Evolutionary Bioinformatics
Volume 15: 1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934319871290

Predict Epitranscriptome Targets and Regulatory Functions of N^6-Methyladenosine (m^{6}A) Writers and Erasers

Yiyou Song[1,*], Qingru Xu[1,*], Zhen Wei[1,2], Di Zhen[1], Jionglong Su[2,3], Kunqi Chen[1,4] and Jia Meng[3,5]

[1]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China. [2]Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China. [3]Research Center for Precision Medicine, Xi'an Jiaotong-Liverpool University, Suzhou, China. [4]Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK. [5]Institute of Integrative Biology, University of Liverpool, Liverpool, UK.

ABSTRACT: Currently, although many successful bioinformatics efforts have been reported in the epitranscriptomics field for N^6-methyladenosine (m^{6}A) site identification, none is focused on the substrate specificity of different m^{6}A-related enzymes, ie, the methyltransferases (writers) and demethylases (erasers). In this work, to untangle the target specificity and the regulatory functions of different RNA m^{6}A writers (METTL3-METT14 and METTL16) and erasers (ALKBH5 and FTO), we extracted 49 genomic features along with the conventional sequence features and used the machine learning approach of random forest to predict their epitranscriptome substrates. Our method achieved reasonable performance on both the writer target prediction (as high as 0.918) and the eraser target prediction (as high as 0.888) in a 5-fold cross-validation, and results of the gene ontology analysis of their preferential targets further revealed the functional relevance of different RNA methylation writers and erasers.

KEYWORDS: N^6-methyladenosine (m^{6}A), target prediction, epitranscriptome, random forest, RNA methylation

RECEIVED: July 19, 2019. **ACCEPTED:** July 31, 2019.

TYPE: Machine Learning for Multi-omics Data Integration - Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been supported by National Natural Science Foundation of China (31671373), Jiangsu University Natural Science Program (16KJB180027), XJTLU Key Program Special Fund (KSF-T-01), and Jiangsu Six Talent Peak Program (XYDXX-118).

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

CORRESPONDING AUTHORS: Kunqi Chen, Department of Biological Sciences, Research Center for Precision Medicine, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China. Email: kunqi.chen@xjtlu.edu.cn

Jia Meng, Department of Biological Sciences, Research Center for Precision Medicine, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.
Email: jia.meng@xjtlu.edu.cn

Introduction

Posttranscriptional RNA modifications are important mechanisms that act on all kinds of RNAs, leading to their increased structural and functional diversity.[1] There are at least 100 kinds of RNA modifications,[2] among which N^6-methyladenosine (m^{6}A) is currently the most prevalent and intensively studied due to its wide impacts.[3] It regulates many essential biological processes including neuronal differentiation, obesity, and messenger RNA (mRNA) stability.[4-6] The m^{6}A RNA methylation is a reversible mark, which is deposited by methyltransferases (or the writers), including METTL3 (methyltransferase-like 3), METTL14 (methyltransferase-like 14), METTL16 (methyltransferase-like 16), and so on, and is removed by demethylases (or the erasers), including FTO (fat mass and obesity–associated protein) and ALKBH5 (ALKB homolog 5).

The writers of RNA m^{6}A modification are protein complexes containing catalytic components METTL3, METTL14, and METTL16, which all have the class I methyltransferase domain. METTL3 is the first identified m^{6}A relevant methyltransferase that has S-adenosylmethionine (SAM)-binding activity.[7] Afterward, METTL14 was discovered as the second methyltransferase that has a methyltransferase domain sharing 22% sequence identity with METTL3.[8] While individual METTL3 or METTL14 exhibits comparably weak catalytic activity in vitro, the METTL3-METTL14 complex has higher catalytic capacity.[9,10] In addition, the crystal structure of METTL3-METTL4 complex suggested that only METTL3 binds with SAM and METTL14 plays a structural role for substrate recognition.[8,11] Thus, the heterodimeric METTL3-METTL14 complex was considered a catalytic domain of m^{6}A methyltransferase. Recently, METTL16 has been identified as another catalytically active m^{6}A mRNA methyltransferase.[12] The METTL16 is similar to METTL3 in structure, but has some unique elements, such as unique αB helix in the Rossmann fold.[13] In addition, these 2 catalytically active m^{6}A mRNA methyltransferases have different roles to play in biological processes. For example, the METTL14 and METTL3 modulate cell cycle progression of cortical neural progenitor cells[14] and depletion of METTL3 or METTL14 promotes tumor progression by enhancing the growth of glioblastoma stem cells.[15] METTL16 can recognize hairpin and methylated adenosine in the U6 snRNA, which regulates the expression of MAT2A.[16]

FTO and ALKBH5 are 2 currently identified m^{6}A-specific RNA demethylases (erasers).[4,17] Although FTO is able to act as a demethylase on another substrate, N^3-methylthymidine

[*] Y.S. and Q.X. contributed equally to this work.

(m^3T), its efficiency is much lower than working on m^6A substrates.[18,19] Although both FTO and ALKBH5 can target specifically RNA m^6A,[19,20] the 2 differ significantly on many levels. For example, at the molecular level, FTO has an amino-terminal AlkB-like domain, a carboxy-terminal domain with a novel fold, and an extra loop that covers on one side of conserved jelly-roll motif.[21] ALKBH5 is a member of the 2-oxoglutarate (2OG) and ferrous iron-dependent nucleic acid oxygenase (NAOX) families, it has a double-stranded β-helix core fold and the active metal site is coordinated by an HXD. . .H motif along with 3 water molecules.[22] In addition, their reaction pathways seem to be different: m^6A is directly converted by ALKBH5 to adenosine; 2 intermediates, *N*^{6}-hydroxymethyladenosine (hm^6A) and *N*^{6}-formyladenosine (fm^6A), are observed during demethylation of m^6A sites by FTO.[23,24] FTO and ALKBH5 also play different roles in terms of physiological functions, one is associated with obesity[4] and the other is thought to participate in the formation of sperm.[25] Moreover, FTO is mainly expressed in the brain,[26] in contrast to ALKBH5, which is found in most tissues, particularly in the testes.[17]

Therefore, according to these studies, 2 kinds of catalytically active m^6A mRNA methyltransferases and 2 demethylases exist that have distinct structures and participate in different biological functions. It would be very interesting to know what the preferential target sites of METTL3-METTL14 complex, METTL16, FTO, and ALKBH5 are and their downstream biological processes. Experimental approaches are effective for testing their functional relevance under a specific experimental condition, such as using different cell lines or testing different treatments. Due to limited detectability, it is not possible to detect target sites on very lowly expressed genes, which is the intrinsic limitation of wet lab–based approach, such as ParCLIP. To unveil comprehensively the epitranscriptome-wide targets of RNA m^6A, we considered using computational approaches.

Currently, the field of bioinformatics has seen the rapid development of new methods and their wide applications in RNA epigenetics. The mammalian m^6A short consensus motif RRACH (where R = A or G; H = A, C, or U) has not been characterized until 2012, when the next-generation sequencing techniques called m^6A-seq or MeRIP-seq (methylated RNA immunoprecipitation sequencing)[27-29] emerged. Thereupon, RMBase and MeTDB have been developed into v2.0, which now can provide millions of m^6A sites in many different species, such as, human, mouse, yeast, and fly.[30,31]

Meanwhile, many successful computational studies have been devised on m^6A site prediction, such as SRAMP, MethyRNA, and RNAMethPre.[32-34] Although there are many good precedents in m^6A site prediction and deposition (database development), there has been no effort made in the substrate prediction of m^6A enzymes. We, therefore, devised a computational tool to study the target specificity of m^6A enzymes. In this study, the predictors were built using the random forest (RF) approach to distinguish the target specificity of the m^6A writers (METTL3-METTL14 complex and METTL16) and the erasers (FTO and ALKBH5), respectively. Although the sequence-derived features were widely used in m^6A site prediction[33,35] and generated reasonably good results, we included additional genome-derived features and achieved substantial improvement in performance.

Matrerials and Methods

The m^6A sites

The transcriptome-wide m^6A sites were extracted from the WHISTLE web server,[36] which used multiple genomic and sequence features to predict the entire epitranscriptome and achieved substantial improvement compared with existing approaches. Please note that all these m^6A sites were originally collected from wet lab experiments[30] and simultaneously supported by the WHISTLE prediction with high confidence. We considered the m^6A sites with probability greater than .6, .7, .8, and .9, which are corresponding to 4 data sets of 98 095, 75 720, 52 687, and 27 646 RNA methylation sites, respectively. In this study, 4 different sets of data were extracted for further analysis. This is because they correspond to different coverage and reliability. A larger set has better coverage of the m^6A epitranscriptome, but may also contain more false m^6A sites that can affect prediction performance. The training data is provided in Supplementary Table 1.

Target sites of the enzymes

The ground truth targets were identified using perturbation experiment, eg, the hypomethylated sites after the knock down of a methyltransferase identified from MeRIP-seq data. Specifically, the raw data were retrieved from GEO (Gene Expression Omnibus; see Table 1), and the FASTQ files were aligned to the reference genome hg19 using hisat2[42] with default settings. The resulting SAM files were then converted to BAM files using samtools with the quality filter –q 30 and the FLAG filter –F 2820. Following that, the number of reads aligned to each individual RNA methylation sites were counted as fragments in R using GenomicAlignment package.[43] For each experiment with regulator perturbation, differential methylation analysis was conducted by DESeq2[44] using the interactive generalized linear model (GLM) design of ~ IP*Treatment, while IP is the indicator vector for the samples being IP, and Treatment is the indicator vector for the samples treated with regulator perturbation. The m^6A sites with the Wald test fdr < 0.05 and the interactive coefficient <0 (>0 for the sample gsc11-ALKBH5-) are treated as the target sites of the regulator. The shared target sites of multiple enzymes, ie, (FTO and ALKBH5) or (M3/M14 vs M16) are considered with ambiguous association and thus excluded from our analysis.

Table 1. GEO data sets used to identify ground truth target sites.

ID	REGULATOR	CELL TYPE	GEO SRA STUDY	PUBLICATION
1	METTL14	A549	SRP039397	Schwartz et al[37]
2	METTL14	Hela	SRP022152	Liu et al[10]
3	METTL14	MonoMac6	SRP103072	Weng et al[38]
4	METTL14	NB4	SRP103072	Weng et al[38]
5	METTL3	A549	SRP039397	Schwartz et al[37]
6	METTL3	AML	SRP099081	Barbieri et al[39]
7	METTL3	Hek293T	SRP039397	Schwartz et al[37]
8	METTL3	Hela	SRP022152	Liu et al[10]
9	METTL16	HEK293A	SRP094637	Pendleton et al[12]
10	ALKBH5	gsc11	SRP067910	Zhang et al[40]
11	FTO	AML	SRP067910	Li et al[41]

Abbreviations: ALKBH5, ALKB homolog 5; FTO, fat mass and obesity–associated protein; GEO, Gene Expression Omnibus; METTL3, methyltransferase-like 3; METTL14, methyltransferase-like 14; METTL16, methyltransferase-like 16; SRA, Sequence Read Archive.

Feature encoding scheme and selection

Sequence-derived features. The nucleotide encoding method according to chemical properties was suggested by Bari et al.[45] In the MethyRNA[32] and M6Apred,[46] this encoding method was applied in the generation of sequence-derived features and achieved good accuracy in the m^{6}A site prediction. In this project, we followed this idea of chemical encoding method to generate sequence-derived features. Specifically, 3 chemical properties of the nucleotides were used to classify adenine (A), cytosine (C), guanine (G), and uracil (U). The first property is ring structures: A and G have 2 ring structures, whereas C and U have only 1 ring. The second property is functional groups. A and C contain amino group, whereas G and U contain the keto group. The third property is the number of hydrogen bonds formed. A and U can form 2 hydrogen bonds during hybridization, whereas G and C can form 3 hydrogen bonds. Based on the 3 structural chemical properties defined above, the *i*th nucleotide from sequence can be encoded by a vector:

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

Thus, A can be marked as (1, 1, 1), C can be marked as (0, 1, 0), G can be marked as (1, 0, 0), and U can be marked as (0, 0, 1). In addition, a feature of the accumulative nucleotide frequency is calculated for each nucleotide position in the sequence. The density of *i*th nucleotide is defined as the sum of all the instances of the *i*th nucleotide before the $i+1$ position. The nucleotide frequency (keep 2 decimal places) is defined by the following formula: $f_i = d_i / i$. Using sequence "AUGGACACU" as an example, the accumulative frequencies for adenine are 1.00 (1/1), 0.40 (2/5), and 0.43 (3/7) at the first, fifth, and seventh sequence positions, respectively, whereas the frequencies for uracil are 0.50 (1/2) and 0.11 (1/9) at the second and ninth sequence positions, respectively. According to the sequence extended 20 bp (base pair) to each side around the m^{6}A sites, they were encoded by the above method, and we obtained a sequence-derived feature with 164 dimensional features for each m^{6}A site.

Genome-derived features

Although the sequence-based features were widely used in the prediction of RNA modification sites, there are potentially other features that can be used.[47] The genomic features have been shown in the WHISTLE project to be effective in the m^{6}A site prediction. A total of 47 genome-derived features were considered for this project (see Table 2). Specifically, genomic features 1 to 17 specify the locations of adenosine sites within the transcript region and their topological properties as dummy variables. To generate features in this category, we used the transcript annotations of hg19 human genome assembly and the GenomicFeatures R/Bioconductor package.[43] Genomic features 18 to 21 define the relative positions of adenosine sites within transcript region, which is calculated based on the distance from the methylated adenine to the 5′ end divided by the total width of the region; the position features are set to 0 if the adenosine sites do not belong to the region. The values of features 22 to 26 are lengths of the transcript region containing the methylated site; if the

Table 2. Genomic features used in the analysis.

ID	NAME	DESCRIPTION	NOTE
1	UTR5	5′ UTR	Dummy variables indicating whether the site overlaps the topological region on the major RNA transcript
2	UTR3	3′ UTR	
3	cds	CDS	
4	Stop_codons	Stop codons flanked by 100 bp	
5	Start_codons	Start codons flanked by 100 bp	
6	TSS	Downstream 100 bp of TSS	
7	TSS_A	Downstream 100 bp of TSS on A	
8	Stop_codons	Stop codons	
9	exon_stop	Exons containing stop codons	
10	alternative_exon	Alternative exons	
11	constitutive_exon	Constitutive exons	
12	internal_exon	Internal exons	
13	long_exon	Long exons (exon length ⩾ 400 bp)	
14	last_exon	Last exons[48]	
15	last_exon_400bp	5′ 400 bp of the last exons[48]	
16	last_exon_sc400	5′ 400 bp of the last exons containing stop codons[48]	
17	intron	Introns	
18	pos_UTR5	Relative position on 5′ UTR	Relative position on the region
19	pos_UTR3	Relative position on 3′ UTR	
20	pos_CDS	Relative position on CDS	
21	pos_exons	Relative position on exon	
22	length_UTR5	5′ UTR length	The region length in base pairs
23	length_UTR3	3′ UTR length	
24	length_cds	CDS length	
25	length_gene_ex	Mature transcript length	
26	length_gene_full	Full transcript length	
27	PC_1bp	PhastCons scores of the nucleotide[49]	Scores related to evolutionary conservation
28	PC_101bp	Average phastCons scores within the flanking 50 bp region[49]	
29	FC_1bp	fitCons scores of the nucleotide[50]	
30	FC_101bp	Average fitCons scores within the flanking 50 bp region[50]	
31	struc_hybridize	Predicted RNA hybridized region[51]	RNA secondary structure
32	struc_loop	Predicted RNA loop region[51]	
33	isoform_num	Isoform number	Attributes of the genes or transcripts
34	exon_num	Exon number	
35	HK_genes	Housekeeping genes[52]	

(Continued)

Table 2. (Continued)

ID	NAME	DESCRIPTION	NOTE
36	sncRNA	sncRNA	RNA annotations related to m6A biology
37	lncRNA	lncRNA	
38	miR_targeted_genes	miRNA-targeted genes[53]	
39	Verified_miRtargets	miRNA-targeted sites verified by experiment[8]	
40	TargetScan	Predicted miRNA targeted sites by TargetScan[9]	
41	HNRNPC_eCLIP	eCLIP data of HNRNPC RNA binding sites[7]	RNA-binding protein annotation from MeTDB database[31]
42	METTL3_TREW	METTL3-binding region[31]	
43	METTL14_TREW	METTL14-binding region[31]	
44	WTAP_TREW	WTAP-binding region[31]	
45	YTHDC1_TREW	YTHDC1-binding region[31]	
46	YTHDF1_TREW	YTHDF1-binding region[31]	
47	YTHDF2_TREW	YTHDF2-binding region[31]	
48	ALKBH5_TREW	ALKBH5-binding region[31]	
49	FTO_TREW	FTO-binding region[31]	

Abbreviations: ALKBH5, ALKB homolog 5; FTO, fat mass and obesity–associated protein; GEO, Gene Expression Omnibus; METTL3, methyltransferase-like 3; METTL14, methyltransferase-like 14; METTL16, methyltransferase-like 16; SRA, Sequence Read Archive.
Features that are directly related to the prediction are not used to avoid overfitting. For example, the features 42 and 43 were not used for writer target prediction, whereas feature 48 and 49 were not used for eraser target prediction.

sites do not belong to the region, the features are set as 0 also. The evolutionary conservation score of the methylated adenosine sites and its flanking regions was measured in features 27 to 30 with 2 metrics of nucleotide conservation: PhastCons score and the fitness consequence scores. Features 31 and 32 represent the RNA secondary structures of transcripts region containing methylated adenine predicted using RNAfold in Vienna RNA package. Features 33 to 35 represent the attributes of the genes or transcripts containing methylated sites. Features 36 to 40 indicate whether the adenosine sites interact with small noncoding RNA, long noncoding RNA, and microRNA, respectively. Finally, features 41 to 49 indicate whether the methylated sites are located within RNA protein-binding regions. For the above features, to avoid ambiguity caused by transcript isoforms, only the primary (longest) transcript of each gene was kept for the extraction of the transcript subregions. The details for genomic features are summarized in Table 2.

Machine learning approach

Machine learning algorithms have been widely used in the field of computational biology. In RNA epigenetics studies, support vector machine (SVM) and RF have been used previously in RNA m^{6}A site prediction,[32,33,46] and both achieved good performances. In this project, the RF algorithm from the R randomForest was used to build predictor models.

Performance evaluation

A 5-fold cross-validation was used for assessing the reliability of the method. In the performance evaluation, the sensitivity (Sn) and specificity (Sp) are defined as follows:

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

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

where *TP, TN, FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively. In addition, prediction performance under different decision thresholds were measured as the receiver operating characteristic (ROC) curve whose *y*-axis is sensitivity and *x*-axis is 1-specificity, the area under ROC curve (AUC) was calculated as the main performance evaluation metrics. In addition, the ACC (overall accuracy) and MCC (Matthews correlation coefficient) were calculated as other indicators to evaluate the reliable of model.

Results and Discussion

Feature selection

Although extensive research in m^{6}A site prediction has demonstrated the effectiveness and reliability of sequence-derived features[32,33,54] and genome-derived features,[36] we seek to, for

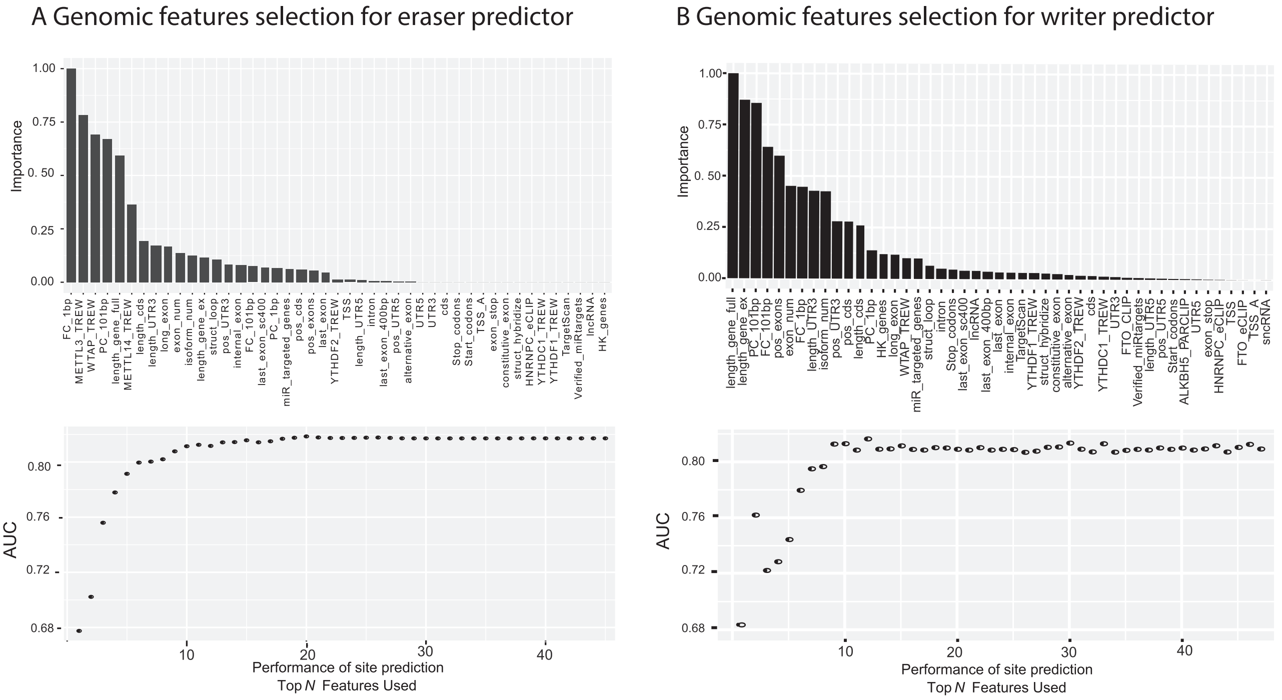

Figure 1. Feature Selection for Predictors. (A) The top 20 genomic features were used for prediction of the targets of erasers, including conservation score, METTL3 targets, etc. (B) The top 15 genomic features were used for prediction of the targets of writers, with the distance to known m^{6}A site as the most important predictive feature, followed by gene length and conservation score.

the first time, use these features to predict the target specificity of m^{6}A enzymes. Due to the abundance of the genomic features, we first performed feature selection to identify the genomic features most relevant to our purpose, which is to improve the reliability of the features and the prediction performance, as well as to save computation time and memory. The feature selection was performed on the set of RNA methylation sites with probability greater than .6.

At the beginning, the Perturb method[55] was used to estimate the relative importance of each genomic feature in the target specificity prediction of eraser targets using the R caret package. To illustrate the relative importance of different features clearly, the measurement results are rescaled and ranked (Figure 1A). According to this rank, relevant AUROC (area under the receiver operating characteristics) figures are generated based on the top N features. We can observe that the best performance was achieved with the top 20 features. Thus, only the top 20 features were used in our prediction model for erasers. Similarly, the same treatment was done on the writer target prediction (Figure 1B), where the best performance was achieved with the top 15 features.

Predictors based on different set of features

Existing computation models overwhelmingly relied on the sequence features. In our prediction model, while it also incorporates features derived from other genomic annotations (see Table 2), It is important to test whether these features contribute to the prediction performance. For this purpose, a 5-fold cross-validation was conducted on the data set of RNA methylation sites with probability greater than .6, and different types of features were used. As shown in Table 3, sequence features were more effective than sequence features in the target prediction for erasers; but not in the case for writers. However, it is consistent for the best performance to be achieved when both sequence and genomic features were incorporated.

Performance on different data sets

In next step, we consider expanding the model to test on all the 4 data sets. As shown in Table 4, the 4 different data sets have different coverage of the m^{6}A epitranscriptome; for erasers, the best target prediction performance was achieved on data set 4, which are RNA methylation sites with probability greater than .9, whereas for writers, the best performance was achieved on data set 3, which are corresponding to the RNA methylation sites with probability greater than .8. To compare with other machine learning approaches, the SVM, Naïve Bayes, decision tree, and GLM were applied to build model. The performances for each method are summarized in Table S2 and were evaluated by the sensitivity, specificity, AUROC, ACC, and MCC.

Biological functions regulated by different enzymes

There are 3585, 4623, 4742, and 4734 sites identified in data set 1 (see Table 4) under the regulation of METTL3-METTL14 complex, METTL16, FTO, and ALKBH5, respectively, which are located on 2149, 2178, 2375, and 2635

Table 3. Performance of predictors based on different features.

FEATURE TYPE	ERASERS (FTO VS ALKBH5)			WRITERS (M3/M14 VS M16)		
	SENSITIVITY	SPECIFICITY	AUROC	SENSITIVITY	SPECIFICITY	AUROC
Sequence	0.789	0.781	0.849	0.656	0.746	0.772
Genome	0.762	0.736	0.827	0.802	0.795	0.886
Both	0.814	0.813	0.887	0.802	0.795	0.889

Abbreviations: ALKBH5, ALKB homolog 5; AUROC, area under the receiver operating characteristics; FTO, fat mass and obesity–associated protein.
This result was achieved on RNA methylation sites with probability greater than .6 with a 5-fold cross-validation.

Table 4. Prediction performance on different data sets (AUROC).

ENZYME TYPE	DATA SET			
	DATA SET 1 ($P^* > .6$) 98 095 SITES	DATA SET 2 ($P > .7$) 75 720 SITES	DATA SET 3 ($P > .8$) 52 687 SITES	DATA SET 4 ($P > .9$) 27 646 SITES
Erasers (FTO vs ALKBH5)	0.873	0.873	0.872	0.888
Writers (M3/M14 vs M16)	0.889	0.888	0.911	0.877

Abbreviations: ALKBH5, ALKB homolog 5; AUROC, area under the receiver operating characteristics; FTO, fat mass and obesity–associated protein.

Four data sets were considered, corresponding to the experiment-validated RNA methylation sites from RMBase and also supported by WHISTLE prediction with probability greater than .6, .7, .8, and .9, respectively. The detailed performance of 5 different classification predictors (RF, SVM, GLM, Naïve Bayes, and decision tree) is presented in Supplementary Table S2.

genes. The biological functions of these targets sites were then annotated with gene ontology enrichment analysis using DAVID website.[56] Figure 2 shows the top 10 mostly enriched biological processes. We can see that different biological processes were enriched in different enzymes. For example, FTO is associated with cell-cell adhesion (5.473E−13), mRNA splicing (2.212E−12), viral process (4.31E−09), whereas the target sites of ALKBH5 are more related to Golgi organization (4.13E−09) and DNA-templated transcription (4.14E−14). METTL16 targets are enriched with genes related to endoplasmic reticulum–associated misfolded protein catabolic process (9.07E−06), regulation of cell cycle (1.15E−04), apoptotic process (6.70E−04) and protein ubiquitination (9.99E−05), whereas METTL3-METTL14 complex preferentially target to genes associated to cell-cell adhesion (2.86E−07), cell division (4.08E−07), and G2/M transition of mitotic cell cycle (3.03E−06). Please see Table S3 for the complete gene ontology enrichment analysis result.

Discussion and Conclusions

Recent progress in RNA modification bioinformatics enabled the precise detection, accurate quantification, differential analysis, and function annotation of m^6A RNA methylation sites in base resolution. RMBase and MetDB collected experimentally validated m^6A sites in multiple species and revealed their potential regulatory functions.[30,31] The exomePeak[57,58] was developed based on Przyborowski and Wilenski's[59,60] method for m^6A site detection and differential methylation analysis from MeRIP-seq data. The computational prediction of m^6A modification sites in different species performed in the works iRNA(m6A)-PseDNC,[61] iRNA-Methyl,[62] m6Apred,[46] RFAthM6A,[63] and BERMP[64] based on machine learning or deep learning approaches. The potential disease relevance and single-nucleotide polymorphism association of m^6A modification were revealed by the m6Avar[65] and m6ASNP[66] by examining whether a disease mutation can alter the potential of RNA methylation status. Meanwhile, complex network method was used in m6Acomet,[67] m6A-Driver,[68] Deepm6A,[69] DRUM,[70] and FunDMDeep-m6A[71] to study the regulatory functions and predict the disease association of m^6A RNA modification.

Here, we have proposed a computational approach for the prediction of the target sites of m^6A enzymes. The computational model proposed relies on 49 genomic features as well as the conventional sequence features. With a model selection step, we showed with a 5-fold cross-validation that the proposed approach achieved relatively good performance in the target prediction for the writers (AUC: 0.918) and erasers (AUC: 0.888). The following gene ontology analysis unveiled the epitranscriptome functional relevance of these enzymes.

The proposed approach suffers from the following limitations. (1) The ground truth target sites were identified from perturbation experiment, in which a target site of a methyltransferase is defined as those whose methylation level decreases when the methyltransferase was knocked down. Obviously, the decrease in methylation level may not be due to direct target but because of a secondary effect. For this

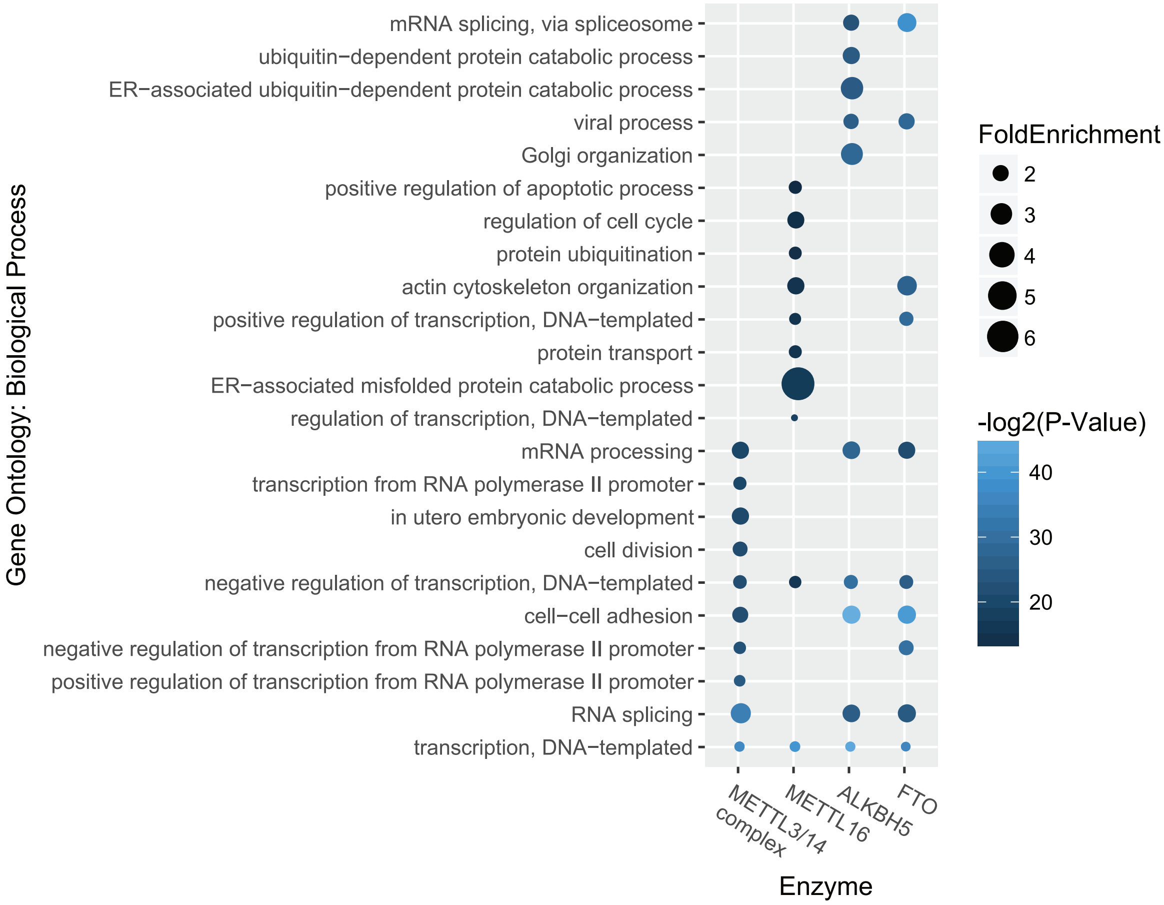

Figure 2. Biological processes enriched in targets of m^{6}A enzymes. Distinct biological processes are enriched in the predicted target sites of different enzymes. Figure shows the top 10 most statistically enriched biological processes associated with the targets of different m^{6}A enzymes.

reason, the ground truth data can be further improved. (2) The features incorporated in the prediction model can be further increased. Although a total of 49 genomic features have been incorporated in our prediction model, the set can be expanded by including, eg, features related to lncRNA, repeat region. Increased feature set can often lead to improved performance. (3) We considered here only a binary classification, which emphasizes the target specificity of different enzymes. However, in practice, it is possible that there are a large number of RNA methylation sites that are simultaneously targeted by both m^{6}A writers (or both m^{6}A erasers) considered in this work. In addition, there are likely to be unknown methyltransferases or demethylases to be discovered and thus are not considered in the prediction models. This would be a difficult question to solve. (4) A better computation method may be used. We used here RF, which is a classic method. Recent development in artificial intelligence, especially deep learning–related approach may achieve better performance.

Author Contributions

JM and KC conceived the idea and designed the research; ZW processed the raw data; YS, QX, and DZ performed the prediction analysis; YS and QX drafted the manuscript first. All authors read, critically revised, and approved the final manuscript.

ORCID iD

Kunqi Chen https://orcid.org/0000-0002-6025-8957

Supplemental Material

Supplemental material for this article is available online.

References

1. Grosjean H. *Fine-Tuning of RNA Functions by Modification and Editing*. Berlin, Germany: Springer; 2005.
2. Boccaletto P, Machnicka MA, Purta E, et al. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res*. 2017;46:D303-D307.
3. Meyer KD, Jaffrey SR. The dynamic epitranscriptome: N^6-methyladenosine and gene expression control. *Nat Rev Mol Cell Biol*. 2014;15:313-326.
4. Jia G, Fu Y, Zhao X, et al. N^6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol*. 2011;7:885-887.
5. Mukobata S, Hibino T, Sugiyama A, et al. m^{6}A acts as a nerve growth factor-gated Ca^{2+} channel in neuronal differentiation. *Biochem Biophys Res Commun*. 2002;297:722-728.
6. Wang X, Lu Z, Gomez A, et al. N^6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*. 2014;505:117-120.
7. Bokar JA, Rath-Shambaugh ME, Ludwiczak R, Narayan P, Rottman F. Characterization and partial purification of mRNA N^6-adenosine methyltransferase from HeLa cell nuclei. Internal mRNA methylation requires a multisubunit complex. *J Biol Chem*. 1994;269:17697-17704.
8. Wang X, Feng J, Xue Y, et al. Structural basis of N^6-adenosine methylation by the METTL3-METTL14 complex. *Nature*. 2016;534:575-578.
9. Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. N^6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol*. 2014;16:191-198.

10. Liu J, Yue Y, Han D, et al. A METTL3-METTL14 complex mediates mammalian nuclear RNA N^6-adenosine methylation. *Nat Chem Biol*. 2014;10:93-95.
11. Sledz P, Jinek M. Structural insights into the molecular mechanism of the m^{6}A writer complex. *Elife*. 2016;5:e18434.
12. Pendleton KE, Chen B, Liu K, et al. The U6 snRNA m^{6}A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell*. 2017;169:824-835.e14.
13. Ruszkowska A, Ruszkowski M, Dauter Z, Brown JA. Structural insights into the RNA methyltransferase domain of METTL16. *Sci Rep*. 2018;8:5311.
14. Yoon KJ, Ringeling FR, Vissers C, et al. Temporal control of mammalian cortical neurogenesis by m^{6}A methylation. *Cell*. 2017;171:877-889.e17.
15. Cui Q, Shi H, Ye P, et al. m^{6}A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep*. 2017;18:2622-2634.
16. Shima H, Matsumoto M, Ishigami Y, et al. S-adenosylmethionine synthesis is regulated by selective N^6-adenosine methylation and mRNA degradation involving METTL16 and YTHDC1. *Cell Rep*. 2017;21:3354-3363.
17. Zheng G, Dahl JA, Niu Y, et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell*. 2013;49:18-29.
18. Jia G, Yang CG, Yang S, et al. Oxidative demethylation of 3-methylthymine and 3-methyluracil in single-stranded DNA and RNA by mouse and human FTO. *FEBS Lett*. 2008;582:3313-3319.
19. Toh JDW, Sun L, Lau LZM, et al. A strategy based on nucleotide specificity leads to a subfamily-selective and cell-active inhibitor of N^6-methyladenosine demethylase FTO. *Chem Sci*. 2015;6:112-122.
20. Yang T, Cheong A, Mai X, Zou S, Woon EC. A methylation-switchable conformational probe for the sensitive and selective detection of RNA demethylase activity. *Chem Commun (Camb)*. 2016;52:6181-6184.
21. Han Z, Matsumoto M, Ishigami Y, et al. Crystal structure of the FTO protein reveals basis for its substrate specificity. *Nature*. 2012;464:1205-1209.
22. Aik W, Scotti JS, Choi H, et al. Structure of human RNA N^6-methyladenine demethylase ALKBH5 provides insights into its mechanisms of nucleic acid recognition and demethylation. *Nucleic Acids Res*. 2014;42:4741-4754.
23. Chen W, Zhang L, Zheng G, et al. Crystal structure of the RNA demethylase ALKBH5 from zebrafish. *FEBS Lett*. 2014;588:892-898.
24. Fu Y, Jia G, Pang X, et al. FTO-mediated formation of N^6-hydroxymethyladenosine and N^6-formyladenosine in mammalian RNA. *Nat Commun*. 2013;4:1798.
25. Tang C, Klukovich R, Peng H, et al. ALKBH5-dependent m^{6}A demethylation controls splicing and stability of long 3′-UTR mRNAs in male germ cells. *Proc Natl Acad Sci U S A*. 2018;115:E325-E333.
26. Gerken T, Girard CA, Tung YC, et al. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science*. 2007;318:1469-1472.
27. Chen K, Lu Z, Wang X, et al. High-resolution N6-methyladenosine (m^{6}A) map using photo-crosslinking-assisted m^{6}A sequencing. *Angew Chem*. 2015;127: 1607-1610.
28. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, et al. Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature*. 2012;485:201-206.
29. 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;149:1635-1646.
30. Xuan J-J, Sun WJ, Lin PH, et al. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res*. 2017;46:D327-D334.
31. Liu H, Wang H, Wei Z, et al. MeT-DB V2.0: elucidating context-specific functions of N^6-methyladenosine methyltranscriptome. *Nucleic Acids Res*. 2018;46:D281-D287.
32. Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N^6-methyladenosine sites. *J Biomol Struct Dyn*. 2017;35:683-687.
33. Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. SRAMP: prediction of mammalian N^6-methyladenosine (m^{6}A) sites based on sequence-derived features. *Nucleic Acids Res*. 2016;44:e91.
34. Xiang S, Liu K, Yan Z, Zhang Y, Sun Z. RNAMethPre: a web server for the prediction and query of mRNA m^{6}A sites. *PLoS One*. 2016;11:e0162707.
35. Chen W, Xing P, Zou Q. Detecting N^6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci Rep*. 2017;7:40242.
36. Chen K, Wei Z, Zhang Q, et al. WHISTLE: a high-accuracy map of the human N^6-methyladenosine (m^{6}A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res*. 2019;47:e41.
37. Schwartz S, Mumbach MR, Jovanovic M, et al. Perturbation of m^{6}A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites. *Cell Rep*. 2014;8:284-296.
38. Weng H, Huang H, Wu H, et al. METTL14 inhibits hematopoietic stem/progenitor differentiation and promotes leukemogenesis via mRNA m^{6}A modification. *Cell Stem Cell*. 2018;22:191-205.e9.
39. Barbieri I, Tzelepis K, Pandolfini L, et al. Promoter-bound METTL3 maintains myeloid leukaemia by m^{6}A-dependent translation control. *Nature*. 2017;552: 126-131.
40. Zhang S, Zhao BS, Zhou A, 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;31:591-606.e6.
41. Li Z, Weng H, Su R, et al. FTO plays an oncogenic role in acute myeloid leukemia as a N^6-methyladenosine RNA demethylase. *Cancer Cell*. 2017;31:127-141.
42. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12:357-360.
43. Lawrence M, Huber W, Pages H, et al. Software for computing and annotating genomic ranges. *Plos Comput Biol*. 2013;9:e1003118.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
45. Bari ATMG, Reaz MR, Choi HJ, Jeong BS. *DNA Encoding for Splice Site Prediction in Large DNA Sequence* Berlin, Gemany: Springer; 2013:46-58.
46. Chen W, Tran H, Liang Z, Lin H, Zhang L. Identification and analysis of the N^6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep*. 2015;5:13859.
47. Xu T, Zheng X, Li B, Jin P, Qin Z, Wu H. A comprehensive review of computational prediction of genome-wide features [published online ahead of print November 16, 2018]. *Brief Bioinform*. doi:10.1093/bib/bby110.
48. Ke S, Pandya-Jones A, Saito Y, 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.
49. Siepel A, Haussler D. Phylogenetic hidden Markov models. In: Nielsen R ed. *Statistical Methods in Molecular Evolution*. New York, NY: Springer; 2005:325-351.
50. Gulko B, Gronau I, Hubisz MJ, Siepel A. Probabilities of fitness consequences for point mutations across the human genome. bioRxiv 006825, 2014.
51. Lorenz R, Bernhart SH, Honer Zu, Siederdissen C, et al. ViennaRNA package 2.0. *Algorithms Mol Biol*. 2011;6:26.
52. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet*. 2013;29:569-574.
53. Chou C-H, Shrestha S, Yang CD, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res*. 2017;46:D296-D302.
54. Li Y-H, Zhang G, Cui Q. PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics*. 2015;31:3362-3364.
55. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Model*. 2003;160:249-264.
56. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2008;4:44-57.
57. Meng J, Cui X, Rao MK, Chen Y, Huang Y. Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics*. 2013;29:1565-1567.
58. Meng J, Lu Z, Liu H, et al. A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods*. 2014;69:274-281.
59. Przyborowski J, Wilenski H. Homogeneity of results in testing samples from Poisson series: with an application to testing clover seed for dodder. *Biometrika*. 1940;31:313-323.
60. Krishnamoorthy K, Thomson J. A more powerful test for comparing two Poisson means. *J Stat Plan Infer*. 2004;119:23-35.
61. Chen W, Ding H, Zhou X, Lin H, Chou KC. iRNA(m^{6}A)-PseDNC: identifying N(6)-methyladenosine sites using pseudo dinucleotide composition. *Anal Biochem*. 2018;561-562:59-65.
62. 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.
63. Wang X, Yan R. RFAthm^{6}A: a new tool for predicting m^{6}A sites in Arabidopsis thaliana. *Plant Mol Biol*. 2018;96:327-337.
64. 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;14:1669-1677.
65. Zheng Y, Nie P, Peng D, et al. m^{6}AVar: a database of functional variants involved in m^{6}A modification. *Nucleic Acids Res*. 2018;46:D139-D145.
66. Jiang S, Xie Y, He Z, et al. m^{6}ASNP: a tool for annotating genetic variants by m^{6}A function. *Gigascience*. 2018;7:1-11.
67. Wu X, Wei Z, Chen K, et al. m^{6}Acomet: large-scale functional prediction of individual m^{6}A RNA methylation sites from an RNA co-methylation network. *BMC Bioinformatics*. 2019;20:223.
68. Zhang S, Zhang S, Liu L, Meng J, Huang Y. m^{6}A-Driver: identifying context-specific mRNA m^{6}A methylation-driven gene interaction networks. *PLoS Comput Biol*. 2016;12:e1005287.
69. Zhang S-Y, Zhang SW, Fan XN, et al. Global analysis of N^6-methyladenosine functions and its disease association using deep learning and network-based methods. *PLoS Comput Biol*. 2019;15:e1006663.
70. Tang Y, Chen K, Wu X, et al. DRUM: inference of disease-associated m^{6}A RNA methylation sites from a multi-layer heterogeneous network. *Front Genet*. 2019;10:266.
71. Zhang S-Y, Zhang S-W, Fan X-N, Zhang T, Meng J, Huang Y. FunDMDeep-m^{6}A: identification and prioritization of functional differential m^{6}A methylation genes. *Bioinformatics*. 2019;35:i90-i98.