

SCIENTIFIC REPORTS



OPEN

Identification and analysis of the N⁶-methyladenosine in the *Saccharomyces cerevisiae* transcriptome

Received: 13 March 2015

Accepted: 10 August 2015

Published: 07 September 2015

Wei Chen^{1,2}, Hong Tran², Zhiyong Liang³, Hao Lin³ & Liqing Zhang²

Knowledge of the distribution of N⁶-methyladenosine (m⁶A) is invaluable for understanding RNA biological functions. However, limitation in experimental methods impedes the progress towards the identification of m⁶A site. As a complement of experimental methods, a support vector machine based-method is proposed to identify m⁶A sites in *Saccharomyces cerevisiae* genome. In this model, RNA sequences are encoded by their nucleotide chemical property and accumulated nucleotide frequency information. It is observed in the jackknife test that the accuracy achieved by the proposed model in identifying the m⁶A site was 78.15%. For the convenience of experimental scientists, a web-server for the proposed model is provided at <http://lin.uestc.edu.cn/server/m6Apred.php>.

The methylation of the 6th nitrogen of Adenosine (N⁶-methyladenosine, m⁶A) is the most prevalent form of RNA modification and is found in all three domains of life¹. m⁶A is catalyzed by an evolutionarily conserved, multi-component enzyme². Unlike adenosine-to-inosine editing, m⁶A does not alter the coding capacity of transcripts^{3,4}. However, it has been demonstrated that m⁶A is associated with a number of key biological processes including mRNA splicing, export, stability, and immune tolerance⁵⁻⁷. Moreover, it has been reported that m⁶A is closely correlated with the mammalian brain development⁸. The regulatory role of m⁶A in cell division has also been reported in plants⁹.

By using high-throughput techniques such as MeRIP-Seq⁸ and m⁶A-seq¹⁰, the distribution of m⁶A has been characterized in the human and mouse transcriptomes⁸. The experimental results revealed that m⁶A sites tend to occur near the stop codon, in 3' UTR, and within long internal exons^{8,11}. The nonrandom distribution of m⁶A sites across the genome is highly conserved from yeasts to humans^{11,12}, suggesting that m⁶A modification is both fundamental and important for organisms. The experimental results also demonstrated that the m⁶A sites identified in the yeast harbored the RGAC (R = A/G) consensus motif², reminiscent of the mammalian RRACU (R = A/G) motif¹¹. Similar to epigenetic DNA and histone modifications, m⁶A modification is also dynamic and reversible, the m⁶A patterns change in different cell types¹⁰ or when cells are stressed¹².

The experimental methods yielded quite encouraging results and did play a role in promoting the research progress on identifying the distribution of m⁶A in the transcriptome. However, resolution of both m⁶A-seq and MeRIP-seq methods is low, only ~24 nt (nucleotide) around the methylated adenosine¹¹. Therefore, experimental methods cannot pinpoint which adenosine residue is actually modified. In addition, current experimental methods are both costly and time consuming. Therefore, it is necessary

¹Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China. ²Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, United States of America. ³Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China. Correspondence and requests for materials should be addressed to W.C. (email: chenweiimu@gmail.com) or L.Z. (email: lqzhang@vt.edu)



Figure 1. Sequence logo of the 10 upstream and 10 downstream nucleotides surrounding m^6A sites.

Features	<i>Sn</i> (%)	<i>Sp</i> (%)	<i>Acc</i> (%)
Ring Structure	69.27	63.43	66.34
Functional Group	70.70	69.90	70.31
Hydrogen Bond	74.18	68.46	71.32
Nucleotide chemical property	75.23	78.02	75.87
Nucleotide chemical property and accumulated nucleotide frequency	79.21	77.04	78.13

Table 1. The predictive results by using different features for m^6A identification.

to develop new methods for studying the distribution and function of m^6A . As excellent complements to experimental techniques, computational methods will speed up genome-wide m^6A detection.

However, to the best of our knowledge, there is no computational tool available for the discovery of m^6A . In the present study, we propose a support vector machine based method to identify the m^6A sites in the *Saccharomyces cerevisiae* genome. By using the nucleotide chemical property and accumulated nucleotide frequency information, the sequence-order effects and nucleotide physicochemical properties are integrated together in the proposed model. In the jackknife test, an overall accuracy of 78.15% is achieved in identifying the m^6A sites in the benchmark dataset. For the convenience of the experimental scientists, a web-server for the proposed model is provided at <http://lin.uestc.edu.cn/server/m6Apred.php>.

Results

Nucleotide preference. In order to understand nucleotide preference surrounding m^6A sites, based on the benchmark dataset, we computed the sequence logos of the 10 upstream and 10 downstream nucleotides using WebLogo¹³. As shown in Fig. 1, besides the well-known consensus motif RGAC (R = A/G) located at -2 to 1 bp relative to the m^6A site (position 0)¹², strong preference of nucleotides in both upstream and downstream sequences surrounding the m^6A site were also observed. The adenines are favored at positions -4 , -3 , and -2 , whereas the uracils are favored at positions from $+2$ to $+4$. In contrast, except for the RGAC (R = A/G) located at -2 to 1 , no exclusive preference of nucleotides was observed surrounding the unmethylated adenosine.

m^6A sites identification. Three cross-validation methods, the sub-sampling (or K-fold cross-validation) test, the independent dataset test, and the jackknife test, are often used to evaluate the quality of a predictor. Among the three methods, the jackknife test is deemed as the least arbitrary and most objective¹⁴ and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors^{15–17}. Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study. In the jackknife test, each sample in the training dataset is in turn singled out as an independent test sample and all the properties are calculated without including the one being identified.

In order to compare the contribution of the features for m^6A site identification, we firstly performed the predictions using individual nucleotide chemical property and their combinations. The predictive results are reported in Table 1. Among the three kinds of nucleotide chemical properties, the hydrogen bond yields the highest predictive accuracy (71.32%), indicating that it has the largest contribution for m^6A site identification. However, the predictive accuracies obtained by using each kind of nucleotide chemical property alone are all lower than that obtained by using all three kinds of nucleotide chemical properties (Table 1).

Considering the observed nucleotide preference surrounding the m^6A sites (Fig. 1) and the above results, the accumulated nucleotide frequency and nucleotide chemical property were combined to encode the sequences in the training dataset. Hence, each 21-bp long sequence in the dataset was represented by an 84 (4×21)-dimensional vector (see Methods) and used as the input of SVM to train the model for identifying m^6A sites. In the jackknife test, the proposed model obtained an accuracy of

Classifier	S_n (%)	S_p (%)	Acc (%)
High	38.22	94.95	66.59
Medium	55.05	90.02	72.54
Low	68.39	84.98	76.68

Table 2. Performance of the proposed model at different thresholds on jackknife test.

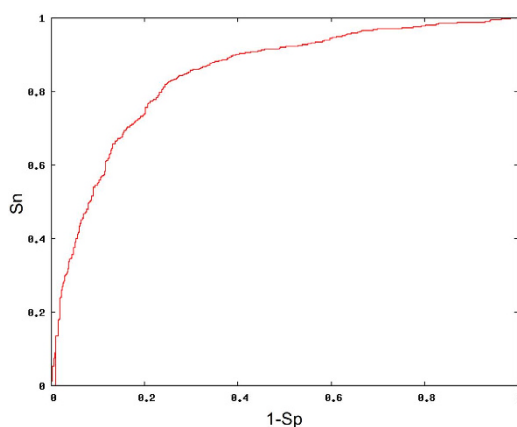


Figure 2. A graphical illustration to show the performance of the model by means of the ROC curve. The vertical coordinate is the true positive rate (S_n) while horizontal coordinate is the false positive rate ($1-S_p$). The area under the ROC curve (AUROC) is 0.84.

78.15% with a sensitivity of 79.21% and specificity of 77.04% (Table 1). The predictive accuracy thus obtained is higher than that obtained by merely using nucleotide chemical properties (75.87%), indicating that nucleotide frequency contributes slightly to the identification of m^6A sites.

As the performance of the proposed model may depend on the threshold, similar to a recent work¹⁸, three thresholds of high, medium and low obtained in jackknife test were selected with the specificity values of 95%, 90% and 85%, respectively. The predictive performances of the proposed model with these different thresholds were reported in Table 2. Meanwhile, in order to provide a graphical illustration to show the performance of the model as its discrimination threshold varied, the ROC curve was plotted in Fig. 2 and an AUROC of 0.84 was obtained.

To ensure that the predictive accuracy is not sensitive to the selection of negative data, we repeated the random sampling procedure ten times and obtained ten random samples of negative datasets for downstream training and prediction. The predictive results of these models for identifying m^6A sites in the jackknife test were reported in Supplementary Table S1. We found that the predictive accuracy is not affected by the selection of negative data.

In addition, the proposed model was also evaluated on the independent testing dataset (see Methods). We found that the proposed model obtained an accuracy of 75.73% with a sensitivity of 53.89% and a specificity of 79.07% for identifying m^6A sites on the testing dataset with the positive-to-negative ratio of 1:10. The precision-recall curve, which plots the corresponding precision-recall pairs over a range of values, was also plotted in Supplementary Figure S1. These results demonstrate the reliability of the model developed in this study.

Comparison with Other classifiers. To further demonstrate the power of the proposed method, we also did some comparative calculations as described below.

First, based on the sequence similarity principle, we used the classic sequence similarity search-based tool BLAST¹⁹ to conduct the jackknife test on the same benchmark dataset. The results thus obtained are given in Table 3, from which we can see that the percentage rate for Acc obtained by BLAST is about 10% lower than the proposed model for m^6A identifications.

Second, we also compared the predictive results of the proposed method with that of four other commonly used classifiers, i.e., Naïve Bayes²⁰, Logistic Function²¹, RBFNetwork²², and Random Forest²³ as implemented in WEKA²⁴. The jackknife test results for identifying m^6A sites in the benchmark dataset for different classifiers were listed in Table 3. It is shown that the sensitivity, specificity, accuracy and AUROC of the proposed SVM model are all higher than that of Naïve Bayes, Logistic Function, and Random Forest. Although the specificity of the proposed method is lower than that of RBFnetwork,

Classifier	<i>Sn</i> (%)	<i>Sp</i> (%)	<i>Acc</i> (%)	AUROC
Blast	70.75	67.55	69.11	–
Naïve Bayes	78.72	70.91	74.81	0.82
Logistic Function	79.32	74.76	77.04	0.83
RBFNetwork	61.18	84.49	72.83	0.79
Random Forest	78.73	64.78	71.75	0.78
SVM	79.21	77.04	78.15	0.84

Table 3. Comparison of different classifiers for m⁶A identification.

Figure 3. A semi-screenshot for the top page of the web-server at <http://lin.uestc.edu.cn/server/m6Apred.php>.

its sensitivity, accuracy, and AUROC are all higher than that of RBFnetwork. Hence, these results suggest that our proposed method is promising and has great potential to become a useful tool for m⁶A identifications.

Web-server

To enable applications of the proposed model and for the convenience of the vast majority of experimental scientists, an online predictor is created. The step-by-step guide on how to use it is provided as the following:

Step 1. Open the web server at <http://lin.uestc.edu.cn/server/m6Apred.php> and you will see the top page on your computer screen, as shown in Fig. 3. Click on the **Read Me** button to see a brief introduction about the predictor and the caveat when using it.

Step 2. On clicking the open circle, the threshold (All, High, Medium, or Low) as reported in Table 2 will be selected. Either type or copy/paste the query RNA sequences into the input box at the center of Fig. 3. The input sequence should be in FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol (“>”) in the first column, followed by lines of sequence data. The words right after the “>” symbol in the single initial line are optional and only used for the purpose of identification and description. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a “>” appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

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 following results will be shown on the screen: the outcome for the 1st query example is: the ‘A’ at position 11 is methylated with a probability of 0.92 and the ‘A’ at position 32 is also methylated with a probability of 0.92. The outcome for the 2nd query sample

is: the A at position 11 is unmethylated with a probability of 0.96. All these results are consistent with the experimental observations.

Step 4. Click on the [Data](#) button to download the datasets used to train and test the model.

Step 5. Click on the [Citation](#) button to find the relevant paper that reports the detailed development and algorithm of the model.

Caveats. Each of the input query sequences must be 21 bp or longer and only contains valid characters: 'A', 'C', 'G', 'U'.

Conclusions

By using the nucleotide chemical property and nucleotide density information, we proposed a support vector machine based model to identify m⁶A sites in the *Saccharomyces cerevisiae* transcriptome. To identify the key features for m⁶A site identification, we compared the predictive results obtained by using different kinds of parameters (Table 1). In comparison with accumulated nucleotide frequency, nucleotide chemical property plays the more important roles for m⁶A site identification. Among the three considered nucleotide chemical properties, the hydrogen bond has the largest contribution for m⁶A site identification, consistent with the recent finding that the hydrogen bond is implicated in formation of RNA secondary structure²⁵ which decreases the m⁶A methylation²⁶.

In addition, we also compared the predictive accuracy of SVM with four other commonly used classification methods for m⁶A site identification. We found that the predictive result of SVM is better than those of Naïve Bayes, Logistic Function and Random Forest. This is likely due to the limited number of the experimentally validated m⁶A sites that used to train the models. Naïve Bayes, Logistic Function and Random Forest require a large number of samples to train, whereas SVM needs fewer training data.

For the convenience of researchers in the scientific community, a web-server for the proposed model is provided. We hope that these results will provide further insights into the understanding of the distribution and function of m⁶A modifications. As the current method is only applicable to *Saccharomyces cerevisiae*, future work will expand to other species to train and improve the model.

Methods

Dataset. By using the m⁶A-seq technique, Schwartz *et al.* identified 1,307 methylated adenine (m⁶A) sites centered around RGAC motifs from 1,183 genes in *Saccharomyces cerevisiae*¹². In order to obtain a high quality training dataset and avoid experiment bias, the 832 m⁶A sites with distances to the detected m⁶A-seq peaks less than 10 bp were selected as positive samples of the training dataset¹². The pairwise sequence similarity within all the positive training samples is less than 85%. The remaining 475 (1,307–832 = 475) m⁶A sites were used to construct the independent testing dataset. The negative samples were obtained by the following steps. By searching *Saccharomyces cerevisiae* genome, we obtained 33,280 adenines centered around the RGAC consensus motif, which were not detected by the m⁶A-seq technique. Therefore, the 33,280 adenines were deemed as nonmethylated adenine. To balance out the numbers between positive and negative samples in model training, we randomly picked 832 samples from the 33,280 non-methylated adenines and used them as negative samples. Following these procedures, we obtained a benchmark dataset including 832 m⁶A site containing sequences and 832 non-m⁶A site containing sequences, respectively. To examine whether the predictive accuracy is sensitive to the selection of negative data, we repeated the random sampling procedure ten times and obtained ten random samples of negative datasets for downstream training and prediction. We also randomly fetched 4,750 negative samples from the ten negative datasets and merged them with the above mentioned 475 samples in the testing dataset. By doing so, an independent testing dataset with the positive-to-negative ratio of 1:10 (475:4,750) was obtained.

It was observed via preliminary trials that when the length of the sequences in the benchmark dataset is 21 bp with the m⁶A in the center, the corresponding predictive results were most promising. Accordingly, all the sequences in the training and testing dataset are 21 bp long and are available at <http://lin.uestc.edu.cn/server/m6Apred.php>.

Sequence encoding. One of the keys in developing a model for identifying genomic attributes is to encode the biological samples with effective expressions. In the present study, nucleotide chemical properties and density information of each nucleotide in RNA sequences were considered.

Chemical property of each nucleotide. There are four different kinds of nucleotides, i.e., adenine (A), guanine (G), cytosine (C) and uracil (U), found in RNA. Each nucleotide has different chemical structure and chemical binding. Shown in Fig. 4, adenine and guanine have two rings, while cytosine and uracil have only one ring. Although RNA is generally single stranded, its biological functions are correlated with the secondary structure. When forming secondary structures, in terms of hydrogen bond, guanine and cytosine have strong hydrogen bonds, whereas adenine and uracil have weak hydrogen bonds. Additionally, in terms of chemical functionality, adenine and cytosine can be classified into the same group, called amino group, while guanine and uracil into the keto group. Therefore, the four kinds of nucleotides can be classified into three different groups in terms of these chemical properties (Table 4).

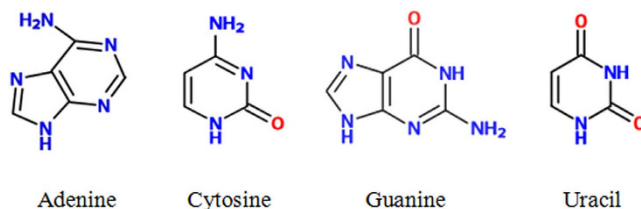


Figure 4. Chemical structure of each nucleotide.

Chemical property	Class	Nucleotides
Ring Structure	Purine	A, G
	Pyrimidine	C, U
Functional Group	Amino	A, C
	Keto	G, U
Hydrogen Bond	Strong	C, G
	Weak	A, U

Table 4. Chemical property of nucleotide in RNA sequence.

In order to include these chemical properties in RNA encoding, we define three coordinates (x, y, z) to represent three chemical groups and assign 1 or 0 values. Hence, each nucleotide $s_i = (x_i, y_i, z_i)$ in the sequence can be encoded by the following formula²⁷.

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

where the coordinate value of each nucleotide is determined by their chemical property of the nucleotide as shown in Table 4. Thus, based on chemical properties, A can be represented by coordinates (1, 1, 1), C can be represented by coordinates (0, 1, 0), G can be represented by coordinates (1, 0, 0), U can be represented by coordinates (0, 0, 1).

Accumulated nucleotide frequency. In order to include the nucleotide frequency information and the distribution of each nucleotide in the RNA sequence, we define the density d_i of any nucleotide s_i at position i in RNA sequence by the following formula²⁶,

$$d_i = \frac{1}{|S_i|} \sum_{j=1}^i f(s_j), \quad f(q) = \begin{cases} 1 & \text{if } s_j = q \\ 0 & \text{other cases} \end{cases} \quad (2)$$

where l is the sequence length, $|S_i|$ is the length of the i -th prefix string $\{s_1, s_2, \dots, s_i\}$ in the sequence, $q \in \{A, C, G \text{ or } U\}$. Suppose an example sequence “UCGUUCAUGG”. The density of ‘U’ is 1 (1/1), 0.5 (2/4), 0.6 (3/5), 0.5 (4/8) at positions 1, 4, 5, and 8, respectively. The density of ‘C’ is 0.5 (1/2), 0.33 (2/6) at positions 2 and 6, respectively. The density of ‘G’ is 0.33 (1/3), 0.22 (2/9), 0.3 (3/10) at positions 3, 9, and 10, respectively. The density of ‘A’ is 0.14 (1/7) at position 7.

By integrating both the nucleotide chemical property and accumulated nucleotide information, the sample sequence “UCGUUCAUGG” can be represented by $\{(0, 0, 1, 1), (0, 1, 0, 0.5), (1, 0, 0, 0.33), (0, 0, 1, 0.5), (0, 0, 1, 0.6), (0, 1, 0, 0.33), (1, 1, 1, 0.14), (0, 0, 1, 0.5), (1, 0, 0, 0.22), (1, 0, 0, 0.3)\}$. By doing so, not only the chemical property was considered, but also the long range sequence order information was incorporated. Therefore, the samples in the benchmark dataset were encoded in terms of both nucleotide chemical property and nucleotide densities.

Support vector machine. The SVM classification algorithm has been widely used in the realm of bioinformatics^{28–30}. Its basic principle is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane with the maximal margin in this space.

In this study, the libsvm-3.18 package was used as an implementation of SVM, which can be downloaded from <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. Because of its effectiveness and speed in nonlinear classification process, the radial basis kernel function (RBF) was selected to perform the prediction. A grid search method was used to optimize the regularization parameter C and kernel parameter γ . The probability score obtained from SVM was used to make predictions.

Performance evaluations. The performance of the model was evaluated using the following metrics: sensitivity (Sn) also named recall, specificity (Sp), precision and accuracy (Acc), which can be expressed as

$$\left\{ \begin{array}{l} Sn = \frac{TP}{TP + FN} \\ Sp = \frac{TN}{TN + FP} \\ Precision = \frac{TP}{TP + FP} \\ Acc = \frac{TP + TN}{TP + FN + TN + FP} \end{array} \right. \quad (3)$$

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

Meanwhile, in order to provide a graphical illustration to show the performance of the model as its discrimination threshold varied, the ROC (receiver operating characteristic) curve was created, where its vertical coordinate is for the true positive rate while horizontal coordinate for the false positive rate. The best possible prediction method would yield a point with the coordinate (0, 1) representing 100% true positive rate and 0 false positive rate or 100% specificity. Therefore, the (0, 1) point is also considered as a perfect classification. A completely random guess would give a point along a diagonal from the point (0, 0) to (1, 1). The AUROC (area under the ROC curve) is often used to indicate the performance quality of a binary classifier: the value 0.5 of AUROC is equivalent to random prediction while 1 of AUROC represents a perfect one.

References

- Cantara, W. A. *et al.* The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–201, doi: 10.1093/nar/gkq1028 (2011).
- Bokar, J. A., Shambaugh, M. E., Polayes, D., Matera, A. G. & Rottman, F. M. Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. *RNA* **3**, 1233–1247 (1997).
- Dai, Q. *et al.* Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and N6-methyladenosine. *Nucleic Acids Res.* **35**, 6322–6329, doi: 10.1093/nar/gkm657 (2007).
- Levanon, E. Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol.* **22**, 1001–1005, doi: 10.1038/nbt996 (2004).
- Kariko, K., Buckstein, M., Ni, H. & Weissman, D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* **23**, 165–175, doi: 10.1016/j.immuni.2005.06.008 (2005).
- Jia, G. *et al.* N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol.* **7**, 885–887, doi: 10.1038/nchembio.687 (2011).
- Nilsen, T. W. Molecular biology. Internal mRNA methylation finally finds functions. *Science* **343**, 1207–1208, doi: 10.1126/science.1249340 (2014).
- Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635–1646, doi: 10.1016/j.cell.2012.05.003 (2012).
- 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* **20**, 1278–1288, doi: 10.1105/tpc.108.058883 (2008).
- Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nat Protoc.* **8**, 176–189, doi: 10.1038/nprot.2012.148 (2013).
- Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201–206, doi: 10.1038/nature11112 (2012).
- Schwartz, S. *et al.* High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **155**, 1409–1421, doi: 10.1016/j.cell.2013.10.047 (2013).
- Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190, doi: 10.1101/gr.849004 (2004).
- Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol.* **273**, 236–247, doi: 10.1016/j.jtbi.2010.12.024 (2011).
- Chen, W., Feng, P. & Lin, H. Prediction of replication origins by calculating DNA structural properties. *FEBS Lett.* **586**, 934–938, doi: 10.1016/j.febslet.2012.02.034 (2012).
- Chen, W. & Lin, H. Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information. *Biochem Biophys Res Commun.* **401**, 382–384, doi: 10.1016/j.bbrc.2010.09.061 (2010).
- Guo, S. H. *et al.* iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **30**, 1522–1529, doi: 10.1093/bioinformatics/btu083 (2014).
- Pan Z. C. *et al.* Systematic analysis of the *in situ* crosstalk of tyrosine modifications reveals no additional natural selection on multiply modified residues. *Sci Rep.* **4**, 7331, doi: 10.1038/srep07331 (2014).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol.* **215**, 403–410, doi: 10.1016/S0022-2836(05)80360-2 (1990).
- Feng, P. M., Ding, H., Chen, W. & Lin, H. Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput Math Methods Med.* **2013**, 530696, doi: 10.1155/2013/530696 (2013).
- Freedman, D. A. *Statistical Models: Theory and Practice.* (Cambridge University Press, 2009).
- Mountrakis, G. & Zhuang, W. Integrating local and global error statistics for multi-scale RBF network training: an assessment on remote sensing data. *PLoS One* **7**, e40093, doi: 10.1371/journal.pone.0040093 (2012).
- Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

24. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479–2481, doi: 10.1093/bioinformatics/bth261 (2004).
25. Lu, X. J., Olson, W. K. & Bussemaker, H. J. The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex. *Nucleic Acids Res.* **38**, 4868–4876, doi: 10.1093/nar/gkq155 (2010).
26. Narayan, P., Ludwiczak, R. L., Goodwin, E. C. & Rottman, F. M. Context effects on N6-adenosine methylation sites in prolactin mRNA. *Nucleic Acids Res.* **22**, 419–426 (1994).
27. Golam Bari, A. T. M., Rokeya Reaz, M. & Jeong, B. S. DNA Encoding for Splice Site Prediction in Large DNA Sequence. *MATCH Communications in Mathematical and in Computer Chemistry* **71**, 241–258 (2014).
28. Liu, B. *et al.* Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **30**, 472–479, doi: 10.1093/bioinformatics/btt709 (2014).
29. Chen, W., Feng, P. M., Lin, H. & Chou, K. C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **41**, e68, doi: 10.1093/nar/gks1450 (2013).
30. Feng, P. M., Chen, W., Lin, H. & Chou, K. C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem.* **442**, 118–125, doi: 10.1016/j.ab.2013.05.024 (2013).

Acknowledgements

This work was supported by National Nature Scientific Foundation of China (Nos. 61100092 and 61202256), Nature Scientific Foundation of Hebei Province (No. C2013209105), Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

Author Contributions

W.C. and L.Z. conceived and designed the experiments; H.T. and Z.L. analyzed the m⁶A-seq data; H.L. and Z.L. implemented SVM and created the back end server; W.C. and L.Z. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chen, W. *et al.* Identification and analysis of the N⁶-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci. Rep.* **5**, 13859; doi: 10.1038/srep13859 (2015).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>