Mol Genet Genomics (2016) 291:2225–2229
DOI 10.1007/s00438-016-1243-7

METHODS PAPER

# Identifying $N^6$-methyladenosine sites in the *Arabidopsis thaliana* transcriptome

**Wei Chen[1] · Pengmian Feng[2] · Hui Ding[3] · Hao Lin[3]**

Received: 24 July 2016 / Accepted: 27 August 2016 / Published online: 2 September 2016


**Abstract** $N^6$-Methyladenosine ($m^6A$) plays important roles in many biological processes. The knowledge of the distribution of $m^6A$ is helpful for understanding its regulatory roles. Although the experimental methods have been proposed to detect $m^6A$, the resolutions of these methods are still unsatisfying especially for *Arabidopsis thaliana*. Benefitting from the experimental data, in the current work, a support vector machine-based method was proposed to identify $m^6A$ sites in *A. thaliana* transcriptome. The proposed method was validated on a benchmark dataset using jackknife test and was also validated by identifying strain-specific $m^6A$ sites in *A. thaliana*. The obtained predictive results indicate that the proposed method is quite promising. For the convenience of experimental biologists, an online webserver for the proposed method was built, which is freely available at http://lin.uestc.edu.cn/server/M6ATH. These results indicate that the proposed method holds a potential to become an elegant tool in identifying $m^6A$ site in *A. thaliana*.

**Keywords** $m^6A$ · Ring structure · Hydrogen bond · Chemical functionality · Support vector machine

Communicated by S. Hohmann.

✉ Wei Chen
chenweiimu@gmail.com

✉ Hao Lin
hlin@uestc.edu.cn

[1] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China

[2] School of Public Health, North China University of Science and Technology, Tangshan 063000, China

[3] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics and Center for Information in Biomedicine, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

## Introduction

Among the ~150 post-transcriptional modifications of RNA, $N^6$-methyladenosine ($m^6A$) is the most prevalent one and has been discovered from bacteria to *Homo sapiens* (Cantara et al. 2011). Recent studies have demonstrated that $m^6A$ is a dynamic and reversible modification (Jia et al. 2011; Liu et al. 2014). $m^6A$ can be installed and erased by $m^6A$ methyltransferases and demethylases (Jia et al. 2011; Liu et al. 2014), respectively. It has been found that $m^6A$ impacts a variety of biological events, such as mRNA splicing and stability (Nilsen 2014), RNA localization and degradation (Meyer and Jaffrey 2014), stem cell pluripotency (Chen et al. 2015a), and cell differentiation and reprogramming (Geula et al. 2015). Therefore, the detection of $m^6A$ is helpful for the revealing its biological functions.

Based on high-throughput experiments, $m^6A$ profiles are available for *Saccharomyces cerevisiae* (Schwartz et al. 2013), *H. sapiens* (Dominissini et al. 2012; Linder et al. 2015), *Mus musculus* (Dominissini et al. 2012), and *Arabidopsis thaliana* (Luo et al. 2014). Recently, Jaffrey et al. provided the single nucleotide resolution profile of the $m^6A$ sites for human using the miCLIP technique (Linder et al. 2015). However, since the high-throughput experimental identifications of $m^6A$ sites rely on next-generation sequencing-based techniques, they are still unable to exactly point out which adenosine is methylated in most

species. Therefore, accurate and base-resolution methods are highly desirable to determine the exact $m^6A$ sites.

The experimental methods yield quite encouraging results and provide unprecedented opportunities to construct computational $m^6A$ site predictors which are excellent complements to experimental techniques. In the last 2 years, a series of computational tools have been developed for *S. cerevisiae* (Chen et al. 2015b, c), *H. sapiens* (Chen et al. 2016b; Zhou et al. 2016), and *M. musculus* (Chen et al. 2016b; Zhou et al. 2016), respectively. However, as far as we know, there is no computational tool available for identifying $m^6A$ sites in plant. Keeping this in mind, in the present study, a support vector machine-based method was proposed to identify $m^6A$ sites in *A. thaliana*.

## Materials and methods

### Dataset construction

Using $m^6A$-targeted antibody coupled with high-throughput sequencing, Luo and his colleagues obtained 7489 $m^6A$ peaks in Can-0 strain and 6094 $m^6A$ peaks in Hen-16 strain of *A. thaliana* (Luo et al. 2014). Among them, 4317 $m^6A$ peaks were detected in both Can-0 and Hen-16 strains and were called common $m^6A$ peaks. Since most of the $m^6A$ peaks contain the motif RRACH (where R stands for purine, A stands for $m^6A$, and H stands for a non-guanine base) (Luo et al. 2014), we collected segments that have the RRACH at the center from the 4317 common $m^6A$ peak containing sequences.

To reduce the homology bias, sequences with more than 60 % sequence similarity were removed using the CD-HIT program (Fu et al. 2012). Thus, we obtained 394 $m^6A$ site containing sequences and selected as positive samples. Preliminary trials indicated that when the length of the segments is 25 nt with the $m^6A$ in the center, the highest predictive results could be obtained. Accordingly, the positive samples are all with the length of 25 nucleotides. The negative samples were collected by choosing the 25-nt long sequences satisfying the rule that the adenosine in the center was not experimentally confirmed as $m^6A$. It is easy to notice that the number of negative samples is dramatically larger than that of positive ones. To deal with the unbalanced numbers between positive and negative samples in model training, 394 sequences were randomly picked out to form the negative samples.

### Representation of RNA sequences

Nucleotide chemical property and nucleotide composition have been successfully used to identify post-transcriptional RNA modifications (Chen et al. 2016a, c). Thus, they were used to encode RNA sequences in the present work. Below is the brief elaboration on how to encode RNA sequences using nucleotide chemical property and nucleotide composition.

RNA is made up of adenine (A), guanine (G), cytosine (C), and uracil (U). These bases have different chemical properties. In terms of ring structures, A and G have two rings, while C and U are pyrimidines that have one ring. When forming secondary structures, C and G form strong hydrogen bonds, whereas A and U form weak hydrogen bonds. In terms of chemical functionality, A and C can be classified into the amino group while G and U into the keto group (Chen et al. 2016a, c). Therefore, three coordinates $(x, y, z)$ were used to represent the chemical properties of the four nucleotides and were assigned 1 or 0 values (Chen et al. 2015c). If the $x$ coordinate stands for the ring structure, $y$ for the hydrogen bond, and $z$ for the chemical functionality, nucleotide in RNA sequence can be encoded by $(x_i, y_i, z_i)$, where

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

Thus, nucleotides A, C, G, and U can be transferred to the coordinates (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively.

To integrate the information of the sequence neighbor surrounding $m^6A$, the density $d_i$ of any nucleotide $n_j$ at position $i$ in a sequence was defined as follows:

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{l} f(n_j), \quad f(n_j) = \begin{cases} 1 & \text{if} \quad n_j = q \\ 0 & \text{other cases} \end{cases}, \tag{2}$$

where $l$ is the sequence length, $|N_i|$ is the length of the $i$th prefix string $\{n_1, n_2, \ldots, n_i\}$ in the sequence, and $q \in \{A, C, G, U\}$.

Therefore, the sequence with a length of $l$ will be encoded by a $(4 \times l)$-dimensional vector. For example, the sequence "AGCGUAAC" can be represented by {1, 1, 1, 1, 1, 0, 0, 0.5, 0, 0, 1, 0.33, 1, 0, 0, 0.5, 0, 1, 0, 0.2, 1, 1, 1, 0.33, 1, 1, 1, 0.43, 0, 0, 1, 0.25}. Accordingly, each 25-bp long sequence in the benchmark dataset can be represented by a 100 $(4 \times 25)$-dimensional vector.

### Support vector machine

As a smart machine learning algorithm, support vector machine (SVM) has been widely used to build models in computational genomics and proteomics (Chen et al. 2013, 2014b; Lin et al. 2013; Cao et al. 2014a, b). Therefore, in the current study, the LibSVM package 3.18 was used to

perform the predictions. The popular radial basis function (RBF) was chosen as the kernel of SVM, where the regularization parameter $C$ and kernel parameter $\gamma$ were optimized using grid search, and their actual values thus obtained for the current study were $C = 0.5$ and $\gamma = 0.0078125$.

### Metrics for validation and evaluation

The performance of the proposed method was evaluated using sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthew's correlation coefficient (MCC), which are expressed as

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\,\%, \tag{3}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\,\%, \tag{4}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \times 100\,\%, \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}, \tag{6}$$

where TP represents the number of the correctly recognized m$^6$A-containing sequences, TN represents the number of the correctly recognized non-m$^6$A-containing sequences, FP represents the number of non-m$^6$A-containing sequences recognized as m$^6$A-containing sequences, and FN represents the number of m$^6$A-containing sequences recognized as non-m$^6$A-containing sequences, respectively.

Moreover, to objectively examine the performance of the proposed predictor, both the ROC (receiver operating characteristic) curve and the precision–recall curve were plotted. The former plots the true positive rate (sensitivity) against the false positive rate (specificity), and the latter plots precision (the fraction of TP in all predicted positives) against recall (sensitivity).

## Results

### Identification of m$^6$A sites

As demonstrated by Eqs. 28–32 in a recent review (Chou 2011), the jackknife test is deemed as the least arbitrary and most objective cross-validation method and has been increasingly adopted by researchers to examine the quality of various computational models (Chen et al. 2012, 2014a; Feng et al. 2014a, b). Therefore, the jackknife test was used to examine the performance of the proposed model. In the jackknife test, the proposed method obtained an Acc of 84.39 % with Sn of 68.78 % and Sp of 100 % for identifying m$^6$A sites, Table 1. Moreover, the ROC curve and precision–recall curve were plotted in Fig. 1. As shown in Fig. 1, the AUROC and AUPRC are 0.85 and 0.87, respectively, indicating the reliability of the proposed model in identifying m$^6$A sites in *A. thaliana*.

**Table 1** Comparison of different methods for identifying m6A by the jackknife test in *Arabidopsis thaliana*

| Method | Sn (%) | Sp (%) | Acc (%) | MCC |
|---|---|---|---|---|
| Naïve Bayes | 71.57 | 91.88 | 81.73 | 0.65 |
| Random Forest | 76.65 | 78.68 | 77.66 | 0.55 |
| J48 | 74.62 | 70.30 | 72.46 | 0.45 |
| SVM | 68.78 | 100.00 | 84.39 | 0.72 |

### Cross-strain validation

Both Can-0 and Hen-16 strain possess their own specific m$^6$A sites that associated with gene activation (Luo et al. 2014). Since the proposed model was trained based on the common m$^6$A sites of Can-0 and Hen-16 strains, it is interesting to see its performances on identifying the strain-specific m$^6$A sites. To this end, we obtained 266 Can-0-specific and 195 Hen-16-specific m$^6$A containing sequences from Luo et al.'s (2014) work, which did not overlap (1-nucleotide) any peak in any two replicates of the other strain. All these sequences are also 25-nt long with the m$^6$A in the center and with the sequence similarity less than 60 %.

The model was then applied to identify the Can-0 and Hen-16 specific m$^6$A sites, respectively. We found that the proposed model could accurately identify 198 m$^6$A sites from the 266 Can-0 specific m$^6$A sites with the Acc of 74.43 % and 144 m$^6$A sites from the 195 Hen-16 specific m$^6$A sites with the Acc of 73.84 %, respectively.

### Comparison with other methods

To further testify its superiority, we compared the performance of the proposed method with that of the other state-of-the-art classifiers, i.e., Naïve Bayes, Random Forest and J48 Tree as implemented in WEKA (Frank et al. 2004). The jackknife test results of different classifiers for identifying m$^6$A sites were reported in Table 1.

Although Sn of the proposed method is lower than those of Naïve Bayes, Random Forest and J48 Tree, its Sp, Acc, and MCC are all higher than those of Naïve Bayes, Random Forest and J48 Tree, indicating that the proposed SVM-based model can be effectively used to identify m$^6$A in *A. thaliana*.

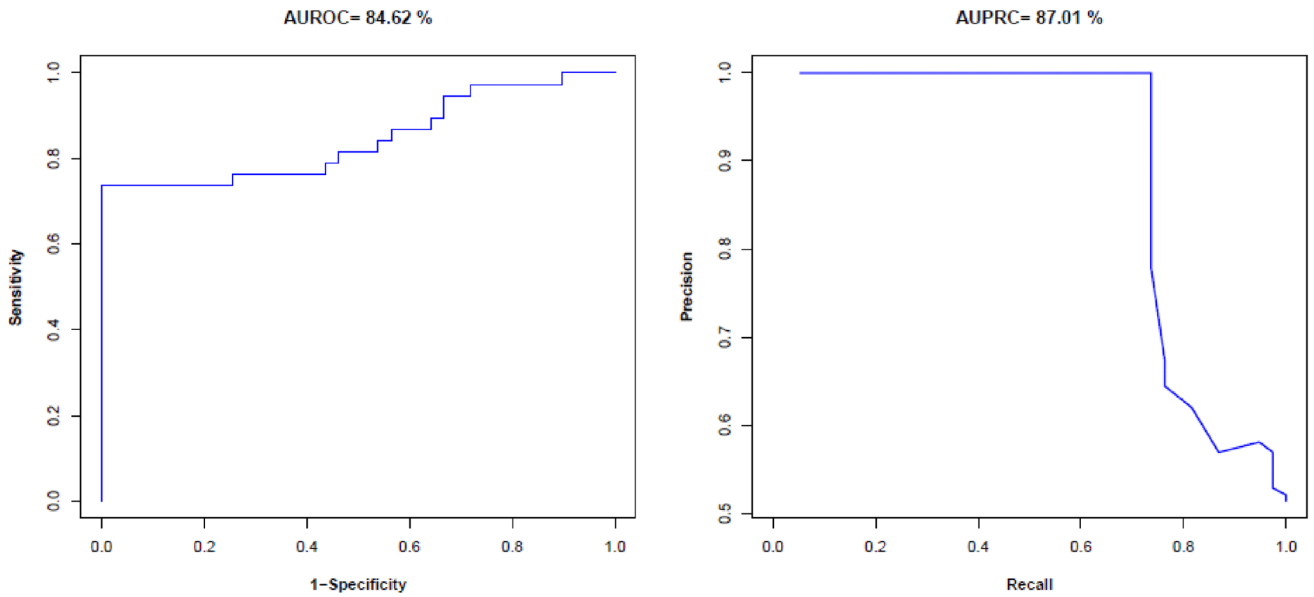


**Fig. 1** Graphical illustration to show the performance of the proposed method for identifying $m^6A$ sites in *A. thaliana*. The performances are illustrated by means of the ROC curves (*left*) and precision–recall curves (*right*)

## Webserver

For the convenience of scientific community, a freely accessible online webserver was established. The user guide on how to use it is given bellow.

**Step 1.** Open the webserver at http://lin.uestc.edu.cn/server/M6ATH, and its top page will be shown as in Fig. 2.

**Step 2.** Either type or copy/paste the query RNA sequences into the input box at the center of Fig. 2.

**Step 3.** Click on the 'Submit' button to see the predicted result. For example, if use the query RNA sequences in the 'Example' window as the input, the outcomes are as follows: A at position 13 in the first and second query sequences are $m^6A$; none of the A in the third and fourth query sequences is $m^6A$. All these results are fully consistent with the experimental observations.

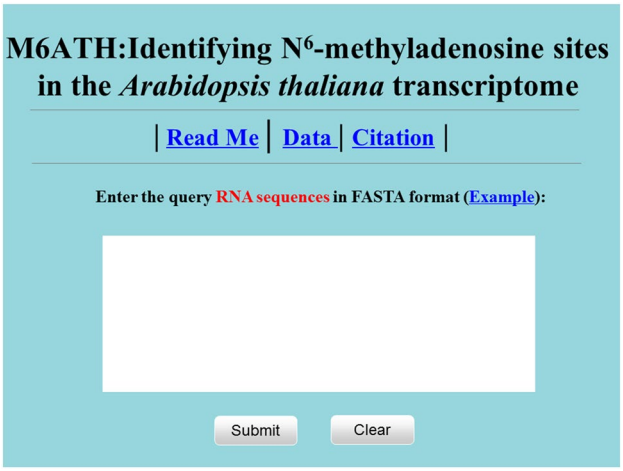


**Fig. 2** Semi-screenshot for the top page of the webserver which is available at http://lin.uestc.edu.cn/server/M6ATH

## Discussions

Benefitting from the high-throughput sequencing data, in the present work, we proposed a computational method to identify $m^6A$ sites in *A. thaliana*, in which RNA sequences were encoded by nucleotide chemical properties and nucleotide composition. In the jackknife test, the proposed method obtained an overall accuracy of 84.39 % for identifying $m^6A$ sites in *A. thaliana*. It is encouraging that the proposed method is also quite good in identifying the strain-specific $m^6A$ sites.

To further demonstrate its performance on the problem of identifying $m^6A$ sites in *A. thaliana*, comparisons were carried out between the proposed method and the other state-of-the-art classifiers. We found that our proposed SVM-based model outperforms other classifiers for identifying $m^6A$ in *A. thaliana*. To enhance the value of the actual applications of the proposed model, a webserver was established at http://lin.uestc.edu.cn/server/M6ATH by which users can easily obtain their desired results.

It has not escaped our notice that the current method is also suitable for identifying $m^6A$ sites in other plants, once the experimental data that can be used to train the models are available. Therefore, it is anticipated that

our method will become a useful tool for identifying $m^6A$ and other post-transcriptional modifications in *A. thaliana*.

**Acknowledgments** This work was supported by Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), China Postdoctoral Science Foundation (No. 2015M582533), the Scientific Research Foundation of the Education Department of Sichuan Province (No. 2015JY0100), and the Fundamental Research Funds for the Central Universities, China (Nos. ZYGX2015J144, ZYGX2015Z006).

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants performed by any of the authors.

## References

Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF (2011) The RNA modification database, RNAMDB: 2011 update. Nucleic Acids Res 39:D195–D201

Cao R, Wang Z, Cheng J (2014a) Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. BMC Struct Biol 14:13

Cao R, Wang Z, Wang Y, Cheng J (2014b) SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinform 15:120

Chen W, Feng P, Lin H (2012) Prediction of replication origins by calculating DNA structural properties. FEBS Lett 586:934–938

Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41:e68

Chen W, Feng PM, Deng EZ, Lin H, Chou KC (2014a) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem 462:76–83

Chen W, Feng PM, Lin H, Chou KC (2014b) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Res Int 2014:623149

Chen T, Hao YJ, Zhang Y, Li MM, Wang M, Han W, Wu Y, Lv Y, Hao J, Wang L, Li A, Yang Y, Jin KX, Zhao X, Li Y, Ping XL, Lai WY, Wu LG, Jiang G, Wang HL, Sang L, Wang XJ, Yang YG, Zhou Q (2015a) m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. Cell Stem Cell 16:289–301

Chen W, Feng P, Ding H, Lin H, Chou KC (2015b) iRNA-Methyl: identifying *N*(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 490:26–33

Chen W, Tran H, Liang Z, Lin H, Zhang L (2015c) Identification and analysis of the *N*(6)-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. Sci Rep 5:13859

Chen W, Feng P, Tang H, Ding H, Lin H (2016a) Identifying 2′-*O*-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. Genomics 107:255–258

Chen W, Tang H, Ye J, Lin H, Chou KC (2016b) iRNA-PseU: identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 5:e332

Chen W, Tang H, Lin H (2016) MethyRNA: a web server for identification of N6-methyladenosine sites. J Biomol Struct Dyn. doi:10.1080/07391102.2016.1157761

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, Sorek R, Rechavi G (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 485:201–206

Feng P, Jiang N, Liu N (2014a) Prediction of DNase I hypersensitive sites by using pseudo nucleotide compositions. Sci World J 2014:740506

Feng P, Lin H, Chen W, Zuo Y (2014b) Predicting the types of J-proteins using clustered amino acids. Biomed Res Int 2014:935719

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS, Ben-Haim MS, Eyal E, Yunger S, Pinto Y, Jaitin DA, Viukov S, Rais Y, Krupalnik V, Chomsky E, Zerbib M, Maza I, Rechavi Y, Massarwa R, Hanna S, Amit I, Levanon EY, Amariglio N, Stern-Ginossar N, Novershtern N, Rechavi G, Hanna JH (2015) Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. Science 347:1002–1006

Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, Yi C, Lindahl T, Pan T, Yang YG, He C (2011) $N^6$-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol 7:885–887

Lin H, Chen W, Ding H (2013) AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PLoS One 8:e75726

Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods 12:767–772

Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, Dai Q, Chen W, He C (2014) A METTL3-METTL14 complex mediates mammalian nuclear RNA $N^6$-adenosine methylation. Nat Chem Biol 10:93–95

Luo GZ, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, Liu J, Chen K, Jia G, Bergelson J, He C (2014) Unique features of the m6A methylome in *Arabidopsis thaliana*. Nat Commun 5:5630

Meyer KD, Jaffrey SR (2014) The dynamic epitranscriptome: $N^6$-methyladenosine and gene expression control. Nat Rev Mol Cell Biol 15:313–326

Nilsen TW (2014) Molecular biology. Internal mRNA methylation finally finds functions. Science 343:1207–1208

Schwartz S, Agarwala SD, Mumbach MR, Jovanovic M, Mertins P, Shishkin A, Tabach Y, Mikkelsen TS, Satija R, Ruvkun G, Carr SA, Lander ES, Fink GR, Regev A (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell 155:1409–1421

Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q (2016) SRAMP: prediction of mammalian $N^6$-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 44:e91