predictor available so far. Thus, in this study we would like to propose the first prediction framework, m5UPred, which can be utilized for high-accuracy identification of m^5U site from RNA sequences. A

Received 7 July 2020; accepted 25 September 2020;
https://doi.org/10.1016/j.omtn.2020.09.031.

Correspondence: Bowen Song, Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China.
E-mail: bowen.song@liverpool.ac.uk

Table 1. Performance Evaluation of m5UPred by Independent Testing Dataset

Mode	Algorithms	Sn (%)	Sp (%)	ACC (%)	MCC	AUC
Full transcript	SVM	86.06	84.72	85.39	0.708	0.933[a]
	RF	85.18	83.44	84.31	0.687	0.929
	NB	86.06	61.53	73.80	0.491	0.845
	GLM	83.03	80.45	81.74	0.635	0.897
Mature mRNA	SVM	86.99	89.07	88.03	0.761	0.951[a]
	RF	81.30	97.56	89.43	0.799	0.951
	NB	92.36	46.91	69.63	0.441	0.862
	GLM	85.57	81.54	83.56	0.672	0.915

We randomly selected 80% of experimentally validated m5U sites as training dataset and the performance of predictors was evaluated by the remaining 20% of m5U sites as independent testing data.
[a]SVM achieved best performance among all classifier with AUC of 0.933 and 0.951 for full transcript mode and mature mRNA mode.

Table 2. Prediction Performance Using Cross Validation and Independent Testing Dataset

Dataset	Testing Method	Sn (%)	Sp (%)	ACC (%)	MCC	AUC
Full transcript	cross validation	87.59	89.04	88.32	0.767	0.956
	independent test set	87.90	88.80	88.35	0.767	0.956
Mature mRNA	cross validation	88.64	91.18	89.91	0.798	0.956
	independent test set	87.44	91.95	89.70	0.795	0.954

80% of experimentally validated m5U sites were used for training, while its performance was evaluated by the remaining 20% as independent testing data.

user-friendly web server has also been developed and made free access publicly at https://www.xjtlu.edu.cn/biologicalsciences/m5u. We anticipate that our newly proposed model, m5UPred, could make the best use of limited experimentally detected data and facilitate the research of m^5U site modification by providing an alternative computational approach.

RESULTS

Determine the Best Machine Learning Algorithm Used for m^5U Site Prediction

In order to determine the best classifier for constructing m5UPred, the performances of different classifiers were tested on the independent dataset (see Table 1). SVM achieved area under the receiver operating characteristic curve (AUC) of 0.933 for full transcript mode and 0.951 for mature mRNA mode, which was the best among all classifiers and used to build m5UPred.

Performance Evaluation of m5UPred by Benchmark and Independent Testing Dataset

The prediction performance of m5UPred was evaluated by 5-fold cross-validation and an independent testing dataset, respectively (see Table 2). Using SVM as the final classifier, m5UPred was developed and tested using the same datasets that were previously applied for algorithm selection. In addition, the tune length of SVM in caret package was set to 9, with a random grid search for optimization. When evaluated by the independent testing data, m5UPred showed good generalization capability and achieved 0.767 Matthews correlation coefficient (MCC) with 88.35% accuracy and 0.956 AUC for full transcript mode and 0.795 MCC with 89.70% accuracy and 0.954 AUC for mature mRNA mode.

Performance Evaluation of m5UPred by Cross-Technique and Cross-Cell-Type Validation

It was shown previously that positive samples of RNA modification captured by different techniques may have different overall patterns.[30] To test the performance and robustness of our newly proposed predictor in finding m^5U sites generated from different techniques and cell types, we further evaluated the performance of m5UPred by cross-technique and cross-cell-type validation (Table 3). When tested by the independent dataset generated from another technique or cell type, m5UPred achieved AUC of 0.882 and 0.922 for cross-technique validation, and AUC of 0.899 and 0.926 for cross-cell-type validation, under full transcript mode and mature mRNA mode, respectively. It shows robustness and reliability of m5UPred.

Web Implementation

To facilitate the access of our model by experimental researchers, a web server has been developed using Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS), and Hypertext Preprocessor (PHP) and is accessible at https://www.xjtlu.edu.cn/biologicalsciences/m5u. It allows users to submit query RNA sequences in FASTA format with over 41 nt in length and a uridine in the center for analysis. The web server will evaluate the possibility of m^5U modification in the given sequences, which return all putative m^5U sites with download function (Figure 1).

DISCUSSION

In this study, we extracted m^5U modification sites from a data source generated by two sequencing methods and two cell types. A high-accuracy predictor was built by using a SVM model to predict the m^5U modification sites. Satisfactory prediction performance was observed, with an average AUC of 0.956 for full transcript and 0.954 for mature mRNA when it was evaluated by an independent test set with 20% randomly selected sites from the dataset. The model was further tested by cross-technique and cross-cell-type validation. The results show that the positive m^5U sites generated from different techniques and cell types share some common features, and these features can be captured by our m5UPred by extracting sequence-derived information. To help understand the sequence composition, we analyzed full transcript sequences from different cell types and sequencing methods by DREME[31] motif discovery tool. The motifs identified are provided in Figure S1 and are consistent with the tRNA T-loop motif of GTTCG/AA proposed by the data source paper.[16] A web server was developed based on our model and made available to the public to assist the prediction of m^5U sites by other researchers.

Table 3. Cross-Technique and Cross-Cell-Type Validation

			Cross-Technique Validation			Cross-Cell-Type Validation		
Mode	Testing Method	Evaluation Metric	miCLIP-Seq	FICC-Seq	Average	HEK293	HAP1	Average
Full transcript	cross validation	Sn (%)	86.70	89.80	88.25	86.26	89.67	87.96
		Sp (%)	86.83	91.37	89.10	87.19	90.48	88.84
		ACC (%)	86.76	90.58	88.67	86.72	80.15	83.44
		MCC	0.735	0.812	0.773	0.735	0.901	0.818
		AUC	0.946	0.966	0.956	0.942	0.969	0.955
	independent dataset	Sn (%)	75.36	56.48	65.92	82.79	57.77	70.28
		Sp (%)	89.23	90.10	89.67	89.62	90.21	89.92
		ACC (%)	82.29	73.29	77.79	86.20	73.99	80.10
		MCC	0.652	0.495	0.574	0.726	0.507	0.617
		AUC	0.910	0.853	0.882	0.941	0.857	0.899
Mature mRNA	cross validation	Sn (%)	88.34	94.14	91.24	89.86	95.32	92.59
		Sp (%)	90.52	98.04	94.28	91.13	96.71	93.92
		ACC (%)	89.43	96.09	92.76	90.50	96.02	93.26
		MCC	0.789	0.922	0.856	0.810	0.920	0.865
		AUC	0.962	0.992	0.977	0.964	0.987	0.975
	independent dataset	Sn (%)	90.07	38.81	64.44	95.41	39.01	67.21
		Sp (%)	90.86	97.51	94.19	88.93	98.12	93.53
		ACC (%)	90.46	68.16	79.31	92.17	68.57	80.37
		MCC	0.809	0.449	0.629	0.845	0.461	0.653
		AUC	0.970	0.873	0.922	0.981	0.871	0.926

As previously shown in Table 2, the experimentally validated m5U sites were further separated by high-throughput sequencing techniques and cell types, under full transcript and mature mRNA mode, respectively. Independent tests by the other technique or cell type are employed to evaluate the performance additional to 5-fold cross validation. When dataset miCLIP_F was used for training, its performance was tested by 5-fold cross-validation and an independent dataset of FICC_F. Similarly, while dataset FICC_F was used for training, miCLIP_F served as an independent test set. The same testing scheme was used for datasets separated by cell types as well. The performance of different sequencing techniques and cell types are then averaged. The performance evaluation of m5UPred by cross-technique and cross-cell-type validation using different machine learning classifiers is listed in Table S1.

It is worth noting that although m^5U sites were generated under different conditions, they are still from a single source; this may pose an over-fitting problem, and the performance of the model may be over-estimated. We calculated the false discovery rate (FDR) and false omission rate (FOR) at different thresholds and the true positive rate (TPR). The result confirmed our previous statement about over-estimation, and the result is provided in the Supplemental Information. We will keep updating our model when more data are available. When comparing the crosslink peaks from two sequencing methods of the raw data, only around 35% of sites with over 5 crosslink peaks are identified by both methods;[16] therefore, the true-positive m^5U sites will need to be further confirmed by more experimental data. We noticed that tRNA modifications take the dominant position in the dataset we used for training, which was also confirmed by the data source paper. Considering the biological structures and functional difference of varied RNA molecules, further study will need to be performed to optimize the model using optimized window size, secondary structures, modification motif, and genome information to further improve the model robustness and generalization ability.

MATERIALS AND METHODS

Training and Testing Datasets

Benchmark and Independent Testing Dataset Used for m^5U Site Prediction

The positive samples (m^5U sites) were obtained from the recently published single-nucleotide resolution m^5U sequencing data,[16] and the sequencing results were generated by FICC-seq and miCLIP-seq technologies on two cell lines, HEK293 and HAP1 (Table 4). Data were downloaded from Gene Expression Omnibus (GEO), with the GEO accession number GEO: GSE109183. Previous studies[32,33] showed RNA sequence 41 nt in length with an RNA modification site in middle provided the most promising prediction result. Thus, we adopt this formula and designed our positive dataset by generating 41 nt sequences with experimental identified U sites in the center. Unmodified uridine sites located on the same transcripts of the positive m^5U sites were randomly selected, and 10 negative datasets were generated. By combining each of these 10 negative sets with the positive data, 10 separate datasets were constructed with a 1:1 positive-to-negative ratio. Their prediction performances were averaged during the evaluation to reduce batch variance.

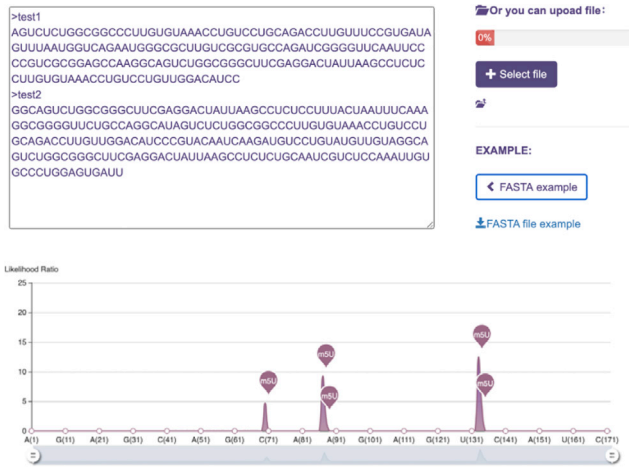

Figure 1. Input and Output of m5UPred Web Server

The input of m5UPred is FASTA sequences. The user can paste the sequences into a text box or provide a FASTA file. m5UPred will predict the m^5U site possibility using our m5UPred model and returns predicted m^5U sites. Results can be downloaded in table format.

The performance of our predictor was evaluated under two modes, namely full transcript mode and mature mRNA mode. For the full transcript mode, the positive and negative m^5U sites located in both exonic and intronic regions are all considered to generate the data, whereas, in the mature mRNA mode, only positive and negative m^5U sites located on mature mRNA transcripts are employed. Positive m5U sites in different RNA families are listed in the Supplemental Information. For performance evaluation, 80% of the dataset was randomly selected as a benchmark training dataset, while the remaining 20% was used as independent testing data.

Dataset Separated by High-Throughput Sequencing Techniques and Cell Types

Besides randomly selected training and testing data from experimentally validated m^5U sites, we applied two more strictly evaluation methods to validate the performance of the proposed predictor. The experimentally validated m^5U sites were further divided by high-throughput sequencing technique and cell type (Table 5), from which the cross-technique and cross-cell-type testing were applied for performance evaluation under full transcript and mature mRNA modes, respectively. The performance was also evaluated by 5-fold cross-validation and an independent test set. But rather than randomly selecting 20% of the total data as a test set, we generated test sets by different techniques and cell types. Concretely, for cross-technique validation, when dataset miCLIP_F or miCLIP_M were used for training, their performance was tested by an independent dataset of FICC_F or FICC_M. Then, we switched the train and test set and used dataset FICC_F or FICC_M for training and miCLIP_F or miCLIP_M for independent testing. The validation scheme was applied for the dataset separated by cell types as well.

Table 4. Base-Resolution Dataset Used for m^5U Site Prediction

GEO Accession	Technique	Cell Line	No. of Sites		Source
			Full	Mature	
GSE109183	miCLIP-Seq	HEK293	3,696	1,232	[15]
	FICC-Seq	HAP1			

80% of the data were randomly selected for training, while the remaining 20% were retained as an independent testing dataset for performance evaluation. All positive and negative sequences used in this project can be freely downloaded from https://www.xjtlu.edu.cn/biologicalsciences/m5u. and are provided in the Supplemental Information.

Feature Extraction

High classification accuracy of many machine learning algorithms largely relies on sequence encoding strategies for feature extraction of RNA sequences. To achieve the best performance, the chemical properties of nucleotides and their distribution information are employed for sequence feature extraction in this study. This strategy was adapted from previous work by Bari et al.,[34] which was originally used for splicing site prediction from DNA sequence and lately is being widely used in encoding RNA sequences for RNA modification prediction.[35–37]

Nucleotide Density

The nucleotide density represents the distribution and frequency information of nucleotides at each position. The density (d_i) of a nucleotide N at i^{th} position can be calculated by the number (n) of N occurred before $(i+1)^{th}$ position divided by i: $d_i = n/i$. So, for a sequence "AUAGUCAUAA," the density of A is 1, 0.67, 0.43, 0.44, and 0.50 at the 1^{st}, 3^{rd}, 7^{th}, 9^{th}, and 10^{th} positions, respectively. Similarly, U is 0.50, 0.40, and 0.38 at positions 2, 5, and 8; C is 0.17 at position 6; and G is 0.25 at position 4 (Table 6).

Nucleotide Chemical Property

The encoding scheme of nucleotide chemical property was designed based on different chemical structures of four RNA nucleotides. Four building blocks of RNA, namely adenosine (A), uridine (U), guanosine (G), and cytosine (C), are categorized into different groups depending on the number of ring structures (two for A and G and one for C and T), existence of an amino group (A and C) or keto group (G

Table 5. Dataset Separated by Sequencing Techniques and Cell Types

Separation Method	Mode	Condition	Site No.	Dataset	Total No.
Technique	full	miCLIP-Seq	2,225	miCLIP_F	3,696
		FICC-Seq	1,471	FICC_F	
	mature	miCLIP-Seq	823	miCLIP_M	1,232
		FICC-Seq	409	FICC_M	
Cell type	full	HEK293	2,467	HEK293_F	3,696
		HAP1	1,229	HAP1_F	
	mature	HEK293	868	HEK293_M	1,232
		HAP1	364	HAP1_M	

Table 6. Calculation of Nucleotide Density for a Sample Sequence

Sequence	A	U	A	G	U	C	A	U	A	A
No. of same nucleotide accumulated	1	1	2	1	2	1	3	3	4	5
Position	1	2	3	4	5	6	7	8	9	10
Nucleotide density	1	0.5	0.67	0.25	0.4	0.17	0.43	0.38	0.44	0.5

and T), and strong (C and G) or weak (A and T) hydrogen bonds. Based on this information, together with the nucleotide density information, nucleotide N at i^{th} position from sequence S (with length l) can be represented by the formula $Ni = \{x_i, y_i, z_i, d_i\} (i = 1, 2, 3, \ldots l)$ which satisfies the following equations:

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

Concretely, A, C, G, and U can be encoded as vectors (1,1,1, d_i), (0,1,0, d_i), (1,0,0, d_i), and (0,0,1, d_i), respectively. So, each of the nucleotides in the RNA sequences will be transferred into four numeric values; thus, each of our 41 nt RNA sequences will be encoded into a 164-dimension vector.

Evaluation Matric

Five metrics were employed to evaluate the performance our model, namely Sn (Sensitivity), Sp (Specificity), MCC, ACC (overall accuracy), and AUC, with the equations below:

$$Sn = \frac{TP}{TP + FN} \quad (2)$$

$$Sp = \frac{TN}{TN + FP} \quad (3)$$

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

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5)$$

in which TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. A threshold of 0.5 was applied to calculate Sn and Sp. R package ROCR[38] was used to evaluate the performance of the model.

Choice of Machine Learning Classifier

SVM, random forest (RF), generalized linear model (GLM), and Naive Bayes (NB) are the most popular machine learning classifiers in RNA modification prediction and have been widely used for different site predictions.[24,27,35] We evaluated the performance of these algorithms by an independent testing dataset, since the evaluation by cross-validation may over-estimate the performance of models.[39] The R package caret was used to construct machine learning models, and all parameters were set by default for primitive evaluation. The results are shown in Table 1. SVM achieved the best results among all four algorithms, which was selected as our final algorithm to build our m^5U site prediction model.

SVM

SVM is one of the most powerful yet flexible algorithms in bioinformatics research and has been used for many applications (e.g., mammalian microRNA target prediction,[40] predicting the subcellular location of proteins,[41] and RNA modification site prediction[17,42]). The final decision of SVM is determined by a representation of different classes in a hyperplane, which can be used for varied classification and regression tasks. In this study, we used the R package caret to construct our SVM model with a non-linear radial basis function. The prediction performance was evaluated by 5-fold cross-validation and an independent test set. Initially, all parameters were set by default to test the performance of the model, and then tune length was set to 9, with a random grid search for optimization. Random grid search has been proved by a previous study with better efficiency than manually set grid searches.[43]

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2020.09.031.

AUTHOR CONTRIBUTIONS

B.S. conceived the idea; K.C. and B.S. collected and processed the data; J.J. and B.S. constructed the predictive model and designed the experiments; Y.T. built the website; and J.J. and B.S. drafted the manuscript. All authors read, critically revised and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (31671373) and the XJTLU Key Program Special Fund (KSF-E-51). This work is partially supported by the AI University Research Centre through XJTLU Key Programme Special Fund (KSF-P-02).

REFERENCES

1. Boccaletto, P., Machnicka, M.A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T.K., de Crécy-Lagard, V., Ross, R., Limbach, P.A., Kotter, A., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. *46* (D1), D303–D307.

2. Batista, P.J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., Bouley, D.M., Lujan, E., Haddad, B., Daneshvar, K., et al. (2014). m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. Cell Stem Cell *15*, 707–719.
3. Delaunay, S., and Frye, M. (2019). RNA modifications regulating cell fate in cancer. Nat. Cell Biol. *21*, 552–559.
4. Yang, C. (2020). ToxPoint: Dissecting Functional RNA Modifications in Responses to Environmental Exposure-Mechanistic Toxicology Research Enters a New Era. Toxicol. Sci. *174*, 1–2.
5. Jonkhout, N., Tran, J., Smith, M.A., Schonrock, N., Mattick, J.S., and Novoa, E.M. (2017). The RNA modification landscape in human disease. RNA *23*, 1754–1769.
6. Yu, Y.T., Terns, R.M., and Terns, M.P. (2004). Mechanisms and functions of RNA-guided RNA modification. In Fine-Tuning of RNA Functions by Modification and Editing, H. Grosjean, ed. Top. Curr. Genet., *12* (Springer), pp. 223–262.
7. He, C. (2010). Grand challenge commentary: RNA epigenetics? Nat. Chem. Biol. *6*, 863–865.
8. Bujnicki, J.M., Feder, M., Ayres, C.L., and Redman, K.L. (2004). Sequence-structure-function studies of tRNA:m5C methyltransferase Trm4p and its relationship to DNA:m5C and RNA:m5U methyltransferases. Nucleic Acids Res. *32*, 2453–2463.
9. Ny, T., and Björk, G.R. (1980). Cloning and restriction mapping of the trmA gene coding for transfer ribonucleic acid (5-methyluridine)-methyltransferase in Escherichia coli K-12. J. Bacteriol. *142*, 371–379.
10. Urbonavičius, J., Jäger, G., and Björk, G.R. (2007). Amino acid residues of the Escherichia coli tRNA(m5U54)methyltransferase (TrmA) critical for stability, covalent binding of tRNA and enzymatic activity. Nucleic Acids Res. *35*, 3297–3305.
11. Nordlund, M.E., Johansson, J.O.M., von Pawel-Rammingen, U., and Byström, A.S. (2000). Identification of the TRM2 gene encoding the tRNA(m5U54)methyltransferase of Saccharomyces cerevisiae. RNA *6*, 844–860.
12. Powell, C.A., and Minczuk, M. (2020). TRMT2B is responsible for both tRNA and rRNA m^5U-methylation in human mitochondria. RNA Biol. *17*, 451–462.
13. Chang, Y.H., Nishimura, S., Oishi, H., Kelly, V.P., Kuno, A., and Takahashi, S. (2019). TRMT2A is a novel cell cycle regulator that suppresses cell proliferation. Biochem. Biophys. Res. Commun. *508*, 410–415.
14. Guo, G., Wang, H., Shi, X., Ye, L., Yan, K., Chen, Z., Zhang, H., Jin, Z., and Xue, X. (2020). Disease Activity-Associated Alteration of mRNA m^5 C Methylation in $CD4^+$ T Cells of Systemic Lupus Erythematosus. Front. Cell Dev. Biol. *8*, 430.
15. Wang, Y., Pang, C., Li, X., Hu, Z., Lv, Z., Zheng, B., and Chen, P. (2017). Identification of tRNA nucleoside modification genes critical for stress response and development in rice and Arabidopsis. BMC Plant Biol. *17*, 261.
16. Carter, J.M., Emmett, W., Mozos, I.R., Kotter, A., Helm, M., Ule, J., and Hussain, S. (2019). FICC-Seq: a method for enzyme-specified profiling of methyl-5-uridine in cellular RNA. Nucleic Acids Res. *47*, e113.
17. Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., Su, J., de Magalhães, J.P., Rigden, D.J., and Meng, J. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res. *47*, e41.
18. Zhou, Y., Zeng, P., Li, Y.H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. *44*, e91.
19. Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. RNA *25*, 205–218.
20. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2015). iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. *490*, 26–33.
21. Qiang, X., Chen, H., Ye, X., Su, R., and Wei, L. (2018). M6AMRFS: Robust prediction of n6-methyladenosine sites with sequence-based features in multiple species. Front. Genet. *9*, 495.
22. Li, J., Huang, Y., Yang, X., Zhou, Y., and Zhou, Y. (2018). RNAm5Cfinder: A Web-server for Predicting RNA 5-methylcytosine (m5C) Sites Based on Random Forest. Sci. Rep. *8*, 17299.
23. Lv, H., Zhang, Z.M., Li, S.H., Tan, J.X., Chen, W., and Lin, H. (2020). Evaluation of different computational methods on 5-methylcytosine sites identification. Brief. Bioinform. *21*, 982–995.
24. Zhang, M., Xu, Y., Li, L., Liu, Z., Yang, X., and Yu, D.J. (2018). Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. Anal. Biochem. *550*, 41–48.
25. Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.
26. He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. BMC Bioinformatics *19*, 306.
27. Li, Y.H., Zhang, G., and Cui, Q. (2015). PPUS: a web server to predict PUS-specific pseudouridine sites. Bioinformatics *31*, 3362–3364.
28. Song, B., Tang, Y., Wei, Z., Liu, G., Su, J., Meng, J., and Chen, K. (2020). PIANO: A Web Server for Pseudouridine-Site (Ψ) Identification and Functional Annotation. Front. Genet. *11*, 88.
29. Liu, K., and Chen, W. (2020). iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics *36*, 3336–3342.
30. Song, B., Tang, Y., Chen, K., Wei, Z., Rong, R., Lu, Z., Su, J., de Magalhães, J.P., Rigden, D.J., and Meng, J. (2020). m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. Bioinformatics *36*, 3528–3536.
31. Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.
32. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.C. (2017). iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. Mol. Ther. Nucleic Acids *7*, 155–163.
33. Feng, P., Ding, H., Chen, W., and Lin, H. (2016). Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. Mol. Biosyst. *12*, 3307–3311.
34. Bari, A.T.M.G., Reaz, M.R., and Choi, H.J.J.B. (2013). DNA Encoding for Splice Site Prediction in Large DNA Sequence. In Database Systems for Advanced Applications. Lecture Notes in Computer Science, B. Hong, X. Meng, L. Chen, W. Winiwarter, and W. Song, eds. (Springer), pp. 46–58.
35. Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). IRNA-2OM: A sequence-based predictor for identifying 2′-O-Methylation Sites in Homo sapiens. J. Comput. Biol. *25*, 1266–1277.
36. Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016). RAMPred: Identifying the N1-methyladenosine sites in eukaryotic transcriptomes. Sci. Rep. *6*, 31080.
37. Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N^6-methyladenosine sites. J. Biomol. Struct. Dyn. *35*, 683–687.
38. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.
39. Baron, G. (2016). Comparison of Cross-Validation and Test Sets Approaches to Evaluation of Classifiers in Authorship Attribution Domain. ISCIS 2016, pp. 81–89.
40. Liu, H., Yue, D., Chen, Y., Gao, S.J., and Huang, Y. (2010). Improving performance of mammalian microRNA target prediction. BMC Bioinformatics *11*, 476.
41. Yu, C.-S., Lin, C.-J., and Hwang, J.-K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. *13*, 1402–1406.
42. Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018). BERMP: a cross-species classifier for predicting m^6A sites by integrating a deep learning algorithm and a random forest approach. Int. J. Biol. Sci. *14*, 1669–1677.
43. James, B., and Yoshua, B. (2012). Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res. *13*, 281–305.

OMTN, Volume 22

Supplemental Information

m5UPred: A Web Server for the Prediction of RNA 5-Methyluridine Sites from Sequences

Jie Jiang, Bowen Song, Yujiao Tang, Kunqi Chen, Zhen Wei, and Jia Meng

Table S1. Overlapped Positive m5U Sites Between Different Techniques and Cell Types

Separation Method	Mode	Condition	Site #	Overlap #	Total #
Technique	Full	miCLIP-Seq	2,225	521	3,696
		FICC-Seq	1,471		
Cell type	Full	HEK293	2,467	732	3,696
		HAP1	1,229		

Table S2. Positive m5U Sites in Different RNAs Families

Gene region	CDS	Intergenic	Intronic	ncRNA_intronic	Ipstream	3'UTR	5'UTR
Number of sites	175	1,880	654	232	277	50	29

Table S3. Performance Evaluation of m5UPred by Cross-technique and Cross-cell Type Validation Using Different Machine Learning Classifiers

	Mode	Classifier	Sn (%)	Sp (%)	ACC (%)	MCC	AUC
Cross-technique validation	Full transcript	SVM	75.87%	85.47%	80.67%	0.616	**0.896**
		RF	69.61%	85.08%	77.35%	0.554	0.870
		NB	80.99%	57.76%	69.37%	0.400	0.785
		GLM	77.51%	81.07%	79.29%	0.586	0.876
	Mature mRNA	SVM	88.48%	89.05%	88.77%	0.775	**0.962**
		RF	81.42%	92.69%	87.05%	0.746	0.955
		NB	91.83%	52.81%	72.32%	0.485	0.855
		GLM	89.63%	84.43%	87.03%	0.742	0.949
Cross-cell type validation	Full transcript	SVM	80.13%	85.98%	83.06%	0.662	**0.918**
		RF	73.05%	85.63%	79.34%	0.592	0.890
		NB	81.64%	59.12%	70.38%	0.420	0.805
		GLM	80.31%	81.64%	80.98%	0.620	0.894
	Mature mRNA	SVM	93.87%	86.15%	90.01%	0.803	**0.970**
		RF	89.15%	92.36%	90.76%	0.816	0.969
		NB	91.54%	53.08%	72.31%	0.483	0.860
		GLM	93.63%	81.04%	87.34%	0.753	0.953

Note: We randomly selected 80% of experimentally validated m5U sites as training dataset and the performance of predictors were evaluated by the rest of 20% of m5U sites as independent testing data.

Table S4. Whole Dataset Performance evaluation

Full	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.960	0.960	0.955	0.953	0.950	0.957	0.950	0.954	0.961	0.956	0.956
MCC	0.784	0.781	0.770	0.755	0.747	0.769	0.748	0.773	0.781	0.763	0.767
SEN	89.72%	87.82%	88.23%	87.69%	86.20%	89.45%	86.60%	87.01%	89.17%	87.14%	87.90%
SPE	88.63%	90.26%	88.77%	87.82%	88.50%	87.42%	88.23%	90.26%	88.90%	89.17%	88.80%
ACC	89.17%	89.04%	88.50%	87.75%	87.35%	88.43%	87.42%	88.63%	89.04%	88.16%	88.35%

Table S5. Whole Dataset Performance evaluation

Mature	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.957	0.959	0.937	0.956	0.960	0.954	0.943	0.960	0.955	0.961	0.954
MCC	0.786	0.794	0.765	0.793	0.773	0.805	0.789	0.814	0.819	0.811	0.795
SEN	86.18%	86.99%	85.37%	89.84%	87.40%	89.02%	87.40%	87.80%	87.40%	86.99%	87.44%
SPE	92.28%	92.28%	91.06%	89.43%	89.84%	91.46%	91.46%	93.50%	94.31%	93.90%	91.95%
ACC	89.23%	89.63%	88.21%	89.63%	88.62%	90.24%	89.43%	90.65%	90.85%	90.45%	89.70%

Table S6. miCLIP_F_train&FICC_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.908	0.915	0.907	0.912	0.912	0.908	0.910	0.916	0.904	0.911	0.910
MCC	0.632	0.650	0.651	0.663	0.649	0.659	0.637	0.683	0.637	0.662	0.652
SEN	74.03%	76.00%	75.80%	75.32%	73.62%	75.39%	74.64%	77.36%	75.12%	76.27%	75.36%
SPE	88.51%	88.51%	88.72%	90.21%	90.35%	89.80%	88.44%	90.35%	88.04%	89.33%	89.23%
ACC	81.27%	82.26%	82.26%	82.77%	81.99%	82.60%	81.54%	83.85%	81.58%	82.80%	82.29%

[a] miCLIP_F dataset as train dataset and FICC_F dataset as an independent dataset

Table S7. miCLIP_M_train&FICC_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.972	0.971	0.967	0.978	0.966	0.969	0.970	0.984	0.961	0.966	0.970
MCC	0.824	0.822	0.822	0.854	0.800	0.792	0.797	0.844	0.743	0.798	0.809
SEN	89.98%	91.93%	91.93%	91.20%	90.22%	89.98%	88.75%	92.42%	86.31%	88.02%	90.07%
SPE	92.42%	90.22%	90.22%	94.13%	89.73%	89.24%	90.95%	91.93%	88.02%	91.69%	90.86%
ACC	91.20%	91.08%	91.08%	92.67%	89.98%	89.61%	89.85%	92.18%	87.16%	89.85%	90.46%

[a] miCLIP_M dataset as train dataset and FICC_M dataset as an independent dataset

Table S8. FICC_F_train&miCLIP_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.852	0.853	0.851	0.852	0.859	0.844	0.860	0.858	0.849	0.851	0.853
MCC	0.487	0.504	0.494	0.493	0.494	0.476	0.509	0.500	0.488	0.502	0.495
SEN	56.58%	57.75%	56.40%	55.87%	56.76%	55.24%	56.45%	58.11%	54.97%	56.63%	56.48%
SPE	89.44%	89.93%	90.07%	90.38%	89.89%	89.48%	91.24%	89.39%	90.61%	90.61%	90.10%
ACC	73.01%	73.84%	73.24%	73.12%	73.33%	72.36%	73.84%	73.75%	72.79%	73.62%	73.29%

[a] FICC_F dataset as train dataset and miCLIP_F dataset as an independent dataset

Table S9. FICC_M_train&miCLIP_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.881	0.884	0.866	0.877	0.868	0.883	0.881	0.857	0.868	0.863	0.873
MCC	0.441	0.460	0.424	0.475	0.438	0.457	0.448	0.432	0.468	0.445	0.449
SEN	40.10%	38.40%	37.67%	42.77%	35.72%	40.58%	36.33%	38.40%	40.10%	38.03%	38.81%
SPE	96.35%	98.42%	96.60%	97.08%	98.42%	97.08%	98.66%	96.72%	98.06%	97.69%	97.51%
ACC	68.23%	68.41%	67.13%	69.93%	67.07%	68.83%	67.50%	67.56%	69.08%	67.86%	68.16%

[a] FICC_M dataset as train dataset and miCLIP_M dataset as an independent dataset

Table S10. HEK293_F_train&HAP1_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.940	0.940	0.933	0.940	0.940	0.947	0.946	0.942	0.943	0.943	0.941
MCC	0.715	0.729	0.721	0.720	0.729	0.737	0.734	0.717	0.720	0.737	0.726
SEN	83.48%	82.18%	82.83%	83.40%	83.65%	82.91%	82.51%	81.61%	80.55%	84.78%	82.79%
SPE	87.96%	90.48%	89.10%	88.53%	89.18%	90.56%	90.64%	89.83%	91.05%	88.85%	89.62%
ACC	85.72%	86.33%	85.96%	85.96%	86.41%	86.74%	86.57%	85.72%	85.80%	86.82%	86.20%

[a] HEK293_F dataset as train dataset and HAP1_F test dataset as an independent dataset

Table S11. HEK293_M_train&HAP1_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.981	0.985	0.981	0.982	0.979	0.981	0.982	0.980	0.978	0.976	0.981
MCC	0.842	0.830	0.863	0.871	0.829	0.848	0.848	0.851	0.827	0.842	0.845
SEN	95.33%	95.88%	94.51%	95.05%	95.05%	95.33%	95.88%	96.15%	95.60%	95.33%	95.41%
SPE	88.74%	86.81%	91.76%	92.03%	87.64%	89.29%	88.74%	88.74%	86.81%	88.74%	88.93%
ACC	92.03%	91.35%	93.13%	93.54%	91.35%	92.31%	92.31%	92.45%	91.21%	92.03%	92.17%

[a] HEK293_M dataset as train dataset and HAP1_M test dataset as an independent dataset

Table S12. HAP1_F_train&HEK293_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.854	0.848	0.860	0.848	0.850	0.857	0.855	0.861	0.861	0.872	0.857
MCC	0.504	0.493	0.505	0.487	0.510	0.509	0.508	0.504	0.524	0.530	0.507
SEN	55.17%	58.94%	56.51%	56.06%	57.80%	57.44%	58.82%	59.02%	59.99%	58.01%	57.77%
SPE	91.77%	88.20%	90.88%	89.83%	90.43%	90.56%	89.50%	89.06%	89.99%	91.85%	90.21%
ACC	73.47%	73.57%	73.69%	72.94%	74.12%	74.00%	74.16%	74.04%	74.99%	74.93%	73.99%

[a] HAP1_F dataset as train dataset and HEK293_F test dataset as an independent dataset

Table S13. HAP1_M_train&HEK293_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.863	0.877	0.870	0.867	0.875	0.869	0.894	0.854	0.861	0.878	0.871
MCC	0.450	0.470	0.456	0.478	0.454	0.463	0.457	0.448	0.446	0.483	0.461
SEN	41.24%	38.02%	39.06%	42.28%	39.06%	40.44%	36.87%	36.52%	37.21%	39.40%	39.01%
SPE	96.31%	99.19%	97.81%	97.58%	97.70%	97.58%	98.96%	98.62%	98.16%	99.31%	98.12%
ACC	68.78%	68.61%	68.43%	69.93%	68.38%	69.01%	67.91%	67.57%	67.68%	69.35%	68.57%

[a] HAP1_M dataset as train dataset and HEK293_M test dataset as an independent dataset

Table S14. Performance Evaluation by FDR and FOR at Different Thresholds

Mode	Threshold	TPR	FDR	FOR
Full Transcript	0.1	0.985	0.322	0.027
	0.2	0.970	0.262	0.043
	0.3	0.958	0.212	0.053
	0.4	0.916	0.181	0.095
	0.5	0.876	0.140	0.127
	0.6	0.838	0.119	0.155
	0.7	0.773	0.089	0.197
	0.8	0.685	0.061	0.248
	0.9	0.463	0.037	0.354
Mature mRNA	0.1	0.972	0.269	0.042
	0.2	0.931	0.196	0.082
	0.3	0.911	0.164	0.098
	0.4	0.862	0.142	0.139
	0.5	0.846	0.107	0.147
	0.6	0.821	0.082	0.162
	0.7	0.776	0.073	0.192
	0.8	0.720	0.048	0.225
	0.9	0.557	0.021	0.310

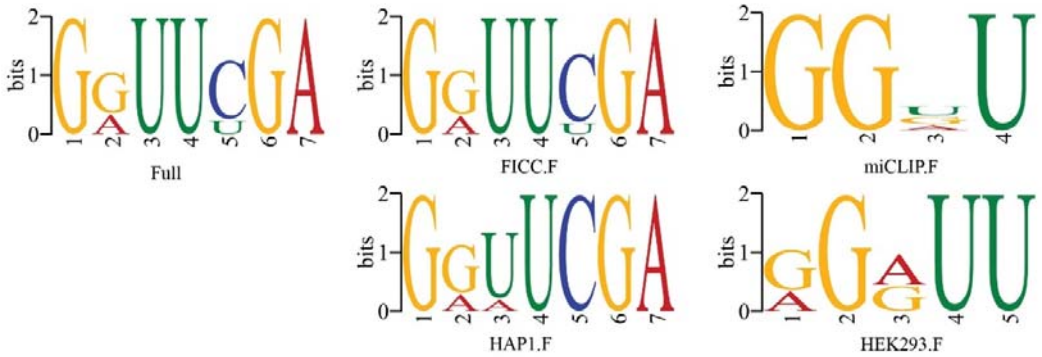

Figure S1. Motif analysis of positive m5U sites generated from different cell types and sequencing methods