





## TECHNICAL NOTE

# m6ASNP: a tool for annotating genetic variants by m<sup>6</sup>A function

Shuai Jiang<sup>1,2,†</sup>, Yubin Xie<sup>2,†</sup>, Zhihao He<sup>2,†</sup>, Ya Zhang<sup>2</sup>, Yuli Zhao<sup>2</sup>, Li Chen<sup>2</sup>, Yueyuan Zheng<sup>2</sup>, Yanyan Miao<sup>2</sup>, Zhixiang Zuo <sup>1,\*</sup> and Jian Ren <sup>1,2,3,\*</sup>

<sup>1</sup>Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou 510060, China, <sup>2</sup>State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China and <sup>3</sup>Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China

\*Correspondence address. Jian Ren. Tel/Fax: +86 20 87342325; E-mail: [zuozyhx@sysucc.org.cn](mailto:zuozyhx@sysucc.org.cn)  <http://orcid.org/0000-0002-4161-1292>; Zhixiang Zuo. E-mail: [renjian.sysu@gmail.com](mailto:renjian.sysu@gmail.com)  <http://orcid.org/0000-0002-2492-2689>; Jian Ren.

<sup>†</sup>Contributed equally.

## Abstract

**Background:** Large-scale genome sequencing projects have identified many genetic variants for diverse diseases. A major goal of these projects is to characterize these genetic variants to provide insight into their function and roles in diseases. N6-methyladenosine (m<sup>6</sup>A) is one of the most abundant RNA modifications in eukaryotes. Recent studies have revealed that aberrant m<sup>6</sup>A modifications are involved in many diseases. **Findings:** In this study, we present a user-friendly web server called “m6ASNP” that is dedicated to the identification of genetic variants that target m<sup>6</sup>A modification sites. A random forest model was implemented in m6ASNP to predict whether the methylation status of an m<sup>6</sup>A site is altered by the variants that surround the site. In m6ASNP, genetic variants in a standard variant call format (VCF) are accepted as the input data, and the output includes an interactive table that contains the genetic variants annotated by m<sup>6</sup>A function. In addition, statistical diagrams and a genome browser are provided to visualize the characteristics and to annotate the genetic variants. **Conclusions:** We believe that m6ASNP is a very convenient tool that can be used to boost further functional studies investigating genetic variants. The web server “m6ASNP” is implemented in JAVA and PHP and is freely available at [60].

**Keywords:** N6-methyladenosine (m<sup>6</sup>A); variant annotation; variant effect prediction; random forest

## Introduction

Due to rapid improvements in high-throughput sequencing technology, the cost and time requirements of these technologies have been greatly reduced, which has triggered the explosive growth of high-throughput sequencing data associated with various diseases. The major goal of these high-throughput sequencing studies is to identify disease-causing variants. However, distinguishing the few disease-causing variants from the majority of passenger variants remains a major challenge. Com-

putational methods that accurately interpret and prioritize the large amount of variants are urgently needed.

Many types of variants have different effects on the function of genes. Nonsynonymous variants, which alter the amino acids in a protein sequence, are among the most studied classes of variants. Alterations in the protein sequence can cause protein dysfunction due to a variety of different mechanisms. For example, variants in critical sites of the catalytic domain may affect protein catalytic functions [1] and variants in amino acids

**Received:** 20 December 2017; **Revised:** 7 February 2018; **Accepted:** 22 March 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

critical to the protein structure may affect protein-protein interactions [2], protein stability [3], and other important features [4]. Moreover, certain amino acid changes can affect post-translational modification, such as phosphorylation [5, 6], lysine modification [7], and glycosylation [8]. Currently, most bioinformatics tools mainly focus on interpreting nonsynonymous variants. For example, SIFT [9] and PolyPhen-2 [10] can predict the tolerance of nonsynonymous variants through sequence conservation; several tools, such as PhosphoSNP [11] and MIMP [12], predict whether amino acid changes affect post-translational modifications.

Compared to nonsynonymous variants, synonymous variants are neglected by most studies investigating diseases, particularly studies investigating tumors [13]. These variants are understudied because they do not alter the amino acid sequence of a protein and are considered "silent" variants. These variants are treated as "neutral" variants in evolutionary studies. However, growing evidence suggests that synonymous variants also affect the function of genes and cause various diseases [14]. Synonymous variants can result in abnormal post-transcriptional regulation, such as mRNA splicing [15], stability [16], and translation speed [17]. Many studies have shown that abnormalities in post-transcriptional regulation are closely related to genetic diseases and complex diseases [18–20]. Several bioinformatics tools that predict the effect of variants on post-transcriptional regulation are available, such as MutPred Splice [21] and SILVA [22], which primarily focus on mRNA splicing.

The post-transcriptional modification of mRNA is also an important post-transcriptional regulatory mechanism, and N6-methyladenosine (m<sup>6</sup>A) modification is among the most highly abundant in post-transcriptional modification [23], which regulates the metabolic processes of most RNA, including the splicing [24], stability [25], and translation of mRNA [26]. m<sup>6</sup>A modification is closely related to multiple diseases. Recently, FTO, an m<sup>6</sup>A demethylase, has been found to play an important role in the development of recessive lethality syndrome [27]. Abnormal m<sup>6</sup>A regulation can lead to individual developmental retardation [28], head malformations [27], mental retardation [29], brain dysfunction [30], and cardiac malformations [31]. More recently, increasing evidence has shown that dysregulation of m<sup>6</sup>A modification was closely related to cancer development. It was shown that abnormal modification of m<sup>6</sup>A and its regulators can lead to leukemia [32], prostate cancer [33], breast cancer [34, 35], bladder cancer [36], and liver cancer [37]. Therefore, it is important to evaluate the effect of variants on m<sup>6</sup>A modification, providing new perspective for understanding the variants, particularly for synonymous variants, thus helping to find more disease-causing variants.

A number of bioinformatics tools have been developed for predicting m<sup>6</sup>A sites, most of which are based on sequence characteristics. IRNA-methyl [38] and pRNAm-PC [39] used a support vector machine to construct a prediction model based on the distribution sequence characteristics. SRAMP [40] is a random forest-based tool trained on the single-nucleotide resolution m<sup>6</sup>A sites from miCLIP-Seq experiments [41, 42]. However, these tools are not specifically designed to deal with the variant data to evaluate the effects of the variants on m<sup>6</sup>A modification. It is highly desirable to develop a specific tool for predicting the effects of variant on m<sup>6</sup>A modification.

Here, we developed an accurate m<sup>6</sup>A site prediction tool that is superior to other similar tools. Based on the m<sup>6</sup>A site prediction tool, we constructed a web server called "m<sup>6</sup>ASNP" that is dedicated to predicting if methylation status of an m<sup>6</sup>A site is

altered by variants around the site. We then applied m<sup>6</sup>ASNP to the variants collected from dbSNP.

## Data collection

To construct the prediction model, we first obtained the single-base-resolution m<sup>6</sup>A sites from two recently published miCLIP experiments. We collected 16,079 human m<sup>6</sup>A sites from Linder et al. [41] and 43,155 human m<sup>6</sup>A sites from Ke et al. [42]. Specifically, in Ke's paper, two tissue samples from mouse are also tested, from which we collected 8,748 and 30,078 N6-methyladenosines in liver and brain, respectively. We then combined these datasets to obtain a nonredundant dataset that contains 55,548 sites in human and 36,192 sites in mouse. For the human model, we used 35,871 nonredundant m<sup>6</sup>A sites as the positive training set, and the remaining 19,677 m<sup>6</sup>A sites were used as the positive test set. Similarly, for the mouse model, 25,334 m<sup>6</sup>A sites were preserved as the positive training set, and another 10,858 m<sup>6</sup>A sites were used as the positive test set. The negative datasets were generated according to the distribution of the positive sets. Because the majority of m<sup>6</sup>A sites conformed to a DRACH motif, we first defined the potential m<sup>6</sup>A sites as adenine sites that conform to the AC motif. Using the positive datasets as references, we extracted the nonmethylated adenines that were followed by a cytosine in the same exon as the negative dataset. From the human genome, we extracted 1904,016 adenine sites as the negative training set, while the negative test set consisted of 1,286,588 adenine sites. In the case of the mouse genome, 1,519,570 adenine sites were extracted as the negative training set and 625,600 adenine sites were constructed as the negative test set (Supplementary Data).

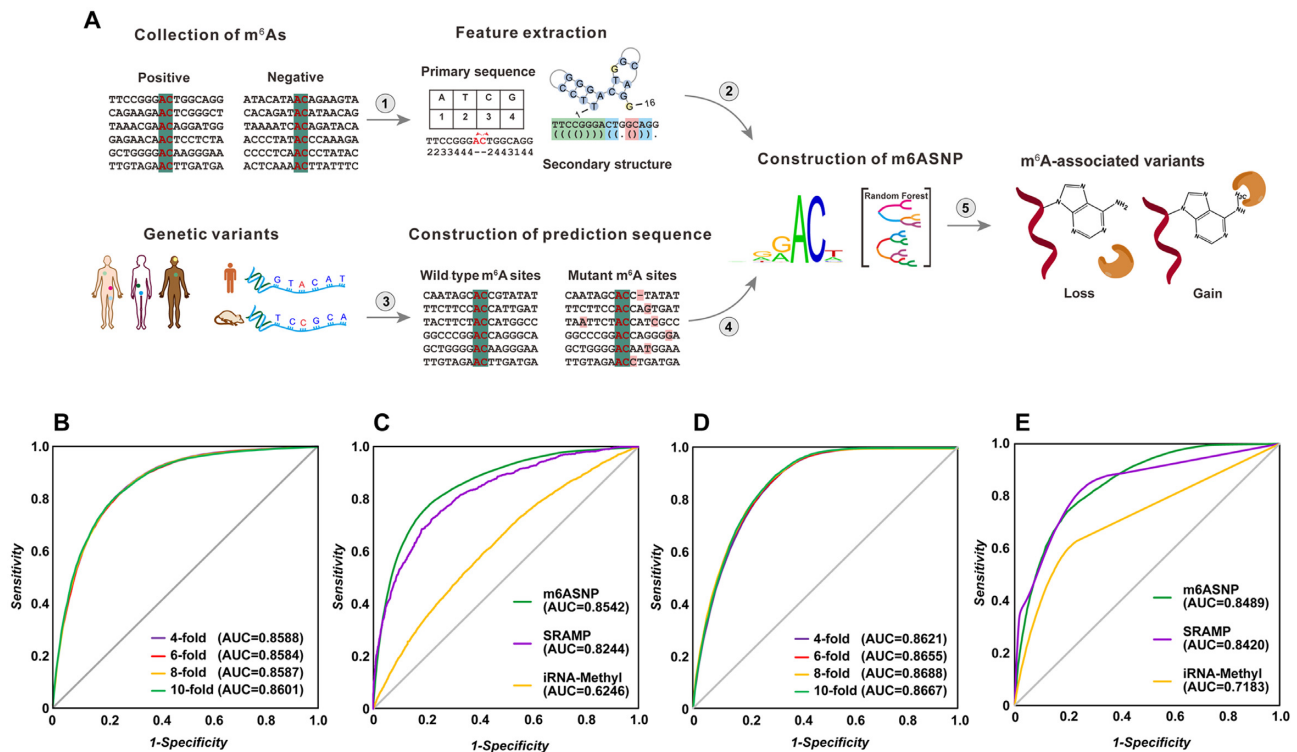
To decipher the potential applications of m<sup>6</sup>ASNP, we further collected a complete set of genetic variants from dbSNP for human and mouse. The single-nucleotide variations (SNVs) within the exonic regions were preserved for subsequent analysis. A total of 13,079,416 and 2,668,046 SNVs were collected in human and mouse, respectively. To investigate the potential role of these SNVs in reshaping the m<sup>6</sup>A event, m<sup>6</sup>A sites from two miCLIP-seq studies [41, 42], two PA-m<sup>6</sup>A-seq experiments [43], and 244 MeRIP-seq samples were integrated. Using m<sup>6</sup>ASNP, we also predicted the potential m<sup>6</sup>A-associated variants from the above dataset. In addition, a transcriptome-wide prediction was also performed. Overall, 311,706 and 40,308 m<sup>6</sup>A-associated variants were obtained from human and mouse, respectively.

In order to identify the potential roles of m<sup>6</sup>A-associated variants in post-transcriptome regulation, the RNA-binding protein (RBP) binding sites from starBase2 [44] and CLIPdb [45], the miRNA-RNA interactions from starBase2, and the canonical splice sites (GT-AG) from Ensembl annotations were collected. In addition, we also obtained a large number of disease-associated single-nucleotide polymorphism (SNPs) from different datasets (GWAS catalog [46], Johnson and O'Donnel [47], dbGAP [48], GAD [49], and ClinVar [50]) to perform disease-association analysis.

## Results

### Construction of m<sup>6</sup>ASNP

As illustrated in Fig. 1A, m<sup>6</sup>ASNP was developed using a random forest algorithm (see Methods section for more details). In order to evaluate the contribution of different encoding features, we first computed the mean decrease of Gini impurity (also known as Gini importance) for the human and mouse model. The distribution plot of Gini importance in different features showed



**Figure 1:** The construction of m6ASNP. A) The computational pipeline for identifying m<sup>6</sup>A-associated variants. (1) The single-nucleotide-resolution data were collected from recently published mCLIP-seq experiments. (2) The primary sequence and secondary structure features were extracted for subsequent model training process. (3) Genetic variants, such as somatic variants or germline SNPs, were input into the computation pipeline. (4) The flanking sequence around the potential m<sup>6</sup>A residue was constructed for both wild-type and mutant samples based on the inputted variants. (5) The loss and gain variants were predicted according to the above data. B) On the human model, 4-, 6-, 8-, and 10-fold cross-validation was performed. C) The performance comparison was made between m6ASNP and other state-of-the-art tools on the human test set. D) The evaluation results of 4-, 6-, 8-, and 10-fold cross-validation in mouse model. E) The performance comparison between m6ASNP and other state-of-the-art tools on the mouse test set.

that the primary sequence was the most effective feature for predicting potential m<sup>6</sup>A sites. Nucleotides in the DRACH motif around the N6-methyladenosine were dominated for classification (Supplementary Fig. S1A). However, secondary structures were still observed to contribute the prediction of m<sup>6</sup>A sites. Further evaluation on the prediction capability of primary sequence and secondary structure indicated that the addition of structural features to the sequence features can improve the accuracy and robustness of both models (Supplementary Fig. S1B). Therefore, in the final model of both human and mouse, we combined those features to obtain a better performance. Next, to evaluate the performance of m6ASNP, 4-, 6-, 8-, and 10-fold cross-validations were performed on both the human and mouse models. In both species, the area under the curves of all the validations were close and larger than 0.84 (Fig. 1B and D), indicating that m6ASNP is an accurate and robust predictor. To further assess the prediction capability in unknown data, we then compared m6ASNP with the two other publicly available predictors, iRNA-Methyl and SRAMP, in the independent test set. As a result, the performance of m6ASNP was found to be superior to all other predictors in both the human and mouse models (Fig. 1C and E).

To balance the prediction accuracy, we selected three thresholds with high, medium, and low stringencies for classification based on the evaluation result from 10-fold cross-validation. The high, medium, and low thresholds were selected by controlling the false-positive rate at 0.05, 0.1, and 0.15, respectively. Table 1 presents the detailed performance under these three selected

thresholds. In general, the high threshold provides the most stringent criterion and is usually used in large-scale prediction. The medium threshold is a balanced criterion and may be appropriate for most cases. The low threshold is the loosest criterion. When users expect to retain as many potential sites as possible, this threshold would be the best option.

### Usage of m6ASNP

In m6ASNP, a standard variant call format (VCF) or a simplified tab delimited file are supported as input data (Fig. 2A). As an example, we applied m6ASNP to the “common and clinical” variants VCF file obtained from ClinVar that contains 7,397 variants. The predicted m<sup>6</sup>A-associated variants are presented in an interactive table (Fig. 2B). Of the 7,397 variants, 206 are predicted to affect the m<sup>6</sup>A modification, either functional gain or loss of modification. The web server will conduct a comprehensive annotation and statistical analysis for all the predicted m<sup>6</sup>A-associated variants. The m<sup>6</sup>A-associated variants from ClinVar are mainly enriched in enzyme-binding and DNA-binding gene ontology (GO) molecular functions (Fig. 2C). The sequence logos are presented to show the changes of gained and lost m<sup>6</sup>A sites between the reference and mutant sequences (Fig. 2D). The “GGACU” motif is more obvious in mutant sequences compared to reference sequences for functional gain variants. While for functional loss variants, the “GGACU” motif is less noticeable in mutant sequences. A circos plot is presented to have an overview of all the m<sup>6</sup>A-associated variants (Fig. 2E).

**Table 1:** Prediction performance from 10-fold cross-validation under high, medium, and low thresholds

Threshold	Human					Mouse				
	Ac	Sn	Sp	MCC	Pr	Ac	Sn	Sp	MCC	Pr
High	0.7235	0.2781	0.9461	0.3158	0.7208	0.7154	0.2477	0.9492	0.2894	0.7092
Medium	0.7487	0.4497	0.8981	0.3973	0.6882	0.7465	0.4467	0.8964	0.3918	0.6832
Low	0.7589	0.5837	0.8465	0.4439	0.6554	0.7591	0.5956	0.8409	0.4471	0.6518

### Characteristics of m<sup>6</sup>A-associated variants predicted by m6ASNP

We further applied m6ASNP to all the variants in dbSNP. As a result, we obtained 133,394 functional gain and 214,884 functional loss m<sup>6</sup>A-associated variants. Among these m<sup>6</sup>A-associated variants, 6,235 located at or near the m<sup>6</sup>A sites from miCLIP experiments and 55,381 located at or near the m<sup>6</sup>A sites from MeRIP-Seq experiments. To characterize m<sup>6</sup>A-associated variants predicted by m6ASNP, we performed a systematic comparison between m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A-associated variants (non-m<sup>6</sup>A variants). We found that m<sup>6</sup>A-associated variants were enriched in protein-coding genes (dbSNP147, 95.77%; dbSNP146, 92.12%) and significantly concentrated in CDS and 3'UTR (Supplementary Fig. S2A and Table S1). Interestingly, in both CDS and UTR regions, m<sup>6</sup>A-associated variants were more conserved than non-m<sup>6</sup>A variants (Fig. 3A). For those conserved m<sup>6</sup>A-associated variants, a significant portion was synonymous compared to all conserved variants (Fig. 3B,  $P < 0.0001$ , hypergeometric test). To further explain the functional role of m<sup>6</sup>A-associated variants, we divided the predicted m<sup>6</sup>A-associated variants into two groups: the functional gain and functional loss variants. The conservation analysis was performed on these two groups, and the results were compared to non-m<sup>6</sup>A variants in both CDS and UTR regions (Supplementary Fig. S3A). Strikingly, in most cases, the functional loss variants were found to be more conservative compared to the gain variants, suggesting that the loss of existing m<sup>6</sup>A sites may undergo stronger selective pressure than the gain mutations on potential adenylate sites. Moreover, m<sup>6</sup>A-associated variants were predicted to be more deleterious than non-m<sup>6</sup>A variants in both the CDS and UTR regions (Fig. 3C, 2-tailed population test). Again, for the predicted data, the functional loss variants appeared to have a higher deleteriousness compared to the functional gain variants and the non-m<sup>6</sup>A variants (Supplementary Fig. S3B). Taken together, we conclude that m<sup>6</sup>A-associated variants, especially the functional loss variants, may have important roles and could be driven by positive selection in mammalian genomes. Furthermore, there were more m<sup>6</sup>A-associated variants located near the splice sites relative to the non-m<sup>6</sup>A variants, mostly distributed in the 20–30 bp flanking region of the splicing sites, implying that the variants were likely to affect RNA splicing as the means of changing the m<sup>6</sup>A levels (Fig. 3D). Moreover, the m<sup>6</sup>A-associated variants preferentially locate in genes with multiple transcripts (Supplementary Fig. S2B). These results were in agreement with the findings reported by Xiao et al. [24].

### m<sup>6</sup>A-associated variants in disease

Genome-wide association studies (GWAS) have revealed many disease-related variants. However, the pathogenesis mechanisms for most of these disease-related variants had not been known. We found 1,919 m<sup>6</sup>A-associated variants from human dbSNP were recorded either in GWAS studies or the ClinVar

database. These 1,919 m<sup>6</sup>A-associated variants were related to various diseases, including cardiovascular phenotype, muscular dystrophy, tuberous sclerosis syndrome, and cancer. Of them, hereditary cancer (436 variants, 22.74%,  $P = 2.27 \times 10^{-30}$ , Chi-squared test), Familial breast cancer (96 variants, 5.01%;  $P = 8.33 \times 10^{-9}$ , Chi-squared test) and hereditary nonpolyposis colorectal cancer (73 variants, 3.81%;  $P = 5.5 \times 10^{-5}$ , Chi-squared test) were the top enriched disease types (Supplementary Table S2). Our findings provide insights into the potential pathogenesis mechanism for many disease-related variants whose functions were not clear before.

Synonymous variants have been neglected in most previous studies of disease. Since m6ASNP can be used to predict the effect of both nonsynonymous and synonymous variants, this tool could significantly supplement the function of current annotating tools that mainly focus on nonsynonymous variants. Indeed, among the m<sup>6</sup>A-associated variants predicted by m6ASNP, 59.86% and 25.67% are synonymous variants in mouse dbSNP and human dbSNP, respectively. By using m6ASNP, we identified many m<sup>6</sup>A-associated synonymous variants that have been shown to be disease related. For instance, rs139362268, a synonymous variant of PALB2, is related to breast cancer and pancreatic cancer. Interestingly, we observed that rs139362268 occurred in the m<sup>6</sup>A site of PALB2, in which m<sup>6</sup>A peaks were detected in six MeRIP-Seq experiments (Supplementary Fig. S4A). We speculated that the cancer-related synonymous variant rs139362268 might be functional through dysregulation of m<sup>6</sup>A modification.

### m<sup>6</sup>A-associated variants in post-transcriptional regulation

It has been reported that m<sup>6</sup>A sites could recruit RBPs that play critical roles in post-transcriptional regulations [52]. We systematically examined the genomic positional relationship between m<sup>6</sup>A-associated variants and RBPs to determine whether m<sup>6</sup>A-associated variants function through RBPs. We found the m<sup>6</sup>A-associated variants were significantly enriched in RBP-binding regions compared to the non-m<sup>6</sup>A variants (Supplementary Fig. S4B). More than 50% of the human m<sup>6</sup>A-associated variants were located within RBP-binding regions. We found 19 RBPs were significantly overlapped with the regions having m<sup>6</sup>A-associated variants (Supplementary Table S3). As expected, the m<sup>6</sup>A reader YTHDF2 and m<sup>6</sup>A eraser ALKBH5 were significantly overlapped with the regions having m<sup>6</sup>A-associated variants compared to the randomly selected regions. Moreover, GO annotations demonstrated that these RBPs are enriched in RNA splicing, RNA translation, and miRNA regulation (Supplementary Table S3). Among them, SFRS1, a known splicing factor, is reportedly involved in alternative splicing and is co-localized with ALKBH5 in a demethylation-dependent manner, suggesting it might participate in the regulation of RNA methylation [53].

It has been reported that m<sup>6</sup>A sites are enriched in miRNA target sites and regulated by miRNAs [54]. Consistent with this,

A

**Data input** paste or upload

```
##fileformat=VCFv4.2
##fileDate=20161202
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
Sample0
17 10584116 . C T 50 PASS . GT 1/1
16 89350178 . G A 50 PASS . GT 1/1
17 78302157 . C A 50 PASS . GT 1/1
17 41133071 . T C 50 PASS . GT 1/1
1 145586403 . G T 50 PASS . GT 1/1
10 13653653 . A G 50 PASS . GT 1/1
3 194361768 . C T 50 PASS . GT 1/1
```

or select local files to upload.

0%

**+ Select Files**

Examples:

◀ VCF example ▶ Tab example

\* VCF or Tab format supported  
\* Paste file size < 500 KB

**Clear**

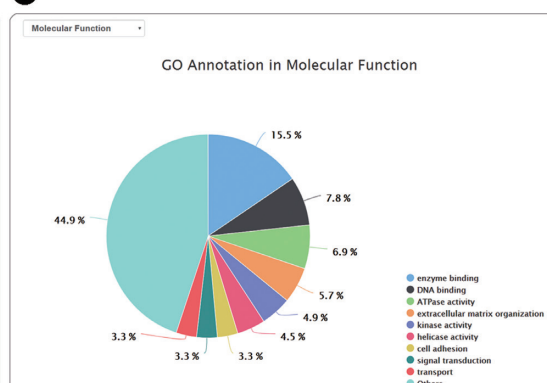
B

**Download** Search

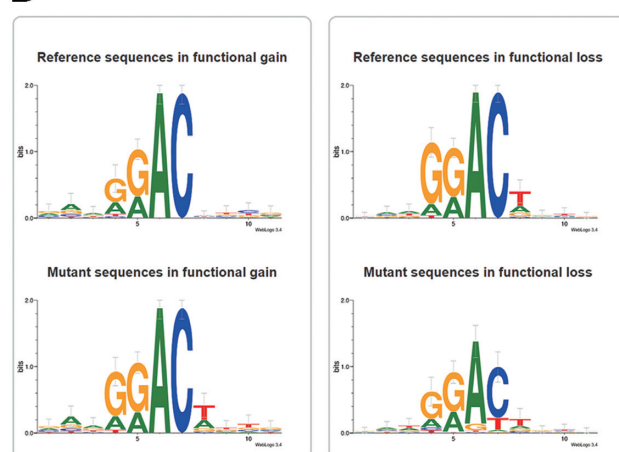
Gene Symbol	Position	Reference Sequence	Score	Mutation Event	Related Mutations
PKP2	chr12: 33049589-	Ref: TCTGGGCGAGAGATCTGGGCACTGACAGCTCCAGCTGGCGCTGCCCTCCGAGGC Mut: TCTGGGCGAGAGATCTGGGCACTGACAGCTCCAGCTGGCGCTGCCCTCCGAGGC	-0.185	Functional Loss	<a href="#">Details</a>
PRKAG2	chr7: 151257642-	Ref: ATATATATGTTGGGATATATTTCCCTGTCGAGATCTGCAAGCTGATCTCAGCAGAC Mut: ATATATATGTTGGGATATATTTCCCTGTCGAGATCTGCAAGCTGATCTCAGCAGAC	0.021	Functional Gain	<a href="#">Details</a>
ART4	chr12: 14993438-	Ref: TATAAATATGAGCTACACCCAGAGGAACTGGTTGAGTTGAGTCACTGGGAACTCT Mut: TATAAATATGAGCTACACCCAGAGGAACTGGTTGAGTTGAGTCACTGGGAACTCT	-0.200	Functional Loss	<a href="#">Details</a>
DNAH5	chr5: 13883173-	Ref: ACAGTAACAGTGCCTCTACATGAGCAGACAGTTTGGCCATTTTCGGGCAAGGTCAC Mut: ACAGTAACAGTGCCTCTACATGAGCAGACAGTTTGGCCATTTTCGGGCAAGGTCAC	-0.021	Functional Loss	<a href="#">Details</a>
PRICKLE1	chr12: 42853873-	Ref: CCGATGCCACTTCCGATATAGGCTGAGACCCAGGAATGAATCGTTTCTGGGACTCTA Mut: CCGATGCCACTTCCGATATAGGCTGAGACCCAGGAATGAATCGTTTCTGGGACTCTA	0.562	Functional Gain	<a href="#">Details</a>
LRRK2	chr12: 40702990+	Ref: GTACCTTGCCTGCTATGACCTCAGCAGAGGAGCGGCTGAGTTGATGACATGAGGCTTGG Mut: GTACCTTGCCTGCTATGACCTCAGCAGAGGAGCGGCTGAGTTGATGACATGAGGCTTGG	0.028	Functional Gain	<a href="#">Details</a>
CYP11B1	chr2: 38298397-	Ref: AGGCAGAAATTGGATCAGGTCGTGGGGAGGGACCGTCTGCTTGTATGGGTGACAGCCCAA Mut: AGGCAGAAATTGGATCAGGTCGTGGGGAGGGACCGTCTGCTTGTATGGGTGACAGCCCAA	-0.100	Functional Loss	<a href="#">Details</a>

Showing 1 to 10 of 206 rows 10 rows per page

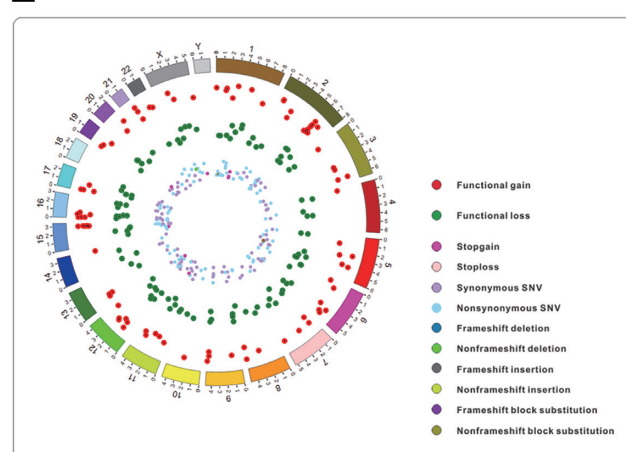
C



D



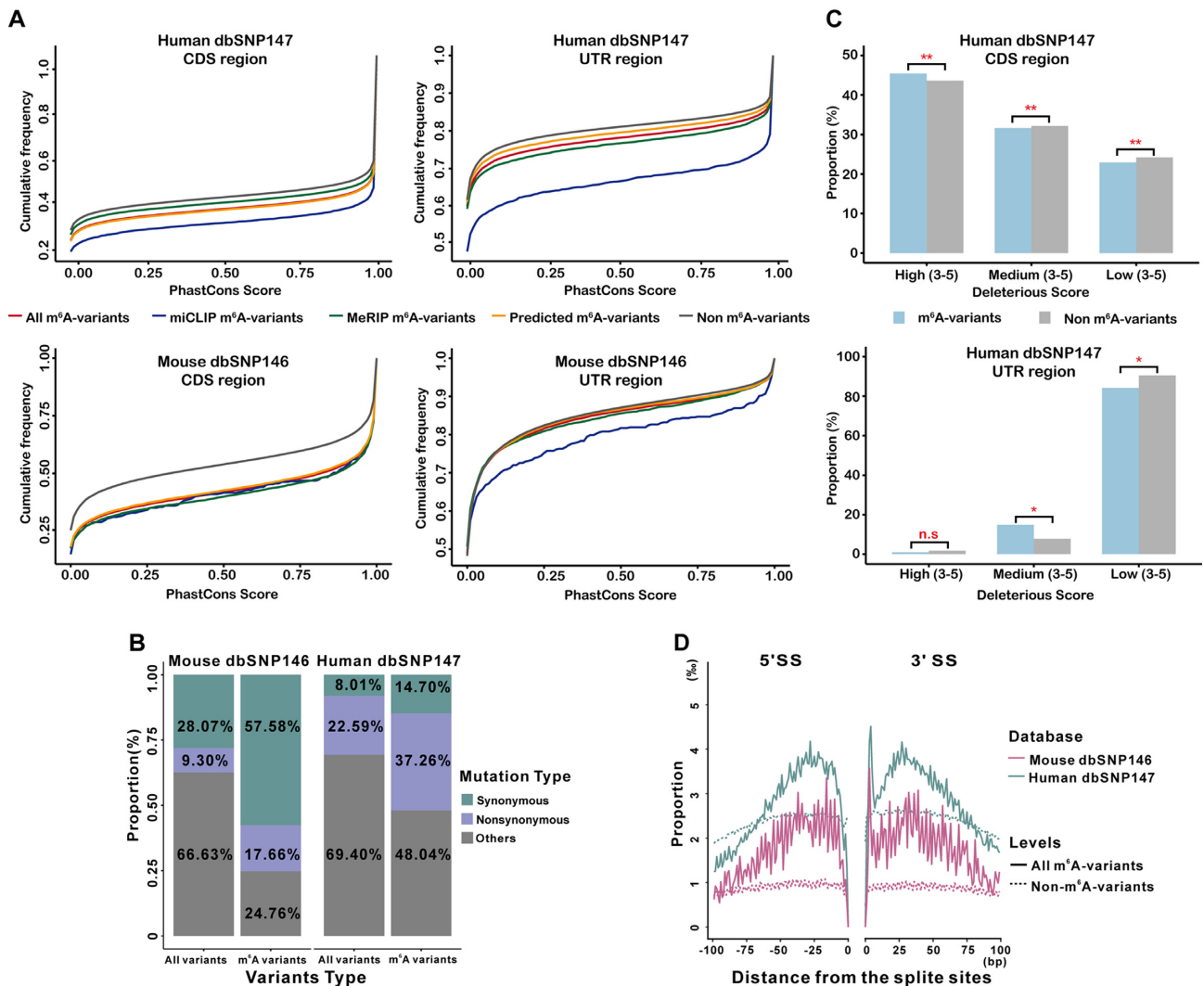
E



**Figure 2:** A snapshot of the m6ASNP web server. A) The main interface. Variants can be input as standard VCF format or tab-delimited flat format. A file uploading module was implemented to support large-scale prediction of m<sup>6</sup>A-associated variants. B) The prediction results were listed in the interactive table, allowing fast retrieval of the result data. C) The gene ontology annotation was performed on the predicted m<sup>6</sup>A-associated variants. D) To present the alterations of the m<sup>6</sup>A motif, the sequence logos were generated automatically for both functional gain and loss variants. E) The gain and loss m<sup>6</sup>A-associated variants, as well as the original SNPs, were illustrated in the circo plot at a genomic level by the BioCircos [51] library.

we found m<sup>6</sup>A-associated variants predicted by m6ASNP occurred significantly more frequently in miRNA target sites than the non-m<sup>6</sup>A variants (Supplementary Fig. S4C). The miRNAs with a significant number of m<sup>6</sup>A-associated variants are listed in Supplementary Table S4. Among them, miR-132-3p and miR-

212-3p were mainly expressed in the brain and played critical roles in neuronal functions as well as circadian clock entrainment [55], which is consistent with m<sup>6</sup>A function [56]. Interestingly, m<sup>6</sup>A-associated variants related to miR-132-3p and miR-



**Figure 3:** Characteristics of m<sup>6</sup>A-associated variants predicted by m6ASNP. A) The cumulative distribution function of phastCons score for different levels of m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants in mouse dbSNP and human dbSNP. B) Proportional distribution of different variant types for the conserved m<sup>6</sup>A-associated variants. C) Proportional distribution of the m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants at three deleterious levels predicted by a combination of five variant function predictors. A 2-tailed test of the population proportion was used to assess significance. D) Proportional distribution of m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants at different distances from the splicing sites.

212-3p were identified in both human and mouse, suggesting a conservation of function in these variants.

## Discussion

There is growing evidence that aberrant m<sup>6</sup>A modification is a potential pathogenesis mechanism in many diseases including cancer, which suggests the variants that disrupt m<sup>6</sup>A modification might cause diseases. However, currently there is still a lack of methodology for annotating variants from high-throughput sequencing studies by m<sup>6</sup>A function. To address this, we developed a novel computation model, m6ASNP, that is dedicated to predicting the variants that disrupt m<sup>6</sup>A modification. Using m6ASNP, we performed further functional analysis on m<sup>6</sup>A-associated variants. By integrating dataset regarding RBP-binding regions, miRNA-targets and splicing sites, m6ASNP can help to reveal the potential relationship among variants, m<sup>6</sup>A modification, and other post-transcriptional regulation. Also, in the disease-association analysis, more than 2,000 disease-

related variants that may be linked with alterations of m<sup>6</sup>A modification were identified. This finding further proves that m6ASNP is a promising tool for studying the potential role of m<sup>6</sup>A variants in clinical investigation.

In conclusion, m6ASNP is a useful computational web server for annotating variants by m<sup>6</sup>A function. m6ASNP will serve as a supplemental method to run in parallel with other annotating tools to comprehensively predict the function of the variants, for both synonymous and nonsynonymous, in the high-throughput sequencing studies of diseases.

## Methods

### Construction of m<sup>6</sup>A site prediction model

The sequences of the flanking regions 30 nucleotides upstream and downstream of a given m<sup>6</sup>A residue were extracted. To transform the primary sequences to numeric vectors, each nucleotide was encoded by four distinct variables. In total, 60 nu-

meric variables were generated for a single m<sup>6</sup>A residue. As reported in recent studies [57, 58], specific RNA secondary structures around the potential adenosines can affect the enzymatic process of RNA methylation. We therefore added secondary structure features to our prediction model. Using the Nussinov algorithm [59], we first predicted the secondary structure for each m<sup>6</sup>A residue and marked the structure state (paired or not paired) with a bracket or dot. For example, a given m<sup>6</sup>A nucleotide with the sequence TTCGGGACTGGCAGG could be represented as (((()))((.))). Next, we extracted the secondary structure triplet, formed by the structure state of the three adjacent nucleotides obtained from the predicted RNA structure. The number of occurrences of each triplet in the sequence was counted and normalized to produce a 27-dimension feature vector. Combining all the primary sequences and secondary structure features, we constructed an 87-dimension vector for each m<sup>6</sup>A residue. These vectors were subsequently used as the input for a random forest classifier for training and prediction.

The random forest classifier for human and mouse were trained separately on the above-collected training set. The tree number was optimized as 500 and the features used for each splitting were set to 9. To assess the performance, we used 4-, 6-, 8-, and 10-fold cross-validation on the training set. The additional test set was also applied in our study to evaluate the robustness. The sensitivity, specificity, and Matthew's correlation coefficient were used to measure the predictor's performance.

### Construction of m<sup>6</sup>ASNP

Based on the m<sup>6</sup>A site prediction model, we then developed a computational pipeline to predict the effect of variants on m<sup>6</sup>A modification. First, variants were mapped to known transcripts. The wild-type and mutant form of the transcript sequences were then generated for m<sup>6</sup>A site prediction. For an m<sup>6</sup>A site that occurred in the wild-type transcript and disrupted in the mutant transcript, we defined it as an m<sup>6</sup>A-associated loss variant. The m<sup>6</sup>A-associated gain variant was conversely formed. To measure the altered degree of m<sup>6</sup>A modifications, equation 1 was defined as follows:

$$S = \ln \left( \frac{RF\_Score_{wild-type}}{RF\_Score_{mutant}} \right) \quad (1)$$

where S denotes the alteration score that quantitatively represented the degree of m<sup>6</sup>A alterations between reference and mutant samples and RF\_Score is the predicted score of a given m<sup>6</sup>A site from the random forest model. Obviously, alteration scores higher than 0 represent m<sup>6</sup>A-gain alterations, while scores lower than 0 represent m<sup>6</sup>A-loss alterations. In some m<sup>6</sup>A-associated loss variants, alteration scores were assigned to MAX, which means that the core AC motif is destroyed by genetic variants, leading to complete losses of m<sup>6</sup>A at those sites.

To provide convenience to the research community, we developed a web server called "m6ASNP" to specifically predict the effect of variants on m<sup>6</sup>A modification. m6ASNP was implemented using JAVA and PHP and is freely accessible at [60].

### Derivation of the m<sup>6</sup>A-associated variants

Based on miCLIP-seq, PA-m<sup>6</sup>A-seq, and MeRIP-seq data, we then combined them with the SNV data from dbSNP and performed m<sup>6</sup>A-association prediction using m6ASNP. Following the same procedure proposed in our previously published work [61], we

constructed three confidence levels of annotations of m<sup>6</sup>A-associated variants for subsequent analysis.

The first annotation was the high-confidence-level data that contained the m<sup>6</sup>A-associated variants derived from miCLIP-seq and PA-m<sup>6</sup>A-seq experiments. Notably, the PA-m<sup>6</sup>A-seq can only detect m<sup>6</sup>A signal in a resolution of ~23 nt. Therefore, in order to obtain precise modification sites, we scanned through all the peak regions and extracted adenosine sites that conformed to DRACH motif as the final m<sup>6</sup>A sites. On this basis, we retained the variants that located near the m<sup>6</sup>A sites as the m<sup>6</sup>A-associated variants.

The second annotation was the medium-confidence-level data. We first downloaded all the published MeRIP-seq data from the GEO database. According to the standard analysis pipeline for MeRIP-seq data, we applied MACS2 [62], MeTPeak [63], and Meyer's method [64] to identify the m<sup>6</sup>A peaks in each study separately. In general, in MeRIP-seq experiments, if a given region is identified as enriched in most of the adopted methods, it is more likely to be a true modification signal. Therefore, to obtain reliable m<sup>6</sup>A peaks, a tool called MSPC [65] was then applied to construct consensus peaks from the above three methods. In those consensus peaks, we then applied m6ASNP to predict m<sup>6</sup>A-associated variants that significantly change the DRACH motif.

The third annotation was the low-confidence-level data, where we used the whole transcriptome sequences for prediction. With a high threshold, m6ASNP will predict the potential m<sup>6</sup>A-associated variants from all collected genetic variants.

In summary, we constructed 13,703 high-confidence-level, 54,222 medium-confidence-level, and 243,880 low-confidence-level m<sup>6</sup>A-associated variants for human. Another 935 high-confidence-level, 9,404 medium-confidence-level, and 17,739 low-confidence-level data were also constructed for mouse.

### Annotation of m<sup>6</sup>A-associated variants

All the identified m<sup>6</sup>A-associated variants were annotated by the transcript structure, including the CDS, 3' UTR, 5' UTR, start codon, and stop codon. For the annotation of noncoding RNA DASHR [66], miRBase (version 21) (miRBase, [RRID:SCR.003152](#)) [67], GtRNAdb [68], and piRNABank [69] were used. To test whether the m<sup>6</sup>A-associated variants were more preferentially distributed in specific transcript structures, we calculated the proportion of variants that located in a given transcript structure. In order to avoid bias, only the variants that were annotated in mRNA were used, and the proportion in 5'-UTR, CDS, and 3'-UTR were calculated. A 2-tailed proportion test was then adopted to compare the proportion difference between m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants. In addition, in order to evaluate their conservation scores and deleteriousness, we further annotated the m<sup>6</sup>A-associated variants by ANNOVAR (updated 1 February 2016) (ANNOVAR, [RRID:SCR.012821](#)) [70]. To avoid any bias, we only preserved those variants located in mRNA for analysis and compared the conservative and deleterious differences between m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants in the same exon. As the selective pressures were quite different in protein-coding sequences and untranslated regions, the above comparison was carried out separately for the CDS and UTR regions. Specifically, the conservation scores were calculated by phastCons with 100-way and 60-way gene conservation profiles for the human and mouse, respectively [71]. The deleteriousness of each variants was measured by integrating the prediction results from five pieces of software (SIFT [72], PolyPhen2 HVAR [10], PolyPhen2 HDIV [10], LRT [73], and

FATHMM [74]). We defined an aggregate score by counting the number of above-listed methods that consider an SNV to be deleterious. A deleterious score of 0 means that the variant is predicted to be tolerated in all methods, while a deleterious score of 5 means that the corresponding variant is predicted to be deleterious in all five predictors. As a result, the aggregate score may range from 0 to 5, and a higher score indicates a higher probability of deleterious.

### Disease-association analysis

A linkage disequilibrium (LD) analysis was performed for each GWAS disease-associated SNP. We used Haploview (Haploview, [RRID:SCR\\_003076](https://www.broadinstitute.org/haploview)) to obtain the LD mutations using a parameter  $r^2 > 0.8$  in at least one of the four populations from CHB, CEU, JPT, and TSI. Then, we selected all m<sup>6</sup>A-associated variants by mapping the variants to GWAS disease-associated SNPs and their LD mutations. Moreover, we collected ClinVar data to annotate the m<sup>6</sup>A-associated variants with specific functions.

### Post-transcriptional regulation association analysis

First, the m<sup>6</sup>A-associated variants were intersected with the collected RBP regions for the same sample. We matched all m<sup>6</sup>A-associated variants with miRNA targets to obtain the m<sup>6</sup>A-associated variants that potentially impacted the miRNA-target interactions. Additionally, we extracted 100 bp upstream of the 5' splicing sites and 100 bp downstream of the 3' splicing sites. Subsequently, we matched the m<sup>6</sup>A-associated variants to these regions to obtain the splicing sites affected by the m<sup>6</sup>A-associated variants.

### Identification of significant RBPs and miRNAs

To determine whether the m<sup>6</sup>A-associated variants were significantly enriched in RBP regions, an empirical evaluation was performed for each RBP. Using YTHDF2 as an example, the process may be described as follows.

First, we calculated the number of m<sup>6</sup>A-associated variants within the YTHDF2-binding regions (defined as  $N_{RBP}$ ). Second, because certain m<sup>6</sup>A-associated variants randomly occur within the YTHDF2-binding regions, we estimated the background count of m<sup>6</sup>A-associated variants for YTHDF2 (defined as  $N_B$ ). Thus, we extracted the longest transcript for each gene from the gene annotation files. The weight of the  $i$ th gene was defined as follows:

$$w(i) = \frac{L(i)}{\sum_{i=0}^n L(i)} \quad (2)$$

$$\sum_{i=0}^n w(i) = 1 \quad (3)$$

where  $n$  was the total number of genes annotated and  $L(i)$  was the length (bp) of the  $i$ th gene. Then, we extracted the same-length reads of all YTHDF2-binding regions, which was defined as  $N_B$ , using weighted random sampling of all transcripts collected above. We repeated this procedure 50,000 times and then obtained the frequency  $F_{RBP}$  when  $N_B$  was greater than  $N_{RBP}$  in the cycle. This frequency may be regarded as an estimation of the probability that observing  $N_B$  greater than  $N_{RBP}$  in random condition. Next, the Benjamini-Hochberg method was applied

to control the false positives. An adjusted  $F_{RBP}$  less than 0.05 was considered a small probability event, suggesting that the m<sup>6</sup>A-associated variants were more likely to occur in the RBP-binding regions of YTHDF2. All significant RBPs are listed in Supplementary Table S2. Certain significant miRNAs, which are listed in Supplementary Table S3, were obtained by performing a similar analysis of miRNA targets.

### Availability of supporting source code and requirements

Project name: m6ASNP

Project home page: <https://m6asnp.renlab.org>  
<https://github.com/RenLabBioinformatics/m6ASNP>  
 RRID:SCR\_016048

Operating system(s): platform independent  
 Programming language: PHP, java, javascript  
 License: GPLv3

### Availability of supporting data

The training data and test data collected from Linder *et al.* and Ke *et al.* are available in the supplementary data. These and snapshots of the code are also available in the GigaScience GigaDB repository [75].

### Additional files

**Supplementary Figure S1:** The feature contribution of the human and mouse model. Distribution plot of the feature's Gini importance for both (A) human and (B) mouse model. The prediction capabilities of different combination of features for (C) human and (D) mouse model.

**Supplementary Figure S2:** A systematic comparison of the m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants. (A) Proportional distribution of the variants at different m<sup>6</sup>A confidence levels and non-m<sup>6</sup>A variants located in the CDS and 3' UTR. A two-tailed test of population proportion was performed to assess significance. (B) Boxplots show the gene isoforms of the m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants in different databases. One-sided Wilcoxon signed-rank test was performed to determine the significance. “\*\*” indicates a significance level of  $P \leq 0.01$ , while “\*” indicates  $P \leq 0.05$ .

**Supplementary Figure S3:** The characteristics of m6A-associated variants predicted by m6ASNP. (A) The conservation differences between functional gain and functional loss variants. (B) The comparison of mutation deleteriousness between functional gain and functional loss variants.

**Supplementary Figure S4:** Association analysis of m6A-associated variants. (A) An example of m<sup>6</sup>A-associated variants in disease. The red rectangle in exon 10 represents a synonymous mutation in PALB2, i.e., rs139362268, while the green rectangle represents the m<sup>6</sup>A site. The 1 to 6 numbering indicates the different samples, followed by HepG2, GM12878, Momo-mac-6, HeLa, shMETTL14 in A549 and shGFP in A549. MeRIP-seq peak tracks of input, and the IP samples were scaled to the same level and colored in red and blue. (B-C) Proportional distribution of different levels of m<sup>6</sup>A-associated variants and non-m<sup>6</sup>A variants located within the RBP-binding regions and miRNA target regions. A two-tailed test of population proportion was performed to assess significance. “\*\*” indicates a significance level of  $P \leq 0.01$ , while “\*” indicates a significance level of  $P \leq 0.05$ .

**Supplementary Table S1:** The distribution characteristics of m6A-associated variants in different transcript structures.

**Supplementary Table S2:** Significant disease phenotypes in m6A-associated.

**Supplementary Table S3:** Significant RBPs in m6A-associated variants Supplementary Table S3. Significant RBPs in m6A-associated variants.

**Supplementary Table S4:** Significant miRNAs in m6A-associated variants.

**Supplementary Data 1.** Single nucleotide resolution m6A sites (Training data, hg19).

## Abbreviations

GO, Gene ontology; GWAS, Genome-wide association study; LD, Linkage disequilibrium; m<sup>6</sup>A, N<sup>6</sup>-methyladenosine; RBP, RNA-binding protein; SNP, single nucleotide polymorphism; VCF, Variant call format.

## Ethics approval and consent to participate

Not applicable.

## Disclosure statement

The authors declare that they have no competing interests.

## Funding

This work was supported by grants from the National Key Research and Development Program [2017YFA0106700]; National Natural Science Foundation of China [31 771 462, 81 772 614, 31 471 252, 31 500 813, 91 753 137 and U1611261]; Guangdong Natural Science Foundation [2014TQ01R387, 2014A030313181 and 2017A030313134]; China Postdoctoral Science Foundation [2017M622864]; Fundamental Research Funds for the Central Universities [No. 17lgpy106].

## Authors' contributions

ZZ and JR conceived, designed, and supervised all phases of the project. YX and SJ developed the prediction model. YX and ZH designed and implemented the Web server. Y.L.Z., YZ, Y.Y.Z., LC and YM performed data analysis. ZZ, YX and JR wrote the manuscript. All authors read and approved the final manuscript.

## References

- Carvalho S, Catarino TA, Dias AM, et al. Preventing E-cadherin aberrant N-glycosylation at Asn-554 improves its critical function in gastric cancer. *Oncogene* 2016;**35**(13):1619–31.
- Gonfloni S, Williams JC, Hattula K, et al. The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src. *The EMBO J* 1997;**16**(24):7261–71.
- Selezneva AI, Walden WE, Volz KW. Nucleotide-specific recognition of iron-responsive elements by iron regulatory protein 1. *J Mol Biol* 2013;**425**(18):3301–10.
- Zhang B, Deng L, Qian Q, et al. A missense mutation in the transmembrane domain of CESA4 affects protein abundance in the plasma membrane and results in abnormal cell wall biosynthesis in rice. *Plant Molecular Biology* 2009;**71**(4–5):509–24.
- Heald R, McKeon F. Mutations of phosphorylation sites in lamin A that prevent nuclear lamina disassembly in mitosis. *Cell* 1990;**61**(4):579–89.
- Xu Y, Gray A, Hardie DG, et al. A novel, de novo mutation in the PRKAG2 gene: infantile-onset phenotype and the signaling pathway involved. *Am J Phy Heart and Circulatory Phy* 2017;**313**(2):H283–H92.
- McCabe MT, Graves AP, Ganji G, et al. Mutation of A677 in histone methyltransferase EZH2 in human B-cell lymphoma promotes hypertrimethylation of histone H3 on lysine 27 (H3K27). *Proc Nat Acad Sci USA* 2012;**109**(8):2989–94.
- Liu X, Gao J, Sun Y, et al. Mutation of N-linked glycosylation in EpCAM affected cell adhesion in breast cancer cells. *Biol Chem* 2017;**398**(10):1119–26.
- Sim NL, Kumar P, Hu J, et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 2012;**40**(Web Server Issue):W452–7.
- Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature Methods* 2010;**7**(4):248–9.
- Ren J, Jiang C, Gao X, et al. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Molecular & cellular proteomics: MCP* 2010;**9**(4):623–34.
- Wagih O, Reimand J, Bader GD. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nature Methods* 2015;**12**(6):531–3.
- Supek F, Minana B, Valcarcel J, et al. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 2014;**156**(6):1324–35.
- Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nature Rev Genetics* 2011;**12**(10):683–91.
- Parmley JL, Chamary JV, Hurst LD. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution* 2006;**23**(2):301–9.
- Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology* 2005;**6**(9):R75.
- Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 2008;**134**(2):341–52.
- Roundtree IA, Evans ME, Pan T, et al. Dynamic RNA modifications in gene expression regulation. *Cell* 2017;**169**(7):1187–200.
- Feigerlova E, Battaglia-Hsu SF. Role of post-transcriptional regulation of mRNA stability in renal pathophysiology: focus on chronic kidney disease. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 2017;**31**(2):457–68.
- Kiebler MA, Scheiffele P, Ule J. What, where, and when: the importance of post-transcriptional regulation in the brain. *Frontiers in neuroscience* 2013;**7**:192.
- Mort M, Sterne-Weiler T, Li B et al. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology* 2014;**15**(1):R19.
- Pruesse E, Quast C, Knittel K et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**(21):7188–96.
- Fu Y, Dominissini D, Rechavi G, et al. Gene expression regulation mediated through reversible m(6)A RNA methylation. *Nature Reviews Genetics* 2014;**15**(5):293–306.
- Xiao W, Adhikari S, Dahal U, et al. Nuclear m(6)A

- Reader YTHDC1 regulates mRNA splicing. *Molecular Cell* 2016;**61**(4):507–19.
25. Wang Y, Li Y, Toth JJ, et al. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology* 2014;**16**(2):191–8.
  26. Meyer KD, Patil DP, Zhou J, et al. 5' UTR m(6)A promotes Cap-independent translation. *Cell* 2015;**163**(4):999–1010.
  27. Boissel S, Reish O, Proulx K, et al. Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. *American Journal of Human Genetics* 2009;**85**(1):106–11.
  28. Daoud H, Zhang D, McMurray F, et al. Identification of a pathogenic FTO mutation by next-generation sequencing in a newborn with growth retardation and developmental delay. *Journal of Medical Genetics* 2016;**53**(3):200–7.
  29. Jonkhout N, Tran J, Smith MA, et al. The RNA modification landscape in human disease. *RNA (New York, NY)* 2017;**23**(12):1754–69.
  30. McGuinness DH, McGuinness D. m6a RNA methylation: the implications for health and disease. *Journal of Cancer Science and Clinical Oncology* 2014;**1**(1, 2394–6520).
  31. Fawcett KA, Barroso I. The genetics of obesity: FTO leads the way. *Trends in Genetics* 2010;**26**(6):266–74.
  32. Li Z, Weng H, Su R, et al. FTO plays an oncogenic role in acute myeloid leukemia as a N6-methyladenosine RNA demethylase. *Cancer Cell* 2017;**31**(1):127–41.
  33. Lewis SJ, Murad A, Chen L, et al. Associations between an obesity related genetic variant (FTO rs9939609) and prostate cancer risk. *PloS One* 2010;**5**(10):e13485.
  34. Zhang C, Samanta D, Lu H, et al. Hypoxia induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m(6)A-demethylation of NANOG mRNA. *Proc Natl Acad Sci U S A* 2016;**113**(14):E2047–56.
  35. Zhang C, Zhi W, Lu H, et al. Hypoxia-inducible factors regulate pluripotency factor expression by ZNF217- and ALKBH5-mediated modulation of RNA methylation in breast cancer cells. *Oncotarget* 2016;**7**(40):64527–42.
  36. Zhang Z, Zhang G, Kong C et al. METTL13 is downregulated in bladder carcinoma and suppresses cell proliferation, migration and invasion. *Scientific Reports* 2016;**6**:19261.
  37. Ma JZ, Yang F, Zhou CC et al. METTL14 suppresses the metastatic potential of hepatocellular carcinoma by modulating N6-methyladenosine-dependent primary MicroRNA processing. *Hepatology (Baltimore, Md)* 2017;**65**(2):529–43.
  38. Chen W, Feng P, Ding H et al. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Analytical Biochemistry* 2015;**490**:26–33.
  39. Liu Z, Xiao X, Yu DJ et al. pRNA-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physicochemical properties. *Analytical Biochemistry* 2016;**497**:60–7.
  40. Zhou Y, Zeng P, Li YH et al. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res* 2016;**44**(10):e91.
  41. Linder B, Grozhik AV, Olarerin-George AO et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature Methods* 2015;**12**(8):767–72.
  42. Ke S, Alemu EA, Mertens C et al. A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes & Development* 2015;**29**(19):2037–53.
  43. Chen K, Lu Z, Wang X et al. High-resolution N(6)-methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. *Angewandte Chemie (International ed in English)* 2015;**54**(5):1587–90.
  44. Li JH, Liu S, Zhou H et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014;**42**(Database issue):D92–7.
  45. Yang YC, Di C, Hu B et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* 2015;**16**:51.
  46. Welter D, MacArthur J, Morales J et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**(Database Issue):D1001–6.
  47. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. *BMC Medical Genetics* 2009;**10**:6.
  48. Mailman MD, Feolo M, Jin Y et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* 2007;**39**(10):1181–6.
  49. Becker KG, Barnes KC, Bright TJ et al. The genetic association database. *Nature Genetics* 2004;**36**(5):431–2.
  50. Landrum MJ, Lee JM, Benson M et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**(D1):D862–8.
  51. Cui Y, Chen X, Luo H et al. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics (Oxford, England)* 2016;**32**(11):1740–2.
  52. Liu J, Yue Y, Han D et al. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nature Chemical Biology* 2014;**10**(2):93–5.
  53. Zheng G, Dahl JA, Niu Y et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Molecular Cell* 2013;**49**(1):18–29.
  54. Chen T, Hao YJ, Zhang Y et al. m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell stem cell* 2015;**16**(3):289–301.
  55. Wanet A, Tacheny A, Arnould T et al. miR-212/132 expression and functions: within and beyond the neuronal compartment. *Nucleic Acids Res* 2012;**40**(11):4742–53.
  56. Fustin JM, Doi M, Yamaguchi Y et al. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* 2013;**155**(4):793–806.
  57. Roost C, Lynch SR, Batista PJ et al. Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *Journal of the American Chemical Society* 2015;**137**(5):2107–15.
  58. Cao G, Li HB, Yin Z et al. Recent advances in dynamic m6A RNA modification. *Open Biology* 2016;**6**(4):160003.
  59. Eddy SR. How do RNA folding algorithms work? *Nature Biotechnology* 2004;**22**(11):1457–8.
  60. m6ASNP website <http://m6asnp.renlab.org>.
  61. Zheng Y, Nie P, Peng D et al. m6AVar: a database of functional variants involved in m6A modification. *Nucleic Acids Res* 2018;**46**(D1):D139–D45.
  62. Zhang Y, Liu T, Meyer CA et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 2008;**9**(9):R137.
  63. Cui X, Meng J, Zhang S et al. A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data. *Bioinformatics (Oxford, England)* 2016;**32**(12):i378–i85.
  64. Meyer KD, Saletore Y, Zumbo P et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 2012;**149**(7):1635–46.
  65. Jalili V, Matteucci M, Masseroli M et al. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics (Oxford, England)* 2015;**31**(17):2761–9.
  66. Leung YY, Kuksa PP, Amlie-Wolf A et al. DASHR: database of small human noncoding RNAs. *Nucleic Acids Res*

- 2016;**44**(D1):D216–22.
67. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;**42**(Database Issue):D68–73.
  68. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* 2016;**44**(D1):D184–9.
  69. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res* 2008;**36**(Database Issue):D173–7.
  70. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**(16):e164.
  71. Siepel A, Bejerano G, Pedersen JS et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 2005;**15**(8):1034–50.
  72. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* 2009;**4**(7):1073–81.
  73. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome research* 2009;**19**(9):1553–61.
  74. Shihab HA, Gough J, Cooper DN et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation* 2013;**34**(1):57–65.
  75. Jiang S, Xie Y, He Z et al. Supporting data for “m6ASNP: a tool for annotating genetic variants by m6A function” Giga-Science Database 2018, <http://dx.doi.org/10.5524/100428>.