



RESEARCH PAPER

 OPEN ACCESS 

M6ADD: a comprehensive database of m⁶A modifications in diseases

Dianshuang Zhou^{a,†}, Hongli Wang^{a,†}, Fanqi Bi^{a,†}, Jie Xing^{a,†}, Yue Gu^a, Cong Wang^a, Menyan Zhang^a, Yan Huang^a, Jiaqi Zeng^a, Qiong Wu^a, and Yan Zhang^{a,b}

^aSchool of Life Science and Technology, Computational Biology Research Center, Harbin Institute of Technology, Harbin, Heilongjiang, China; ^bState Key Laboratory of Respiratory Disease, Guangzhou Medical University, Guangzhou, China

ABSTRACT

N⁶-methyladenosine (m⁶A) modification is an important regulatory factor affecting diseases, including multiple cancers and it is a developing direction for targeted disease therapy. Here, we present the M6ADD (m⁶A-diseases database) database, a public data resource containing manually curated data on potential m⁶A-disease associations for which some experimental evidence is available; the related high-throughput sequencing data are also provided and analysed by using different computational methods. To give researchers a tool to query the m⁶A modification data, the M6ADD was designed as a web-based comprehensive resource focusing on the collection, storage and online analysis of m⁶A modifications, aimed at exploring the associations between m⁶A modification and gene disorders and diseases. The M6ADD includes 222 experimentally confirmed m⁶A-disease associations, involving 59 diseases from a review of more than 2000 published papers. The M6ADD also includes 409,229 m⁶A-disease associations obtained by computational and statistical methods from 30 high-throughput sequencing datasets. In addition, we provide data on 5239 potential m⁶A regulatory proteins related to 24 cancers based on network analysis prediction methods. In addition, we have developed a tool to explore the function of m⁶A-modified genes through the protein–protein interaction networks. The M6ADD can be accessed at <http://m6add.edbc.org/>.

ARTICLE HISTORY

Received 15 November 2020
Revised 25 March 2021
Accepted 31 March 2021

KEYWORDS

M⁶a modification; diseases; experimentally confirmed data; high-throughput sequencing data

Introduction

There are currently more than 150 known chemical RNA modifications and m⁶A is the most common among them [1,2]. This post-transcriptional RNA modification is dynamic and reversible, regulated by methylases, demethylases, and proteins that preferentially recognize m⁶A modification [3–5]. At present, more than 20 m⁶A regulatory proteins have been discovered, known as RNA methylation writers, erasers and readers. These specific proteins can dynamically regulate m⁶A in cells, causing m⁶A to affect the structure and various functions of mRNA [6]. Common m⁶A proteins include methyltransferase like 3 (*METTL3*), methyltransferase like 14 (*METTL14*), FTO alpha-ketoglutarate dependent dioxygenase (*FTO*), alkB homolog 5 (*ALKBH5*), YTH N⁶-methyladenosine RNA binding protein 1 (*YTHDF1*) and YTH N⁶-methyladenosine RNA binding protein 2 (*YTHDF2*). The regulatory proteins affect mRNA metabolism by installing, removing, and selectively combining m⁶A modifications. For example, the analysis of mRNA expression levels between *METTL3* knockdown cells and controls showed that m⁶A modification stabilized mRNA. When the m⁶A of the transcripts was lost, the expression level of the transcripts decreased accordingly [7]. m⁶A modification plays an important role in a variety

of biological processes related to transcriptional regulation, such as RNA shearing [8], translation [9], mRNA stability, etc [10]. About one-third of mammalian mRNAs undergo m⁶A modification. Each mRNA has an average of 3–5 m⁶A modifications, and many m⁶A sites have evolved between human and mouse [11].

m⁶A modification is an important factor affecting disease and participates in the processes of many types of cancers [12]. *METTL14* regulates its mRNA targets (such as *MYB* and *MYC*) through m⁶A modification to play a carcinogenic role in acute myeloid leukaemia (AML) [13]. The *METTL3*/*YTHDF2* m⁶A axis degraded *SETD7* and *KLF4*, which promotes the development of bladder cancer [14]. m⁶A modification is significantly related to the abnormal expression of oncogenes, which increases the complexity of the gene regulatory mechanism [11]. *METTL3* is highly expressed in AML and plays a key role in AML cell survival and leukaemia progression in an m⁶A dependent manner by promoting the translation of target mRNAs [15]. The primary miRNA labelled with m⁶A can be distinguished by *hnrnpA2B1*, which interacts with *DGCR8* to promote primary miRNA processing [16]. In liver cancer cells, the reader protein *YTHDF2* can bind to an m⁶A site on the 3'-UTR of *EGFR*, resulting in a decrease in *EGFR* expression and inhibiting tumour cell proliferation [17]. In lung cancer cells, the

CONTACT Yan Zhang  zhangtyo@hit.edu.cn  School of Life Science and Technology, Computational Biology Research Center, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

[†]Dianshuang Zhou, Hongli Wang, Fanqi Bi and Jie Xing contributed equally to this work

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

writer protein METTL3 can positively regulate the expression of EZH2 through m⁶A modification, affecting the progression of lung cancer cells [18]. m⁶A modifications have also been found on noncoding RNAs(ncRNAs), and can affect their expression and function. The interaction between the long noncoding RNA (lncRNA) GAS5-AS and ALKBH5 affects the GAS5 m⁶A modification and enhances GAS5 stability [19]. The eraser protein ALKBH5 can change the m⁶A level of the lncRNA NEAT1, promoting the invasion and metastasis of gastric cancer cells [20]. The m⁶A modification of RNA can also affect other diseases. FTO is believed to be the first m⁶A demethylase related to obesity [21] and has also been linked to diabetes [22] and heart failure [23]. These studies show that m⁶A modification of RNA can affect the course of diseases and may be a potential treatment direction for disease.

Methylated RNA immunoprecipitation sequencing (MeRIP-seq) is a common method for detecting m⁶A modifications using high-throughput sequencing technology. Some high-throughput data using MeRIP-seq have been generated that focus on m⁶A modifier proteins interfering with changes at the cellular level, provide the characteristics of m⁶A modifications in tissue cells, and lay the foundation for further interpretation of m⁶A function in disease. The integration of m⁶A data and other public data such as Gene Expression Omnibus (GEO) [24], Sequence Read Archive (SRA) [25] and ENCODE [26] has contributed to some of the existing m⁶A databases. MeT-DB v2.0 [27], RMBase v2.0 [28], CVM6A [29] and REPIC [30] collected transcriptome-

wide m⁶A peaks from multiple species by using published MeRIP-seq data. M6AVar [31] combined m⁶A with variants that may affect the m⁶A function to explore the influence of m⁶A-related variants on post-transcriptional regulation. M6A2target [32] collected information about the WERs(writers, erasers and readers) of m⁶A and their targets, including data collected from the literature and high-throughput sequencing data.

Here, we have developed the M6ADD database, which includes manually curated data on potential m⁶A-disease associations for which some experimental evidence is available and data obtained from m⁶A high-throughput sequencing data analysis, aiming to explore the relationship between m⁶A modifications and gene disorders and diseases (Fig. 1). The M6ADD contains 222 experimentally confirmed m⁶A disease relationship pairs (m⁶A modified gene-disease) of 185 human and 37 mouse. We screened the m⁶A modified region of difference between normal and disease samples from the sequencing data of 30 kinds of 16 diseases through two calculation methods and provided statistically evaluated results. The m⁶A-disease data includes the m⁶A genomic location, disease name, m⁶A regulatory protein, regulation mode, tissue/cell line, experimental method and data source. We also predicted potential m⁶A regulatory proteins in 24 cancers based on The Cancer Genome Atlas (TCGA) data [33]. In addition, we also developed a PPI network tool to obtain protein interaction networks for m⁶A genes of interest, and inferred the functions of these genes.

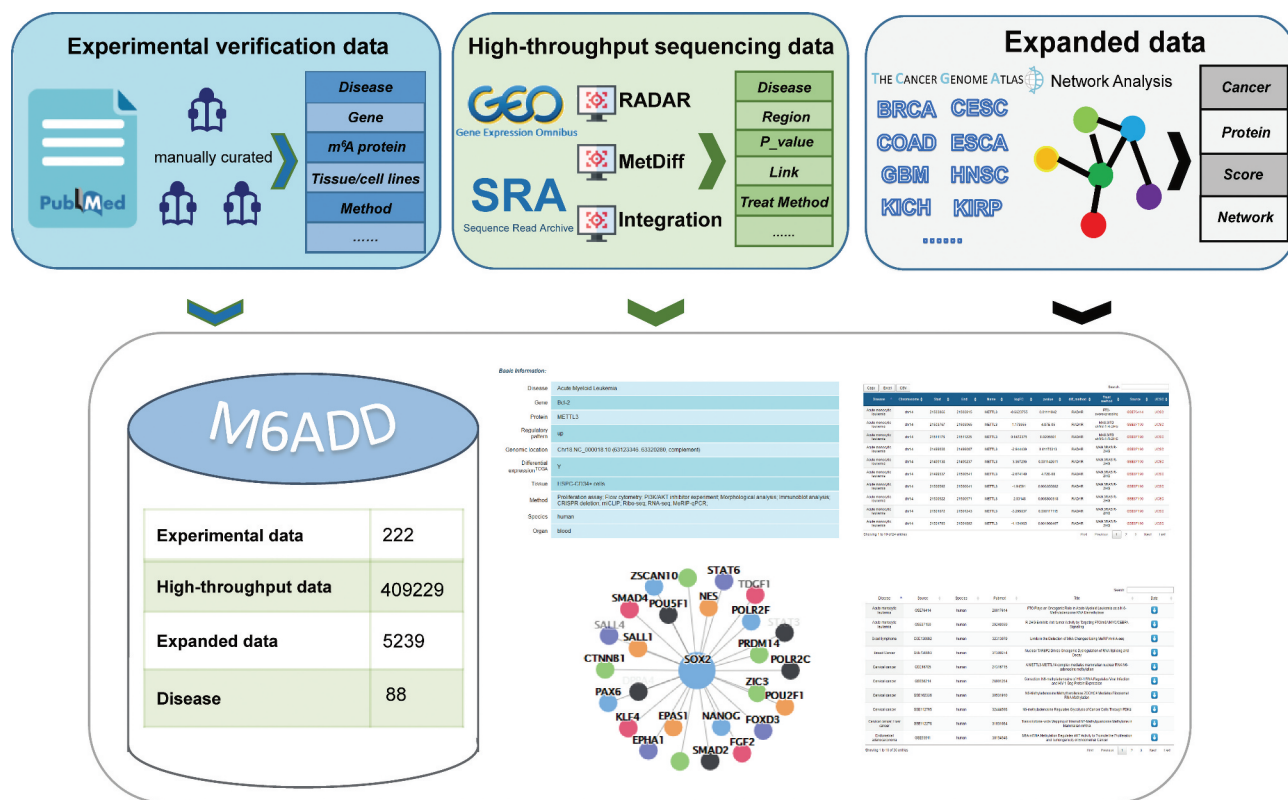


Figure 1. Data sources and overall design of the M6ADD.

Materials and methods

Collection of experimentally confirmed m⁶A-disease data

Based on the currently available m⁶A data related to diseases in public resources, we divided the data collected in the database into experimentally confirmed m⁶A-disease data and high-throughput sequencing m⁶A-disease data.

For the experimentally confirmed data, we searched the list of keywords in PubMed database [34]: ‘m⁶A’ and ‘N⁶-methyladenosine methylome’, and explored a total of nearly 3000 documents. We selected the literature on m⁶A-disease associations that had been experimentally confirmed, and we manually extracted the important data. The potential relationship between m⁶A modifications and diseases was investigated, in the selected papers, using experimental methods such as western blotting and RT-qPCR, which have strong credibility. To ensure the quality of the data, each piece of data was inspected at least twice. For each article, we extracted the diseases, genes, m⁶A regulatory proteins, regulatory mode, gene position in the genome, experimental tissues and cell lines (such as HEK293T, H1299, A549 cell lines), and experimental methods (such as western blot; real-time quantitative reverse transcription-PCR and RNA immunoprecipitation assays), species, disease organs, a brief description of the m⁶A and disease regulatory mechanisms, and the literature source. In particular, if the disease matched one of the cancer types in TCGA, we also utilized from the Gene expression profiling interactive analysis (GEPIA) tool [35] to retrieve some information from TCGA, for example whether the gene is a differentially expressed in that cancer. Finally, through manual mining of the literature, we obtained a total of 222 experimentally confirmed m⁶A-disease associations, including 59 diseases, 20 organs, 20 m⁶A regulatory proteins and 100 genes.

Collection of high-throughput sequencing m⁶A-disease data

For the high-throughput sequencing data, we retrieved and downloaded the raw data of m⁶A-disease high-throughput sequencing in the GEO database and the SRA database. We obtained 30 sets of high-throughput data by MeRIP-seq technology, including a total of 225 samples, covering 21 tissues and cell lines and obtained 409,229 m⁶A-disease associations (m⁶A modification site-disease). For each dataset, we used a unified method to process the original data. We used FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) to perform quality control on the original MeRIP-seq sequencing data, and used hisat2 [36] and the reference genome (hg38) for sequence alignment. Subsequently, samtools [37] was used to convert the SAM file generated by the alignment into a BAM file, and the sorted BAM file was used as the input file for the next step of the analysis. In the second step, we used two methods for the differential analysis of the MeRIP-seq data (RADAR [38] and MeTDiff [39]) to identify the differential m⁶A regions, and map these regions to the corresponding genes (hg38). We sorted the results obtained by these two methods. Each data contained disease, the

differential m⁶A region (chromosome, start position and end position), gene, logFC, P_value, sample processing method, data sources and a link to the UCSC Genome Browser [40] in the differential m⁶A area.

In addition to the results calculated by these two independent methods, we also developed a calculation method that integrates the results of these two methods. We divided the differentially methylated regions identified by the two methods into an up-regulated group (logFC>0) and a down-regulated group (logFC<0) to integrate them. The formula was as follows:

Where:

- logFC_{RADAR} represents the logFC value from RADAR;
- logFC_{MeTDiff} represents the logFC value from MeTDiff;

$$\text{Mean}_{\text{MeTDiff}} = \begin{cases} \sum_{i=1}^{M_1} \log FC_i / M_1; \log FC > 0 \\ \sum_{j=1}^{M_2} \log FC_j / M_2; \log FC < 0 \end{cases}$$

$$\text{Mean}_{\text{RADAR}} = \begin{cases} \sum_{i=1}^{R_1} \log FC_i / R_1; \log FC > 0 \\ \sum_{j=1}^{R_2} \log FC_j / R_2; \log FC < 0 \end{cases}$$

$$\text{ALL_mean}_1 = \left(\sum_{i=1}^{R_1} \log FC_i + \sum_{j=1}^{M_1} \log FC_j \right) / (R_1 + M_1); (\log FC > 0)$$

$$\text{ALL_mean}_2 = \left| \left(\sum_{i=1}^{R_2} \log FC_i + \sum_{j=1}^{M_2} \log FC_j \right) \right| / (R_2 + M_2); (\log FC < 0)$$

Where R1 and R2 represent the number of differentially up-regulated and down-regulated regions, respectively, identified by RADAR and M1 and M2 represent the number of differentially-up-regulated and down-regulated regions, respectively, identified by MeTDiff.

Collection of Protein-Protein Interaction (PPI) data related to cancer

We collected additional data from other data sources to help users explore in depth the potential relationship between m⁶A and disease. We selected the highest confidence (>0.9) interactions from the STRING v11 database [41], and obtained a total of 11,967 genes and 486,872 interactions.

We downloaded the expression profile data of 24 cancers from the TCGA database. After preprocessing the data, differential expression analysis of cancer samples and normal samples was performed (Log₂|FC|>1 and P_value<0.01) by using DESeq2 [42]. We integrated protein interaction networks derived from the STRING database and human protein reference database [43], mapped the genes differentially expressed in cancer as nodes to the integrated protein interaction network, and extracted one-step neighbour sites of these nodes as sub-networks. By searching m⁶A-related literatures, manually mining m⁶A regulatory proteins from these literatures. A total

of 27 m⁶A regulatory factors were summarized. We extracted the expression data of these m⁶A regulatory factors in cancer, calculated the Pearson correlation coefficient between the m⁶A regulatory factors and differentially expressed genes, and retained gene pairs with a correlation coefficient greater than 0.4. We then mapped the m⁶A regulatory factor as a seed gene to the sub-network, extracted the shortest distance between the seed gene and the differentially expressed genes in the network, and standardized it with the formula:

$$X^* = (X - X_{min}) / (X_{max} - X_{min})$$

where X is the distance between two nodes and X_{max} and X_{min} are the maximum and minimum values of the shortest distance in the entire network, respectively.

We defined the weighting formula for edges in the weighted network as:

$$W = 1 - (1 - R) * L$$

where W is the weight, R is the correlation between nodes, and L is the shortest standardized distance between two nodes.

Taking the one-step neighbour sub-network as the seed network, the weighting formula was used to calculate the weights of the m⁶A regulatory factors and the differentially expressed genes, and a list of weights between the m⁶A regulatory factors and the differentially expressed genes was obtained. Then we calculated the average of the minimum weights obtained from 1000 perturbations of the network as the cut-off and removed gene pairs lower than the cut-off in the weight list obtained in the first calculation. The remaining gene pairs were added to the one-step neighbour sub-network, thereby obtaining a reconstructed weight network. We used Cytoscape software [44] and MCODE tool [45] to mine the network, and screened for functional modules containing m⁶A regulatory factors. We sorted out the

modules of m⁶A regulatory factors corresponding to each cancer, visualized the modules network, screened out the genes that are differentially expressed in the cancers contained in the modules, and ranked the correlation weights of the genes and their corresponding m⁶A regulatory factors. Finally, 5239 cancer-related m⁶A regulatory protein prediction data were obtained.

Results

Exploring the functions of m⁶A related genes in disease

The M6ADD provides a large number of curated data from literature and related high-throughput sequencing results on potential m6A-disease associations for which some experimental evidence is available. Among the experimentally confirmed data, most of the data indicated the genes modified by m⁶A. In the high-throughput data, each differential m⁶A region identified based on the calculation method contained the corresponding gene. In order to facilitate users to explore the functions of the m⁶A-related genes of interest, M6ADD provides the ‘m⁶A-Net’ interface. Users can input the differential m⁶A modification gene included in the database to obtain the protein interaction network of the gene. A link to the DAVID website [46] is provided at the bottom of the interface to perform functional enrichment analysis on the obtained protein interaction network. As an example, we take the differential m⁶A modification gene SOX2, which is related to acute leukaemia, liver cancer, and cervical cancer. Clicking ‘Submit’ gives the SOX2 protein interaction network diagram and allows the picture to be exported (Fig. 2A). Below the network diagram, there are three links to DAVID to show functional annotations. Clicking on the links opens the DAVID website and allows annotation of SOX2 and the gene set that interacts with SOX2. The DAVID results show that SOX2 and the gene set

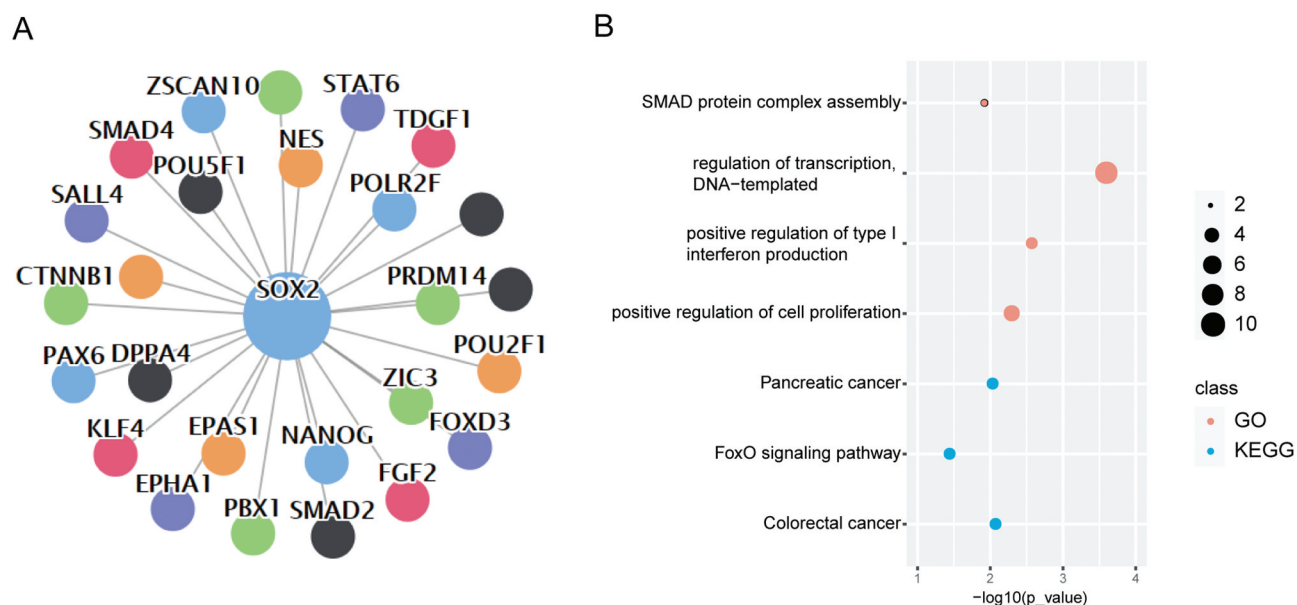


Figure 2. Application of the m⁶A-Net page. (A) PPI network diagram obtained by searching for SOX2 on the m⁶A-Net page. (B) Functional enrichment analysis of SOX2 and the gene set interacting with SOX2. SOX2 and the interacting genes are enriched in functions and pathways closely related to cancer.

interacting with SOX2 are significantly enriched in a number of cancer-related functional nodes and pathways, including 'DNA as template transcriptional regulation (GO: 0006355)', 'Positive regulation of cell proliferation (GO: 0008284)', 'SMAD protein complex assembly (GO: 0007183)', 'Positive regulation of type I interferon production (GO: 0032481)', 'FoxO signaling pathway (hsa04068)', 'Pancreatic cancer pathway (hsa05212)', 'Colorectal cancer pathway (hsa05210)', etc. (Fig. 2B). This shows that SOX2 is a gene closely related to a variety of cancers, and it would be worthwhile to carry out more in-depth research on SOX2 in a variety of carcinogenic mechanism.

Prediction of potential regulatory proteins in the m⁶A modification process

The M6ADD provides 5239 data on potential m⁶A regulatory proteins related to 24 cancers based on network analysis prediction methods (Fig. 3A). Each predicted protein related to a specific cancer has a prediction score. A higher score indicates a greater potential that the protein is the m⁶A regulatory protein for that cancer. We found that among the 5239 data on potential m⁶A regulatory proteins associated with 24 cancers, some proteins were predicted in multiple cancers. There were 87 proteins predicted in more than four cancers. Functional enrichment analysis of these 87 proteins

showed that these were significantly enriched in functions and pathways related to tumour cell proliferation and differentiation, including DNA replication, mitosis, cell differentiation, chromatin organization and cell proliferation (Fig. 3B). This indicates that these proteins may be common cancer proteins and could be related to underlying the biological processes and mechanisms of multiple cancers.

Among the 24 cancers, cholangiocarcinoma (CHOL) had the most m⁶A regulatory proteins according to the algorithm, with a total of 1,253 predicted data. Liver cancer (LIHC) had 546 predicted m⁶A regulatory protein data, renal cell carcinoma (KICH) had 527 m⁶A regulatory protein data, and pancreatic cancer (PAAD) had the least data, with nine m⁶A regulatory protein data. Taking renal papillary cell carcinoma as an example, there are currently five known m⁶A regulatory proteins, FMR1, HNRNPC, IGF2BP2, RBM15B and YTHDC1. Based on our prediction method, 34 potential m⁶A regulatory proteins related to renal papillary cell carcinoma were identified. Seven of these genes are included as cancer genes by the Cancer Gene Census [47], including BRCA1, MYH11, AXIN1, FAS, CDH1, GATA2, and MDM2. Using the hypergeometric distribution test, the P_{value} was less than 0.01 (Fig. 3C). This indicates that the predicted proteins provided by the M6ADD might be potentially cancer-related m⁶A regulatory proteins, which could provide new research directions for the future.

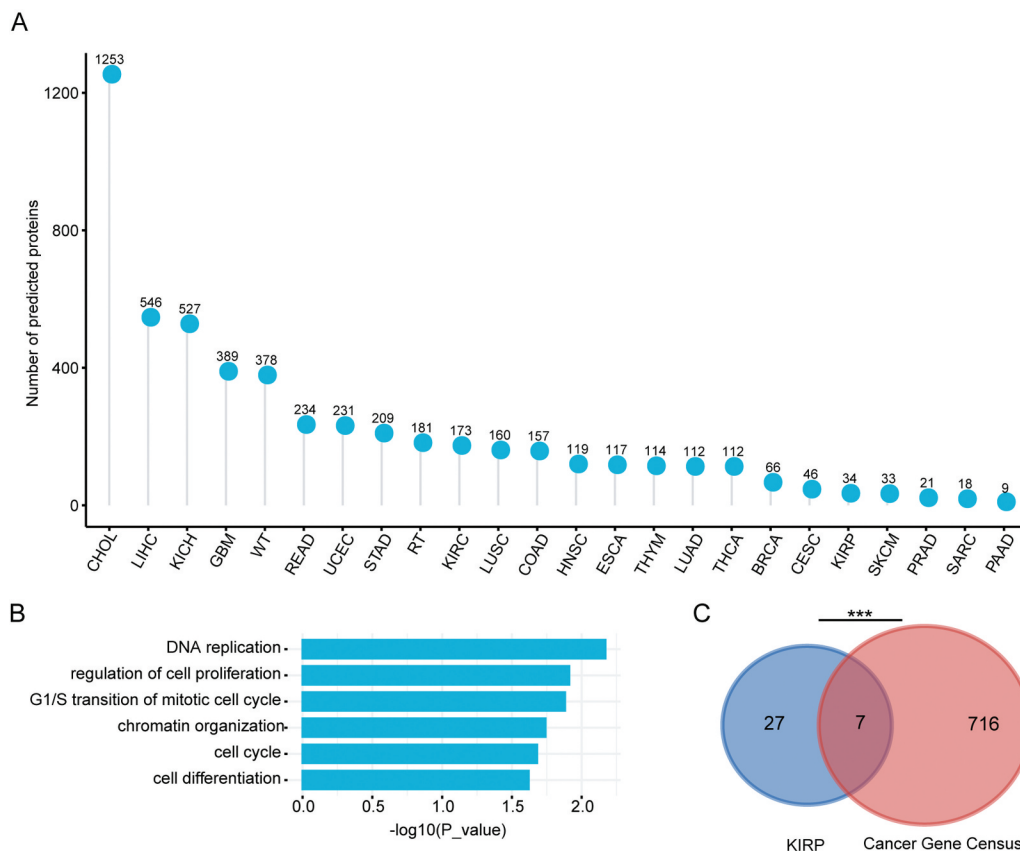


Figure 3. The M6ADD predicts the potential m⁶A regulatory proteins of various cancers. (A) The number of m⁶A regulatory protein in each cancer predicted in the M6ADD. (B) Functional annotation of 87 regulatory proteins predicted in a variety of cancers. (C) The overlap of seven predicted proteins in KIRP and CGC. Hypergeometric testing shows that the P_{value} is less than 0.01, indicating that the prediction result has a strong correlation with the cancer gene set.

Differential methylation analysis of m⁶A based on experimental conditions

We found that the samples were processed by different experimental methods before performing MeRIP-seq in the high-throughput data. The most common processing method for samples was to knockdown m⁶A regulatory protein expression, such as METTL3, METTL14 or METTL1 knockdown. Five sets of samples were processed with METTL3 knockdown, and the corresponding diseases were cervical cancer, endometrial adenocarcinoma, glioblastoma, liver cancer and type 2 diabetes. We selected the two cancers with the most data (cervical cancer, 1974 genes, and endometrial adenocarcinoma, 638 genes) to compare the integration results and found that a total of 107 genes were identified as differential m⁶A modification genes in both cancers (Fig. 4A). Hypergeometric analysis showed that the two gene sets had a strong correlation ($P < 0.01$). This shows that although the samples are different, the use of the same sample processing method gives a high similarity between the results calculated.

Previous studies have shown that m⁶A modification has strong tissue or cell specificity. Comparison of the m⁶A peak enrichment shows the strongest correlation between the same tissues or cells, even when different experiments have been performed. The M6ADD contains 30 sets of disease-related high-throughput data processing results for 16 diseases, of which liver cancer and cervical cancer have the largest number of studies. Liver cancer involves six studies (GSE37002, GSE90642, GSE102620, GSE110320, GSE134630 and GSE112276) and cervical cancer involves five studies (GSE112276, GSE46705, GSE86214, GSE102336 and GSE112795). In the six liver cancer studies, the ‘Heat_Shock’, ‘UV irradiation’, ‘shMETTL14’, ‘METTL3 knockdown’, ‘SETD2 knockdown’, ‘WTAP knockdown’, ‘KIAA1429 knockdown’ and ‘Induce EMT’ were used to process the samples. We found that using data from samples with different processing methods to calculate the difference in m⁶A modification resulted in larger differences. We have

used the integrated result to illustrate this phenomenon, because the integrated result contains the different m⁶A modifications identified by the two calculation methods. The first step was to identify the differences in the m⁶A-modified regions and the number of aligned genes. Among the six studies, five studies used HepG2 cells. Among them, the number of differential m⁶A regions and the number of genes in the comparison obtained by using the ‘UV irradiation’ method to process the sample data (GSE37002) were the largest, at 19,399 and 8369, respectively. When using the ‘SETD2 knockdown’ method to process the sample data (GSE134630), the number of differential m⁶A regions and the number of genes in the comparison were the least, at 54 and 50, respectively. In particular, in GSE110320 data set study, the researchers used knockdown of METTL14, METTL3, SETD2, and WTAP and then performed m⁶A-seq. We found that a total of 223 differential m⁶A modification genes were identified in total among four samples types, but the number of differential m⁶A modification genes in two types of samples was 23, and the number of differential m⁶A modification genes in three types of samples was 7, and the number of differential m⁶A modification genes in 4 types of samples were not identified (Fig. 4B). This shows the use of different methods of samples processing will lead to large differences in the identification of disease-related m⁶A genes, even when using the same cell line or tissue. It indicates that researchers need to pay attention to the processing methods used for each sample when using disease-related m⁶A high-throughput data in the GEO database. Each of the high-throughput sequencing datasets included in the M6ADD contain data on the sample processing methods, allowing users to select appropriate data.

Web interface and application

The M6ADD provides a user-friendly interface to help researchers explore m⁶A-disease data. The M6ADD contains several interfaces such as ‘Browse’, ‘Search’, ‘Download’,

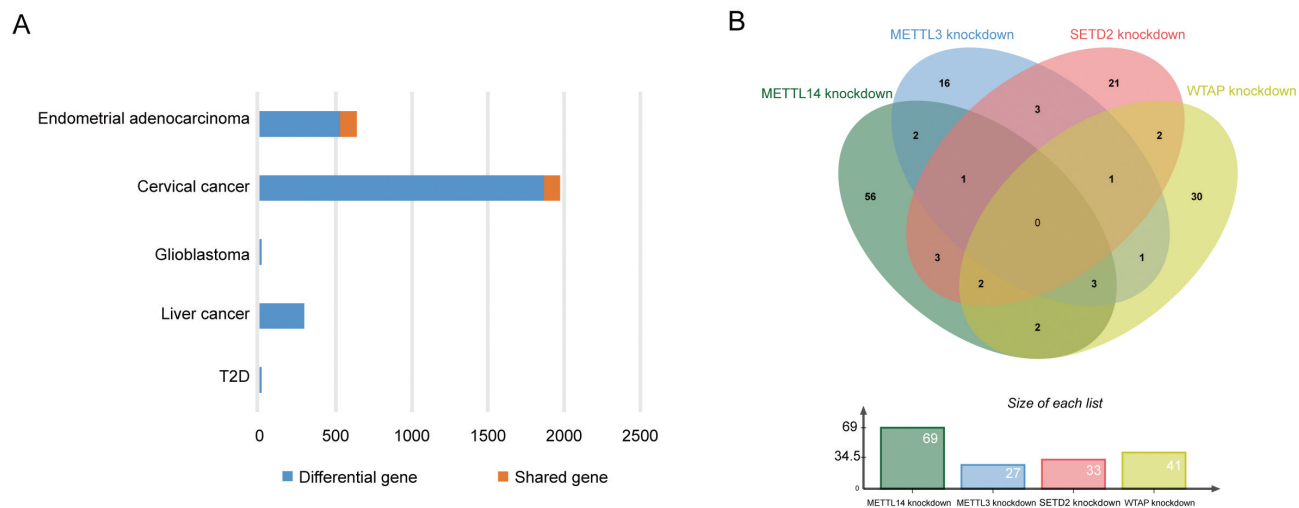


Figure 4. Processing methods have a great influence on the results. (A) Results of the same experimental method in different cancer samples. (B) Results of the same experimental sample under different methods.

'm⁶A-Net', 'm⁶A-Regulator' and 'Help'. In the 'Browse' interface, users can click on a specific 'gene', 'disease' or 'protein' to obtain the corresponding data. The Browse page divides the data into two parts: 'Experiment' and 'Sequencing'. In the Experiment section, there are three columns: gene, disease, and protein. By clicking on the word in each column, you get the corresponding data on the right side of the interface. The Sequencing section contains gene columns and has the same function as the Experiment section. In the 'Search' interface, the data search is divided into two parts, 'Search for experimental verification data' and 'Search for high-throughput sequencing data'. In the 'Search for experimental verification data' section, users can perform a combined search by selecting species and organs and entering the names of diseases, proteins, and genes. In the 'Search for high-throughput sequencing data' section, users can search by selecting the method of calculating the difference in m⁶A and entering the names of diseases and genes. In the download interface, the M6ADD provides all experimental confirmation data and high-throughput sequencing data. It also describes each set of high-throughput data and provides the results of the three dataset calculation methods. In the 'm⁶A-Net' interface, users can obtain the corresponding PPI network by entering a disease-related m⁶A gene, and can perform functional annotations. On the 'm⁶A-Regulator' interface, users can select cancer type and known m⁶A regulatory proteins to obtain potential cancer m⁶A regulatory proteins provided by the M6ADD. The help interface contains pictures and text descriptions of all the main interface functions in the database. A detailed usage for M6ADD is available on the interface of help. Users can better understand the function of browse, query, and download through reading the help interface.

The M6ADD provides a large amount of disease-focused m⁶A data. For example, users can search for 'Glioblastoma' in 'Search for experimental verification data' to obtain all current experimentally confirmed glioblastoma data. The result includes the target gene, m⁶A protein, regulation mode, whether it is differentially expressed in TCGA, organs, and a link to all data. Clicking 'detail' opens all the information about the data, which is divided into two categories, the basic information, and the data source (Fig. 5A). We found that many studies have confirmed that METTL3 is related to glioblastoma, and METTL3 could affect SOX2, SRSFs and ADAM19 to affect the biological process of glioblastoma. Selecting 'Integration' to search for 'Glioblastoma' in 'Search for high-throughput sequencing data' gives 2265 different m⁶A data, including P_value, sample processing method, integrated score, UCSC genome browser links, and other data. At the same time, users can click the button at the top left of the table to export the data (Fig. 5B). In 'm⁶A-Net', users can search for the gene SOX2 to get a protein interaction network including SMAD2, PBX1, EPAS1 and other genes, and can annotate the gene set through the link on the web page. In the 'm⁶A-Regulator' interface, users can select 'Glioblastoma' and the regulatory protein 'YTHDC1' to obtain the predicted 15 m⁶A regulatory proteins related to glioblastoma, and click on the 'Network Diagram' to get the glioblastoma Predict m⁶A regulatory protein network diagram (the yellow nodes in the network are differentially expressed genes

in cancer, and the green nodes are non-differentially expressed genes) (Fig. 5C).

Discussion and conclusions

Increasing attention is currently being paid to m⁶A modification, which is the most abundant modification on RNA and can regulate the expression of specific genes. Less than 100 studies were published in 2013, while nearly 1,000 studies were published in 2020, showing that m⁶A modification is becoming a new epigenetic research focus. m⁶A modification has been confirmed to be closely related to a variety of diseases, especially cancers. Many m⁶A regulatory proteins are expected to become potential cancer treatment targets. Interference with m⁶A regulatory proteins can affect the growth and survival of tumour cells. For example, meclufenamic acid (MA) can inhibit FTO (an m⁶A eraser) and thus inhibit the growth of GBM cells. As m⁶A-related studies increase, more m⁶A-related targets for cancer treatment will be discovered. We developed the M6ADD to make better use of existing resources, focusing on collecting data between m⁶A modification and diseases, and providing new m⁶A regulators predicted by a web-based method.

Each piece of data, in addition to providing direct data on the relationship between m⁶A modification and disease, also provides other important data to show mechanistic insight and experimental evidence. The experimentally confirmed data provides information on whether the m⁶A modified target gene was differentially expressed in TCGA data. In the high-throughput sequencing data, since the algorithm can identify the m⁶A peak region, we provide a link to the UCSC for this region, and users can obtain other genome-related information for this region through the UCSC website. In the predicted cancer m⁶A regulatory protein data, we provide the network results we identified and visualize the network content graphically. The functional enrichment analysis of our predicted data and comparison with known cancer genes also shows that our method has high accuracy. The M6ADD provides an independent tool to help users quickly obtain specific gene interaction proteins and the function of a gene set. For example, searching for the gene 'MYC' and the disease 'Cervical cancer' in the experimentally confirmed data shows that MYC has been confirmed to be related to cervical cancer and is regulated by the IGF2BP1/2/3 and FTO proteins. Searching for the gene 'MYC' and the disease 'Cervical cancer' in the high-throughput data section shows that MYC has been calculated to be related to cervical cancer. The MYC interaction proteins and their functions can be obtained using the m⁶A-Net tool. The functional results show that this gene set is significantly enriched in multiple cancer-related pathways, including the FOXO signalling pathway which is being considered as a therapeutic target for cancer treatment.

With the advancement of sequencing technology and experimental methods, more data will be generated in the future. These studies will provide opportunities for further expansion of the M6ADD. We will work to update the M6ADD data set, including both experimentally confirmed

A

Search for experimental verification data

Species: human mouse

Organ:

Disease:

Protein:

Gene:

Glioblastoma	ADAM19	METTL3	down	N	brain	detail
Glioblastoma	ADAM19	METTL14	down	N	brain	detail
Glioblastoma	SOX2	METTL3	up	Y	brain	detail
Glioblastoma	SRSFs	METTL3	up	N	brain	detail

Basic Information:

Disease: Glioblastoma

Gene: ADAM19

Protein: METTL3

B

Search for high-throughput sequencing data

Differential Method: RADAR MetDiff Integration

Disease:

Gene:

Disease	Chromosome	Start	End	Name	log ₂ C	pvalue	score	Transcript	Source	UCSC
Glioblastoma	chr3	138571109	138571317	CEP70	-2.83	2.25e-131	2.46442432026794	METTL14 knockdown	GSE94808	UCSC
Glioblastoma	chr12	51821191	51821290	FKBP2	-2.12	2.25e-131	2.09702457950016	METTL14 knockdown	GSE94808	UCSC
Glioblastoma	chr11	124766946	124767047	MSANTD2	-1.49	2.25e-131	1.77102199223803	METTL3 knockdown	GSE94808	UCSC
Glioblastoma	chr9	20413886	20413986	MLL3	-1.29	2.25e-131	1.66781910737387	METTL14 knockdown	GSE94808	UCSC
Glioblastoma	chr6	28436148	28436248	ZSCAN23	4.09	2.25e-131	0.742281080076675	METTL14 knockdown	GSE94808	UCSC
Glioblastoma	chr12	27963343	27963443	PTHLH	3.1	2.25e-131	0.3717678503838	METTL3 knockdown	GSE94808	UCSC
Glioblastoma	chr6	31733871	31733920	CLIC1	3.737669618	0	0.27397720323466	METTL14 knockdown	GSE94808	UCSC
Glioblastoma	chr16	71449393	71449493	ZNF23	2.85	2.25e-131	0.20312756020974	METTL3 knockdown	GSE94808	UCSC
Glioblastoma	chr8	60853285	60853334	CHD7	3.295838966	5.17e-12	0.10458113835758	METTL14 knockdown	GSE94808	UCSC
Glioblastoma	chr6	31624468	31624517	PRRC2A	3.238678452	1.25e-11	0.587045402016803	METTL14 knockdown	GSE94808	UCSC

Showing 1 to 10 of 2,265 entries

C

Cancer name:

Regulatory protein:

Disease	Verification	Protein	Ythc	Ythc
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	ANKRD3	1
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	TECTB	1
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	MYH8	1
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	MAS1	1
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	POU5F1B	1
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	RAB9B	0.932031471
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	FAM72D	0.928873944
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	ANKRD18B	0.919334967
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	PARDA6	0.910980774
Colon adenocarcinoma(COAD)	Network Diagram	FNRL1	RERG	0.91051172

Showing 1 to 10 of 157 entries

Figure 5. A schematic workflow of the M6ADD. (A) Search applications and results for experimental verification data. (B) Search applications and results for high-throughput sequencing data. (C) Search applications and results for predicted m⁶A regulatory protein data.

data and high-throughput data. We plan to update once every 6 months. In addition, we will develop new calculation and visualization tools to help researchers better use the data.

In summary, the M6ADD provides a convenient and useful data resource for researchers interested in studying the relationship between m⁶A modification and disease. The M6ADD presents a global view of m⁶A modification functions in human diseases and will help researchers to discover more cancer biomarkers and treatment targets.

Disclosure statement

No potential conflict of interest was reported by the author(s)

Funding

We thank the support of National Natural Science Foundation of China [grant number 61972116], Applied Technology Research and Development Plan of Heilongjiang Province [grant number GA20C018].

Data availability statements

Our pipeline and code are freely available on GitHub (<https://github.com/linjian-wgzg/m6ADD>) and M6ADD is hosted at <http://m6add.edbc.org/>.

References

- [1] Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat Rev Genet.* 2017;18(5):275–291.
- [2] Mathlin J, Le Pera L, Colombo T. A census and categorization method of epitranscriptomic marks. *Int J Mol Sci.* 2020;21(13):21.
- [3] Zhou Z, Lv J, Yu H, et al. Mechanism of RNA modification N6-methyladenosine in human cancer. *Mol Cancer.* 2020;19(1):104.
- [4] Yang Y, Hsu PJ, Chen YS, et al. Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res.* 2018;28(6):616–624.
- [5] Meyer KD, Jaffrey SR. Rethinking m(6)A readers, writers, and erasers. *Annu Rev Cell Dev Biol.* 2017;33:319–342.
- [6] Zaccara S, Ries RJ, Jaffrey SR. Reading, writing and erasing mRNA methylation. *Nat Rev Mol Cell Biol.* 2019;20(10):608–624.

- [7] Dominissini D, Moshitch-Moshkovitz S, Schwartz S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*. 2012;485(7397):201–206.
- [8] Zhao X, Yang Y, Sun BF, et al. FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res*. 2014;24(12):1403–1419.
- [9] Yang Y, Fan X, Mao M, et al. Extensive translation of circular RNAs driven by N(6)-methyladenosine. *Cell Res*. 2017;27(5):626–641.
- [10] Abakir A, Giles TC, Cristini A, et al. N(6)-methyladenosine regulates the stability of RNA:DNA hybrids in human cells. *Nat Genet*. 2020;52(1):48–55.
- [11] Huang H, Weng H, Chen J. m(6)A modification in coding and non-coding RNAs: roles and therapeutic implications in cancer. *Cancer Cell*. 2020;37(3):270–288.
- [12] Zhao W, Qi X, Liu L, et al. Epigenetic regulation of m(6)A modifications in human cancer. *Mol Ther Nucleic Acids*. 2020;19:405–412.
- [13] Weng H, Huang H, Wu H, et al. METTL14 inhibits hematopoietic stem/progenitor differentiation and promotes leukemogenesis via mRNA m(6)A modification. *Cell Stem Cell*. 2018;22(2):191–205 e9.
- [14] Xie H, Li J, Ying Y, et al. METTL3/YTHDF2 m(6) A axis promotes tumorigenesis by degrading SETD7 and KLF4 mRNAs in bladder cancer. *J Cell Mol Med*. 2020;24:4092–4104.
- [15] Vu LP, Pickering BF, Cheng Y, et al. The N(6)-methyladenosine (m(6)A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat Med*. 2017;23(11):1369–1376.
- [16] Alarcon CR, Goodarzi H, Lee H, et al. HNRNPA2B1 is a mediator of m(6)A-dependent nuclear RNA processing events. *Cell*. 2015;162(6):1299–1308.
- [17] Zhong L, Liao D, Zhang M, et al. YTHDF2 suppresses cell proliferation and growth via destabilizing the EGFR mRNA in hepatocellular carcinoma. *Cancer Lett*. 2019;442:252–261.
- [18] Chen WW, Qi JW, Hang Y, et al. Simvastatin is beneficial to lung cancer progression by inducing METTL3-induced m6A modification on EZH2 mRNA. *Eur Rev Med Pharmacol Sci*. 2020;24(8):4263–4270.
- [19] Wang X, Zhang J, Wang Y. Long noncoding RNA GAS5-AS1 suppresses growth and metastasis of cervical cancer by increasing GAS5 stability. *Am J Transl Res*. 2019;11(8):4909–4921.
- [20] Zhang J, Guo S, Piao HY, et al. ALKBH5 promotes invasion and metastasis of gastric cancer by decreasing methylation of the lncRNA NEAT1. *J Physiol Biochem*. 2019;75(3):379–389.
- [21] Jia G, Fu Y, Zhao X, et al. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol*. 2011;7(12):885–887.
- [22] Shen F, Huang W, Huang JT, et al. Decreased N(6)-methyladenosine in peripheral blood RNA from diabetic patients is associated with FTO expression rather than ALKBH5. *J Clin Endocrinol Metab*. 2015;100:E148–54.
- [23] Mathiyalagan P, Adamiak M, Mayourian J, et al. FTO-dependent N(6)-methyladenosine regulates cardiac function during remodeling and repair. *Circulation*. 2019;139:518–532.
- [24] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(D1):D991–5.
- [25] Kodama Y, Shumway M, Leinonen R, et al. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res*. 2012;40(D1):D54–6.
- [26] Davis CA, Hitz BC, Sloan CA, et al. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 2018;46(D1):D794–D801.
- [27] Liu H, Wang H, Wei Z, et al. MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res*. 2018;46(D1):D281–D7.
- [28] Xuan JJ, Sun WJ, Lin PH, et al. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res*. 2018;46(D1):D327–D34.
- [29] Han Y, Feng J, Xia L, et al. CVM6A: a visualization and exploration database for m(6)as in cell lines. *Cells*. 2019;8(2):168.
- [30] Liu S, Zhu A, He C, et al. REPIC: a database for exploring the N(6)-methyladenosine methylome. *Genome Biol*. 2020;21(1):100.
- [31] 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.
- [32] Deng S, Zhang H, Zhu K, et al. M6A2Target: a comprehensive database for targets of m6A writers, erasers and readers. *Brief Bioinform*. 2020. DOI:10.1093/bib/bbaa055.
- [33] Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19(1A):A68–77.
- [34] Coordinators NR. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2018;46(D1):D8–D13.
- [35] Tang Z, Li C, Kang B, et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. 2017;45(W1):W98–W102.
- [36] Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37(8):907–915.
- [37] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
- [38] Zhang Z, Zhan Q, Eckert M, et al. RADAR: differential analysis of MeRIP-seq data with a random effect model. *Genome Biol*. 2019;20(1):294.
- [39] Cui X, Zhang L, Meng J, et al. MeTDiff: a novel differential RNA methylation analysis for MeRIP-seq data. *IEEE/ACM Trans Comput Biol Bioinform*. 2018;15(2):526–534.
- [40] Haeussler M, Zweig AS, Tyner C, et al. The UCSC genome browser database: 2019 update. *Nucleic Acids Res*. 2019;47(D1):D853–D8.
- [41] Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607–D13.
- [42] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
- [43] Keshava Prasad TS, Goel R, Kandasamy K, et al. Human protein reference database—2009 update. *Nucleic Acids Res*. 2009;37:D767–72.
- [44] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–2504.
- [45] Rivera CG, Vakil R, Bader JS. NeMo: network module identification in cytoscape. *BMC Bioinformatics*. 2010;11(Suppl 1):S61.
- [46] Huang Da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44–57.
- [47] Sondka Z, Bamford S, Cole CG, et al. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*. 2018;18(11):696–705.