Zhang *et al. BMC Bioinformatics* (2020) 21:447
https://doi.org/10.1186/s12859-020-03787-w

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

REW-ISA: unveiling local functional blocks in epi-transcriptome profiling data via an RNA expression-weighted iterative signature algorithm

Lin Zhang[1,2], Shutao Chen[1,2], Jingyi Zhu[1,2], Jia Meng[3] and Hui Liu[1,2*]

*Correspondence:
hui.liu@cumt.edu.cn
[1] Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China
Full list of author information is available at the end of the article

Abstract

Background: Recent studies have shown that N^6-methyladenosine (m^{6}A) plays a critical role in numbers of biological processes and complex human diseases. However, the regulatory mechanisms of most methylation sites remain uncharted. Thus, in-depth study of the epi-transcriptomic patterns of m^{6}A may provide insights into its complex functional and regulatory mechanisms.

Results: Due to the high economic and time cost of wet experimental methods, revealing methylation patterns through computational models has become a more preferable way, and drawn more and more attention. Considering the theoretical basics and applications of conventional clustering methods, an RNA Expression Weighted Iterative Signature Algorithm (REW-ISA) is proposed to find potential local functional blocks (LFBs) based on MeRIP-Seq data, where sites are hyper-methylated or hypo-methylated simultaneously across the specific conditions. REW-ISA adopts RNA expression levels of each site as weights to make sites of lower expression level less significant. It starts from random sets of sites, then follows iterative search strategies by thresholds of rows and columns to find the LFBs in m^{6}A methylation profile. Its application on MeRIP-Seq data of 69,446 methylation sites under 32 experimental conditions unveiled 6 LFBs, which achieve higher enrichment scores than ISA. Pathway analysis and enzyme specificity test showed that sites remained in LFBs are highly relevant to the m^{6}A methyltransferase, such as METTL3, METTL14, WTAP and KIAA1429. Further detailed analyses for each LFB even showed that some LFBs are condition-specific, indicating that methylation profiles of some specific sites may be condition relevant.

Conclusions: REW-ISA finds potential local functional patterns presented in m^{6}A profiles, where sites are co-methylated under specific conditions.

Keywords: m^{6}A methylation, Iterative signature algorithm, Biclustering

Background

N^6-methyladenosine (m^{6}A), which refers to the methylation of the adenosine bases at the nitrogen-6 position, is the most abundant post-transcriptional modification present in mRNAs and long non-coding RNAs [1]. It has been found to function in various pathways related to mRNA stability [2], DNA damage [3], differentiation [4], circadian clock [5], neurogenesis [6], immunity [7], anti-tumor activity [7], learning and memory [8], sex determination [9], heat shock response [10], etc. With the emergence of high-throughput sequencing technologies, Methylated Immunoprecipitation sequencing (MeRIP-Seq, or m^{6}A-seq) [11, 12], researchers have been able to examine the dynamics and various functions of m^{6}A in human, mouse, yeast, rice and other species [2, 5, 13–15].

The m^{6}A methylation has been found to be governed or mediated by relevant enzymes, i.e., writers (METTL3/METTL14/WTAP complex [15, 16], KIAA1429 [17], VIRMA [18], RBM15 [19], ZC3H13 [20], etc.,), erasers (FTO [21], ALKBH5 [22], etc.,) as well as readers (YTH family [2, 10, 23–25], IGF2BP1-3 [26], eIF3 [27], etc.,). However, due to the complexity of life, the detailed regulatory circuit of RNA methylate remains uncharted, and it is believed to be more complex than enzyme induced mechanisms.

Till this day, several clustering methods have been proposed to identify co-methylation patterns in MeRIP-Seq data, trying to elucidate the functional mechanisms of m^{6}A methylation. Liu et al. used four different clustering approaches, such as K-means, hierarchical clustering, Bayesian factor regression model as well as nonnegative matrix factorization to unveil the co-methylation patterns [15, 28]. To our knowledge, they revealed the linkage between the global co-methylation patterns embedded in epi-transcriptomic data for the first time. Cui et al. proposed MeTCluster to uncover the potential patterns of m^{6}A methylation. It utilized a hierarchical graphical model to depict the reads counts, suggesting m^{6}A functions could be location specific [29]. We have previously proposed an infinite beta binomial mixture model based on Dirichlet Process (DPBBM) to unveil the co-methylation patterns embedded in MeRIP-Seq data [30]. All the above-mentioned methods focused on clustering methylation sites under all conditions. Current studies have shown that on average 3–5 m^{6}A RNA methylation sites position on each mRNA in human genome [31, 32]. It is conceivable that some specific sites may co-methylate under a subset of experimental conditions. Thus, clustering of sites over all conditions may miss biological meaningful information. On the one hand, sites sharing the same regulatory factor are more likely to co-methylate together; on the other hand, sites residing on the genes that belong to the same pathway may exhibit co-methylation patterns over subsets of experiments. Therefore, we aim to find some local functional blocks (LFBs), where sites are hyper-methylated or hypo-methylated simultaneously across the specific conditions in the same LFB, to unveil the local function patterns in m^{6}A methylation profile.

Biclustering methods have been widely used to identify co-expressed genes under subsets of conditions in large scale microarray data [33–37]. Ihmels proposed an iterative signature algorithm (ISA) [34] to seek for biclusters, where subsets of co-regulated genes and conditions were selected by iterative searching procedure [35]. Murali et al. proposed Xmotifs algorithm, which takes a discretized gene expression matrix as input, to find co-expression patterns, where genes share the same expression level [36]. Prelić et al. proposed a Bimax method, which takes a binarized gene expression matrix as input, to find potential co-expression patterns [37]. The preprocess of discretization of input data results in serious information loss. When profiled by MeRIP-Seq technology, the quantification of RNA methylation level needs to be estimated from two complementary integer measurements indicating site reads count from input and IP samples. Conventionally, m^{6}A methylation level is achieved by simple division operation, which calculates the percentage of site reads in IP sample over the total site reads of input and IP samples. However, it is not always accurate. Even if sites show the same percentage in value, their methylation levels maybe quite different due to their different RNA expression level. To be more specific, if the RNA expression level is very low, there may exist noise, which makes the percentage less confident. Therefore, we proposed herein an RNA Expression Weighted Iterative Signature Algorithm (REW-ISA), which adopted RNA expression level as weight to weaken the confidence sites, then followed an iterative search strategy through rows and columns to seek for LFBs. During the LFB searching strategy, each potential LFB is identified by column threshold (defined as T_C) and row threshold (defined as T_R). T_C and T_R are updated automatically according to Standard Deviation within Clusters (SDwC) and Average Similarity within Clusters (ASwC) metrics iteratively. SDwC indicates the closeness of each element in each LFB while ASwC indicates the correlation of each condition pair in each LFB.

REW-ISA was implemented on simulated data as well as real MeRIP-Seq induced m^{6}A methylation level matrix to find potential LFBs. On simulated data, Score of Bi-Clustering (SoBC) metric was followed to evaluate the identification performance of LFBs. On real data, Gene Ontology (GO) analysis and enzyme specificity test were in the next conducted to validate the identified LFBs. As a result, REW-ISA can find LFBs that cover collaboratively hyper-methylated sites under specific conditions.

Results

Performance evaluation

In this study, we applied ISA as well as REW-ISA for simulated data biclustering for performance comparison. As is known, intersection over union (IoU) is a widely used evaluation metric in object detection, which is define as

$$\mathrm{IoU} = \frac{A_o}{A_U} \tag{1}$$

where A_O represents the intersection between the obtained LFBs and ground truth, while A_U indicates the union of the obtained LFBs and ground truth. For example, suppose there are s LFBs embedded in simulated data, and n LFBs are obtained by clustering algorithm. In addition, let $\boldsymbol{G} = \{g_1, \ldots, g_s\}$ indicate whether there is uncovered LFB matching the s real LFBs respectively. At initialization, all elements in $\boldsymbol{G}$ are 0. So, we can calculate the IoU between each obtained LFB and the real LFBs. To be more specific, the IoU metric for the i-th obtained LFB and real one is achieved, and its maximum value is regarded as the final score of the i-th LFB, which is indicated by IoU_{id}, representing the i-th uncovered LFB matches the d-th real LFB best. Thus, $g_d = 1$. For all the n identified LFBs, the average of IoU_{id} with $i = 1$, ..., n, indicated as IoU_{mean} hereafter can be achieved. Since the number of obtained LFBs may differ from real, IoU_{mean} metric may not be sufficient for performance evaluation. Therefore, SoBC is defined to evaluate the agreement between REW-ISA obtained LFBs and ground truth.

$$\mathrm{SoBC} = \frac{r}{\max(s, n)} \mathrm{IoU}_{mean} \tag{2}$$

where r indicates the number of ones in $\boldsymbol{G}$. Thus, $r \leq \min(s, n)$. As SoBC approaches 1, the performance of biclustering is better.

In REW-ISA clustering procedure, T_C and T_R are key parameters for clustering stringency, and SDwC, ASwC scores are introduced to determine suitable T_C and T_R. The mean and standard deviation of each LFB are combined in SDwC by (3).

$$\mathrm{SDwC} = \frac{\sqrt{\sum\limits_{k=1}^{N} \frac{1}{m_k \cdot n_k} \sum\limits_{i=1}^{m_k} \sum\limits_{j=1}^{n_k} \left(w_{kij} p_{kij} - \overline{\boldsymbol{W}_k \boldsymbol{P}_k}\right)^2}}{N} \tag{3}$$

where N indicates the number of algorithm obtained LFBs, $\boldsymbol{P_k}$ is the m^{6}A methylation level of the k-th LFB, $\boldsymbol{W_k}$ is the RNA expression weight for k-th LFB, m_k and n_k are the number of sites and conditions in the k-th LFB, w_{kij} is the RNA expression level of the i-th site under condition j in the k-th LFB, p_{kij} is the methylation level of the i-th site under condition j in the k-th LFB, $\overline{\boldsymbol{W_k P_k}} = (1/m_k n_k) \sum_{i=1}^{\boldsymbol{m_k}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{n_k}} \boldsymbol{w_{kij} p_{kij}}$. Thus, SDwC represents the standard deviation of methylation levels in each LFB.

ASwC is regarded as another concern for T_C and T_R selection. The pearson correlation between condition a and b in the k-th LFB is first calculated as r_{kab},

$$r_{kab} = \frac{\sum_{t=1}^{m_k} \left[\left(w_{kta} p_{kta} - \overline{\boldsymbol{W_{ka} P_{ka}}} \right) \left(w_{ktb} p_{ktb} - \overline{\boldsymbol{W_{kb} P_{kb}}} \right) \right]}{\sqrt{\sum_{t=1}^{m_k} \left(w_{kta} p_{kta} - \overline{\boldsymbol{W_{ka} P_{ka}}} \right)^2} \sqrt{\sum_{t=1}^{m_k} \left(w_{ktb} p_{ktb} - \overline{\boldsymbol{W_{kb} P_{kb}}} \right)^2}} \tag{4}$$

where $\boldsymbol{W_{ka}}$ and $\boldsymbol{W_{kb}}$ represent the RNA expression level under condition a and b in the k-th LFB, $\boldsymbol{P_{ka}}$ and $\boldsymbol{P_{kb}}$ represent the RNA methylation level under condition a and b, $\overline{\boldsymbol{W_{ka} P_{ka}}} = (1/m_k) \sum_{i=1}^{\boldsymbol{m_k}} \boldsymbol{w_{kia} p_{kia}}$, $\overline{\boldsymbol{W_{kb} P_{kb}}} = (1/m_k) \sum_{i=1}^{\boldsymbol{m_k}} \boldsymbol{w_{kib} p_{kib}}$. Then, ASwC is defined as

$$\mathrm{ASwC} = \frac{\sqrt{\sum_{k=1}^{N} \frac{2}{n_k (n_k - 1)} \sum_{a=1}^{n_k} \sum_{b=1, b \neq a}^{n_k} r_{kab}}}{N} \tag{5}$$

where n_k is the number of conditions in the k-th LFB. Thus, ASwC indicates how the involved sites co-methylate between conditions in each LFB.

Our original intention is to better reveal the biological functional mechanisms of the co-methylation modules based on transcriptome data, larger SDwC and smaller ASwC metrics are preferred to get larger LFBs with more implicit information.

Simulated data

For performance evaluation, a simulated RNA methylation dataset of size 1000×15 was generated from a mixture of 4 beta-binomial distributions, corresponding to three biclustering blocks and the background (Fig. 1a). The overall distribution characteristics of the simulated data were set similar to that of the real MeRIP-Seq data (Fig. 1b,

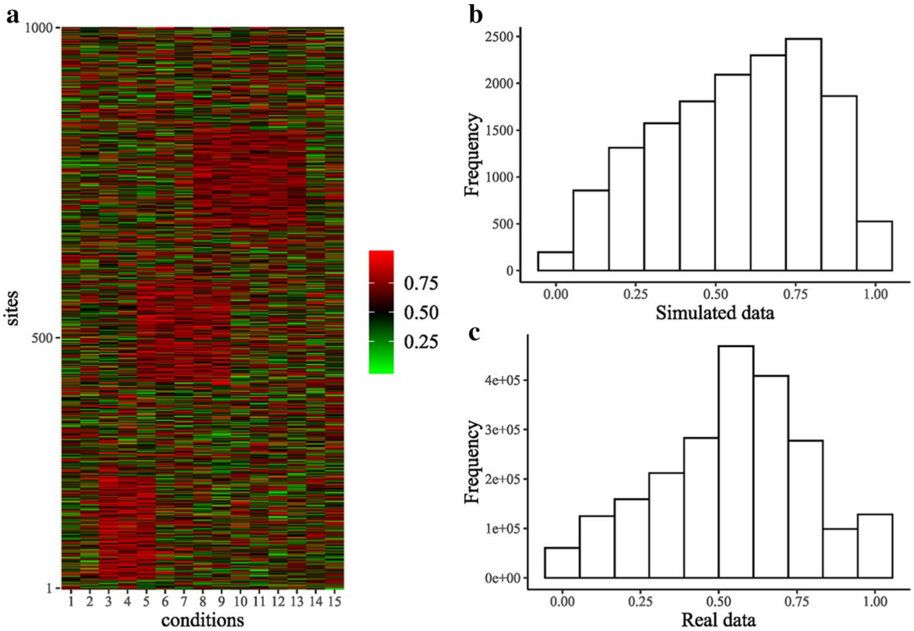

Fig. 1 Comparison of statistical characteristics between simulated data and real data. **a** Heatmap of simulated data. Rows are corresponding to m^{6}A sites while columns represent conditions. **b** Histogram of simulated data. **c** Histogram of real data

c) to mimic real scenarios. The matched methylation expression data, which is used as the "weight" in the proposed algorithm, was directly calculated from the simulated RNA methylation dataset as previously described.

As is known, T_C and T_R are defined for subset selection along columns and rows in conventional ISA. They are also defined in REW-ISA, and play a decisive role in LFB stringency. Since the methylation level matrix is $p \gg n$, it is intuitive that the conditions should be selected more carefully, thus T_C asks for more cautious exercise, and detailed explanations are given in method section.

In REW-ISA, a grid search method was followed for parameter optimization. The range of T_R is 0.1–5 with step size 0.1, and the range of T_C is 0.05–3 with step size 0.05. Their upper bounds are recommended a large value, then adapted automatically in following procedures. For each $\{T_R, T_C\}$ setting, 40 experiments were conducted to ensure the robustness. After repeated experiments, the mode of LFB number was adopted as final number of LFBs for each threshold pair, as shown in Fig. 2.

According to Fig. 2, the range of T_R and T_C can be shrunk. To be more specific, the thresholds that obtain the largest number of LFBs were first reached, then filter out all the combinations with larger T_R or T_C. This is because, in the abandoned combinations, REW-ISA can also get the same number of LFBs with smaller T_R and T_C. However, smaller T_R and T_C remain more rows and columns in each LFB, which may unveil more useful biological information. Thus, the range of T_R becomes 0.1–2.5, while the range of T_C becomes 0.05–1.15.

We can see from Fig. 2 that when T_R and T_C are close to their lower limits, only a few LFBs can be achieved with very large scale, which cannot uncover implicit information for functions. Thus, REW-ISA raises the lower bound of T_R and T_C appropriately. Suppose the

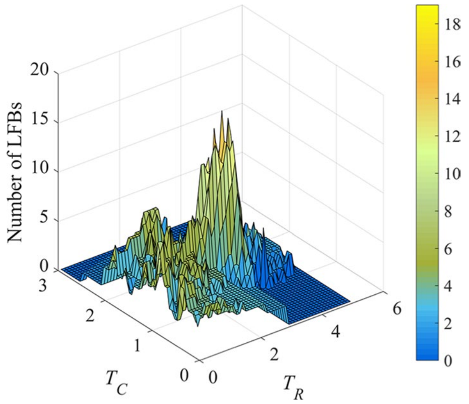

Fig. 2 The number of LFBs identified by REW-ISA with T_R set from 0.1 to 5 at step size of 0.1, T_C set from 0.05 to 3 at step size of 0.05. The color scale indicates the number of LFBs obtained

matrix of LFB number obtained under different threshold setting is $\boldsymbol{L} \in \mathbb{R}^{rn\times cn}$ with T_R and T_C adjusted, where rn represents the number of T_Rs and cn represents the number of T_Cs considered. Min–max normalization of $\boldsymbol{L}$ is performed to obtain $\boldsymbol{L^{norm}} \in \mathbb{R}^{rn\times cn}$, and then the variance of each row and column in $\boldsymbol{L^{norm}}$ is calculated. The variance of the i-th row in $\boldsymbol{L^{norm}}$ is vr_i $(1\leq i\leq rn)$, and the variance of the j-th column is vc_j $(1\leq j\leq cn)$.

$$vr_i = \frac{\sum_{j=1}^{cn}\left(l_{ij}^{norm} - \frac{1}{cn}\sum_{j=1}^{cn} l_{ij}^{norm}\right)^2}{cn} \tag{6}$$

$$vc_j = \frac{\sum_{i=1}^{rn}\left(l_{ij}^{norm} - \frac{1}{rn}\sum_{i=1}^{rn} l_{ij}^{norm}\right)^2}{rn} \tag{7}$$

Furthermore, the mean values of elements in $\boldsymbol{vr}$ and $\boldsymbol{vc}$ are calculated as vr_{mean} and vc_{mean}, respectively. We set $i' = \min\{i : vr_i \geq vr_{mean}\}$ and $j' = \min\{j : vc_j \geq vc_{mean}\}$, and then drop out the first $i'-1$ values of the T_R and the first $j'-1$ values of the T_C from consideration. Thus, the range of T_R further becomes 0.3–2.5, while the range of T_C becomes 0.1–1.15.

After shrinking the range of T_R and T_C, the matrix of LFB numbers under each threshold pair setting is updated to be $\boldsymbol{L'} \in \mathbb{R}^{rn'\times cn'}$, where $rn' = rn - i' + 1$ and $cn' = cn - j' + 1$. Then, within the selected threshold range of T_R and T_C, a sliding window of size $\eta_r \times \eta_c$ is used to help find more stable selections of T_R and T_C. The value of η_r and η_c are selected by Eqs. (8) and (9),

$$\eta_r = 2 \times \left\lceil \frac{0.1}{step_r} \right\rceil + 1 \tag{8}$$

$$\eta_c = 2 \times \left\lceil \frac{0.1}{step_c} \right\rceil + 1 \tag{9}$$

Table 1 Statistics of the number of LFBs obtained in the simulated data

ln	2	3	4	5	6	7	9	14
f_{ln}	0.133	**0.445**	0.244	0.093	0.057	0.017	0.003	0.008

The bold value in the table is the maximum value of the frequency

ln represents the number of LFBs obtained; and f_{ln} represents the frequency at which the number of LFBs is *ln*

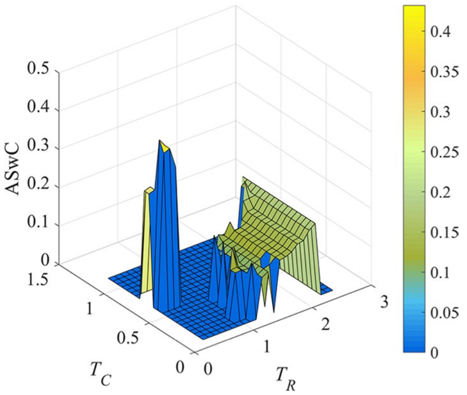

Fig. 3 ASwC metrics of LFBs identified by REW-ISA, the number of LFBs obtained by each threshold pair is 3. The color scale indicates the value of ASwC

where $step_r$ represents the variable step size of T_R, $step_c$ represents that of T_C, and $\lceil\cdot\rceil$ is round up to integer operation. The sliding window is obvious to cover odd number of rows and columns, which makes the thresholds value of interest locate in the center of the sliding window. Specifically, for the element l'_{ij} locating in the *i*-th $(1 \leq i \leq rn')$ row and the *j*-th $(1 \leq j \leq cn')$ column of $\boldsymbol{L'}$, the mode of the values covered by sliding window is calculated, then compared to the center value l'_{ij}. If they are equal, the threshold setting is maintained for further consideration, and $ls'_{ij} = 1$ is recorded in the matrix $\boldsymbol{LS} \in \mathbb{R}^{rn' \times cn'}$. Otherwise, $ls'_{ij} = 0$. It is worth noting that when sliding the elements on the boundary of $\boldsymbol{L'}$, we only select the effective elements in the sliding window to test the stability.

Through $\boldsymbol{L'}$ and the stable score matrix $\boldsymbol{LS}$, the threshold pairs with stable LFB number can be screened out. Let $\boldsymbol{S}=\boldsymbol{L'} \times \boldsymbol{LS}$, $\boldsymbol{S} \in \mathbb{R}^{rn' \times cn'}$, the threshold pairs corresponding to non-zero elements in $\boldsymbol{S}$ are stable threshold combinations. After filtering out 0 in $\boldsymbol{S}$, the number of obtained LFBs in $\boldsymbol{S}$ are counted to provide frequency f_{ln}, where *ln* represents the number of LFBs, and the result is shown in Table 1.

According to the statistics of the number of stable LFBs, that is, Table 1, we find that the number of LFBs with the highest frequency is 3. Therefore, we can reasonably conclude that there is a total of 3 LFBs in the simulated data.

Furthermore, the threshold pairs in $\boldsymbol{LS}$ are filtered according to the number of obtained LFBs is 3. At the same time, we update the ASwC calculated under effective settings of T_R and T_C. The ASwC value obtained by effective threshold pair is shown in Fig. 3.

Because ASwC is used to measure the similarity between columns in LFBs, the smaller ASwC is, the more information is contained in each LFB. Since smaller ASwC can help retrieve more biological meaningful information, we calculate the mean of ASwC achieved under effective threshold pairs, and remain the threshold pairs with which ASwC scores are less than the mean score to further shrink the range of T_R and T_C.

Within the narrowed T_R and T_C range, the SDwC values of each threshold pair are further compared, and the result is shown in Fig. 4.

As shown in Fig. 4, the optimal value of T_R and T_C, where SDwC gets its maximum indicating the loose and information abundance of each LFB, are 1.2 and 0.35 respectively. Since smaller threshold may find the larger LFB, REW-ISA chooses smaller T_R and T_C when the maximum value of SDwC is achieved under multiple pairs of T_R and T_C.

In a word, it is suggested that the upper bound initialization of T_R and T_C be set to larger values, and REW-ISA can automatically shrink the range. Besides, the step size of T_R and T_C during grid search has no effect on final result. The REW-ISA thresholds optimization algorithm is in the following.

Algorithm 1: Thresholds optimization process of REW-ISA

```
Input: Methylation level matrix P, weight matrix W and initialize the range of $T_R$ and $T_C$
Output: The optimal $T_R$ and $T_C$, and the number of LFBs determined by the above $T_R$ and $T_C$
Step1: Run REW-ISA within the initial threshold range, obtain the threshold pair that generates the most LFBs,
    and then shrink the range of $T_R$ and $T_C$
Step2: Within the threshold range after contraction, the LFB number matrix L′, stable score matrix LS, stable LFB
    number matrix S, compactness SDwC and ASwC are calculated
Step3: Count the frequency of the number of LFBs in the matrix S
Step4: According to the maximum frequency, select the corresponding optimal number of LFBs
Step5: The ASwC value corresponding to each threshold pair is calculated, and the thresholds ranges are further
    reduced
Step6: Select the optimal $T_R$ and $T_C$ according to the maximum SDwC within the selected thresholds range
Return The optimal $T_R$ and $T_C$, and the number of LFBs determined by the optimal $T_R$ and $T_C$
```

To validate the automatic parameter selection procedure of T_R and T_C, we investigated the SoBC of ISA and REW-ISA identified LFBs with varying T_Cs (0.1 to 1.15) and T_Rs (0.3 to 2.5), as shown in Fig. 5.

As shown Fig. 5b, in REW-ISA, the optimal value of T_R and T_C locate between 1.2–1.4 and 0.1–0.65 respectively on simulated data. This is consistent with the beforementioned parameter selection procedure given in Algorithm 1. Besides, we also found that the SoBC

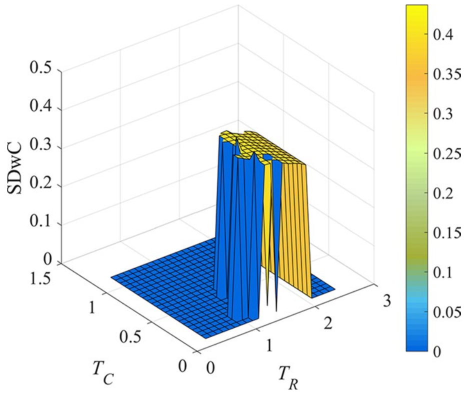

Fig. 4 SDwC metrics of LFBs identified by REW-ISA, the values of T_R and T_C are their respective shrinking threshold ranges. The color scale indicates the value of SDwC

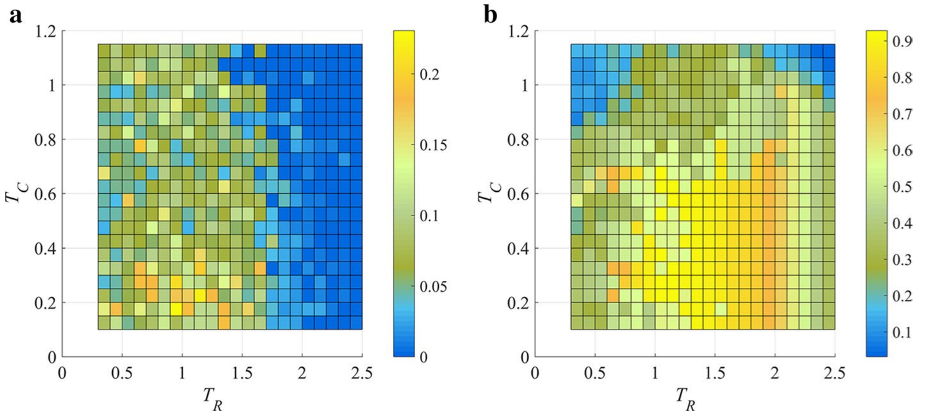

Fig. 5 Heatmap of SoBC with T_C varying from 0.05 to 1.5, and T_R varying from 0.1 to 2.5. **a** Heatmap of SoBC obtained by ISA. **b** Heatmap of SoBC obtained by REW-ISA. The color scale indicates SoBC score

metric of REW-ISA can reach around 0.9, while that of ISA is around 0.2, implying that the LFBs uncovered by REW-ISA is more effective than ISA.

Real data

A total of 69,446 human m^6A sites identified by six base-resolution mi-CLIP and m^6A-CLIP experiments were obtained by WHISTLE project [38–42]. However, mi-CLIP and m^6A-CLIP only report the positioning of m^6A sites, but do not provide the methylation level of each site. The information of methylation level still comes from MeRIP-Seq data. To be more specific, 32 samples in 10 publicly human m^6A MeRIP-Seq data sets were collected [2, 5, 11, 17, 43–47], and most of them can be retrieved from MeTDBV2.0 database [48]. The detailed description of data was given in additional file (see Additional file 1: Table S1).

As is known, MeRIP-Seq data profiles the m^6A epi-transcriptome by input and IP data. Thus, we first followed [42, 49] to quantify methylation level of each site. The biological replicates of the same cell line from the same experiment were merged, and the methylation level of the combined samples is essentially the average of all the biological replicates. All the sequencing data were downloaded in SRA format from Gene Expression Omnibus, and the reads were aligned to human reference genome hg19 with Tophat2 (with default settings as *read-mismatches* = 2, *read-gap-length* = 2, *read-edit-dist* = 2, *min-anchor* = 8, *min-intron-length* = 50 and *max-intron-length* = 500,000) for Fragments Per Kilobase of transcript per Million (FPKM) statistics [50].

The methylation level was then quantified by calculating the ratio of fold enrichment of reads in IP sample over the total of IP and input samples. To be more specific, let t_{ij} representing FPKM of the i-th site in IP sample under the j-th condition, and h_{ij} representing FPKM of the i-th site in input sample under the j-th condition. Let $\boldsymbol{P}$ indicate the methylation level matrix, the methylation level of the i-th site under the j-th condition p_{ij} can be calculated following (10).

$$p_{ij} = \frac{t_{ij}+\alpha}{t_{ij} + h_{ij}+2\alpha} \tag{10}$$

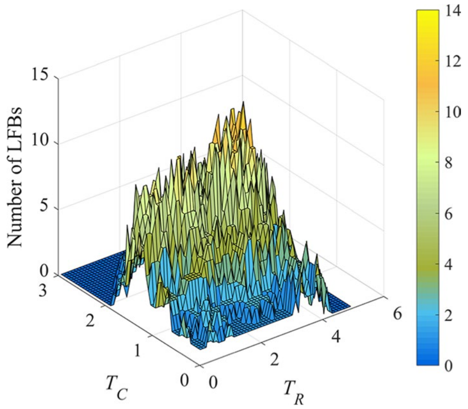

Fig. 6 The number of LFBs identified by REW-ISA with T_R ranges in 0.1–5 with step size 0.1, T_C ranges in 0.05–3 with step size 0.05. The color scale indicates the number of LFBs obtained

Table 2 Statistics of the number of LFBs obtained in the real data

ln	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
f_{ln}	0.06	0.12	0.06	0.16	**0.22**	0.08	0.08	0.07	0.09	0.05	0.01

The bold value in the table is the maximum value of the frequency

ln represents the number of LFBs obtained; and f_{ln} represents the frequency at which the number of LFBs is *ln*

where α is a very small value, aiming to avoid *NaN* where FPKM of both IP and input samples are zeros, and p_{ij} resides in (0,1).

We also constructed the weight matrix $\boldsymbol{W}$ corresponding to $\boldsymbol{P}$,

$$w_{ij} = \log(t_{ij} + h_{ij} + 1) \tag{11}$$

where 1 was added to ensure $w_{ij} \geq 0$. With the employment of $\boldsymbol{W}$, the less confident sites with lower expression level are weakened for further biclustering analysis.

Then, REW-ISA is conducted based on $\boldsymbol{P}$ and $\boldsymbol{W}$. Within the range of T_R being 0.1–5 with step size 0.1, and T_C being 0.05–3 with step size 0.05, T_R and T_C are optimized through grid search method. The experiments were repeated 10 times for each parameter setting.

As shown in Fig. 6, maximum number of LFBs is 14. The upper bounds of T_R and T_C are 3.5 and 1.35 respectively. Furthermore, the variances of each row and column in the LFB number matrix are calculated according to (6) and (7), and then the mean values of row variances and column variances are calculated respectively. The first elements which is larger than the above mean values are selected from the obtained row and column variance vectors, and the corresponding T_R and T_C are the new lower bounds of T_R and T_C. Based on the above process, the lower bounds of T_R and T_C are set to 1.3 and 0.6. The statistics of the number of LFBs obtained under different T_R and T_C are shown in Table 2. Thus, the number of LFBs is preferred to be 6.

Based on the threshold pairs that achieve 6 LFBs, ASwC scores are presented in Fig. 7.

Then, the mean of ASwC scores is calculated, and the threshold pairs that get greater ASwC than the mean is further filtered. For the remained threshold pairs, their corresponding SDwC values are calculated, as shown in Fig. 8.

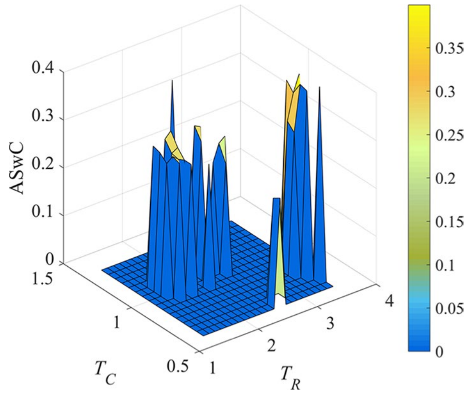

Fig. 7 ASwC metrics of LFBs identified by REW-ISA, the number of LFBs obtained by the threshold pairs is 6. The color scale indicates the value of ASwC

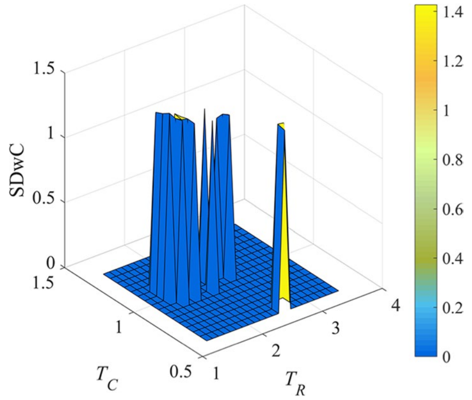

Fig. 8 SDwC metrics of LFBs identified by REW-ISA, the values of T_R and T_C are their respective shrinking threshold ranges. The color scale indicates the value of SDwC

Based on Fig. 8, T_R and T_C are selected to be 1.6 and 1 as optimal, where the largest SDwC appears.

To further explore the biological relevance of the reported six LFBs, we first annotated the Entrez Gene ID and Gene Symbol of genes corresponding to each site in each LFB, then conducted pathway and GO enrichment analysis. Six KEGG pathways known to be regulated by RNA methylation [3, 11, 21] were selected to validate whether a pathway is significantly enriched in a specific LFB using Fisher's exact test. The output *p*-value shows the significance of association between the obtained LFBs and the biological pathways with multiple hypothesis corrections.

We could see from Table 3 that obtained LFB1, LFB2 and LFB3 are significantly enriched in fatty acid metabolism. Fatty acids are a substance of the aliphatic group, and the efficacy and function of fatty acids are mainly supplemented for human absorption. Also, studies have found that fatty acids play important roles in regulating metabolism, growth and development and cell differentiation. LFB2 and LFB3 are also enriched in p53 pathway, which consists of a network responding to a variety of intrinsic and extrinsic stress signals that impact upon cellular homeostatic mechanisms, disrupting DNA replication, chromosome segregation and cell division, etc.

Table 3 Pathway analysis of REW-ISA obtained LFBs

ID	Number of sites	Enrichment statistics	KEGG pathways					
			Apoptosis	DNA repair	Fatty acid metabolism	p53 pathway	UV response down	UV response Up
LFB1	4780	OR	0.9939	0.5541	0.2156	1.5781	2.2294	1.4360
		p-value	1.0000	0.1891	**0.0113**	0.0655	**0.0032**	0.1887
		FDR	1.0000	0.3219	0.0665	0.1573	0.0437	0.3219
LFB2	4834	OR	1.0623	0.5375	0.2092	1.7047	2.3044	1.2789
		p-value	0.7558	0.1470	**0.0114**	**0.0251**	**0.0019**	0.4122
		FDR	0.8245	0.2940	0.0665	0.1006	0.0437	0.5712
LFB3	4899	OR	0.9241	0.8954	0.3044	1.8062	2.2083	1.4464
		p-value	1.0000	0.8738	**0.0336**	**0.0129**	**0.0041**	0.1967
		FDR	1.0000	0.9253	0.1209	0.0665	0.0437	0.3219
LFB4	5440	OR	0.7568	0.8152	0.4260	1.4756	1.6945	1.3718
		p-value	0.4760	0.6583	0.0531	0.0775	**0.0453**	0.2341
		FDR	0.6119	0.7406	0.1469	0.1744	0.1469	0.3664
LFB5	5713	OR	0.7447	0.8027	0.4624	1.5126	1.4158	1.2249
		p-value	0.4125	0.5729	0.0629	0.0526	0.1856	0.4756
		FDR	0.5712	0.6874	0.1573	0.1469	0.3219	0.6119
LFB6	4714	OR	0.7329	0.7878	0.2251	1.5574	2.1795	1.3776
		p-value	0.5174	0.6185	**0.0162**	0.0834	**0.0049**	0.3094
		FDR	0.6423	0.7183	0.0729	0.1766	0.0437	0.4641

The values with *p*-value less than 0.05 are shown in bold in the table

OR stands for odds ratio; *p*-value is evaluated by Fisher's exact test; and the FDR is calculated with BH method

[51, 52]. As is known, Gamma and UV irradiation could also result in DNA damage [53]. LFB6 is shown to be significantly enriched in both Fatty Acid Metabolism and UV response Down pathways, implying that LFB6 is composed of methylation sites that are relevant to life development process from fatty acid metabolism affected by UV stimulation. For LFB5, it is not enriched to any of the six KEGG pathways, indicating that LFB5 may have other implicit biological significance, and further analysis is carried out in the next.

The GO enrichment analysis was then conducted by clusterProfiler Bioconductor package [54] for each obtained LFB, with p-value cutoff set as 0.05 and q-value cutoff 0.2. For all the GO terms enriched by genes in each LFB, the negative log transform of p-value was employed as their enrichment scores.

$$s_i = -\log(p_i) \tag{12}$$

where p_i is the p-value of the i-th GO term.

In fact, GO terms with more enriched genes may not show higher enrichment scores. As to LFBs, the proportion of genes that involved in LFB is also an important factor. It is conceivable that we adhere a weight to better describe the contribution of each GO term. The weight of the i-th GO term is defined as m_i/M, where m_i is the number of genes of the i-th GO term enriched in this LFB and M is the total number of genes in this LFB. Therefore, WE_score is defined as (13).

$$\text{WE\_score} = \frac{s_1 m_1/M + s_2 m_2/M + \cdots + s_l m_l/M}{m_1/M + m_2/M + \cdots + m_l/M + m_{non}\ /\ M} \tag{13}$$

where l is the number of GO terms that enriched in each LFB, m_{non} is the number of genes covered by LFB but not enriched by any GO term. The higher the WE_score is, the more biologically significant the LFBs are [55]. Three state of the art algorithms, Xmotifs [36], Bimax [37], ISA [35] were conducted for biclustering on real data in comparison to REW-ISA. Besides, subsets with different number of sites and conditions were selected randomly as LFBs from real data. WE_score for all the obtained LFBs by all the algorithms are given in Fig. 9.

It can be seen from Fig. 9 that the REW-ISA algorithm is effective, and the result is consistent with many related research results [37, 56]. The average WE_score of LFBs inferred by REW-ISA is 13.2% higher than that of ISA, which implied more biological significance of REW-ISA. Besides, LFBs identified by the four algorithms achieve significantly higher WE_score than random one, which also indicates the biological significance.

We further examined whether the identified LFBs show enzyme's substrate specificity. 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^{6}A methyltransferases. Therefore, we investigated the association between each LFB and four m^{6}A methyltransferases, including METTL3, METTL14, WTAP as well as KIAA1429. For this purpose, 12,643 METTL3 targeted sites, 7689 METTL14 targeted sites, 13,124 WTAP-targeted sites and 399 KIAA1429 targeted RNA methylation sites were first identified by TREW [48], as shown in Table 4.

Then, the association between the sites in each LFB and m^{6}A methyltransferases target sites was further evaluated by Fisher's exact test. The reported p-value indicates the significance of association between sites and methyltransferase target sites. As shown in Table 5, all the four m^{6}A methyltransferases targeted sites in the six obtained LFBs are significantly enriched (FDR < 0.05), which means the LFBs obtained

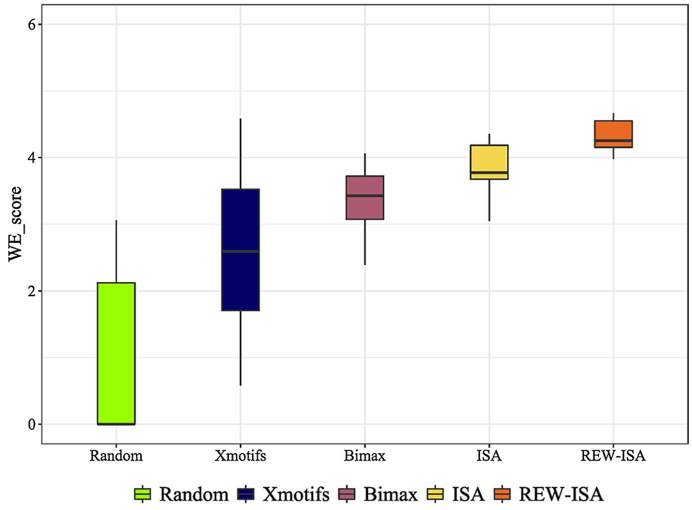

Fig. 9 WE_score of LFBs obtained by random gene groups and four biclustering algorithms

Table 4 Number of m^{6}A methyltransferase target sites in each LFB

ID	Methyltransferase component			
	METTL3	METTL14	WTAP	KIAA1429
LFB1	3605	3916	4303	267
LFB2	3728	3989	4342	287
LFB3	3888	4207	4490	296
LFB4	4432	4569	4991	345
LFB5	4712	4635	5255	344
LFB6	3728	3989	4342	287

Table 5 Enzyme specificity analysis of REW-ISA obtained LFBs

ID	Number of sites	Enrichment statistics	Methyltransferase component			
			METTL3	METTL14	WTAP	KIAA1429
LFB1	4780	OR	2.5617	14.7698	9.8933	2.4626
		p-value	5.87E−184	0	0	2.71E−33
		FDR	7.83E−184	0	0	2.71E−33
LFB2	4834	OR	2.8308	15.4638	9.6866	2.6610
		p-value	1.28E−221	0	0	1.19E−40
		FDR	1.92E−221	0	0	1.25E−40
LFB3	4899	OR	3.2549	20.2677	12.1375	2.7250
		p-value	7.73E−278	0	0	1.80E−43
		FDR	1.24E−277	0	0	2.05E−43
LFB4	5440	OR	3.7806	17.8603	12.4793	2.9442
		p-value	0	0	0	4.22E−56
		FDR	0	0	0	5.33E−56
LFB5	5713	OR	4.0817	14.6430	12.9915	2.7716
		p-value	0	0	0	1.09E−50
		FDR	0	0	0	1.30E−50
LFB6	4714	OR	2.6572	18.1924	11.2583	2.7263
		p-value	3.71E−194	0	0	2.82E−42
		FDR	5.23E−194	0	0	3.07E−42

by REW-ISA were indeed the collaboratively hyper-methylated sites under specific conditions.

In LFB1, LFB2, LFB3 and LFB6, Venn diagrams of the sites, conditions and functional annotations of genes that selected sites involved in each LFB reside on were shown in Fig. 10. As shown in Fig. 10a, it was obvious that this four LFBs contain 12,971 identical methylation sites. From the perspective of conditions, the conditions involved in LFB1 were all covered by LFB3 and LFB6, while LFB3 and LFB6 contain two conditions that were not contained by LFB1, respectively, as shown in Fig. 10b. It is also worth mentioning that for LFB2, all the conditions included in it are from human liver hepatocellular cells (HepG2) cell lines, indicating some LFBs may be condition specific. Although the conditions contained in LFB1 and LFB3 are very similar, they still contain over one hundred unshared functional annotations, which may be due to site differences between them, as shown in Fig. 10c.

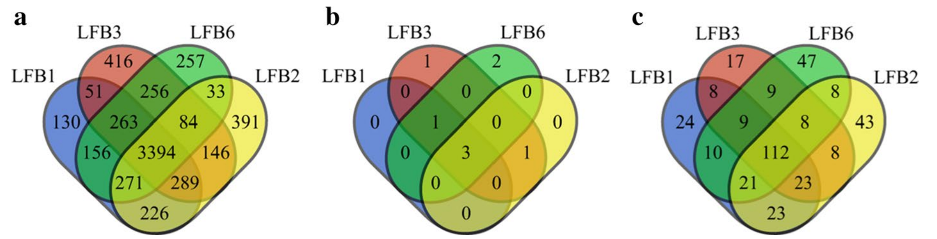

Fig. 10 Venn diagrams for obtained LFBs. **a** Venn diagrams of sites in LFB1, LFB2, LFB3 and LFB6. **b** Venn diagrams of conditions in LFB1, LFB2, LFB3 and LFB6. **c** Venn diagrams of functional annotations of genes that selected sites in LFB1, LFB2, LFB3 and LFB6 reside on

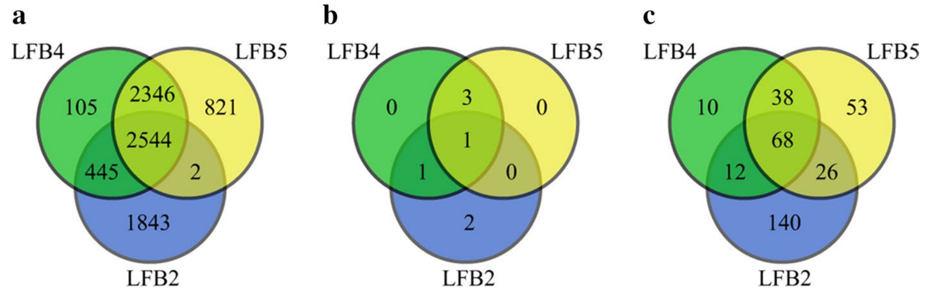

Fig. 11 Venn diagrams for obtained LFBs. **a** Venn diagrams of sites in LFB2, LFB4 and LFB5. **b** Venn diagrams of conditions in LFB2, LFB4 and LFB5. **c** Venn diagrams of functional annotations of genes that selected sites in LFB2, LFB4 and LFB5 reside on

Since LFB2 was found enriched in three KEGG pathways previously, and might be condition specific, LFB2 was compared with LFB4 and LFB5 for further study, and the Venn diagram is shown in Fig. 11.

It can be seen from Fig. 11 that although LFB2, LFB4 and LFB5 share 2544 sites, they share only one condition, which leads to 203 functional annotations that are not shared at all, indicating that the three LFBs may play different roles in m^{6}A methylation.

We further investigated the functions of LFB2 in detail, as shown in Fig. 12.

Some genes that sites in LFB2 reside on are found to be involved in m^{6}A-related pathways, such as Ras protein signal transduction [57], macromolecule methylation [58], peptidyl-lysine modification [59], histone modification [60] and covalent chromatin modification [61], implying LFB2 may further help elaborate the functional mechanisms of m^{6}A methylation. Besides, some pathways, such as response to heat [10, 62], are found to be significantly enriched in LFB2, which is also consistent with previous analysis that LFB2 covers conditions with HepG2 cells that exposed to ultraviolet radiation, heat shock, hepatocyte growth factor (HGF; also known as scatter factor (SF)), and interferon-γ.

Since LFB5 is not enriched in any of the six KEGG pathways in the previous analysis, the functionality of LFB5 is similarly examined in further detail, and the result is shown in Fig. 13.

We can see that LFB5 is mainly enriched in functional annotations related to histone and lysine modification, in which the modification form is mainly acetylation modification. M^{6}A modification in RNA has been found to be determined by histone

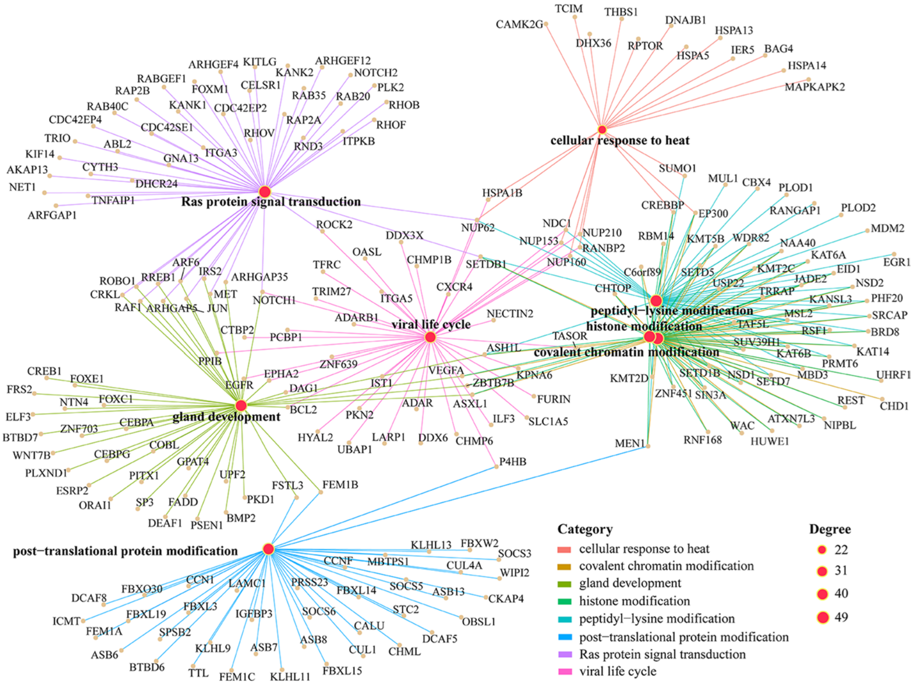

Fig. 12 The functional relationship diagram obtained from the analysis of the genes related to the LFB2 using KEGG pathway. Degree represents the number of genes enriched by KEGG pathway

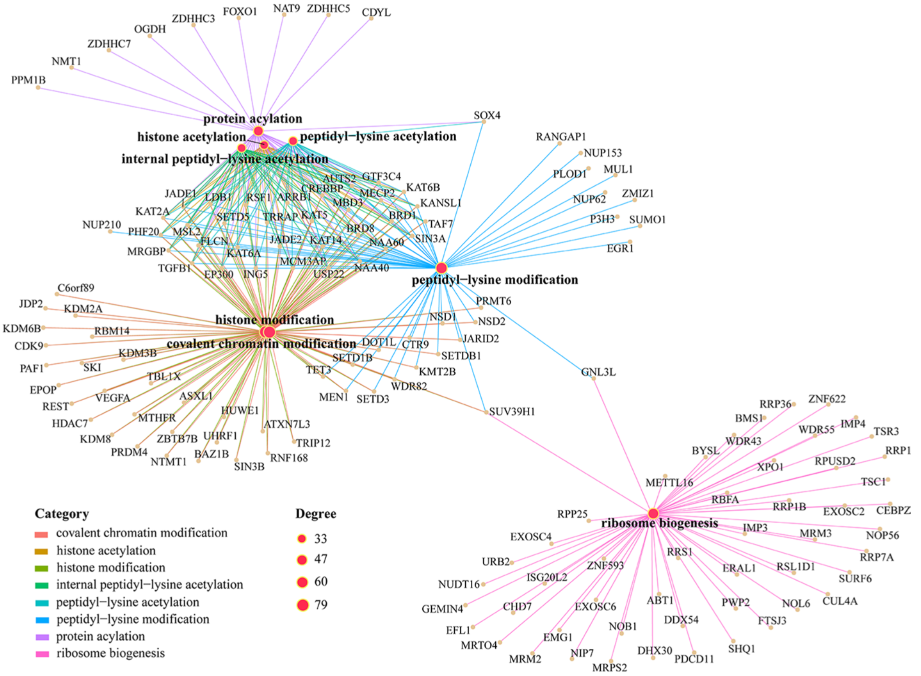

Fig. 13 The functional relationship diagram obtained from the analysis of the genes related to the LFB5 using KEGG pathway. Degree represents the number of genes enriched by KEGG pathway

modification [63]. Therefore, the genes contained in LFB5 may help uncover the relationship between histone and lysine modification and m^{6}A methylation.

Discussion

More and more studies have shown that m^{6}A RNA methylation plays an extremely important role in a variety of biological processes. Moreover, the functions of m^{6}A methylation have been revealed by more and more researchers. Through the study of m^{6}A methylation, we could understand the pathogenesis of the disease at post-transcriptional level, which would help us build a more comprehensive understanding of life process such as disease mechanisms. However, unveiling the functional m^{6}A methylation sites through biological experiments is time-consuming and expensive, so it is very necessary to develop some effective computational algorithms to predict potential functional m^{6}A sites. In this paper, we developed an RNA expression weighted ISA method, REW-ISA, to uncover the potential local methylation patterns across subsets of condition. REW-ISA approached 6 LFBs based on MeRIP-Seq data from 10 cell lines under 32 different conditions. Further GO analysis and some specificity tests show that REW-ISA obtained LFBs can find hyper-methylated local functional patterns, which are highly relevant with conditions.

REW-ISA could achieve reliable biclustering patterns because of its adoption of RNA expression level. For the m^{6}A methylation level matrix, the level was drawn based on the ratio between IP and input samples, and there are no additional supplements for RNA expression level. To be more specific, the methylation levels of sites of high expression level should be more confident than those of low expression level since the reads count statistics in low expression sites may come from noise, which makes them unconfident. By incorporation of RNA expression level, sites with very low abundance of reads count will be assigned very low weight, thus, excluded for consideration of biclustering. Of course, REW-ISA still has some deficiencies that needs to be improved in the future. REW-ISA seek for LFBs of hyper-methylated sites under subsets of conditions based on methylation level, which is achieved by simple division operation. However, it may lead to information loss during the division operation. The information carried by both input and IP samples should be more than the methylation level and RNA expression level. In the future, we will develop new computational model to overcome these limitations.

Conclusions

With comparison with conventional ISA method, we believe that our test suggests REW-ISA as a simple but effective tool for local functional pattern recognition tasks. Through the experiments, we also showed that REW-ISA is also feasible for real-world applications with similar issues as local pattern analysis problem in m^{6}A methylation profiles.

Methods

In conventional ISA method, rows and columns of data are standardized first, and subsets of rows and columns are updated iteratively according to their own thresholds. However, in REW-ISA, we propose to import weights to enhance the confidence of methylation level estimation, so the min–max normalization was employed instead of z-score normalization. The methylation level matrix $\boldsymbol{P} \in \mathbb{R}^{p \times n}$ turns into $\boldsymbol{P^R}$ after row min–max normalization, and turns into $\boldsymbol{P^C}$ after column min–max normalization. The flowchart of REW-ISA is shown in Fig. 14. In general, REW-ISA consists of two steps. The first step aims to form the methylation level and weight matrix for all sites under all conditions. The second part conducts iteratively selection of subsets of rows and columns for LFBs. With the employment of $\boldsymbol{W} \in \mathbb{R}^{p\times n}$, the contribution of sites showing similar methylation level may be distinguishable due to their different expression level.

For subset selection of columns, the updated subsets are achieved following (14).

$$\begin{cases} e^C_{\boldsymbol{U'}v} = \frac{1}{|\boldsymbol{U'}|}\sum_{u\in \boldsymbol{U'}}\left(w_{uv}\cdot p^R_{uv}\right) & v \in \boldsymbol{V} \\ \boldsymbol{V'} = \left\{ v \in \boldsymbol{V} \;:\; |e^C_{\boldsymbol{U'}v} - \frac{1}{|\boldsymbol{V}|}\sum_{v\in \boldsymbol{V}} e^C_{\boldsymbol{U'}v}| > \frac{T_C}{\sqrt{|\boldsymbol{U'}|}} \right\} \end{cases} \tag{14}$$

where $\boldsymbol{V}$ is the column set of $\boldsymbol{P}$, p^R_{uv} refers to the u-th site under v-th condition in $\boldsymbol{P^R}$, w_{uv} is the RNA expression level of the u-th site under v-th condition. In Eq. (14), $e^C_{U'v}$ is calculated based on $\boldsymbol{P^C}$ and $\boldsymbol{W}$ for column subset selection, but only the conditions involved in $\boldsymbol{U'}$ are considered. Higher T_C setting will result in less conditions in LFB.

Then, the subsets of rows are updated following (15).

$$\begin{cases} e^R_{u\boldsymbol{V'}} = \frac{1}{|\boldsymbol{V'}|}\sum_{v\in \boldsymbol{V'}}\left(w_{uv}\cdot p^C_{uv}\right) & u \in \boldsymbol{U} \\ \boldsymbol{U'} = \left\{ u \in \boldsymbol{U} \;:\; |e^R_{u\boldsymbol{V'}} - \frac{1}{|\boldsymbol{U}|}\sum_{u\in \boldsymbol{U}} e^R_{u\boldsymbol{V'}}| > \frac{T_R}{\sqrt{|\boldsymbol{V'}|}} \right\} \end{cases} \tag{15}$$

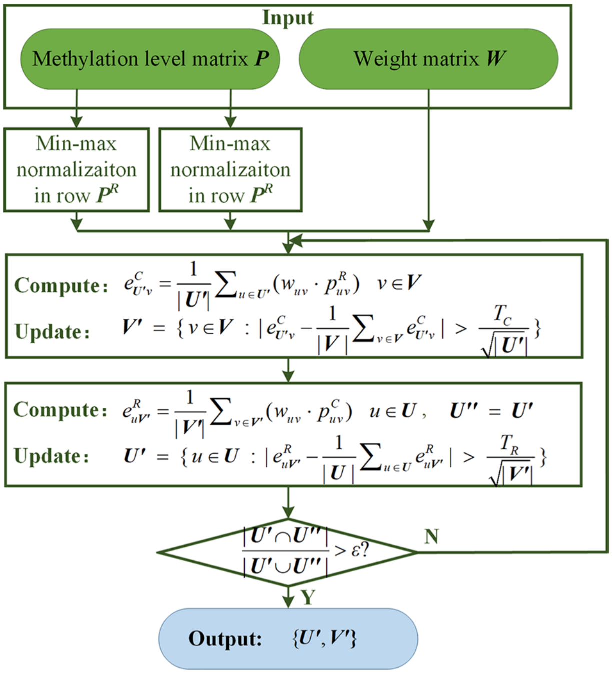
Input
Methylation level matrix P
Weight matrix W
Min-max normalizaiton in row P^R
Min-max normalizaiton in row P^R
Compute: e^C_{U'v} = 1/|U'| Σ_{u∈U'}(w_{uv} · p^R_{uv}) v∈V
Update: V' = {v∈V : |e^C_{U'v} − 1/|V| Σ_{v∈V} e^C_{U'v}| > T_C/√|U'|}
Compute: e^R_{uV'} = 1/|V'| Σ_{v∈V'}(w_{uv} · p^C_{uv}) u∈U , U'' = U'
Update: U' = {u∈U : |e^R_{uV'} − 1/|U| Σ_{u∈U} e^R_{uV'}| > T_R/√|V'|}
|U'∩U''| / |U'∪U''| > ε?
N
Y
Output: {U', V'}

Fig. 14 The flowchart of REW-ISA consists of two steps: The first step prepares the methylation level matrix ***P*** and weight matrix ***W***; the second step iteratively updates subsets for LFBs. The iterative update refers to the iterative selection along columns and rows

where $\boldsymbol{U}$ is the row set of $\boldsymbol{P}$, p^{C}_{uv} refers to the u-th site under v-th condition in $\boldsymbol{P^{C}}$, $e^{R}_{uV'}$ is calculated based on $\boldsymbol{P^{R}}$ and $\boldsymbol{W}$ for row subset selection, but only the sites involved in $\boldsymbol{V'}$ are considered. Higher T_R results in less sites in LFB.

Since $\boldsymbol{P}$ is $p \gg n$, it is intuitive that LFBs will cover more sites than conditions. Thus, $|\boldsymbol{U'}| \gg |\boldsymbol{V'}|$. The selection of sites needs to be more strict, thus T_R is recommended to be larger for less sites inclusion for each LFB. On the contrary, there are not that much conditions in $\boldsymbol{P}$, thus, T_C is recommended to be smaller for loose constrain of conditions in each LFB.

In the parameter selection procedure, it is recommended that the upper bound of T_R and T_C be set larger values, and the algorithm can automatically shrink the thresholds range. Although the larger upper bound may introduce computation load, it is still acceptable since no LFBs can be achieved under large thresholds setting. The optimization of T_R and T_C is reached by grid search. With optimized T_R and T_C, a subset of sites is randomly selected as $\boldsymbol{U'}$, and then the subset of conditions $\boldsymbol{V'}$ is selected according to (14). $\boldsymbol{U'}$ and $\boldsymbol{V'}$ are updated iteratively by (14) and (15) until convergenece is satisfied.

$$\frac{|\boldsymbol{U'} \cap \boldsymbol{U''}|}{|\boldsymbol{U'} \cup \boldsymbol{U''}|} \leq \varepsilon \tag{16}$$

where ε is the default convergence criteria, $\boldsymbol{U''}$ represents the subset of sites in previous iteration, and $\boldsymbol{U'}$ represents the subset of site in current iteration.

The implementation of REW-ISA following the above definition is summarized in the following.

Algorithm 2: REW-ISA biclustering algorithm

```
Input: Methylation site $V$, conditions $U$, methylation level matrix $P$, weight matrix $W$ and converge threshold $\varepsilon$
Output: A series of LFBs ($U'$, $V'$)
Step1: Construct row normalized matrix $P^R$, construct column normalized matrix $P^C$
Step2: Given the pre-defined range for $T_R$ and $T_C$, get the automatically optimized parameter settings
Step3: Under the optimized parameters $T_R$ and $T_C$, initialize the sites subset $U'$ and update $U'$ and $V'$ iteratively
  until the convergence condition is met
Step4: Report $U'$ and $V'$
Return A series of LFBs ($U'$, $V'$)
```

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03787-w.

Additional file 1. Table S1: The detailed information of real data. The information of real data. Experiments with light green background were not included in MeTDBV2.0 yet. Experiment ID: The index of experiments; Cell line: The cell line that MeRIP-Seq profiled. Expr_name in MeTDB V2.0: The retrieval information in MeTDBV2.0. If the data was not included in MeTDBV2.0, the GEO accession numbers were provided instead (indicated as light green background). Treatment: The treatment applied for the experiment. Reference: The title of the source reference. Reference ID: Reference number in the article.

Abbreviations

m^{6}A: N^6-mthyladenosine; REW-ISA: RNA Expression Weighted Iterative Signature Algorithm; LFB: Local functional block; MeRIP-Seq: Methylated RNA immunoprecipitation sequencing; DPBBM: Dirichlet process based beta binomial mixture model; ISA: Iterative signature algorithm; SDwC: Standard deviation within clusters; ASwC: Average similarity within clusters; SoBC: Score of bi-clustering; GO: Gene ontology; IoU: Intersection over union; FPKM: Fragments Per Kilobase of transcript per Million; HGF: Hepatocyte growth factor; HepG2: Human liver hepatocellular cells.

Acknowledgements
Not applicable.

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

Funding
This work has been supported by Fundamental Research Funds for the Central Universities (Research Projects No. 2015XKQY21 to HL), the National Natural Science Foundation of China (Research Projects Nos. 61971422 to LZ, 31871337 to HL). 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.

Availability of data and materials
The code implemented to perform the analysis is deposited at https://github.com/cst-cumt/REW-ISA. The detailed information of real data consists of MeRIP-Seq data from 10 cell lines under 32 kinds of treatments, as listed in Additional file 1: Table S1. For each type of treatment, MeRIP-Seq gets input and IP sequences respectively. Among the 32 conditions, 30 can be retrieved from MeTDBV2.0 at https://180.208.58.19/metdb_v2, and the other two can be retrieved by GEO accession numbers (SRR5080301-SRR50312 and SRR5239086-SRR5239109).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China. [2] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China. [3] Department of Biological Sciences, AI University Research Center, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China.

Received: 17 February 2020 Accepted: 29 September 2020
Published online: 09 October 2020

References

1. Long W-L, Guo H, Sheng J, Song R-H, Xu Y. Role of m6A RNA methylation in tumorigenesis and development. Biotechnol Bull. 2019;6:25.
2. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature. 2014;505(7481):117–20.
3. Xiang Y, Laurent B, Hsu C-H, Nachtergaele S, Lu Z, Sheng W, Xu C, Chen H, Ouyang J, Wang S, et al. RNA m6A methylation regulates the ultraviolet-induced DNA damage response. Nature. 2017;543(7646):573–6.
4. Shay G, Sharon M-M, Dan D, Abed AlFatah M, Nitzan K, Mali S-D, Vera H. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. Science (New York). 2015;6225(347):1.
5. Fustin J-M, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka Masaki S, Kakeya H, Manabe I, et al. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell. 2013;155(4):793–806.
6. Yoon K-J, Ringeling FR, Vissers C, Jacob F, Pokrass M, Jimenez-Cyrus D, Su Y, Kim N-S, Zhu Y, Zheng L, et al. Temporal control of mammalian cortical neurogenesis by m6A methylation. Cell. 2017;171(4):877-889.e817.
7. Han D, Liu J, Chen C, Dong L, Liu Y, Chang R, Huang X, Liu Y, Wang J, Dougherty U, et al. Anti-tumour immunity controlled through mRNA m6A methylation and YTHDF1 in dendritic cells. Nature. 2019;566(7743):270–4.
8. Shi H, Zhang X, Weng Y-L, Lu Z, Liu Y, Lu Z, Li J, Hao P, Zhang Y, Zhang F, et al. m6A facilitates hippocampus-dependent learning and memory through YTHDF1. Nature. 2018;563(7730):249–53.
9. Haussmann IU, Bodi Z, Sanchez-Moran E, Mongan NP, Archer N, Fray RG, Soller M. m6A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. Nature. 2016;540(7632):301–4.
10. Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian S-B. Dynamic m6A mRNA methylation directs translational control of heat shock response. Nature. 2015;526(7574):591–4.
11. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012;485(7397):201–6.
12. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
13. Hess ME, Hess S, Meyer KD, Verhagen LA, Koch L, Brönneke HS, Dietrich MO, Jordan SD, Saletore Y, Elemento O. The fat mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry. Nat Neurosci. 2013;16(8):1042–8.

14. Schwartz S, Agarwala SD, Mumbach MR, Jovanovic M, Mertins P, Shishkin A, Tabach Y, Mikkelsen TS, Satija R, Ruvkun G. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell. 2013;155(6):1409–21.
15. Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X. A METTL3–METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat Chem Biol. 2014;10(2):93–5.
16. Ping X-L, Sun B-F, Wang L, Xiao W, Yang X, Wang W-J, Adhikari S, Shi Y, Lv Y, Chen Y-S. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. Cell Res. 2014;24(2):177–89.
17. Schwartz S, Mumbach MR, Jovanovic M, Wang T, Maciag K, Bushkin GG, Mertins P, Ter-Ovanesyan D, Habib N, Cacchiarelli D. Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites. Cell Rep. 2014;8(1):284–96.
18. Yue Y, Liu J, Cui X, Cao J, Luo G, Zhang Z, Cheng T, Gao M, Shu X, Ma H. VIRMA mediates preferential m6A mRNA methylation in 3′ UTR and near stop codon and associates with alternative polyadenylation. Cell discovery. 2018;4(1):1–17.
19. Patil DP, Chen C-K, Pickering BF, Chow A, Jackson C, Guttman M, Jaffrey SR. m6A RNA methylation promotes XIST-mediated transcriptional repression. Nature. 2016;537(7620):369–73.
20. Wen J, Lv R, Ma H, Shen H, He C, Wang J, Jiao F, Liu H, Yang P, Tan L. Zc3h13 regulates nuclear RNA m6A methylation and mouse embryonic stem cell self-renewal. Mol Cell. 2018;69(6):1028–38.
21. Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, Yi C, Lindahl T, Pan T, Yang Y-G. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol. 2011;7(12):885.
22. Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang C-M, Li CJ, Vågbø CB, Shi Y, Wang W-L, Song S-H. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol Cell. 2013;49(1):18–29.
23. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. N6-methyladenosine modulates messenger RNA translation efficiency. Cell. 2015;161(6):1388–99.
24. Xiao W, Adhikari S, Dahal U, Chen Y-S, Hao Y-J, Sun B-F, Sun H-Y, Li A, Ping X-L, Lai W-Y. Nuclear m6A reader YTHDC1 regulates mRNA splicing. Mol Cell. 2016;61(4):507–19.
25. Xu C, Wang X, Liu K, Roundtree IA, Tempel W, Li Y, Lu Z, He C, Min J. Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. Nat Chem Biol. 2014;10(11):927–9.
26. Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, Zhao BS, Mesquita A, Liu C, Yuan CL. Recognition of RNA N6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat Cell Biol. 2018;20(3):285–95.
27. Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, Elemento O, Pestova TV, Qian S-B, Jaffrey SR. 5′ UTR m6A promotes cap-independent translation. Cell. 2015;163(4):999–1010.
28. Liu L, Zhang SW, Zhang YC, Liu H, Zhang L, Chen R, Huang Y, Meng J. Decomposition of RNA methylome reveals co-methylation patterns induced by latent enzymatic regulators of the epitranscriptome. Mol Biosyst. 2015;11(1):262–74.
29. Cui X, Meng J, Zhang S, Rao MK, Chen Y, Huang Y. A hierarchical model for clustering m6A methylation peaks in MeRIP-seq data. BMC Genom. 2016;17(7):520.
30. Zhang L, He Y, Wang H, Liu H, Huang Y, Wang X, Meng J. Clustering count-based RNA methylation data using a nonparametric generative model. Curr Bioinform. 2019;14(1):11–23.
31. Liu H, Flores MA, Meng J, Zhang L, Zhao X, Rao MK, Chen Y, Huang Y. MeT-DB: a database of transcriptome methylation in mammalian cells. Nucl Acids Res. 2014;43:D197.
32. Sun W, Li J, Liu S, Wu J, Zhou H, Qu L, Yang J. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucl Acids Res. 2015;44:D259–65.
33. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2018;19(4):575–92.
34. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. Nat Genet. 2002;2002:1089–1089.
35. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E. 2003;67:031902.
36. Murali T, Kasif S. Extracting conserved gene expression motifs from gene expression data. In: Biocomputing 2003. Singapore: World Scientific; 2002, p. 77–88.
37. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006;22(9):1122–9.
38. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods. 2015;12(8):767–72.
39. Ke S, Alemu EA, Mertens C, Gantman EC, Fak JJ, Mele A, Haripal B, Zucker-Scharff I, Moore MJ, Park CY. A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. Genes Dev. 2015;29(19):2037–53.
40. Vu LP, Pickering BF, Cheng Y, Zaccara S, Nguyen D, Minuesa G, Chou T, Chow A, Saletore Y, MacKay M. The N6-methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. Nat Med. 2017;23(11):1369.
41. Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vågbø CB, Geula S, Hanna JH, Black DL, Darnell JE, Darnell RB. m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. Genes Dev. 2017;31(10):990–1006.
42. Chen K, Wei Z, Zhang Q, Wu X, Rong R, Lu Z, Su J, de Magalhaes JP, Rigden DJ, Meng J. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucl Acids Res. 2019;47(7):e41–e41.
43. Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, Tu BP, Conrad NK. The U6 snRNA m6A methyltransferase METTL16 regulates SAM synthetase intron retention. Cell. 2017;169(5):824–35.
44. Barbieri I, Tzelepis K, Pandolfini L, Shi J, Millán-Zambrano G, Robson SC, Aspris D, Migliori V, Bannister AJ, Han N. Promoter-bound METTL3 maintains myeloid leukaemia by m6A-dependent translation control. Nature. 2017;552(7683):126–31.

45. Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, Huang H, Nachtergaele S, Dong L, Hu C. FTO plays an oncogenic role in acute myeloid leukemia as a N6-methyladenosine RNA demethylase. Cancer Cell. 2017;31(1):127–41.
46. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell. 2012;149(7):1635–46.
47. Batista PJ, Molinie B, Wang J, Qu K, Zhang J, Li L, Bouley DM, Lujan E, Haddad B, Daneshvar K. m6A RNA modification controls cell fate transition in mammalian embryonic stem cells. Cell Stem Cell. 2014;15(6):707–19.
48. Liu H, Wang H, Wei Z, Zhang S, Hua G, Zhang S-W, Zhang L, Gao S-J, Meng J, Chen X. MeT-DB V20: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. Nucl Acids Res. 2018;46(D1):D281–7.
49. Wu X, Wei Z, Chen K, Zhang Q, Su J, Liu H, Zhang L, Meng J. m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network. BMC Bioinform. 2019;20(1):223.
50. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.
51. Harris SL, Levine AJ. The p53 pathway: positive and negative feedback loops. Oncogene. 2005;24(17):2899–908.
52. Vogelstein B, Fau-Lane D, Lane D, Fau-Levine AJ, Levine AJ. Surfing the p53 network. Nature. 2000;408(6810):307–10.
53. Levine AJ, Hu W, Feng Z. The P53 pathway: what questions remain to be explored? Cell Death Differ. 2006;13(6):1027–36.
54. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics J Integr Biol. 2012;16(5):284–7.
55. Li L, Guo Y, Wu W, Shi Y, Cheng J, Tao S. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. BioData Min. 2012;5(1):8.
56. Richards AL, Holmans P, O'Donovan MC, Owen MJ, Jones L. A comparison of four clustering methods for brain expression microarray data. BMC Bioinform. 2008;9(1):490.
57. Yang F, Jin H, Que B, Chao Y, Zhang H, Ying X, Zhou Z, Yuan Z, Su J, Wu B. Dynamic m6A mRNA methylation reveals the role of METTL3-m6A-CDCP1 signaling axis in chemical carcinogenesis. Oncogene. 2019;38(24):4755–72.
58. Steitz TA, Moore PB. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem Sci. 2003;28(8):411–8.
59. Zhang C, Chen Y, Sun B, Wang L, Yang Y, Ma D, Lv J, Heng J, Ding Y, Xue Y. m6A modulates haematopoietic stem and progenitor cell specification. Nature. 2017;549(7671):273–6.
60. Lin X, Chai G, Wu Y, Li J, Chen F, Liu J, Luo G, Tauler J, Du J, Lin S. RNA m6A methylation regulates the epithelial mesenchymal transition of cancer cells and translation of Snail. Nat Commun. 2019;10(1):1–13.
61. Liu J, Dou X, Chen C, Chen C, Liu C, Xu MM, Zhao S, Shen B, Gao Y, Han D. N6-methyladenosine of chromosome-associated regulatory RNA regulates chromatin state and transcription. Science. 2020;367(6477):580–6.
62. Yu J, Li Y, Wang T, Zhong X. Modification of N6-methyladenosine RNA methylation on heat shock protein expression. PLoS ONE. 2018;13(6):e0198604.
63. Huang H, Weng H, Zhou K, Wu T, Zhao BS, Sun M, Chen Z, Deng X, Xiao G, Auer F. Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. Nature. 2019;567(7748):414–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.