Briefings in Bioinformatics, 22(6), 2021, 1–17

https://doi.org/10.1093/bib/bbab088
Problem Solving Protocol

ConsRM: collection and large-scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome

Bowen Song†, Kunqi Chen†, Yujiao Tang†, Zhen Wei, Jionglong Su, João Pedro de Magalhães, Daniel J. Rigden and Jia Meng

Corresponding author: Jia Meng, Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; Institute of Systems, Molecular and Integrative Biology, University of Liverpool, L7 8TX, Liverpool, United Kingdom. E-mail: jia.meng@xjtlu.edu.cn
†These authors contributed equally to this work.

Abstract

Motivation N6-methyladenosine (m^6A) is the most prevalent RNA modification on mRNAs and lncRNAs. Evidence increasingly demonstrates its crucial importance in essential molecular mechanisms and various diseases. With recent advances in sequencing techniques, tens of thousands of m^6A sites are identified in a typical high-throughput experiment, posing a key challenge to distinguish the functional m^6A sites from the remaining 'passenger' (or 'silent') sites. **Results:** We performed a comparative conservation analysis of the human and mouse m^6A epitranscriptomes at single site resolution. A novel scoring framework, ConsRM, was devised to quantitatively measure the degree of conservation of individual m^6A sites. ConsRM integrates multiple information sources and a positive-unlabeled learning framework, which integrated genomic and sequence features to trace subtle hints of epitranscriptome layer conservation. With a series validation experiments in mouse, fly and zebrafish, we showed that ConsRM outperformed well-adopted conservation scores (phastCons and phyloP) in distinguishing the conserved and unconserved m^6A sites. Additionally, the m^6A sites with a higher ConsRM score are more likely to be functionally important. An online database was developed containing the conservation metrics of 177 998 distinct human m^6A sites to support conservation analysis and functional prioritization of individual m^6A sites. And it is freely accessible at: https: //www.xjtlu.edu.cn/biologicalsciences/con.

Key words: conservation analysis; N6-methyladenosine (m^6A); genome analysis; scoring framework

Bowen Song received a Master of Research degree from University of Liverpool. He is currently a PhD student in the Institute of Systems, Molecular and Integrative Biology, University of Liverpool. His research interests are bioinformatics and RNA modifications.
Kunqi Chen received a PhD degree from University of Liverpool. He is currently a research professor at Key Laboratory of Ministry of Education of Gastrointestinal Cancer, School of Basic Medical Science, Fujian Medical University, Fuzhou, China. His research interests are bioinformatics and database construction.
Yujiao Tang received a Bachelor of Science degree from Xi'an Jiaotong-Liverpool University. She is currently a PhD student at University of Liverpool with research interests in bioinformatics databases.
Zhen Wei is an assistant professor at Department of Biological Science, Xi'an Jiaotong-Liverpool University. His research interests are computational biology and data mining.
Jionglong Su is an associated professor at Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University. His research interests are Artificial Intelligence and data mining.
João Pedro de Magalhães is a professor at Institute of Ageing & Chronic Disease, University of Liverpool. His work focuses primarily on the biology and genetics of aging, combining experimental and computational methods.
Daniel J. Rigden is a professor at Institute of Systems, Molecular and Integrative Biology, University of Liverpool. His research interests are structural biology and bioinformatics.
Jia Meng is an associated professor at Department of Biological Sciences, Xi'an Jiaotong-Liverpool University. His research interests are epitranscriptome and bioinformatics.

Submitted: 11 December 2020; **Received (in revised form):** 4 February 2021

Introduction

As sequencing technologies have advanced, the prevalence and abundance of RNA modifications in the mammalian epitranscriptome have become increasingly evident [1–3]. Today, more than 150 distinct biochemical modifications have been identified [4], among which N^6-methyladenosine (m^6A) has become a research hotspot for its significance in essential molecular mechanisms of eukaryotic species [5, 6]. This non-cap methylation of the adenosine base at the nitrogen-6 position was firstly detected in the 1970s [7], and its functions have been intensively studied during the past few years.

As the most abundant RNA modification on mRNAs and lncRNAs, m^6A RNA methylation plays important roles in a series of essential biological processes. Tens of thousands of m^6A RNA methylation sites have been detected in the transcriptome, suggesting a wide-ranging effect of this important modification on the regulation of gene expression [8]. The translation efficiency was found to be affected by m^6A modification [9, 10], for example, m^6A impacts translation extension via modulating the anticodon pairing rate of both tRNA and mRNA [9]. It also recruits CCR4-NOT complex and is involved in the regulation of histone modifications [11]. Studies also reported various biological functions that m^6A modification participates, including but not limited to translation, response to heat shock [12], DNA damage [13] and embryonic development [14–16]. Meanwhile, m^6A dysregulation is found to be associated with various diseases: for example, the m^6A demethylase FTO is closely related to the development of recessive lethality syndrome [17]. More recent studies have also revealed that abnormal m^6A regulation may impact cancer development, such as breast cancer [18, 19], prostate cancer [20] and liver cancer [21].

Three categories of m^6A-related proteins have been identified—the writers, erasers and readers. m^6A is produced co-transcriptionally by the methyltransferase complexes (MTC), also termed as 'writers', which include methyltransferase-like 3 (METTL3) catalytic subunit [22], methyltransferase-like 14 (METTL14) [23], methyltransferase-like 16 (METTL16) [24], VIRMA [25], ZC3H13 [26], RNA-binding motif 15 (RBM15) [5] and Wilms tumor 1-associated protein (WTAP) [27]. METTL3 was first identified as a member of the MTC [28] and was reported to be highly conserved in eukaryotes from human to yeast [29] with versatile and important biological functions. In mouse, METTL3 is closely related to cell division, reprogramming and differentiation [16, 30]. Besides, METTL3 was found to be responsible for spermatogenesis through meiosis regulation and spermatogonial differentiation [31]. Inducer of meiosis 4 (IME4) is the ortholog of METTL3 in *Saccharomyces cerevisiae* and *Drosophila melanogaster*. Deletion of IME4 in *Saccharomyces cerevisiae* and Drosophila leads to defects in sporulation [32] and lethal phenotype [33], respectively. Also, in *Arabidopsis thaliana*, dysregulation of MT-A (ortholog of METTL3) causes embryonic lethal [34]. In zebrafish embryos, knockdown of METTL3 and WTAP results in the defection of tissue differentiation [27]. m^6A erasers like fat mass and obesity-associated protein (FTO) [35] and AlkB homolog 5 (ALKBH5) [36] remove m^6A methylation, demonstrating that m^6A methylation is a reversible process. ALKBH5 is also responsible for the demethylation activity in mouse: ALKBH5-deficient mice were found to have increasing level of mRNA m^6A, which leads to impaired fertility [36]. Additionally, ALKBH9B and ALKBH10B [37], which belong to the AlkB family, act as the m^6A demethylases in Arabidopsis [38] and revert m^6A to adenosine. Probing the function of m^6A readers is key to understanding how m^6A modification regulates gene expression. Readers include YTH (YT512-B homology) domain family [39], eukaryotic initiation factor 3 (EIF3) [5], insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs) [40] and heterogeneous nuclear ribonucleoprotein (HNRNPA2B1 and HNRNPC) [41]. These reader proteins recognize and bind specifically to m^6A-modified RNAs, leading to the implementation of biological functions and distinct destinies of target RNAs [42]. Among them, YTH family members were found to be highly conserved in different species, such as humans, Drosophila, yeast and Arabidopsis [5], with a YTH consensus domain for m^6A recognition. Taken together, the reversible m^6A and its regulators were widely found and conserved among various species, indicating the broad biological roles of this dynamic process on RNAs.

Due to the advances of high-throughput sequencing approaches developed for transcriptome-wide mapping of the m^6A RNA modification, tens of thousands of m^6A sites can be identified simultaneously with a single high-throughput experiment [43–48]. The antibody-based sequencing approach MeRIP-Seq (or m^6A-Seq) provides m^6A-containing regions with a resolution of around 100-nt [43, 44]. Since its invention in 2012, it has been widely applied and successfully identified m^6A methylation in more than 30 organisms among various species. Besides MeRIP-Seq (or m^6A-Seq), the precise location of m^6A modification can be profiled transcriptome-wise with a variety of advanced sequencing approaches of single base-resolution, including antibody-based methods (m^6A-CLIP [48], miCLIP [45], PA-m^6A-Seq [47], m^6ACE-seq [49]), enzyme-based methods (m^6A-REF-seq [50]), fusion domain-based method (MAZTER-seq [51], DART-seq [52]) and substrate alternation-based method (m^6A-Label-seq [53]). These experimental approaches, together, offered valuable information concerning the position of m^6A RNA modification in different species and under various biological contexts.

Databases of m^6A RNA methylation sites have been developed to allow users to query and analyze this information [54–59]. Among them, MetDB [59] and RMBase [58] each collected around 400 000 unique m^6A sites in human. However, although m^6A has been unambiguously demonstrated to be functionally important, it is unlikely that all (or most) m^6A sites are functionally important, raising the question of how to distinguish the functional m^6A sites from the remaining 'passenger' (or 'silent') sites caused by off-target effects of m^6A methyltransferases. This is especially important as today's high-throughput sequencing techniques typically identify huge numbers of m^6A sites.

We seek to address this challenge from the perspective of evolution. It is known that positive (or purifying) selection reflects how evolutionary forces shape biological process or features, and hence functionally important elements that increase organismal fitness are more likely to be conserved during evolution [60]. The idea of inferring functionality from evolutionary evidence has been used quite extensively in functional genomics for prioritizing various biological elements, such as phosphorylation or structure prediction [61–65]. The degree of conservation may be considered as a priori indicator of functionality; however, how and to what degree can we analyze the cross-species conservation of an arbitrary m^6A RNA methylation site, and use it as an evaluation metric for functional discrimination remains elusive. Previous studies of the m^6A epitranscriptome have reported the overall conserved landscape between human and mouse [43, 44], the evolution of m^6A modification among primates [66] and the important link of m^6A between genetic and phenotypic variation [67]. A recent study focused on the dynamic m^6A methylation profiles across

various human tissues, and reported that m^{6}A modification on different genic locations may subject to different selection pressure [68]. Besides, Liu *et al*. [69] found that a significant proportion of m^{6}A sites, especially those located on protein coding regions, are not evolutionarily conserved and are likely nonfunctional. Meanwhile, they analyzed m^{6}A sites detected in *S. cerevisiae* and S. mikatae and showed that the number of m^{6}A sites shared by the two species was significantly greater than expected by chance, indicating a small proportion of yeast m^{6}A modifications are conserved and likely functional. However, the similar analysis between human and mouse m^{6}A sites was unavailable due to lack of m^{6}A sequencing data profiled at base-resolution level. These studies have focused on the overall conservation of the epitranscriptome without paying attention to the conservation of individual m^{6}A sites. Very recently, the m^{6}A-Atlas database [70] provided a binary label (conserved or not conserved) for the conservation of individual m^{6}A sites based on existing epitranscriptome data, but it didn't quantitatively assess the degree of conservation of m^{6}A sites, and more importantly, neither can it computationally predict the potential conservation (not observable from existing data) of individual m^{6}A sites (see Supplementary Table S1 for a detailed comparison). Since the coverage of the epitranscriptome from existing data is quite limited, the conserved m^{6}A sites reported from m^{6}A-Atlas [70] are likely to be incomplete. Computational approaches that can predict the conserved m^{6}A sites not yet confirmed from existing data would be highly desired at current stage of epitranscriptome studies when high-quality epitranscriptome data are not abundantly available in species other than human.

We present here ConsRM, a resource for conservation analysis of m^{6}A RNA methylation sites. It has three key features: a novel scoring framework for quantitatively measuring the conservation of individual m^{6}A sites, an online database collecting 177 998 human m^{6}A RNA methylation sites along with their conservation scores and a web server that helps evaluate the conservation of any newly detected user-provided list of human m^{6}A sites. To explore whether the conserved m^{6}A sites reported by ConsRM differ from less-conserved positions in terms of functionality, a series of comparisons was performed, including germline versus somatic mutations, mutations, deleterious levels, disease-associated analysis and RNA-binding protein interactions, etc. Results showed that the m^{6}A sites scored higher in our system were clearly more likely to be conserved in different species, and are more strongly associated with various biological functions, suggesting the effectiveness of our approach in distinguishing functional m^{6}A sites from silent modifications from the perspective of conservation.

Materials and methods

Evaluate the conservation of individual m^{6}A site with ConsRM score

To systematically evaluate the degree of conservation of individual m^{6}A site, we developed a detailed scoring mechanism, the ConsRM score, by integrating information from six different sources, including positional mapping, tissue-specific mapping, support from multiple studies, sequence similarity, machine learning modeling and genome conservation.

Positional mapping

Positional mapping concerns whether the RNAs transcribed from conserved loci of human and mouse are equally m^{6}A modifiable at corresponding positions. To evaluate positional mapping, we collected a total of 177 998 and 110 959 m^{6}A sites in human and mouse transcriptome, respectively, from 46 datasets generated from six different m^{6}A profiling techniques (Supplementary Sheet S1). It is worth mentioning that we considered only techniques with base-resolution in our analysis to reduce false positive m^{6}A sites. Although MetDB [59] and RMBase [58] both hold a larger collection of human m^{6}A sites, our previous study reported that there existed a substantial proportion of false positive records, which were induced from their motif-based m^{6}A-seq (MeRIP-seq) data analysis pipeline [71]. For each human m^{6}A site (based on hg19 genome assembly), its corresponding coordinate in mouse transcriptome (based on mm10 genome assembly) was identified using the UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The conservation scores of human m^{6}A sites were assigned for 1 mark if their corresponding coordinates in mouse transcriptome were also m^{6}A modifiable. Besides precise positional mapping, we also checked the nearby regions for possible imprecise mapping [61], and assigned 0.8, 0.6, 0.4 and 0.2 mark for human m^{6}A sites if an m^{6}A modification were detected at 1 bp, 2 bp, 3 bp and 4 bp distance from their corresponding mouse coordinates. Imprecise mapping assumes that the m^{6}A sites located very close to each other may have similar and mutually replaceable functions but subject to some penalty.

Tissue-specific mapping

Tissue-specific mapping concerns whether the RNAs transcribed from the conserved locus of human and mouse are simultaneously m^{6}A modified at the same tissue. Currently, m^{6}A sites were successfully identified under 12 tissues from human transcriptome, and for mouse, this number slightly dropped to nine (Supplementary Sheet S1). Among them, four tissues were shared by both species, including brain, liver, kidney and embryonic stem cell (ESC). In tissue-specific mapping, 1 additional mark was assigned for the m^{6}A sites observed in the same tissue of human and mouse.

Supports from multiple studies

Supports from multiple studies concern whether an m^{6}A site can be detected by multiple m^{6}A profiling studies. It was shown previously that RNA modification sites captured by different high-throughput sequencing techniques may exhibit different overall patterns and capture different technical bias [72–75], indicating that different techniques may have their own technical preference. We suspected the m^{6}A sites detected under multiple studies are more reliable with less possibility of being a false positive signal. For this reason, we assigned 1, 0.8, 0.6, 0.4 or 0.2 mark for m^{6}A sites that can be detected by more than 6, 5–6, 3–4, 2 or 1 m^{6}A profiling datasets.

Sequence similarity

Sequence similarity of m^{6}A surrounding bases was also considered. Specifically, sequences were extracted from the 11 bp flanking windows centered on a human m^{6}A and its corresponding base in the mouse transcriptome, respectively. Following a previous example [61], we considered here a customized motif-specific scoring matrix (MSSM) that assigns scores to each position of the paired sequences, with the m^{6}A-forming motif DRACH being assigned with higher weight. Marks were assigned to identical base or transition (changing of a purine nucleotide

Table 1. Motif-specific scoring matrix (MSSM)

Positions		−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5
m^6A motif					D	R	A	C	H			
Score (max: 32)	No change (I)	2	2	2	4	4	4	4	4	2	2	2
	Transition (S)	1	1	1	3	3	3	3	3	1	1	1
	Transversion (V)	0	0	0	0	0	0	0	0	0	0	0
Example	Human seq	T	G	T	A	A	A	C	A	G	A	G
(total: 28)	Mouse seq	G	G	C	A	G	A	G	A	G	G	G
	Comparison	V	I	S	I	S	I	V	I	I	S	I
	Score	0	2	1	4	3	4	0	4	2	1	2

Note: The score obtained from sequence similarity of the two sequences is: 23/32 = 0.71875.

to another purine, or a pyrimidine nucleotide to another pyrimidine, A↔G or C↔T), but not for transversion (changes between purine and pyrimidine). For example, a human m^6A site was detected at position 77 776 162 of positive strand on chromosome 15 by m^6A-REF-Seq [76], the sequences within the 11 bp flanking windows of this m^6A site and its corresponding locus in mouse transcriptome are TGTAAACAGAG and GGCAGAGAGGG, respectively, with A stands for the centered m^6A in human or its corresponding residual in mouse sequence. The cumulative score of this paired sequence is 15 (4 + 3 + 4 + 0 + 4) within the motif region, and 8 (0 + 2 + 1 + 2 + 1 + 2) for the rest of the regions. The score from sequence similar was then 0.71875, which was calculated from (15 + 8)/32 (see Table 1). Besides, 0 mark was assigned for a paired sequence if another base (not A) was observed on the corresponding position in mouse transcriptome.

A machine learning model

A machine learning model was applied for inferring the conserved m^6A sites. The conservation of all m^6A sites was evaluated by a newly proposed predictor, which extracted various domain knowledge derived from genomic features [77, 78] (Supplementary Tables S2 and S3) as well as conventional sequence features, i.e. chemical properties [79, 80] and nucleotide density [81–83] (Supplementary Table S4). For training purpose, we retained sites that are m^6A-modifiable in both human and mouse transcriptome as the positive dataset P. It is worth noting that, although existing studies have accumulated a large amount of epitranscriptome data, the issue of false negative still remains, i.e. due to the incompleteness of the data, the sites not observed to be conservable according to existing data may still be conserved. Therefore, the rest of samples (after excluding positive set P) should be considered as the unlabeled dataset U (rather than the negative set). And the positive-unlabeled learning (PU learning) strategy was applied to find the most reliable negative samples. Specifically, following a previous study [84], the PU learning process was divided into the following two steps. First, we randomly selected the same number of positive samples as the positive dataset P from the unlabeled dataset U, and assigned this subset as the negative dataset N with 1:1 positive–negative ratio. The datasets P and N were then used to train an SVM-based predictor, and the category probabilities of the unlabeled dataset U were predicted. The top 1% of the unlabeled U with the highest probability are likely to be unidentified positive samples, and are excluded from dataset U. This process was repeated 10 times, from which the most reliable negative dataset RN was retained. Next, the dataset RN and P were used to train the final prediction model.

Support Vector Machine (SVM) has been shown to be a quite effective machine learning algorithm in the field of computation biology and achieved good performance previously in various site prediction studies [83, 85, 86]. The R language interface of LIBSVM [87] was used in our study to develop the final prediction model, and the radial basis function was set as kernel following the default setting for other parameters.

For performance evaluation, we randomly selected 80% of dataset P as positive training data, while the rest of 20% was used for independent testing. Initially, for each positive site, 10 negative data were selected from RN. Later, 10 independent predictors were constructed with a balanced 1:1 positive–negative ratio, and their prediction results were averaged. Consequently, two prediction frameworks were developed to evaluate the conservation degree of m^6A sites in human and their corresponding positions of mouse transcriptome, respectively. A 5-fold cross-validation was also performed on training dataset. The prediction accuracy was represented by the receiver operating characteristic curve (ROC curve) (sensitivity against 1-specificity), and the area under ROC curve (AUROC) was calculated as the main performance evaluation metric for its nonparametric characteristics. Moreover, the sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthews correlation coefficient (MCC) were also presented for performance evaluation, specifically:

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{3}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

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

Genome conservation

The phastCons 100-way conservation scores were calculated for human genome derived from genome-wide multiple alignments with 99 other vertebrate species. It was integrated into our scoring framework to evaluate the conservation degree of each human m^6A site from a more general perspective. The scores were generated by R package phastCons100way. UCSC.hg19 [88], with the mark ranging from 0 to 1.

Taken together, the ConsRM score was calculated for each experimentally validated human m^6A site by taking the average

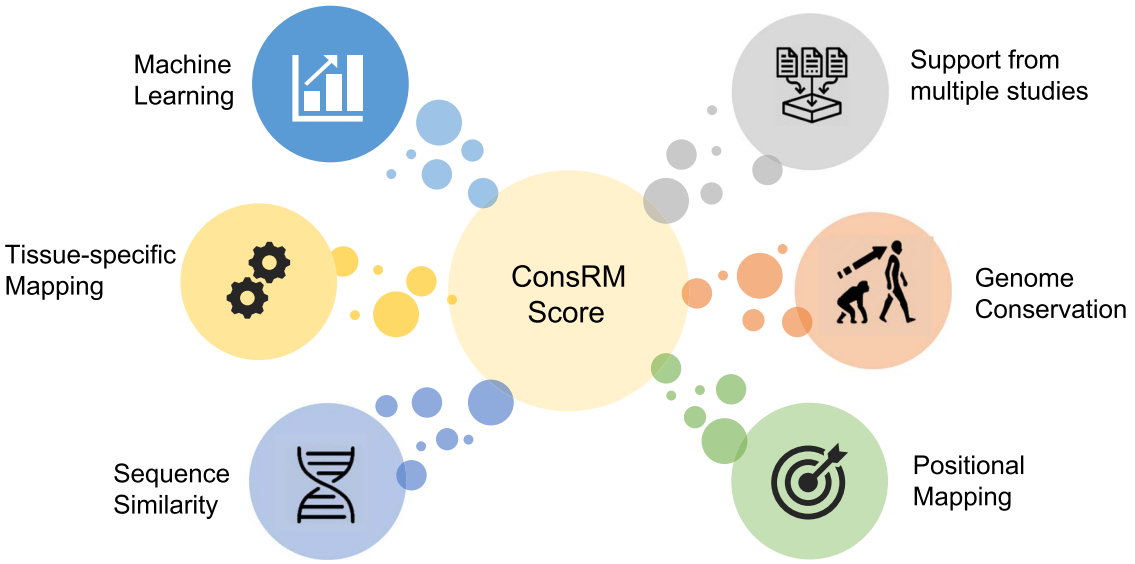

Figure 1. Construction of the ConsRM Score. The ConsRM score was designed to quantitatively measure the degree of conservation of individual m^{6}A RNA methylation sites. It integrates information from six different sources, including the positional mapping, tissue-specific mapping of m^{6}A-epitranscriptomes between human and mouse, the sequence similar, genome conservation, the support of multiple m^{6}A studies and a machine learning model.

of the scores derived from the six aspects mentioned above, and ranges from 0 to 1. The ConsRM score can effectively quantify the evolutionary conservation of individual m^{6}A sites and provide insights into their functional potentials. The overall design of the conservation scoring framework is summarized in Figure 1.

Functional differentiation of the conserved and unconserved m^{6}A sites

In order to test if the m^{6}A sites with a higher ConsRM score are more likely to be functional, a series of experiments were performed involving various biological data. Specifically, 37 094 832 germline mutations and 2684 788 somatic mutations were obtained from dbSNP [89] and the TCGA database (TCGA v15.0) [90], respectively. Only the single-nucleotide variations localized on exonic regions were retained for subsequent analysis (Supplementary Sheet S2). The deleterious level of each genetic mutation was predicted by SIFT [91], PolyPhen2 HVAR [92], PolyPhen2HDIV [92], LRT [93] and FATHMM [94] using the ANNOVAR package [95]. The disease-associated tagSNPs were derived from GWAS catalog [96], Johnson and O'Donnell [97] and the ClinVar database [98], and were used to decipher the potential relationship between disease pathogenesis and m^{6}A conservation level. Besides, since m^{6}A modification was found to recruit RNA-binding proteins that are closely associated with posttranscriptional regulations [23], we then checked whether the conserved m^{6}A sites are more likely to localize within the RBP binding sites collected from STARBASE2 [99], especially for the m^{6}A reader proteins YTHDF1, YTHDF2, YTHDF3, YTHDC1 and YTHDC2 (Supplementary Table S5). Lastly, the experimentally validated m^{6}A sites identified in rat (genome assembly: rn6) and zebrafish (danRer10) were collected (Supplementary Sheet S1), and used to test whether the human m^{6}A sites with higher ConsRM score were more likely to be conserved (as m^{6}A) in a third species.

Construction of ConsRM website

The website interface of ConsRM online platform was constructed using Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP), with MySQL tables exploited for the storage of the metadata. The multiple statistical diagrams were presented by EChars, and Jbrowse genome browser [100] was employed for interactive exploration and visualization of relevant genome coordinate-based records.

Results

Justifying elements of ConsRM score

A significant number of m^{6}A sites are conserved between human and mouse

We first tried to identify the conserved m^{6}A sites between human and mouse. By comparing the conserved loci of human and mouse, we found that, among the total of 177 998 human m^{6}A RNA methylation sites, 22 359 (12.56%) are shared in the mouse transcriptome, i.e. the corresponding bases in the mouse transcriptome are also m^{6}A-modifiable according to the data collected (Supplementary Sheet S1). These conserved m^{6}A sites between human and mouse transcriptome were referred as m^{6}A^{++} sites in the following text for simplicity.

Of interest is whether the proportion of m^{6}A^{++} sites (m^{6}A sites observed to be conserved) is statistically significant. For this purpose, we randomly generated 177 998 pseudo m^{6}A sites from the As within the DRACH consensus motifs of the same m^{6}A-carrying transcripts in human, identified their corresponding genome coordinates in the mouse transcriptome, and compared them with the mouse m^{6}A sites. The process was repeated 1000 times. The testing results showed that only a very small number of pseudo human m^{6}A sites were observed to conserved in mouse (Figure 2A). The result showed the number of modifiable m^{6}A sites shared between human and mouse transcriptome (22 359) is significantly greater than expected by chance (497), suggesting that a majority of the m^{6}A^{++} sites are results of purifying selection, and thus likely to be functional. It is worth noting that, after excluding the 22 359 m^{6}A^{++} sites, the remaining 155 639 human m^{6}A sites are considered unlabeled sites (rather than unconserved sites), we and refer them as m^{6}A^{+-} sites in the next for simplicity. Some of them may still be conserved but

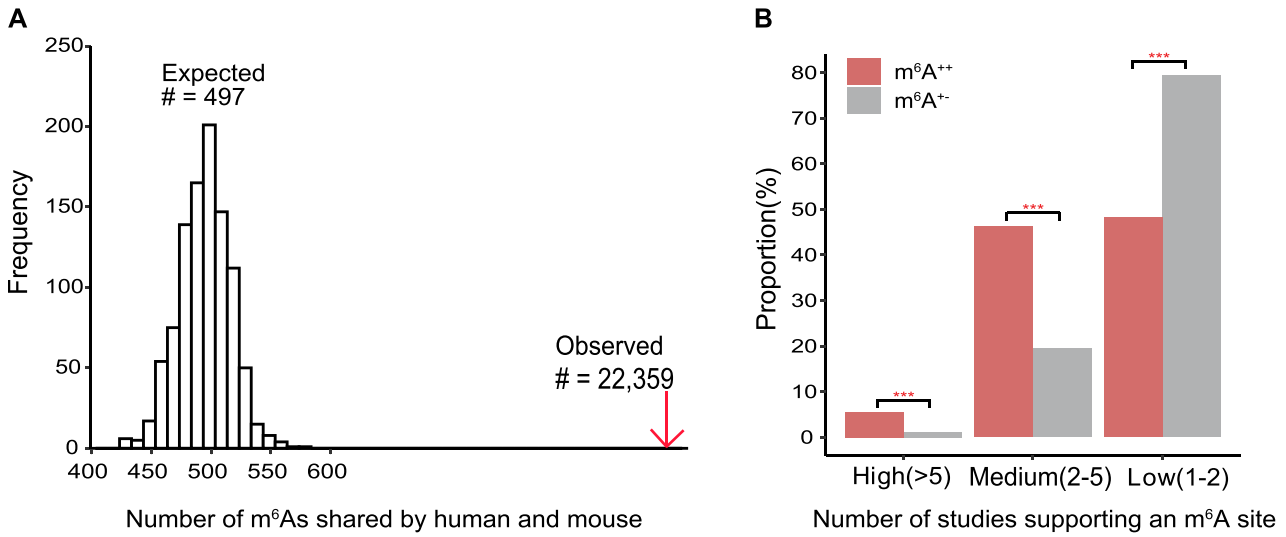

Figure 2. Comparison between characteristics of m^6A^{++} and m^6A^{+-} sites. A. The number of m^6A sites shared between human and mouse transcriptome (22 359) is significantly greater than expected by chance (497), suggesting that a majority of the m^6A^{++} sites are results of purifying selection, and thus likely to be functional. B. m^6A^{++} sites have a much higher proportion to be captured by multiple datasets, with high-record level (1232 m^6A^{++} sites, 5.51%, $P < 0.001$, Chi-squared test) and medium-record level (10 335 m^6A^{++} sites, 46.22%, $P < 0.001$, Chi-squared test), compared with m^6A^{+-} sites (high-record level: 1736 m^6A^{+-} sites, 1.12%; medium-record level: 30,385 m^6A^{+-} sites, 19.52%).

not positively reported from our analysis due to limited coverage of the mouse epitranscriptome.

The conserved m⁶A sites are more likely to be captured by multiple studies

A number of m^6A profiling techniques have been developed, and it was shown previously that the detection of RNA modification sites is sensitive to the specific profiling technique used [72–75]. It may be reasonable to speculate that the conserved m^6A sites (m^6A^{++} sites) can be more robustly detected by multiple m^6A profiling studies due to their prominence presence. To test this hypothesis, we systematically evaluated the record time (the number of studies supporting a specific m^6A site) between 22 359 m^6A^{++} sites and the rest of 155 639 unlabeled m^6A sites (m^6A^{+-} sites). Interestingly, we observed that m^6A^{++} sites have a much higher proportion in both high-record level (1232 m^6A^{++} sites, 5.51%; $P < 0.001$, Chi-squared test) and medium-record level (10 335 m^6A^{++} sites, 46.22%; $P < 0.001$, Chi-squared test), compared with m^6A^{+-} sites (high-record level: 1736 m^6A^{+-} sites, 1.12%; medium-record level: 30 385 m^6A^{+-} sites, 19.52%; Figure 2B), suggesting that record time can be used to distinguish more conserved m^6A sites among all detected m^6A sites, and it was thus integrated into our scoring metrics.

Performance evaluation of the machine learning component

Two machine learning models were developed to predict the degree of conservation of individual m^6A sites in human and mouse transcriptome, respectively, using the positive dataset m^6A^{++} sites and reliable negative dataset under the positive-unlabeled learning framework. They integrate both genomic features and the conventional sequence-derived features for enhanced predictive capability (Please refer to the detailed description in the section MATERIALS AND METHODS).

The performance of the newly constructed machine learning predictors was evaluated by a 5-fold cross validation and an independent testing dataset. As shown in Table 2, our predictors achieved reasonable prediction performance with AUROC of 0.840 and 0.829 in the prediction of the conserved m^6A sites for human and mouse, respectively, without relying on additional epitranscriptome profiling data. The predictive performance of our models is higher in human (0.840) than in mouse (0.829), which might be because of more complete data collection and predictive feature construction for human machine learning model. The prediction results were then integrated with other evidence to construct a more reliable indicator for the conservation of RNA methylation site.

Feature ranking was performed to identify the most effective genomic features used for labeling the conserved m^6A sites in human and mouse transcriptome, respectively (Supplementary Figure S1). The top two most critical features for both human and mouse models are PC_101bp (average phastCons scores within the flanking 101 bp) and PC_1bp (phastCons scores of the nucleotide), suggesting that epitranscriptome conservation partially contributed to the conservation of genome. Besides, among the top 10 most important features, three other features were shared by both human and mouse models, including clust_A_f100 (count of neighboring A within 201 nt window), clust_A_f1000 (count of neighboring A within 2001 nt window) and long_exon (exon length > = 400 bp), indicating the importance of clustering effects and long exons in identifying the conserved m^6A sites.

In addition, sequence similarity, tissue-specific mapping and phastCons 100-way conservation scores calculated for human genome were previous applied for conservation-related studies [61, 88, 101]. These elements were also used to extract conservation information of m^6A RNA modification sites and incorporated into ConsRM score.

Assessing the conservation of individual RNA methylation sites with ConsRM score

The ConsRM score was devised to quantify the degree of conservation of individual m^6A RNA methylation sites by integrating the scores inferred from six different sources, including positional mapping, tissue-specific mapping, supports from multiple

Table 2. Performance evaluation of proposed machine learning approaches

Species	Testing method	Evaluation metrics				
		Sn	Sp	ACC	MCC	AUROC
Human	Cross-validation	0.769	0.747	0.758	0.516	0.840
	Independent testing	0.772	0.747	0.759	0.519	0.840
Mouse	Cross-validation	0.732	0.757	0.745	0.489	0.832
	Independent testing	0.726	0.759	0.742	0.485	0.829

Note: 80% of m^6A^{++} sites were used for training, while its performance was tested on the rest of 20% m^6A^{++} sites as independent testing dataset. We assigned 10 m^6A^{+-} sites, obtained from reliable negative (RN) dataset, for each m^6A^{++} sites as the negative data. A total of 10 independent predictors were constructed with balanced 1:1 positive–negative ratio with their performance averaged to fully take advantage of the unbalanced positive and negative data.

techniques, sequence similarity, machine learning modeling and genome conservation (see Figure 3). As these scores should all contribute positively to the conservation of m^6A sites, and they are positively correlated with each other. The most correlated pair of sources are positional mapping (*b) and machine learning modeling (*f) with Pearson's correlation of 0.75.

It is worth noting that the conservation of m^6A modification is different from the conservation of the genome. The conservation of an m^6A site strictly requires the conservation of the relevant genomic locus, but the opposite is not true. A brief comparison of the genomic conservation (phastCons score) and the epitranscriptome conservation (ConsRM score) can be found in Figure 3 (*a versus *g), with Pearson's correlation of 0.67. In fact, the score from machine learning model is most correlated (Pearson's correlation of 0.93) to the ConsRM score, with sequence similarity being ranked at the second place (Pearson's correlation of 0.75). It is important to note that, although machine learning model alone could report scores highly correlated to ConsRM, it should not be used to replace ConsRM score, because ConsRM integrates significant amount of information at single site level, e.g. direct validation of conservation from positional mapping (*b) and tissue-specific mapping (*c) reported by existing epitranscriptome datasets.

The newly proposed ConsRM scoring framework was applied to all the 177 998 m^6A RNA methylation sites in human, based on which, these sites were further stratified into three groups according to their conservation level, i.e. the top 30% (53 399) sites of high conservation, 30–60% (53 399) sites of medium level of conservation, and last 40% (71 200) sites of low conservation. In order to examine the potential functionality associated with conservation (or test the reliability of the proposed scoring mechanism in identifying functional m^6A sites), several experiments were performed in the following. We showed that ConsRM score can effectively predict the conserved m^6A sites not yet revealed by existing studies by comparing to the epitranscriptome datasets in mouse, rat and zebrafish, and the m^6A sites with higher ConsRM scores are more likely to be functionally important, i.e. less likely to be affected by germline mutations and more likely to fall within the binding regions of various RNA-binding proteins, especially m^6A readers.

ConsRM score can effectively evaluate conservation degree of individual m^6A sites, with implications of its functionality

ConsRM predicts more conserved m^6A sites not yet supported by existing studies, compared with genomic conservation scores phastCons and phyloP

It is worth noting that although a large number of m^6A sites have been identified in human and mouse, it is reasonable to believe that the unveiled m^6A epitranscriptome is still incomplete. Thus, it is likely that many genuine m^6A sites remain to be uncovered: in other words, a significant proportion of the current m^6A^{+-} sites may turn out to be conserved (actually m^6A^{++} sites) as more data are accumulated. It is therefore interesting to test whether the newly developed ConsRM score has the potential to predict the conserved m^6A sites not supported by existing datasets.

For this purpose, all mouse m^6A datasets identified by m^6A-CLIP-seq (mouse experiments 9–14, Supplementary Sheet S1) were completely excluded when calculating the ConsRM score of human m^6A sites, and using as independent testing purpose. We focused on only the m^6A^{+-} sites in this testing experiment, assuming that m^6A^{+-} sites with higher ConsRM score are more m^6A conserved and thus more likely to be identified in the independent testing datasets. Similarly, the human m^6A^{+-} sites were classified into three groups according to their ConsRM score (top 30%, 30–60% and last 40%), and the RNA methylation status of their corresponding coordinates in mouse transcriptome was examined by comparing with the omitted six new mouse datasets.

Interestingly, we observed that 8180 (19.60% of the conserved As) m^6A^{+-} sites from high conservation level can now be mapped to experimentally observed m^6A sites in the new mouse m^6A-CLIP-seq datasets, compared with 2377 (5.07% of the conserved As) from the medium conservation level and 247 (1.31% of the conserved As) m^6A^{+-} sites from low conservation level, respectively (Table 3). For performance comparison, we performed this analysis using phastCons100way.UCSC.hg19 [88] and phyloP100way.UCSC.hg19 [102], respectively, which are well-known genomic conservation scores that presenting a high coverage, near base-resolution of nucleotide conservation. Similarly, the human m^6A^{+-} sites were also classified into three groups by phastCons and phyloP, respectively. Although it is clear that, although phastCons and phyloP can also convey some information related to epitranscriptome layer conservation, ConsRM is significantly more effective (phastCons: 6275, 14.93% of conserved m^6As, $P < 0.001$, Chi-squared test; phyloP: 5468, 11.96% of conserved m^6As, $P < 0.001$, Chi-squared test, compared with ConsRM: 8180, 19.60%). Meanwhile, at medium and low levels, phastCons and phyloP made more false positive predictions (m^6A with a low conservation score later confirmed to be conserved by new experiments). Taken together, these results suggest that the proposed ConsRM framework is more effectively in distinguishing the conserved and unconserved m^6A sites by directly learning from the partially available m^6A epitranscriptomes. Please note that we have excluded the impact of genome conservation in the analysis by considering only the conserved As in the above analysis. Because the unconserved loci were not considered in this analysis, the

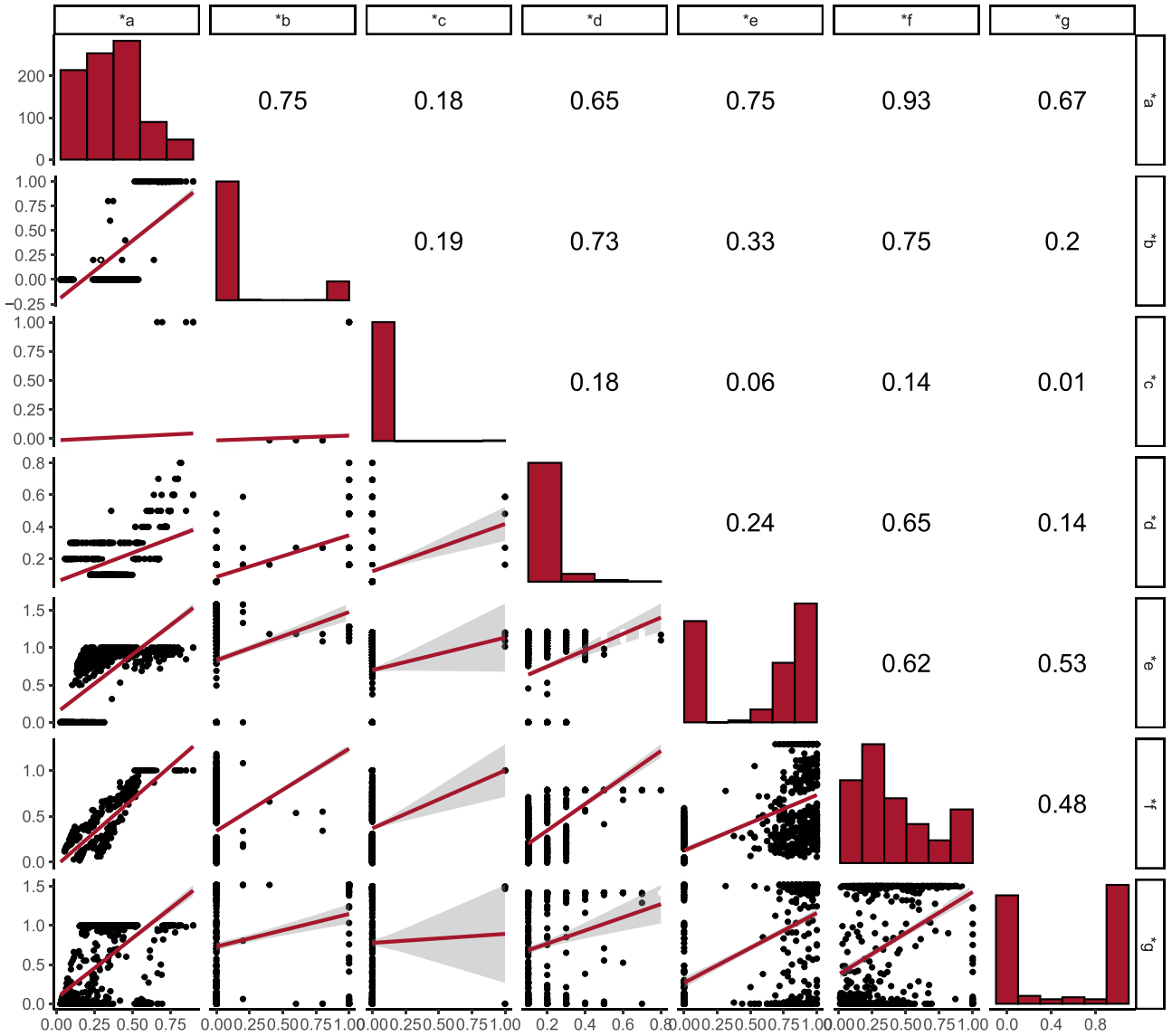

Figure 3. ConsRM score for quantifying the degree of conservation of individual m^{6}A sites. ConsRM score (*a) integrates the information from six different sources, including positional mapping (*b), tissue-specific mapping (*c), supports from multiple studies (*d), sequence similarity (*e), machine learning modeling (*f) and genome conservation (*g). We showed the overall distribution pattern of individual source with histogram at the diagonals. The scatter plots and the Pearson's Correlations of two arbitrary sources were shown in the corresponding position of the lower left and upper right triangle, respectively.

impact of cross-species genome conservation was eliminated. Please see Figure 4 for detailed schematic representation of this experiment.

m^{6}A sites with higher ConsRM scores are more likely to be m^{6}A sites in a third species

We further examined whether the highly scored m^{6}A sites are more likely to be conserved in a third species, with the hypothesis that evolutionarily conserved m^{6}A modifications have a greater probability to be functional. A total of 6348 rat and 63 998 zebrafish m^{6}A sites were collected for this purpose (Supplementary Sheet S1). We also collected the sites that are conserved as A but not known to be methylated as the negative control.

Consistent with previous testing results, the m^{6}A sites with higher ConsRM scores are more likely to be A and m^{6}A in a third species (see Table 4). Specifically, 51 403 (97.17%) human m^{6}A sites from high conservation level are As in rat, and 746 (1.39%) are also experimentally validated m^{6}A sites, compared with m^{6}A sites from medium conservation level (83.59% and 0.55% are or m^{6}A sites, respectively) and m^{6}A sites of low conservation level (31.25% are As, 0.11% are m^{6}A sites). A very similar pattern was observed in zebrafish with 45.91%, 35.36% and 6.26% of human m^{6}A sites are A for the sites of high, medium and low conservation level, respectively, and 3.80%, 1.25%, 0.18 are m^{6}A sites.

Again, to exclude the impact of genome conservation, we used the human m^{6}A sites that are As (rather than m^{6}A) in the third species as a negative control, and calculated the ratio between the conserved m^{6}A sites and other As. As shown in Table 4, this ratio still increases clearly with the conservation score from 0.35% (for the low conservation set) to 1.45% (for high conservation set), and a similar pattern can be observed in zebrafish from 2.83% (for lowly conserved m^{6}A sites) to 8.27% (for highly conserved m^{6}A sites), suggesting the ConsRM score

Table 3. ConsRM predicts the conserved m^{6}A sites not supported by existing datasets

Conservation level	Site #	Method	# of m^{6}A^{+-} sites	# of A found in independent mouse dataset (%)	# of newly confirmed m^{6}A $^{++}$ sites in independent mouse dataset (%)	Ratio of newly confirmed m^{6}A and A
High	53 399	ConsRM	41 772	41 743 (99.93%)	8108 (19.58%)	19.60%
		PhastCons	46 479	42 028 (90.42%)	6275 (13.50%)	14.93%
		phyloP	47 455	45 716 (96.34%)	5468 (11.52%)	11.96%
Medium	53 399	ConsRM	53 399	46 925 (87.88%)	2377 (4.45%)	5.07%
		PhastCons	50 357	35 611 (70.72%)	2796 (5.55%)	7.85%
		phyloP	49 595	33 116 (66.77%)	3351 (6.76%)	10.12%
Low	71 200	ConsRM	71 200	18 828 (26.44%)	247 (0.35%)	1.31%
		PhastCons	69 535	29 857 (42.94%)	1661 (2.39%)	5.56%
		phyloP	69 321	28 664 (41.35%)	1913 (2.76%)	6.67%

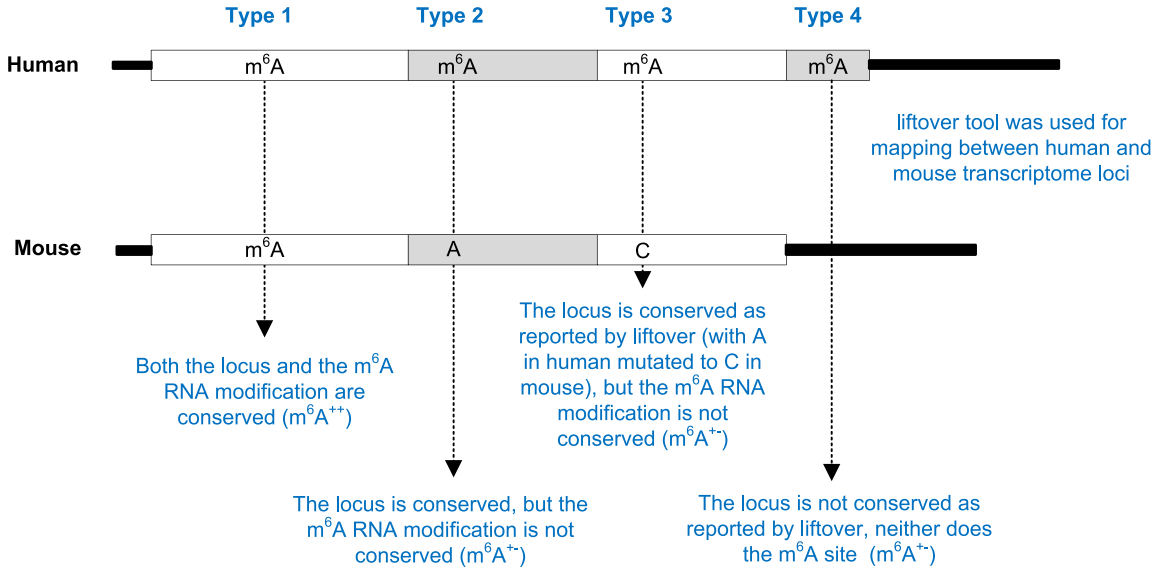

Figure 4. Schematic representation of the mouse experiment. These exist four different types of m^{6}A sites according to their conservation status. All mouse m^{6}A datasets generated by m^{6}A-CLIP-seq were used as independent testing data, and were excluded when calculating the ConsRM score for human m^{6}A sites. We then find the corresponding coordinates of human m^{6}A^{+-} sites in mouse transcriptome, and compare them with the independent mouse dataset to assess the capability of different scoring frameworks in predicting the previously unknown conserved m^{6}A sites. To elimiate the confounding factor of genome conservation, we considered primarily the ratio between type 1 m^{6}A^{++} sites and the type 2 m^{6}A^{+-} sites when reporting the results (last column in Table 3).

Table 4. Conservation of human m^{6}A sites with different ConsRM score in rat and zebrafish

ConsRM Score	#	Rat				Zebrafish			
		A	m^{6}A	Ratio	P-value	A	m^{6}A	Ratio	P-value
High	53 399	51 403 (97.17%)	746 (1.39%)	1.45%	High-Medium ***	24 517 (45.91%)	2028 (3.80%)	8.27%	High-Medium ***
Medium	53 399	44 606 (83.59%)	292 (0.55%)	0.66%	Medium-Low ***	18 884 (35.36%)	669 (1.25%)	3.54%	Medium-Low **
Low	71 200	22 253 (31.25%)	79 (0.11%)	0.35%	High-Low ***	4456 (6.26%)	126 (0.18%)	2.83%	High-Low ***

Note: '***' indicates a significance level of $P < 0.01$, while '**' indicates $P < 0.05$.

successfully captured the epitranscriptome conservation. Taken together, these results confirmed the effectiveness of the newly proposed ConsRM framework in assessing the degree of conservation of individual RNA methylation sites in human and mouse, and extrapolating that knowledge to other species.

It is worth mentioning that a higher proportion of human m^{6}A sites are observed to be conserved in zebrafish compared with rat. This probably resulted from the incomplete and imbalanced data collection (6348 rat m^{6}A sites and 63 998 zebrafish m^{6}A sites, see Supplemental Sheet S1). Currently, epitranscriptome data generated from base-resolution technique are still highly scarce. It is reasonable to speculate that the observed degree of conservation should still increase substantially as more data are available.

SNP density analysis clearly differentiated m^{6}A sites of different ConsRM score

After showing that the proposed scoring system can effectively distinguish the conserved and unconserved m^{6}A sites, we next

tested whether there exists a clear association between conservation and biological function.

Previous work suggests that mutations in m^{6}A consensus motifs are suppressed in human cancer cells [103]. We first calculated the SNP density for m^{6}A sites of different conservation levels by checking if the m^{6}A modifications can be destroyed by a genetic mutation occurring within the 5 bp flanking windows centered with the m^{6}A sites (or alter the m^{6}A-forming motif DRACH). Since the conserved m^{6}A sites are likely to be functional and thus subject to purifying selection, their SNP density was predicted to vary from that of unconserved one. Two types of genetic mutations were considered: germline and somatic. Germline mutations occur in gametes and can be passed onto offspring, while somatic mutations result from the damage to genes in an individual cell during a person's life and are thus not inheritable. Notably, as shown in Table 5, the density of both somatic and germline mutations differs in m^{6}A sites of different conservation levels. Interestingly, distinct patterns were observed between germline and somatic mutations. While the density of germline mutation decreases slightly with increasing conservation of m^{6}A sites (with 3.14%, 2.93% and 2.57% for the m^{6}A sites with low, medium and high conservation level, respectively), somatic mutation exhibited an increasing trend (with 5.22%, 13.34% and 16.83% for the m^{6}A sites with low, medium and high conservation level, respectively). These results suggest that the more conserved m^{6}A sites were less affected by germline mutations when compared with the less conserved sites, and the cancer-caused somatic mutations are more likely to occur around the conserved m^{6}A sites and thus destroy the relevant function. This suggests that these most conserved m^{6}A sites may play a more critical role in various biological processes compared with other sites. Meanwhile, we found that although only a very small number of germline mutations were related to highly conserved m^{6}A sites likely as the result of purifying selection, a higher proportion of them were predicted to have highly deleterious consequences, compared with germline variants localized to the less conserved m^{6}A sites. Lastly, all germline and somatic mutations associated with different conservation levels of m^{6}A sites were mapped with the disease-causing TagSNPs obtained from Johnson and O'Donnell, GWAS catalog, and ClinVar database. We observed that germline mutations associated with highly scored m^{6}A sites overlapped more frequently with disease-causing TagSNPs, indicating that m^{6}A sites with a higher ConsRM score have a larger propensity to cause disease, presumably because they are more likely to be intimately involved in functional regulatory events. Taken together, we systemically evaluated the conservation degree of a large amount of experimentally validated m^{6}A modifications, using our newly developed scoring system. The testing results obtained from germline and somatic mutations suggested that m^{6}A sites with higher ConsRM score are generally more functionally important, which in turn proved that the ConsRM score is able to quantify the conservation of m^{6}A sites, or identify functional m^{6}A RNA methylation sites.

Conserved m^{6}A sites may mediate more RNA-binding protein interactions

It was found previously that m^{6}A RNA methylation may recruit RNA-binding proteins that are closely associated with posttranscriptional regulations [23]. We examined possible interactions between RNA binding proteins and m^{6}A sites of different conservation levels by checking whether they localized within the regions of RBP binding sites (Table 6). Most m^{6}A sites from the highly conserved group are related to at least one RNA binding protein (46 668 m^{6}A sites, 87.48%). This proportion dropped to 63.67% for m^{6}A sites with lowest conservation level. Moreover, the same tendency was also observed for five different types of m^{6}A reader. For example, 12 089 (22.66%) m^{6}A sites from the high conservation group were localized in the binding regions of YTHDF1, while only 5839 (8.21%) from the low conservation group were associated with this m^{6}A reader. Thus, this finding supports the reliability of the proposed scoring mechanism for finding biological meaningful m^{6}A sites.

Case study

Glioblastoma (GBM) is as the most devastating primary malignant brain tumor in human, as recurrence is nearly inevitable even after modern surgery and highly aggressive therapies [104]. Many studies have shown that FOXM1 is overexpressed and plays a key role in GBM proliferation and regulation [105–108]. In a more recent study, Zhang *et al*. [109] and colleagues uncovered and provided insight into the important roles of m^{6}A modification in GBM. Specifically, ALKBH5, an m^{6}A eraser, is elevated in GBM and hence facilitates the expression of oncogene FOXM1 through its demethylation activity. We therefore examined the distribution and conservation degree of m^{6}A methylation on gene FOXM1 using our ConsRM score. In total, we found 33 experimental validated m^{6}A sites identified by five high-throughput sequencing techniques using non-tumor cell lines (Figure 5). Among them, 17 m^{6}A sites (51.52%) are classified into high conservation level (with maximal ConsRM score: 0.7667, top 1%). We then extracted all the human genes of a similar size (± 5% size of gene FOXM1) using R package TxDb.Hsapiens.UCSC.hg19.knownGene. As shown in Supplementary Sheet S3, among the 490 genes of similar size as FOXM1, we observed an average of 6.722 m^{6}A sites with 15.82% of them being classified as highly conserved, compared to FOXM1 with 33 m^{6}A sites ($P = 0.00816$) and 51.52% ($P = 0.08367$) are of high conservation level (Figure 5). Consistent with previous studies, our finding suggested that the human oncogene FOXM1 is highly associated with m^{6}A modification.

Functional characterization of the most conserved m^{6}A sites

We further extracted the top 1000 most conserved m^{6}A sites (with the largest ConsRM score), and examined their putative functional relevance with Gene Ontology Enrichment Analysis of their hosting genes. We showed in Figure 6A the top 20 biological processes enriched with the most conserved m^{6}A sites, many of which have been confirmed to be closely associated with m^{6}A RNA methylation in recent studies, including but not limited to: chromatin remodeling [110, 111] (covalent chromatin modification, P-value = 3.69E-05; chromatin remodeling, P-value = 9.87E-05),splicing [39] (mRNA splicing via spliceosome, *P*-value = 1.37E-04, RNA splicing, *P*-value = 4.09E-04), DNA damage [13] (cellular response to DNA damage stimulus, *P*-value = 9.47E-04), synapse [112] (postsynaptic density protein 95 clustering, *P*-value = 1.56E-03, layer formation in cerebral cortex, *P*-value = 1.97E-03, protein localization to synapse, *P*-value = 1.97E-03). The complete results are available in Supplementary Excel Sheet S4. Meanwhile, we also plotted the overall distribution of the most conserved m^{6}A sites on mRNAs. Consistent with our knowledge of this modification, these sites were also enriched near the stop codons of the mRNAs (Figure 6B).

Table 5. SNP density analysis for m^{6}A sites with different degrees of conservation

Mutation type	Conservation level	Number of m^{6}A sites	SNPs within +−2 bp (motif)		SNPs with deleterious level > 3		Disease- associated SNPs	
Germline mutation	High	53 399	1374 (2.57%)	High-Medium***	45 (3.28%)	High-Medium**	115 (8.37%)	High-Medium-
	Medium	53 399	1567 (2.93%)	Medium-Low**	32 (2.04%)	Medium-Low***	125 (7.98%)	Medium-Low**
	Low	71 200	2238 (3.14%)	High-Low***	7 (0.31%)	High-Low***	133 (5.94%)	High-Low***
Somatic mutation	High	53 399	8986 (16.83%)	High-Medium***	1656 (18.43%)	High-Medium***	42 (0.47%)	High-Medium-
	Medium	53 399	7126 (13.34%)	Medium-Low***	1040 (14.59%)	Medium-Low***	22 (0.31%)	Medium-Low-
	Low	71 200	3718 (5.22%)	High-Low***	164 (4.41%)	High-Low***	11 (0.29%)	High-Low-

Note: The effectiveness of ConsRM in distinguishing disease-associated somatic mutation is not significant maybe partially due to very limited number of such mutations. '***' indicates a significance level of $P < 0.01$, '**' indicates $P < 0.05$, while '-' indicates $P > 0.1$. High-Medium represents a significance level between high and medium group, Medium-Low for medium and low group, and High-Low for high and low group.

Table 6. Number of RBP-related m^{6}A sites

Conservation Level	RNA-binding proteins					
	All RBPs	YTHDF1	YTHDF2	YTHDF3	YTHDC1	YTHDC2
High	46 668 (87.48%) H-M***	12 089 (22.66%) H-M***	13 939 (26.19%) H-M***	2499 (4.68%) H-M***	6393 (11.98%) H-M***	427 (0.80%) H-M***
Medium	41 169 (77.17%) M-L***	5254 (9.85%) M-L***	7131 (13.37%) M-L***	1223 (2.29%) M-L-	3602 (6.75%) M-L***	313 (0.59%) M-L***
Low	45 288 (63.67%) H-L***	5839 (8.21%) H-L***	8608 (12.10%) H-L***	1546 (2.17%) H-L***	3153 (4.43%) H-L***	474 (0.67%) H-L*

Note: H-M represents a significance level between high and medium group, M-L for medium and low group and H-L for high and low group. '***' indicates a significance level of $P < 0.01$, '**' indicates $P < 0.05$, '*' indicates $P < 0.1$, while '-' indicates $P > 0.1$.

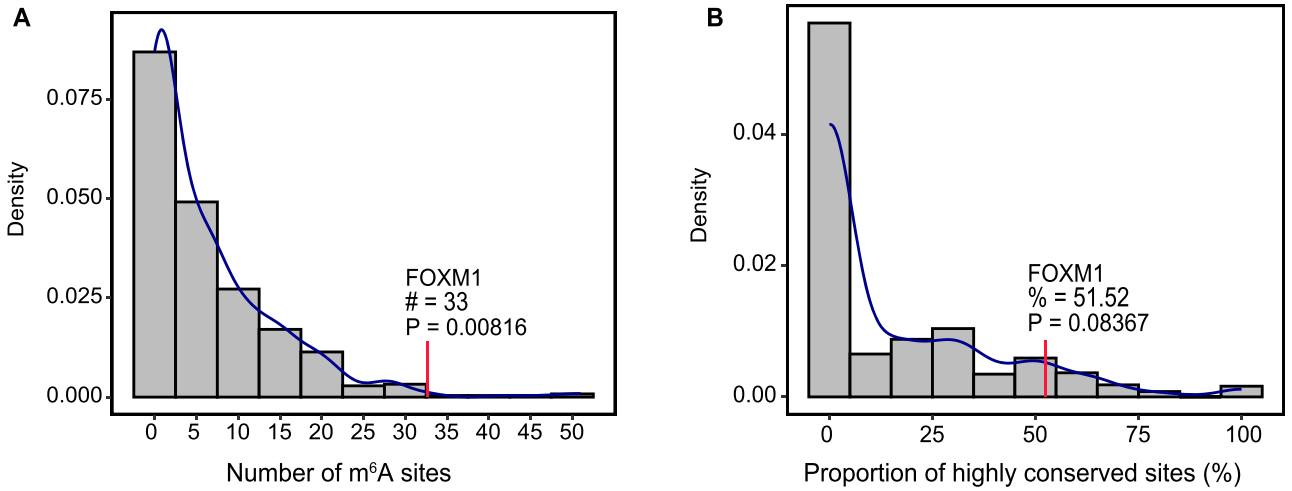

Figure 5. Comparison of m^{6}A distribution and characteristics between human oncogene FOXM1 and other genes of similar size. A. For the 490 genes of similar size as FOXM1 (± 5% size), an average of 6.722 m^{6}A sites were observed, compared with human oncogene FOXM1 (33 sites, $P = 0.00816$). B. An average, 15.82% of m^{6}A sites were classified as highly conserved, this ratio increased to 51.52% ($P = 0.08367$) for FOXM1. More details can be found in Supplementary Sheet S3.

Integrating multiple epitranscriptomes with MultiScore framework

The ConsRM framework previously defined considered only the conservation between human and mouse epitranscriptome for the following reasons: first, only limited epitranscriptome data are currently available in other species. A fairly comprehensive epitranscriptome data collection was performed in this study, and we have collected 27 human samples and 19 mouse samples; while only three rat samples and six zebrafish samples were found (see Supplementary Excel Sheet S1). Limited number of samples directly mean limited epitranscriptome coverage, and may lead to increased biological and technical bias in further analysis. Secondly, the ConsRM score based on well-characterized epitranscriptomes of human and mouse alone already produced reasonable results that are superior to

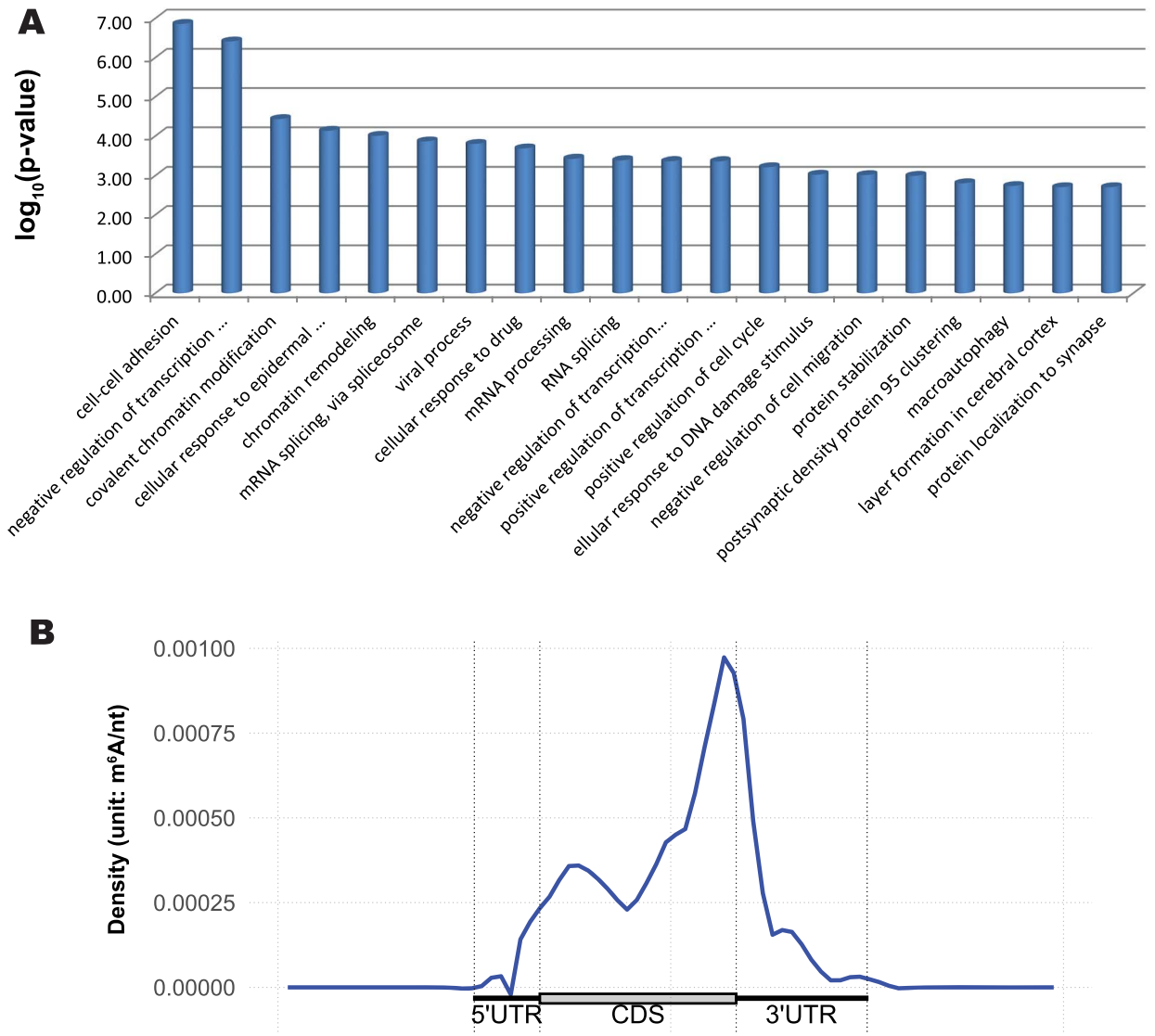

Figure 6. Characterization of the most conserved m^{6}A sites. A. Biological processes associated with the most conserved m^{6}A sites. The complete enrichment analysis results are available in Supplementary Excel Sheet S4. The GO analysis results were obtained by using DAVID web server [113]. B. Distribution of the most conserved m^{6}A sites on mRNAs. Consistent with our understandings of this modification, these most conserved m^{6}A sites were also shown to be enriched near the stop codons of the mRNA. There may be a second peak after the start codon, but it is not clear whether this is real pattern or simply noise. The metagene plot was generated using MetaTX R package [114] under its default setting to measure the absolute density of m^{6}A sites on standardized mRNA model in the presence of isoform transcripts with the unit number of m^{6}A sites per nucleotide of mRNA.

the well-adopted conservation metrics (phyloP and phastCons), as was shown in our tests involving mouse, rat and zebrafish datasets.

Nevertheless, a more general framework that can integrate all available epitranscriptomes in different species is desirable. For this purpose, a more general metrics termed ConsRM-MultiScore was proposed by extending our original ConRM formulation. It averages multiple pair-wise conservation scores calculated between the epitranscriptome to be evaluated and all other epitranscriptomes in different species to obtain a more general overview via the integration of the all the available epitranscriptome datasets. The ConsRM-MultiScore is also provided in ConsRM website, which was calculated from all four species considered in this study, i.e. human, mouse, rat and zebrafish. However, since the epitranscriptomes of rat and zebrafish are likely of limited coverage due to small number of experimental samples available, which may produce technical and biological bias, the score was not used as the primary metrics in the ConsRM database. However, as more and more epitranscriptome datasets are produced in species other than human and mouse, the MultiScore framework is likely to be more important in the near future.

The ConsRM website

To effectively share our findings, the ConsRM website was developed to serve as a centralized online platform for deciphering the evolutionary conservation of individual m^{6}A RNA methylation sites (see Figure 7). It features a comprehensive collection of 177 998 experimentally validated human m^{6}A RNA methylation sites along with their ConsRM scores and various conservation-related metrics, which directly reflect their evolutionary conservation and potential functionality (Figure 7A–C). Additionally, a web server was constructed for calculating the ConsRM

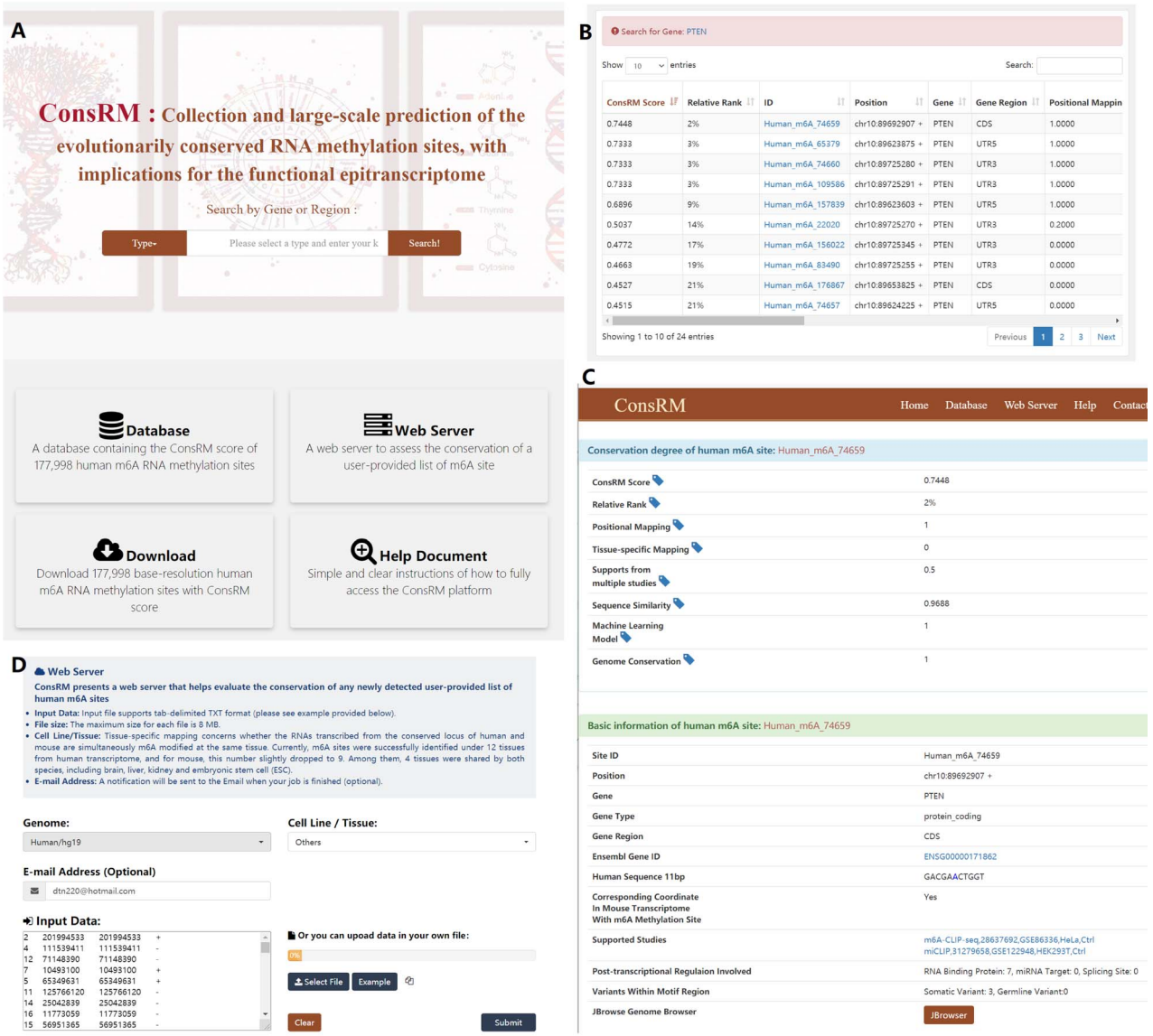

Figure 7. Contents of ConsRM online platform. A. Homepage of ConsRM website, user can query the conservation metrics of 177 998 experimentally validated human m^6A sites by gene or region. B. An example. Search for 'PTEN' returns 24 m^6A sites located on the PTEN transcripts. Four of them are highly conserved, among the top 5% most conserved m^6A sites. C. The detailed conservation metrics and other regulatory information of the most conserved m^6A site on PTEN, including detailed experiment information and posttranscriptional regulations involved such as miRNA target, splicing site and RNA-binding protein interaction. D. A web server was built to calculate the ConsRM score for any newly identified human m^6A sites provided by user. A notification Email can be sent after the conservation assessment job is completed.

scores for any newly detected human m^6A RNA methylation sites provide by users (Figure 7D), with the statistical significance of the conservation score being assessed by comparing to all the experimentally validated m^6A sites collected in the ConsRM database via the relative ranking, e.g. top 1% of all experimentally validated m^6A sites for extremely conserved sites or 80% for lowly conserved sites. The m^6A RNA methylation sites were also annotated to specify their interaction with posttranscriptional machinery (RBP-binding, microRNA interaction and splicing sites), with the JBrowse Genome Browser properly set to enable the direct exploration of the genomic regions of interests, and detailed help documents were also provided. All the materials presented in ConsRM web can be freely downloaded. The ConsRM website is freely available at: https: //www.xjtlu.edu.cn/biologicalsciences/con. We believe ConsRM will make a very useful resource for conservation and functional studies of m^6A RNA methylation sites.

Conclusions and discussions

N6-methyladenosine (m^6A) is the most prevalent and abundant RNA modification on mRNAs and lncRNAs with increasing evidence supporting its crucial importance in essential molecular mechanisms and various diseases. Given that functionally important m^6A sites increase organismal fitness and hence are more likely to be conserved during evolution, we developed a novel scoring framework to quantitatively measure the degree of conservation of individual m^6A RNA methylation

sites. It integrates positional mapping, tissue-specific mapping, supported from multiple techniques, sequence similarity, machine learning modeling and genome conservation.

Importantly, we showed that the newly developed scoring framework can effectively predict the conserved and unconserved human m^{6}A sites in mouse and a third species, after excluding the confounding factor of genome conservation. By directly learning from the incomplete m^{6}A epitranscriptomes, ConsRM outperformed well-adopted conservation scores such as phastCons and phyloP in assessing epitranscriptome conservation. Further analysis revealed that germline mutations were less likely to affect the more conserved m^{6}A sites compared with the unconserved sites, while an opposite trend was observed on the somatic mutations, suggesting that dysregulation of the conserved m^{6}A sites was more likely to be associated with disease pathogenesis. Importantly, the conserved m^{6}A sites were also more likely to fall within the binding regions of various RNA-binding proteins, especially m^{6}A readers.

Due to limited data availability, tissue-specific mapping was not emphasized in the current version of ConsRM. Among the 56 epitranscriptomes we collected in this study, there exist only four pairs (totally eight) of matched tissues in different species, and all of them are between human and mouse. Instead of considering only matched tissues, the ConsRM framework acknowledges conservation in a broader sense, i.e. m^{6}A co-occurred on the homologous loci of different species, including in unmatched tissue types. We showed in Supplementary Table S6 that, although not as strong as the conservation in matched tissues (fold enrichment of 68.17), strong conservation was also observed between the epitranscriptomes of unmatched tissues (averaged fold enrichment of 62.11), suggesting that general conservation, which does not require matching of tissue type, is also of critical biological importance. Nevertheless, tissue specificity may play a more important role in epitranscriptome conservation analysis as increasing datasets are accumulated for matched tissues in different species.

To share our findings more effectively, we developed a user-friendly online platform ConsRM (https: //www.xjtlu.edu.cn/biologicalsciences/con), which contains 177 998 human m^{6}A RNA methylation sites along with their conservation metrics for conservation analysis and general functional prioritization. A web server was also provided to assess the conservation of a user-provided list of m^{6}A sites. To the best of our knowledge, this is the first efforts trying to unveil the functionality of every individual m^{6}A sites from the conservation perspective.

Future routes to improve the ConsRM scoring could include: (1) ConsRM incorporated only the epitranscriptome data from human and mouse without using data from other species such as rat and fly. This is primarily due to the limited availability of epitranscriptome data of high-quality. Currently, epitranscriptome data obtained from base-resolution technique are still very scarce for species other than human and mouse. It should be interesting how the proposed ConsRM framework can be extended to multiple species when epitranscriptome datasets are more abundantly available. (2) The proposed ConsRM framework was trained with only conservation labels, i.e. the presence or absence of the m^{6}A sites at the matched locus in human and mouse transcriptome. Considering that conservation and functionality are often highly correlated, additional evidence can be integrated into the ConsRM scoring framework for enhancing its capability of assessing and predicting the conservation and functionality of individual RNA methylation sites, including but not limited to the binding sites of RNA-binding proteins, the splicing sites and the SNPs.

Key Points

- In this study, we performed a comparative conservation analysis of the human and mouse m^{6}A epitranscriptomes at single site resolution.
- A novel scoring framework, ConsRM, was devised to quantitatively measure the degree of conservation of individual m^{6}A sites. ConsRM integrates multiple information sources and a positive-unlabeled machine learning framework, which integrated genomic and sequence features to trace subtle hints of epitranscriptome layer conservation.
- We showed that ConsRM outperformed well-adopted conservation scores (phastCons and phyloP) in distinguishing the conserved and unconserved m^{6}A sites. Additionally, the m^{6}A sites with a higher ConsRM score are more likely to be functionally important.
- We further developed an online database containing the conservation metrics of 177 998 distinct human m^{6}A sites to support conservation analysis and functional prioritization of individual m^{6}A sites. And it is freely accessible at: https: //www.xjtlu.edu.cn/biologicalsciences/con.

Data availability

The data underlying this article are available via https://www.xjtlu.edu.cn/biologicalsciences/con, and in its online supplementary material.

Author contributiion

J.M., D.J.R., J.P.M., J.S. and Z.W. conceived the idea and initialized the project; B.S. designed the research plan; K.C. collected and processed the experimentally validated m^{6}A sites; B.S. performed analysis related to conservation degree, machine learning model, and validation; Y.T. and B.S. built the website; B.S. drafted the manuscript. All authors read, critically revised and approved the final manuscript.

Supplementary data

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

Funding

This work has been supported by the National Natural Science Foundation of China (31671373); XJTLU Key Program Special Fund (KSF-T-01). This work is partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02).

Conflict of Interest

None declared.

References

1. Garcias Morales D, Reyes JL. A birds'-eye view of the activity and specificity of the mRNA m(6) a methyltransferase complex. *Wiley Interdiscip Rev RNA* 2021;**12**(1):e1618.
2. Chen YS, *et al*. Dynamic transcriptomic m(5) C and its regulatory role in RNA processing. *Wiley Interdiscip Rev RNA* 2021;e1639. https://doi.org/10.1002/wrna.1639.
3. McCown PJ, *et al*. Naturally occurring modified ribonucleosides. *Wiley Interdiscip Rev RNA* 2020;**11**(5):e1595.
4. Boccaletto P, *et al*. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res* 2018;**46**(D1):D303–7.
5. Meyer KD, Jaffrey SR. Rethinking m(6)a readers, writers, and erasers. *Annu Rev Cell Dev Biol* 2017;**33**:319–42.
6. Niu Y, *et al*. N6-methyl-adenosine (m6A) in RNA: an old modification with a novel epigenetic function. *Genomics Proteomics Bioinformatics* 2013;**11**(1):8–17.
7. Desrosiers R, Friderici K, Rottman F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci U S A* 1974;**71**(10):3971–5.
8. Dominissini D, *et al*. Topology of the human and mouse m 6 a RNA methylomes revealed by m 6 A-seq. *Nature* 2012;**485**(7397):201.
9. Wang X, *et al*. N-6-methyladenosine modulates messenger RNA translation efficiency. *Cell* 2015;**161**(6):1388–99.
10. Slobodin B, *et al*. Transcription impacts the efficiency of mRNA translation via co-transcriptional N6-adenosine methylation. *Cell* 2017;**169**(2):326, e12–37.
11. Huang H, *et al*. Histone H3 trimethylation at lysine 36 guides m(6)a RNA modification co-transcriptionally. *Nature* 2019;**567**(7748):414–9.
12. Zhou J, *et al*. Dynamic m(6)a mRNA methylation directs translational control of heat shock response. *Nature* 2015;**526**(7574):591–4.
13. Xiang Y, *et al*. RNA m(6)a methylation regulates the ultraviolet-induced DNA damage response. *Nature* 2017;**543**(7646):573–6.
14. Hao J, *et al*. The perturbed expression of m6A in parthenogenetic mouse embryos. *Genet Mol Biol* 2019;**42**(3):666–70.
15. Wang Y, *et al*. N(6)-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. *Nat Neurosci* 2018;**21**(2):195–206.
16. Wang Y, *et al*. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol* 2014;**16**(2):191–8.
17. Boissel S, *et al*. Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. *Am J Hum Genet* 2009;**85**(1):106–11.
18. Zhang C, *et al*. Hypoxia induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m(6)A-demethylation of NANOG mRNA. *Proc Natl Acad Sci U S A* 2016;**113**(14):E2047–56.
19. Zhang C, *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**(40):64527–42.
20. Lewis SJ, *et al*. Associations between an obesity related genetic variant (FTO rs9939609) and prostate cancer risk. *PLoS One* 2010;**5**(10):e13485.
21. Ma JZ, *et al*. METTL14 suppresses the metastatic potential of hepatocellular carcinoma by modulating N(6) -methyladenosine-dependent primary MicroRNA processing. *Hepatology* 2017;**65**(2):529–43.
22. Schumann U, Shafik A, Preiss T. METTL3 gains R/W access to the Epitranscriptome. *Mol Cell* 2016;**62**(3):323–4.
23. Liu J, *et al*. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat Chem Biol* 2014;**10**(2):93–5.
24. Schwartz S, *et al*. Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5 ' sites. *Cell Rep* 2014;**8**(1):284–96.
25. Yue Y, *et al*. VIRMA mediates preferential m(6)a mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov* 2018;**4**:10.
26. Wen J, *et al*. Zc3h13 regulates nuclear RNA m(6)a methylation and mouse embryonic stem cell self-renewal. *Mol Cell* 2018;**69**(6):1028, e6–38.
27. Ping XL, *et al*. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res* 2014;**24**(2):177–89.
28. Bokar JA, *et al*. Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. *RNA* 1997;**3**(11):1233–47.
29. Yue Y, Liu J, He C. RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev* 2015;**29**(13):1343–55.
30. Geula S, *et al*. Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science* 2015;**347**(6225):1002–6.
31. Xu K, *et al*. Mettl3-mediated m(6)a regulates spermatogonial differentiation and meiosis initiation. *Cell Res* 2017;**27**(9):1100–14.
32. Clancy MJ, *et al*. Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res* 2002;**30**(20):4509–18.
33. Hongay CF, Orr-Weaver TL. Drosophila inducer of MEiosis 4 (IME4) is required for notch signaling during oogenesis. *Proc Natl Acad Sci U S A* 2011;**108**(36):14855–60.
34. Zhong S, *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**(5):1278–88.
35. Jia G, *et al*. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol* 2011;**7**(12):885–7.
36. Zheng G, *et al*. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell* 2013;**49**(1):18–29.
37. Duan HC, *et al*. ALKBH10B is an RNA N(6)-Methyladenosine demethylase affecting Arabidopsis floral transition. *Plant Cell* 2017;**29**(12):2995–3011.
38. Martinez-Perez M, *et al*. Arabidopsis m(6)a demethylase activity modulates viral infection of a plant virus and the m(6)a abundance in its genomic RNAs. *Proc Natl Acad Sci U S A* 2017;**114**(40):10755–60.
39. Haussmann IU, *et al*. M(6)a potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature* 2016;**540**(7632):301–4.
40. Muller S, *et al*. IGF2BP1 promotes SRF-dependent transcription in cancer in a m6A- and miRNA-dependent manner. *Nucleic Acids Res* 2019;**47**(1):375–90.
41. Zhao BS, Roundtree IA, He C. Post-transcriptional gene regulation by mRNA modifications. *Nat Rev Mol Cell Biol* 2017;**18**(1):31–42.
42. Tang Y, *et al*. DRUM: inference of disease-associated m(6)a RNA methylation sites from a multi-layer heterogeneous network. *Front Genet* 2019;**10**:266.

43. Dominissini D, *et al*. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;**485**(7397):201–6.
44. Meyer KD, *et al*. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 2012;**149**(7):1635–46.
45. Linder B, *et al*. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 2015;**12**(8):767–72.
46. Ke S, *et al*. A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev* 2015;**29**(19):2037–53.
47. Chen K, *et al*. High-resolution N(6) -methyladenosine (m(6) a) map using photo-crosslinking-assisted m(6) a sequencing. *Angew Chem Int Ed Engl* 2015;**54**(5):1587–90.
48. Gjoneska E, *et al*. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 2015;**518**(7539):365–9.
49. Koh CW, Goh YT, Goh WS. Atlas of quantitative single-base-resolution N 6-methyl-adenine methylomes. *Nat Commun* 2019;**10**(1):1–15.
50. Zhang Z, *et al*. Single-base mapping of m6A by an antibody-independent method. *Sci Adv* 2019;**5**(7):eaax0250.
51. Garcia-Campos MA, *et al*. Deciphering the 'm(6)a code' via antibody-independent quantitative profiling. *Cell* 2019;**178**(3):731, e16–47.
52. Meyer KD. DART-seq: an antibody-free method for global m 6 A detection. *Nat Methods* 2019;**16**(12):1275–80.
53. Shu X, *et al*. A metabolic labeling method detects m(6)a transcriptome-wide at single base resolution. *Nat Chem Biol* 2020;**16**(8):887 -+.
54. Nie F, *et al*. RNAWRE: a resource of writers, readers and erasers of RNA modifications. *Database (Oxford)* 2020;**2020**:baaa049.
55. Liu S, *et al*. REPIC: a database for exploring the N(6)-methyladenosine methylome. *Genome Biol* 2020;**21**(1):100.
56. Deng S, *et al*. M6A2Target: a comprehensive database for targets of m6A writers, erasers and readers. *Brief Bioinform* 2020;1–11.
57. Han Y, *et al*. CVm6A: a visualization and exploration database for m(6)as in cell lines. *Cell* 2019;**8**(2):168.
58. Xuan JJ, *et al*. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res* 2018;**46**(D1):D327–34.
59. Liu H, *et al*. MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res* 2018;**46**(D1):D281–7.
60. Graur D, Zheng Y, Azevedo RB. An evolutionary classification of genomic function. *Genome Biol Evol* 2015;**7**(3): 642–5.
61. Malik R, Nigg EA, Korner R. Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics* 2008;**24**(12):1426–32.
62. Johnson JR, *et al*. Prediction of functionally important Phospho-regulatory events in Xenopus laevis oocytes. *PLoS Comput Biol* 2015;**11**(8):e1004362.
63. Xiao Q, *et al*. Prioritizing functional phosphorylation sites based on multiple feature integration. *Sci Rep* 2016;**6**:24735.
64. Waterhouse A, *et al*. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;**46**(W1):W296–303.
65. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007;**23**(15):1875–82.
66. Ma L, *et al*. Evolution of transcript modification by N(6)-methyladenosine in primates. *Genome Res* 2017;**27**(3):385–92.
67. Zhang Z, *et al*. Genetic analyses support the contribution of mRNA N(6)-methyladenosine (m(6)a) modification to human disease heritability. *Nat Genet* 2020;**52**(9):939–49.
68. Zhang H, *et al*. Dynamic landscape and evolution of m6A methylation in human. *Nucleic Acids Res* 2020;**48**(11):6251–64.
69. Liu Z, Zhang J. Most m6A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Mol Biol Evol* 2017;**35**(3):666–75.
70. Tang Y, *et al*. m6A-atlas: a comprehensive knowledge-base for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res* 2021;**49**(D1):D134–43.
71. Chen K, *et al*. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res* 2019;**47**(7):e41.
72. Adachi H, De Zoysa MD, Yu Y-T. *Post-transcriptional pseudouridylation in mRNA as well as in some major types of noncoding RNAs*. In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 2018;**1862**(3):230–239.
73. Zaringhalam M, Papavasiliou FN. Pseudouridylation meets next-generation sequencing. *Methods* 2016;**107**:63–72.
74. Hussain S, *et al*. Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol* 2013;**14**(11):215.
75. Capitanchik C, *et al*. How do you identify m(6) a methylation in transcriptomes at high resolution? A comparison of recent datasets. *Front Genet* 2020;**11**(398):398.
76. Zhang Z, *et al*. Single-base mapping of m(6)a by an antibody-independent method. *Sci Adv* 2019;**5**(7):eaax0250.
77. Chen X, *et al*. RNA methylation and diseases: experimental results, databases, web servers and computational models. *Brief Bioinform* 2019;**20**(3):896–917.
78. Song B, *et al*. PSI-MOUSE: predicting mouse Pseudouridine sites from sequence and genome-derived features. *Evol Bioinform Online* 2020;**16**:1176934320925752.
79. Chen W, *et al*. iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. *Mol Ther Nucleic Acids* 2018;**11**:468–74.
80. Bari ATMG, *et al*. DNA encoding for splice site prediction in large DNA sequence. *Springer Berlin Heidelberg* 2013;46–58.
81. Yang H, *et al*. iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. *J Comput Biol* 2018;**25**(11):1266–77.
82. Chen W, *et al*. RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes. *Sci Rep* 2016;**6**:31080.
83. Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N6-methyladenosine sites. *J Biomol Struct Dyn* 2017;**35**(3):683–7.
84. Zeng X, *et al*. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief Bioinform* 2020;**21**(4):1425–36.
85. Song B, *et al*. m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. *Bioinformatics* 2020;**36**(11): 3528–36.
86. Xiang S, *et al*. RNAMethPre: a web server for the prediction and query of mRNA m6A sites. *PLoS One* 2016;**11**(10):e0162707.
87. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**(3):1–27.

88. Siepel A, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**(8):1034–50.
89. Sherry ST, *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**(1):308–11.
90. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* 2015;**19**(1A):A68.
91. 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**(7):1073–81.
92. Adzhubei IA, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4):248–9.
93. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009;**19**(9):1553–61.
94. Shihab HA, *et al*. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;**34**(1):57–65.
95. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**(16):e164–4.
96. Buniello A, *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2018;**47**(D1):D1005–12.
97. Johnson AD. C.J. O'Donnell, *An open access database of genome-wide association results*. *BMC Med Genet* 2009;**10**:6.
98. Landrum MJ, *et al*. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2015;**44**(D1):D862–8.
99. Li J-H, *et al*. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2013;**42**(D1):D92–7.
100. Buels R, *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 2016;**17**:66.
101. Ma L, *et al*. Evolution of transcript modification by N6-methyladenosine in primates. *Genome Res* 2017;**27**(3):385–92.
102. Pollard KS, *et al*. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**(1):110–21.
103. An M, Wang H, Zhu Y. Mutations in m6A consensus motifs are suppressed in the m6A modified genes in human cancer cells. *PLoS One* 2020;**15**(8):e0236882.
104. Wen PY, Kesari S. Malignant gliomas in adults. *N Engl J Med* 2008;**359**(5):492–507.
105. Kim SH, *et al*. EZH2 protects glioma stem cells from radiation-induced cell death in a MELK/FOXM1-dependent manner. *Stem Cell Reports* 2015;**4**(2):226–38.
106. Schonberg DL, *et al*. Preferential iron trafficking characterizes glioblastoma stem-like cells. *Cancer Cell* 2015;**28**(4):441–55.
107. Zhang N, *et al*. FoxM1 promotes beta-catenin nuclear localization and controls Wnt target-gene expression and glioma tumorigenesis. *Cancer Cell* 2011;**20**(4):427–42.
108. Li Y, Zhang S, Huang S. FoxM1: a potential drug target for glioma. *Future Oncol* 2012;**8**(3):223–6.
109. Zhang S, *et al*. M(6)a demethylase ALKBH5 maintains Tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program. *Cancer Cell* 2017;**31**(4):591, e6–606.
110. Wang D, *et al*. DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Res* 2021.
111. Liu J, *et al*. N (6)-methyladenosine of chromosome-associated regulatory RNA regulates chromatin state and transcription. *Science* 2020;**367**(6477):580–6.
112. Merkurjev D, *et al*. Synaptic N(6)-methyladenosine (m(6)a) epitranscriptome reveals functional partitioning of localized transcripts. *Nat Neurosci* 2018;**21**(7):1004–14.
113. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2008;**4**(1):44–57.
114. Wang Y, *et al*. MetaTX: deciphering the distribution of mRNA-related features in the presence of isoform ambiguity, with applications in epitranscriptome analysis. *Bioinformatics* 2020.