RESEARCH

Open Access

m6A-TCPred: a web server to predict tissue-conserved human m^{6}A sites using machine learning approach

Gang Tu[1], Xuan Wang[1,2*], Rong Xia[3] and Bowen Song[4]

*Correspondence: xuan.wang@liverpool.ac.uk

[1] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[2] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L7 8TX, UK
[3] Department of Financial and Actuarial Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[4] Department of Public Health, School of Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China

Abstract

Background: N6-methyladenosine (m^{6}A) is the most prevalent post-transcriptional modification in eukaryotic cells that plays a crucial role in regulating various biological processes, and dysregulation of m^{6}A status is involved in multiple human diseases including cancer contexts. A number of prediction frameworks have been proposed for high-accuracy identification of putative m^{6}A sites, however, none have targeted for direct prediction of tissue-conserved m^{6}A modified residues from non-conserved ones at base-resolution level.

Results: We report here m6A-TCPred, a computational tool for predicting tissue-conserved m^{6}A residues using m^{6}A profiling data from 23 human tissues. By taking advantage of the traditional sequence-based characteristics and additional genome-derived information, m6A-TCPred successfully captured distinct patterns between potentially tissue-conserved m^{6}A modifications and non-conserved ones, with an average AUROC of 0.871 and 0.879 tested on cross-validation and independent datasets, respectively.

Conclusion: Our results have been integrated into an online platform: a database holding 268,115 high confidence m^{6}A sites with their conserved information across 23 human tissues; and a web server to predict the conserved status of user-provided m^{6}A collections. The web interface of m6A-TCPred is freely accessible at: www.rnamd.org/m6ATCPred.

Keywords: m^{6}A modification, Machine learning, Web server, Support vector machine, Gene ontology

Introduction

In recent years, the field of RNA modification has gained significant prominence, with its origins dating back to the groundbreaking proposal in 2008. To date, over 300 distinct RNA modifications have been identified, with more than 170 of them undergoing extensive investigation [1, 2]. These modifications exert profound influences on RNA molecules, impacting their structural conformation, stability, and functional attributes. Notably, the N6-methyladenosine (m^{6}A) modification has garnered substantial attention due to its prevalence within mRNA and its pivotal role in various biological pathways.

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

m^6A modification stands out as the most prevalent and comprehensively studied post-transcriptional modification within mRNA, almost across the entire transcriptome. This prevalence is particularly pronounced in higher eukaryotic organisms. The significance of m^6A modification extends to its far-reaching impact on diverse aspects of biological development, spanning hematopoietic development [3], reproductive processes [4], central nervous system functioning [5] and the regulation of cancer pathways [6]. Consequently, the precise identification of m^6A methylation sites has become increasingly imperative in the realm of biological research.

Currently, researchers mainly employ two high-throughput sequencing techniques to profile the m^6A sites: MeRIP-seq and miCLIP-seq. MeRIP-seq capitalizes on m^6A-specific antibodies to immune-precipitate small RNA fragments following splicing and reverse transcription. Subsequently, the cDNA fragments are sequenced, unveiling the location and extent of m^6A enrichment [7]. Besides MeRIP-seq, miCLIP-seq utilizes UV light-induced cross-linking to introduce specific mutational features, which enables the precise identification of the m^6A residues in RNA molecules [8]. Nonetheless, the reliability of both methods has been tested extensively, revealing susceptibility to multiple influencing factors, leading to occasional inaccuracies and instability. These factors include antibody specificity, domain fusion, and various statistical approaches aimed at mitigating technical noise [9–11]. Beyond these technical concerns, the considerable labor, material, and time costs further intensify the challenges faced by researchers.

To address these challenges, computational databases [12–17] and *in silicon* efforts [18–25] focusing on various biological domains have been developed. For m^6A RNA methylation, tools like *SRAMP* [26] and *iRNA toolkits* [27–29] were established, drawing on a variety of sequence-derived data. The *WHISTLE* [30], emerged as a high-precision predictor, pioneering the integration of domain knowledge and genomic features into m^6A prediction frameworks. More recently, deep learning-based methodologies have also demonstrated their prowess in m^6A detection [31–33].While these advancements have significantly enhanced in silico identification of modified residues. To date, no research has been made to predict conserved m^6A sites across multiple human tissues, despite their established biological significance [34–36]. The ConsRM [37] firstly quantified the conservation degree of base-resolution m^6A sites between human and mouse transcriptomes. The evolutionary conservation in influenza was researched through potential m^6A sites based on DRACH motif [38].

The identification of tissue-conserved m^6A sites assumes paramount importance due to their resistance to interference from extraneous factors and their inherent stability, rendering them invaluable indicators for quantifying m^6A expression levels. Moreover, most existing prediction tools are predominantly based on sequence information and do not incorporate annotations regarding potential post-transcriptional regulatory functions, which are instrumental in elucidating functional consequences. Therefore, the efficacy of these predictors remains subject to limitations.

Here, we present m6A-TCPred, a web server to predict tissue-conserved m^6A sites in human. By learning and testing the m^6A datasets identified from 23 human tissues, the newly integrated framework m6A-TCPred (see Fig. 1) provides a high-accuracy mapping of tissue-conserved human m^6A sites (an average value of AUROC 0.879). Additionally, our results have been integrated into an online platform: a database to hold 268,115 high

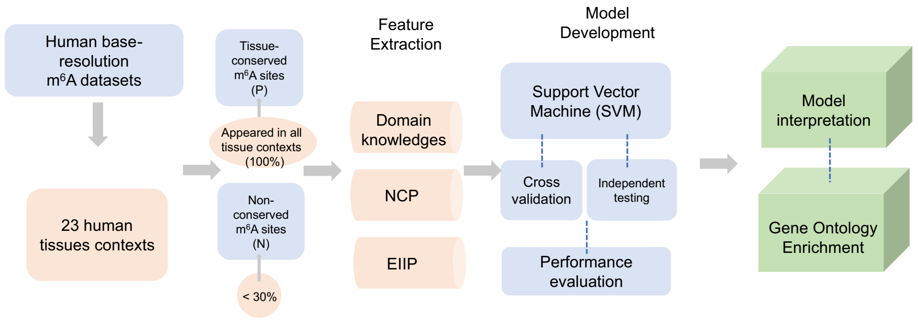

Human base-resolution m⁶A datasets
23 human tissues contexts
Tissue-conserved m⁶A sites (P)
Appeared in all tissue contexts (100%)
Non-conserved m⁶A sites (N)
< 30%
Feature Extraction
Domain knowledges
NCP
EIIP
Model Development
Support Vector Machine (SVM)
Cross validation
Independent testing
Performance evaluation
Model interpretation
Gene Ontology Enrichment

Fig. 1 The framework of m6A-TCPred. The entire datasets from 23 human tissue were further classified into conserved and normal datasets. The model constructed on the genomic and sequence features were trained and evaluated through cross validation and independent testing. The model was further illustrated by interpretation and gene ontology enrichment

confidence m^{6}A sites with their conserved information across 23 human tissues; and a web server to predict the conserved status of user-uploaded m^{6}A datasets. The m6A-TCPred is freely accessible at: www.rnamd.org/m6ATCPred.

Materials and methods

Training and testing dataset

m6A-TCPred was proposed to predict the tissue-conserved m^{6}A methylation sites. The entire datasets were all high-confidence experimentally validated m^{6}A sites (identified in at least two independent studies) collected from m^{6}A-Atlas database [39]. To comprehensively expand the prediction range, a total of 268,115 base-resolution m^{6}A sites were first filtered from m^{6}A-Atlas (with record time > 2). Next, by checking with m^{6}A-containing regions identified from 23 different tissue contexts, we defined the m^{6}A sites simultaneously appeared in all tissue contexts as tissue-conserved m^{6}A sites (dataset P), while the m^{6}A sites appeared in less than 30% of tissue contexts as tissue-specific m^{6}A sites (dataset N). Importantly, the m^{6}A sites that did not appear in any tissue context were excluded to avoid potential bias. Totally, the dataset P contains 10,424 m^{6}A sites, while dataset N includes 54,949 m^{6}A sites across 23 human tissues. To maximize the use of dataset P, the dataset N was further split into 10 sub-datasets and then developed into 10 models in 1:1 Positive-to-Negative ratio with positive datasets to achieve average performance. For performance evaluation, 80% of the dataset was randomly selected as training data, while the rest of 20% was used for independent testing. Please refer to Additional file 1 for the detailed dataset collected to develop the prediction framework.

Feature extraction

Sequence-derived features. Encoding approaches based on sequence information have been broadly applied and achieved good performance in prediction [40–42]. Our new model also adopted the encoding strategy consisting of two parts: Chemical Properties of nucleotides (NCP) and Electron–Ion Interaction Pseudopotential (EIIP).

The first encoding was originally employed in the prediction of DNA sequence splicing site and its efficiency has been confirmed in RNA modification prediction [43, 44]. It depends on the structural differences of chemical properties from ring structure, functional groups and hydrogen bonds. Specifically, one ring structure exists between cytosine and uracil, as well as two ring structures between adenine and guanine. Adenine and cytosine have amino groups, while ketone group is carried between guanine and uracil. Guanine and cytosine, connected by three hydrogen bonds, have stronger binding ability than adenine and cytosine with two hydrogen bonds. As a result of these three structural chemical properties, the encoding of *i*-th nucleotide of given sequence S will be conducted as Vector $S_i = (X_i, Y_i, Z_i)$:

$$X_i = \begin{cases} 1 \;\, if \;\, S_i \in \{A, G\} \\ 0 \;\, if \;\, S_i \in \{C, U\} \end{cases}, \quad Y_i = \begin{cases} 1 \;\, if \;\, S_i \in \{A, C\} \\ 0 \;\, if \;\, S_i \in \{G, U\} \end{cases}, \quad Z_i = \begin{cases} 1 \;\, if \;\, S_i \in \{A, U\} \\ 0 \;\, if \;\, S_i \in \{C, G\} \end{cases} \quad (1)$$

Therefore, A, C, G and U can be encoded by vectors (1,1,1), (0,1,0), (1,0,0) and (0,0,1) respectively.

The Electron–Ion Interaction Pseudopotential (EIIP) was calculated from the delocalized electrons energy from amino acids and nucleotides [45]. This encoding strategy was originally used in the DNA sequences to locate exons and gradually promoted to RNA sequences field [46]. In EIIP method, each nucleotide in RNA sequence was standing for a numeric value to represents its EIIP energy. Specifically, the EIIP for nucleotide A, G, C, U is 0.1260, 0.0806, 0.1340 and 0.1335, respectively.

Genome-derived features. To capture the distinct characteristics of tissue-conserved m^6A sites across human tissues, we extracted 54 additional genomic features from m6ALogisticModel package (generation R code in Additional file 5). These features were selected to accurately represent the topological attributes of the modified residues.1–15 are dummy variable features that indicate whether the tissue-conserved m^6A sites are located within specific transcriptome regions with unique topological properties. All genomic features were generated using the R/Bioconductor package and the hg19 TxDb transcriptome annotation package [47]. In addition, only the primary (longest) transcript of each gene was kept to avoid ambiguities arising from transcript isoforms and extract transcriptional sub-regions [48]. Features 16–19 provide actual valued information on the relative transcript region position, including 3’UTR, 5’UTR, and the whole transcriptome. Sites falling outside these regions are assigned a value of zero. Features 20–27 detail the length of the transcript region containing the modification site, with a value of zero for sites not belonging to the region. Features 28–31 describe the distance between the adenine site and the splicing junction 5’end or 3’end. The distance to the closest tissue-conserved m^6A site in the training data is calculated to quantify the clustering effect of conserved m^6A sites. Features 32 and 33 provide the evolutionary conservation score of the conserved adenosine site and its flanking regions, calculated using Phast-Cons score [49] to assess the conservation level of potential nucleotide sequences. The RNA structure surrounding the conserved adenine site is included in features 34–35, predicted using the RNAfold Vienna RNA package [50]. Features 36–43 illustrate the structural function of m^6A regulatory binding complexes, including readers, writers, and erasers. Features 44–48 encompass genomic properties of genes or transcripts containing conserved m^6A sites. Features 49–51 indicate the z-score of GC content, while features 52–54 provide omics information, such as microRNA target sites. Detailed information of each genomic feature can be found in Table S1 in Additional file 3.

Machine learning approach used for prediction of tissue-conserved m^{6}A site

The Support Vector Machine (SVM) approach is a data-driven machine learning algorithm widely utilized for classification tasks. Notably, SVM outperforms artificial neural networks in scenarios involving extensive genomic data, yielding lower error rates [51]. This technique has been previously applied to identify biomarkers from gene expression data, explore protein interactions [52], pinpoint therapeutic cancer targets, and achieve genome-wide recognition using diverse high-throughput datasets [53].In this research, the prediction model was constructed on R studio interface of LIBSVM [54], with radial basis function serving as kernel. The other parameters were employed by default.

Performance evaluation of tissue-conserved m^{6}A site prediction

To comprehensively evaluate the performance of our predictor, the SVM classifier was examined through fivefold cross-validation on training datasets and tested from 10 independent testing datasets.

To comprehensively evaluate the prediction model performance, five metrics were introduced as the indicators to examine the reliability. Receiver Operating Characteristic (ROC) curve (sensitivity against 1-specificity) and Area Under ROC Curve (AUROC) are the primary performance evaluation indicators. The other four metrics, Sensitivity (Sn), Specificity (Sp), Matthew's Correlation Coefficient (MCC) and Overall Accuracy (ACC) were also employed to quantify and test the model reliability. The entire process was conducted in *R language*.

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

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

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

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

where the TP, TN, FP and FN respectively represent true positive; true negative; false positive and false negative.

Website construction

The predictor website platform is based on Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP), as well as the MySQL tables for metadata storage.

Table 1 Performance evaluation of tissue-conserved m^{6}A sites of using different sequence strategy and algorithms

Sequence Strategy	**Algorithm**	**Independent Testing**				
		Sn	**Sp**	**ACC**	**MCC**	**AUROC**
NCP + ND	SVM	0.598	0.589	0.593	0.187	0.628
	NB	0.655	0.465	0.560	0.123	0.585
	GLM	0.587	0.581	0.584	0.169	0.621
NCP + EIIP	SVM	0.624	0.614	0.619	0.239	0.669
	NB	0.707	0.469	0.585	0.186	0.636
	GLM	0.604	0.619	0.612	0.224	0.660
EIIP + PseKNC	SVM	0.641	0.600	0.620	0.240	0.663
	NB	0.921	0.188	0.555	0.162	0.635
	GLM	0.604	0.606	0.605	0.210	0.648

The SVM (support vector machine) represent binary classification method. NB refers to the naïve bayes classification method. GLM (generalized linear model) is linear regression model. The following sequence encoding strategy, NCP refers to the nucleotide chemical property [43]. ND is nucleotide density [55]. EIIP refers to electron–ion interaction pseudopotential (EIIP) [45]. PseKNC refers to Pseudo K-tuple nucleotide composition [56]

Table 2 Prediction performance using cross validation and independent test dataset

Model	**Testing method**	**Sn**	**Sp**	**ACC**	**MCC**	**AUC**
m6A-TCPred	Cross-validation	0.795	0.789	0.792	0.584	0.871
	Independent testing	0.806	0.796	0.801	0.603	0.879

Results

Determine the best machine learning sequence strategy and algorithm for tissue-conserved m^{6}A site prediction

Although the sequence encoding strategy has already achieved a little reliability in prediction. In order to achieve the best classifier to construct the m6A-TCPred, we combined it with genomic features and tested its performances under different sequence encoding strategies and classifiers on independent dataset (see Table 1). The highest AUROC score is 0.669 when the encoding strategy adopts NCP + EIIP under SVM classifier, representing the best performance among all sequence encoding and algorithm.

Performance evaluation of tissue-conserved m^{6}A predictor by benchmark and independent testing

All the features are normalized and converted into numerical matrix between 0 and 1. The final tissue-conserved m^{6}A prediction model was constructed based on combination of sequence and genomic features.

To comprehensively evaluate the model, ten independent datasets were randomly extracted from negative datasets and integrated with the positive data for average performance evaluation. The prediction model achieved the average value of AUROC 0.879, as well as AUROC 0.871 of fivefold cross-validation (Table 2), suggesting reliability in model distinguishment.

Feature ranking and functional characterization of tissue-conserved m^{6}A sites

The feature ranking illustrates the performance efficiency of all features when model processing, which points out the contribution of various features to identifying tissue-conserved m^{6}A sites. The Fig. 2A listed the top 15 most effective features in predicting tissue-conserved m^{6}A sites, from which we can observed that exon regions may have strong associations in distinguishing tissue-conserved m^{6}A sites from non-conserved ones, especially for long exon regions (> 400 bp). Additionally, feature selection was performed. When using the top 24 genomic features, the model exhibited the best performance with an AUROC of 0.89. As more features were added, the model performance slightly decreased, ranging between AUROC of 0.87 and 0.88 (Figure S1 in Additional file 4). Taken together, this result indicated that feature overfitting has only very limited impact on model performance.

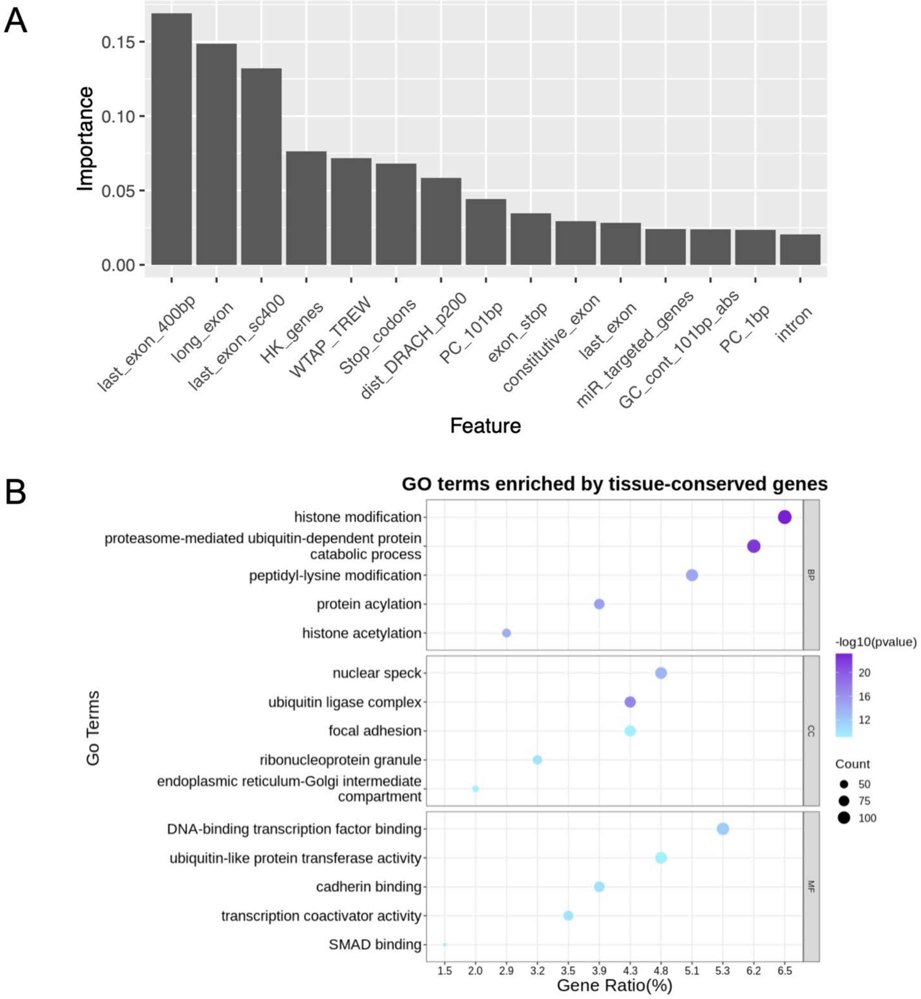


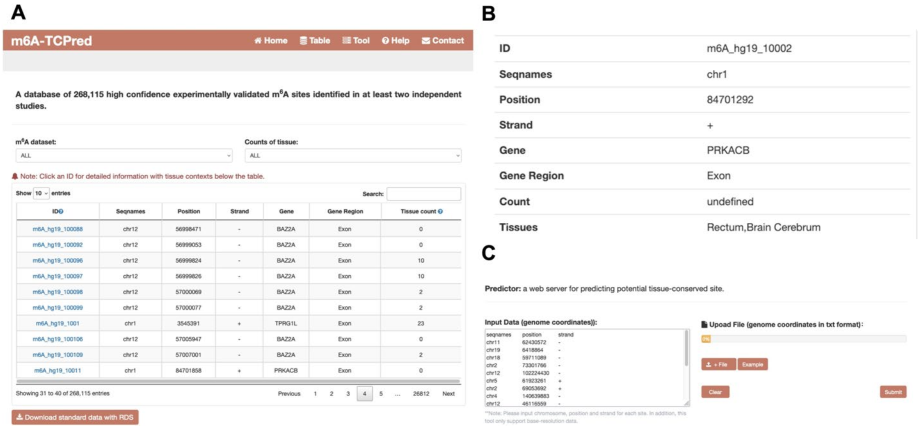
Fig. 2 Model interpretation. **A** The top 15 of most contributing features. Features are output based on machine learning model recognition capabilities. **B** Gene enrichment analysis of the tissue-conserved m^{6}A sites. BP is the biological process; CC is cell component and MF represents molecular function. The gene identification is obtained by using R package RMAnno. Gene ontology analysis is conducted by R package ClusterProfiler [59]

We further examined the biological characterization of the tissue-conserved m^{6}A sites. Specifically, over 260,000 m^{6}A sites were extracted and more than 10,000 of them identified as conserved m^{6}A sites in human tissue for examining their putative functional relevance with Gene Ontology (GO) Enrichment Analysis. In Fig. 2B, we present the top five enriched items in biological pathways, cellular composition, and molecular functions, respectively. Many of these enriched items have previously been confirmed to have strong associations with m^{6}A methylation. Among them, most of them have been studied in several research. For histone modification, the consumption of H3K36me3 has been proved to decrease the abundance of m^{6}A sites [57]. For transcription factor, the m6A-mediated regulation of JUN and JUNB TFs are critical in gene regulation network [58]. A more comprehensive list of the results from the Gene Ontology analysis is available in Additional file 2 (Fig. 2).

Additionally, to further explore the molecular features of tissue-conserved m^{6}A sites from normal ones, we conducted the motif analysis for both of them. The result (Figure S2 in Additional file 4) showed that the sequences of tissue-conserved m^{6}A sites and normal m^{6}A sites follow the pattern of DRACH motif. Meanwhile, no significant differences were observed between tissue-conserved and non-conserved m^{6}A sites, which was consistent with our finding that sequence-based information alone cannot effectively used for classification.

Web server implementation

To enhance the practicality and accessibility of our prediction model, we have developed a user-friendly web server, which is accessible at http://www.rnamd.org/m6ATCPred. Figure 3A exhibits all m^{6}A site datasets incorporated in our model. The selection menus provide users with the flexibility to filter conserved, non-conserved sites, or examine the tissue counts for each site. Detailed information for each site is easily accessible by clicking on the respective ID (see Fig. 3B). Figure 3C illustrates the functionality of our web server, it allows users to upload their m^{6}A coordinates, while the possibility of input data



A database of 268,115 high confidence experimentally validated m^{6}A sites identified in at least two independent studies.

m^{6}A dataset: ALL

Counts of tissue: ALL

Note: Click an ID for detailed information with tissue contexts below the table.

Show 10 entries Search:

ID	Seqnames	Position	Strand	Gene	Gene Region	Tissue count
m6A_hg19_100088	chr12	56998471	-	BAZ2A	Exon	0
m6A_hg19_100092	chr12	56999053	-	BAZ2A	Exon	0
m6A_hg19_100096	chr12	56999824	-	BAZ2A	Exon	10
m6A_hg19_100097	chr12	56999826	-	BAZ2A	Exon	10
m6A_hg19_100098	chr12	57000069	-	BAZ2A	Exon	2
m6A_hg19_100099	chr12	57000077	-	BAZ2A	Exon	2
m6A_hg19_1001	chr1	3545391	+	TPRG1L	Exon	23
m6A_hg19_100106	chr12	57005947	-	BAZ2A	Exon	0
m6A_hg19_100109	chr12	57007001	-	BAZ2A	Exon	2
m6A_hg19_10011	chr1	84701858	+	PRKACB	Exon	0

Showing 31 to 40 of 268,115 entries Previous 1 2 3 4 5 … 26812 Next

ID	m6A_hg19_10002
Seqnames	chr1
Position	84701292
Strand	+
Gene	PRKACB
Gene Region	Exon
Count	undefined
Tissues	Rectum,Brain Cerebrum

Predictor: a web server for predicting potential tissue-conserved site.

Input Data (genome coordinates)):

```
seqnames  position   strand
chr11     62430572   -
chr19     6418864    -
chr18     59711089   -
chr2      73301766   -
chr12     102224430  -
chr5      61923261   +
chr2      69053692   +
chr4      140639883  -
chr12     46116559   -
```

**Note: Please input chromosome, position and strand for each site. In addition, this tool only support base-resolution data.

Upoad File (genome coordinates in txt format):

0%

Fig. 3 The database and web server of m6A-TCPred. **A** The database exhibited the information of all m^{6}A sites. The user can select conserved, non-conserved sites and tissue counts or search according to their own needs. **B** The detailed information of one m^{6}A site. **C** The web server allows users to submit their own m^{6}A coordinates. The m6A-TCPred will predict conservative probability of each site

is evaluated, and each related site is classified into conserved/non-conserved. The free download function is also available.

Conclusion

m^6A methylation stands as one of the most pivotal RNA modifications, with its substantial abundance in mRNA and integral role in biological processes garnering significant attention. Despite this, the precise identification and localization of conserved m^6A sites in human tissues has remained a largely unexplored domain, primarily due to limitations associated with experimental methodologies. In response to these challenges, our research has developed m6A-TCPred, a computational predictor that integrates sequence feature information and genomic features. In contrast to existing predictors, our model excels in the accurate identification of tissue-conserved m^6A sites and their discrimination from non-conserved sites across multiple human tissues. Meanwhile, regrading whether tissue-conserved sites are influenced by housekeeping gene, we evaluated the model performance by adjusting datasets that does not contain housekeeping gene. Our result (Table S2 in Additional file 3) showed there is no significant difference compared with original AUROC. Therefore, the m6A-TCPred didn't have any bias on housekeeping gene.

Our research findings are exciting and the achievements are mainly concentrated on three aspects. Firstly, the m6A-TCPred is a high-accuracy predictor, demonstrating the high efficiency in predicting tissue-conserved m^6A sites. Through fivefold cross-validation and independent testing, our model achieves an impressive AUROC score of 0.879, surpassing the previous limitations associated with sequence encoding. To ensure the broad accessibility of our research, we have integrated the entire model into a user-friendly website. This resource is open to all, enabling individuals to submit genome coordinate files and use our predictor for tissue-conserved m^6A site predictions. We anticipate that this tool will serve as a powerful resource for researchers delving into the intricacies of conserved m^6A sites in human tissue. Secondly, the model ranking and GO analysis provides biological characterizations of tissue-conserved m^6A sites. It identified the potential functional regions with high probability of conserved m^6A sites. The GO analysis provides a connection between conserved m^6A sites and biological functions and some of the content has been mentioned in relevant research. Thirdly, the whole dataset and its relevant annotations were integrated into a website, which is the first collection about tissue-conserved m^6A sites.

It is essential to acknowledge certain limitations. The presence of bias in our training datasets, stemming from inherent limitations in experimental techniques, may influence the model's performance. Further, sample size constraints might not capture the full spectrum of potential outcomes. As we continue our research, we remain committed to improving the model's reliability with the incorporation of the latest sequencing data. Additionally, we recognize that the accuracy of our predictions is contingent on the availability of additional information, such as RNA secondary structure, free energy, RNA type, and more.

Abbreviations

m^6A	N6-methyladenosine
NCP	Nucleotide chemical property

EIIP	Electron–ion interaction pseudopotential
SVM	Support vector machine
Sn	Sensitivity
Sp	Specificity
MCC	Matthews correlation coefficient
ACC	Overall accuracy
AUROC	Area under the ROC curve
GO	Gene ontology
BP	Biological process
CC	Cell component
MF	Molecular function
HTML	Hyper text markup language
CSS	Cascading style sheets
PHP	Hypertext preprocessor

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05738-1.

Additional file 1. Supplementary Table S1. m6A sites distributions in different tissues

Additional file 2. Supplementary Table S2. The detailed information of Gene Ontology analysis

Additional file 3. Table S1. Genome-derived features from m6A datasets. **Table S2.** The model performance of non-HKG m6A sites

Additional file 4. Figure S1. Feature selection of genome-derived features of m6ATCPred. **Figure S2**. The motif analysis of conserved m6A sites. A) The motif of tissue-conserved m6A residues. B) The motif of normal m6A restudies.

Additional file 5. #Genomic Feature genertaion.

Acknowledgements

Not applicable.

Author contributions

X.W initialized this project and collected the dataset. G.T performed the entire research. X.W constructed the website. R.X helped the proofreading. G.T, X.W and B.S wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

XJTLU Key Program Special Fund [KSF-E-51 and KSF-P-02]; Scientific Research Foundation of Nanjing University of Chinese Medicine (Grant No. 013038030001). This work is Supported by the Supercomputing Platform of Xi'an Jiaotong-Liverpool University.

Availability of data and materials

The raw data used to construct m6A-TCPred was collected from the m6A-Atlas v2.0: http://rnamd.org/m6a. The detailed information of m6A-TCPred dataset can be found in Supplementary information Additional file 1 and is freely accessible at: http://www.rnamd.org/m6ATCPred.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 6 November 2023 Accepted: 11 March 2024
Published online: 25 March 2024

References

1. Arzumanian VA, Dolgalev GV, Kurbatov IY, Kiseleva OI, Poverennaya EV. Epitranscriptome: review of top 25 most-studied RNA modifications. Int J Mol Sci 2022;23(22).
2. Liu Q, Chen J, Wang Y, Li S, Jia C, Song J, Li F. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief Bioinf 2021;22(3).
3. Hart SM, Foroni L. Core binding factor genes and human leukemia. Haematologica. 2002;87(12):1307–23.

4. Qi ST, Ma JY, Wang ZB, Guo L, Hou Y, Sun QY. N6-methyladenosine sequencing highlights the involvement of mRNA methylation in oocyte meiotic maturation and embryo development by regulating translation in Xenopus laevis. J Biol Chem. 2016;291(44):23020–6.
5. Hess ME, Hess S, Meyer KD, Verhagen LA, Koch L, Bronneke HS, Dietrich MO, Jordan SD, Saletore Y, Elemento O, et al. The fat mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry. Nat Neurosci. 2013;16(8):1042–8.
6. Liu J. Regulation of gene expression by N6-methyladenosine in cancer. Trends Cell Biol. 2019;29(6):487–99.
7. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012;485(7397):201–6.
8. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods. 2015;12(8):767–72.
9. Hawley BR, Jaffrey SR. Transcriptome-wide mapping of m6A and m6Am at single-nucleotide resolution using miCLIP. Curr Protoc Mol Biol. 2019;126(1): e88.
10. McIntyre ABR, Gokhale NS, Cerchietti L, Jaffrey SR, Horner SM, Mason CE. Limits in the detection of m(6)A changes using MeRIP/m(6)A-seq. Sci Rep. 2020;10(1):6590.
11. Meng J, Lu Z, Liu H, Zhang L, Zhang S, Chen Y, Rao MK, Huang Y. A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. Methods. 2014;69(3):274–81.
12. Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, de Crecy-Lagard V, Ross R, Limbach PA, Kotter A et al: MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res 2018;46(D1):D303-D307.
13. Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, Qu LH, Yang JH: RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res 2018;46(D1):D327–D334.
14. Bao X, Zhang Y, Li H, Teng Y, Ma L, Chen Z, Luo X, Zheng J, Zhao A, Ren J, et al. RM2Target: a comprehensive database for targets of writers, erasers and readers of RNA modifications. Nucleic Acids Res. 2023;51(D1):D269–79.
15. Song B, Wang X, Liang Z, Ma J, Huang D, Wang Y, de Magalhaes JP, Rigden DJ, Meng J, Liu G et al: RMDisease V2.0: an updated database of genetic variants that affect RNA modifications with disease and trait implication. Nucleic Acids Res 2022.
16. Luo X, Li H, Liang J, Zhao Q, Xie Y, Ren J, Zuo Z. RMVar: an updated database of functional variants involved in RNA modifications. Nucleic Acids Res. 2021;49(D1):D1405–12.
17. Wang X, Zhang Y, Chen K, Liang Z, Ma J, Xia R, de Magalhaes JP, Rigden DJ, Meng J, Song B: m7GHub V2.0: an updated database for decoding the N7-methylguanosine (m7G) epitranscriptome. Nucleic Acids Res 2023.
18. Qiu WR, Jiang SY, Xu ZC, Xiao X, Chou KC. iRNAm 5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget. 2017;8(25):41178–88.
19. Chen W, Song X, Lv H, Lin H. iRNA-m2G: identifying N(2)-methylguanosine sites based on sequence-derived information. Mol Ther Nucleic Acids. 2019;18:253–8.
20. Zhai J, Song J, Cheng Q, Tang Y, Ma C. PEA: an integrated R toolkit for plant epitranscriptome analysis. Bioinformatics. 2018;34(21):3747–9.
21. Liang Z, Zhang L, Chen H, Huang D, Song B. m6A-Maize: weakly supervised prediction of m(6)A-carrying transcripts and m(6)A-affecting mutations in maize (Zea mays). Methods 2021.
22. Körtel N, Rücklé C, Zhou Y, Busch A, Hoch-Kraft P, Sutandy FXR, Haase J, Pradhan M, Musheev M, Ostareck D et al. Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. Nucleic Acids Res 2021.
23. Xiong Y, He X, Zhao D, Tian T, Hong L, Jiang T, Zeng J. Modeling multi-species RNA modification through multi-task curriculum learning. Nucleic Acids Res 2021.
24. Wang C, He Z, Jia R, Pan S, Coin LJ, Song J, Li F. PLANNER: a multi-scale deep language model for the origins of replication site prediction. IEEE J Biomed Health Inform 2024.
25. Li F, Fan C, Marquez-Lago TT, Leier A, Revote J, Jia C, Zhu Y, Smith AI, Webb GI, Liu Q, et al. PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. Brief Bioinform. 2020;21(3):1069–79.
26. Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. 2016;44(10): e91.
27. Chen W, Ding H, Zhou X, Lin H, Chou KC. iRNA(m6A)-PseDNC: Identifying N(6)-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem. 2018;561–562:59–65.
28. Chen W, Feng P, Ding H, Lin H, Chou KC. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. 2015;490:26–33.
29. Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics. 2020;36(11):3336–42.
30. Liu L, Song B, Chen K, Zhang Y, de Magalhaes JP, Rigden DJ, Lei X, Wei Z. WHISTLE server: a high-accuracy genomic coordinate-based machine learning platform for RNA modification prediction. Methods 2021.
31. Zou Q, Xing P, Wei L, Liu B. Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. RNA. 2019;25(2):205–18.
32. Chen Z, Zhao P, Li F, Wang Y, Smith AI, Webb GI, Akutsu T, Baggag A, Bensmail H, Song J. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform. 2020;21(5):1676–96.
33. Huang D, Song B, Wei J, Su J, Coenen F, Meng J: Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data. Bioinformatics 2021.
34. Song B, Chen K, Tang Y, Wei Z, Su J, Magalhães JPd, Rigden DJ, Meng J. ConsRM: collection and large-scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome. Brief Bioinf 2021.

35. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell. 2012;149(7):1635–46.
36. Ma L, Zhao B, Chen K, Thomas A, Tuteja JH, He X, He C, White KP. Evolution of transcript modification by N(6)-methyladenosine in primates. Genome Res. 2017;27(3):385–92.
37. Song B, Chen K, Tang Y, Wei Z, Su J, de Magalhaes JP, Rigden DJ, Meng J. ConsRM: collection and large-scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome. Brief Bioinform 2021;22(6).
38. Bayoumi M, Munir M. Evolutionary conservation of the DRACH signatures of potential N6-methyladenosine (m(6)A) sites among influenza A viruses. Sci Rep. 2021;11(1):4548.
39. Liang Z, Ye H, Ma J, Wei Z, Wang Y, Zhang Y, Huang D, Song B, Meng J, Rigden DJ et al: m6A-Atlas v20: updated resources for unraveling the N6-methyladenosine (m6A) epitranscriptome among multiple species. Nucleic Acids Res;2023.
40. Xiong Y, He X, Zhao D, Tian T, Hong L, Jiang T, Zeng J. Modeling multi-species RNA modification through multi-task curriculum learning. Nucleic Acids Res. 2021;49(7):3719–34.
41. Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N(6)-methyladenosine sites. J Biomol Struct Dyn. 2017;35(3):683–7.
42. Li F, Leier A, Liu Q, Wang Y, Xiang D, Akutsu T, Webb GI, Smith AI, Marquez-Lago T, Li J, et al. Procleave: predicting protease-specific substrate cleavage sites by combining sequence and structural information. Genom Proteom Bioinf. 2020;18(1):52–64.
43. Bari ATMG, Reaz MR, Choi H-J, Jeong B-S. DNA encoding for splice site prediction in large DNA sequence. In: Database Systems for Advanced Applications: 2013// 2013; Berlin, Heidelberg. Springer Berlin Heidelberg: 46–58.
44. 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(11):1266–77.
45. Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation. 2006;1(6):197–202.
46. Jiang J, Song B, Chen K, Lu Z, Rong R, Zhong Y, Meng J. m6AmPred: Identifying RNA N6, 2′-O-dimethyladenosine (m6Am) sites based on sequence-derived information. Methods. 2022;203:328–34.
47. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8): e1003118.
48. Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vågbø CB, Geula S, Hanna JH, Black DL, Darnell JE Jr, Darnell RB. m(6)A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. Genes Dev. 2017;31(10):990–1006.
49. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
50. Lorenz R, Bernhart SH, Honer Z, Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6:26.
51. Byvatov E, Schneider G. Support vector machine applications in bioinformatics. Appl Bioinf. 2003;2(2):67–77.
52. Chen L, Xuan J, Riggins RB, Clarke R, Wang Y. Identifying cancer biomarkers by network-constrained support vector machines. BMC Syst Biol. 2011;5:161.
53. Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, Moffat J, Kim PM. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Med. 2014;6(7):57.
54. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2(3):Article 27.
55. Jiang J, Song B, Tang Y, Chen K, Wei Z, Meng J. m5UPred: a web server for the prediction of RNA 5-methyluridine sites from sequences. Mol Ther Nucleic Acids. 2020;22:742–7.
56. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014;456:53–60.
57. Huang H, Weng H, Zhou K, Wu T, Zhao BS, Sun M, Chen Z, Deng X, Xiao G, Auer F, et al. Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. Nature. 2019;567(7748):414–9.
58. Suphakhong K, Terashima M, Wanna-Udom S, Takatsuka R, Ishimura A, Takino T, Suzuki T. m6A RNA methylation regulates the transcription factors JUN and JUNB in TGF-beta-induced epithelial-mesenchymal transition of lung cancer cells. J Biol Chem. 2022;298(11): 102554.
59. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.