frontiers in Genetics

BRIEF RESEARCH REPORT
published: 28 May 2021
doi: 10.3389/fgene.2021.654820

REW-ISA V2: A Biclustering Method Fusing Homologous Information for Analyzing and Mining Epi-Transcriptome Data

Lin Zhang[1,2], *Shutao Chen*[1,2], *Jiani Ma*[1,2], *Zhaoyang Liu*[1,2] *and Hui Liu*[1,2]*

[1] *Engineering Research Center of Intelligent Control for Underground Space, China University of Mining and Technology, Ministry of Education, Xuzhou, China,* [2] *School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China*

Background: Previous studies have shown that N6-methyladenosine (m^6A) is related to many life processes and physiological and pathological phenomena. However, the specific regulatory mechanism of m^6A sites at the systematic level is not clear. Therefore, mining the RNA co-methylation patterns in the epi-transcriptome data is expected to explain the specific regulation mechanism of m^6A.

Methods: Considering that the epi-transcriptome data contains homologous information (the genes corresponding to the m^6A sites and the cell lines corresponding to the experimental conditions), rational use of this information will help reveal the regulatory mechanism of m^6A. Therefore, based on the RNA expression weighted iterative signature algorithm (REW-ISA), we have fused homologous information and developed the REW-ISA V2 algorithm.

Results: Then, REW-ISA V2 was applied in the MERIP-seq data to find potential local function blocks (LFBs), where sites are hyper-methylated simultaneously across the specific conditions. Finally, REW-ISA V2 obtained fifteen LFBs. Compared with the most advanced biclustering algorithm, the LFBs obtained by REW-ISA V2 have more significant biological significance. Further biological analysis showed that these LFBs were highly correlated with some signal pathways and m^6A methyltransferase.

Conclusion: REW-ISA V2 fuses homologous information to mine co-methylation patterns in the epi-transcriptome data, in which sites are co-methylated under specific conditions.

Keywords: m^6A methylation, homologous information, iterative signature algorithm, biclustering, unsupervised learning

OPEN ACCESS

Edited by:
Giovanni Nigita, The Ohio State University, United States

Reviewed by:
Rui Henriques, Universidade de Lisboa, Portugal
Jie Jiang, Xi'an Jiaotong-Liverpool University, China
Rathipriya R, Periyar University, India

***Correspondence:**
Hui Liu
hui.liu@cumt.edu.cn

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Genetics

Received: *17 January 2021*
Accepted: *28 April 2021*
Published: *28 May 2021*

Citation:
Zhang L, Chen S, Ma J, Liu Z and Liu H (2021) REW-ISA V2: A Biclustering Method Fusing Homologous Information for Analyzing and Mining Epi-Transcriptome Data. Front. Genet. 12:654820. doi: 10.3389/fgene.2021.654820

INTRODUCTION

At present, researchers have identified more than 170 different chemical modifications in RNA (Frye et al., 2018). N^6-methyladenine (m^6A) is the most common and abundant post-transcriptional RNA modification in mRNAs and long non-coding RNAs (Fu et al., 2014), and its methylation occurs at the sixth position of nitrogen atoms of adenosine. Studies have shown

that m^6A is involved in some RNA metabolic processes such as mRNA transcription, translation, nucleation, splicing and degradation (Ping et al., 2014; Lin et al., 2016; Deng et al., 2018). Besides, m^6A also plays an important role in the early development of eukaryotic cells, sex determination, antiviral immunity, brain development, and directed differentiation of hematopoietic stem cells (Zhang et al., 2017, 2019a). In addition to the above biological processes, m^6A modification is also related to many pathological phenomena, such as leukemia, glioma and hepatocellular carcinoma (Lachén-Montes et al., 2016; Chai et al., 2019).

The m^6A methylation in RNA is a dynamic and reversible process regulated by methyltransferases and demethylases. Since the main role of m^6A methyltransferases is to catalyze RNA to produce m^6A methylation modifications, these enzymes are often called "writers." The most common m^6A writer is composed of core components METTL3, METTL14, WTAP, and other subunits (Liu et al., 2014; Ping et al., 2014). On the contrary, m^6A demethylases mainly mediate m^6A demethylation, so these enzymes are also known as "erasers." The common erasers are FTO, AKLBH5, and so on (Jia et al., 2011). Studies have shown that m^6A has a series of biological functions because many RNA binding proteins mediate it. These binding proteins can specifically recognize m^6A methylated adenosine on RNA, so these proteins are often referred to as "readers." The common readers include protein YT521-B homologous (YTH) domain family (Meyer et al., 2015), etc. In recent years, with the development of methylated RNA immunoprecipitation sequencing (MeRIP-seq, or m^6A-seq) technology (Dominissini et al., 2012; Meng et al., 2014), many m^6A experimental data continue to emerge, which makes it possible to analyze m^6A in the whole transcriptome. However, since there are a few enzymes, such as m^6A writers, erasers and readers only, each enzyme may regulate a large number of m^6A sites. In other words, the methylation level of m^6A site regulated by the same enzyme may share the same pattern, which is called the co-methylation pattern of m^6A.

Till this day, some researchers have used clustering methods to study the co-methylation patterns in epi-transcriptome data, trying to clarify the functional mechanism of m^6A methylation. Based on MeRIP-seq data, Liu et al. used *k*-means clustering, hierarchical clustering, Bayesian factor regression model and non-negative matrix decomposition to cluster m^6A sites (Liu et al., 2015). To better fit the distribution of epi-transcriptome data, Zhang et al. proposed an infinite beta binomial mixture model based on Dirichlet Process (DPBBM) to reveal the co-methylation patterns (Zhang et al., 2019b). Besides, our previously proposed RNA Expression Weighted Iterative Signature Algorithm (REW-ISA) (Zhang et al., 2020) applied biclustering to the analysis of epi-transcriptome data for the first time. However, the above methods only used the read counts of the m^6A sites of the IP sample and the input sample in MeRIP-seq data. They did not fully consider the homologous information of sites and experimental conditions. Homology is a central concept in comparative biology, in which the most basic meaning of homology is to have a common ancestor. The homologous information of MeRIP-seq data can be divided into two categories: the genes corresponding to the m^6A sites and the cell lines (or environments) corresponding to the experimental conditions. Appropriate use of the above-mentioned homologous information will help discover potential local functional blocks (LFBs) and better reveal the m^6A regulatory mechanism. Besides, although some of the most advanced biclustering methods have been developed, such as runibic (Wang et al., 2016; Orzechowski et al., 2018a), EBIC (Orzechowski et al., 2018b), QUBIC2 (Xie et al., 2020) and RecBic (Liu et al., 2020), their goal is to identify the trend-preserving biclusters. However, when mining m^6A co-methylation pattern, it is expected to obtain locally hyper-methylated biclusters, so these new methods are not applicable.

Therefore, we proposed an improved RNA expression weighted iterative signature algorithm (REW-ISA V2), which fuses the homologous information of sites and experimental conditions in the iterative search for LFBs. Consistent with the previous method, each potential LFB is identified by the row threshold (defined as T_R) and column threshold (defined as T_C) during the LFB searching strategy. It is important to note that REW-ISA V2 updates T_R and T_C's selection process, optimizing the selection of thresholds through the built-in rich constraint framework. According to the previous study (Henriques et al., 2015, 2017), REW-ISA V2 is a non-deterministic greedy algorithm, which can be used to find hyper-methylated biclusters. Besides, REW-ISA V2 can obtain these overlapping LFBs when there is overlap between the LFBs implied in the input data.

To verify the effectiveness of the fusion of homologous information, REW-ISA V2 was applied to the collected MERIP-seq data to find potential LFBs. The obtained LFBs were further analyzed by the Gene Ontology (GO) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, and enzyme-specific experiments, in an attempt to reveal the possible regulatory mechanism of m^6A. As a result, REW-ISA V2 can better find potential LFBs with high methylation levels in the epi-transcriptome data.

METHODS

Pre-processing of Real Data

As is known, MeRIP-Seq data profiles the m^6A epi-transcriptome by IP and input samples. Thus, we first need to follow (Chen et al., 2019) and (Wu et al., 2019) to quantify the information of m^6A sites. Specifically, after downloading the sequencing data from Gene Expression Omnibus (GEO) in SRA format, the Tophat2 (Kim et al., 2013) needs to be used to compare the sequencing data reads with the human reference genome, and finally obtain the Fragments Per Kilobase of transcript per Million (FPKM) statistics.

To mine the potential LFBs in the epi-transcriptome data, only the FPKM statistical information of IP and input samples are not enough. It is necessary to calculate the m^6A methylation level of each m^6A site under each experimental condition. Let m denote the total number of m^6A sites and n denote the total number of conditions. Therefore, according to the REW-ISA, the methylation level matrix $P \in \mathbb{R}^{m \times n}$ and the RNA expression level

matrix $W \in \mathbb{R}^{m\times n}$ can be further calculated using the IP sample and the input samples, as shown in (1, 2).

$$p_{ij} = \frac{t_{ij}+\alpha}{t_{ij}+h_{ij}+2\alpha}, \tag{1}$$

$$w_{ij} = \log(t_{ij}+h_{ij}+1). \tag{2}$$

In (1) and (2), t_{ij} represents the FPKM of the i-th m^{6}A site under the j-th condition in the IP sample, and h_{ij} represents the FPKM of the i-th m^{6}A site under the j-th condition in the input sample. Besides, α in (1) is a very small value, aiming to avoid *NaN* where FPKM of both IP and input samples are zeros. The purpose of introducing the RNA expression level is to provide a confidence level for m^{6}A methylation level in further biclustering analysis.

REW-ISA V2

To eliminate the effect of global sites or conditions on $\boldsymbol{P}$, REW-ISA V2 performs standard normalization on the whole, rows and columns of $\boldsymbol{P}$ in turn to eliminate the global effect, as shown in (3–5). $\boldsymbol{P^{nw}}$, $\boldsymbol{P^{nr}}$, and $\boldsymbol{P^{nc}}$ represent the matrices obtained after whole normalization, row normalization, and column normalization, respectively.

$$p_{ij}^{nw} = \frac{p_{ij} - \text{mean}(\boldsymbol{P})}{\max(\boldsymbol{P}) - \min(\boldsymbol{P})}, \tag{3}$$

$$p_{ij}^{nr} = \frac{P_{ij}^{nw} - \text{mean}(\boldsymbol{P_{i\cdot}^{nw}})}{\max(\boldsymbol{P_{i\cdot}^{nw}}) - \min(\boldsymbol{P_{i\cdot}^{nw}})}, \tag{4}$$

$$p_{ij}^{nc} = \frac{p_{ij}^{nr} - \text{mean}(\boldsymbol{P_{\cdot j}^{nr}})}{\max(\boldsymbol{P_{\cdot j}^{nr}}) - \min(\boldsymbol{P_{\cdot j}^{nr}})}. \tag{5}$$

In (3–5), mean(·) represents calculating the mean value, max(·) represents calculating the maximum value, and min(·) represents calculating the minimum value. $\boldsymbol{P_{i\cdot}^{nw}}$ represents the i-th row in $\boldsymbol{P^{nw}}$, and $\boldsymbol{P_{\cdot j}^{nr}}$ represents the j-th column in $\boldsymbol{P^{nr}}$. Then min-max normalization is performed on the overall data to generate $\boldsymbol{P^t}$, which will facilitate subsequent combination with RNA expression level.

$$p_{ij}^{t} = \frac{p_{ij}^{nc} - \min(\boldsymbol{P^{nc}})}{\max(\boldsymbol{P^{nc}}) - \min(\boldsymbol{P^{nc}})}. \tag{6}$$

For the RNA expression level matrix $\boldsymbol{W}$, since its distribution fluctuates with the MeRIP-seq data, it is necessary to perform the min-max normalization on $\boldsymbol{W}$ to generate $\boldsymbol{W^t}$, which acts as confidence matrix for $\boldsymbol{P^t}$.

$$w_{ij}^{t} = \frac{w_{ij} - \min(\boldsymbol{W})}{\max(\boldsymbol{W}) - \min(\boldsymbol{W})}. \tag{7}$$

Suppose that k-1 ($2 \le k \le K$) LFBs have been found, and the k-th LFB is currently being searched. Assuming that the k-th LFB is $\boldsymbol{B_k}$, the site indicator ρ_k and the condition indicator κ_k are used to indicate the sites and conditions contained in $\boldsymbol{B_k}$. Specifically, the site indication ρ_{ik} is one if the i-th site is present in $\boldsymbol{B_k}$ (zero otherwise). The condition indication κ_{jk} is one if the j-th condition is present in $\boldsymbol{B_k}$ (zero otherwise). The average methylation level μ_k^p and average expression level μ_k^w of $\boldsymbol{B_k}$ can be further calculated, as shown in (8, 9), respectively.

$$\mu_k^p = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}^t \rho_{ik}\kappa_{jk}}{\sum_{i=1}^{m}\rho_{ik}\sum_{j=1}^{n}\kappa_{jk}}, \tag{8}$$

$$\mu_k^w = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}^t \rho_{ik}\kappa_{jk}}{\sum_{i=1}^{m}\rho_{ik}\sum_{j=1}^{n}\kappa_{jk}}. \tag{9}$$

Each time a LFB is found, the average methylation level and average expression level of the LFB should to be removed from $\boldsymbol{P^t}$ and $\boldsymbol{W^t}$. The purpose of removing is to prevent the algorithm from falling into a loop looking for a strong LFB. Let residual matrix $\boldsymbol{P^{(k)}}$ represent the methylation level matrix after eliminating the μ^p of the first k-1 LFBs,

$$p_{ij}^{(k)} = p_{ij}^t - \sum_{z=1}^{k-1}(\mu_z^p \rho_{iz}\kappa_{jz}). \tag{10}$$

Then, $\boldsymbol{P^{(k)}}$ turns into $\boldsymbol{P^{R(k)}}$ after row min-max normalization and turns into $\boldsymbol{P^{C(k)}}$ after column min-max normalization. Similarly, let $\boldsymbol{W^{(k)}}$ represent the RNA expression level matrix after eliminating the μ^w of the first k-1 LFBs,

$$w_{ij}^{(k)} = w_{ij}^t - \sum_{z=1}^{k-1}(\mu_z^w \rho_{iz}\kappa_{jz}). \tag{11}$$

After obtaining the above $\boldsymbol{P^{R(k)}}$, $\boldsymbol{P^{C(k)}}$ and $\boldsymbol{W^{(k)}}$, combined with the homologous information of sites and conditions, the algorithm begins to search for LFBs iteratively. The algorithm running from a randomly selected site's subset $\boldsymbol{U'}$ and updates the conditions' subset $\boldsymbol{V'}$ according to (12).

$$\begin{cases} e_{\mathbf{U'}v}^C = \frac{1}{|\boldsymbol{U'}|}\sum_{u\in\mathbf{U'}}(w_{uv}^{(k)} \cdot p_{uv}^{R(k)}) & v \in \boldsymbol{V} \\ t_{\mathbf{U'}v}^C = \left|\rho(P_{\mathbf{U'}v}^t \cdot W_{\mathbf{U'}v}^t, \frac{\sum_{b\in\boldsymbol{H_v^C}}(P_{\mathbf{U'}b}^t \cdot W_{\mathbf{U'}b}^t)}{|\boldsymbol{H_v^C}|})\right| & v \in \boldsymbol{H_v^C},\ \boldsymbol{H_v^C} \in \boldsymbol{V}, \\ \boldsymbol{V'} = \{v \in \boldsymbol{V} : \left|e_{\mathbf{U'}v}^C \cdot t_{\mathbf{U'}v}^C - \frac{1}{|\boldsymbol{V}|}\sum_{v\in\boldsymbol{V}} e_{\mathbf{U'}v}^C \cdot t_{\mathbf{U'}v}^C\right| > \frac{T_C}{\sqrt{|\boldsymbol{U'}|}}\} \end{cases} \tag{12}$$

where $\boldsymbol{V}$ is the conditions set of $\boldsymbol{P^t}$, refers to the u-th site under the v-th condition in $\boldsymbol{P^R}$, is the RNA expression level of the u-th site under v-th condition, $\boldsymbol{H_v^C}$ represents the subset of homologous conditions corresponding to the v-th condition. $\rho(\cdot)$ represents to calculate Pearson similarity, $|\cdot|$ represents to calculate absolute value (or module). Besides, $\boldsymbol{T_C}$ is a hyperparameter, and its function is to select the subset of conditions $\boldsymbol{V'}$. In (12), $e_{\mathbf{U'}v}^C$ is calculated based on $\boldsymbol{P^{R(k)}}$ and $\boldsymbol{W^{(k)}}$, which represents the average methylation level score of the v-th condition combined with the confidence of the expression level. $t_{\mathbf{U'}v}^C$ is calculated based on $\boldsymbol{P^t}$ and $\boldsymbol{W^t}$, representing the average similarity score of the v-th condition relative to its homologous conditions subset. In the process of calculating $e_{\mathbf{U'}v}^C$ and, only the sites involved in $\boldsymbol{U'}$ are considered.

Then, the subsets of sites are updated following (13).

$$\begin{cases} e_{u\mathbf{V'}}^R = \frac{1}{|\boldsymbol{V'}|}\sum_{v\in\mathbf{V'}}(w_{uv}^{(k)} \cdot p_{uv}^{C(k)})\ u \in \boldsymbol{U} \\ t_{u\mathbf{V'}}^R = \left|\rho(P_{u\mathbf{V'}}^t \cdot W_{u\mathbf{V'}}^t, \frac{\sum_{a\in\boldsymbol{H_u^R}}(P_{a\mathbf{V'}}^t \cdot W_{a\mathbf{V'}}^t)}{|\boldsymbol{H_u^R}|})\right| & u \in \boldsymbol{H_u^R}, \boldsymbol{H_u^R} \in \boldsymbol{U}, \\ \boldsymbol{U'} = \{u \in \boldsymbol{U} : \left|e_{u\mathbf{V'}}^R \cdot t_{u\mathbf{V'}}^R - \frac{1}{|\boldsymbol{U}|}\sum_{u\in\boldsymbol{U}} e_{u\mathbf{V'}}^R \cdot t_{u\mathbf{V'}}^R\right| > \frac{T_R}{\sqrt{|\boldsymbol{V'}|}}\} \end{cases} \tag{13}$$

where $\boldsymbol{U}$ is the sites set of $\boldsymbol{P^t}$, refers to the u-th site under the v-th condition in $\boldsymbol{P^C}$, $\boldsymbol{H_u^R}$ represents the subset of homologous sites corresponding to the u-th site. Besides, T_R is a hyperparameter, and its function is to update the subset of sites $\boldsymbol{U'}$. $e^R_{u\boldsymbol{V'}}$ represents the average methylation level score of the u-th site combined with the confidence of the expression level. $t^R_{u\boldsymbol{V'}}$ represents the average similarity score of the u-th site relative to its homologous sites subset. In the process of calculating $e^R_{u\boldsymbol{V'}}$ and $t^R_{u\boldsymbol{V''}}$, only the conditions involved in $\boldsymbol{V'}$ are considered.

Using the preset hyperparameters T_R and T_C, $\boldsymbol{U'}$ and $\boldsymbol{V'}$ are updated iteratively by (12, 13) until convergence is satisfied (or the maximum number of preset iterations is reached). The convergence condition is shown in (14).

$$\frac{|\boldsymbol{U'} \cap \boldsymbol{U''}|}{|\boldsymbol{U'} \cup \boldsymbol{U''}|} \geq \varepsilon \tag{14}$$

where ε is the default convergence criteria, and its value is slightly <1. $\boldsymbol{U''}$ represents the site's subset in the previous iteration, and $\boldsymbol{U'}$ represents its subset in the current iteration. If the algorithm converges within the maximum number of iterations, it means that the k-th LFB, $\boldsymbol{B_k} = \{\boldsymbol{U'}, \boldsymbol{V'}\}$ has been found. The flow chart of searching for the k-th LFB by REW-ISA V2 is shown in **Figure 1**.

Then the algorithm will return to (8) and continue to look for the next LFB. Conversely, if the convergence condition of (14) is not satisfied when the algorithm reaches the maximum number of iterations, REW-ISA V2 will automatically terminate and output all previously obtained LFBs. We recommend setting ε to 0.99 and the maximum number of iterations not <50. The closer the value of ε is to 1 and the greater the maximum number of iterations, the more accurate the LFBs obtained by REW-ISA V2. The REW-ISA V2 algorithm based on R language can be downloaded freely from https://github.com/labiip/REWISAV2.

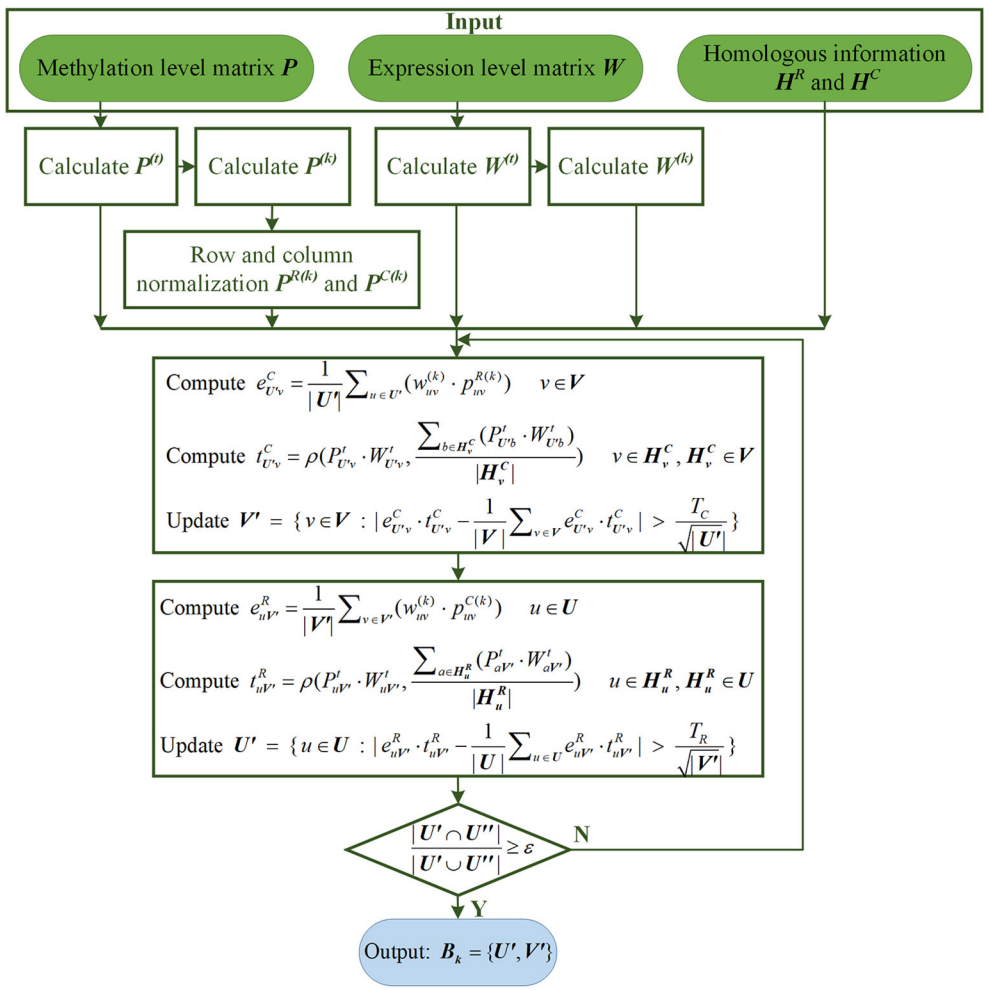

FIGURE 1 | The flow chart of searching for the k-th LFB by REW-ISA V2.

Enrichment Constraint Framework

It can be seen from (12, 13) that the selection of T_R and T_C will greatly affect the biological significance of the obtained LFBs. Therefore, based on Meng et al. (2009), we introduced a grid search-based enrichment constraint framework for the algorithm to optimize T_R and T_C selection further. For LFBs obtained under different T_R and T_C combinations, we need to extract the genes corresponding to the m^{6}A sites in each LFB and then perform GO analysis based on "clusterProfiler" (Yu et al., 2012) for each LFB. For the range of T_R and T_C, we recommend setting it between 0 and 3, and the step size is 0.1. On this basis, the range of specific thresholds should be appropriately adjusted according to the input real data. Assuming that a LFB is obtained, the number of genes corresponding to the m^{6}A site contained in it is M. The number of GO terms obtained by GO analysis of the LFB is l. Then the weighted enrichment score (WE_score) (Li et al., 2012) of this LFB can be calculated by (15).

$$\begin{cases} s_i = -\log(p_i) \\ \text{WE\_score} = \frac{s_1 m_1/M + s_2 m_2/M + \cdots + s_l m_l/M}{m_1/M + m_2/M + \cdots + m_l/M + m_{non}/M}, \end{cases} \tag{15}$$

where p_i is the p-value of the i-th GO term, m_i is the number of genes of the i-th GO term enriched, m_{non} is the number of genes covered by LFB but not enriched by any GO term. The higher the WE_score, the stronger the biological significance of this LFB.

However, as the number of genes corresponding to the sites in LFB increases, WE_score will also show an increasing trend, as shown in **Supplementary Figure 1**. Therefore, only using WE_score to evaluate the biological significance of obtained LFBs is not perfect, and the number of genes corresponding to the sites in LFBs also needs to be considered. Assume that the data analyzed contain a total of M_{all} genes, and further assume an obtained LFB is containing M genes and record its WE_score as W_m. We randomly select M genes from all genomes, and their WE_score is recorded as W_{rm}. The relative promotion rate (RPR) of WE_score can be further calculated, as shown in (16).

$$\text{RPR} = \frac{M(W_m - W_{rm})}{M_{all} W_{rm}}. \tag{16}$$

The larger the RPR is, the larger the area of the obtained LFB is, and the more biological significance of the obtained LFB is. On the one hand, in the actual process of mining LFBs, we hope to get more LFBs. On the other hand, we hope to get LFBs with rich biological significance. Therefore, the number of LFBs obtained by each pair of threshold combinations is obtained by grid search under different T_R and T_C combinations. The threshold combinations corresponding to the maximum number of LFBs are selected. Then, the average RPR of the LFBs is calculated based on the selected combination of T_R and T_C. Finally, the optimal T_R and T_C are the threshold combinations corresponding to the maximum average RPR.

RESULTS

We collected 32 samples from 10 publicly human m^{6}A MeRIP-seq datasets (Dominissini et al., 2012; Meyer et al., 2012; Fustin et al., 2013; Batista et al., 2014; Schwartz et al., 2014; Wang et al., 2014; Barbieri et al., 2017; Li et al., 2017; Pendleton et al., 2017) to mine potential LFBs, most of which can be retrieved from the MeT-DB V2.0 database (Liu et al., 2018). **Table 1** summarizes the MeRIP-seq real data set used in this project. Then, calculate the corresponding $\boldsymbol{P}$ and $\boldsymbol{W}$ through (1) and (2), and perform REW-ISA V2. Within the range of T_R being 0.1-2 with step size 0.1, and T_C being 0.1–2 with step size 0.1, T_R and T_C are optimized through the enrichment constraint framework. The experiments were repeated ten times for each parameter setting. Although optimizing T_R and T_C based on the gathered biological significance may produce biased results. However, this process provides guidance for the selection of T_R and T_C. Finally, under the optimal T_R of 0.4 and the optimal T_C of 0.7, a total of fifteen LFBs are obtained. The number of m^{6}A sites, the number of genes corresponding to m^{6}A sites and the number of conditions contained in these LFBs are shown in **Supplementary Table 1**.

For the above-mentioned real data set, Bimax (Prelić et al., 2006), Xmotifs (Murali and Kasif, 2003), Plaid (Lazzeroni and Owen, 2002), ISA (Bergmann et al., 2003), REW-ISA (Zhang et al., 2020), FBCwPlaid (Chen et al., 2021), runibic (Orzechowski et al., 2018a), and QUBIC2 (Xie et al., 2020) were all included for comparison with REW-ISA V2. To make the LFBs obtained by the above methods have significant biological significance, the parameters of these methods have been appropriately adjusted. For each LFB obtained by each method, the two enrichment indicators, WE_score and RPR, were both calculated for evaluation. The comparison results are shown in **Figures 2A,B**, respectively. As can be seen from **Figure 2A**, the average WE_score of the LFBs obtained by the REW-ISA V2 algorithm is higher than that of ISA and REW-ISA, which indicates that the fusion of homologous information is effective for mining LFBs. Although the average WE_score of LFBs obtained by REW-ISA V2 is lower than that of the FBCwPlaid algorithm, there are significant differences in RPR between the two methods. After further analysis of the LFBs, we found that this was caused by the size of LFBs found by REW-ISA V2 was smaller than that found by the FBCwPlaid algorithm. In other words, the LFBs found by REW-ISA V2 had higher enrichment scores with fewer corresponding genes. Besides, we can find that runbic and QUBIC2 do not perform well in the task of m^{6}A hyper-methylation pattern recognition. It may be due to the following two points. On the one hand, the two algorithms mainly identify the trend-preserving biclusters, which is different from the hyper-methylation bicluster. On the other hand, the LFBs obtained are generally small. This also reflects the need of developing biclustering methods for epi-transcriptome data. In a word, the average RPR of LFBs inferred by REW-ISA V2 is significantly higher than that of other biclustering algorithms, which means that the LFBs obtained by REW-ISA V2 may be more biologically significant.

To further explore the biological significance of the obtained LFBs, we selected four LFBs with more sites from the fifteen LFBs. As can be seen from **Supplementary Table 1**, for the four selected LFBs, they cover 1,256, 1,619, 824, and 1,148 genes, respectively. An important feature of any biclustering is the

TABLE 1 | MeRIP-seq datasets used in the study.

ID	GEO accession	Cell line	Treatment	Source
1–4	SRR456542–SRR456549, SRR456551–SRR456557	HepG2	UV, HGF, IFN, UT	Dominissini et al., 2012
5–6	SRR903368–SRR903379	U2OS	CTL, DAA	Fustin et al., 2013
7–10	SRR847358–SRR847377	HeLa	Ctrl, METTL3-, METTL14-, WTAP-	Liu et al., 2014
11–12	SRR1182582–SRR1182590	ES/NPC	hNPC, hESC	Schwartz et al., 2014
13–18	SRR1182591–SRR1182596, SRR494613–SRR494618, SRR5080301–SRR50312	HEK293	Ctrl, WTAP-, METTL3-, METTL16-	Meyer et al., 2012; Schwartz et al., 2014; Pendleton et al., 2017
19–21	SRR1182597–SRR1182602	OKMS	D0, D5_WITH_DOX, D5_WO_DOX	Schwartz et al., 2014
22–26	SRR1182603–SRR1182630	A549	Ctrl, METTL3-, METTL14-, WTAP-, KIAA1429-	Schwartz et al., 2014
27–28	SRR3066062–SRR3066069	AML	Ctrl, FTO+	Li et al., 2017
29–30	SRR5239086–SRR5239109	AML2	Ctrl, METTL3-	Barbieri et al., 2017
31–32	SRR1035213–SRR1035224	ESC	T0, T48	Batista et al., 2014

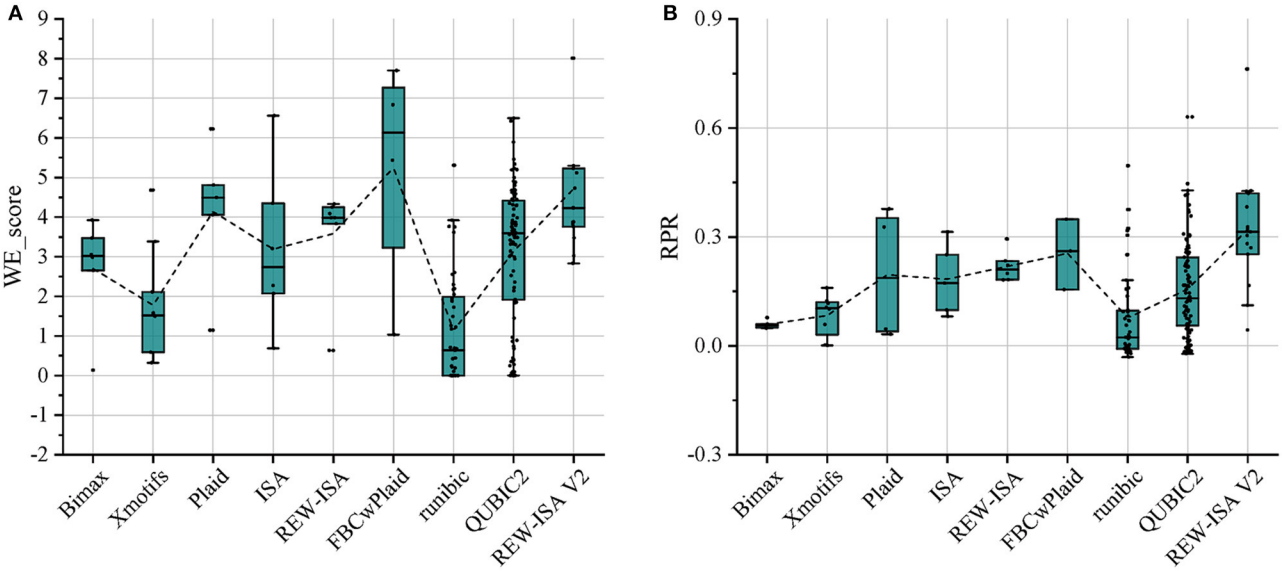

FIGURE 2 | Compare the LFBs obtained by the nine methods through two enrichment evaluation indicators. **(A)** The WE_score comparison of LFBs obtained by each method. **(B)** The RPR comparison of LFBs obtained by each method. The connecting line between the boxes is the mean connecting line.

identified subsets of conditions, so the conditions contained in the four selected LFBs are explored in detail, as shown in **Supplementary Table 2**. The methylation level heatmaps of the four selected LFBs are shown in **Figure 3**. For the KEGG pathway analysis, six KEGG pathways known to be regulated by RNA methylation were selected (Dominissini et al., 2012; Xiang et al., 2017), such as apoptosis, DNA repair, fatty acid metabolism, etc. Then, Fisher's exact test was used to verify whether each LFB was significantly enriched in some specific pathways. The output *p*-value shows the correlation between the four LFBs obtained and six biological pathways, as well as the importance of multiple hypothesis correction. We could see from **Supplementary Table 3** that the four selected LFBs are significantly enriched in the ultraviolet (UV) response up. Although the enrichment degree of LFB2 is lower than that of the other three LFBs in the UV response up, its enrichment in the apoptosis is significantly higher than that of the other three LFBs, indicating that LFB2 may further affect apoptosis through some other m^{6}A-related pathways. Besides, LFB1, LFB3, and LFB4 are also significantly enriched in

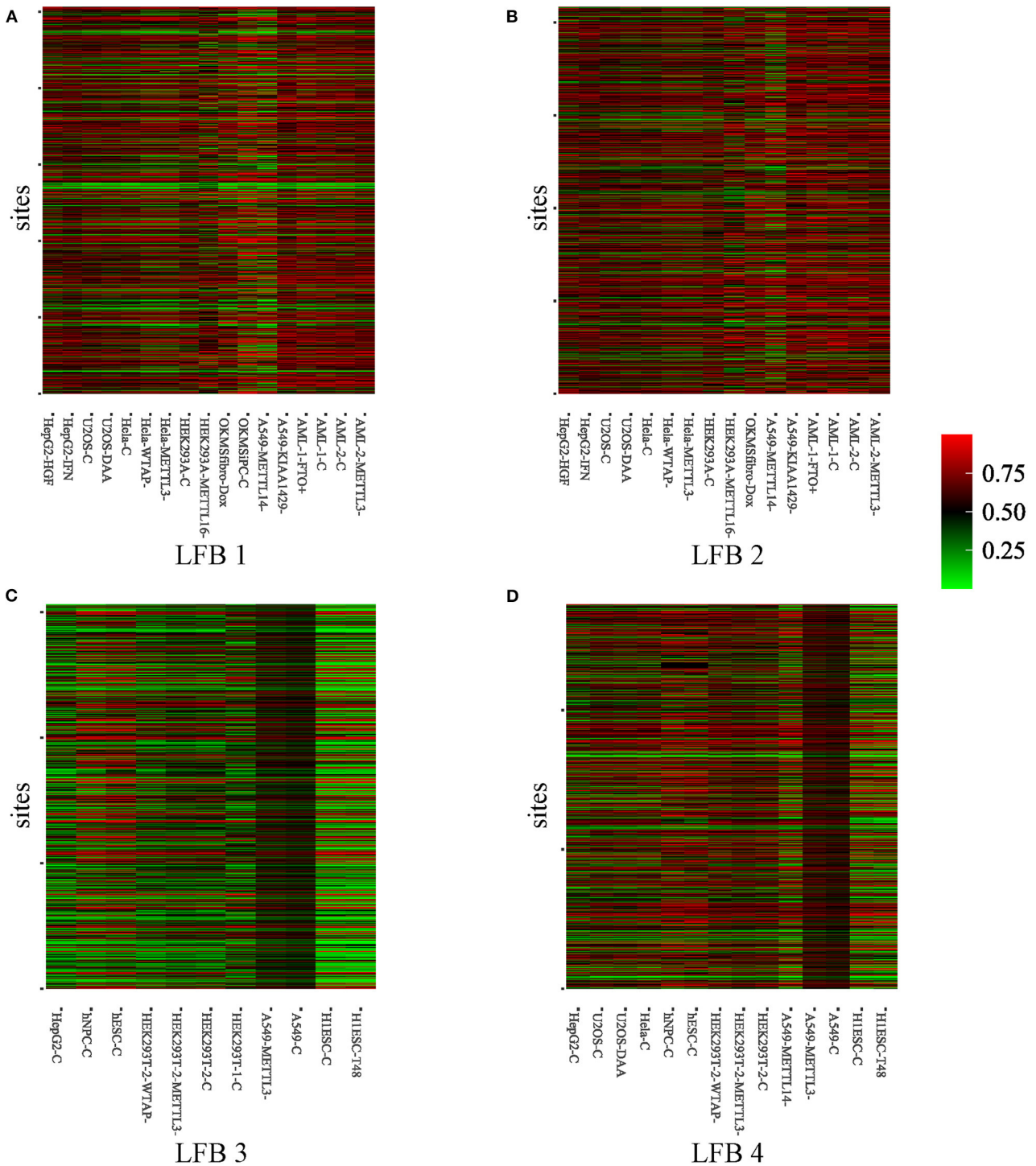

FIGURE 3 | Heatmaps of methylation level of the four selected LFBs. **(A)** The methylation level heatmap of the LFB1. **(B)** The methylation level heatmap of the LFB2. **(C)** The methylation level heatmap of the LFB3. **(D)** The methylation level heatmap of the LFB4.

DNA repair, which may be related to DNA damage caused by ultraviolet radiation. Since m^6A has been proved to be related to stem cell differentiation and cancer progression (Batista et al., 2014), there is a reasonable explanation for enriching LFB1 and LFB3 in fatty acid metabolism. As the main components of neutral fat, phospholipids and glycolipids, fatty acids can meet various body needs and regulate metabolism, growth and development (Azain, 2004). The p53 pathway enriched

in LFB4 indicates that LFB4 may be related to stress signal, regulation of intracellular homeostasis, chromosome segregation, and cell division (Harris and Levine, 2005). Through the above analysis, it is not difficult to see that the LFBs obtained by REW-ISA V2 have more significant biological significance than the randomly selected LFB. Therefore, an in-depth analysis of the LFBs obtained may help reveal the specific regulatory mechanism of m^6A.

To check whether the detected LFBs have biological significance, we further conducted the enzymes substrate specificity experiments on the four selected LFBs. Since LFB covers hyper-methylated sites and conditions, the sites and conditions involved in each LFB are more likely to be the target sites of m^6A methyltransferases. Therefore, we studied the association between each selected LFB and four m^6A methyltransferases, including METTL3, METTL14, WTAP as well as KIAA1429. For this purpose, 38,845 METTL3 targeted gene sites, 19,099 METTL14 targeted gene sites, 35,144 WTAP targeted gene sites, and 1,784 KIAA1429 targeted gene sites included in the real data were first identified by TREW tool (Liu et al., 2018). After REW-ISA V2, we summarized the distribution of target RNA methylation sites involved in each LFB (**Supplementary Table 4**). Then, the association between the sites in each selected LFB and m^6A methyltransferases target sites was further evaluated by Fisher's exact test. The experimental enrichment results are shown in **Supplementary Table 5**, where *p*-value indicates the significance of association between sites and methyltransferase target sites. The results showed that the sites contained in the four selected LFBs were significantly enriched in the target sites of the four methyltransferases. This means that under specific conditions, the LFBs obtained by REW-ISA V2 were indeed the collaboratively hyper-methylated sites, which will help biologists to further study the specific regulation mechanism of m^6A. The detailed analysis process and results can be obtained in the **Supplementary Materials**.

DISCUSSION

Although more and more studies have shown that the modification of m^6A in RNA is related to many important biological functions, the specific regulatory mechanism of m^6A is still unclear. To quickly and effectively predict potential functional m^6A sites from the epi-transcriptome data, it is important to develop some computational algorithms, which will help us have a more comprehensive understanding of m^6A-related life processes. Based on REW-ISA, in this article, we developed REW-ISA V2 to better reveal the potential local co-methylation patterns across subsets of conditions. REW-ISA V2 was implemented on the real MeRIP-seq data, and a total of 15 LFBs were obtained. Further comparison and analysis show that, compared with other biclustering algorithms, the LFBs obtained by REW-ISA V2 has more significant biological significance.

REW-ISA V2 could obtain reliable biclustering patterns because of the use of homologous information. More specifically, the sites' methylation levels corresponding to the same gene will show a similar trend with a high probability. Similarly, conditions derived from the same cell line will have similar trends in all sites. Therefore, the rational use of homologous information will help to better mine local co-methylation patterns. Of course, REW-ISA V2 still has some deficiencies that need to be improved in the future. First of all, REW-ISA V2 uses simple multiplication to fuse homologous information, which inevitably introduces noise at the same time. Secondly, because the database on which GO analysis depends is incomplete, the enrichment constraint framework designed is prone to human error. Finally, the enrichment constraint framework built into REW-ISA V2 usually takes a long time. In the future, we will use BSig (Henriques and Madeira, 2018) to better evaluate the obtained LFBs and develop a new computational model to overcome these limitations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LZ and SC built the architecture for REW-ISA V2, designed and implemented the experiments, analyzed the result, and wrote the paper. JM conducted the experiments, analyzed the result, and revised the paper. ZL and HL supervised the project, analyzed the result, and revised the paper. All authors read, critically revised, and approved the final manuscript.

FUNDING

This work has been supported by Fundamental Research Funds for the Central Universities (Grant No. 2019ZDPY15 to LZ). The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.654820/full#supplementary-material

REFERENCES

Azain, M. (2004). Role of fatty acids in adipocyte growth and development. *J. Anim. Sci.* 82, 916–924. doi: 10.1093/ansci/82.3.916

Barbieri, I., Tzelepis, K., Pandolfini, L., Shi, J., Millán-Zambrano, G., Robson, S. C., et al. (2017). Promoter-bound METTL3 maintains myeloid leukaemia by m6A-dependent translation control. *Nature* 552, 126–131. doi: 10.1038/nature24678

Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., et al. (2014). m6A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* 15, 707–719. doi: 10.1016/j.stem.2014.09.019

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E* 67:031902. doi: 10.1103/PhysRevE.67.031902

Chai, R., Wu, F., Wang, Q., Zhang, S., Zhang, K., Liu, Y., et al. (2019). m6A RNA methylation regulators contribute to malignant progression and have clinical prognostic impact in gliomas. *Aging* 11, 1204–1225. doi: 10.18632/aging.101829

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47:e41. doi: 10.1093/nar/gkz074

Chen, S., Zhang, L., Lu, L., Meng, J., and Liu, H. (2021). FBCwPlaid: a functional bi-clustering analysis of epi-transcriptome profiling data via a weighted plaid model. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2021.3049366

Deng, X., Su, R., Feng, X., Wei, M., and Chen, J. (2018). Role of N6-methyladenosine modification in cancer. *Curr Opin. Genet. Dev.* 48, 1–7. doi: 10.1016/j.gde.2017.10.005

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Frye, M., Harada, B. T., Behm, M., and He, C. (2018). RNA modifications modulate gene expression during development. *Science* 361, 1346–1349. doi: 10.1126/science.aau1646

Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.* 15, 293–306. doi: 10.1038/nrg3724

Fustin, J.-M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., et al. (2013). RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* 155, 793–806. doi: 10.1016/j.cell.2013.10.026

Harris, S. L., and Levine, A. J. (2005). The p53 pathway: positive and negative feedback loops. *Oncogene* 24, 2899–2908. doi: 10.1038/sj.onc.1208615

Henriques, R., Antunes, C., and Madeira, S. C. (2015). A structured view on pattern mining-based biclustering. *Pattern Recognit.* 48, 3941–3958. doi: 10.1016/j.patcog.2015.06.018

Henriques, R., Ferreira, F. L., and Madeira, S. C. (2017). BicPAMS: software for biological data analysis with pattern-based biclustering. *BMC Bioinformatics* 18:82. doi: 10.1186/s12859-017-1493-3

Henriques, R., and Madeira, S. C. (2018). BSig: evaluating the statistical significance of biclustering solutions. *Data Min. Knowl. Discov.* 32, 124–161. doi: 10.1007/s10618-017-0521-2

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., et al. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* 7, 885–887. doi: 10.1038/nchembio.687

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Lachén-Montes, M., González-Morales, A., De Morentin, X. M., Pérez-Valderrama, E., Ausín, K., Zelaya, M. V., et al. (2016). An early dysregulation of FAK and MEK/ERK signaling pathways precedes the β-amyloid deposition in the olfactory bulb of APP/PS1 mouse model of Alzheimer's disease. *J. Proteom.* 148, 149–158. doi: 10.1016/j.jprot.2016.07.032

Lazzeroni, L., and Owen, A. (2002). Plaid models for gene expression data. *Stat. Sinica* 12, 61–86. doi: 10.1109/ITW.2002.1115477

Li, L., Guo, Y., Wu, W., Shi, Y., Cheng, J., and Tao, S. (2012). A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. *BioData Min.* 5:8. doi: 10.1186/1756-0381-5-8

Li, Z., Weng, H., Su, R., Weng, X., Zuo, Z., Li, C., et al. (2017). FTO plays an oncogenic role in acute myeloid leukemia as a N6-methyladenosine RNA demethylase. *Cancer Cell* 31, 127–141. doi: 10.1016/j.ccell.2016.11.017

Lin, S., Choe, J., Du, P., Triboulet, R., and Gregory, R. I. (2016). The m6A methyltransferase METTL3 promotes translation in human cancer cells. *Mol. Cell* 62, 335–345. doi: 10.1016/j.molcel.2016.03.021

Liu, H., Wang, H., Wei, Z., Zhang, S., Hua, G., Zhang, S., et al. (2018). MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res.* 46, D281–D287. doi: 10.1093/nar/gkx1080

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., et al. (2014). A METTL3–METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* 10, 93–95. doi: 10.1038/nchembio.1432

Liu, L., Zhang, S., Zhang, Y., Liu, H., Zhang, L., Chen, R., et al. (2015). Decomposition of RNA methylome reveals co-methylation patterns induced by latent enzymatic regulators of the epitranscriptome. *Mol. BioSyst.* 11, 262–274. doi: 10.1039/C4MB00604F

Liu, X., Li, D., Liu, J., Su, Z., and Li, G. (2020). RecBic: a fast and accurate algorithm recognizing trend-preserving biclusters. *Bioinformatics* 36, 5054–5060. doi: 10.1093/bioinformatics/btaa630

Meng, J., Gao, S., and Huang, Y. (2009). Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics* 25, 1521–1527. doi: 10.1093/bioinformatics/btp235

Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., et al. (2014). A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods* 69, 274–281. doi: 10.1016/j.ymeth.2014.06.008

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., et al. (2015). 5′ UTR m6A promotes cap-independent translation. *Cell* 163, 999–1010. doi: 10.1016/j.cell.2015.10.012

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Murali, T., and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.* 8, 77–88. doi: 10.1142/9789812776303_0008

Orzechowski, P., Pańszczyk, A., Huang, X., and Moore, J. H. (2018a). runibic: a Bioconductor package for parallel row-based biclustering of gene expression data. *Bioinformatics* 34, 4302–4304. doi: 10.1093/bioinformatics/bty512

Orzechowski, P., Sipper, M., Huang, X., and Moore, J. H. (2018b). EBIC: an evolutionary-based parallel biclustering algorithm for pattern discovery. *Bioinformatics* 34, 3719–3726. doi: 10.1093/bioinformatics/bty401

Pendleton, K. E., Chen, B., Liu, K., Hunter, O. V., Xie, Y., Tu, B. P., et al. (2017). The U6 snRNA m6A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell* 169, 824–835.e14. doi: 10.1016/j.cell.2017.05.003

Ping, X., Sun, B., Wang, L., Xiao, W., Yang, X., Wang, W., et al. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res.* 24, 177–189. doi: 10.1038/cr.2014.3

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., et al. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22, 1122–1129. doi: 10.1093/bioinformatics/btl060

Schwartz, S., Mumbach, M. R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G. G., et al. (2014). Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites. *Cell Rep.* 8, 284–296. doi: 10.1016/j.celrep.2014.05.048

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Wang, Z., Li, G., Robinson, R. W., and Huang, X. (2016). UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data. *Sci. Rep.* 6:23466. doi: 10.1038/srep23466

Wu, X., Wei, Z., Chen, K., Zhang, Q., Su, J., Liu, H., et al. (2019). m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network. *BMC Bioinformatics* 20:223. doi: 10.1186/s12859-019-2840-3

Xiang, Y., Laurent, B., Hsu, C.-H., Nachtergaele, S., Lu, Z., Sheng, W., et al. (2017). RNA m6A methylation regulates the ultraviolet-induced DNA damage response. *Nature* 543, 573–576. doi: 10.1038/nature21671

Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., et al. (2020). QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics* 36, 1143–1149. doi: 10.1093/bioinformatics/btz692

Yu, G., Wang, L., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, C., Chen, Y., Sun, B., Wang, L., Yang, Y., Ma, D., et al. (2017). m6A modulates haematopoietic stem and progenitor cell specification. *Nature* 549, 273–276. doi: 10.1038/nature23883

Zhang, C., Fu, J., and Zhou, Y. (2019a). A review in research progress concerning m6A methylation and immunoregulation. *Front. Immunol.* 10:922. doi: 10.3389/fimmu.2019.00922

Zhang, L., Chen, S., Zhu, J., Meng, J., and Liu, H. (2020). REW-ISA: unveiling local functional blocks in epi-transcriptome profiling data via an RNA expression-weighted iterative signature algorithm. *BMC Bioinformatics* 21:447. doi: 10.1186/s12859-020-03787-w

Zhang, L., He, Y., Wang, H., Liu, H., Huang, Y., Wang, X., et al. (2019b). Clustering count-based RNA methylation data using a nonparametric generative model. *Curr. Bioinform.* 14, 11–23. doi: 10.2174/1574893613666180601080008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.