Hindawi
BioMed Research International
Volume 2018, Article ID 2075173, 15 pages
https://doi.org/10.1155/2018/2075173

*Research Article*

# Enhancing Epitranscriptome Module Detection from m$^6$A-Seq Data Using Threshold-Based Measurement Weighting Strategy

**Kunqi Chen**,[1,2] **Zhen Wei**,[1,2] **Hui Liu**,[3] **João Pedro de Magalhães**,[2] **Rong Rong**,[1,4] **Zhiliang Lu**,[1,4] **and Jia Meng**[1,4]

[1]*Department of Biological Sciences, RCPM, URCHT, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China*
[2]*Institute of Ageing & Chronic Disease, University of Liverpool, L7 8TX, Liverpool, UK*
[3]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China*
[4]*Institute of Integrative Biology, University of Liverpool, L7 8TX, Liverpool, UK*

Correspondence should be addressed to Jia Meng; jia.meng@xjtlu.edu.cn

Received 27 December 2017; Accepted 27 May 2018; Published 14 June 2018

Academic Editor: Angelo Ciaramella



To date, with well over 100 different types of RNA modifications associated with various molecular functions identified on diverse types of RNA molecules, the epitranscriptome has emerged to be an important layer for gene expression regulation. It is of crucial importance and increasing interest to understand how the epitranscriptome is regulated to facilitate different biological functions from a global perspective, which may be carried forward by finding biologically meaningful epitranscriptome modules that respond to upstream epitranscriptome regulators and lead to downstream biological functions; however, due to the intrinsic properties of RNA molecules, RNA modifications, and relevant sequencing technique, the epitranscriptome profiled from high-throughput sequencing approaches often suffers from various artifacts, jeopardizing the effectiveness of epitranscriptome modules identification when using conventional approaches. To solve this problem, we developed a convenient measurement weighting strategy, which can largely tolerate the artifacts of high-throughput sequencing data. We demonstrated on real data that the proposed measurement weighting strategy indeed brings improved performance in epitranscriptome module discovery in terms of both module accuracy and biological significance. Although the new approach is integrated with Euclidean distance measurement in a hierarchical clustering scenario, it has great potential to be extended to other distance measurements and algorithms as well for addressing various tasks in epitranscriptome analysis. Additionally, we show for the first time with rigorous statistical analysis that the epitranscriptome modules are biologically meaningful with different GO functions enriched, which established the functional basis of epitranscriptome modules, fulfilled a key prerequisite for functional characterization, and deciphered the epitranscriptome and its regulation.

## 1. Introduction

In the exploration of epigenetic modifications of RNA that has lasted for 5 decades, more than 100 types of posttranscriptional chemical RNA modifications have been identified [1]. Among these modifications, $N^6$-methyladenosine (m$^6$A) is the most abundant type of RNA modifications that steers or participates in various biological functions including circadian clock [2], translation [3, 4], cortical neurogenesis [5], microRNA processing [6], Drosophila sex determination [7, 8], T cell homeostasis [9], RNA-protein interaction [10], and RNA stability [11, 12]. It also plays an important role in DNA damage response [13], heat shock response [14], and the resolution of naïve pluripotency towards differentiation [15]. As RNA methylation participates in many fundamental cellular processes, it is closely related to many types of disease, such as cancer [16, 17] and virus infection [18]. It has been shown that m$^6$A demethylase ALKBH5 maintains the tumorigenicity of glioblastoma stem-like cells by programming cell proliferation [19]; the m$^6$A demethylase FTO plays as an oncogene in Acute Myeloid Leukemia [20]; and the m$^6$A methyltransferase METTL3 controls myeloid differentiation of normal

hematopoietic and leukemia cells [21]. Mutations of the RNA methylation enzymes are linked to colon cancer and endometrial cancer [22]. Due to the importance of RNA $m^6A$ modification to biological regulation and health, it is of crucial importance and increasing interest to study how the epitranscriptome is shaped to regulate relevant biological processes.

There are a large number of RNA $m^6A$ sites enriched near stop codon, on 3'UTRs and on long exons of the transcriptome [23]. It was originally reported in 2012 that there exist over 12,000 $m^6A$ sites on 7676 mammalian genes that contain $m^6A$ [24, 25]. Due to the limitation of sequencing depth, context-specific expression, and dynamics of RNA $m^6A$ sites, the actual number of $m^6A$ sites in the human epitranscriptome is likely to be much larger. There are more than 0.3~0.4 million predicted unique $m^6A$ sites reported in the human epitranscriptome according to two recent bioinformatics databases RMBase [26] and MetDB [27], which are collected by merging MeRIP-Seq data from published studies, although many of these $m^6A$ sites may exist under very few conditions (tissue/cell types/treatment) or even false positive due to the way the sites are searched; i.e., an unmodified a residual, which conforms the RRACH motif, was false positively reported by the MeRIP-Seq technique due to its proximity to real $m^6A$ sites [24, 25].

The $m^6A$ modification is directly deposited or erased by relevant enzymes, i.e., RNA $m^6A$ methyltransferase (writer) and demethylase (eraser), which are accountable to the observed landscape of $m^6A$ epitranscriptome in cells. The most well studied $m^6A$ methyltransferase is a complex [28–30] composed of at least four proteins, including METTL3, METTL14, WTAP, and KIAA1429 [29, 31–33]. It has been shown that METTL3 functions catalytically, while the other proteins mainly serve as regulatory units that mediate the substrate specificity of the methyltransferase complex [34–36]. The fat mass and obesity associated protein (FTO) was identified in 2011 as the first known $m^6A$ demethylase [37]. Moreover, the protein ALKBH5, derived from the same protein family (ALKB) of FTO, was identified as a second $m^6A$ demethylase that impacts RNA metabolism and mouse fertility [38]. Very recently, METTL16 is identified as another RNA $m^6A$ writer that targets pre-mRNAs and noncoding RNAs [39].

Although there are likely to be additional $m^6A$-relevant enzymes yet discovered by people, the total number of primary $m^6A$-regulating genes is likely to be much less than the total number of $m^6A$ sites in the epitranscriptome. Due to the substrate specificity of $m^6A$-relevant enzymes, epitranscriptome modules are naturally formed when a larger number of $m^6A$ sites are regulated by a small number of regulators; i.e., the $m^6A$ sites that share the same regulator will exhibit similar methylation pattern across different experiment conditions, reflecting the catalytic efficacy of their common regulator under respective conditions. The concept of regulatory module has been used extensively in the field of bioinformatics. For example, a transcriptional module of 148 genes that are downregulated during differentiation has been functionally associated with self-renewal [40]. A transcriptional module of 4382 genes is identified to be associated with cell cycle from a time course data with 24 samples in yeast using state space models [41]. In DNA methylation data analysis, modules in the epigenome have been associated with ageing effects [42] and alcohol use disorders [43]. In studies of lncRNA, coexpression of a gene and an lncRNA is often a strong indication for functional relevance of the two and has been used for predicting the functions of novel lncRNAs [44, 45]. Given the aforementioned examples in transcriptomics, epigenomics, and genomics, because the methylation sites of the same epitranscriptome module are coregulated across different experiment conditions, it is reasonable to speculate that they are functionally related as well, i.e., participating in the same or related biological processes and pathways.

Previously, the studies of epitranscriptome module are mainly restricted to the study of substrate specificity of the epitranscriptome enzymes. Through the perturbation of $m^6A$ writers, Regev lab identified two distinct classes of $m^6A$ sites based on whether they depend on WTAP, a key regulator of the METTL13-METTL14 writer complex [33]. Liu et al. performed four different clustering approaches to 3274 preselected RNA methylation sites and identified an epitranscriptome module that is likely to be mediated by the $m^6A$ demethylase FTO [46]. As increasing significance and biological functions of RNA $m^6A$ modifications are revealed by recent studies, it is of growing necessity to understand the epitranscriptome regulation. The study of epitranscriptome modules provided a viable venue to achieve it.

Currently, the most popular high-throughput sequencing approach for profiling RNA methylome is methylated RNA immunoprecipitation sequencing (MeRIP-seq or $m^6A$-seq) [24, 25]. From technical perspective, $m^6A$-seq may be considered as a marriage of RNA-seq and ChIP-seq technique, where the methylation signal is obtained by sequencing the immunoprecipitated RNA fragments with anti-$m^6A$ antibody (the IP sample), and the control background is generated using all the input RNA fragments (the input sample). A major difficulty faced by computational biologists when searching for the epitranscriptome modules is to deal with the artifacts in epitranscriptome high-throughput sequencing data, which is mainly due to the context-specific gene expression, the limitation of sequencing depth. Constrained by the detection ability, it is always very difficult to accurately quantify the methylation level of very lowly expressed genes. For example, if the reads count of a specific methylation site in the IP sample is $t$ and the reads count in the paired input sample is $c$, without taking into account the difference in sequencing depth, a natural measurement for the methylation level of this site $m$ is

$$m = \frac{t}{t + c}, \tag{1}$$

where $m \in [0, 1]$. This way of quantifying methylation level has been widely used in DNA methylation analysis in the form of beta-value [47]. However, this approach can be problematic in RNA methylation data analysis when dealing with very lowly expressed genes. For example, while a methylation site with $t = 100$ and $c = 0$ is likely to be highly methylated

($m = 1$) a methylation site with $t = 1$ and $c = 0$ may not ($m = 1$). As a matter of fact, there is barely any signal for the latter case to make any reliable estimation, although the estimated methylation level is 1. Different from DNA methylation data, where the background is homogenous and the background reads coverage is expected to be the same across the entire genome, there exists rather prominent heterogeneity in the reads coverage of the transcriptome and epitranscriptome data; i.e., there are usually a small number of highly expressed genes coupled with a very large number of lowly expressed genes, whose methylation signal is too weak to be estimated reliably, which severely limits the performance of computational approaches based on this measurements. It is necessary to develop strategy that can take advantage of the estimated methylation level together with its reliability.

To unlock the full potentials of epitranscriptome sequencing data, we designed a convenient measurement weighting strategy to incorporate the measurements together with their reliability as the weight. Under this scheme, unreliable measurements that are supported by relatively small number of reads are given less weight in the computation model, while measurements supported by a large number of reads are given more weight. In this way, even if some measurements are not accurate, because smaller weights are assigned to them, the final computation results are still likely to be robust. We will show in the next how to use the weighting strategy in a hierarchical clustering approach to find epitranscriptome modules with weighted Euclidean distance and show the performance improvement compared with the same method but without using the weighting scheme.

## 2. Method

Considering we have the methylation profile of $N$ methylation sites obtained from $S$ experimental conditions and the reads count of the $n$-th methylation site under $s$-th condition in the IP sample is $t_{n,s}$, the reads count of the $n$-th methylation site under $s$-th condition in the input sample is $c_{n,s}$. The size factors of the IP and input samples of the $s$-th condition are $d_{s,t}$ and $d_{s,c}$, respectively, which reflects the sequencing depth (or library size) of the sample, which may be estimated using geometric mean or other approaches. Based on previous definition, the estimated methylation level of the $n$-th methylation site under $s$-th condition is

$$m_{n,s} = \frac{t_{n,s}/d_{s,t}}{t_{n,s}/d_{s,t} + c_{n,s}/d_{s,c}} = \frac{t_{n,s}d_{s,c}}{t_{n,s}d_{s,c} + c_{n,s}d_{s,t}} \tag{2}$$

where $n \in \{1, 2, \cdots, N\}$ and $s \in \{1, 2, \cdots, S\}$. When give the estimated RNA methylation profiles, it is fairly easy to apply hierarchical clustering approach to search for epitranscriptome modules. A typical measurement people use to measure the similarity of two methylation profiles is the Euclidean distance, where the distance between the $i$-th and $j$-th methylation sites $d(i, j)$ may be calculated as follows:

$$d(i, j) = \sqrt{\sum_{s=1}^{S} \left[ \left( m_{i,s} - m_{j,s} \right)^2 \right]} \tag{3}$$

Smaller $d(i, j)$ suggests that the two sites share a very similar methylation profile across different experimental conditions, belong to the same epitranscriptome module, may be regulated by the same epitranscriptome regulator, and may be functional relevant based on previous experience in genomics analysis. However, the distance measurement by Euclidean distance can be seriously affected by a few unreliable measurements estimated from a small number of reads and then may seriously jeopardize the clustering results. To fully take advantage of the potentials of the epitranscriptome sequencing data, we consider here a weighting strategy by using the weighted Euclidean distance. Specifically, the weighted Euclidean distance between the $i$-th and $j$-th methylation sites $d_w(i, j)$ may be calculated as follows:

$$d_w(i, j) = \sqrt{\sum_{s=1}^{S} \left[ w_{s,i,j} \left( m_{i,s} - m_{j,s} \right)^2 \right]} \tag{4}$$

where $w_{s,i,j} > 0$ and $\sum_{s=1}^{S} w_{s,i,j} = 1$. The weight is determined by a function of the reads counts $w_{s,i,j} = f_w(t_{i,s}, c_{i,s}, t_{j,s}, c_{j,s})$. In this formulation, the weight assigned to the $s$-th experimental condition $w_{s,i,j}$ should reflect the reliability of this sample. If the measurements obtained under this sample were estimated from a small number of reads $(t_{i,s}, c_{i,s}, t_{j,s}, c_{j,s})$, the relevant part of the result may not be reliable and a smaller weight should be assigned.

Although it is conceptually easy to depict the desired properties of the weight function $w_{s,i,j} = f_w(t_{i,s}, c_{i,s}, t_{j,s}, c_{j,s})$, there still exist different ways to define the function and it is still an open question how to choose a proper weighting strategy according to the data. In this manuscript, we consider the following 2 weighting schemes:

(i) logarithm-based approach, in which,

$$\begin{aligned} \widetilde{w}_{s,i,j} &= f_w \left( t_{i,s}, c_{i,s}, t_{j,s}, c_{j,s} \right) \\ &= \log \left( t_{i,s} + c_{i,s} + t_{j,s} + c_{j,s} + 1 \right) \end{aligned} \tag{5}$$

$$w_{s,i,j} = \frac{\widetilde{w}_{s,i,j}}{\sum_{\forall s} \widetilde{w}_{s,i,j}} \tag{6}$$

In this approach, the weight of a sample increases logarithmically with the number of reads. Conceivably, it is reasonable to assume that there exists a big difference in terms of reliability between measurements supported by 2 and 200 reads but only very minor difference between those supported by 1002 and 1202 reads.

(ii) threshold-based approach, in which

$$\begin{aligned} \widetilde{w}_{s,i,j} &= f_w \left( t_{i,s}, c_{i,s}, t_{j,s}, c_{j,s} \right) \\ &= \begin{cases} 1 & \text{when } \left( t_{i,s} + c_{i,s} + t_{j,s} + c_{j,s} \right) \geq \alpha \\ \beta & \text{when } \left( t_{i,s} + c_{i,s} + t_{j,s} + c_{j,s} \right) < \alpha \end{cases} \end{aligned} \tag{7}$$

$$w_{s,i,j} = \frac{\widetilde{w}_{s,i,j}}{\sum_{\forall s} \widetilde{w}_{s,i,j}} \tag{8}$$

Table 1: Datasets in the study.

| Dataset ID | Tissue/Cell | Treatment | # Sample (IP & input) | Source |
|---|---|---|---|---|
| 1 | HepG2 | | 4 & 3 | [25] |
| 2 | HepG2 | UV | 1 & 1 | [25] |
| 3 | HepG2 | HS | 1 & 1 | [25] |
| 4 | HepG2 | HGF | 1 & 1 | [25] |
| 5 | HepG2 | IFN | 1 & 1 | [25] |
| **6** | Human Brain | | 1 & 1 | [25] |
| **7** | HEK293T | | 3 & 3 | [56] |
| **8** | U2OS | | 3 & 3 | [2] |
| **9** | U2OS | DAA | 3 & 3 | [2] |

This approach is conceptually similar to the "threshold method" [52] with two parameters $\alpha$ and $\beta$. When the total number of reads mapped to the two sites in sample $s$ is greater than the threshold $\alpha$, we consider they are fairly accurate measurements and assign a normal weight 1, while a smaller weight $\beta$ is assigned to the sample when the measurements of methylation level are not reliable, with $0 \leq \beta \leq 1$. In practice, it is necessary to further optimize the two parameters $\alpha$ and $\beta$ *ad hoc* with respective to the datasets used.

With two different measurement weighting strategies defined as previously, we will next compare the proposed approach on real data with conventional approach without measurement weighting.

## 3. Result

*3.1. RNA Methylation Sequencing Data.* The datasets used in the following analysis were from published studies and downloaded directly from Gene Expression Omnibus (GEO) in SRA format. The data profiles the m$^6$A epitranscriptome in HEK293, HepG2, U2OS, and human brain under different treatments (see Table 1). The reads are aligned to human reference genome assembly (hg19) with the default setting of Tophat2 [53]. Subsequently, the epitranscriptome (all the RNA m$^6$A methylation sites under different conditions) was retrieved using exomePeak [54] with UCSC gene annotation and default settings by following a previous approach [46]. Furtherly, the reads count of every RNA methylation site ($t_{n,s}$ and $c_{n,s}$ for $\forall n \in \{1, 2, \cdots, N\}$ and $\forall s \in \{1, 2, \cdots, S\}$) is retrieved using R/Bioconductor packages [55]. The biological replicates obtained from the same condition are merged together, and the total number of reads is used to estimate the size factor of samples ($d_{s,t}$ and $d_{s,c}$ for $\forall s \in \{1, 2, \cdots, S\}$). The methylation ratio of all sites is estimated according to (2). The estimated methylation level is then quantile normalized to remove possible batch effect. Only the sites that show strong dynamics are retained for further analysis; this is achieved by selecting the RNA methylation sites with larger variance in methylation level. Finally, the methylation profile of each methylation site is standardized by subtracting the mean and divided by its standard deviation to ensure all sites contribute equally to the analysis.

*3.2. Comparing Different Weighting Schemes Using True Sample Labels.* A major limitation for assessing the performance of different weighting schemes in a clustering approach for RNA methylation analysis is the lack of ground truth. The epitranscriptome regulation is complicated, and it is not clear which group of RNA methylation sites shares a common regulator across different experimental conditions. To this end, an alternative approach is considered by taking advantage of the sample labels. There are 6 samples including two triplicates profiling the m$^6$A epitranscriptome in human U2OS cell line with or without DAA treatment (see Table 1, Dataset ID 8 & 9), and if a clustering approach is applied to the 6 samples, it should retrieve two distinct groups corresponding to the DAA and control conditions in the experiment setting. Conceivably, when two different weight schemes are used in the clustering analysis, the one that returns more consistent results with experimental setting suggests a better performance.

The sample label is used as group truth for clustering analysis in the first experiment. Specifically, a total of 916 small datasets, each containing the methylation profiles of 6 samples and 30 RNA methylation sites adjacent with each other in genomic coordinates, are generated by splitting the original high-throughput dataset (Dataset ID 8 & 9 in Table 1), and a hierarchical clustering classifier using Euclidean distance with different weighting schemes (no weighting, logarithm-based and threshold-based) was applied to the small datasets to group the samples into 2 clusters, and the clustering results are then compared to the true sample labels for assessing the clustering performance on all the 916 small datasets. In this analysis, the parameters of threshold-based approach ($\alpha$ and $\beta$) were arbitrarily set to 0.03 to 0.45, respectively, without necessary optimization. Instead of using a specific threshold value for $\beta$, we use here a relative quantile value, where 0.45 is corresponding to the 45% quantile for reads count of all the measurements. As is shown in Table 2, given that there are a total of 6 samples from groups, the probability to obtain a correct clustering result by random is only 3.2%. The RNA methylation profile contains clustering information. Correct clustering results may be obtained for more than 26% of times when the standard approach is applied, which assigns all measurements with equal weight. Additionally, the clustering performance can be further improved by taking advantage

Table 2: Percentage of correct clustering.

| Clustering Approach | # Trial | # Correct Result | % Correct Result |
|---|---|---|---|
| Random Guess | 916 | 30 | 3.22% |
| No weighting | | 241 | 26.31% |
| Logarithm-based weighting | | 294 | 32.06% |
| Threshold-based weighting | | 344 | 37.60% |

of the proposed weighting strategy (32% and 37.6%), which shows the proposed measurement weighting scheme can significantly improve the clustering performance. Please note that the correct percentage is relative low because we used a very stringent criteria; i.e., the clustering results are considered correct if and only if all the samples are clustered correctly. These performances are about ten times more accurate than that achieved from a random classifier (3.22%), which suggests that the clustering results are statistically meaningful. Additionally, we show in Supplementary Materials that the proposed threshold-based weighting scheme is equally applicable when using M-value to quantify the RNA methylation status (see Table S1) and it is also useful when using squared Euclidean distance or City Block to measure the similarity of RNA methylation profiles (see Tables S2 and S3).

In the previous result, the 916 small datasets were generated by splitting the complete high-throughput dataset, and the RNA methylation sites of the same small dataset are adjacent to each other on the genome, which may possess systematic correlation that influences the clustering result. To eliminate this bias, we consider a more random test in the next. Specifically, 916 small datasets with 30 random selected methylation sites are generated, to which clustering analysis using different weighting strategies was applied and the clustering performance was assessed again using the true sample labels. The analysis was repeated for 100 times, and the results are shown in Figure 1. The proposed weighting strategies consistently improve the clustering performance. We also tested the cases when M-value, squared Euclidean, or City Block is used to quantify the RNA methylation status or the similarity of RNA methylation profiles. As is shown in Supplementary Materials Figure S1, consistent improvement in clustering performance is observed when the proposed weight scheme is implemented.

In the previous study, we tested a case when there are 30 RNA methylation sites available for the clustering analysis. We study next the influence of dimension size on clustering performance by changing the number of sites included in the analysis. As shown in Figure 2, the clustering performance increases as the dimension (number of RNA methylation sites) increases, and the clustering method using the weighting strategies consistently outperforms the one that does not use it. In all setting tested, threshold-based weight strategy provides the best clustering performance and the logarithm-based weighting strategy also outperforms the one that does not use measurement weighting. It is now rather clear that many measurements are not accurate and need to be penalized in some way in the analysis. We also tested the cases when M-value, squared Euclidean, or City Block is used to quantify the RNA methylation status or the similarity of RNA methylation profiles. As is shown in Supplementary Materials Figure S2, very similar results are observed. The proposed approach can consistently improve clustering performance when different quantification methods or distance measurements are used.

*3.3. Parameter Optimization for Threshold-Based Weight Strategy.* With previous results, the threshold-based measurement weighting strategy has shown superior performance; this method will be our focus in the next section. It is worth mentioning that the parameters of this method; i.e., the threshold $\alpha$ and weight $\beta$ are still not sufficiently optimized. It is important to further fine-tune these two parameters for the best possible performance. Till this end, we consider here a 2-D grid search, where all combinations of $\alpha$ and $\beta$ are tested, with the threshold parameter $\in$ [0, 0.05, 0.15, 0.25, 0.35, 0.45, 0.50, 0.55, 0.65, 0.75, 0.85, 0.95] and the weight parameter $\beta \in$ [1E-4, 5E-4, 2.5E-3, 1.35E-2, 0.03, 0.045, 0.06, 0.09, 0.135, 0.15, 0.3, 0.75, 1, 1.5, 7.5]. Please note that when $\alpha = 0$ or $\beta = 1$, no measurements will be penalized and the weighting strategy will be essentially the same as standard approach without measurement weighting. When $\beta > 1$, a larger weight will be assigned to the measurements that are less accurate, which is expected to damage the clustering performance. This setting is used in the analysis as a negative control. Similar to before, the performance is tested on small random datasets with different number of RNA methylation sites ($N \in$ [10, 20, 30, 40, 50, 60, 70, 80, 80, 90, 100]). After repeating the analysis 100 times, the average clustering performance under each possible combination of setting is summarized in Figure 3. We can see that the performance patterns on dataset of different sizes are similar. Better clustering performance was achieved when setting a relative small weight parameter $\beta$ and a medium threshold parameter $\alpha$. A large weight parameter $\beta$, which assigns a larger weight to less accurate measurement, always undermines the clustering performance, just as previously expected. When comparing the results achieved on datasets of different size, the optimal threshold parameter $\alpha$ increases as the data size increases. This observation is reasonable, because compared with small dataset, a larger dataset can afford to lose more unreliable measurements in the analysis.

*3.4. Quality Assessment of Epitranscriptome Modules with Gene Ontology Analysis.* We demonstrated with previous analysis the effectiveness of measurement weighting strategy in clustering analysis of biological samples by referring to the sample labels as the ground truth. It is important to test whether the proposed measurement weighting strategy is equally useful in the search of epitranscriptome modules, i.e., clustering the RNA methylation sites into different groups where the sites belong to the same group show consistent hyper- or hypomethylation states under different experiment conditions, suggesting the sharing of a common regulator.

There are a total of 42,758 methylation sites identified from 9 experimental conditions in the datasets. Data preprocessing was firstly performed, in which we aim to select

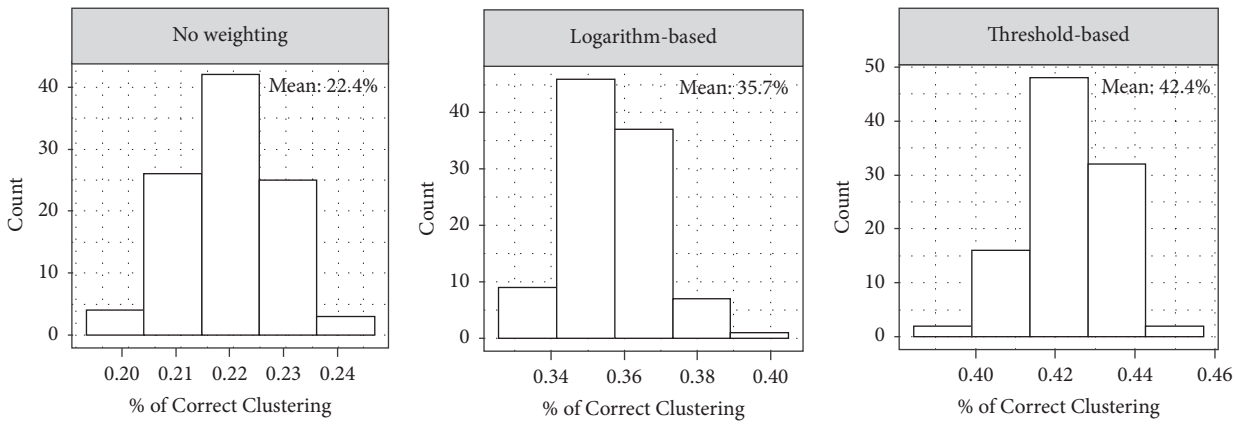


FIGURE 1: **Performance of method using or not using measurement weighting strategy**. When datasets of 30 RNA methylation sites and 6 samples are using for clustering analysis, the proposed two weighting strategies always lead to performance improvement.

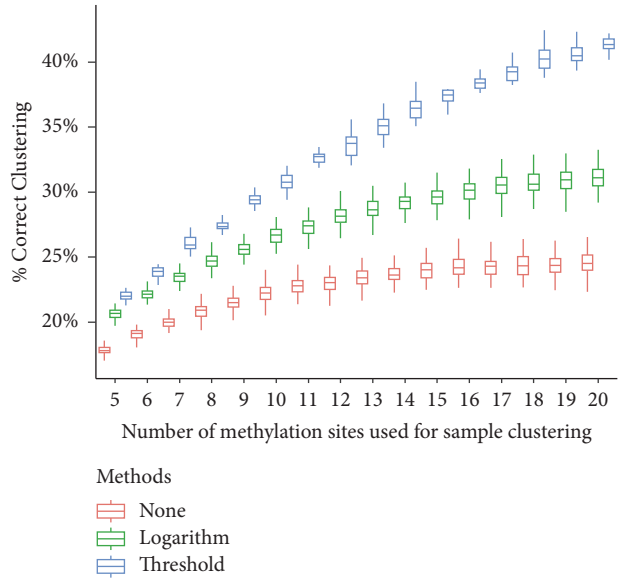


FIGURE 2: **Impact of dimension size**. Small datasets of different number of RNA methylation sites are generated, to which the clustering approach was applied with or without sample weighting strategies. The clustering performance increases as the dimension (number of RNA methylation sites) increases, and the clustering method using the weighting strategies consistently outperforms the one that does not use it.

the assured RNA methylation sites with substantial dynamics in the methylation level across different experimental conditions. We selected the top 20,000 RNA methylation sites with the largest average methylation level and then the top 10,000 sites with the largest variance in methylation level among the previously selected sites. These sites show strong methylation signal and strong dynamics in the data analyzed, which are likely to capture the epitranscriptome modules induced by epitranscriptome regulators.

Before applying to the threshold-based measurement weighting strategy to the data, it is necessary to optimize its parameters *ad hoc*. To do it, small random datasets of 9 dimensions, which is the dimension of the real data used in the clustering analysis, are generated and a 2D grid search for the optimal parameters of the threshold-based weighting approach was performed as described previously. As shown in Figure 4, our result suggests that the optimal clustering result is achieved when setting $\alpha = 0.45$ and $\beta = 0.09$, which will be used in the following analysis. Please note that, under this setting, around 45% of measurements are assigned with minimal weight in the analysis, most of which are likely to be located on very lowly expressed genes, whose methylation status cannot be reliably estimated. This is consistent with our knowledge that only around half of all the genes are expressed in a specific cell type [57]. Although penalizing these RNA methylation sites may inevitably repress some patterns, our experiments suggest the overall effect is to enhance the aggregation patterns of epitranscriptome modules and thus contribute to the clustering analysis.

A major difficulty for assessing the quality of the identified epitranscriptome module is the lack of ground truth. Although there exists bioinformatics database MetDB [58] supporting the query about epitranscriptome regulation of RNA methylation sites by enzymes, this evidence has not been properly integrated and a specific regulation may be supported by only a single study, lacking consistency between different experiments. Additionally, the known enzyme genes, including RNA methyltransferases METTL3, METTL14, WTAP and demethylase FTO, and ALKBH5, although have the potential may not actually play a leading regulatory role or induce an epitranscriptome module and it is very likely that there exist additionally still unknown regulators of the $m^6A$ epitranscriptome, such as the newly identified RNA $m^6A$ methyltransferase METTL16 [39, 59]. For the aforementioned reasons, it is difficult to provide a ground truth to assess the identified epitranscriptome modules; we thus consider an alternative approach by using gene ontology (GO) as a guidance; i.e., for two epitranscriptome modules that consist of the same number of genes, the one that has more GO terms more significantly enriched is more biologically meaningful and thus more likely to represent a true epitranscriptome module than the other one [60]. It is

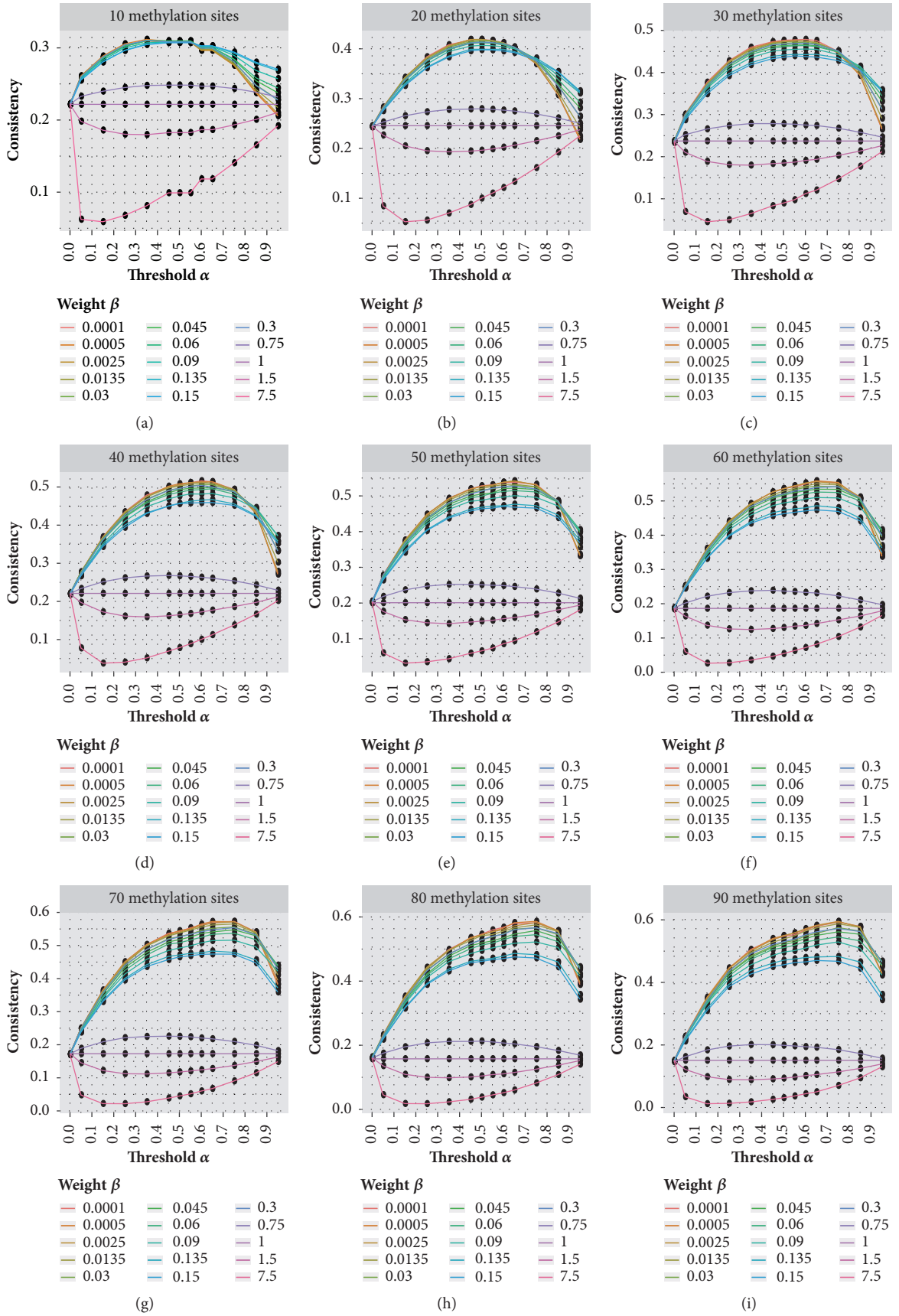


Figure 3: Continued.

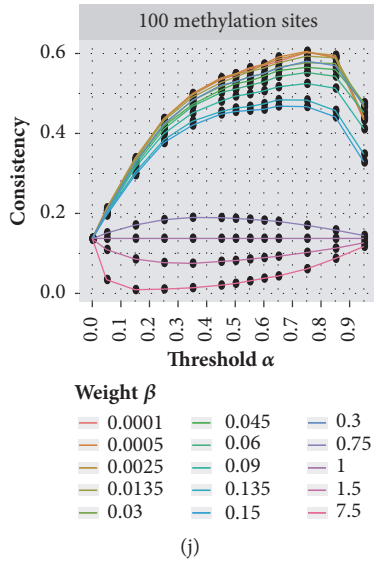


FIGURE 3: **Parameter optimization for the threshold-based method**. Small datasets are generated by sampling randomly from real RNA methylation dataset, to which clustering analysis used the threshold-based weighting strategy with different parameters and the clustering performance was evaluated by comparing to true sample labels. Better clustering performance was achieved when setting a relative small value for weight parameter $\beta$ and a medium value for threshold parameter $\alpha$.

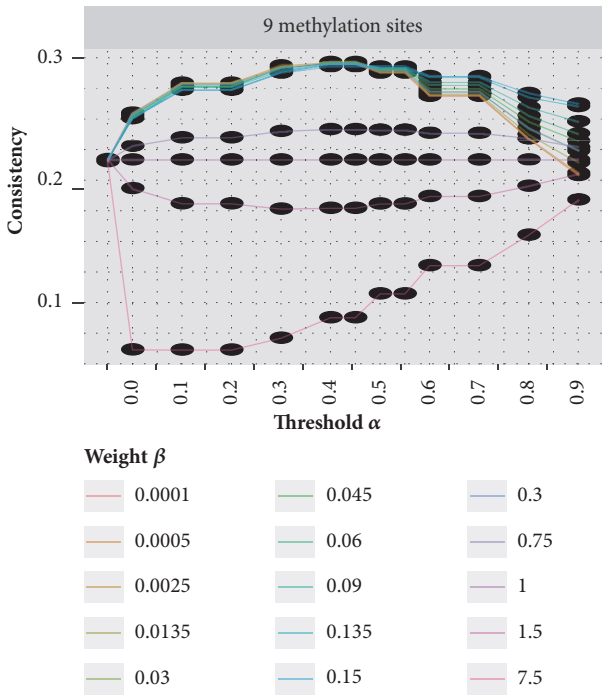


FIGURE 4: **Parameter optimization for the threshold-based method on real data**. Small datasets of the same dimension size as real dataset were generated, to which clustering analysis used the threshold-based weighting strategy with different parameters. Optimal clustering result on a dataset of 9 measurements is achieved when setting $\alpha = 0.45$ and $\beta = 0.09$, which will be adopted in the following clustering analysis on real data.

important to note that the two modules in comparison need to be of the same size, because a larger group is a lot more likely to have more GO terms enriched in it compared with a smaller group. Additionally, because the epitranscriptome modules identified from different approaches are likely to be of different size, in practice it is still difficult to compare the results from different methods under the aforementioned scheme. To solve this problem, we proposed an alternative indirect approach, in which the epitranscriptome modules identified from clustering analysis are directly compared with random modules of the size using GO analysis. If the modules identified from one approach are a lot more likely to be more biological meaningful than the random modules, while that identified from a different approach is less likely to be more biological meaningful than the random modules, then the former approach, which generated more biological meaningful results, may be considered superior to the latter one.

Specifically, 3,000 out of the total of 10000 RNA methylation sites after preprocessing are randomly selected, to which hierarchical clustering analysis was applied with or without measurement weighting strategy. The gene ontology enrichment analysis was conducted based on human gene ontology annotation database downloaded from R package org.Hs.eg.db [61] on Bioconductor. All of the three GO categories (BP, CC, and MF) were used in the enrichment analysis with the 3,000 random selected methylation sites set as the background. When calculating the biological significance of a specific epitranscriptome module, the GO terms with more than 1,000 counts in the background are considered too general and thus discarded from the analysis. The $p$ values were calculated from one sided hypergeometric test for each GO term using customized R script, and the top 20 GO terms with the most significant $p$ values were treated with negative logarithm and added together as the measurement of the biological significance of a specific module. As previously described, the clustering results (epitranscriptome modules

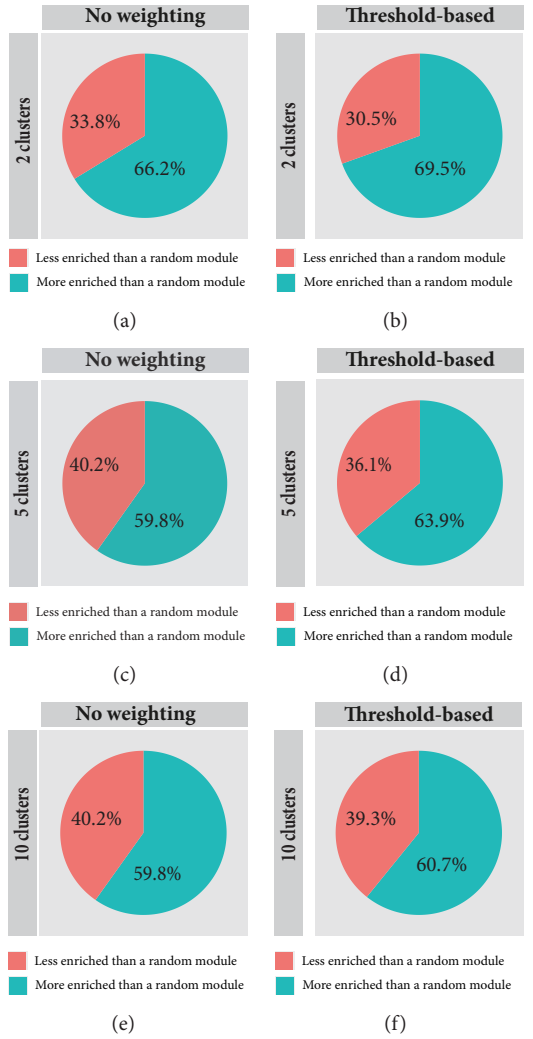


FIGURE 5: **Comparing epitranscriptome module detection based on biological significance**. The epitranscriptome modules identified from clustering analysis are always more likely to be biologically meaningful than the random modules, and this is true for clustering analysis using the measurement weighting strategy (66.2%, 59.8%, and 59.8% when $k = 2$, 5, and 10, respectively). The results obtained with measurement weighting scheme consistently outperform those obtained without measurement weighting (69.5% vs 66.2 when $k = 2$, 63.9% vs 59.8% when $k = 5$, and 60.7% vs 59.8% when $k = 10$), suggesting the proposed threshold-based measurement weighting strategy is helpful to improve clustering result and find more biological meaningful epitranscriptome modules. Clustering analysis with or without measurement weighting strategy was applied to 3000 random selected RNA methylation sites, and the epitranscriptome modules identified are compared with random group of genes of the same size in terms of biological significance using gene ontology enrichment analysis. Using bootstrap sampling approach, the analysis was repeated for 100 times and the results are summarized in this figure.

identified) of different approaches are then compared indirectly via random gene set of the same size. Please note that we used in this analysis only a fraction (3000 sites) rather than all the 10000 RNA methylation site, which is essentially the bootstrap sampling strategy for achieving a more robust results. The previous analysis was repeated 100 times to rule out the possible impact of randomness. Because the optimal number of clusters is not available, we tested 3 different settings, i.e., the number of clusters $k = 2$, 5, and 10.

As is shown in Figure 5, the epitranscriptome modules identified from clustering analysis are always more likely to be biologically meaningful than the random modules and this is true for clustering analysis using the measurement weighting strategy (66.2%, 59.8%, and 59.8% when $k = 2$, 5, and 10, respectively) or not using the measurement weighting strategy (69.5%, 63.9%, and 60.7% when $k = 2$, 5, and 10, respectively), suggesting that the epitranscriptome module not only contains a number of RNA methylation sites whose methylation states are coregulated but also carries some biological significance that can be captured using gene ontology analysis. It is the first time to be proved true on real RNA methylation data with rigorous statistical analysis that the regulatory functions are enriched in epitranscriptome modules. The results obtained with measurement weighting scheme consistently outperform those obtained without measurement weighting (69.5% vs 66.2 when $k = 2$, 63.9% vs

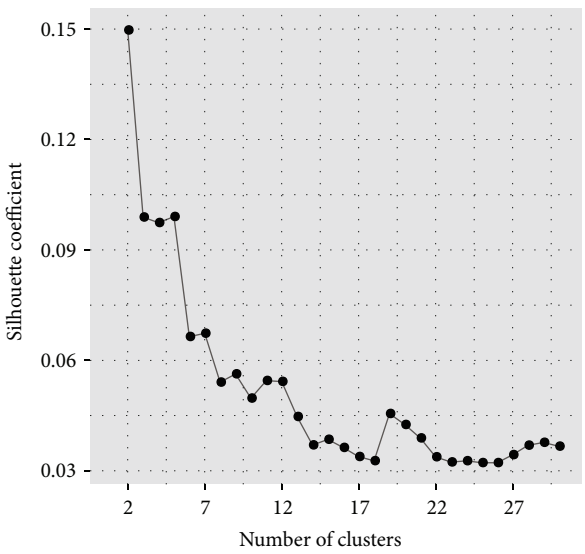


FIGURE 6: **Silhouette coefficient on the epitranscriptome data**. The Silhouette coefficient was used to assess the quality of clustering result when a different number of clusters are used ($k$). However, the largest value obtained is only 0.15, suggesting there is no clear evidence to support a specific model.

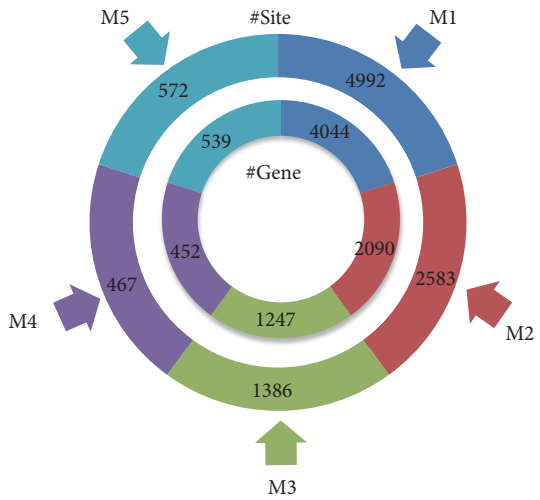


FIGURE 7: **Epitranscriptome modules identified from hierarchical clustering analysis**. Hierarchical clustering analysis of the RNA methylome identified 5 epitranscriptome modules with 4492, 2538, 1386, 467, and 572 sites, respectively, which are located on 4044, 2090, 1247, 452, and 539 genes.

59.8% when $k = 5$, and 60.7% vs 59.8% when $k = 10$), suggesting the proposed threshold-based measurement weighting strategy enhanced the clustering result and helped to find more biological meaningful epitranscriptome modules.

*3.5. The Biological Functions of Epitranscriptome Modules.* We, next, seek to explore the biological meanings of true epitranscriptome modules using the proposed measurement weighting strategy. Before clustering analysis is applied, we firstly try to use Silhouette approach [62] on all the preprocessed RNA methylation data to determine an optimal number of clusters. As shown in Figure 6, the largest Silhouette coefficient value obtained is only 0.15, suggesting there is no clear evidence to support a specific model (number of clusters). This is reasonable because that the epitranscriptome regulation is complex with multiple regulators and a single RNA methylation site can be regulated by multiple regulators simultaneously. Additionally, the epitranscriptome data is highly noisy due to the impact of transcriptome regulation and bias in sequencing. Even with the proposed approach, we may still miss true epitranscriptome modules or capture false positive patterns. Because an optimal number of clusters could not be determined, we set arbitrarily $k = 5$. The number was chosen to be not too small or too large for downstream functional analysis of the epitranscriptome modules identified.

We then applied hierarchical clustering ($k = 5$) with threshold-based measurement weighting strategy ($\alpha = 0.45$ and $\beta = 0.09$) to the entire preprocessed data to search for epitranscriptome modules. As shown in Figure 7, clustering analysis identified 5 epitranscriptome modules with 4492, 2538, 1386, 467, and 572 sites, respectively, which are located on 4044, 2090, 1247, 452, and 539 genes. It is possible that multiple RNA methylation sites located on the same gene belong to the same or different epitranscriptome modules.

The five identified epitranscriptome modules (M1-M5) are then functionally annotated using DAVID website [51] to explore their biological relevance (the complete results are available in Supplement Materials Table S4.). Distinct KEGG pathways are enriched in the modules. Notably, Huntington's disease, Parkinson's disease, Alzheimer's disease, and synaptic vesicle cycle are all enriched in the identified epitranscriptome module M2, which is consistent with our understanding of the role of RNA methylation in neurological diseases [5, 63]. The circadian rhythm pathway, which has been shown to be regulated via the epitranscriptome [2], is enriched in epitranscriptome module M2. Many cancer related pathways are also overrepresented in different epitranscriptome modules, including, transcriptional misregulation in cancer, signaling pathways regulating pluripotency of stem cells, basal cell carcinoma and microRNAs in cancer enriched in M1, pathways in cancer, and small cell lung cancer enriched in M5. Besides, pathways related to obesity, such as insulin signaling pathway and nonalcoholic fatty liver disease, are also enriched in epitranscriptome module M2, suggesting a possible relationship with FTO, which is the first obesity-related gene identified from GWAs analysis [48] and the first known RNA $m^6A$ demethylase [49]. Figure 8 shows the most enriched KEGG pathways of each epitranscriptome module.

Besides the pathway-based enriched analysis, DAVID also reveals the association between gene ontology (GO) terms and the identified epitranscriptome modules (the complete results are available in Supplement Materials Table S4.). Figure 9 shows the top 10 mostly enriched GO functions related to biological process. Interestingly, epitranscriptome module M1 is enriched with functions related to transcription (positive regulation of transcription); and M2 is enriched with positive regulation of apoptotic process, cell cycle arrest,

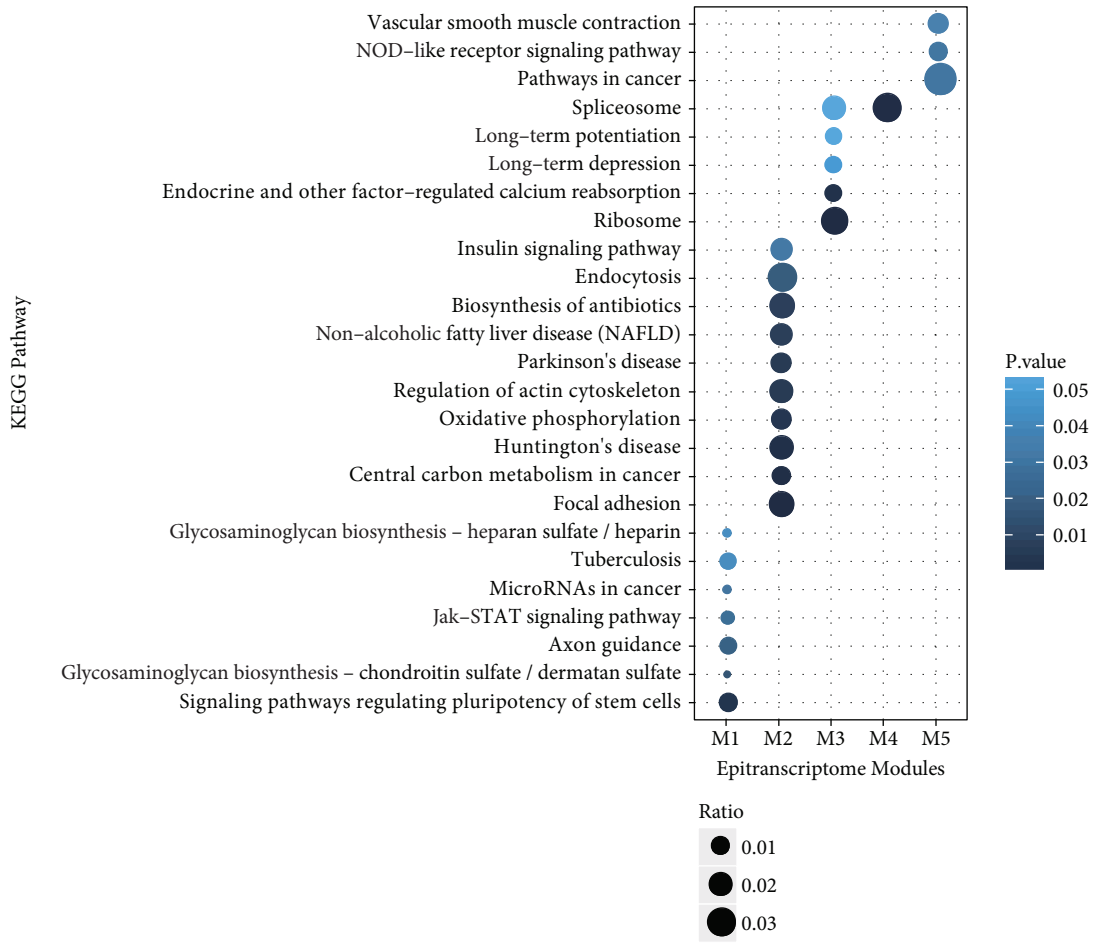


Figure 8: **KEGG pathways enriched in epitranscriptome modules**. Distinct pathways are enriched in different epitranscriptome modules. Interestingly and consistent with our understanding, insulin signaling pathway and nonalcoholic fatty liver disease are both enriched in epitranscriptome module M2 [48, 49] and axon guidance is enriched in module M1 [50]. Figure shows the top 10 most statistically enriched KEGG pathways in the identified epitranscriptome modules from DAVID [51]. Less terms are shown if there are less pathways enriched with significance level 0.05 using default setting of DAVID.

regulation of defense response to virus, and viral process; M3 is enriched with mRNA/rRNA processing, RNA splicing, DNA methylation, and translation.

To test whether the identified epitranscriptome module is potentially induced by the activity of RNA methylation enzymes, we firstly identified the WTAP-dependent methylation sites in 3 different cell lines (A549, Hela, and HEK293T) using exomePeak R/Bioconductor package by performing differential RNA methylation analysis on MeRIP-seq data obtained from WTAP knockdown and wild type conditions [31, 33]. We then compared the identified WTAP target sites and the 5 identified epitranscriptome modules. Interestingly, we found that epitranscriptome module 1 is significantly enriched in WTAP preferential target sites under all 3 conditions (A549 cell line: Odds Ratio= 3.1910, p value = 5.87E-45; HeLa cell line: Odds Ratio= 3.7395, p value = 3.94E-28; and HEK293T cell line: Odds Ratio= 2.3401, p value = 8.22E-23), suggesting it is very likely to be mediated by WTAP, a very important component of $m^6A$ RNA methyltransferase protein complex [33].

## 4. Conclusion

Due to the impact of context-specific gene expression and limitation of sequencing depth, the epitranscriptome data is highly noise and it is usually difficult to accurately quantify the methylation level of very lowly expressed genes using conventional approaches developed for ChIP-seq or RNA-seq. In order to more accurately capture the epitranscriptome modules, which reflects the regulation imposed via epitranscriptome layer, we propose to use measurement weighting strategy to penalize the measurements that are less accurate due to weak signal in sequencing data. In this study, two different types of weighted schemes (logarithm-based & threshold-based) are developed. A 2D grid search was performed to further optimize the parameters of threshold-based approach. When the proposed measurement weighting strategy is applied under a hierarchical clustering approach, we show in real data that compared with conventional approach without a measurement weighting scheme, the proposed approach can indeed help to improve the classification performance

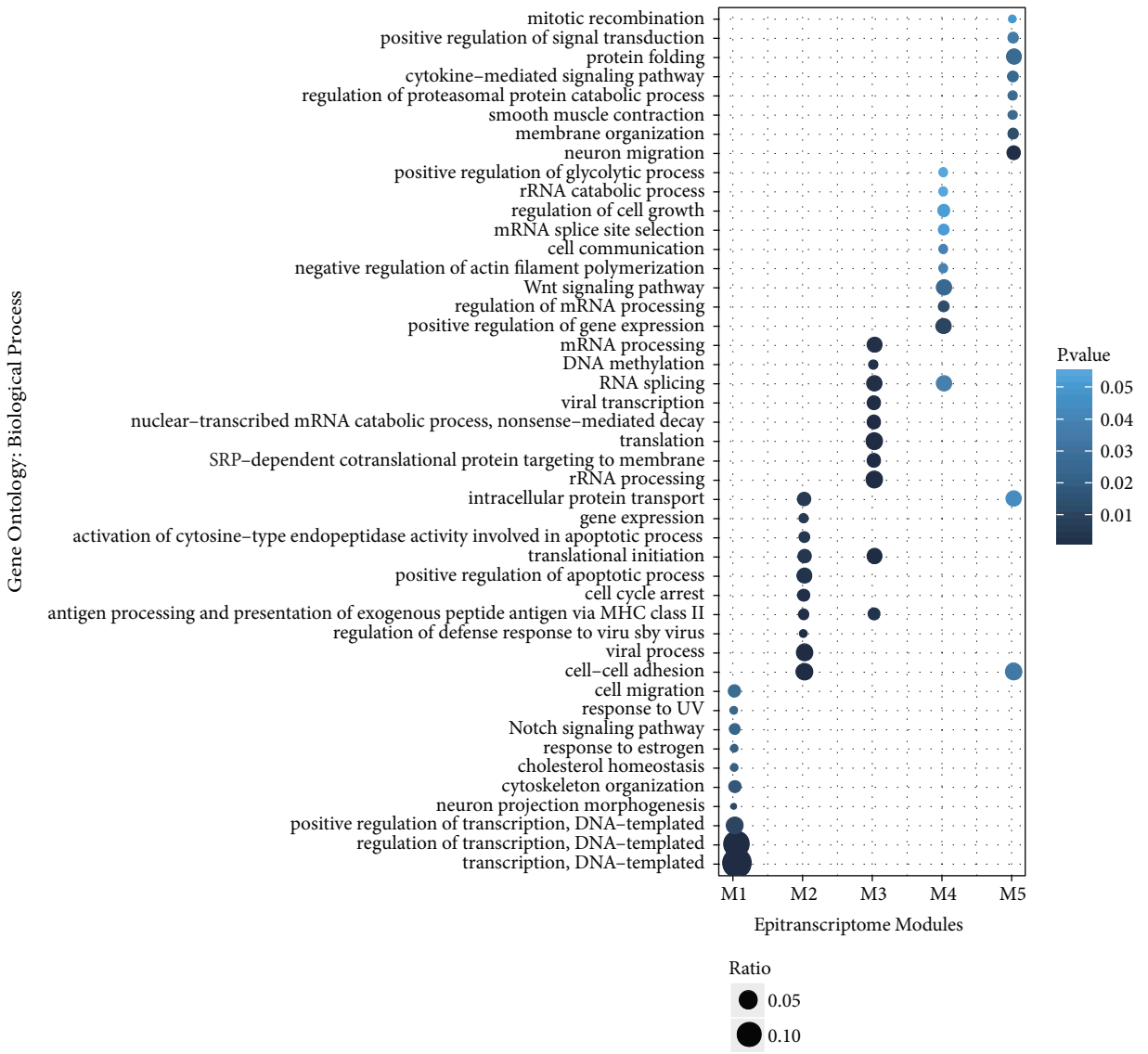


FIGURE 9: **Biological processes enriched in epitranscriptome modules**. Distinct biological processes are enriched in different epitranscriptome modules. Figure shows the top 10 most statistically enriched biological processes in the identified epitranscriptome modules from DAVID [51].

and identify more biologically meaningful epitranscriptome modules. When applied to the real dataset using the optimal parameters determined from a training process, 5 epitranscriptome modules are identified from real data with distinct biological functions linked to recent studies in the field, suggesting the potential usage of the proposed method.

The proposed method is the first approach developed for dealing with RNA $m^6A$ epitranscriptome sites with low reads coverage in a clustering analysis. Although demonstrated under a hierarchical clustering analysis framework with Euclidean distance, the proposed measurement weighing strategy is conceptually easy and can be conveniently extended to another computational analysis related to distance measurement concerning the epitranscriptome and RNA methylation, such as, K-means, K-nearest neighbor methods, and Pearson correlation, related to RNA $m^5C$ methylome. For example, we show in the Supplementary Materials that the proposed threshold-based weighting scheme is equally applicable when using squared Euclidean distance or City Block to measure the similarity of RNA methylation profiles. The approach clearly pointed out that many measurements from high-throughput sequencing data may not be accurate and need to be handled carefully to keep as much information as possible and at the same time avoid possible contamination in signal.

It is worth mentioning that using 100 repeated experiments in a bootstrap sampling analysis, we show that the epitranscriptome modules are more likely to be biologically meaningful than a random group of genes of the same size in terms of gene ontology analysis. As far as we know, this

is the first time to show with robust statistical analysis (rather than in a single isolated example) that the biological functions are enriched in epitranscriptome modules. Previously, epitranscriptome modules are considered the induced pattern of epitranscriptome regulators and are expected to emerge when a large number of RNA methylation sites are regulated by a small number of regulators [46], which explains the generation mechanism of epitranscriptome modules. Our results suggest that, besides the generation mechanism, the epitranscriptome modules also directly regulate corresponding biological functions, which justifies the regulatory aims of epitranscriptome modules. Our work established the functional basis of epitranscriptome modules, which fulfilled a key prerequisite for further functional characterization and deciphered the epitranscriptome and its regulation.

The study still has a number of limitations that may be improved. Firstly, the proposed threshold-based approach relies on two parameters that need to be optimized in data analysis. In practice, the most suitable values of the two parameters are likely to vary on different datasets, which may not be easy to determine in lack of appropriate training dataset. It would be nice to develop an easy-to-use parameter optimizing procedure for the proposed threshold-based approach or propose a nonparametric method. Secondly, due to the data availability and the lack of clear evidence for the optimal number of clusters, we explored the biological functions of epitranscriptome modules using only 9 samples and set the number of clusters $k = 5$; additionally, the clustering structure used assumes that the clusters identified are mutually exclusive; i.e., a methylation site can only belong to a single cluster. In practice, it is important to include more samples, using different number of clusters and different clustering structures such as biclustering to capture other potentially interesting epitranscriptome patterns. Thirdly, this study takes advantage of only the numeric patterns embedded in $m^6A$-seq data [24, 25] but not data from other techniques such as CLIP-based approach [64] that may capture the direct target substrate of RNA methylation-related enzymes. An integrative analysis of multiple data types that address both the generation mechanism and the regulatory aims of the epitranscriptome modules is highly desired to paint a global picture of the epitranscriptome. It would be very interesting to see how a specific epitranscriptome enzyme, e.g., FTO, regulates a specific biological function via modulating the methylation status of thousands of substrate genes.

## Abbreviation

| | |
|---|---|
| $m^6A$: | N6-methyladenosine |
| MeRIP-Seq: | Methylated RNA immunoprecipitation sequencing |
| IP: | Immunoprecipitation |
| GO: | Gene ontology |
| BP: | Biological process. |

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Authors' Contributions

Jia Meng, Rong Rong, Zhiliang Lu, and João Pedro de Magalhães conceived the idea and designed the research; Kunqi Chen, Zhen Wei, and Hui Liu implemented the analysis. Kunqi Chen drafted the manuscript. All authors read, critically revised, and approved the final manuscript. This work has also been supported by National Natural Science Foundation of China [31671373 and 61401370], Jiangsu University Natural Science Program [16KJB180027], and Jiangsu Science and Technology Program [BK20140403].

## Acknowledgments

The authors thank computational support from the UTSA Computational Systems Biology Core, funded by the National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health.

## Supplementary Materials

*Supplementary 1.* In terms of the cases when M-value, squared Euclidean, or City Block is used to quantify the RNA methylation status or the similarity of RNA methylation profiles, the results are shown in the Supplementary Materials (Tables S1-S3 and Figures S1, S2).

*Supplementary 2.* The complete gene set enrichment analysis result of the 5 epitranscriptome modules identified from real dataset is available in Supplementary Materials Table S4.

## References

[1] P. Boccaletto, M. A. Machnicka, E. Purta et al., "MODOMICS: a database of RNA modification pathways. 2017 update," *Nucleic Acids Research*, 2017.

[2] J. M. Fustin, M. Doi, Y. Yamaguchi et al., "XRNA-methylation-dependent RNA processing controls the speed of the circadian clock," *Cell*, vol. 155, no. 4, pp. 793–806, 2013.

[3] F. Liu, W. Clark, G. Luo et al., "ALKBH1-Mediated tRNA Demethylation Regulates Translation," *Cell*, vol. 167, no. 3, pp. 816–828.e16, 2016.

[4] X. Wang, B. S. Zhao, and I. A. Roundtree, "N(6)-methyladenosine modulates messenger RNA translation efficiency," *Cell*, vol. 161, no. 6, pp. 1388–1399, 2015.

[5] K.-J. Yoon, F. R. Ringeling, C. Vissers et al., "Temporal Control of Mammalian Cortical Neurogenesis by m6A Methylation," *Cell*, vol. 171, no. 4, pp. 877–889.e17, 2017.

[6] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, "N6-methyladenosine marks primary microRNAs for processing," *Nature*, vol. 519, no. 7544, pp. 482–485, 2015.

[7] I. U. Haussmann, Z. Bodi, E. Sanchez-Moran et al., "M6 A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination," *Nature*, vol. 540, no. 7632, pp. 301–304, 2016.

[8] T. Lence, J. Akhtar, M. Bayer et al., "M6A modulates neuronal functions and sex determination in Drosophila," *Nature*, vol. 540, no. 7632, pp. 242–247, 2016.

[9] H.-B. Li, J. Tong, S. Zhu et al., "M6A mRNA methylation controls T cell homeostasis by targeting the IL-7/STAT5/SOCS pathways," *Nature*, vol. 548, no. 7667, pp. 338–342, 2017.

[10] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan, "N6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions," *Nature*, vol. 518, no. 7540, pp. 560–564, 2015.

[11] J. Mauer, X. Luo, A. Blanjoie et al., "Reversible methylation of m6 Am in the 5′ cap controls mRNA stability," *Nature*, 2016.

[12] X. Wang, Z. Lu, A. Gomez et al., "N 6-methyladenosine-dependent regulation of messenger RNA stability," *Nature*, vol. 505, no. 7481, pp. 117–120, 2014.

[13] Y. Xiang, B. Laurent, C.-H. Hsu et al., "RNA m6 A methylation regulates the ultraviolet-induced DNA damage response," *Nature*, vol. 543, no. 7646, pp. 573–576, 2017.

[14] J. Zhou, J. Wan, X. Gao, X. Zhang, S. R. Jaffrey, and S.-B. Qian, "Dynamic m6 A mRNA methylation directs translational control of heat shock response," *Nature*, vol. 526, no. 7574, pp. 591–594, 2015.

[15] S. Geula, S. Moshitch-Moshkovitz, D. Dominissini et al., "m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation," *Science*, vol. 347, no. 6225, pp. 1002–1006, 2015.

[16] S. Wang, C. Sun, J. Li et al., "Roles of RNA methylation by means of N6-methyladenosine (m6A) in human cancers," *Cancer Letters*, vol. 408, pp. 112–120, 2017.

[17] X. Deng, R. Su, X. Feng, M. Wei, and J. Chen, "Role of N6-methyladenosine modification in cancer," *Current Opinion in Genetics & Development*, vol. 48, pp. 1–7, 2018.

[18] N. S. Gokhale and S. M. Horner, "RNA modifications go viral," *PLoS Pathogens*, vol. 13, no. 3, Article ID e1006188, 2017.

[19] S. Zhang, B. S. Zhao, A. Zhou et al., "m6A Demethylase ALKBH5 Maintains Tumorigenicity of Glioblastoma Stem-like Cells by Sustaining FOXM1 Expression and Cell Proliferation Program," *Cancer Cell*, vol. 31, no. 4, pp. 591–606.e6, 2017.

[20] S. Nachtergaele, L. Dong, C. Hu, X. Qin, L. Tang et al., "FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N 6-Methyladenosine RNA Demethylase," *Cancer Cell*, vol. 31, pp. 1–15, 2017.

[21] L. P. Vu, B. F. Pickering, Y. Cheng et al., "The N 6 -methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells," *Nature Medicine*, vol. 23, no. 11, pp. 1369–1376, 2017.

[22] V. Stojković and D. G. Fujimori, "Mutations in RNA methylating enzymes in disease," *Current Opinion in Chemical Biology*, vol. 41, pp. 20–27, 2017.

[23] G. Cao, H. Li, Z. Yin, and R. A. Flavell, "Recent advances in dynamic m6A RNA modification," *Open Biology*, vol. 6, no. 4, p. 160003, 2016.

[24] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons," *Cell*, vol. 149, no. 7, pp. 1635–1646, 2012.

[25] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz et al., "Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq," *Nature*, vol. 484, no. 7397, pp. 201–206, 2012.

[26] J.-J. Xuan, W.-J. Sun, P.-H. Lin, K.-R. Zhou, S. Liu, L.-L. Zheng et al., "RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data," *Nucleic Acids Research*, 2017.

[27] H. Liu, M. A. Flores, J. Meng et al., "MeT-DB: a database of transcriptome methylation in mammalian cells," *Nucleic Acids Research*, 2014.

[28] J. A. Bokar, M. E. Rath-Shambaugh, R. Ludwiczak, P. Narayan, and F. Rottman, "Characterization and partial purification of mRNA N6-adenosine methyltransferase from HeLa cell nuclei: Internal mRNA methylation requires a multisubunit complex," *The Journal of Biological Chemistry*, vol. 269, no. 26, pp. 17697–17704, 1994.

[29] P. Narayan and F. M. Rottman, "An in vitro system for accurate methylation of internal adenosine residues in messenger RNA," *Science*, vol. 242, no. 4882, pp. 1159–1162, 1988.

[30] J. A. Bokar, M. E. Shambaugh, D. Polayes, A. G. Matera, and F. M. Rottman, "Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase," *RNA*, vol. 3, no. 11, pp. 1233–1247, 1997.

[31] J. Liu, Y. Yue, D. Han et al., "A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation," *Nature Chemical Biology*, vol. 10, no. 2, pp. 93–95, 2014.

[32] X.-L. Ping, B.-F. Sun, L. Wang et al., "Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase," *Cell Research*, vol. 24, no. 2, pp. 177–189, 2014.

[33] S. Schwartz, M. R. Mumbach, M. Jovanovic et al., "Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites," *Cell Reports*, vol. 8, no. 1, pp. 284–296, 2014.

[34] P. Wang, K. A. Doxtader, and Y. Nam, "Structural Basis for Cooperative Function of Mettl3 and Mettl14 Methyltransferases," *Molecular Cell*, vol. 63, no. 2, pp. 306–317, 2016.

[35] H. Yin, H. Wang, W. Jiang, Y. Zhou, and S. Ai, "Electrochemical immunosensor for N6-methyladenosine detection in human cell lines based on biotin-streptavidin system and silver-SiO2 signal amplification," *Biosensors and Bioelectronics*, 2016.

[36] K. I. Zhou and T. Pan, "Structures of the m6A methyltransferase complex: two subunits with distinct but coordinated roles," *Molecular Cell*, vol. 63, no. 2, pp. 183–185, 2016.

[37] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang et al., "N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO," *Nature Chemical Biology*, vol. 7, no. 12, pp. 885–887, 2011.

[38] G. Zheng, J. A. Dahl, Y. Niu et al., "ALKBH5 is a aammalian RNA demethylase that impacts RNA metabolism and mouse fertility," *Molecular Cell*, vol. 49, no. 1, pp. 18–29, 2013.

[39] A. S. Warda, J. Kretschmer, P. Hackert et al., "Human METTL16 is a N6-methyladenosine (m6A) methyltransferase that targets pre-mRNAs and various non-coding RNAs," *EMBO Reports*, 2017.

[40] G. Hu, J. Kim, Q. Xu, Y. Leng, S. H. Orkin, and S. J. Elledge, "A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal," *Genes & Development*, vol. 23, no. 7, pp. 837–848, 2009.

[41] O. Hirose, R. Yoshida, S. Imoto et al., "Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models," *Bioinformatics*, vol. 24, no. 7, pp. 932–942, 2008.

[42] S. Horvath, Y. Zhang, P. Langfelder et al., "Aging effects on DNA methylation modules in human brain and blood tissue," *Genome Biology*, vol. 13, no. 10, p. R97, 2012.

[43] F. Wang, H. Xu, H. Zhao, J. Gelernter, and H. Zhang, "DNA co-methylation modules in postmortem prefrontal cortex tissues of European Australians with alcohol use disorders," *Scientific Reports*, vol. 6, Article ID 19430, 2016.

[44] Q. Liao, C. Liu, X. Yuan et al., "Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network," *Nucleic Acids Research*, vol. 39, no. 9, pp. 3864–3878, 2011.

[45] X. Guo, L. Gao, Q. Liao et al., "Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks," *Nucleic Acids Research*, 2012.

[46] L. Liu, S. Zhang, Y. Zhang et al., "Decomposition of RNA methylome reveals co-methylation patterns induced by latent enzymatic regulators of the epitranscriptome," *Molecular BioSystems*, vol. 11, no. 1, pp. 262–274, 2015.

[47] P. Du, X. Zhang, C.-C. Huang et al., "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, article 587, 2010.

[48] K. A. Fawcett and I. Barroso, "The genetics of obesity: FTO leads the way," *Trends in Genetics*, vol. 26, no. 6, pp. 266–274, 2010.

[49] G. Jia, Y. Fu, X. Zhao et al., "N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO," *Nature Chemical Biology*, vol. 7, no. 12, pp. 885–887, 2011.

[50] J. Yu, M. Chen, H. Huang et al., "Dynamic m6A modification regulates local translation of mRNA in axons," *Nucleic Acids Research*, 2017.

[51] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[52] M. Dehmer, *Applied statistics for network biology: methods in systems biology*, Wiley-Blackwell, Weinheim, Germany, 2011.

[53] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, article R36, 2013.

[54] J. Meng, Z. Lu, H. Liu et al., "A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package," *Methods*, vol. 69, no. 3, pp. 274–281, 2014.

[55] M. Lawrence, W. Huber, H. Pagès et al., "Software for computing and annotating genomic ranges," *PLoS Computational Biology*, vol. 9, no. 8, Article ID e1003118, 2013.

[56] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons," *Cell*, vol. 149, no. 7, pp. 1635–1646, 2012.

[57] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell*, Garland Science, New York, NY, USA, 4th edition, 2002.

[58] H. Liu, H. Wang, Z. Wei et al., "MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome," *Nucleic Acids Research*, 2017.

[59] H. Shima, M. Matsumoto, Y. Ishigami et al., "S-Adenosylmethionine synthesis is regulated by selective N6-adenosine methylation and mRNA degradation involving METTL16 and YTHDC1," *Cell Reports*, vol. 21, no. 2, pp. 3354–3363, 2017.

[60] J. Meng, S.-J. Gao, and Y. Huang, "Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules," *Bioinformatics*, vol. 25, no. 12, pp. 1521–1527, 2009.

[61] H. Shi, X. Wang, Z. Lu et al., "YTHDF3 facilitates translation and decay of N 6-methyladenosine-modified RNA," *Cell Research*, vol. 27, no. 3, pp. 315–328, 2017.

[62] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[63] L. Li, L. Zang, F. Zhang et al., "Fat mass and obesity-associated (FTO) protein regulates adult neurogenesis," *Human Molecular Genetics*, vol. 26, no. 13, pp. 2398–2411, 2017.

[64] M. Hafner, M. Landthaler, L. Burger et al., "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP," *Cell*, vol. 141, no. 1, pp. 129–141, 2010.