Journal Pre-proofs

Database

m6A-TSHub: Unveiling the Context-specific m^{6}A Methylation and m6A-affecting Mutations in 23 Human Tissues

Bowen Song, Daiyun Huang, Yuxin Zhang, Zhen Wei, Jionglong Su, João Pedro de Magalhães, Daniel J. Rigden, Jia Meng, Kunqi Chen

PII: S1672-0229(22)00114-0
DOI: https://doi.org/10.1016/j.gpb.2022.09.001
Reference: GPB 661

To appear in: *Genomics, Proteomics & Bioinformatics*

Received Date: 12 October 2021
Revised Date: 19 August 2022
Accepted Date: 2 September 2022

Please cite this article as: B. Song, D. Huang, Y. Zhang, Z. Wei, J. Su, J. Pedro de Magalhães, D.J. Rigden, J. Meng, K. Chen, m6A-TSHub: Unveiling the Context-specific m^{6}A Methylation and m6A-affecting Mutations in 23 Human Tissues, *Genomics, Proteomics & Bioinformatics* (2022), doi: https://doi.org/10.1016/j.gpb.2022.09.001

m6A-TSHub: Unveiling the Context-specific m^{6}A Methylation and m6A-affecting Mutations in 23 Human Tissues

Bowen Song[1,2,3], Daiyun Huang[4,5,*], Yuxin Zhang[4], Zhen Wei[4,6], Jionglong Su[7], João Pedro de Magalhães[6], Daniel J. Rigden[3], Jia Meng[3,4,8], Kunqi Chen[1,*]

[1]*Key Laboratory of Gastrointestinal Cancer (Fujian Medical University), Ministry of Education, School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350004, China*

[2]*Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

[3]*Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom*

[4]*Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

[5]*Department of Computer Science, University of Liverpool, Liverpool L69 7ZB, United Kingdom*

[6]*Institute of Ageing & Chronic Disease, University of Liverpool, Liverpool L69 7ZB, United Kingdom*

[7]*School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

[8]*AI University Research Centre, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

[*] Corresponding authors.

E-mail: kunqi.chen@fjmu.edu.cn (Chen K), daiyun.huang@liverpool.ac.uk (Huang D).

Running title: *Song B et al / m6A-TSHub for Context-specific m^{6}A Epitranscriptome*

Total word counts (from 'Introduction' to 'Discussion'): 5589

Total figures: 6

Total tables: 3

Total supplementary figures: 2

Total supplementary tables: 1

Total supplementary files: 4

Abstract

As the most pervasive epigenetic marker present on mRNA and lncRNA, N^6-methyladenosine (m^6A) RNA methylation has been shown to participate in essential biological processes. Recent studies revealed the distinct patterns of m^6A methylome across human tissues, and a major challenge remains in elucidating the tissue-specific presence and circuitry of m^6A methylation. We present here a comprehensive online platform m6A-TSHub for unveiling the context-specific m^6A methylation and genetic mutations that potentially regulate m^6A epigenetic mark. m6A-TSHub consists of four core components, including (1) m6A-TSDB: a comprehensive database of 184,554 functionally annotated m^6A sites derived from 23 human tissues and 499,369 m^6A sites from 25 tumor conditions, respectively; (2) m6A-TSFinder: a web server for high-accuracy prediction of m^6A methylation sites within a specific tissue from RNA sequences, which was constructed using multi-instance deep neural networks with gated attention; (3) m6A-TSVar: a web server for assessing the impact of genetic variants on tissue-specific m^6A RNA modification; and (4) m6A-CAVar: a database of 587,983 TCGA cancer mutations (derived from 27 cancer types) that were predicted to affect m^6A modifications in the primary tissue of cancers. The database should make a useful resource for studying the m^6A methylome and genetic factor of epitranscriptome disturbance in a specific tissue (or cancer type). m6A-TSHub is accessible at: www.xjtlu.edu.cn/biologicalsciences/m6ats.

KEYWORDS: N^6-methyladenosine (m^6A); Context-specific analysis; Cancer mutations; Genome analysis; Functional annotation

Introduction

Among the more than 150 distinct chemical modifications naturally decorating cellular RNAs [1], N^6-methyladenosine (m^6A) is the most pervasive marker present on mRNA and lncRNA, and has been associated with a number of essential biological functions and processes [2,3], including mRNA stability [4], splicing [5], translation [6,7], heat shock [8], DNA damage [9], and embryonic development [10]. Increasing evidence has indicated a critical role of m^6A dysregulation in various human diseases, especially multiple cancers, such as breast cancer [11,12] and prostate cancer [13]. For example, inhibition of an m^6A methyltransferase (*METTL13*) could be used as a potential therapeutic strategy against acute myeloid leukemia [14].

Developed in 2012, m^6A-seq (MeRIP-seq) was the first whole transcriptome m^6A profiling approach [15,16]. It relies on antibody-based enrichment of the m^6A signal, enabling the identification of m^6A-containing regions with a resolution of around 100 nt. Currently, m^6A-seq is still the most popular m^6A profiling approach and has been applied in more than 30 different organisms. Besides m^6A-seq, recent advances in integration of Ultra Violet cross-linking, enzymatic activity and domain fusion have offered improved even base-resolution m^6A detection through techniques such as, miCLIP/m^6A-CLIP-seq [17,18], m^6A-REF-seq [19] and DART-seq [20]. However, compared with m^6A-seq, these approaches require more complicated experimental procedures and have therefore been applied in fewer biological contexts.

To date, more than 120 computational approaches have been developed for the computational identification of RNA modifications [21,22] from the primary RNA sequences. These include the iRNA toolkits [23−31], MultiRM [32], DeepPromise [22], RNAm5CPred [33], SRAMP [11], Gene2vec [34], PEA [35], PPUS [36], WHISTLE [37], m5UPred [38], WeakRM frameworks [39,40], m6ABoost [41], PULSE [42], m6AmPred [43], BERMP [44], and MASS [45]. Together, these efforts have greatly advanced our understanding of multiple RNA modifications in different RNA regions and in various species (see recent reviews [22, 46–48]). A number of epitranscriptome databases have been constructed. MODOMICS collects the pathways related to more than 150 different RNA modifications [1]. RMBase [49], m5C-Atlas [50], and m^6A-Atlas [51] assembled millions of experimentally validated m^6A and m^5C sites. REPIC was established as a comprehensive atlas for exploring the association between m^6A RNA methylation and chromatin modifications [52]. ConsRM provides the conservation score of individual m^6A sites at the base resolution, which can be used to differentiate the functionally important and 'passenger' m^6A sites [53]. m6A2Target compiled the target molecules of m^6A methyltransferases, demethylases, and binding proteins [54]. This work has extended our knowledge of the functional epitranscriptome, and greatly facilitated relevant research. Special efforts have also been made to explore the effects of genetic variants on RNA modifications and their association with various

diseases. m6AVar [55] was the first database that focused on the genetic factors related to epitranscriptome disturbance. It documented more than 400,000 m^6A-affecting genetic variants, which were further labeled with disease and phenotype associations identified from genome-wide association studies (GWAS) analysis. This prediction framework was improved and later applied to eight other RNA modifications (m^5C, m^1A, m^5U, Ψ, m^6Am, m^7G, and 2'-O-Me, and A-to-I) by RMVar [56] and RMDisease [57]. These above databases systematically revealed the general association between epitranscriptome layer dysregulation and various diseases (see a recent review [58]).

Existing computational approaches for epitranscriptome analysis have been quite successful in providing a large quantity of useful information; however, most of them failed to consider the tissue-specificity of m^6A epi-transcriptome [59,60]. Indeed, a recent study by Liu et al. unveiled distinct tissue-specific signatures of the m^6A epitranscriptome in human and mouse [61], which are induced by context-specific expression of m^6A regulators (methyltransferases, demethylases, and RNA binding proteins) [62] and genetic drivers [63]. Nevertheless, most existing approaches for RNA modification site prediction completely ignore the context-specificity of the epitranscriptome and simply assume a single model for different tissues, undermining their accuracy and applicability. To the best of our knowledge, the only three approaches that clearly support the identification of tissue-specific m^6A methylation are im6A-TS-CNN [64], iRNA-m6A [65], and TS-m6A-DL [66], all covering only three human tissue types (brain, liver, and heart). Similarly, when screening for the genetic variants that can affect RNA modifications, previous work assumes a consistent influence in different tissues (see Table S1 for a detailed description and comparison). However, since different epitranscriptome patterns were observed among different tissues, genetic mutations that can alter m^6A methylation in one tissue may not necessarily function similarly in a different tissue. Likewise, there are significant differences in incidence, mortality and molecular signatures across cancer originating from different tissues [67,68]. It is therefore highly desirable to develop approaches that could take full advantage of the tissue-specific RNA methylation profiles so as to make more reliable predictions with respect to a specific tissue type [69]. This is particularly critical for studying the epitranscriptome circuits of diseases that are explicitly associated with a specific tissue, such as, cancers.

To address this issue, we present here a comprehensive online platform m6A-TSHub for unveiling the context-specific m^6A methylation and m^6A-affecting mutations in 23 human tissues. m6A-TSHub consists of four core components, (1) m6A-TSDB: a database for 184,554 experimentally validated m^6A-containing peaks (m^6A sites) derived from 23 distinct human normal tissues and 499,369 m^6A-containing peaks (m^6A sites) from 25 matched tumor conditions, extracted from 233 m^6A-seq samples, respectively. (2) m6A-TSFinder: an integrated online server for the prediction of tissue-specific m^6A

modifications in 23 human tissues, built upon a gated attention-based multi-instance deep neural network. (3) m6A-TSVar: a web server for systemically assessing the tissue-specific impact of genetic variants on m^{6}A RNA modification in 23 human tissues. (4) m6A-CAVar: a database of 587,983 The Cancer Genome Atlas (TCGA) cancer mutations (derived from 27 cancer types) that may lead to the gain or loss of m^{6}A sites in the corresponding cancer-originating tissues.

In addition, the m^{6}A-associated variants were also annotated with their potential post-transcriptional regulatory roles, including RNA binding protein (RBP) binding regions, microRNA targets, and splicing sites, along with their known disease and phenotype linkage integrated from GWAS catalog [70] and ClinVar databases [71]. The m6A-TSHub is freely accessible at www.xjtlu.edu.cn/biologicalsciences/m6ats, and should be a useful resource for studying the m^{6}A methylome and genetic basis of epitranscriptome disturbance with respect to a specific cancer type or tissue. The overall design of m6A-TSHub is shown in **Figure 1**.

Data collection and processing

Data resource (m6A-TSDB)

We collected the epitranscriptome profiles of 23 healthy human tissues, from which the tissue-specific RNA methylation patterns were learned using deep neural networks. Specifically, the raw sequencing data of 78 m6A-seq samples were downloaded directly from Gene Expression Omnibus (GEO) repository of National Center for Biotechnology Information (NCBI) [72] and National Genomics Data Center (NGDC) [73] (Table S2). Adaptors and low-quality nucleotides were removed by Trim Galore [74], followed by quality control using FastQC. The processed reads were then aligned to the reference genome GRCh37/hg19 by HISAT2 [75]. The m^{6}A enriched regions (peaks) located on transcripts were detected by exomePeak2 [76] using its default setting with GC contents corrected. In total, m^{6}A profiling samples from 23 healthy human tissues (184,554 m^{6}A-containing peaks) were processed. We filtered all obtained m^{6}A enriched regions to retain peaks with at least one DRACH consensus motif and used these peak regions containing tissue-specific m^{6}A signals as positive data. Negative data was randomly collected from non-peak regions located on the same transcript of the corresponding positive data, and cropped to balance the length and number between positive and negative regions (with a positive to negative ratio of 1:1). The genomic sequences of both positive and negative regions were then extracted for developing the tissue-specific m^{6}A prediction model.

To evaluate the effect of cancer somatic variants on m^{6}A methylation in their originating tissues, a total of 2,587,191 cancer somatic variants from 27 different cancer types were obtained from TCGA

(release version v27.0-fix) [77] (Table S3). Meanwhile, 155 m^{6}A-seq samples profiling the epitranscriptome (499,369 m^{6}A-containing peaks) of 25 cancer cell lines (corresponding to 17 tissue types) were also obtained using the same data processing pipeline (Table S2), which were used for the validation of the predicted effects on m^{6}A methylation of the variants (detailed in the following).

Learning tissue-specific m^{6}A methylation with deep neural networks (m6A-TSFinder)

The purpose of weakly supervised learning is to develop predictive models by learning from weakly labeled data, such as m^{6}A peaks of low resolution detected by the m6A-seq (or MeRIP-seq) technique [15,16]. Unlike supervised learning based on single-nucleotide resolution data, it works for the case where only coarse-grained labels (indicating whether a genome bin contains an m^{6}A site) are available for these peaks of various lengths. We previously proposed a general weakly supervised learning framework WeakRM [78], which takes labels at the sequence level (rather than a nucleotide level) as input and predicts the sub-regions that are most likely to contain the RNA modification. As a simplified illustration shown in **Figure 2**, the m6A-TSFinder framework is divided into several sub-sections. Firstly, multi-instance learning treats each entire RNA sequence as a 'bag', with multiple 'instances' within the 'bag' determined by a fixed-length sliding window. Previous studies have shown that a 40−50 nt context region is sufficient for modification predictions. Therefore, in m6A-TSFinder, a sliding window of 50 nt was used, which was also helpful in improving the prediction resolution. Secondly, the RNA instances were fed into the m6A-TSFinder model using one-hot encoding, which is widely used in deep learning-based models. The extracted instances pass through the same feature extraction module (the weights of the network are shared in this module) and output instance-level features. The network architecture of the feature extraction section used in m6A-TSFinder includes the first convolutional layer to capture motifs, a max-pooling layer to remove weak features and enlarge the receptive field, a dropout layer that prevents overfitting in training, and a second convolutional layer which learns local dependencies among motifs. In order to further improve the performance of the model, in m6A-TSFinder, we use a long short-term memory (LSTM) layer to replace the second convolutional layer, so that the model can learn the long-range dependence of the motif while maintaining local dependence. Lastly, gated attention was used as the score function to obtain bag-level probabilities from multiple instance-level features. The gated attention module consists of three fully connected layers. The first two layers learn hidden representations from the instance features using tanh and sigmoid activation functions. Their element-wise multiplication is then sent to the third fully connected layer, which learns the similarity between the product and a context feature vector and outputs an attention score for each instance. The score is further normalized using

the softmax function, so that the weights of all instances add up to 1. The weighted summation of instance features is treated as the bag-level feature and used to output the final probability score. Together, our model can be trained end-to-end using the binary cross-entropy loss calculated by the bag-level label. Our model was trained using the Adam optimizer under the Tensorflow framework. The learning rate was initially set to 1E−4, and gradually decayed to 1E−5 during the training process of 20 epochs. It is worth mentioning that when the number of instances is consistently set to 1, the weight of the instance is always 1, and the label becomes the instance level. In that case, the gated attention module is degraded, and the network becomes a strong supervised learning framework with two feature extraction layers.

Decoding the tissue-specific effect of variants on m^{6}A methylation (m6A-TSVar & m6A-CAVar)

Similar to previous studies [55,56,79,80], a cancer somatic variant is defined as a tissue-specific m^{6}A variant if it could lead to the gain or loss of m^{6}A methylation in a specific tissue. The tissue-specific inference was made possible by our deep neural network model m6A-TSFinder. Specifically, the predicted tissue-specific m^{6}A variants were further classified into three confidence levels − low, medium, and high (**Figure 3**).

Low confidence level

An m^{6}A-associated variant with a low confidence level was defined directly by the tissue-specific prediction model. For example, a synonymous somatic variant (Chr5:92929473, positive strand, C>T, TCGA barcode: TCGA-49-6742-01A-11D-1855-08) was extracted from The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) project, which was then predicted to eliminate the methylation of an experimentally validated m^{6}A-containing region (Chr5:92929314−92929786, positive strand) originally detected in human lung tissue [61].

Medium confidence level

The m^{6}A variants in medium confidence level are those that can be verified on available epitranscriptome data from cancer samples originating from the matched tissue. Following the low confidence level mentioned above, by checking the m^{6}A-containing regions reported in lung adenocarcinoma cancer cell line A549 [81] and H1299 [82], we confirmed that no m^{6}A peaks were further observed in A549 and H1299 for the variant-affected region (Chr5:112176059−112176334,

positive strand). Consequently, this LUAD somatic variant was upgraded to a 'medium' confidence level in the m6A-CAVar database. It is worth noting that the predicted m^{6}A dynamics in m6A-CAVar were systematically validated using available epitranscriptome datasets from the matched healthy and cancer samples, providing another layer of quality assurance from real omics datasets: existing approaches only use those datasets to provide the m^{6}A site information without searching for potential evidence of m^{6}A status switching.

High confidence level

Only a very small number of variants have been clearly associated with diseases and phenotypes unveiled from GWAS analysis, and are known as disease-TagSNPs. These variants exhibited their clinical significance and are very likely to be functionally important. Thus, m^{6}A variants of 'high' confidence level were defined as the validated m^{6}A variants that can also be mapped to disease-TagSNPs extracted from ClinVar [71] and GWAS catalog [70], while those not validated were referred to as 'critical'.

Additionally, the association level (AL) between an SNP and m^{6}A RNA modification was defined as follows:

$$\mathrm{AL} = \begin{cases} 2P_{SNP} - 2\max\left(0.5, P_{WT}\right) & \text{for gain} \\ 2P_{WT} - 2\max\left(0.5, P_{SNP}\right) & \text{for loss} \end{cases} \quad (1)$$

where P_{WT} and P_{SNP} represent the probability of m^{6}A RNA modification for the wild-type and mutated sequences, respectively. The AL ranges from 0 to 1, with 1 indicating the maximum impact on m^{6}A methylation. The statistical significance was assessed by comparing the ALs of all mutations, with which the upper bound of the *P* value can be calculated from its absolute ranking. The m^{6}A-associated variants with AL > 0.4 and $P < 0.1$ were retained. We also considered the possibility of a variant destroying a part of (but not an entire) m^{6}A peak. For peaks wider than 500 nt, the impacts were also evaluated on the 200 nt flanking regions of the variant.

The predicted m^{6}A variants were then validated on the epitranscriptome datasets from the matched health and cancer samples. We consider a prediction validated by omic data if the matched dynamics of m^{6}A sites were observed under the healthy tissue and the cancer samples with the same tissue origin. It may be worth noting that, omic data was only used to inform the prediction of m^{6}A sites in previous studies [55,56,79,80]; however, our analysis also relies on it to confirm the predicted

disturbance of m^{6}A status between the health and cancer conditions. This extra layer of confirmation directly from available omic datasets should effectively enhance the reliability of our database.

Functional annotation

The identified m^{6}A variants were annotated with various information, including transcript region (coding sequence, three prime untranslated region, five prime untranslated region, start codon, and stop codon), gene annotation (gene symbol, gene type, and Ensembl gene ID), evolutionary conservation (phastCons 60-way), deleterious level by SIFT [83], PolyPhen2 HVAR [84], PolyPhen2HDIV [84], LRT [85] and FATHMM [86] using the ANNOVAR package [87], absolute ranking by comparing to the ALs of all mutations (top 1% and top 5%), and TCGA sample information (TCGA case ID, TCGA barcode, TCGA sample count, and sample total variant number). A total of 177,998 high-confidence m^{6}A sites detected using base-resolution technology previously were collected and used to pinpoint the precise location of the mediated m^{6}A sites within the variant-affected regions (Table S4). In addition, aspects of the post-transcriptional machinery that can be mediated by m^{6}A methylation were also annotated, including RBP binding regions from POSTAR2 [88], miRNA-RNA interaction from miRanda [89] and starBase2 [90], and splicing sites from UCSC [91] annotation with GT-AG role. Furthermore, to unveil potentially related pathogenesis, any association between disease and m^{6}A variants was extracted from the GWAS catalog [70] and ClinVar [71] databases.

Database and web interface implementation

Hyper text markup language (HTML), cascading style sheets (CSS), and hypertext preprocessor (PHP) were applied to construct the m6A-TSHub web interface. All metadata was stored using MySQL tables. Besides, EChars was exploited to present statistical diagrams, and the Jbrowse genome browser [92] was included for interactive exploration and visualization of relevant records for genome regions of interest.

Database content and usage

Collection of m^{6}A sites from 23 normal human tissues and 25 cancer cell lines in m6A-TSDB

In m6A-TSDB, a total of 184,554 and 499,369 m^{6}A-containing peaks were collected from 23 normal human tissues and 25 cancer samples, respectively. Among them, 17 out of 25 tumor samples have

the m^{6}A profiles of their matched primary tissues. The m^{6}A enriched peaks were called using exomePeak2 [76] with GC-correction function after mapping the processed reads to human reference genome version hg19. It is worth mentioning that, for a more complete m^{6}A epitranscriptome landscape view, a total of 177,998 base-resolution m^{6}A sites collected from 27 datasets using six different m^{6}A profiling techniques were integrated and used to pinpoint the precise location of the mediated m^{6}A sites within all tissue-specific m^{6}A peaks (Table S4). In addition, all m^{6}A-containing peaks were labeled with information showing whether these sites were affected by cancer somatic variants and potential involved post-transcriptional regulations. All data collected in the m6A-TSDB can be freely downloaded or shared.

Performance evaluation and model interpretation of tissue-specific m^{6}A site prediction m6A-TSFinder

The performance of tissue-specific m^{6}A site predictors was evaluated using 10-fold cross-validation and independent testing. For each distinct human tissue, we randomly selected 15% of experimentally validated m^{6}A sites and used them as an independent testing dataset. For 10-fold cross-validation, the training data was randomly divided into 10 groups with the same number of positive and negative peaks. The prediction performance of each tissue-specific predictor is shown in **Table 1**. In general, the prediction accuracy for most tissues (20 out of the total 23 tissues) is in line with conventional approaches for m^{6}A site prediction under strong supervision with base-resolution datasets, which typically reported a prediction performance between 0.8 and 0.85 in terms of the area under ROC curve (AUROC) [22,93]. The performance for kidney (AUROC = 0.718), bone marrow (AUROC = 0.757) and brainstem (AUROC = 0.789) was somewhat worse, but the reasons are not very clear. In addition, in order to find the recurring sequence patterns preferred by each tissue-specific m^{6}A prediction model, we further divided the peaks into instances of length (l = 50) and extracted the consensus motifs from instances with predicted values higher than 0.5 using integrated gradient and TF-Modisco, under each tissue model, respectively. By trimming the overall letter frequencies with three gaps and two mismatches allowed, we identified one consistence motif under all tissue models (Figure S1), which was matched to the known m^{6}A consensus motif DRACH. Please refer to Figure S1 for details.

Performance compared with existing approaches

We further compared the performance of the proposed m6A-TSFinder with existing m^{6}A predictors specifically targeted at the tissue level. Dao et al. previously developed an Support Vector Machine

(SVM)-based model (iRNA-m6A) for m^{6}A identification in the human brain, liver, and kidney [65]. Later, im6A-TS-CNN [64] and TS-m6A-DL [66] further improved prediction performance by applying a convolutional neural network, using the same training and testing datasets provided in Dao's work. It is worth mentioning that the training and testing datasets used in their work contain positive and negative sequences fixed to 41 nt length with m^{6}A sites or unmethylated adenosines in the center. These models learn to capture discriminative sequence patterns at positions a fixed distance from the target adenosine. When making predictions, the well-trained models take the centered adenosine and its surrounding sequences and return the probability that the central adenosine is methylated. When only low-resolution data are available, sequence lengths vary from 100 nt to hundreds, and methylation is not fixed at the center of the sequence. Therefore, the pre-set requirements of these base-resolution models (TS-m6A-DL, im6A-TS-CNN, and iRNA-m6A) cannot be fulfilled, making it difficult to fairly evaluate their performance on low-resolution data. Furthermore, the only three tissue-specific base-resolution datasets originate from m6A-REF-seq, which can only detect m^{6}A in NNACA, whereas the 23 low-resolution considered in this work contain m^{6}A from broader sequence contexts. Inconsistencies between data further limit direct comparisons between models. Nevertheless, we apply m6A-TSFinder to the same training and testing datasets of the three base-resolution models to show performance and fair comparisons when base-resolution data is available. Specifically, as described in the data collection and processing section, the prediction of m^{6}A from fixed-length sequences centered at the target site can be considered a special case of m6A-TSFinder, where each input sequence is treated as a single instance. As shown in **Table 2**, when tested on the independent dataset, m6A-TSFinder outperformed the three competing methods in two of the three tissues tested (brain and liver) and achieved the best average performance (AUROC of 0.8593). The improvement may be due to the application of the LSTM layer after the convolutional layer, which enables the model to learn the long-range dependencies between the motifs. In addition, by learning from the low-resolution datasets, we expanded the human tissues supported from 3 to 23, which could significantly facilitate future research focusing on the dynamics of m6A methylome across different tissues.

Assessing the impact of genetic variants on tissue-specific m^{6}A sites by m6A-TSVar

The m6A-TSVar web server was designed to assess the impact of genetic variants on tissue-specific m^{6}A RNA methylation using deep neural networks. The collected experimentally validated m^{6}A peaks from 23 human tissues were integrated. The changes in the probability of m^{6}A methylation affected by mutations were calculated, with the returned value of AL indicating how extreme the impact on

m^{6}A methylation was. To our best knowledge, the m6A-TSVar is the first web server for exploring m^{6}A-affecting variants within a specific tissue by integrating the tissue-specific m^{6}A patterns.

Screening for cancer variants that affect m^{6}A in their primary tissues in m6A-CAVar

In m6A-CAVar, the cancer somatic variants from 27 TCGA projects were extracted. Their impacts on m^{6}A RNA modification in the corresponding 23 healthy human tissues were evaluated and then systematically validated using 17 paired normal and tumor samples. A total of 587,983 cancer somatic variants were predicted to affect the m^{6}A methylation status in their originating tissues (the 'low' confidence level group). Among them, the dynamic m^{6}A status induced by 122,473 variants was observed on the available epitranscriptome profiles (the 'medium' confidence level group), and 1718 confirmed m^{6}A-variants were known to be associated with diseases and other phenotypes from GWAS analysis (the 'high' confidence level group) (**Table 3**). Please refer to data collection and processing for more details related to the definition of different confidence groups.

Deciphering the tissue-specificity of cancer m^{6}A variants

Of interest is whether m^{6}A variants function in different cancer-originating tissues. For this purpose, we calculated the proportion of m^{6}A variants that function in different numbers of tissues, and the results suggested that most m^{6}A-associated cancer variants are tissue- and cancer-specific (93.25%), while only around 1.17% are functional in the originating tissues of more than three types of cancers (**Figure 4**A). The consistency is much higher at the gene level. Only around 16.59% of m^{6}A variant carrying genes are associated with a single tissue. More than 60.29% were shared in more than three tissue types (Figure 4B), suggesting some common epitranscriptome layer circuitry at the gene level in different cancers. We further examined the proportion of shared m^{6}A variant-carrying genes between two different tissues. As shown in Figure 4C, most tissues, *e.g.*, skin and stomach, strongly correlate with each other. However, tissues like the heart, testis, and thyroid showed a rather weak association with other tissues, which may suggest more tissue-specific epitranscriptome circuitry for cancers originating in those tissues.

We finally identified the m^{6}A variant-carrying genes that are associated with the most TCGA cancer types. Only experimentally validated m^{6}A variants (medium confidence level and above) were considered here for a more reliable analysis. Top of the list was *CENPF,* where variants may change its m^{6}A methylation status in the primary tissue of 15 cancer types, followed by *DST*, *MKI67,* and *PLEC*,

which were all related to 14 cancer types (detailed in Table S5). Among them, the roles in epitranscriptome regulation of *CENPF*, *MKI67*, and *PLEC* have been indicated previously in glioblastoma [94], breast cancer [95], and pancreatic cancer [96], respectively.

Enhanced web interface and application

The m6A-TSHub features a user-friendly web interface with multiple useful functions, including databases and online servers, which enable users to fast query databases, upload their own custom jobs, and download all m^{6}A-related information at the tissue level. The collected functional m^{6}A-affecting variants can be queried by a human body diagram according to their primary tissues (**Figure 5**A), as well as by different cancer types along with further filters (*e.g.*, gene type, m^{6}A status, confidence level, and disease association; Figure 5B). The query function also returns several categories of useful information, including TCGA project names [77], tumor-growth tissues, genes, chromosome regions, COSMIC ID [97], and disease phenotypes (Figure 5C). The details of tissue-specific m^{6}A peaks collected in m6A-TSDB (Figure 5D) and cancer m^{6}A-associated variants in m6A-CAVar (Figure 5E) can be viewed by clicking the site or variant ID, along with annotated disease-association regulations (Figure 5F). Furthermore, online servers allow for the identification of m^{6}A sites and m^{6}A-associated variants within user-defined regions, with 23 types of human tissues to be selected (Figure 5G and H). A genome browser is available for interactive exploration of the genome regions of interest, including the human gene annotation track, 23 normal tissue tracks, 25 cancer cell line tracks, single-base m^{6}A epitranscriptome landscape track, and post-transcriptional regulation tracks. All metadata provided in the m6A-TSHub can be freely downloaded, alone with server scripts provided to run the prediction tools locally (required language: R and Python). Users can refer to the 'help' and 'download' page for more detailed guidance and instructions.

Utility case study 1: *PIK3CA* variant in colon cancer

Previous studies have reported that m^{6}A RNA modification plays an important role in colon cancer [61,98−100]. The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) project [77] presented a large number of somatic variants identified from various colon adenocarcinoma samples. However, it is still unclear which single genetic variant may lead to m^{6}A dysregulation. In m6A-CAVar, a somatic variant at Chr3: 178952085 (A > T) on PIK3CA identified from TCGA-COAD project (TCGA barcode: TCGA-AA3821-01A-01W-0995-10) was predicted to erase the m^{6}A methylation of a region (Chr3: 178951888−178952363, positive strand). The m^{6}A methylation was observed in healthy human colon,

but disappeared in the colon adenocarcinoma cancer cell line HCT116 [101]. This somatic variant is also recorded in the COSMIC database from colon tumor samples under the legacy identifier of COSM776, and reported to be associated with 27 submitted interpretations and evidence in the ClinVar database [71], including PIK3CA-related overgrowth spectrum (ClinVar accession: RCV000201235.1), breast adenocarcinoma (ClinVar accession: RCV000014629.5), and pancreatic adenocarcinoma (ClinVar accession: RCV000417557.1). Taken together, these observations strongly support the functional importance of this variant. Additionally, the m^{6}A-associated variant falls within the binding regions of two RNA binding proteins (TARDBP and NUDT21), whose interaction may be regulated by the loss of m^{6}A methylation in the cancer condition, providing some putative downstream regulatory consequences of the variant.

Utility case study 2: *PLEC* variant in glioblastoma

Glioblastoma (GBM) is the most aggressive type of brain tumor and is associated with rising mortality. The roles of m^{6}A regulators in this disease have been previously indicated [102−105]. A somatic cancer variant on *PLEC* was identified from the TCGA-GBM project (TCGA barcode: TCGA-06-5416-01A-01D-1486-08) at Chr8: 144991388 (C > T). This cancer variant was predicted to lead to a gain of an m^{6}A site on a previously un-methylated region in a healthy human cerebrum. Indeed, an m^{6}A site was detected in this region from malignant GBM tumor cell line U-251. This mutation has a record in ClinVar database (ClinVar accession: RCV000177727.1). Screening for potential post-transcriptional regulations revealed that the cancer variant falls within the target binding regions of six RNA binding proteins, including the m^{6}A reader YTHDF1, which are known to bind m^{6}A-containing RNAs and promote cancer stem cell properties of GBM cells [106]. It should be of immediate interest to ask whether the methylation of *PLEC* regulates its interaction with YTHDF1 and other RBPs, and what the functional consequences are.

Utility case study 3: *EGFR* variant in lung cancer

The associations between m^{6}A RNA modification and human lung cancers have been well studied. The m^{6}A eraser *FTO* may be a prognostic factor in The Cancer Genome Atlas Lung Squamous Cell Carcinoma (TCGA-LUSC) [107], and the m^{6}A writer METTL3 regulates *EGFR* expression to promote cell invasion of human lung cancer cells [82]. The m6A-CAVar database can be used to explore the role of m^{6}A variants of *EGFR* in lung cancers. We first search by gene name '*EGFR*' on the front page of the m6A-CAVar database, then filter the results and keep only records related to lung tissue, which retains

a total of 10 cancer m^{6}A-associated variants from two lung cancer types (**Figure 6**A and B). Alternatively, the users can query all recorded m^{6}A-associated variants that function in lung tissue by simply clicking the relevant part from the human body diagram (Figure 6C). More details can be accessed by clicking the variant ID. For example, if we check further details of an m^{6}A-gain variant from the TCGA-LUAD project at Chr7: 55259515 (T > G), we can see that this variant is recorded in the ClinVar database and is relevant to eight disease conditions, including lung cancers (Figure 6D), which may suggest potential cancer pathogenesis originates in the epitranscriptome layer.

Discussion and perspectives

The context-specific expressions and functions of m^{6}A regulations have been repeatedly reported in existing studies [59−63], suggesting the involvement of the tissue-specific m^{6}A methylome in essential biological processes and multiple disease mechanisms. Besides, the associations between RNA methylation levels and the activities of RNA methylation regulators were clearly unveiled, reporting that there exist some condition-specific RNA co-methylation patterns (a group of RNA m^{6}A methylation sites whose methylation levels go up and down together) [108−110]. These co-methylation patterns are enriched by the substrate targets of m^{6}A regulators and thus are probably regulated by specific m^{6}A methyltransferase or demethylase.

Here we present m6A-TSHub, a comprehensive online platform for unveiling the context-specific m^{6}A methylation and m^{6}A-affecting mutations in 23 human tissues and 25 tumor conditions. In m6A-TSHub, a total of 184,554 and 499,369 m^{6}A sites derived from 23 normal human tissues and 25 matched tumor samples were collected (m6A-TSDB), from which some potential patterns for the tissue-specific m^{6}A modification sites were revealed (*e.g.*, heart-enriched gene *RYR2* and *PXDNL*; Figure S2). Based on these collected data, 23 distinct m^{6}A prediction models were built at the tissue level using deep neural networks (m6A-TSFinder). In addition, to elucidate the genetic factor of epitranscriptome dysregulation, m6A-CAVar identified a total of 587,983 cancer somatic mutations that may alter the m^{6}A status in corresponding cancer originating tissues and annotated them with various functional annotations, including features relating to post-transcriptional regulations (RBP binding regions, microRNA targets, and splicing sites), disease and phenotype association, as well as other useful genomic information (transcript structure, phastCons, and deleterious level) to provide a more comprehensive overview. We also provide a web server m6A-TSVar for assessing the effect of genetic variants on m^{6}A methylation in a specific tissue.

While most of the existing approaches for RNA modification site prediction ignore the tissue-specific signatures of m^{6}A methylation, by taking advantage of existing tissue-specific epitranscriptome data, our method can predict the m^{6}A methylation within a specific tissue. Compared with existing approaches for tissue-specific m^{6}A methylation site prediction [64−66], our approach m6A-TSFinder achieved a higher prediction performance (Table 2) and hugely expanded the number of supported tissue types from 3 to 23 (Table 1).

Compared with existing approaches for decoding the epitranscriptome impact of genetic variants, m6A-CAVar has the following two major advantages. First, m6A-CAVar relies on a finer prediction model (m6A-TSFinder) that appreciates the specific pattern of RNA methylomes across different tissues. By directly learning from the epitranscriptome profiles in 23 healthy human tissues, m6A-CAVar is able to evaluate the tissue-specific impact of cancer somatic variants on m^{6}A modification in their originating tissue, providing a more detailed picture of the genome-epitranscriptome association. This improves on existing approaches that ignore the distinct signatures of RNA methylation across different tissues and thus fail to address tissue-specific effects. Second, the predicted m^{6}A dynamics in m6A-CAVar were systematically validated using available epitranscriptome datasets from the matched healthy and cancerous samples, providing another layer of quality assurance from real omics datasets. In contrast, existing approaches use those datasets only to provide the m^{6}A site information without searching for potential evidence of m^{6}A status switching.

To date, epitranscriptome data is still rather scarce. Due to the limited availability of datasets, matched healthy tissue and cancer m^{6}A profiling samples are only available for 14 out of the total 27 cancer types, prohibiting a more thorough validation of the predicted results. Furthermore, a substantial discrepancy has been observed among different RNA modification profiling approaches due to technical biases [111−114], which can produce additional inaccuracy. Currently, context-specific epitranscriptome prediction is only possible for a small number of conditions (cell line, tissue type, treatment) with data [64−66]. However, the m6A-TSHub framework will be further expanded when epitranscriptome datasets are more abundantly available for more comprehensive and less biased screening of context-specific m^{6}A-variants, along with linking the tissue-specific epitranscriptome patterns with other important cancer-associated factors such as human aging [67,115]. Besides, the current version of m6A-TSHub was built on human genome assemble hg19. A LiftOver file from hg19 to hg38 was provided on the ‘download’ page, and the next version of the database will be updated based on the latest genome assembly. Particularly promising is the recent development in Nanopore direct RNA sequencing technology that enables simultaneous identification of multiple RNA modifications with simplified sample preparation procedures [116−124].

Data availability

The data underlying this article are available via www.xjtlu.edu.cn/biologicalsciences/m6ats. The online versions of the m6A-TSFinder and m6A-TSVar web server are available via www.xjtlu.edu.cn/biologicalsciences/m6ats by clicking the 'tool' section. The local version and project codes can be accessed on the 'download' page.

CRediT author statement

Bowen Song: Methodology, Data curation, Software, Visualization, Writing-original draft. **Daiyun Huang:** Software, Supervision. **Yuxin Zhang:** Visualization. **Zhen Wei:** Resources, **Jionglong Su:** Visualization. **João Pedro de Magalhães:** Visualization. **Daniel J. Rigden:** Writing-review & editing. **Jia Meng:** Conceptualization, Supervision, Writing-review & editing, Funding acquisition. **Kunqi Chen:** Conceptualization, Resources, Writing-review & editing, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 32100519 and 31671373), Scientific Research Foundation for Advanced Talents of Fujian Medical University (Grant No. XRCZX2021019), and XJTLU Key Program Special Fund (Grant Nos. KSF-T-01, KSF-E-51, and KSF-P-02). The authors also acknowledge the researcher of the various resources mentioned in the manuscript to share their data, especially for tissue-specific m^6A sequencing data.

ORCID

0000-0002-8586-0573 (Bowen Song)

0000-0002-3067-7165 (Daiyun Huang)

0000-0003-1900-6712 (Yuxin Zhang)

0000-0002-1614-9614 (Zhen Wei)

0000-0001-5360-6493 (Jionglong Su)

0000-0002-6363-2465 (João Pedro de Magalhães)

0000-0002-7565-8937 (Daniel J. Rigden)

0000-0003-3455-205X (Jia Meng)

0000-0002-6025-8957 (Kunqi Chen)

References

[1] Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, et al. MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res 2018;46:D303–7.

[2] Meyer KD, Jaffrey SR. Rethinking m(6)A readers, writers, and erasers. Annu Rev Cell Dev Biol 2017;33:319–42.

[3] Niu Y, Zhao X, Wu YS, Li MM, Wang XJ, Yang YG. N^6-methyl-adenosine (m^6A) in RNA: an old modification with a novel epigenetic function. Genomics Proteomics Bioinformatics 2013;11:8–17.

[4] Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, et al. N6–methyladenosine–dependent regulation of messenger RNA stability. Nature 2014;505:117–20.

[5] Mendel M, Delaney K, Pandey RR, Chen K-M, Wenda JM, Vågbø CB, et al. Splice site m6A methylation prevents binding of U2AF35 to inhibit RNA splicing. Cell 2021;184:3125–42.e25.

[6] Choe J, Lin S, Zhang W, Liu Q, Wang L, Ramirez-Moya J, et al. mRNA circularization by METTL3-eIF3h enhances translation and promotes oncogenesis. Nature 2018;561:556–60.

[7] Barbieri I, Tzelepis K, Pandolfini L, Shi J, Millan-Zambrano G, Robson SC, et al. Promoter-bound METTL3 maintains myeloid leukaemia by m(6)A-dependent translation control. Nature 2017;552:126–31.

[8] Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian SB. Dynamic m(6)A mRNA methylation directs translational control of heat shock response. Nature 2015;526:591–4.

[9] Xiang Y, Laurent B, Hsu CH, Nachtergaele S, Lu Z, Sheng W, et al. RNA m(6)A methylation regulates the ultraviolet-induced DNA damage response. Nature 2017;543:573–6.

[10] Zhong S, Li H, Bodi Z, Button J, Vespa L, Herzog M, et al. MTA is an Arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. Plant Cell 2008;20:1278–88.

[11] 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:e91.

[12] Zhang C, Zhi WI, Lu H, Samanta D, Chen I, Gabrielson E, et al. Hypoxia-inducible factors regulate pluripotency factor expression by ZNF217- and ALKBH5-mediated modulation of RNA methylation in breast cancer cells. Oncotarget 2016;7:64527–42.

[13] Lewis SJ, Murad A, Chen L, Davey Smith G, Donovan J, Palmer T, et al. Associations between an obesity related genetic variant (FTO rs9939609) and prostate cancer risk. PLoS One 2010;5:e13485.

[14] Yankova E, Blackaby W, Albertella M, Rak J, De Braekeleer E, Tsagkogeorga G, et al. Small molecule inhibition of METTL3 as a strategy against myeloid leukaemia. Nature 2021;593:597–601.

[15] Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 2012;485:201–6.

[16] 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;149:1635–46.

[17] 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;12:767–72.

[18] Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vagbo CB, Geula S, et al. 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;31:990–1006.

[19] Zhang Z, Chen LQ, Zhao YL, Yang CG, Roundtree IA, Zhang Z, et al. Single-base mapping of m(6)A by an antibody–independent method. Sci Adv 2019;5:eaax0250.

[20] Meyer KD. DART-seq: an antibody-free method for global m(6)A detection. Nat Methods 2019;16:1275–80.

[21] Liu L, Song B, Ma J, Song Y, Zhang SY, Tang Y, et al. Bioinformatics approaches for deciphering the epitranscriptome: recent progress and emerging topics. Comput Struct Biotechnol J 2020;18:1587–604.

[22] Chen Z, Zhao P, Li F, Wang Y, Smith AI, Webb GI, et al. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform 2020;21:1676–96.

[23] Qiu WR, Jiang SY, Sun BQ, Xiao X, Cheng X, Chou KC. iRNA–2methyl: identify RNA 2'-O-methylation sites by incorporating sequence–coupled effects into general PseKNC and Ensemble classifier. Med Chem 2017;13:734–43.

[24] 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;25:1266–77.

[25] 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;561–62:59–65.

[26] 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.

[27] Qiu WR, Jiang SY, Xu ZC, Xiao X, Chou KC. iRNAm5C–PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget 2017;8:41178–88.

[28] 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.

[29] Chen W, Feng P, Song X, Lv H, Lin H. iRNA–m7G: identifying N(7)-methylguanosine sites by fusing multiple features. Mol Ther Nucleic Acids 2019;18:269–74.

[30] Tahir M, Tayara H, Chong KT. iRNA–PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components. J Theor Biol 2019;465:1–6.

[31] Chen W, Tang H, Ye J, Lin H, Chou KC. iRNA–PseU: identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 2016;5:e332.

[32] Song Z, Huang D, Song B, Chen K, Song Y, Liu G, et al. Attention-based multi–label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat Commun 2021;12:4011.

[33] Fang T, Zhang Z, Sun R, Zhu L, He J, Huang B, et al. RNAm5CPred: prediction of RNA 5–methylcytosine sites based on three different kinds of nucleotide composition. Mol Ther Nucleic Acids 2019;18:739–47.

[34] Zou Q, Xing P, Wei L, Liu B. Gene2vec: gene subsequence embedding for prediction of mammalian N(6)-methyladenosine sites from mRNA. RNA 2019;25:205–18.

[35] Zhai J, Song J, Cheng Q, Tang Y, Ma C. PEA: an integrated R toolkit for plant epitranscriptome analysis. Bioinformatics 2018;34:3747–9.

[36] Li YH, Zhang G, Cui Q. PPUS: a web server to predict PUS-specific pseudouridine sites. Bioinformatics 2015;31:3362–4.

[37] Chen K, Wei Z, Zhang Q, Wu X, Rong R, Lu Z, et al. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m^6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res 2019;47:e41.

[38] 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.

[39] 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;37:i222–30.

[40] 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 2022;203:226–32.

[41] Körtel N, Rücklé C, Zhou Y, Busch A, Hoch-Kraft P, Sutandy FXR, et al. Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. Nucleic Acids Res 2021;49:e92.

[42] He X, Zhang S, Zhang Y, Lei Z, Jiang T, Zeng J. Characterizing RNA pseudouridylation by convolutional neural networks. Genomics Proteomics Bioinformatics 2021;19:815–33.

[43] Chen Z, Zhao P, Li C, Li F, Xiang D, Akutsu T, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Res 2021;49:e60.

[44] Huang Y, He N, Chen Y, Chen Z, Li L. BERMP: a cross-species classifier for predicting m(6)A sites by integrating a deep learning algorithm and a random forest approach. Int J Biol Sci 2018;14:1669–77.

[45] Xiong Y, He X, Zhao D, Tian T, Hong L, Jiang T, et al. Modeling multi-species RNA modification through multi–task curriculum learning. Nucleic Acids Res 2021;49:3719–34.

[46] Zhu X, He J, Zhao S, Tao W, Xiong Y, Bi S. A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. Brief Funct Genomics 2019;18:367–76.

[47] Lv H, Zhang Z-M, Li S-H, Tan J-X, Chen W, Lin H. Evaluation of different computational methods on 5–methylcytosine sites identification. Brief Bioinform 2019;21:982–95.

[48] Chen X, Sun YZ, Liu H, Zhang L, Li JQ, Meng J. RNA methylation and diseases: experimental results, databases, web servers and computational models. Brief Bioinform 2019;20:896–917.

[49] Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, et al. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res 2018;46:D327–34.

[50] Ma J, Song B, Wei Z, Huang D, Zhang Y, Su J, et al. m5C–Atlas: a comprehensive database for decoding and annotating the 5-methylcytosine (m5C) epitranscriptome. Nucleic Acids Res 2022;50:D196–203.

[51] Tang Y, Chen K, Song B, Ma J, Wu X, Xu Q, et al. m6A–Atlas: a comprehensive knowledgebase for unraveling the N6–methyladenosine (m^6A) epitranscriptome. Nucleic Acids Res 2020;49:D134–43.

[52] Liu S, Zhu A, He C, Chen M. REPIC: a database for exploring the N(6)-methyladenosine methylome. Genome Biol 2020;21:100.

[53] Song B, Chen K, Tang Y, Wei Z, Su J, Magalhães JPd, et al. ConsRM: collection and large–scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome. Brief Bioinform 2021;22:1–17.

[54] Deng S, Zhang H, Zhu K, Li X, Ye Y, Li R, et al. M6A2Target: a comprehensive database for targets of m6A writers, erasers and readers. Brief Bioinform 2020;22:1–11.

[55] Zheng Y, Nie P, Peng D, He Z, Liu M, Xie Y, et al. m6AVar: a database of functional variants involved in m6A modification. Nucleic Acids Res 2018;46:D139–45.

[56] Luo X, Li H, Liang J, Zhao Q, Xie Y, Ren J, et al. RMVar: an updated database of functional variants involved in RNA modifications. Nucleic Acids Res 2021;49:D1405–12.

[57] Chen K, Song B, Tang Y, Wei Z, Su J, Meng J. RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis. Nucleic Acids Res 2020;49:D1396–404.

[58] Ma J, Zhang L, Chen S, Liu H. A brief review of RNA modification related database resources. Methods 2022;203:342–53.

[59] Begik O, Lucas MC, Liu H, Ramirez JM, Mattick JS, Novoa EM. Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. Genome Biol 2020;21:97.

[60] Zhang H, Shi X, Huang T, Zhao X, Chen W, Gu N, et al. Dynamic landscape and evolution of m6A methylation in human. Nucleic Acids Res 2020;48:6251–64.

[61] Liu J, Li K, Cai J, Zhang M, Zhang X, Xiong X, et al. Landscape and regulation of m(6)A and m(6)Am methylome across human and mouse tissues. Mol Cell 2020;77:426–40 e6.

[62] An S, Huang W, Huang X, Cun Y, Cheng W, Sun X, et al. Integrative network analysis identifies cell-specific trans regulators of m6A. Nucleic Acids Res 2020;48:1715–29.

[63] Xiong X, Hou L, Park YP, Molinie B, Gregory RI, Kellis M. Genetic drivers of m(6)A methylation in human brain, lung, heart and muscle. Nat Genet 2021;53:1–10.

[64] Liu K, Cao L, Du P, Chen W. im6A–TS–CNN: Identifying the N(6)-methyladenine site in multiple tissues by using the convolutional neural network. Mol Ther Nucleic Acids 2020;21:1044–9.

[65] Dao FY, Lv H, Yang YH, Zulfiqar H, Gao H, Lin H. Computational identification of N6-methyladenosine sites in multiple tissues of mammals. Comput Struct Biotechnol J 2020;18:1084–91.

[66] Abbas Z, Tayara H, Zou Q, Chong KT. TS–m6A–DL: tissue-specific identification of N6-methyladenosine sites using a universal deep learning model. Comput Struct Biotechnol J 2021;19:4619–25.

[67] Chatsirisupachai K, Lesluyes T, Paraoan L, Van Loo P, de Magalhaes JP. An integrative analysis of the age-associated multi-omic landscape across cancers. Nat Commun 2021;12:2345.

[68] Silva AS, Wood SH, van Dam S, Berres S, McArdle A, de Magalhaes JP. Gathering insights on disease etiology from gene expression profiles of healthy tissues. Bioinformatics 2011;27:3300–5.

[69] Pei G, Hu R, Jia P, Zhao Z. DeepFun: a deep learning sequence-based model to decipher non–coding variant effect in a tissue- and cell type-specific manner. Nucleic Acids Res 2021;49:W131–9.

[70] Buniello A, MacArthur JA L, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI–EBI GWAS catalog of published genome–wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2018;47:D1005–12.

[71] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2015;44:D862–8.

[72] Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2004;32:13–6.

[73] Members C-N, Partners. Database resources of the national genomics data center, China national center for bioinformation in 2021. Nucleic Acids Res 2021;49:D18–28.

[74] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 2011;17:10–2.

[75] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37:907–15.

[76] Meng J, Lu Z, Liu H, Zhang L, Zhang S, Chen Y, et al. A protocol for RNA methylation differential analysis with MeRIP-seq data and exomePeak R/Bioconductor package. Methods 2014;69:274–81.

[77] Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 2015;19:A68.

[78] 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;37:i222–30.

[79] Song B, Tang Y, Chen K, Wei Z, Rong R, Lu Z, et al. m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. Bioinformatics 2020;36:3528–36.

[80] Sun WJ, Li JH, Liu S, Wu J, Zhou H, Qu LH, et al. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res 2016;44:D259–65.

[81] Schwartz S, Mumbach MR, Jovanovic M, Wang T, Maciag K, Bushkin GG, et al. Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. Cell Rep 2014;8:284–96.

[82] Lin S, Choe J, Du P, Triboulet R, Gregory RI. The m(6)A methyltransferase METTL3 promotes translation in human cancer cells. Mol Cell 2016;62:335–45.

[83] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.

[84] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.

[85] Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009;19:1553–61.

[86] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 2013;34:57–65.

[87] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high–throughput sequencing data. Nucleic Acids Res 2010;38:e164.

[88] Zhu Y, Xu G, Yang YT, Xu Z, Chen X, Shi B, et al. POSTAR2: deciphering the post–transcriptional regulatory logics. Nucleic Acids Res 2018;47:D203–11.

[89] Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife 2015;4:e05005.

[90] Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-seq data. Nucleic Acids Res 2013;42:D92–7.

[91] Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol 2013;9:e1003118.

[92] Buels R, Yao E, Diesh CM, Hayes RD, Holmes IH. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 2016;17:66.

[93] Zhang SY, Zhang SW, Zhang T, Fan XN, Meng J. Recent advances in functional annotation and prediction of the epitranscriptome. Comput Struct Biotechnol J 2021;19:3015–26.

[94] Zhang S, Zhao BS, Zhou A, Lin K, Zheng S, Lu Z, et al. m6A demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem–like cells by sustaining FOXM1 expression and cell proliferation program. Cancer Cell 2017;31:591–606. e6.

[95] Hou Y-Y, Cao W-W, Li L, Li S-P, Liu T, Wan HY, et al. MicroRNA-519d targets MKi67 and suppresses cell growth in the hepatocellular carcinoma cell line QGY-7703. Cancer Lett 2011;307:182–90.

[96] Geng Y, Guan R, Hong W, Huang B, Liu P, Guo X, et al. Identification of m6A–related genes and m6A RNA methylation regulators in pancreatic cancer and their association with survival. Ann Transl Med 2020;8:387.

[97] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47:D941–7.

[98] Liu T, Li C, Jin L, Li C, Wang L. The prognostic value of m6A RNA methylation regulators in colon adenocarcinoma. Med Sci Monit 2019;25:9435–45.

[99] Yang Z, Wang T, Wu D, Min Z, Tan J, Yu B. RNA N6-methyladenosine reader IGF2BP3 regulates cell cycle and angiogenesis in colon cancer. J Exp Clin Cancer Res 2020;39:203.

[100] Guo T, Liu DF, Peng SH, Xu AM. ALKBH5 promotes colon cancer progression by decreasing methylation of the lncRNA NEAT1. Am J Transl Res 2020;12:4542–9.

[101] Ni W, Yao S, Zhou Y, Liu Y, Huang P, Zhou A, et al. Long noncoding RNA GAS5 inhibits progression of colorectal cancer by interacting with and triggering YAP phosphorylation and degradation and is negatively regulated by the m(6)A reader YTHDF3. Mol Cancer 2019;18:143.

[102] Kim SH, Joshi K, Ezhilarasan R, Myers TR, Siu J, Gu C, et al. EZH2 protects glioma stem cells from radiation-induced cell death in a MELK/FOXM1–dependent manner. Stem Cell Reports 2015;4:226–38.

[103] Schonberg DL, Miller TE, Wu Q, Flavahan WA, Das NK, Hale JS, et al. Preferential iron trafficking characterizes glioblastoma stem–like cells. Cancer Cell 2015;28:441–55.

[104] Zhang N, Wei P, Gong A, Chiu WT, Lee HT, Colman H, et al. FoxM1 promotes beta–catenin nuclear localization and controls Wnt target-gene expression and glioma tumorigenesis. Cancer Cell 2011;20:427–42.

[105] Li Y, Zhang S, Huang S. FoxM1: a potential drug target for glioma. Future Oncol 2012;8:223–6.

[106] Yarmishyn AA, Yang Y-P, Lu K-H, Chen Y-C, Chien Y, Chou S-J, et al. Musashi-1 promotes cancer stem cell properties of glioblastoma cells via upregulation of YTHDF1. Cancer Cell Int 2020;20:1–15.

[107] Liu J, Ren D, Du Z, Wang H, Zhang H, Jin Y. m(6)A demethylase FTO facilitates tumor progression in lung squamous cell carcinoma by regulating MZF1 expression. Biochem Biophys Res Commun 2018;502:456–64.

[108] Liu L, Zhang SW, Zhang YC, Liu H, Zhang L, Chen R, et al. Decomposition of RNA methylome reveals co-methylation patterns induced by latent enzymatic regulators of the epitranscriptome. Mol Biosyst 2015;11:262–74.

[109] Zhang L, Chen S, Ma J, Liu Z, Liu H. REW–ISA V2: a biclustering method fusing homologous information for analyzing and mining epi-transcriptome data. Front Genet 2021;12:654820.

[110] Chen S, Zhang L, Lu L, Meng J, Liu H. FBCwPlaid: a functional bi-clustering analysis of epi-transcriptome profiling data via a weighted plaid model. IEEE/ACM Trans Comput Biol Bioinform 2022;19(3):1640-1650.

[111] Adachi H, De Zoysa MD, Yu Y-T. Post-transcriptional pseudouridylation in mRNA as well as in some major types of noncoding RNAs. Biochim Biophys Acta Gene Regul Mech 2019;1862:230–9.

[112] Zaringhalam M, Papavasiliou FN. Pseudouridylation meets next-generation sequencing. Methods 2016;107:63–72.

[113] Hussain S, Aleksic J, Blanco S, Dietmann S, Frye M. Characterizing 5-methylcytosine in the mammalian epitranscriptome. Genome Biol 2013;14:215.

[114] Capitanchik C, Toolan–Kerr P, Luscombe NM, Ule J. How do you identify m(6) A methylation in transcriptomes at high resolution? A comparison of recent datasets. Front Genet 2020;11:398.

[115] Chatsirisupachai K, Palmer D, Ferreira S, de Magalhaes JP. A human tissue-specific transcriptomic analysis reveals a complex relationship between aging, cancer, and cellular senescence. Aging Cell 2019;18:e13041.

[116] Gao Y, Liu X, Wu B, Wang H, Xi F, Kohnen MV, et al. Quantitative profiling of N(6)-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using nanopore direct RNA sequencing. Genome Biol 2021;22:22.

[117] Smith MA, Ersavas T, Ferguson JM, Liu H, Lucas MC, Begik O, et al. Molecular barcoding of native RNAs using nanopore sequencing and deep learning. Genome Res 2020;30:1345–53.

[118] Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Holzer M, et al. Direct RNA nanopore sequencing of full–length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res 2019;29:1545–54.

[119] Begik O, Lucas MC, Pryszcz LP, Ramirez JM, Medina R, Milenkovic I, et al. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. Nat Biotechnol 2021;39:1278–91.

[120] Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, et al. Accurate detection of m(6)A RNA modifications in native RNA sequences. Nat Commun 2019;10:4079.

[121] McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, et al. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun 2019;10:579.

[122] Price AM, Hayer KE, McIntyre ABR, Gokhale NS, Abebe JS, Della Fera AN, et al. Direct RNA sequencing reveals m(6)A modifications on adenovirus RNA are necessary for efficient splicing. Nat Commun 2020;11:6016.

[123] Pryszcz LP, Novoa EM. ModPhred: an integrative toolkit for the analysis and storage of nanopore sequencing DNA and RNA modification data. Bioinformatics 2021;38:257–60.

[124] Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, et al. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol 2021;39:1394–402.

Figure legends

Figure 1 The overall design of m6A-TSHub

By integrating 184,554 m^6A sites detected from 23 different healthy human tissues (m6A-TSDB), a deep learning framework that learns tissue-specific RNA methylation patterns was developed (m6A-TSFinder). The effect of genetic variants on tissue-specific m6A sites was then evaluated (m6A-TSVar). A total of 587,983 cancer somatic mutations were predicted to be able to affect m^6A methylation of RNA in their corresponding cancer-originating tissues. The predicted m^6A-affecting SNPs were then systematically validated using available cancer epitranscriptome datasets, and then functionally annotated with disease and phenotype association from GWAS, along with features relating to the post-transcriptional machinery (microRNA target sites, splicing sites, and RNA binding protein binding sites) that are potentially mediated by m^6A modification (m6A-CAVar). A web interface was constructed to enable the exploration, query, online analysis, and download of relevant results and data. GWAS, genome-wide association study. RBP, RNA binding protein.

Figure 2 A simplified graphic illustration of the proposed m6A-TSFinder framework

Figure 3 Workflow of how to determine the confidence level of m^{6}A variants

Three types of confidence levels were applied. The cancer-driving somatic variants were extracted from TCGA-projects, and mapped to the m^{6}A profiling samples derived from corresponding tumor-growth tissues. A tissue-specific weakly supervised model was then applied to obtain m^{6}A-associated variants labeled in low confidence level. m^{6}A profiling samples from tumor-growth tissues were then used for validation of the prediction results, and the validated portion was classified into medium confidence level. Lastly, all variants with medium confidence level were annotated with disease information from ClinVar and GWAS, and then classified into the high confidence group. Lung tissue, healthy and cancerous, is used as an example here. The same protocol was followed for all 23 tissues. GWAS, genome-wide association study.

Figure 4 Tissue-specificity of cancer m^{6}A variants

A. The proportion of m^{6}A variants that are shared among different tissues. Most m^{6}A-associated variants (93.25%) were identified in only 1 tissue, with 3.86%, 1.7%, and 1.17% identified in 2, 3, and more than 3 tissues, respectively. **B.** The proportion of m^{6}A variant-carrying genes shared among tissues. The consistency is much higher at the gene level. Most m^{6}A variants-carrying genes are shared among multiple tissues, with only 16.59% associated to one tissue type. **C.** The pairwise association of tissues in terms of proportion of shared m^{6}A variant carrying genes. Most tissues are significantly correlated. The exceptions are heart, adrenal gland, lymph nodes, bone marrow, testis, and thyroid.

Figure 5 Enhanced web interface

A. A human body diagram is available for querying cancer somatic m^{6}A-associated variants in their originating tissues. **B.** Users can query the associated variants by cancer type. **C.** Users can also query the associated variants disease, region, gene symbol, COSMIC and Rs ID, and further filter the returned results. **D.** Details of tissue-specific m^{6}A peaks collected in m6A-TSDB. **E.** Details of cancer-related m^{6}A-associated variants. **F.** Details of disease annotation involved. **G.** The online-tools provided for analysis

of user-uploaded files, including assessing m^{6}A-associated variants in tissues (m6A-TSVar). **H.** The online-tool for identifying tissue-specific m^{6}A sites (m6A-TSFinder).

Figure 6 Case study on gene *EGFR*

A. Searching for the gene '*EGFR'* in m6A-CAVar database returns a total of 10 m^{6}A variants identified in two lung cancer types, the details of which can be viewed by clicking the m6A-CAVar ID. **B.** Users can further filter the search results in specific cancer types. **C.** A human body map is provided on the front page of m6A-CAVar website, which enables quick positioning of cancer m^{6}A-associated variants functions at a specific tissue. **D.** The disease and phenotype association of recorded m^{6}A variant.

Tables

Table 1 Performance evaluation of tissue-specific m6A model

Table 2 Performance comparison between m6A-TSFinder and competing approaches on independent dataset (AUROC)

Table 3 Tissue-specific m^{6}A cancer variants collected in m6A-CAVar

Supplementary material

Figure S1 Motif captured under each tissue-specific m6A prediction model

The consensus motifs from instances with higher than average weights were extracted using TF-MoDISco, under each tissue model, respectively. To sum up, we identified one consistence motif GGACU under all tissue models, which was matched to the known m6A consensus motif DRACH.

Figure S2 m6A patterns captured under specific human gene

A. human gene *RYR2* encodes a ryanodine receptor found in cardiac muscle sarcoplasmic reticulum, this gene was biased expressed in heart and brain. We found three m6A sites located on gene *RYR2* from heart samples, compared with two m6A sites from brain, one from liver, one from ovary, and one from uterus, respectively. **B.** for human gene *PXDNL* (biased expression in heart), we observed only one tissue-specific m6A sites from heart sample. **C.** and **D.** human gene *HMGCS2* and *C6* were both reported to be biased expressed in liver. We found one m6A peak located on gene *HMGCS2* and gene *C6*, respectively, identified from human liver sample.

Table S1 Distinct features of m6A-TSHub compared with existing resources

Table S2 m6A-seq samples

Table S3 TCGA sources

Table S4 Single-based samples

Table S5 Gene-TCGA associations

Table 1 Performance evaluation of tissue-specific m^{6}A model

Tissue type	10-fold cross-validation				Independent testing			
	Accuracy	Precision	MCC	AUROC	Accuracy	Precision	MCC	AUROC
Lung	0.764	0.835	0.536	0.843	0.775	0.761	0.55	0.853
Bladder	0.758	0.760	0.517	0.836	0.766	0.750	0.532	0.848
Colon	0.740	0.770	0.482	0.810	0.744	0.730	0.490	0.810
Lymph nodes	0.771	0.797	0.544	0.844	0.78	0.735	0.570	0.844
Cerebrum	0.745	0.799	0.495	0.827	0.758	0.768	0.515	0.834
Cerebellum	0.715	0.718	0.432	0.798	0.72	0.731	0.441	0.801
Hypothalamus	0.733	0.724	0.467	0.799	0.746	0.74	0.493	0.811
Brainstem	0.727	0.742	0.454	0.764	0.721	0.713	0.443	0.789
Kidney	0.685	0.694	0.369	0.755	0.647	0.628	0.297	0.718
Bone marrow	0.694	0.634	0.391	0.757	0.698	0.721	0.397	0.757
Liver	0.742	0.747	0.484	0.805	0.737	0.717	0.476	0.803
Ovary	0.730	0.710	0.464	0.814	0.726	0.722	0.453	0.812
Prostate	0.752	0.779	0.507	0.819	0.759	0.736	0.521	0.830
Soft tissues	0.766	0.855	0.544	0.855	0.771	0.775	0.543	0.858
Skin	0.750	0.850	0.511	0.835	0.773	0.753	0.547	0.857
Stomach	0.772	0.820	0.549	0.852	0.77	0.764	0.539	0.848
Corpus uteri	0.722	0.656	0.452	0.813	0.734	0.715	0.470	0.822
Adrenal gland	0.737	0.771	0.474	0.804	0.741	0.716	0.485	0.817
Heart	0.778	0.824	0.558	0.854	0.772	0.759	0.546	0.846
Rectum	0.747	0.725	0.496	0.826	0.767	0.747	0.536	0.828
Testis	0.743	0.770	0.489	0.810	0.731	0.734	0.463	0.804
Thyroid gland	0.765	0.805	0.533	0.845	0.753	0.733	0.509	0.830
Pancreas	0.761	0.770	0.523	0.838	0.751	0.739	0.502	0.834

Note: MCC, matthews correlation coefficient; AUROC, the area under ROC curve.

Table 2 Performance comparison between m6A-TSFinder and competing approaches on independent dataset (AUROC)

	Performance on independent dataset			
	m6A-TSFinder	**TS-m6A-DL**	**im^{6}A-TS-CNN**	**iRNA-m^{6}A**
Brain	0.8132	0.8097	0.8056	0.7845
Liver	0.8850	0.8784	0.8805	0.8681
Kidney	0.8796	0.8802	0.8727	0.8565
Average	0.8593	0.8561	0.8529	0.8364

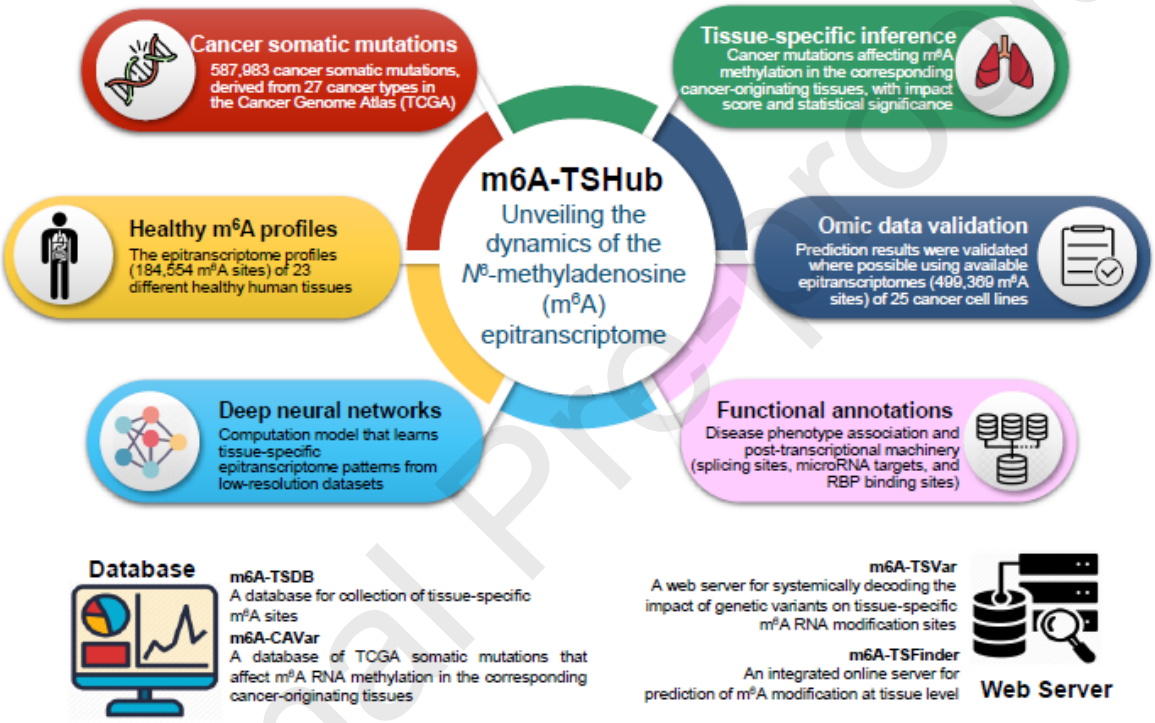
Note: For a fair comparison, the m6A-TSFinder was rebuilt for human brain, liver, and kidney, using the same training and testing datasets applied in the three previous works. The 41nt sequences were considered as one instance and fed into m6A-TSFinder.

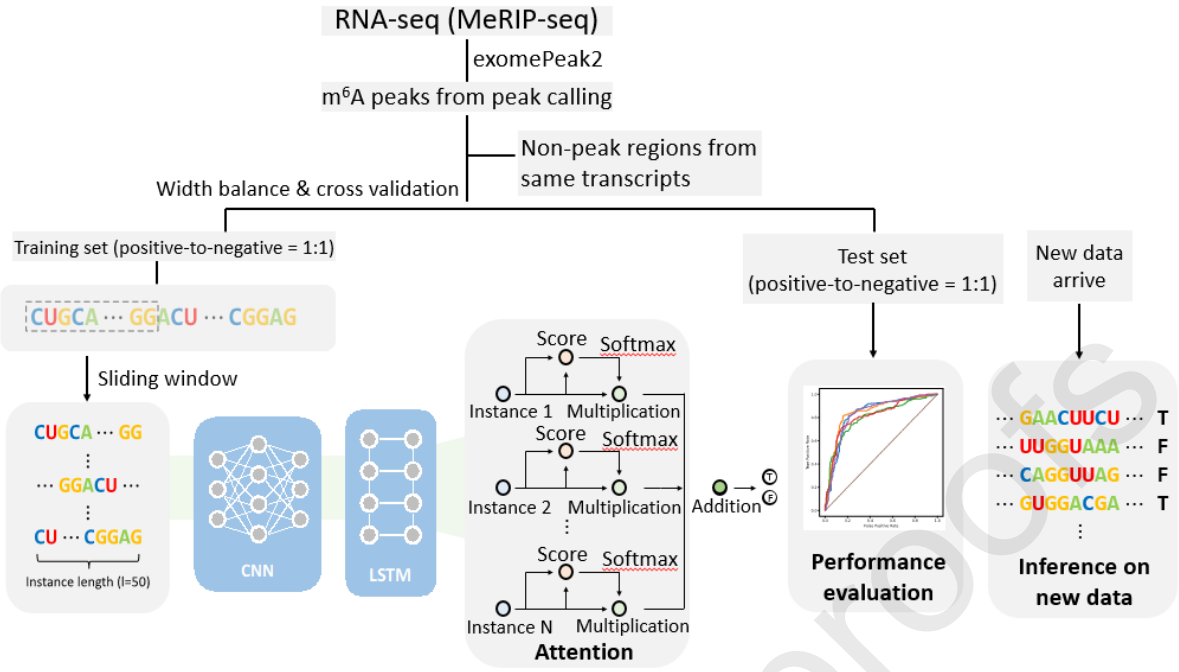
Table 3 Tissue-specific m^{6}A cancer variants collected in m6A-CAVar

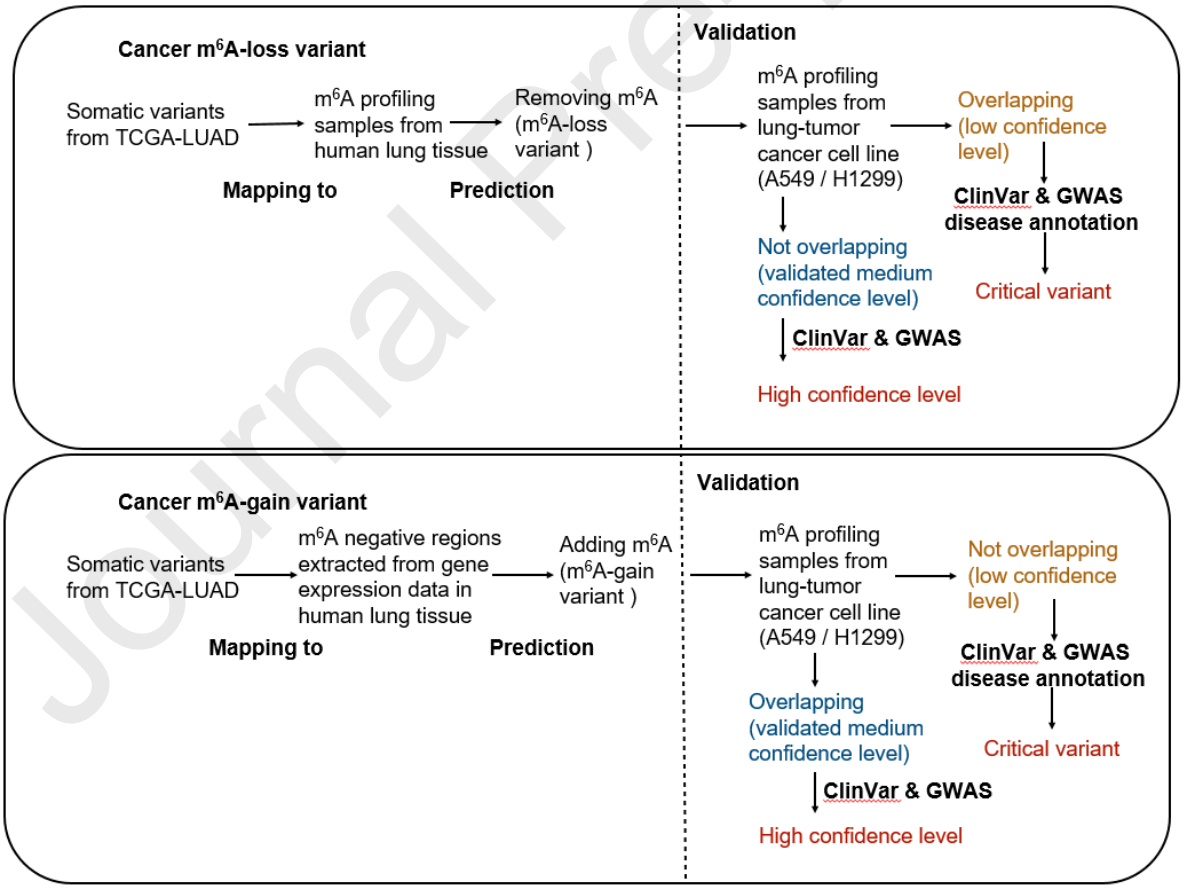
Cancer type	**Primary tissue**	**Matched cancer cell line**	**Variant type**	**Classification**			**Total**
				Low	**Middle**	**High**	
Lung adenocarcinoma (TCGA-LUAD)	Lung	A549, H1299	Gain	27,845	6526	30	34,401
			Loss	1233	1391	2	2626
Bladder urothelial carcinoma (TCGA-BLCA)	Urinary bladder	BCa5637	Gain	25,508	3702	13	29,223
			Loss	3079	1691	6	4776
Colon adenocarcinoma (TCGA-COAD)	Colon	HT29, HCT116	Gain	30,540	8391	82	39,013
			Loss	6	8284	74	8364
Lymphoid neoplasm diffuse large B-cell lymphoma (TCGA-DLBC)	B lymphocyte cell lines	OCI-Ly1	Gain	1189	82	2	1273
			Loss	74	69	0	143
Glioblastoma multiforme (TCGA-GBM)	Cerebrum	U251, GOS-3, PBT003	Gain	8509	3648	47	12,204
			Loss	1453	1181	12	2646
	Cerebellum		Gain	8319	3659	38	12,016
			Loss	1928	1271	4	3203
	Hypothalamus		Gain	6723	3414	27	10,164
			Loss	1522	1482	18	3022
	Brainstem		Gain	7559	3168	40	10,767
			Loss	1374	1451	8	2833
Kidney renal clear cell carcinoma (TCGA-KIRC)	Kidney	iSLK.219	Gain	3844	227	4	4075
			Loss	54	33	0	87
Acute myeloid leukemia (TCGA-LAML)	Hematopoietic stem cells (HSC)	MOLM13, THP1, NOMO-1, MONO-MAC-6, MA9.3ITD	Gain	448	274	0	722
			Loss	3	35	3	41

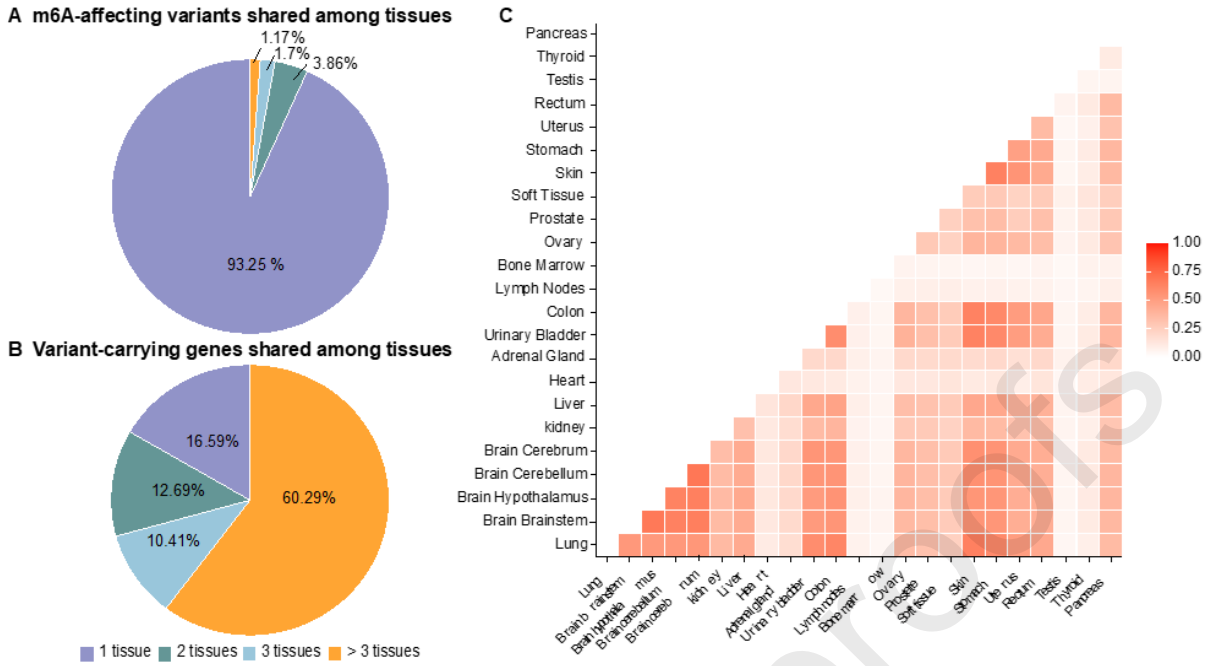
Liver hepatocellular carcinoma (TCGA-LIHC)	Liver	HepG2, Huh7, SMMC7721, HCCLM3	Gain	7416	2511	2	9929
			Loss	18	1765	4	1787
Ovarian serous Cystadenocarcinoma (TCGA-OV)	Ovary	PEO1	Gain	7022	531	0	7553
			Loss	1350	1090	6	2446
Prostate adenocarcinoma (TCGA-PRAD)	Prostate gland	Cd-RWPE-1	Gain	3825	636	6	4467
			Loss	550	288	2	840
Sarcoma (TCGA-SARC)	Soft tissues	U20S	Gain	3592	1324	4	4920
			Loss	373	28	0	401
Skin cutaneous melanoma (TCGA-SKCM)	Skin	Mel624	Gain	79,470	17,177	118	96,765
			Loss	6472	1559	2	8033
Stomach adenocarcinoma (TCGA-STAD)	Stomach	BGC823	Gain	35,438	2202	34	37,674
			Loss	1103	3313	27	4443
Uterine corpus endometrial carcinoma (TCGA-UCEC)	Corpus uteri	HEC-1-A	Gain	80,712	38,242	266	119,220
			Loss	7813	1828	22	9663
Lung squamous cell carcinoma (TCGA-LUSC)	Lung	-	Gain	31,106	-	118	31,224
			Loss	2328	-	2	2330
Mesothelioma (TCGA-MESO)	Lung	-	Gain	595	-	4	599
			Loss	57	-	0	57
	Heart	-	Gain	674	-	5	679
			Loss	102	-	0	102
Brain lower grade glioma (TCGA-LGG)	Cerebrum	-	Gain	6714	-	92	6806
			Loss	1423	-	19	1442
	Cerebellum	-	Gain	6601	-	109	6710
			Loss	1745	-	16	1761
	Hypothalamus	-	Gain	5010	-	77	5087
			Loss	1698	-	11	1709
	Brainstem	-	Gain	5740	-	114	5854
			Loss	1528	-	13	1541
Kidney chromophobe (TCGA-KICH)	Kidney	-	Gain	484	-	9	493
			Loss	9	-	0	9
Kidney renal papillary cell carcinoma (TCGA-KIRP)	Kidney	-	Gain	4028	-	17	4045
			Loss	118	-	0	118
Cholangiocarcinoma (TCGA-CHOL)	Liver	-	Gain	728	-	2	730
			Loss	166	-	2	168
Adrenocortical carcinoma (TCGA-ACC)	Adrenal gland	-	Gain	2285	-	21	2306
			Loss	385	-	3	388
Pheochromocytoma and paraganglioma (TCGA-PCPG)	Adrenal gland	-	Gain	345	-	1	346
			Loss	57	-	0	57
Rectum adenocarcinoma (TCGA-READ)	Rectum	-	Gain	12,433	-	100	12,533
			Loss	1098	-	4	1102
Thymoma (TCGA-THYM)	Heart	-	Gain	520	-	7	527
			Loss	80	-	2	82
Testicular germ cell tumors (TCGA-TGCT)	Testis	-	Gain	405	-	3	408
			Loss	109	-	0	109

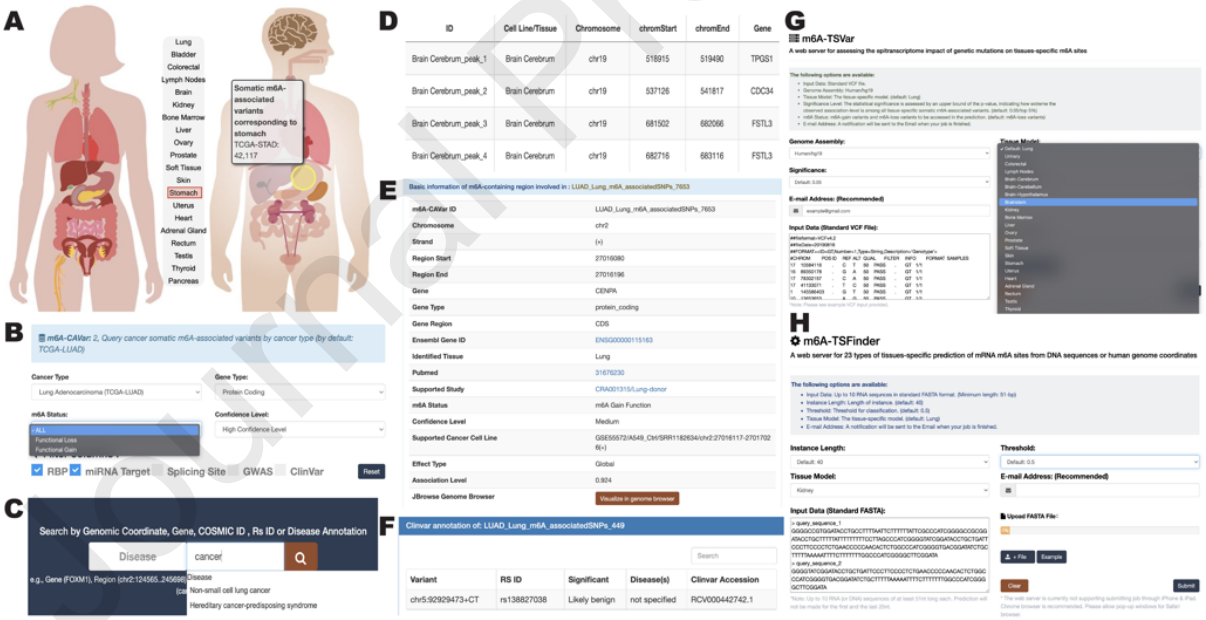
Thyroid carcinoma (TCGA-THCA)	Thyroid gland	-	Gain	992	-	6	998
			Loss	156	-	0	156
Pancreatic adenocarcinoma (TCGA-PAAD)	Pancreas	-	Gain	6473	-	55	6528
			Loss	1236	-	3	1239
Total	-	-	-	463,792	122,473	1718	587,983








A m6A-affecting variants shared among tissues
1.17%
1.7%
3.86%
93.25 %
B Variant-carrying genes shared among tissues
16.59%
12.69%
60.29%
10.41%
1 tissue
2 tissues
3 tissues
> 3 tissues
C
Pancreas
Thyroid
Testis
Rectum
Uterus
Stomach
Skin
Soft Tissue
Prostate
Ovary
Bone Marrow
Lymph Nodes
Colon
Urinary Bladder
Adrenal Gland
Heart
Liver
kidney
Brain Cerebrum
Brain Cerebellum
Brain Hypothalamus
Brain Brainstem
Lung
1.00
0.75
0.50
0.25
0.00


A
Somatic m6A-associated variants corresponding to stomach TCGA-STAD: 42,117
B
m6A-CAVar 2, Query cancer somatic m6A-associated variants by cancer type (by default: TCGA-LUAD)
RBP
miRNA Target
Splicing Site
GWAS
ClinVar
C
Search by Genomic Coordinate, Gene, COSMIC ID , Rs ID or Disease Annotation
Disease
cancer
Non-small cell lung cancer
Hereditary cancer-predisposing syndrome
D
E
F
Clinvar annotation of: LUAD_Lung_m6A_associatedSNPs_449
G
m6A-TSVar
H
m6A-TSFinder
A web server for 23 types of tissues-specific prediction of mRNA m6A sites from DNA sequences or human genome coordinates