UNIVERSITY OF LIVERPOOL

Large-scale functional annotation of individual RNA methylation sites by mining complex biological networks

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Xiangyu Wu

September 2020

Declaration

I confirm that:

- I have read and understood the University's PGR Policy on Plagiarism and Dishonest Use of Data.
- I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.
- I have not copied material from another source nor committed plagiarism nor fabricated, falsified or embellished data when completing the attached material.
- I have not copied material from another source, nor colluded with any other student in the preparation and production of this material.
- If an allegation of suspected academic malpractice is made, I give permission to the University to use source-matching software to ensure that the submitted material is all my own work.

Abstract

Increasing evidences suggest that post-transcriptional RNA modifications regulate essential biomolecular functions and are related to the pathogenesis of various diseases. To date, the study of epitranscriptome layer gene regulation is mostly focused on the function of mediator proteins of RNA methylation limited by laborious experimental procedures, i.e., the readers, writers and erasers. However, there is limited investigation of the functional relevance of individual m^6A RNA methylation sites. To address this, we annotated human m^6A sites in large-scale based on the guilt-by-association principle from complex biological networks. In the first chapter, the network was constructed based on public human MeRIP-Seq datasets profiling the m^6A epitranscriptome under independent experimental conditions. By systematically examining the network characteristics obtained from the RNA methylation profiles, a total of 339,158 putative gene ontology functions associated with 1446 human m^6A sites were identified. These are biological functions that may be regulated at epitranscriptome layer via reversible m^6A RNA methylation. The results were further validated on a soft benchmark by comparing to a random predictor. In the second chapter, another approach was applied to annotate the individual human m^6A sites by integrating the methylation profile, gene expression profile and protein-protein interaction network with guilt-by-association principle. The consensus signals on sites were amplified by multiplying the co-methylation network and the methylation-expression network. The PPI network smoothed the correlation for a query site to gene expression for furthering GSEA functional annotation. In the third chapter, we functionally annotated 18,886 m^6A sites that are conserved between human and mouse from a larger epitranscriptome datasets using

method previously described. Besides, we also completed two side projects related to SARS-CoV-2 viral m^6A site prediction and m^6A site prediction from Nanopore sequencing technology.

Acknowledgements

I would like to thank Dr. Jia Meng, Prof. Zhiliang Lu, Dr. Rong Rong, Dr. Lin Zhang and Prof. Joao Pedro for their supervision of my PhD studies. Dr. Zhen Wei for verifying the reduction of the batch effect and technical bias. Kunqi Chen for building the web server. Qing Zhang, Jialin Ma, Bowen Song and other students in Jia's group for discussion during research. Jing Qian for her support and accompany. Researches in the thesis were supported by National Natural Science Foundation of China [31671373 and 31871337]; Jiangsu University Natural Science Program [16KJB180027]; XJTLU Key Programme Special Fund [KSF-T-01]; Jiangsu Six Talent Peak Program [XYDXX-118].

.

Publications and Author Contribution Statements

For Chapter 1, Jia Meng and Lin Zhang conceived the idea and designed the research; Zhen Wei and Hui Liu collected and processed the raw data; Zhen Wei performed the GC content correction and verified the reduction of the batch effect and technical bias; Xiangyu Wu, Zhen Wei, Qing Zhang and Jionglong Su constructed the network analysis, evaluated the performance, characterized the network, and performed other data analysis tasks; Kunqi Chen built the website. Xiangyu Wu drafted the manuscript.

For Charpter 2, Jia Meng, Rong Rong, Zhiliang Lu and Joao Pedro Magalhaes conceived the idea and designed the research; Kunqi Chen performed the m^6A site prediction; Xiangyu Wu and Qing Zhang performed the network-based functional annotation of individual m^6A sites; Kunqi Chen built the website; Kunqi Chen, Xiangyu Wu, Qing Zhang and Wei Zhen drafted the manuscript.

For Chapter 3, Kunqi Chen conceived the idea and initialized the project; Kunqi Chen and Jiongming Ma collected and processed the epitranscriptome data of eukaryotes and viruses, respectively; Bowen Song generated post-transcriptional annotations, conducted disease-association analysis and conservation analysis; Xiangyu Wu performed GO function prediction; Yujiao Tang designed and built the m^6A-Atlas website. For the SARS-CoV2 project, Jia Meng and Xiangyu Wu conceived the idea and design the research; Qingru Xu and Xiangyu Wu performed the raw data generation and site prediction. Qingru Xu performed most frequent m^6A sites and

stop codon position comparison. For the Nanopore project, Jia Meng, Xiangyu Wu and Jionglong Su conceived the idea and design the research; Xiangyu Wu generated the raw data; Shen Jia, Haochen Luo, Qiheng Gao and Jiaqi Guo performed the machine learning approach.

Related publications include:

1. Wu Xiangyu, Wei Zhen, Chen Kunqi, Zhang Qing, Su Jionglong, Liu Hui, Zhang Lin, Meng Jia*: **m6Acomet: large-scale functional prediction of individual m(6)A RNA methylation sites from an RNA co-methylation network**. *BMC Bioinformatics* 2019, **20**(1):223.
2. Chen Kunqi#, Wei Zhen#, Zhang Qing#, Wu Xiangyu#, Rong Rong, Lu Zhiliang, Su Jionglong, de Magalhaes Joao Pedro, Rigden Daniel J, Meng Jia*: **WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach**. Nucleic Acids Res 2019, 47(7):e41.
3. Tang Yujiao#, Chen Kunqi#, Song Bowen#, Ma Jiongming#, Wu Xiangyu#, Xu Qingru, Wei Zhen, Su Jionglong, Liu Gang, Rong Rong et al: **m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome**. Nucleic Acids Res 2020.
4. Shen Jia, Haochen Luo, Qiheng Gao, Jiaqi Guo, Jionglong Su, Jia Meng and Xiangyu Wu* (2019). **Detection of m6A RNA methylation in Nanopore sequencing data using support vector machine. 2019 12th International Congress on Image and Signal Processing**, BioMedical Engineering and Informatics (CISP-BMEI 2019), Suzhou, China, Oct. 19-21, 2019.

Table of Contents

List of Abbreviations

m^{6}A = N6-methyladenosine

MeRIP-Seq = Methylated RNA immunoprecipitation sequencing

GC = Guanine-cytosine

GO = Gene Ontology

BP = biological process

CC = cellular component

MF = molecular function

MCL = Markov Cluster algorithm

scc = Spearman correlation coefficient

SNP = single nucleotide polymorphism

GSEA = Gene Set Enrichment Analysis

PPI = Protein-Protein interaction

RPKM = Reads Per Kilobase per Million mapped reads \

GMM = graph for methylation-methylation

GME = graph for methylation-expression

FDR = false discovery rate

lncRNA= long non-coding RNA

mRNA = messenger RNA

SVM = support vector machine

Chapter 1 **m^{6}Acomet: large-scale functional prediction of individual m^{6}A RNA methylation sites from an RNA co-methylation network**

Section 1.1 Introduction

N6-methyladenosine (m^{6}A) is one of the most common RNA post-transcriptional chemical modifications. It is formed with an addition of a methyl group at the 6' position of adenosine in RNA [1]. It is abundant in mRNA, snRNA and rRNA among plants, viruses and eukaryotes [2, 3]. In mammals, methyltransferases (m^{6}A writer), such as METTL3, METTL14 and WTAP, together with demethylases (m^{6}A eraser), and YTH domain family of proteins (m^{6}A reader), regulate the complex reverse mechanism of m^{6}A [4]. The m^{6}A was found to influence diverse biological regulations such as RNA stability [5], heat shock response [6], and circadian clock [7] etc. Diseases, such as cancer [8] are proved to be regulated by m^{6}A as well. Current research focuses more on the overall functions or regulations involving m^{6}A. However, the biological function of each individual RNA methylation sites is not exactly known. Although the regulatory roles of several specific methylation sites have been elucidated, it is very expensive to identify the functions of RNA methylation sites with wet-lab experiments. Instead, computational approach may provide a viable venue. It is possible that the functions of each individual RNA methylation sites can be predicted from the statistical evidence such as strong correlation with the expression level of genes whose functions are already known.

The regulatory functions of methylation sites in biological processes are still under research [9-11]. It is conceivable to assume that m^{6}A sites that have similar properties

would share similar biological functions. Indeed, our previous studies showed that the RNA methylation sites consisting of an epitranscriptome module, which is a number of RNA methylation sites whose methylation level are co-regulated across different experimental conditions, are more likely to be functionally enriched compared to a random module [12, 13]. This strongly suggests that the epitranscriptome functions of the RNA methylation sites may be identified based on existing high-throughput sequencing data. It is meaningful to investigate the regulatory role of these sites by constructing the co-methylation network with the guilt-by-association principle. The guilt-by-association is a validated principle in network research, which states that if two patterns share some similar properties, they are most likely to share a connection. To be more specific, gene pairs are more likely to be functionally related if they show similar expression patterns across samples [14]. This principle has been widely applied in lncRNA functional prediction by the protein-protein interaction network [15], co-transcription factor network, and co-expression network [14]. In our research, we suppose that, if both methylation sites are hyper- or hypo-methylated simultaneously across various samples, they will be considered co-methylated and often of related biological interests. In the co-methylation network, each node represents a methylation site, and each edge denotes a strong correlation or anti-correlation between each pair of sites.

The datasets used for generating the methylation level on sites in this program are all produced by the MeRIP-Seq technique [1, 16]. Methylated RNA immunoprecipitation sequencing (MeRIP-Seq) technique was developed to investigate m^6A in epitranscriptome analysis [17]. The mRNAs which contain m^6A

sites are first fragmented into short pieces of ~ 100 bp long, following which fragments with methylation sites are filtered by antibodies in immunoprecipitate as IP samples, while raw fragments are treated as Input control samples [18]. After mapping both the reads of eluted IP sample and control (Input) sample back to the reference genome, the peak-calling or methylation evaluation algorithm will be employed to detect the m^6A peaks for furthering investigation [19]. The mi-CLIP [20] and the m^6A-CLIP [21] were developed recently to generate single-base resolution m^6A profile, and the upcoming data sets were utilized to obtain the m^6A sites directly in the project. The principle of mi-CLIP is to bind the cross-linking RNA-m^6A antibody to specific sites where mutagenesis will occur during reverse transcription of the antibody-bound RNA. Trucations or C-T transitions, which are mutagenesis signatures, can be sequenced to precisely map m^6A sites. The m^6A-CLIP located thousands of m^6A residues using cross-linking immunoprecipitation technique (UV CLIP) with high accuracy since only the m^6A-containing oligonucleotide can attract the m^6A antibody.

Before constructing the co-methylation network, the matrix which gives the methylation level on each site over various samples needs to be constructed. However, preprocessing is required for the raw data due to technical or biological biases. The *DESeq2* [22] is a R package which uses shrinkage estimation for fold changes, and dispersion for gene-level differential analysis with RNA-Seq data. The reproducibility and stability of results are improved by shrinkage estimators after using DESeq2. This algorithm can reduce type-I errors and offer consistent performance on small studies. Guanine-cytosine (GC) content is one of the critical technical variabilities. It was shown to have significant impact on m^6A-seq [23] and

other sequencing techniques such as RNA-seq and ChIP-seq. The CQN algorithm developed by Hansen [24] is aimed to reduce systematic bias in GC content. It combines the robust generalized regression and conditional quantile normalization to improve the precision of gene expression level measurement. In our project, *DESeq2* and *CQN* are applied to estimate the methylation level of each m^6A site.

After building the complex network, cellular modules were identified for further annotation with gene ontology (GO). The GO is a bioinformatics initiation to unify gene product across species [25]. We mainly used GO for annotating on gene sets to describe the functions of a specific gene list. The GO enrichment analysis will determine which GO terms are over represented, generate the GO term list with statistical evidence such as the p-value. The GO terms may be classified into three main categories: biological process (BP), cellular component (CC) and molecular function (MF). To improve the annotation performance, the enriched GO terms will be reduced to generic GO slim terms, by skipping specific fine-grained terms, which is useful when board classifications of function annotation are required [26].

In this project, we computationally predicted the biological functions that are likely to be associated with individual m^6A RNA methylation sites. We used bioinformatics methods such as clustering, network topological analysis, as well as enrichment analysis for functional annotation in this project. The results may be queried directly on a public webserver, which provides predicted functions for each individual RNA methylation sites.

Section 1.2 Methods

Section 1.2.1 mi-CLIP and m^{6}A-CLIP supported m^{6}A sites

A total of 69,446 human m^{6}A sites reported by six mi-CLIP and m^{6}A-CLIP experiments, which profiles the m^{6}A epitranscriptome at base-resolution, were obtained from the WHISTLE project [20, 21, 27-29]. The m^{6}A sites were labeled positive and retained for the following analysis if it embodies the DRACH consensus motifs of m^{6}A modification and were supported by at least two out of the total six samples.

Section 1.2.2 MeRIP-Seq data for quantifying the RNA methylation level

The mi-CLIP and m^{6}A-CLIP report only the location of the methylation site, but do not provide direct quantification of the methylation level of these sites. The information of the methylation level was obtained from MeRIP-Seq data. Specifically, 32 samples in 10 publicly human m^{6}A MeRIP-Seq data sets from published studies were obtained from public database. All these samples contain both IP and Input data, and most of them were selected from the epitranscriptome database MeT-DB [30], with which it is now possible to construct the RNA co-methylation network. The biological replicates under the same cell line and from the same laboratory were merged, and the methylation level of the combined sample is essentially the average of all the biological replicates. Moreover, several outlier samples such as the sample from HepG2 cell line with heat shock treatment were dropped before the construction of the network due to low quality. Table 1 summarizes the data sets used in this project. All the original data were down loaded in SRA format from Gene Expression Omnibus, and the reads were aligned to human reference genome (hg19/GRCh37) with aligner *Tophat2* [31].

Table 1 Datasets used in Section 1.

ID	GEO accession	Cell line	Treatment	Source
1-4	SRR456542-SRR456549, SRR456551-SRR456557	HepG2	UV, HGF, IFN, UT	[17]
5-6	SRR903368-SRR903379	U2OS	CTL, DAA	[32]
7-10	SRR847358-SRR847377	HeLa	Ctrl, METTL14-, METTL3-, WTAP-	[33]
11-12	SRR1182582-SRR1182590	ES/NPC	hNPC, hESC	[34]
13-18	SRR1182591-SRR1182596, SRR494613-SRR494618, SRR5080301-SRR50312	Hek293T, Hek293A	Ctrl, WTAP-, METTL3-, METTL16-	
19-21	SRR1182597-SRR1182602	OKMS	D0, D5_WITH_DOX, D5_WO_DOX	
22-26	SRR1182603-SRR1182630	A549	Ctrl, METTL14-, METTL3-, WTAP-, KIAA1429-	
27-28	SRR3066062-SRR3066069	AML	Ctrl, FTO+	[35]
29-30	SRR5239086-SRR5239109	AML2	Ctrl, METTL3-	[36]
31-32	SRR1035213-SRR1035224	ESC	T0, T48	[37]

Section 1.2.3 Processing the methylation data

The R package DESeq2 [22] was applied to estimate the methylation level at each m^6A site. All the samples were labeled with conditions (IP and Input) and sequence types (Single-end and Paired-end), and the reads count matrix was generated by counting the reads which share overlaps with bins. These bins are 101 bp long with each methylation site located at the center. The methylation level was then quantified by calculating the fold enrichment of reads in the IP sample compared with the input control sample with DESeq2, which uses shrinkage estimation and considers the over-dispersion of reads. This step produces the quantification result of the logarithmic fold change indicating the methylation level of each site. However, we

found that conditions from the same laboratory cell line could not be classified into the same group by hierarchical clustering. We suppose that the GC content, which is the common systematic bias when dealing with RNA-seq data, could be further reduced. Therefore, the output of DESeq2 was first normalized by package CQN [24] to reduce the GC bias. After this additional bias correction, the estimated methylation level after the normalization by CQN does not show any GC content bias, and we can see that the conditions from the same cell line could be clustered together.

Section 1.2.4 Site filtering

The methylation sites need to be filtered due to low estimation accuracy on part of the raw sites. These sites were filtered by the following steps:

i. The methylation level will be masked NA if the expression value is lower than 8, or the count number on (IP + Input) samples of the same site is lower than 50. Throughout all the 32 conditions, sites should be dropped if too many missing values (NA count > 15) occur.

ii. Filtering the neighboring sites helps reduce the influence of replication on functional prediction. If the distance between two sites is too small, e.g., less than 50 bp, due to limited resolution of the m^6A seq technology, it is highly possible that they are located on the same gene and be annotated with the same function. We ought to keep one of them for further annotation. The Spearman correlation between two methylation sites which are located closer than 101 bp is calculated. If the correlation between them is above 0.8, they

may be in fact corresponding to the same m^6A site but incorrectly captured twice due to limited resolution of the m^6A-seq technology. In that case, the site with lower methylation level will be dropped.

iii. Since a larger variance among different conditions indicates more obvious functions, sites with median absolute deviation of methylation level value across different conditions higher than 0.4 will be retained.

After site filtering, the site number was reduced from 69,446 to 13,415, following which quantile normalization was performed to remove potential batch effect.

Section 1.2.5 Construction of the RNA co-methylation network

The RNA methylation data which contains the methylation level of 13,415 sites over 32 conditions was used to construct the RNA co-methylation network. In the beginning, the Spearman correlation between each site pair was computed. Fisher's asymptotic distribution was applied to estimate the p-value of each Spearman correlation coefficient (scc), and the p-value of each site pair was adjusted with Bonferroni method. The scc p-value for each gene pair with Fisher's asymptotic test was implemented with function *corPvalueFisher* in package *WGCNA* [38]. The p-values were adjusted with Bonferroni method with function *mt.rawp2adjp* in package *multtest*. Site pairs with high spearman correlation (correlation value ranked in the top or bottom 10%) and low p-value (lower than 0.05) for their methylation levels are regarded to be significant co-methylation pair. The adjacency matrix was then built to denote the correlation in methylation level between each pair of sites.

To build the network, function *graph.adjacency* in package igraph was applied to create the graphs file, and the degree distribution of this co-methylation network can be visualized. The power-law degree distribution indicates that our co-methylation network is a typical scale-free network [39], which means that the majority of the nodes in the network are connected with few other nodes, while the minority of the hub nodes are connected with plenty of nodes. Moreover, the network topological property will be visualized and analyzed with the professional network investigation software such as Cytoscape [40]. The function *exportNetworkToCytoscape* in package *WGCNA* can export the file from the adjacency matrix for visualization in Cytoscape [40].

The function of each m^6A site was annotated with two different algorithms: hub-based method and module-based method.

Section 1.2.6 The hub-based method

From the degree distribution of our co-methylation network, it is observed that the minority of hub sites are related to large number of methylation sites. Since these hub sites play a significant role in the whole network, it would be of interest to investigate their functions in human biological process. The function neighbors in package igraph [41] helped us find the neighboring sites of each m^6A site.

In hub-based method, the function of the hub methylation site is determined by the enrichment result of its neighbor sites, and only the sites with more than three immediate neighbor sites are treated as the hub sites. A total of 1889 hub sites

remained if only those with more than 3 edges are considered. Before annotating the function, we assume that the functions of the methylation sites are the same as the ones in their corresponding genes. The Entrez Gene ID and Gene Symbol of the gene corresponding to each hub site and their neighboring sites were labeled for annotation. Therefore, the enriched GO BP terms of the neighboring sites on the genes may convey the function of the hub site. In addition, the functions of the gene where the hub-site located can reflect the role of the hub site. We performed the GO BP enrichment analysis on the corresponding gene of the hub site and the corresponding genes of its neighboring sites. The GO slim terms, which are the subset of GO term, were applied to reduce the GO enriched terms. This generic subset is used as the scope of GO Slim. Since the term GO:0008150 (biological process) is too general, it was removed from the analysis as well. The consistent terms between GO Slim BP terms of the hub gene, where each hub site located, and GO Slim BP terms of the neighbor genes, where the neighbor sites of the same hub site located, were treated as the reliably predicted functions of each hub site. To evaluate the prediction performance of the predicted GO terms, the functional enrichment p-value of the slim term (PV) and the number of the enriched slim terms of neighbor sites (GN) are calculated and set as the cutoff parameters. The recall and precision of the prediction performance are defined as **Equation 1** and **Equation 2**.

Equation 1

$$\text{Recall} = \frac{\sum \text{Both known \& predicted GO term number}}{\sum \text{Known GO term number}}$$

Equation 2

$$\text{Precision} = \frac{\sum \text{Both known \& predicted GO term number}}{\sum \text{Predicted GO term number}}$$

GO terms of the hub site, the corresponding neighbors, and their overlap on each PV and GN cutoff were generated separately by previous data. The values of recall and precision for each cutoff were calculated to evaluate the prediction performance.

Section 1.2.7 Permutation on the network

To assess the efficacies of the predicted GO terms, permutation on sites was performed to rebuild the random network. If both the recall and precision values of real network are much higher than that of the random network, the predicted functions in the functional network should be biologically significant. The functions *rewire* and *keeping_degseq* in the ig raph package were used to randomly rewire the edges without creating multiple edges, keeping the degree distribution of the raw graph unchanged without loop edges. The rewiring algorithm substitutes two arbitrary edges in each step ((a, b) and (c, d)) with the edges which are not existed in the raw graph as ((a, d) and (c, b)). The exchanging steps were repeated 100 times for the original graph. After the permutation, the number of neighbors of the same site does not change. This is similarly carried out on the random network as well. Since it is highly possible that the neighboring sites of the same hub site correspond to the same gene in the real network, this might result in a lower overall p-value of terms annotated in the random network. We constrained the neighbor gene number of the same hub site in the random network as that in the real network. The GO and GO slim terms enriched in permutation network together with other parameters were

used to predict the function of methylation sites. The results of the random network were compared to that of the real network in terms of the performance of prediction.

Section 1.2.8 The module-based method

Another way to predict the site function is to investigate the modules in the network. It is common that sites among the same co-methylated module share similar functions. In the module-based method, the Markov Cluster (MCL) algorithm [42] is chosen as the clustering algorithm in grouping methylation sites. MCL is a scalable cluster algorithm, which is based on the stochastic flow in graphs to identify modules with random walk. We transformed the co-methylation network into the MCL input format, which contained the information of two nodes (sites) and the edge weight between nodes. With the inflation value set to 1.4 by default, the modules containing more than 9 sites were identified to be significant modules. These clustered sites will then be annotated according to the Gene Ontology of their hosting genes. The terms of the same site annotated by module-based method and hub-based method were then compared to test the annotation accuracy.

Section 1.3 Results

Section 1.3.1 Selection of raw m^6A sites and normalization

After filtering the methylation sites corresponding to lowly expressed genes with low gene expression level and low read count quantity in IP and Input samples, the raw predicted single-base resolution human m^6A sites were reduced from 69,446 to 36,542. Furthermore, 17,758 sites were discarded to reduce the number of neighboring sites corresponding to the same gene that are very close to each other.

A total of 13,415 sites with relatively higher median absolute deviation (over 0.4) among the remained 18,784 sites were kept. We believe that the m^6A sites remained after selection should be statistically significant. After merging the biological replicates from the same condition together with methylation level estimation by DESeq2, the GC content bias were normalized by CQN. The dendrograms in **Figure 1**, constructed using Euclidean distance as the metric, helped us insight into the joint distribution between samples with and without the CQN normalization. Samples from the same cell line and experiment were labelled with the same color. Samples of the same color were not clustered together in the dendrogram without CQN (**Figure 1a**). In contrast, after the CQN normalization, almost all the conditions from the same cell line were clustered into the same group with highly correlated methylation patterns (**Figure 1b**). This indicates that the GC content biases were removed. Additionally, we tested the relative importance of each individual samples. To test for sample independence, we removed each individual sample from the original 32 samples in the methylation-level matrix to build the adjacency matrix. Since the matrix generated from previous section is filled with “0” and “1” and the dimensions of two matrices are the same as well, we can compare their topological similar by calculating the odds ratio (OR) between the adjacency matrix of original and new one generated with one sample removed. The histogram of odds ratios between adjacency matrices built by all the 32 samples and with one sample removed is shown in **Figure 2a**. All the OR values are large, ranging between 2400 and 2700, which means that the topological connections are 2000 times more likely to be consistent with each other compared with the random permutation. There are no obvious outliers corresponding to samples that will induce substantial topological

changes to the co-methylation network. Besides, although the samples perturbed with m^6A enzymes represent an abnormal kind of methylation profiles and may induce more bias, To prove whether samples with m^6A enzymes (METTL3, METTL14, FTO etc.) perturbation would induce bias to the co-methylation network. We followed similar procedure described previously. As shown in **Figure 2b**, the topological changes induced by samples with enzyme permutation are actually slightly smaller than the other samples, as indicated by higher consistency between the adjacency matrices. Given that the number of MeRIP-seq samples is very limited, we believe it is better to keep all samples for the following analysis.

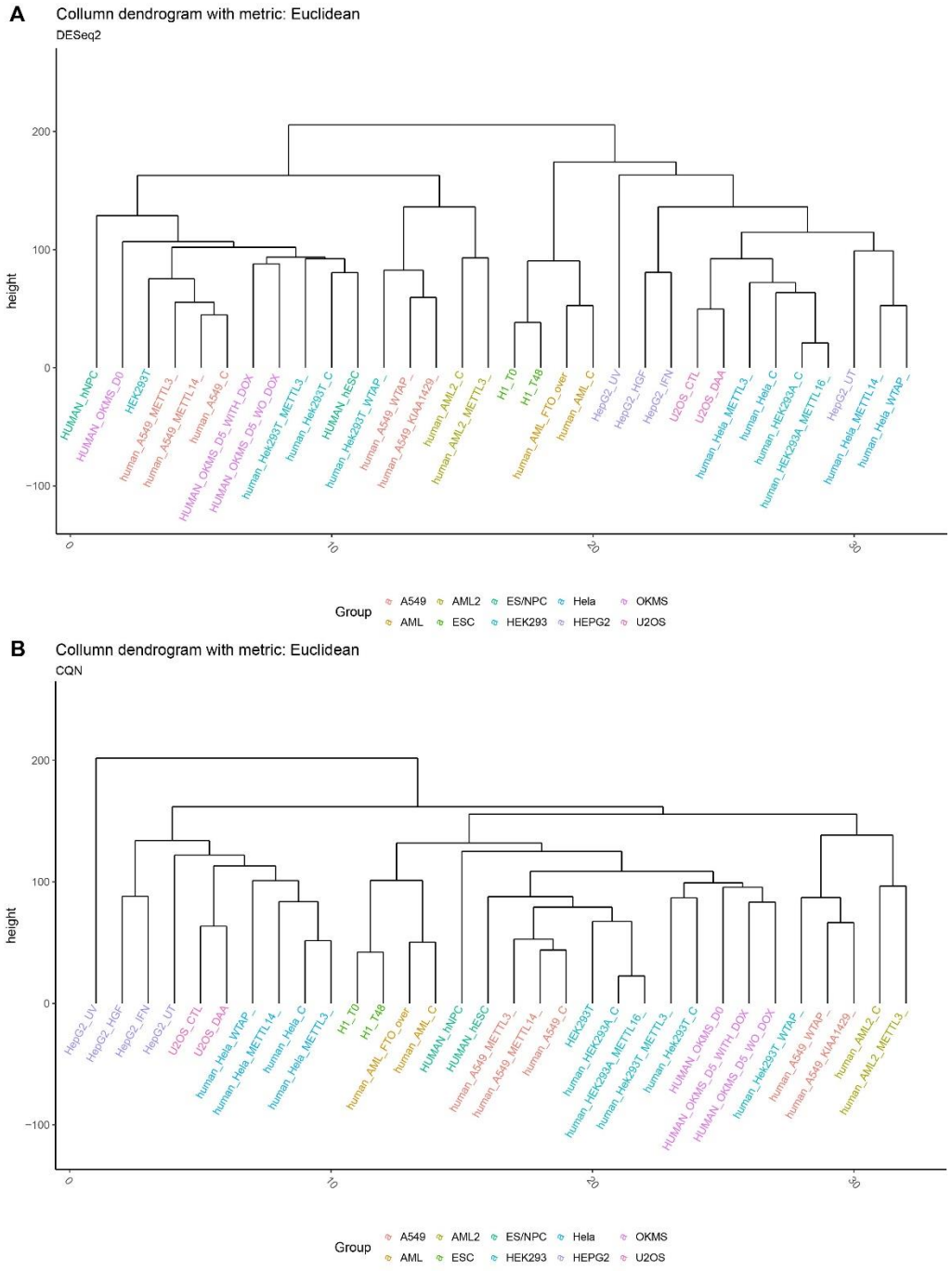

Figure 1 **Correction of technical variability**. **(A)** The clustered dendrogram of samples before applying CQN to remove technical variability. Many highly related samples are not clustered closely. **(B)** The clustered dendrogram of samples after applying CQN to remove technical variability. More related samples are clustered together, suggesting that the application of CQN in the analysis is very effective.

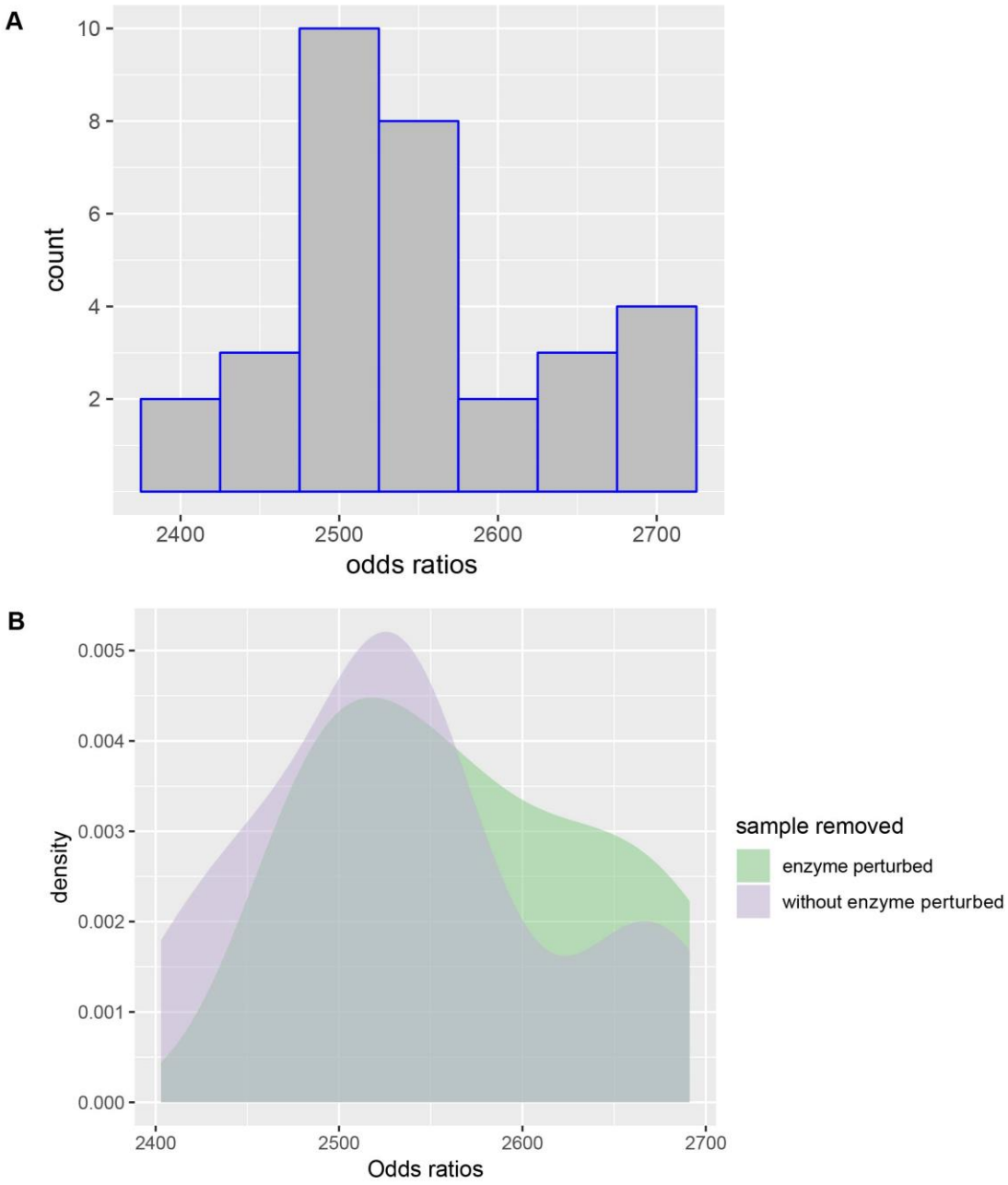

Figure 2 **Outlier identification**. **(A)** The histogram of odds ratios between adjacency matrix built by all the 32 samples and with one sample removed. There are no obvious outliers corresponding to samples that will induce substantial topological changes to the co-methylation network. **(B)** Topological changes induced to the co-methylation network. The topological changes induced to the co-methylation network by samples with enzyme permutation are not bigger than the other samples.

Section 1.3.2 Co-methylation network construction

The methylation data of 13,415 sites under 32 conditions was used to construct the co-methylation network. Since the location of these sites are known, the genes where these sites correspond to (Entrez Gene ID & Gene Symbol) were labeled. A total of 52 sites were dropped due to the absence of relevant gene annotation, and the remaining sites were kept for network construction. The site pair was defined as the co-methylation site pair only if its scc is ranked in the highest or lowest 10% and its adjusted p-value is lower than 0.05. According to the above strategy, the adjacency matrix was constructed to obtain the linkage between site pairs. The function in package igraph transformed the format of the matrix to the igraph format. A network consisting of 18,477 edges and 13,363 nodes was constructed. The constructed network with the most significant functions of four main modules in Cytoscape is shown in **Figure 3**. We observed that majority sites are clustered together in a huge group, where several modules can be identified. Moreover, we obtained small clusters ranging between 2 and 9 sites. To have a better understanding of the network, we looked at the degree distribution (see **Figure 4**), which unveiled that this co-methylation is a typical scale-free network. The scale-free network tallies with most biology networks for its robustness against disruptions. In the scale-free network, highly connected hubs, making up a relatively small number of nodes, will mainly are pivotal in determining the property of the network. The log-log plot gives an almost linear trend, with the degree exponents to be around 2.

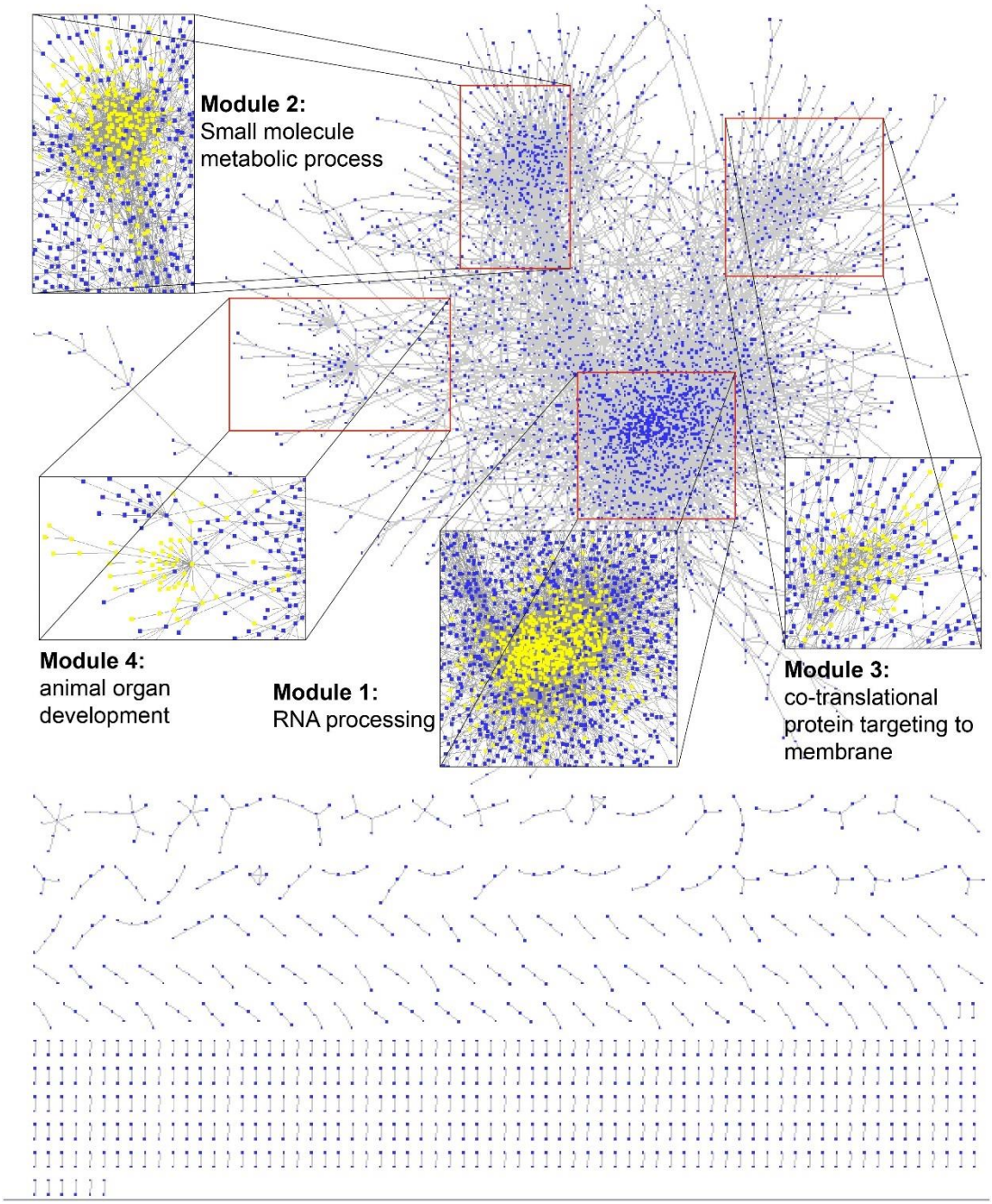

Figure 3 **Visualization of the co-methylation network in Cytoscape**. A total of 18,477 edges and 13,363 nodes make up this co-methylation network. The m^6A sites are represented by blue nodes, and gray lines represent the high positive and negative correlation between each node. Majority of the sites (91.5%) were clustered into a huge module, and few sites share high correlation in methylation level within small modules. Four largest modules were amplificated and labelled in yellow, and the most significant Gene Ontology term of each module was labelled as well.

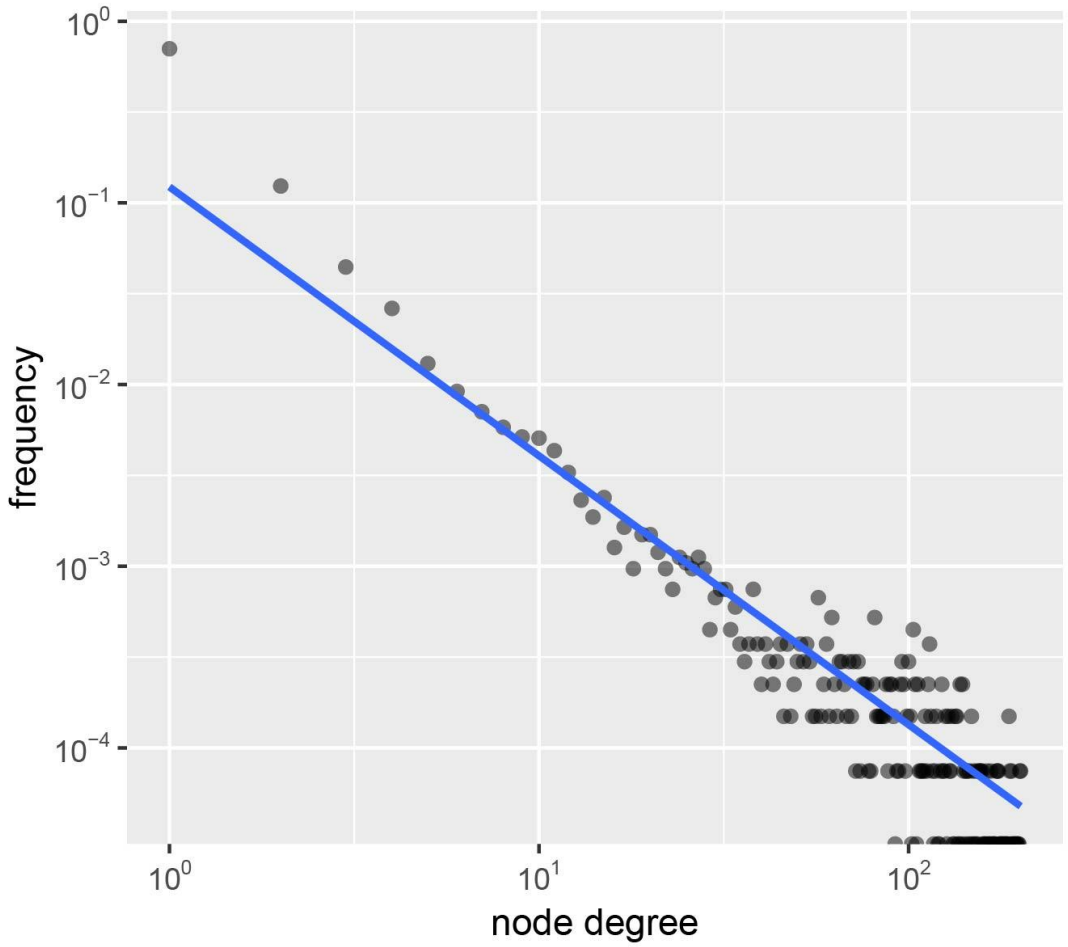

Figure 4 **The degree distribution of the co-methylation network**. The tendency of the degree on the log-log plot fits with power law, and the degree exponent of this network is close to 2, thus the powerlaw degree distribution conforms to scale-free network topology.

Additionally, it should be of great interests to compare the constructed co-methylation network to the co-expression network. For this purpose, we downloaded the human co-expression data (Coexpression version: Has-u.c2–0) from COXPRESdb [43], and built the gene co-expression network with cut-off threshold 0.8, i.e., if the Pearson correlation value between two genes is more than 0.8, the gene pair are considered co-expressed. Meanwhile, a gene-gene co-methylation network was converted from the site-site co-methylation network constructed previously. If two sites are co-methylated, their hosting genes are considered co-methylated as well.

The gene-gene co-methylation network was then compared to the gene-gene co-expression network. However, we failed to observe strong topological correlation between the co-expression and co-methylation networks. Although is still positive, the Pearson correlation of their adjacency matrices is only 2.2E-4, which suggests the epitranscriptome regulatory impact of transcriptional expression may be relatively weak at global level.

Section 1.3.3 Hub-based method

The annotation of methylation sites relies on the functional enrichment in the hosting genes of their neighbor sites according to the guilt-by-association principle. Because the functional information of individual RNA methylation sites is unavailable in existing database, we consider a soft benchmark by assuming that the functions of a sites are similar to that of its hosting genes. In the network, 1899 (14.2%) sites with connections to more than 3 immediate neighbors are defined as hub sites. To evaluate the accuracy of the prediction, we also annotated the predicted functions with the known GO terms of their corresponding gene. Thus, the enriched GO BP terms of genes where these hub sites correspond to were annotated with the Entrez ID using packages *GO.db* and *AnnotationDbi*. The corresponding genes of 1780 hub sites were annotated with GO BP terms. We also annotated all the neighboring sites of each hub site with GO BP terms. Both the annotated terms were reduced to GO Slim BP terms, and the term GO:0008150 (biological process) was excluded in enrichment results because this term almost occurs in every reduced GO Slim term. A total of 1446 sites were annotated with more than one GO slim BP term. The terms occurring as both predicted and known terms were treated as hit terms. Permutation

on sites was performed to construct the random network. In the random network, the GO BP terms and GO BP slim terms of the corresponding genes of the hub-sites were the same as that in the real network, while the predicted terms by their neighbors were different. We defined recall and precision to measure the prediction performance, and two cutoff parameters PV and GN can affect the prediction performance.

We showed in **Figure 5** the relationship between recall and precision values of both real and random networks under different cutoffs of PV (circle size) and GN (facet title). The points in blue are the performance values in the real network, and the points in red are the performance values in the random network. The values of recall and precision in the real network under these cutoffs are much higher than that in random network, which proves that the prediction in the real network should be of biological significance. The recall value is highest (13.8%) when the values of both GN (16) and PV (10^{-1}) cutoff are high. The precision value is highest (15.3%) when the values of both GN (4) and PV (10^{-3}) are low. The recall value is strongly affected by GN, while the precision value is affected more by PV. Therefore, the PV and GN will not be set too low or too high to get the reasonable recall and precision.

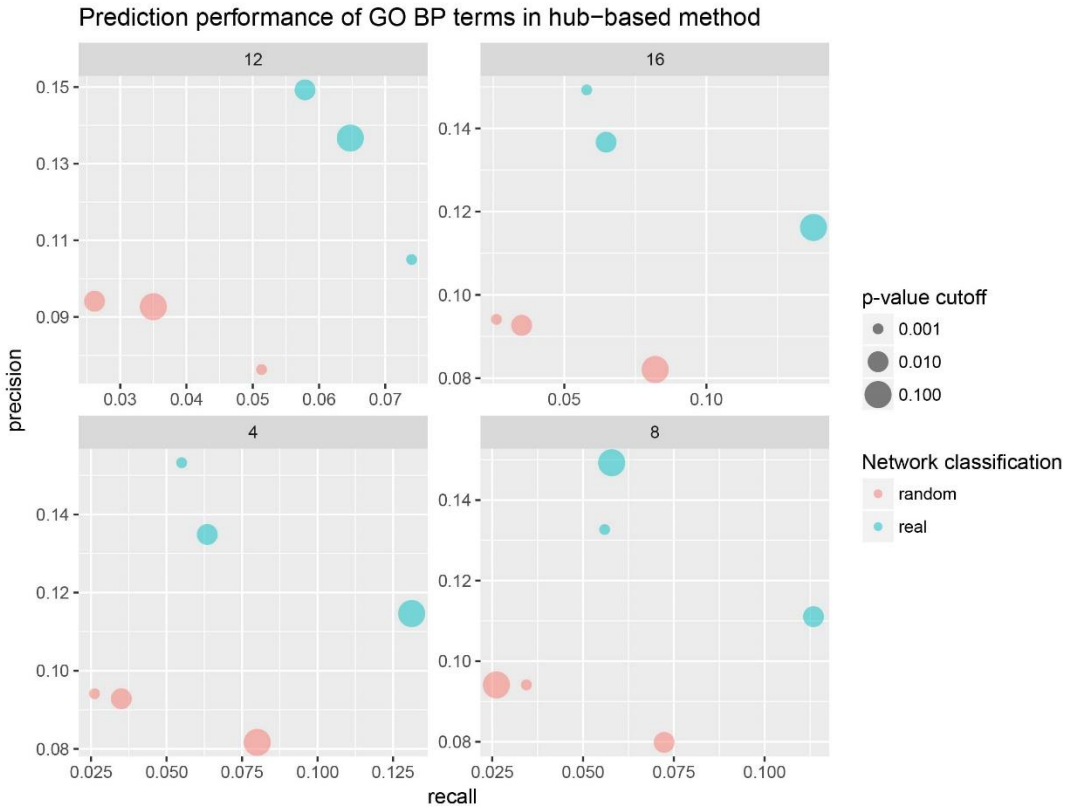

Figure 5 **Performance of hub-based functional prediction.** In the recall-precision plot, blue circles represent values under the real network, and red circles represent values under the random network. The x-axis and y-axis give the respective values of recall and precision. The number labelled as title in each facet represents each GN cutoff. The lower PV cutoff represent the smaller circle in the figure. From the figure, the values of recall and precision in the real network are much higher than the random network with the same cutoff.

Section 1.3.4 Module-based method

It is highly possible that sites within a co-methylated module share similar functions. Therefore, analyzing the corresponding genes of methylation sites in the same module can help us predict the site functions in the module-based method. The igraph object after network construction was set as the input file of the clustering algorithm. After clustering the sites with MCL algorithm (inflation value set to 1.4), 76 modules (2303 sites) containing 10 or more sites were defined as modules, the enrichment analysis of GO BP terms of these modules was performed. All the

modules were enriched with more than one GO BP terms whose p-values are lower than 0.05. After adjusting the p-value by the BH method, 8 modules were significantly (adjusted p-value < 0.05) annotated with at least one GO term. In **Figure 6**, the enriched result of the eight modules in the module-based method is given. The enriched terms in each module are labeled using different colors and columns. The size of points in the dot plot gives an indication of the magnitude of p-values corresponding to the enriched terms. The shape of points there indicates the statistical significance of the terms. The GO BP terms in module 2 (small molecule metabolic process, organonitrogen compound biosynthetic process, etc.) and module 3 (co-translational protein targeting to membrane, protein targeting to ER, etc.) are statistically significant (shown the **Figure 6**)

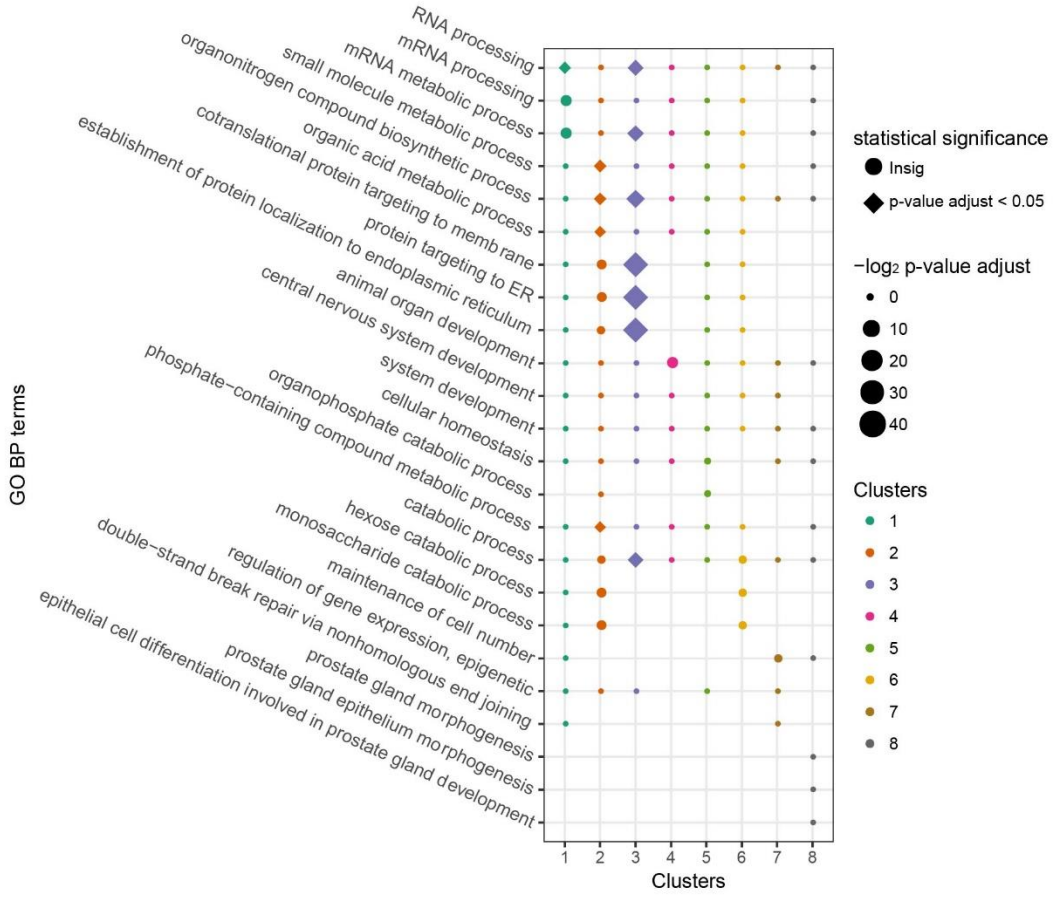

Figure 6 **GO enrichment plot of the eight most significant modules from the module-based method**. The larger the point size means the lower the p-value of the GO term. The shape in rectangle means the significance in statistics (p-value after BH adjustment is lower than 0.05), while the shape in circle means the insignificant term. GO BP terms such as RNA processing as well as small molecule metabolic process are statistically significant.

Section 1.3.5 Overlap of the functional enrichment

To evaluate the prediction accuracies of both methods, we compare the enriched functions of the same site by the two methods. Among the 2303 sites annotated by the module-based method, 1346 (58.4%) sites are annotated in hub-based method. Majority of them (1262, 93.8%) are annotated with one or more GO BP term predicted by both methods, and about 27 GO BP overlap terms occur on each site in average. We also calculate the number of overlap terms in the random network, with

the findings that 61.3% (825) sites are enriched with one or more GO BP term by both methods, and 3.2 overlap terms occur on each site. **Figure 7** is the boxplot which shows the count of overlap terms predicted by the hub-based method and the module-based method on each site. The number of overlapping terms of both prediction methods is higher in the real network than the random network, indicating that the predicted functions annotated by the module-based method are credible.

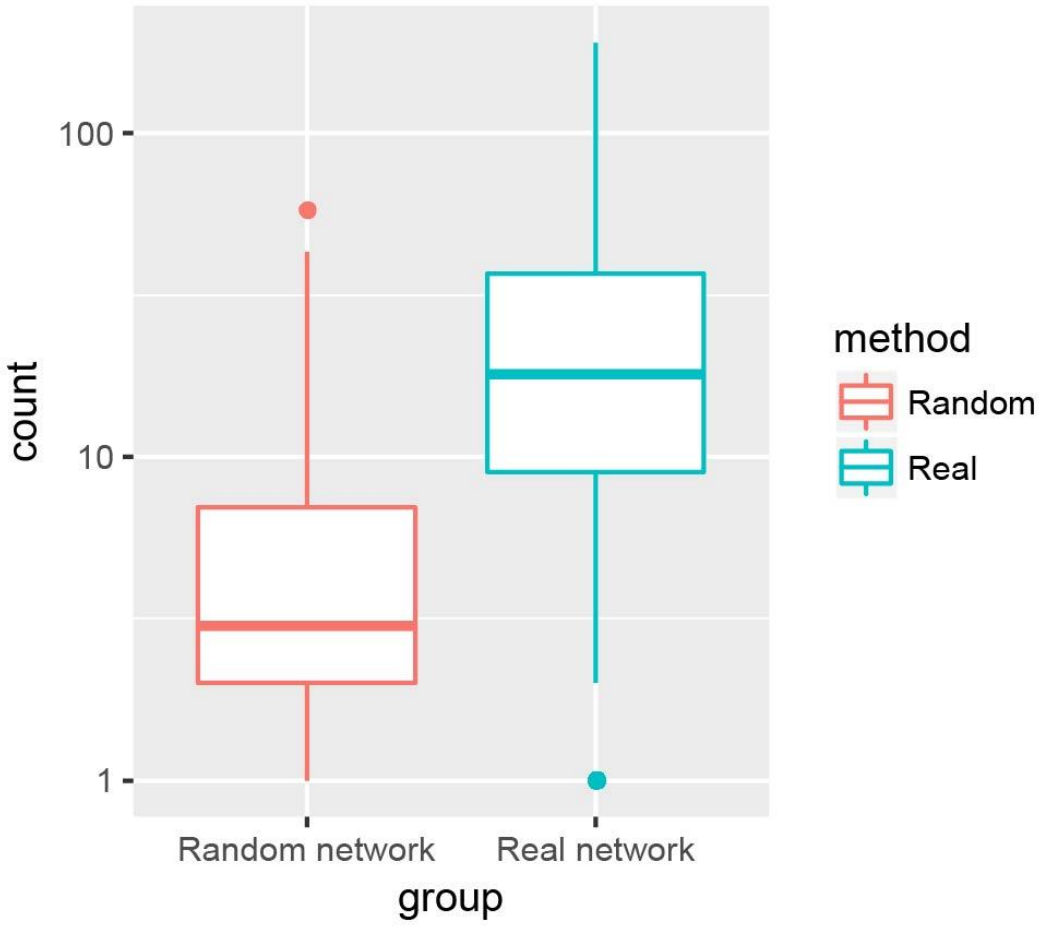

Figure 7 **The enriched terms are more consistent in real network**. The boxplot of the overlap term number in the real network and the random network at the same methylation site. The box in red represents the term count in the random network, and the box in blue represents the term count in the real network. After $\log_{10}$ transforming the term count on y-axis. It is obvious that the overlap terms are much more in the real network (mean 27) than in the random network (mean 3.2).

Section 1.3.6 Database construction

To enable the direct query of the predicted functions associated with individual m^{6}A RNA methylation sites, we constructed a web site m^{6}Acomet, which stands for functional prediction of **m^{6}A** RNA methylation sites from RNA **co-met**ylation network, and is freely available: http://180.208.58.19/m6acomet/. A data table, which contains the necessary information of sites, is provided, including: methylation site ID, position on chromosome, RNA strand, corresponding Gene Symbol, corresponding Gene Entrez ID, count of neighbor sites, count of corresponding genes of neighbor sites, count of GO BP terms of the hub gene, count of GO BP Slim terms of the hub gene, count of GO BP terms of predicted neighbor genes, count of GO BP Slim terms of the predicted neighbor genes, count of hit terms of the two slim term columns, and count of the GO BP terms annotated by module-based method. The detailed information, which includes the exact GO (or GO slim) terms together with the enrichment significance (p-value < 0.05) and its neighboring m^{6}A sites in the RNA co-methylation network, will be shown if the user clicks on the relevant hyperlinks.

Section 1.4 Conclusion

The functional characterization of post-transcriptional modification sites by wet experiments is extremely expensive and laborious. For this reason, we propose a computational framework, for the first time, to predict the putative functions of individual RNA methylation sites from an RNA co-methylation network in large-scale. Specifically, before network construction, the methylation level on each site was estimated and normalized by *DESeq2* and *CQN*. Several systematic biases in GC content and batch effect were adjusted. The raw predicted m^{6}A sites were further

filtered, and only the sites with substantial biological signals were kept for further analysis. The RNA co-methylation network was built from MeRIP-seq data profiling the transcriptome-wide RNA methylation status in 32 experimental conditions. We showed that the co-methylation network exhibit typical scale-free characteristics. The biological functions of each individual RNA methylation sites were then inferred based on the guilt-by-association principle. Two different types of algorithms were developed for functional annotation. We suppose that the regulation role of each m^6A site should be similar to the annotation roles of its corresponding gene. For this purpose, the methylation sites with three or more edges were functionally annotated by the hub-based method. The prediction performances (recall and precision) were defined to assess the predictive efficacies of the real and random networks. The PV and GN were chosen as cutoff parameters to assess the prediction performance. The random network was constructed to compare the prediction performance with that from the real network. By taking advantage of a soft benchmark, our result showed that the recall and precision values of the real network are both higher than that of the random network with various cutoff. In other words, the prediction results in the co-methylation network suggested higher biological significance. In the module-based method, sites from largest modules (module size $\geqslant$ 10) clustered by MCL algorithm were annotated by GO terms. After comparing the enriched terms of the sites annotated by both methods, we found that majority of the sites share overlapping GO terms, suggesting that the functional enrichment in module-based method is reasonable. Functional enrichment by different methods can extend the range of annotation terms and increase the number of predicted sites. The

predictions in some cases by two methods are complementary and coherent, which reinforce the validity in prediction.

It is worthwhile mentioning that the biological function of an RNA methylation site may be different from that of its host gene. The former focused on epitranscriptome layer regulation; while the latter may be regulated through any layers of gene expression regulation, e.g., DNA methylation, post-translation modification, etc. In this work, we focused specifically on the RNA methylation profiles, which is governed by RNA epigenetics regulation and thus echo biological processes regulated at epitranscriptome layer. Although the epitranscriptome modules (or RNA co-methylation pattern) have previously been shown to demonstrate functional relevance of the RNA methylation sites [12, 13], it is, to the best of our knowledge, that we are the first to use this property for functional prediction for individual RNA methylation sites.

The annotation result of the human m^6A sites in our project are presented in an online database m^6Acomet. It supports the query with respect to a biological function or a number of co-methylated RNA methylation sites, and may serve as a source of reference for further biological research.

However, the project still has a few limitations. For example, the annotation rates for all the filtered methylation sites in both methods are not satisfactory. The criteria for the construction of the co-methylation network may be too stringent and could be further optimized; more data sources such as protein-protein interaction, pathways,

can be integrated with the RNA co-methylation network for a more accurate functional annotation.

Chapter 2 **Functional annotation of m^{6}A methylation sites using guilt-by-association principle in WHISTLE**

Section 2.1 Introduction

The function of epitranscriptome has been mainly investigated from their mediator proteins. The m^{6}A epitranscriptome is directly manipulated by three classes of RNA binding proteins, namely the writer (methyltransferase), eraser (demethylase) and reader (m^{6}A specific RNA binding protein), which install, erase or recognize the m^{6}A modification [44]. The key role of m^{6}A in regulation of gene expression regulation has become evident from a number of knockdown and overexpression experiments. Genetic inactivation of m^{6}A writer METTL3 resulted in long-lasting Nanog expression upon differentiation and resulting defective ESC exit [45, 46]. Knockdown of RBM15 or METTL3 was shown to impair XIST-mediated X chromosome inactivation [47]. METTL3 and METTL14 were shown to modulate murine spermatogenesis [27, 48]. The m^{6}A reader YTHDF2 recognizes and reduces the stability of transcripts [5]; while another reader, YTHDF1 promotes translation of targeted transcripts [49]. With the advance of next-generation sequencing [50] and other techniques, the versatile functions of m^{6}A RNA modification have been more comprehensively understood. However, existing functional studies are primarily limited by laborious experimental procedures or limited samples, resulting low statistical significance and low resolution [50]. In spite of this, such methods are only suitable for investigating the molecular functions of individual mediator protein under a specific context, but are sufficient to uncover the functions of individual RNA methylations site under different conditions.

The exploration of epitranscriptomics has greatly benefited from computational biology and bioinformatics. The realization of online databases enabled the annotation of transcriptome-wide RNA modification sites with tissue/cell type specificity, potential association with RNA-binding-protein binding proteins, disease-associated SNPs, miRNA binding sites, differential methylation profile across multiple experiments, etc. [51, 52]. Nonetheless, despite the great efforts being made, there is, to our knowledge, still no available database or computational effort for functional annotation of individual RNA methylation sites under the Gene Ontology framework. Instead of characterizing the functional relevance of individual epitranscriptome mediator protein, the question we are pursuing is: what biological process may be affected when a specific site is (de)methylated, i.e., providing a Gene Ontology-annotated map of the m^6A epitranscriptome at single site resolution using machine learning approach.

The accumulation of high-throughput data and theories on complex networks have furthered the understanding of biology components in context instead of as discrete parts [39, 53]. A network presents a graphical model where the nodes represent the biological components and the edges allow functional information flow between the connecting nodes [54]. There are many successful applications in recent years. Disease gene prioritization in complex disease yielded a fruitful result since single gene abnormalities might only confer a marginal phenotypic defect while the overall malfunction is a collective effect of a myriad of interactors [55]. Random walk with restart was successfully used in drug-target interaction mining [56], and predicted

functions of long-noncoding RNAs derived from a co-expression network [14]. Meanwhile, with the rapid update of genomic information and high-throughput techniques, several major databases, e.g. STRING [57], have been built for the integration of multi-omics data that are freely available to public, which may serve as prior knowledge when modeling the unknown relationships. One of the major challenges of functional prediction is the sparse known site-gene functional connections, which usually serve as the “seed” for many network based algorithm. Therefore we here adopted the philosophy of "guilt-by-association" [58, 59] to build initial edges from data. In particular, if an RNA methylation site has methylation profile that is well correlated to the expression profile of a set of genes, then one may assume that they are functionally related and thus “connected” in the network. MeRIP-seq data are well-controlled and naturally suitable for such purpose, since the Input and IP sample of the m^6A-seq data share common upstream preparation and the Input sample is essentially an equivalent to RNA-seq when quantifying gene expression abundance. When we use a network approach to infer the functions of each m^6A sites, one of the potential challenges is the limited number of datasets. Due to the laborious nature of m^6A-seq preparation [50], the pool is limited in both number of experiments and number of replicates. To obtain a biologically robust network structure, we consider amplifying only the consistent signals and add prior knowledge for regularization purposes. A classical pipeline for disease-associated gene prediction was provided by taking advantage of the network-based information propagation [60], which resulted in more consistent somatic mutation pattern thereby enabling downstream clustering of tumor subtypes [61]. Under this pipeline, the output is an influence confidence vector for a certain site, which is a ranked gene

list that could be used as input to the Gene Set Enrichment Analysis (GSEA). GSEA is based on the Kolmogorov–Smirnov test to identify if the distribution of enrichment scores is significantly different from the null distribution [62]. Compared to the over-representation tests, e.g. the Fisher's exact test, GSEA is non-parametric and considers the rank of genes under certain perturbation (e.g. methylation of a particular site).

Here, the "guilt-by-association" principle was applied to further annotate the functional relevance of each individual RNA methylation site by integrating the gene expression profiles, RNA methylation profiles and PPI networks. Specifically, the m^6A-seq data profiling in 38 different experimental conditions obtained from 11 studies was used for analysis. The expression level for each gene and the methylation level for each site were quantified. The initial co-methylation and methylation-expression network were built from available datasets, and then the consensus signals for each site were amplified by multiplying these two. The correlations for a query site to gene expression were then smoothed using a Protein-Protein interaction (PPI) network [57], and the resulting gene lists are used as input for GSEA algorithm for functional annotation. Several case-studies related to YTH-domain m^6A readers are presented as evidence of the biological insights obtained from the constructed networks.

Section 2.2 Materials & methods

Section 2.2.1 Gene expression level quantification

The gene expression level is quantified from the Input control sample of the MeRIP-seq data, so that we have the matched RNA m^6A methylation profiles as well.

Specifically, the raw sequence data from 11 independent studies corresponding to 38 independent experimental conditions, were downloaded from GEO list in **Table 2** and aligned with HISAT2 [63] to the human genome assembly hg19 downloaded from the illumina iGenomes. The gene expression levels (FPKM) were averaged among biological replicates obtained from the same experimental condition, in which the samples are from different cell lines or subjected to different treatments. The resulting expression profile contains the expression level of 22,687 genes under 38 experimental conditions. The technical bias such as GC content was corrected by conditional quantile normalization approach with the *CQN* R package [24].

Table 2 m^6A-seq data used in Section 2.

ID	Sample Label	SRA Study	GEO accession	Source
1 - 4	HepG2-UV	SRP012098	SRR456542-SRR456543	[17]
	HepG2-HS		SRR456544-SRR456545	
	HepG2-HGF		SRR456546-SRR456547	
	HepG2-IFN		SRR456548-SRR456549	
5	HEK293T-1	SRP007335	SRR494613-SRR494618	[64]
6 - 9	Hela	SRP022152	SRR847358-SRR847361, SRR847370-SRR847373	[33]
	Hela-METTL14-		SRR847362-SRR847365	
	Hela-WTAP-		SRR847366-SRR847369	
	Hela-METTL3-		SRR847374-SRR847377	
10 - 11	U2OS	SRP026127	SRR903368-SRR903370, SRR903374-SRR903376	[32]
	U2OS-DAA		SRR903371-SRR903373, SRR903377-SRR903379	
12 - 13	H1ESC	SRP033229	SRR1035213-SRR1035224	[37]
	H1ESC-T48		SRR1035217-SRR1035220	
14 - 26	hNPC	SRP039397	SRR1182582-SRR1182586	[34]
	hESC		SRR1182587-SRR1182590	
	HEK293T-2-WTAP		SRR1182591-SRR1182592	
	HEK293T-2-METTL3-		SRR1182593-SRR1182594	
	HEK293T-2		SRR1182595-SRR1182596	
	OKMSfibro-Dox		SRR1182597-SRR1182598	
	OKMSfibro		SRR1182599-SRR1182600	
	OKMSiPC		SRR1182601-SRR1182602	
	A549-WTAP-		SRR1182603-SRR1182606, SRR1182625-SRR1182626	
	A549-METTL14-		SRR1182607-SRR1182614, SRR1182635-SRR1182636	
	A549-METTL3-		SRR1182615-SRR1182618, SRR1182629-SRR1182630	
	A549		SRR1182619-SRR1182624, SRR1182633-SRR1182634	
	A549-KIAA1429-		SRR1182627-SRR1182628	
27 - 30	AML-1-FTO+	SRP067910	SRR3066062-SRR3066065	[35]
	AML-1		SRR3066066-SRR3066069	
	gsc11		SRR4310464-SRR4310465, SRR4310468-SRR4310469	[65]
	gsc11-ALKBH5-		SRR4310466-SRR4310467, SRR4310470-SRR4310471	
31 - 32	HEK293A	SRP094637	SRR5080301-SRR5080303, SRR5080307-SRR5080309	[66]
	HEK293A METTL16-		SRR5080304-SRR5080306, SRR5080310-SRR5080312	
33 - 34	AML-2	SRP099081	SRR5239086-SRR5239101	[36]
	AML-2-METTL3-		SRR5239090-SRR5239109	
35 - 38	NB4	SRP103072	SRR5417009,SRR5417011, SRR5419908, SRR5419910	[67]
	MM6		SRR5417010,SRR5417014, SRR5419912, SRR5419914	
	NB4-METTL14-		SRR5417012-SRR5417013, SRR5419909, SRR5419911	
	MM6-METTL14-		SRR5417015-SRR5417016, SRR5419913, SRR5419915	

Section 2.2.2 RNA methylation level quantification

The methylation levels of these sites were estimated from the m^{6}A-seq data, which we previously used to estimate the expression level profiles based on their input control samples as well. The m^{6}A-seq data regards methylation level as the relative abundance between IP (immunoprecipitation by antibody binding to m^{6}A site) and input (normal RNA-seq). The methylation level for each site is calculated as **Equation 3**.

Equation 3

$$m_{i,s} = \log_2\left(n_{i,s,t} / n_{i,s,c}\right)$$

Where $n_{i,s,t}$ and $n_{i,s,c}$ represent the read abundance (in RPKM) of a specific m^{6}A site in the IP and input sample of MeRIP-seq data, respectively; and the methylation level $m_{i,s}$ is estimated using the DESeq2 package with Bayesian shrinkage for more robust quantification of the very lowly expressed genes [22]. Similar to gene expression quantification, the technical bias was corrected by conditional quantile normalization approach [24].

Section 2.2.3 Initial network construction

Filtering is necessary to remove any trivial or confounding signals from the predicted epitranscriptome map. First we filtered the trivial RNA methylation sites whose methylation level was consistently low (mean within 75% percentile). The filtered methylation site in this step may not be a strong (or house-keeping) methylation site and may function only under relatively fewer experimental conditions. After this step,

29,855 sites are retained among the original 82,245 sites supported by at least one of the six base-resolution experiments (see **Table 3**). Then, the top 10,000 sites with largest variance in methylation level were preserved for functional annotation. These sites were selected because more information related to their epitranscriptome function may be revealed through their highly dynamic methylation profiles. Please note that, although we performed transcriptome-wide m^6A site prediction, only part of the m^6A epitranscriptome is functionally annotated. The screening prior to functional annotation is necessary because that it will greatly reduce the computation load while increasing the accuracy and reliability of the annotation by focusing on only the most reliable, dynamic and extensively occurred RNA methylation sites.

Table 3 Base-resolution dataset used in m^6A site prediction.

ID	Cell line	Note	Technique	Source
1	HEK293	abacm antibody	mi-CLIP	[20]
2		sysy antibody		
3	MOLM13		m^6A-CLIP	[27]
4	A549			[21]
5	CD8T			
6	HeLa			[28]

An RNA co-methylation network (GMM, which stands for graph for methylation-methylation) and a methylation-expression network (Pre-Amp GME, which denotes the graph for methylation-expression before signal amplification) are built according to the criteria used in [14]. The spearman correlation between site pairs and site-gene pairs were calculated and p-values were generated through Fisher's asymptotic

test [38], and adjusted by Bonferroni multiple test correction implemented in the R *multtest* package [68]. For both networks, only methylation site pairs with a p-value of 0.05 or less and with a Spearman correlation value ranked in the top or bottom 0.05 percentile were regarded as an edge. Positive and negative correlations were explicitly encoded as "+1" or " -1" in the adjacency matrix. Self-loops were prohibited.

Section 2.2.4 Signal amplification

From multiple knockout experiments biologists have come to understand that, the functional impact of RNA modification is an orchestrated event that involves the action of methyl-transferase on multiple sites. Thus we devised a scheme (**Equation 4**) to amplify the consensus signals, which we believed are of more biological significance.

Equation 4

$$Adj'_{me} = Adj_{mm} \times Adj_{me}$$

where Adj_{me} represents the adjacency matrix of initial methylation-expression network, Adj_{mm} represents the adjacency matrix of initial co-methylation network, and Adj'_{me} gives the adjacency matrix of amplified methylation-expression network (Post-Amp GME, which denotes the graph for methylation-expression after signal amplification). Any site conferring consensus functional impact (either positive or negative correlation) with its neighboring sites in the co-methylation network on a

gene, which should have been more reliably connected to that gene will be strengthened in terms of edge weight. In contrast, if the neighboring sites of a query site have no predominate correlations (an even chimeric combination of positive or negative) to a gene, then the query site may not be convincingly connected to that gene and such edge tends to be diminished in matrix multiplication. An additional merit of signal amplification is to increase the number of edges present in the network, which might be of biological significance since those edges come from consensus behavior of highly related sites.

Section 2.2.5 Network smoothing

We adopted the PRINCE network smoothing framework described by [60]. Starting from the amplified methylation-expression network we want to find an optimal information source in matrix form that could both smooth over the protein interaction network and the known connections, which could be expressed as **Equation 5**.

Equation 5

$$F^t = \alpha W' F^{t-1} + (1-\alpha) Y$$

where α is the relative importance for protein-interaction matrix to the known interactions, and F represents a mapping from a particular site to a gene set with different weights, while the superscript denotes the number of iterations. Y gives the original seed mapping between a site to gene set in amplified GME. W' gives the

normalized version of adjacency matrix of STRING Protein-Protein Interaction network [69], which is calculated as **Equation 6**.

Equation 6

$$W' = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

where W is the weighted adjacency matrix of STRING PPI and D is the diagonal matrix where $D(i,i)$ is the sum of the i-th row in W.

Section 2.2.6 Network randomization

A random network needs to be generated for the purpose of comparing performance. To better simulate biology network we used a node-switching algorithm that preserves the degree of each node. The resulting randomized network is believed to better serve as a null model. In this work, we randomized the co-methylation network by *rewire* function in *igraph* R package [41] and randomized the initial methylation-expression network by *BiRewire* R package [70], which is specifically designed for randomization of a bipartite graph. The final randomized methylation-expression matrix is the product of two randomized version of adjacency matrix of initial co-methylation network and the incidence matrix of the methylation-expression network. Similarly, the randomized network was also smoothed over PPI as described in previous section.

Section 2.2.7 Functional prediction by GSEA

According to the guilt-by-association principle, it is reasonable to assume that, if a specific methylation site exhibit a strong association with a functional gene module (a number of genes participate in a specific biological process), it is likely that the methylation site is functionally related to the module via epitranscriptome layer of gene expression regulation, which directly controls the methylation level of an m^6A site via RNA methyltransferases, demethylases or RNA binding proteins.

The resulting GME network after amplification and smoothing is composed of the association between each individual RNA methylation site and the genes derived from RNA methylation, gene expression and protein-protein interaction data. We then adopted the GSEA method [62] using the R package *clusterProfiler* [71] to predict the Gene Ontology terms that are likely to be associated with a specific RNA methylation site from the ranked gene list extracted from the amplified and smoothed GME network.

Section 2.2.8 Soft benchmark for functional prediction

To overcome the problem of the lack of benchmark for experimentally verified functional sites, we use the function of the gene where a site resides as a soft-benchmark to assess prediction performance. It is also worth to mention that the predictions not supported by the soft benchmark may correspond to previously unknown biological mechanisms that regulated at epitranscriptome layer, and worthy to be explored. Researchers have previously shrunk the predicted GO pool to GO slim terms and compared specificity and precision for true and random network

[14]. However the slim version of GO is clearly not as comprehensive. Here, following previous approaches, we define the prediction "hit" as any term that has semantic similarity measure higher than 0.8 by Wang's graph-based method implemented in *GOSemsim* package to a term annotated to the residing gene [72, 73]. Then, we compared our prediction results with the performance achieved on the random network.

Section 2.2.9 Feature gene selection for YTH-domain readers

To further validate the biological significance of our analysis, we focused on the three well-characterized readers: YTHDC1, YTHDF1 and YTHDF2 to verify if their binding site could be predicted to mediate the known functions of these enzymes in RNA metabolism. First the reader target sites were filtered based on differential methylation upon reader knockout from MeT-DB [30] or Clip mappings from RMBase [52]. For the subset of target site for a certain reader, only the genes that shows consensus influence pattern among all sites were included, by applying the following filters: (1) Retain the genes that have absolute association in smoothed GME over 0.001 among all reader-targeted sites, which ensures consistency of association among all binding site to that reader; (2) For each selected genes, rank sites by association weight in the smoothed GME. (3) Select genes by variance of site ranks. The genes with the lower variance in site rank is retained for subsequent analysis. This will result in three gene sets for each of the three m^6A readers. Fisher's exact test with Benjamini-Hochberg multiple hypothesis correction [74] was then applied for GO molecular function (MF) prediction.

Section 2.3 Results

Section 2.3.1 Functional annotation of individual m^6A sites

We then functionally annotated individual m^6A sites using the guilt-by-association principle from a network constructed from gene expression data, RNA co-methylation data and protein-protein interaction data with gene set enrichment analysis.

Section 2.3.2 Network characterization

Scale-free network structure is desired for most biology networks [39] for its robustness against disruptions. In scale-free networks, the majority of network nodes are of low degree, and only minority of them are important. From **Figure 8**, we could observe a good fit with power law (the black line in each plot). Comparison of two plots revealed that there are significantly more nodes after signal amplification, but the overall scale-free property of biology network is well-preserved. This result supported the potential biological relevance of the network (**Figure 8B**). The signal amplification step is important to strengthen the confident edges between expression and genes, as well as to build new connections that may be of biological significance. As a result, more sites could be annotated after signal amplification, which further supports the design (**Figure 8A**).

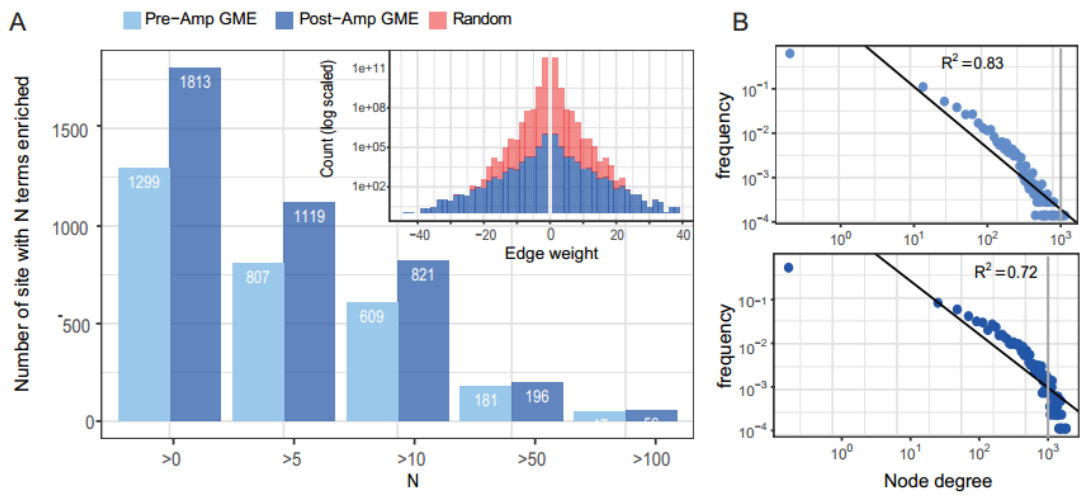

Figure 8 **Effects of network amplification. (A)** Post-Amp GME predicts more sites than Pre-Amp GME, since the amplification expands the edge weight, highlighting the most consistent signals from sites to genes, in sharp contrast to a random network. **(B)** Degree distribution of Pre-Amp GME and Post-Amp GME conforms to scale-free network topology.

Section 2.3.3 Self-gene correlation

As noted in previous sections, we used the function of a gene where a site resides as a "soft-benchmark" for prediction performance evaluation. However under such assumption, the edges that connects an m^{6}A site to its own gene, which could be signified by deviation of Spearman correlation distribution of self-gene with a site compared with permutated correlation distribution, would potentially contribute significantly to the results. To assess this risk, we looked at the Spearman correlation between the methylation level of an m^{6}A site and the expression level of its residing gene (we termed it as "self-gene"), and built the null distribution by permutation and resampling the same number of sites from other genes. **Figure 9** demonstrated that the self-gene correlation distribution is not significantly different from the background null distribution, and self-gene correlation only constitutes a tiny fraction of the methylation-expression networks. This result further corroborates that the

methylation level of each site tends to be independent of the expression level of its residing gene. Therefore, most of the signals that agree well with our assumed "soft-benchmark" come from the biology-driven network topology, instead of self-gene connections.

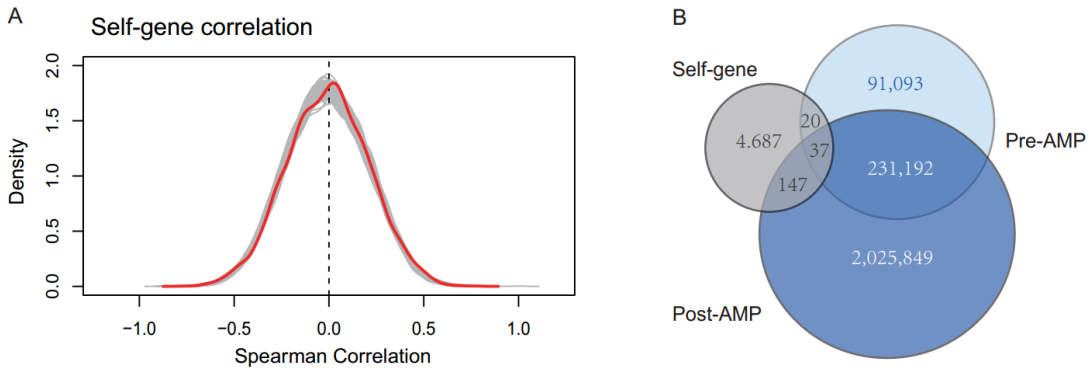

Figure 9 **The effect of self-gene. (A)** The red line indicates the 6,150 self-gene Spearman correlation coefficients, and the grey lines are 1,000 permutations which serves as a null model of spearman correlation distribution. The p-value of Kolmogorov–Smirnov test is higher than 0.1, suggesting self-gene is not over-presented in our constructed network. **(B)** Examination of site-gene overlapping shows that self-gene only marginally overlaps with both Pre-AMP and Post-AMP GME. The given numbers denotes the number of edges belonging to that subset.

Section 2.3.4 Functional enrichment

GSEA is performed for each weighted gene vector, which is a list of genes that are associated to a specific RNA methylation site, to annotate the functional relevance of each individual RNA methylation site via the guilt-by-association principle. To assess the performance, we ask the following two questions: (1) How much biological significance, in terms of the number of sites that would be convincingly annotated under certain significance level, could be deduced from our analysis? (2) How many terms being annotated are shared between the m^6A site annotated and its host gene?

Figure 10 shows that in all of three domains, the enrichment result for true network

is significantly superior to the random counterpart. As the FDR adjusted p-value cutoff increases, more sites tends to be annotated. A similar trend was observed as alpha (the relative importance α) increases, which is not unexpected since α favors stronger information propagation by the biologically meaningful PPI network. The FDR corrected p-value was chosen as cutoff, since the raw p-value is limited in resolution due to our preset number of permutation. Despite the seemingly large adjusted p-value by FDR, the raw p-value with adjusted value less than 0.2 normally is below 0.002. For a reasonable combination of p-value cut-off and α (e.g., $\alpha = 0.5$ and $p_{adj} = 0.25$), one might safely annotate the functions of more than 1,000 sites. To answer the second question, we regard a predicted GO term as a "hit" when it has semantic similarity with any terms annotated for that gene ("soft-benchmark"). The number of hit terms is significantly greater in our method compared to random, which well-demonstrated the biological significance of our network-based approach (**Figure 10B**).

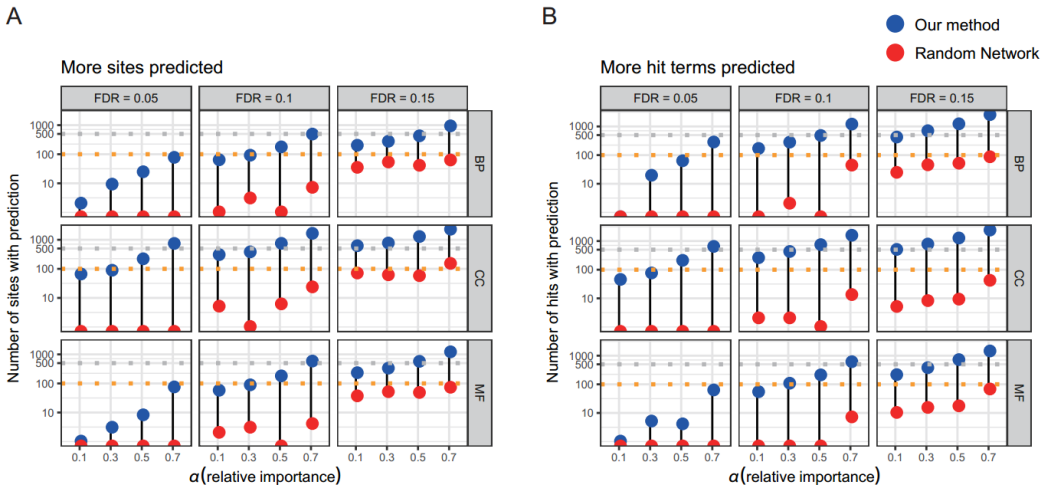

Figure 10 **Performance assessment for GSEA method**. **(A)** The red node represents the number of sites with at least one predicted terms for true matrix, and the blue node represents the randomized counterpart. The number on the grey panel denotes the FDR adjusted p-value. The X-axis in all plots gives the relative importance. **(B)** The total number of "hit" term in true and random network for GO categories, i.e., biological process (BP), cellular compartment (CC) and molecular function (MF).

Section 2.3.5 Case study: the YTH-domain readers

The functions of m^{6}A modification are presumably executed by readers, which are RNA binding proteins that recognizes specifically m^{6}A sites [44]. Among known readers, three YTH-domain readers, YTHDC1, YTHDF1, and YTHDF2 are comparatively well-characterized. YTHDF1 has been shown to interact with translation machinery (eIF3, G3BP1, etc.) and promotes active protein synthesis [49]. YTHDF2 was shown to participate in mRNA decay by redirecting those transcripts into decay sites [5, 75, 76]. YTHDC1 has been shown to promote exon inclusion by selectively recruiting mRNA splicing factors [77-79].

To demonstrate the feasibility of our feature gene selection pipeline, we applied hierarchical clustering of two groups of sites (**in-group**, which is the selected sites; the other is **out-group**, which has the same quantity with in-group but is randomly

sampled from the unselected sites) based on selected features to see our selected features are sufficient to distinguish one from the other. From the following clustering pattern (**Figure 11A**), we could clearly find enrichment of in-group (red) at one end of the graph. YTHDF1-mapped genes forms a densely connected network in STRING database, which further support the functional coherence. The other two reader, YTHDC1 and YTHDF2, shows similar pattern in terms of site-differentiation and network enrichment.

We further characterize each gene set by Fisher's exact test for GO enrichment of molecular functions (MF), and the enrichment results were then corrected by Benjamini–Hochberg procedure [74] and shown in **Figure 11C**. YTHDC1 is known to function in pre-RNA splicing. Enriched GO terms such as "nuclease activity" can be directly linked to splicing. There are also terms related to RNA metabolism, which features "RNA binding" or "methyltransferase activity". These result collectively show that the predicted functions of YTHDC1 related sites agrees well with the function of YTHDC1 itself. YTHDF1 is known to be functional in translation by interacting with translational machinery. We could identify many known components of the translational machinery, including translation initiation factors and aminoacyl-tRNA ligase (which is responsible to add appropriate amino acid to tRNA). An interesting finding is related to microtubule plus-end binding, which is observed in the neuronal growth cone [80]. There are also many components of transcriptional machineries, including nucleoside diphosphate kinase, which is responsible for nucleoside triphosphate (e.g. 5' Guanosine-triphosphate Cap m^7G), and presumed to

serve as a positioning guide for translation [81] by 43S ribosomal preinitiation complex via eIF4F complex.

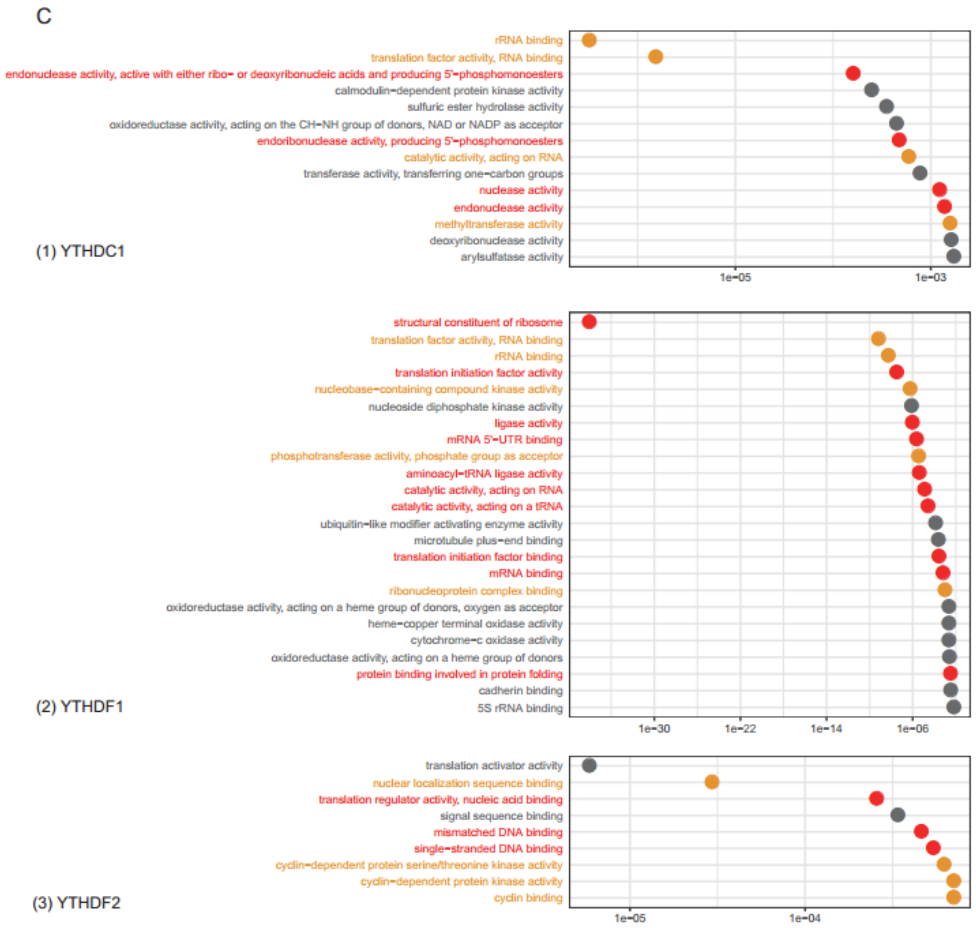

Figure 11 **Functional prediction from YTH-domain reader mapped sites. (A)** The cluster dendrogram shows that YTHDF1-mapped sites and random sites can be well-distinguished from the feature set. **(B)** Selected feature genes for YTHDF1 formed a dense cluster in STRING database. **(C)** The predicted GO MF function of YTHDC1, YTHDF1 and YTHDF2 mapped sites. Grey terms denotes functions that are only marginally correlated with this particular enzyme, orange terms denote the functions that are generally related to m^6A readers, red terms are specifically associated with this m^6A reader.

YTHDF2 is known to mediate RNA decay. The He lab has come up with a unified perspective that YTHDF2-mediated degradation controls the life-time of transcripts while YTHDF1-mdiated translation increases translational efficiency. The overall

effect is fast responsive gene expression [49], which is likely to be observed in cell cycle, where self-perpetuating positive feedback loops are frequent to generate quick responses. The Cyclin related pathway controls processes related to cell cycle. Although there are no well-established experimental data that connects cyclin and YTHDF2, YTHDF2 was observed to contribute to "migration-proliferation dichotomy" and might be responsible for carcinogenesis [82, 83]. Single-stranded or mismatched DNA is the characteristics of cellular stress [84, 85], which is prone to inducing apoptosis which triggers global RNA decay [86, 87]. The role of YTHDF2 related sites in mediating RNA decay can be further substantiated by the frequent appearance of "stabilization" in the biological process (BP) terms of GO. Moreover, the link between YTHDF2 and translation may be established by heat shock protein in cellular stress [88]. Under physiological conditions, FTO functions to remove the m^6A from the 5'UTR from nucleus and YTHDF2 usually localizes in the cytoplasm [44]. However upon heat shock, YTHDF2 was found to translocate into the nucleus to protect m^6A at 5'UTR and assembles m7G cap for translation initiation in eukaryotes. This mechanism is presumed to promote cap-independent translation for heat-shock specific proteins.

Section 2.4 Conclusions

To functionally annotate the potential regulatory impact of each individual RNA methylation site, a high-quality methylation-expression network was constructed, strengthened by an RNA co-methylation network and then smoothed by a protein-protein interaction network. Each vector in the resulting matrix is a ranked gene list which serves as an input for GSEA functional enrichment of each individual site based

on the guilt-by-association principle. By integrating gene expression profiles, RNA methylation profiles and protein interactome with the guilt-by association principle, we functionally annotated the top 10,000 most dynamic m^6A RNA methylation sites with 4,310,949 gene ontology terms, for the first time, at single site resolution, providing a useful reference for researchers who are particularly interested in the biological functions that are regulated at epitranscriptome layer through reversible m^6A methylation.

In-depth biology mining was performed on YTH domain m^6A reader families, and the molecular function enrichment of each gene set corresponds well with the experimental-validated functions of each reader. Please note that, of interests here are the biological functions that exhibit putative association to the epitranscriptome. It is different from an arbitrary functions of the m^6A site-containing gene, which may be regulated through other types of biological regulation, such as, the transcription factor, miRNA, DNA methylation, post-translational modification, etc.

By correlating with the RNA methylation profiles, the functions we predicted here are restricted to those that are likely to be regulated via the epitranscriptome, and should be favorable to the researchers in the RNA epigenetics field. The predicted functions of individual RNA methylation sites are available from our WHISTLE web server: www.xjtlu.edu.cn/biologicalsciences/whistle.

Chapter 3 **Annotating functions of m^{6}A sites conserved between human and mouse in m^{6}A-Atlas and other side projects**

Section 3.1 Introduction

A variety of chemical modifcations are naturally decorated on cellular RNAs, modulating their biogenesis, stability and functions [89]. To date, > 150 types of RNA modifcations have been identifed [90], among which, N6-methyladenosine (m^{6}A) is the most pervasive and the most intensively studied non-cap reversible marker present on eukaryotic mRNAs and lncRNAs [44]. Recent studies suggest that m^{6}A plays a pivotal role during various biological processes including stress [91], heat shock [6] and DNA damage [92], and regulates molecular functions such as RNA–protein interaction [93], RNA stability [5] and translation efficiency [49].

A number of high-throughput experimental approaches have been developed for profling the transcriptome-wide distribution of m^{6}A RNA modifcation, including, most notably, the antibody-based approach m^{6}A-seq (or MeRIP-seq) [17, 64]. With m^{6}A-seq, it is possible to identify condition-specifc m^{6}A sites [94, 95], quantify the m^{6}A methylation levels [96, 97], or compare between experimental condition [95]. Despite the limits of m^{6}A-seq regarding the reproducibility, data quality and mediocre resolution (around 100bp) [23], this technology has been widely applied to characterize the m^{6}A epitranscriptome under various biological contexts in more than 30 organisms since its invention in 2012. Besides m^{6}A-seq, there are also recent techniques such as miCLIP [20] and m^{6}A-CLIP [21], that offer improved or even baseresolution epitranscriptome determination. However, these approaches report

primarily the precise location of m^6A sites in physical, and are unsuitable for quantifcation of m^6A methylation levels.

To date, several bioinformatics websites and databases have been constructed aiming to properly collect, annotate, share and interpret the rapidly growing knowledge in RNA modifcations [30, 52, 90, 98]. Among them, Met-DB [30] is the frst epitranscriptome database collecting m^6A sites on mRNAs and lncRNAs (rather than small RNAs) reported from highthroughput sequencing approaches. The most recent release of Met-DB hosted m^6A sites in seven species collected from 185 m^6A-seq experiments. RMBase [52] is currently the most comprehensive epitranscriptome database containing ~1,397,000 RNA modifcation sites among 13 species including m^6A and other RNA modifcations such as m^5C and m^1A. These works addressed various aspects of RNA modifcations, and together greatly improved our understanding of the epitranscriptome. Nevertheless, the m^6A site collections in existing epitranscriptome databases (Met-DB and RMBase) suffer from limited reliability and only collect binary profiles.

To address these limitations, we constructed m^6A-Altas, a comprehensive knowledgebase for unraveling the N6-methyladenosine (m^6A) epitranscriptome. Compared to existing databases, m^6A-Atlas features a high-confdence collection of reliable m^6A sites identifed from base-resolution technologies only and the quantitative condition-specifc epitranscriptome profles estimated from a large number of high-throughput sequencing samples covering various tissues and cell lines. Conservation analysis is a powerful way for identifying the functionally

important m^{6}A sites [99]. The putative biological functions of individual m^{6}A sites were predicted for the conserved sites between human and mouse according to the guilt-by-association principle, as m^{6}Acomet [100]. This method is based on the association of methylation patterns (or co-methylation) among functionally related m^{6}A sites inferred from the collected quantitative epitranscriptome data.

The co-methylated m^{6}A sites exhibit correlated methylation status under different experimental conditions, it is often reasonable to speculate that they share some common regulators at the epitranscriptome layer and have related biological functions. This is the basic idea of the guilt-by-association principle. The guilt-by-association is a validated principle in network research, which states that if two patterns share some similar properties, they are most likely to share a connection. This principle has been widely applied in lncRNA functional prediction by the protein-protein interaction network [15], co-transcription factor network, and co-expression network [14]. The predicted biological functions of the m^{6}A sites may help generate hypotheses for subsequent experimental validation.

Section 3.2 Methods

Section 3.2.1 Quantifcation of m^{6}A methylation levels

There are a total of 22,359 m^{6}A sites in human that are conserved in mouse with *LiftOver* tool from the UCSC genome browser [101]. The epitranscriptome profiles (RNA methylation levels) of these sites were inferred from the m^{6}A-seq data obtained under 109 experimental conditions with *exomePeak* R/Bioconductor package [95].

Quantile normalization was performed to remove potential batch effects among samples.

Section 3.2.2 Construction of the co-methylation network

Two sites are considered to be co-methylated if their epitranscriptome profiles are significantly correlated (with p-value of the Pearson's correlation less than 0.05 and ranked in the highest or lowest 10%). The linkage between co-methylated site pairs was retained in the adjacency matrix of the co-methylation network with package *igraph* [41].

Section 3.2.3 Functional annotation with hub-based method

As the functions of individual m^6A sites are not available, we rely on the Gene Ontology (GO) of the m^6A-hosting genes, and infer the functions of individual m^6A sites using the hub-based method. In hub-based method, the function of the hub methylation site is determined by the enrichment result of its neighbor sites. Only the sites with more than three immediate neighbor sites are treated as the hub sites. We then use the significantly enrichment GO functions to annotate the hub sites. It may be worth mentioning that the GO terms associated with the hub site-carrying gene were not used in the prediction analysis, so that the predicted functions of m^6A site are independent from its hosting gene. Please also note that the GO prediction was made from the epitranscriptome profiles of individual m^6A site (rather than entire gene), and thus can only predict site-specific GO functions. The functions of methylation for the entire gene are not covered in this analysis.

Section 3.3 Results

Section 3.3.1 Construction of the co-methylation network

The matrix contains the methylation level of 22,359 conserved human m^6A sites under 109 samples was applied to generate the adjacency matrix, which is the co-methylation network as well. the co-methylation network consists of 19,149 nodes and 3,015,649 edges, with the degree distribution of the network in line with a typical scale-free network.

Section 3.3.2 Annotation of conserved m^6A sites with hub-based method

18,886 conserved hub sites with more than three immediate neighboring sites were annotated with GO framework. The Entrez Gene ID and associated GO terms of the genes carrying the neighboring site of a hub site were obtained and enrichment analyzed. With a p-value cutoff of 0.1 for the GO enrichment analysis, a total of 8,570,604 GO terms were associated with the 18,886 conserved hub sites. **Figure 12** shows the density plot of the GO term number. Each hub site is enriched with about 443 GO terms with the enrichment p-value cutoff as 0.1. When there are more than 50 terms associated with a hub m^6A site, the m^6A-Atlas database by default displays the top 50 most enriched GO terms. The annotation results can be visualized in the corresponding human m^6A sites on the m^6A-Atlas web server.

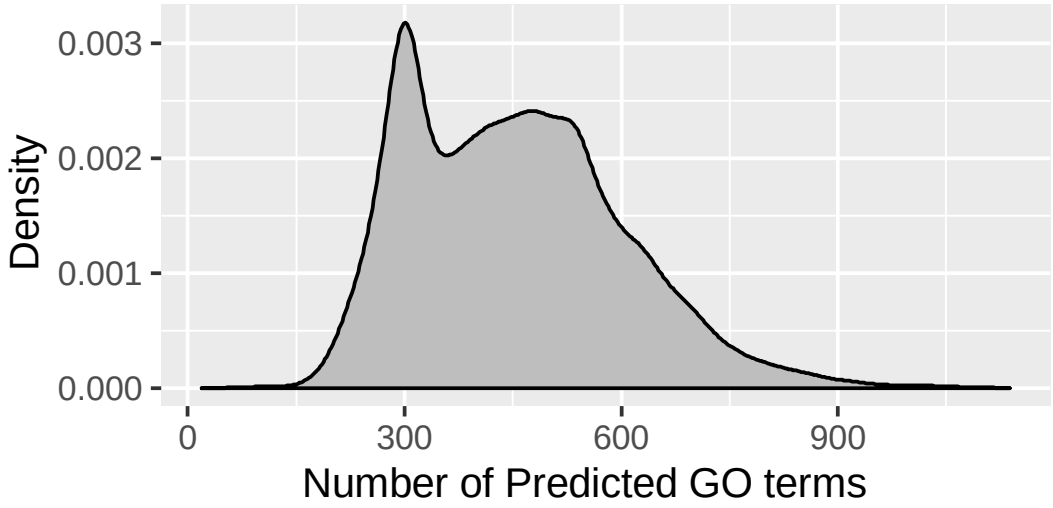

Figure 12 **Number of GO terms predicted to be associated with an individual m^{6}A site**. Most hub sites were enriched with GO terms number ranged from 200 to 700 with a cutoff of 0.1. On average, each hub m^{6}A site is associated with 443 GO terms for the enrichment p-value.

Section 3.4 Conclusion

This work is part of the m^{6}A-Atlas project. The co-methylation network was built to annotate the regulatory function of each human conserved site based on the guilt-by-association principle. With the hub-based approach described in the m^{6}Acomet, the GO framework was applied in functional enrichment of individual site. 18,886 human conserved sites were annotated with 8,570,604 gene ontology terms, and the predicted functions of individual human conserved sites are available from our m^{6}A-Atlas database: www.xjtlu.edu.cn/biologicalsciences/atlas.

Section 3.5 Side project 1: COVID-19 project

Section 3.5.1 Introduction

Three coronaviruses crosses the species barrier to human that bring deadly pneumonia to us since the 21st century: severe acute respiratory syndrome

coronavirus (SARS-CoV) [102], Middle-East respiratory syndrome coronavirus (MERS-CoV) [103], and SARS-CoV-2 [104]. SARS-CoV-2 was firstly discovered in Wuhan, Hubei province of China in December 2019, and was sequenced in January 2020 [105]. SARS-CoV-2, similar to other coronaviruses, is an enveloped virus with a ~30 kb, positive-sense, single-stranded RNA genome. The ongoing outbreak of atypical pneumonia (COVID-2019) caused by SARS-CoV-2 that has spread all over the world affected over 5,900,000 people and killed over 350,000 people as of the end of May. On January 30, 2020, the World Health Organization declared the SARS-CoV-2 epidemic a public health emergency of international concern. The main lethal symptoms of SARS-CoV-2 virus infecting on human immune system include adult respiratory distress syndrome (ARDS) and multiple organ dysfunction syndrome (MODS).

Lu et al. found that RNA m^6A might be an important mechanism that enables virus to escape the attacking of the host cell immune system [106]. Chemical modifications will not change the basic genetic information on sequence, while these modifications might bring new features on virus. We decide to focus on the most abundant RNA modification m^6A. It assumes that viral m^6A and host m^6A sites should have the sequence property, because viral transcripts rely on the enzyme of the host. It should be reasonable to use human m^6A site predictor to predict the m^6A sites on virus strains. Therefore, in this project, we used SRAMP [107] method, which is a m^6A predictor based on species' sequence information, to study the m^6A sites on SARS-CoV-2 by predicting on sequence data. Since many studies indicate that m^6A sites are more likely to enrich near the stop codon position [21], we also investigated the most

frequent appeared 27 m^6A 41bp sequences positional relationship with essential virus genes' stop codon. The results show that two m^6A 41bp sequences are quiet near the stop codon position of spike (S) protein which is the major mediation that assists the coronavirus entry into the host cells by forming the homotrimers protruding on the viral surface [108].

Section 3.5.2 Materials and mothods

451 SARS-CoV-2 strains were downloaded from China National Center for Bioinformation (2019 Novel Coronavirus Resource) (https://bigd.big.ac.cn/ncov) the beginning of April, 2020 in fasta format. SRMAP was used to predict the m^6A sites on SARS-CoV-2, which is a reliable predictor based on the sequence information of the species. The SARS-CoV-2 strain file was imported into SRAMP tool, and the predicted m^6A motifs with its corresponding confidence value were obtained. Then only the high/very high confidence m^6A motif would be retained with the extracted 41nt flanking window where the m^6A site settled at the central position. The most frequently 41bp sequences were detected and ordered by MUSCLE [109] according to the sequence similarity. The most frequent m^6A sites positions were compared with major SARS-CoV-2 stop codon positions to see whether they are specifically enriched near some stop codons.

Section 3.5.3 Results and discussion

SRAMP predicted results show in **Figure 13** the number of m^6A motifs that each strain has and the distribution indicates that most of the strains have 26-27 motifs (81.15%). Overall, 9980 high/very high confidence motifs are retained here. **Table 4** summary

the 27 most frequent motif sequences (41bp), which is ordered by the sequence similarity (MUSCLE). The centered m^{6}A site is in bold font. **Figure 14** demonstrates the comparison of the stop codon (red lines) with those m^{6}A motif sequences (41bp). We could notice that the third red line (left to right) stands for the S protein stop codon position of SARS-CoV-2, and it is very near to two sequences (No. 8: ATGGGAATCTGGAGTAAAAGACTGTGTTGTATTACACAGTT and No. 19: TAATCCTTATGACAGCAAGAACTGTGTATGATGATGGTGCT). These two types of m^{6}A motifs may play an important role in involving the virus infection. within the motif with an underline.

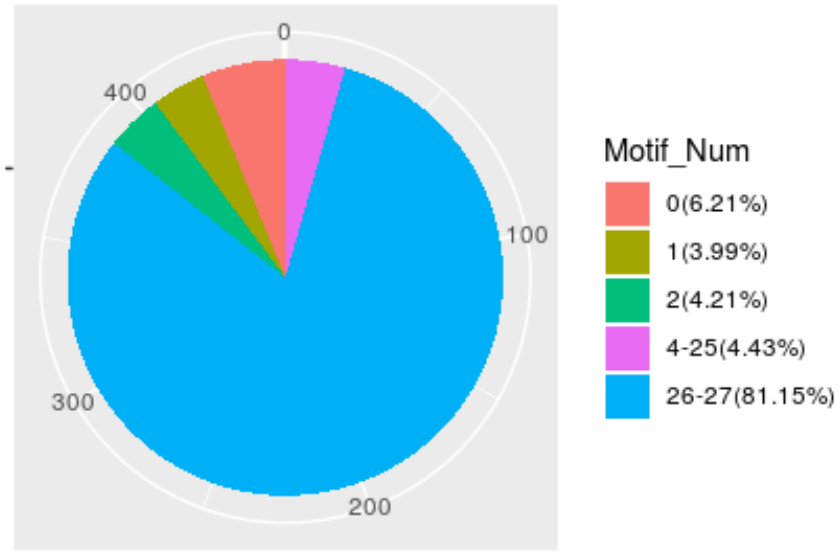

Figure 13 m^{6}A motif number(s) of individual strain. Most strains have 26-27 motifs.

Table 4 The Most Frequent 27 41bp Sequences Ordered By Sequence Similarity.

41bp_seq	Frequency
AAATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACG	386
AACCATTACAGATGCTGTAGACTGTGCACTTGACCCTCTCT	379
AAGCAAAATGTTGGACTGAGACTGACCTTACTAAAGGACCT	383
AGAGTCACATGTTGACACTGACTTAACAAAGCCTTACATTA	383
AGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCG	392
AGTGTGTAACATTAGGGAGGACTTGAAAGAGCCACCACATT	376
ATACCACTTATGTACAAAGGACTTCCTTGGAATGTAGTGCG	388
ATGGGAATCTGGAGTAAAAGACTGTGTTGTATTACACAGTT	377
ATTACATTACACATAAACGAACTTATGGATTTGTTTATGAG	383
CAAGCTGAAAATGTAACAGGACTCTTTAAAGATTGTAGTAA	235
CAAGCTGAAAATGTAACAGGACTTTTTAAAGATTGTAGTAA	149
CACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGG	384
CATACCTGGCATACCTAAGGACATGACCTATAGAAGACTCA	384
CTCTCAGCCTTTTCTTATGGACCTTGAAGGAAAACAGGGTA	373
GCTAACACCTGTACTGAAAGACTCAAGCTTTTTGCAGCAGA	381
GGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAA	407
GTAGCCTCAAAGATTTTGGGACTACCAACTCAAACTGTTGA	383
GTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAG	391
TAATCCTTATGACAGCAAGAACTGTGTATGATGATGGTGCT	383
TGAGAATCTTCACAATTGGAACTGTAACTTTGAAGCAAGGT	384
TGGAGTTCATGCTGGCACAGACTTAGAAGGTAACTTTTATG	383
TGGGTTATCTTCAACCTAGGACTTTTCTATTAAAATATAAT	379
TGTCTGAAGCAAAATGTTGGACTGAGACTGACCTTACTAAA	384
TTAGAAGGTAACTTTTATGGACCTTTTGTTGACAGGCAAAC	383
TTGATAAAGCTGGTCAAAAGACTTATGAAAGACATTCTCTC	383
TTGTTAAGCGTGTTGACTGGACTATTGAATATCCTATAATT	383
TTTCTTTGAGAGAAGTGAGGACTATTAAGGTGTTTACAACA	384

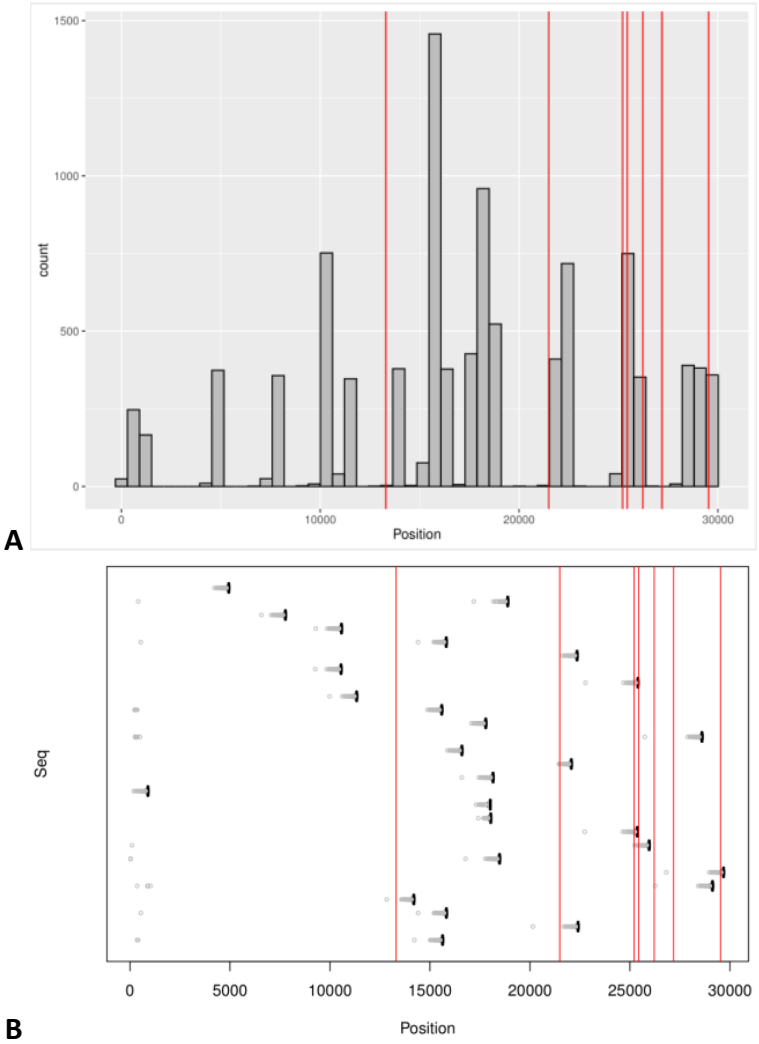

Figure 14 **Position distributions of most frequent motif sequences**. Red lines stand for stop codon positions. (**A**) Overall distribution (**B**) 27 most frequent sequences distribution (sequences in **Table 4**).

Section 3.5.4 Conclusion

To the best of our knowledge, we provided here the first computational prediction study of RNA methylation sites in SARS-CoV-2. According to the predicted results on SARS-CoV-2, we found the majority of each virus has 26-27 m^6A motifs. We also summarized the most frequent 27 motif sequences (41bp) here, and detected two of them are enriched near the S protein stop codon, which is the primary bridge that

enables the virus entering the host cells. The reason why we look at the stop codon position is that the previous m^6A studies stated the preferred enrichment of m^6A in those positions. The special relationship with the critical S protein indicates the potential role that m^6A might play in the virus infection. In the further, we could also explore the m^6A association with other virus parameters such as infection time, location, and additional clinical information to see whether there could be more implications.

Section 3.6 Side project 2: Nanopore project

Section 3.6.1 Introduction

To date, internal RNA modifications are widespread in different classes of RNA and essential in different biological functions such as human pathology, Biomolecule [1, 19, 110, 111]. However, due to the fact that the construct of the RNA sequence is complex, the mapping and quantifying approaches for modified RNA nucleotides still require further progress [16, 112, 113].

Looking through the development of sequencing technology - from Sanger's Method to the next-generation sequencing (NGS) technologies, and then to the third-generation sequencing (TGS) technologies, the TGS technologies has played the most pivotal role in the detection of m^6A modifications in RNA. Through this process, technical limitations, such as labor-intensive workload, high cost, and comparatively short reads per operation, have been subsequently addressed [114]. Liu et al [115] makes the assumption that the base-calling 'errors' happening during the process of in vitro transcription can reflect the current intensity changes and as a result detect

the m^6A. In view of this, we emphasize the meanings of base-calling 'errors' and make the similar assumptions with them. Later, we reproduce the results and validate the performance of SVM algorithms with different kernels on the detection of m^6A RNA modifications in native RNA sequence.

Section 3.6.2 Methodology

Liu et al.'s research [115] reveals that the bias of current intensity owing to RNA modifications causes the errors in base-calling. Furthermore, these errors may be utilized in the detection of m^6A RNA modification. Therefore, in order to verify the detection of m^6A, our experiment process follows three main steps: the generation of the raw data, extraction of features from based-called files and application of SVM models as the classifiers to train the data.

We used the raw data generated by Liu et al. Following their pipeline of processing based-called files, we extract and combine feartures such as base quality, mismatch and deletion frequency. In order to better emulate the RNA sequence model, a total of 4 sequences were designed by splitting the 10kb sequence (the current intensity changes of this sequence have been read by Nanopore before) into smaller sequences of slightly different size (2329bp, 2543bp, 2678bp and 2795bp, which we named 'Curlcake 1', 'Curlcake 2', 'Curlcake 3' and 'Curlcake 4', respectively) by using the software curlake. These four files can highly represent the construct of the RNA sequence and were in-vitro transcribed either in the presence of ATP or N6-Methyladenosine-5'-Triphosphate (m^6ATP). Sequences are performed with two replicates of modified RNA and two with unmodified.

After obtaining the four RNA sequence, we use the EpiNano (https://github.com/enovoa/EpiNano) to perform the pretreatment of the data so as to extract the base-calling features from the file. First, sequence are locally base-called using Albacore 2.1.7 and filted by NanoFilt. Setting '--headcrop 5 --tailcrop 3' to trim the first five nucleotides and the last three from the end. Secondly, convert the 'U' in the sequence to 'T' and map the synthetic sequences using minimap2 [112] with the settings '-ax map-on'. The mapped results are transformed into mpileup format using samtools1.4, and after that calling variants for each single read-to-reference alignment. Variants at each single reference site were extracted from the mapping results. Thus, we need to summarize the variants results into the format 'csv' file using cutomized Python script from [115]. Finally, extract the features from the fast5 file and get a csv file of 16 columns.

SVM is a supervised learning technique for classification by first constructing a hyperplanein a high- or infinite-dimensional space [116]. In the sample space, the hyperplane can be expressed as $\boldsymbol{\omega^T x+b=0}$, where $\boldsymbol{\omega}$=(ω_1;ω_2;ω_3;…;ω_d)is the normal vector of the hyperplane,'b'is the displacement term, which determine the distance between the origin and thehyperplane.Therefore,the distance between any point xin the space and the hyperplane ($\boldsymbol{\omega,b}$) could be written as $\mathbf{r} = \frac{|\boldsymbol{\omega}^{\mathbf{T}}\mathbf{x}+\mathbf{b}|}{\|\boldsymbol{\omega}\|}$. Suppose the hyperplane($\boldsymbol{\omega}, \mathbf{b}$) can classify the sample accurately, namely for $\forall(\mathbf{x_i}, \mathbf{y_i}) \in$ D(sample set), there are inequalities in **Equation 7**. The points closed to the hyperplane are support vector, where the equalities hold. The sum of the

distances between the support vectors and the points (also known as 'margin' is $\frac{2}{\|\omega\|}$, and larger margin means better generalization. Therefore, it is necessary to search for the pair$(\boldsymbol{\omega}, \mathbf{b})$ to maximize $\mathbf{r}$ in **Equation 8**. More specifically, the parameters can be solved by applying the mehod of Lagrange Multiplier in **Equation 9**. When the linear classification is impossible, we should use kernel function to map the original sample space to higher-dimensional space, namely in $\mathbf{f}(\mathbf{x}) = \boldsymbol{\omega}^{\mathbf{T}}\boldsymbol{\phi}(\mathbf{x}) + \mathbf{b}$.

Equation 7

$$\begin{cases} \boldsymbol{\omega}^{\mathbf{T}}\mathbf{x_i} + \mathbf{b} \geq +\mathbf{1}, & \mathbf{y_i} = +\mathbf{1} \\ \boldsymbol{\omega}^{\mathbf{T}}\mathbf{x_i} + \mathbf{b} \leq -\mathbf{1}, & \mathbf{y_i} = -\mathbf{1} \end{cases}$$

Equation 8

$$\max_{\boldsymbol{\omega},\mathbf{b}} \frac{2}{\|\boldsymbol{\omega}\|}$$

$$\mathbf{s.t.}\, \mathbf{y_i}(\boldsymbol{\omega}^{\mathbf{T}}\mathbf{x_i} + \mathbf{b}) \geq \mathbf{1}, \mathbf{i} = \mathbf{1}, \mathbf{2}, \mathbf{3}, \ldots \mathbf{m}.$$

Equation 9

$$\mathbf{L}(\boldsymbol{\omega}, \mathbf{b}, \boldsymbol{\alpha}) = \frac{\mathbf{1}}{\mathbf{2}}\|\boldsymbol{\omega}\|^2 + \sum_{\mathbf{i=1}}^{\mathbf{m}} \boldsymbol{\alpha}_\mathbf{i}\Big(\mathbf{1} - \mathbf{y_i}(\boldsymbol{\omega}^{\mathbf{T}}\mathbf{x_i} + \mathbf{b})\Big)$$

The most commonly used kernel are linear $\boldsymbol{\kappa}(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i^T x_j}$, polynomial $\boldsymbol{\kappa}(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i^T x_j})^{\mathbf{d}}$ and Gaussian kernel $\boldsymbol{\kappa}(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{e}^{\frac{\|\mathbf{x_i}-\mathbf{x_i}\|^2}{2\sigma^2}}$.In this project, we mainly used Linear and Gaussian kernel. After kernel mapping, the subsequent steps are similar.

we obtain four CSV files (since the overall RNA sequence is divided into four parts) after the process of feature extraction. Two of the Excels contain m^6A), and the other two collect the unmodified ribonucleotides ('unm').In order to train the data, we combine the Excels into two parts, each including both 'm^6A' and 'unm' data. We then add a new column to the table called 'Sample' which is used for denoting the situation of the nucleotide (0 as the unmodified and 1 as the modified). We use the SVM classifier to numerically identify each feature. To improve the accuracy of the experiment, we shall use the file1 (rep1) to carry out training and testing independently, and use file2 to conduct the cross-validation check. For the file1, we do the independent testing and training, namely 75% of the data for training and 25% of the data for predicting. On the other hand, file2 is used for cross-validation of the training which we divide the data into five cross to validate the result we obtain from file1. Therefore, the first part focus more on the features while the second part lay emphasis on the performance of kernel function (adjusting the parameters to maximize the performance of the classifiers).

Section 3.6.3 Results

The method of dividing a long RNA sequence into many 5-mer sequences is that we consecutively shift one nucleotide at a time. (e.g., the raw sequence is 'GACGUAG', and then the three 5mer sequences are 'GACGU', 'ACGUA', and 'CGUAG'). Due to the fact that the m^6A modification tends to present in a well-defined RNA motif, RRACH, we focus on the structure of 5-mer. In addition, since the raw data has large numbers of 5-mer centred in G, C, U, we attribute these 5-mers as the control group to make comparison with the m^6A group.

We focus on whether a single feature at position0 is able to classify a given RRACH k-mer into m^6A-modified or unmodified nucleotides. Results show that base quality, deletion frequency and mismatch frequency are able to predict the test with 66-80% accuracy while some other features like current intensity are not strong enough to distinguish the example (50% accuracy). To better compare the AUC score between each single feature, we draw the ROC curve below (**Figure 15**).Though we cannot assert which factor has greatest impact (since the ROC curve intersect with each other), current intensity has a poor performance with the curve almost coinciding with y=x. As a control, we use the same set of features in control k-mers (i.e., those with the same sequence context, but centralize in C, G, U), finding that the features do not distinguish between m^6A-modified and m^6A-unmodfied datasets. To make the difference between m^6A group and control group more obvious, we visualize the feature performance of the control group on the same figure. It is clear that the control group fail to distinguish the RNA and the figure also illustrates the importance of centralizing in 'A' during the detection of m^6A. To improve the performance of the algorithm, we then examined whether a combination of the features will improve the prediction accuracy. Results (**Figure 16**) show that the combination of these three features can achieve the highest accuracy above 83% with an AUC score over ninety percent. (Since the current intensity change is too weak to be a feature, we neglect it in the experiment). Finally, we test whether the inclusion of all the features from the neighbouring positions (-2, -1, 1, 2) will improve the performance of the algorithm. We find out that the inclusion of the neighbourhood will increase the

accuracy to 89-91% (AUC > 95%) (**Figure 17**), suggesting that the feature of neighbourhood should be considered into the experiment.

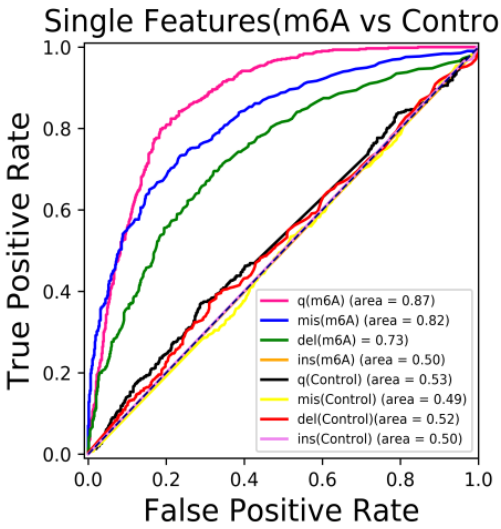

Figure 15 Single Features (m^6A vs Control).

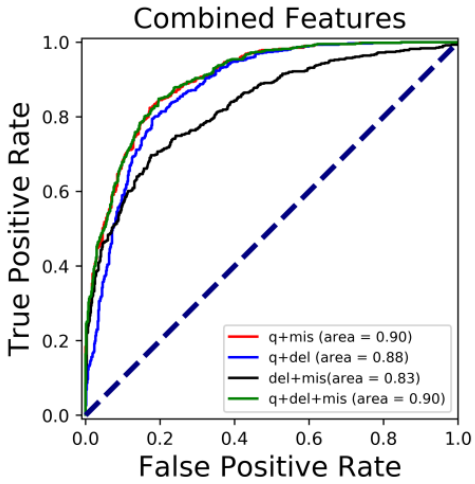

Figure 16 Accuracy of the combined Features.

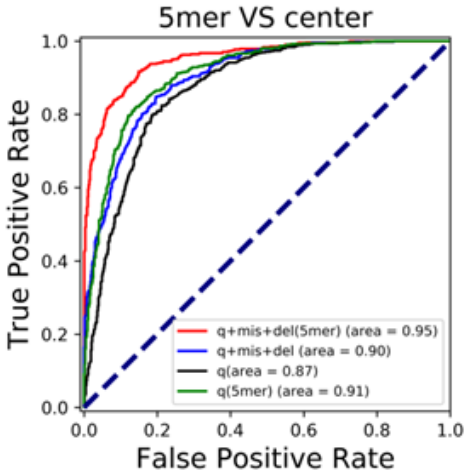

Figure 17 Inclusion of all the features from the neighbouring positions improves the performance.

Therefore, we choose 'mean per-base quality', 'mismatch frequency' and 'deletion frequency' as the input features (since the current intensity change is too weak to be a feature, we neglect it in the following experiment). To optimize the model, we examine the performance of SVM with different kernel and utilize grid search to obtain the optimal parameters. From the perspective of run-time, polynomial kernel is time-consuming, while Linear and Gaussian kernel is relatively time-efficient. The results also show that in terms of accuracy and AUC, the performance of SVM with linear kernel are slightly worse than Gaussian kernel and the following discussion will focus on the Gaussian kernel. The result (**Figure 18**) reveals that when gamma = 0.01 and C = 100, the accuracy of test data attains its maximum which is close to one hundred percent. To evaluate the sensitivity of the model from the data, we apply five-fold cross-validation to assess the performance of the model in terms of accuracy, AUC and recall rate. We find that the accuracy is between 94% and 97%, which is slightly worse than test set and the recall rate is 92-96%. The AUC scores get close to 99% which informs us that the model is relatively strong and robust. The

overall performance is close to Liu's research and our model is even slightly better in term of accuracy.

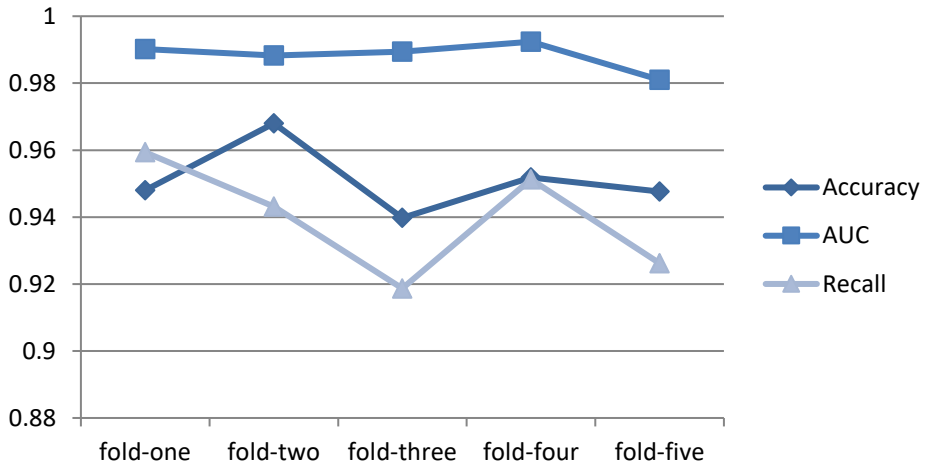

Figure 18 Results of the cross-validation.

Bibliography

1. Fu Yu, Dominissini Dan, Rechavi Gideon, He Chuan: **Gene expression regulation mediated through reversible m(6)A RNA methylation**. *Nature Reviews Genetics* 2014, **15**(5):293-306.
2. Bokar Joseph A., Shambaugh Mary, Polayes D., Matera A. G., Rottman F. M.: **Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase**. *RNA* 1997, **3**(11):1233-1247.
3. Novoa Eva Maria, Mason Christopher E., Mattick John S.: **RNA PROCESSING AND MODIFICATIONS Charting the unknown epitranscriptome**. *Nat Rev Mol Cell Bio* 2017, **18**(6):339-340.
4. Liu Jianzhao, Yue Yanan, Han Dali, Wang Xiao, Fu Ye, Zhang Liang, Jia Guifang, Yu Miao, Lu Zhike, Deng Xin *et al*: **A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation**. *Nat Chem Biol* 2014, **10**(2):93-95.
5. Wang Xiao, Lu Zhike, Gomez Adrian, Hon Gary C, Yue Yanan, Han Dali, Fu Ye, Parisien Marc, Dai Qing, Jia Guifang *et al*: **N6-methyladenosine-dependent regulation of messenger RNA stability**. *Nature* 2014, **505**(7481):117-120.
6. Zhou Jun, Wan Ji, Gao Xiangwei, Zhang Xingqian, Jaffrey Samie R, Qian Shu-Bing: **Dynamic m(6)A mRNA methylation directs translational control of heat shock response**. *Nature* 2015, **526**(7574):591-594.
7. Lokody Isabel: **Gene regulation: RNA methylation regulates the circadian clock**. *Nat Rev Genet* 2014, **15**(1):3.
8. Deng Xiaolan, Su Rui, Feng Xuesong, Wei Minjie, Chen Jianjun: **Role of N(6)-methyladenosine modification in cancer**. *Curr Opin Genet Dev* 2018, **48**:1-7.
9. He Chuan: **Grand challenge commentary: RNA epigenetics?** *Nat Chem Biol* 2010, **6**(12):863-865.
10. Chen Xing, Sun Ya-Zhou, Liu Hui, Zhang Lin, Li Jian-Qiang, Meng Jia: **RNA methylation and diseases: experimental results, databases, Web servers and computational models**. *Briefings in Bioinformatics* 2017, **20**(3):896-917.
11. Chandola Udita, Das Radhika, Panda Binay: **Role of the N6-methyladenosine RNA mark in gene regulation and its implications on development and disease**. *Brief Funct Genomics* 2015, **14**(3):169-179.
12. Chen Kunqi, Wei Zhen, Liu Hui, Magalhães João Pedro de, Rong Rong, Lu Zhiliang, Meng Jia: **Enhancing Epitranscriptome Module Detection from m⁶A-Seq Data Using Threshold-Based Measurement Weighting Strategy**. *BioMed Research International* 2018, **2018**:2075173.
13. Zhang Lin, He Yanling, Wang Huaizhi, Liu Hui, Huang Yufei, Wang Xuesong, Meng Jia: **Clustering Count-based RNA Methylation Data Using a Nonparametric Generative Model**. *Current Bioinformatics* 2018, **13**.
14. Liao Qi, Liu Changning, Yuan Xiongying, Kang Shuli, Miao Ruoyu, Xiao Hui, Zhao Guoguang, Luo Haitao, Bu Dechao, Zhao Haitao *et al*: **Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network**. *Nucleic Acids Res* 2011, **39**(9):3864-3878.
15. Stelzl Ulrich, Worm Uwe, Lalowski Maciej, Haenig Christian, Brembeck Felix H, Goehler Heike, Stroedicke Martin, Zenkner Martina, Schoenherr Anke,

Koeppen Susanne *et al*: **A human protein-protein interaction network: a resource for annotating the proteome**. *Cell* 2005, **122**(6):957-968.

16. Helm Mark, Motorin Yuri: **Detecting RNA modifications in the epitranscriptome: predict and validate**. *Nat Rev Genet* 2017, **18**(5):275-291.
17. Dominissini Dan, Moshitch-Moshkovitz Sharon, Schwartz Schraga, Salmon-Divon Mali, Ungar Lior, Osenberg Sivan, Cesarkas Karen, Jacob-Hirsch Jasmine, Amariglio Ninette, Kupiec Martin *et al*: **Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq**. *Nature* 2012, **485**(7397):201-206.
18. Saletore Yogesh, Meyer Kate, Korlach Jonas, Vilfan Igor D, Jaffrey Samie, Mason Christopher E: **The birth of the Epitranscriptome: deciphering the function of RNA modifications**. *Genome Biol* 2012, **13**(10):175.
19. Lee Mihye, Kim Boseon, Kim V Narry: **Emerging roles of RNA modification: m(6)A and U-tail**. *Cell* 2014, **158**(5):980-987.
20. Linder Bastian, Grozhik Anya V, Olarerin-George Anthony O, Meydan Cem, Mason Christopher E, Jaffrey Samie R: **Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome**. *Nat Methods* 2015, **12**(8):767-772.
21. Ke Shengdong, Alemu Endalkachew A, Mertens Claudia, Gantman Emily Conn, Fak John J, Mele Aldo, Haripal Bhagwattie, Zucker-Scharff Ilana, Moore Michael J, Park Christopher Y *et al*: **A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation**. *Genes Dev* 2015, **29**(19):2037-2053.
22. Love Michael I, Huber Wolfgang, Anders Simon: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
23. Zhang Teng, Zhang Shao-Wu, Zhang Lin, Meng Jia: **trumpet: transcriptome-guided quality assessment of m(6)A-seq data**. *BMC Bioinformatics* 2018, **19**(1):260.
24. Hansen Kasper D., Irizarry Rafael A., Wu Zhijin: **Removing technical variability in RNA-seq data using conditional quantile normalization**. *Biostatistics* 2012, **13**(2):204-216.
25. Consortium The Gene Ontology, Ashburner Michael, Ball Catherine A., Blake Judith A., Botstein David, Butler Heather, Cherry J. Michael, Davis Allan P., Dolinski Kara, Dwight Selina S. *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
26. Consortium Gene Ontology: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D258-261.
27. Vu Ly P, Pickering Brian F, Cheng Yuanming, Zaccara Sara, Nguyen Diu, Minuesa Gerard, Chou Timothy, Chow Arthur, Saletore Yogesh, MacKay Matthew *et al*: **The N(6)-methyladenosine (m(6)A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells**. *Nat Med* 2017, **23**(11):1369-1376.
28. Ke Shengdong, Pandya-Jones Amy, Saito Yuhki, Fak John J, Vagbo Cathrine Broberg, Geula Shay, Hanna Jacob H, Black Douglas L, Darnell James E Jr, Darnell Robert B: **m(6)A 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.

29. Chen Kunqi, Wei Zhen, Zhang Qing, Wu Xiangyu, Rong Rong, Lu Zhiliang, Su Jionglong, de Magalhaes Joao Pedro, Rigden Daniel J, Meng Jia: **WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach**. *Nucleic Acids Res* 2019, **47**(7):e41.
30. Liu Hui, Wang Huaizhi, Wei Zhen, Zhang Songyao, Hua Gang, Zhang Shao-Wu, Zhang Lin, Gao Shou-Jiang, Meng Jia, Chen Xing *et al*: **MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome**. *Nucleic Acids Res* 2018, **46**(D1):D281-D287.
31. Kim Daehwan, Pertea Geo, Trapnell Cole, Pimentel Harold, Kelley Ryan, Salzberg Steven L: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**(4):R36.
32. Fustin J. M., Doi M., Yamaguchi Y., Hida H., Nishimura S., Yoshida M., Isagawa T., Morioka M. S., Kakeya H., Manabe I.: **RNA-methylation-dependent RNA processing controls the speed of the circadian clock**. *Cell* 2013, **155**.
33. Liu Jianzhao, Yue Yanan, Han Dali, Wang Xiao, Fu Ye, Zhang Liang, Jia Guifang, Yu Miao, Lu Zhike, Deng Xin: **A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation**. *Nature chemical biology* 2014, **10**(2):93-95.
34. Schwartz Schraga, Mumbach Maxwellr., Jovanovic Marko, Wang Tim, Maciag Karolina, Bushkin G. Guy, Mertins Philipp, Ter-Ovanesyan Dmitry, Habib Naomi, Cacchiarelli Davide: **Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites**. *Cell Reports* 2014, **8**(1):284-296.
35. Li Zejuan, Weng Hengyou, Su Rui, Weng Xiaocheng, Zuo Zhixiang, Li Chenying, Huang Huilin, Nachtergaele Sigrid, Dong Lei, Hu Chao: **FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N 6 -Methyladenosine RNA Demethylase**. *Cancer Cell* 2016.
36. Barbieri Isaia, Tzelepis Konstantinos, Pandolfini Luca, Shi Junwei, Millán-Zambrano Gonzalo, Robson Samuel C, Aspris Demetrios, Migliori Valentina, Bannister Andrew J, Han Namshik: **Promoter-bound METTL3 maintains myeloid leukaemia by m 6 A-dependent translation control**. *Nature* 2017, **552**(7683):126.
37. Batista Pedro J, Molinie Benoit, Wang Jinkai, Qu Kun, Zhang Jiajing, Li Lingjie, Bouley Donna M, Lujan Ernesto, Haddad Bahareh, Daneshvar Kaveh: **m 6 A RNA modification controls cell fate transition in mammalian embryonic stem cells**. *Cell stem cell* 2014, **15**(6):707-719.
38. Langfelder Peter, Horvath Steve: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.
39. Barabasi Albert-Laszlo, Oltvai Zoltan N: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.
40. Su Gang, Morris John H, Demchak Barry, Bader Gary D: **Biological network exploration with Cytoscape 3**. *Curr Protoc Bioinformatics* 2014, **47**:8 13 11-24.

41. Csardi Gabor, Nepusz Tamas: **The igraph software package for complex network research**. *InterJournal, complex systems* 2006, **1695**(5):1-9.
42. Van Dongen S.: **Graph clustering via a discrete uncoupling process**. *Siam J Matrix Anal A* 2008, **30**(1):121-141.
43. Obayashi Takeshi, Kagaya Yuki, Aoki Yuichi, Tadaka Shu, Kinoshita Kengo: **COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference**. *Nucleic Acids Res* 2019, **47**(D1):D55-D62.
44. Meyer Kate D, Jaffrey Samie R: **Rethinking m(6)A Readers, Writers, and Erasers**. *Annu Rev Cell Dev Biol* 2017, **33**:319-342.
45. Visvanathan Abhirami, Patil Vikas, Arora Anjali, Hegde A S, Arivazhagan A, Santosh V, Somasundaram Kumar: **Essential role of METTL3-mediated m(6)A modification in glioma stem-like cells maintenance and radioresistance**. *Oncogene* 2018, **37**(4):522-533.
46. Wang Yang, Li Yue, Toth Julia I, Petroski Matthew D, Zhang Zhaolei, Zhao Jing Crystal: **N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells**. *Nat Cell Biol* 2014, **16**(2):191-198.
47. Patil Deepak P, Chen Chun-Kan, Pickering Brian F, Chow Amy, Jackson Constanza, Guttman Mitchell, Jaffrey Samie R: **m(6)A RNA methylation promotes XIST-mediated transcriptional repression**. *Nature* 2016, **537**(7620):369-373.
48. Xu Kai, Yang Ying, Feng Gui-Hai, Sun Bao-Fa, Chen Jun-Qing, Li Yu-Fei, Chen Yu-Sheng, Zhang Xin-Xin, Wang Chen-Xin, Jiang Li-Yuan *et al*: **Mettl3-mediated m(6)A regulates spermatogonial differentiation and meiosis initiation**. *Cell Res* 2017, **27**(9):1100-1114.
49. Wang Xiao, Zhao Boxuan Simen, Roundtree Ian A, Lu Zhike, Han Dali, Ma Honghui, Weng Xiaocheng, Chen Kai, Shi Hailing, He Chuan: **N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency**. *Cell* 2015, **161**(6):1388-1399.
50. Dominissini Dan, Moshitch-Moshkovitz Sharon, Salmon-Divon Mali, Amariglio Ninette, Rechavi Gideon: **Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing**. *Nat Protoc* 2013, **8**(1):176-189.
51. Sun Wen-Ju, Li Jun-Hao, Liu Shun, Wu Jie, Zhou Hui, Qu Liang-Hu, Yang Jian-Hua: **RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data**. *Nucleic Acids Research* 2015.
52. Xuan Jia-Jia, Sun Wen-Ju, Lin Peng-Hui, Zhou Ke-Ren, Liu Shun, Zheng Ling-Ling, Qu Liang-Hu, Yang Jian-Hua: **RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data**. *Nucleic Acids Res* 2018, **46**(D1):D327-D334.
53. Bottini Silvia, Pratella David, Grandjean Valerie, Repetto Emanuela, Trabucchi Michele: **Recent computational developments on CLIP-seq data analysis and microRNA targeting implications**. *Brief Bioinform* 2018, **19**(6):1290-1301.
54. Ideker Trey, Nussinov Ruth: **Network approaches and applications in biology**. *PLoS Comput Biol* 2017, **13**(10):e1005771.

55. Wang Xiujuan, Gulbahce Natali, Yu Haiyuan: **Network-based methods for human disease gene prediction**. *Briefings in Functional Genomics* 2011, **10**(5):280-293.
56. Chen Xing, Liu Ming-Xi, Yan Gui-Ying: **Drug-target interaction prediction by random walk on the heterogeneous network**. *Mol Biosyst* 2012, **8**(7):1970-1978.
57. Szklarczyk Damian, Franceschini Andrea, Wyder Stefan, Forslund Kristoffer, Heller Davide, Huerta-Cepas Jaime, Simonovic Milan, Roth Alexander, Santos Alberto, Tsafou Kalliopi P *et al*: **STRING v10: protein-protein interaction networks, integrated over the tree of life**. *Nucleic Acids Res* 2015, **43**(Database issue):D447-452.
58. Leale Guillermo, Baya Ariel Emilio, Milone Diego H, Granitto Pablo M, Stegmayer Georgina: **Inferring Unknown Biological Function by Integration of GO Annotations and Gene Expression Data**. *IEEE/ACM Trans Comput Biol Bioinform* 2018, **15**(1):168-180.
59. Lefever Steve, Anckaert Jasper, Volders Pieter-Jan, Luypaert Manuel, Vandesompele Jo, Mestdagh Pieter: **decodeRNA- predicting non-coding RNA functions using guilt-by-association**. *Database (Oxford)* 2017, **2017**.
60. Vanunu Oron, Magger Oded, Ruppin Eytan, Shlomi Tomer, Sharan Roded: **Associating genes and protein complexes with disease via network propagation**. *PLoS Comput Biol* 2010, **6**(1):e1000641.
61. Hofree Matan, Shen John P, Carter Hannah, Gross Andrew, Ideker Trey: **Network-based stratification of tumor mutations**. *Nat Methods* 2013, **10**(11):1108-1115.
62. Subramanian Aravind, Tamayo Pablo, Mootha Vamsi K., Mukherjee Sayan, Ebert Benjamin L., Gillette Michael A., Paulovich Amanda, Pomeroy Scott L., Golub Todd R., Lander Eric S. *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545-15550.
63. Kim Daehwan, Langmead Ben, Salzberg Steven L: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods* 2015, **12**(4):357-360.
64. Meyer Kate D, Saletore Yogesh, Zumbo Paul, Elemento Olivier, Mason Christopher E, Jaffrey Samie R: **Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons**. *Cell* 2012, **149**(7):1635-1646.
65. Zhang Sicong, Zhao Boxuan Simen, Zhou Aidong, Lin Kangyu, Zheng Shaoping, Lu Zhike, Chen Yaohui, Sulman Erik P, Xie Keping, Bogler Oliver *et al*: **m(6)A Demethylase ALKBH5 Maintains Tumorigenicity of Glioblastoma Stem-like Cells by Sustaining FOXM1 Expression and Cell Proliferation Program**. *Cancer Cell* 2017, **31**(4):591-606 e596.
66. Pendleton Kathryn E, Chen Beibei, Liu Kuanqing, Hunter Olga V, Xie Yang, Tu Benjamin P, Conrad Nicholas K: **The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention**. *Cell* 2017, **169**(5):824-835 e814.
67. Weng Hengyou, Huang Huilin, Wu Huizhe, Qin Xi, Zhao Boxuan Simen, Dong Lei, Shi Hailing, Skibbe Jennifer, Shen Chao, Hu Chao *et al*: **METTL14 Inhibits**

Hematopoietic Stem/Progenitor Differentiation and Promotes Leukemogenesis via mRNA m(6)A Modification. *Cell Stem Cell* 2018, **22**(2):191-205 e199.

68. Pollard Katherine, Dudoit Sandrine, Laan Mark: **Multiple Testing Procedures: the multtest Package and Applications to Genomics**. *Mark J van der Laan* 2005.
69. Szklarczyk Damian, Morris John H, Cook Helen, Kuhn Michael, Wyder Stefan, Simonovic Milan, Santos Alberto, Doncheva Nadezhda T, Roth Alexander, Bork Peer: **The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible**. *Nucleic acids research* 2016:gkw937.
70. Iorio Francesco, Bernardo-Faura Marti, Gobbi Andrea, Cokelaer Thomas, Jurman Giuseppe, Saez-Rodriguez Julio: **Efficient randomization of biological networks while preserving functional characterization of individual nodes**. *BMC Bioinformatics* 2016, **17**(1):542.
71. Yu Guangchuang, Wang Li-Gen, Han Yanyan, He Qing-Yu: **clusterProfiler: an R package for comparing biological themes among gene clusters**. *OMICS* 2012, **16**(5):284-287.
72. Wang James Z, Du Zhidian, Payattakool Rapeeporn, Yu Philip S, Chen Chinfu: **A new method to measure the semantic similarity of GO terms**. *Bioinformatics* 2007, **23**(10):1274-1281.
73. Yu Guangchuang, Li Fei, Qin Yide, Bo Xiaochen, Wu Yibo, Wang Shengqi: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products**. *Bioinformatics* 2010, **26**(7):976-978.
74. Benjamini Yoav, Hochberg Yosef: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995, **57**(1):289-300.
75. Shi Hailing, Wang Xiao, Lu Zhike, Zhao Boxuan S, Ma Honghui, Hsu Phillip J, Liu Chang, He Chuan: **YTHDF3 facilitates translation and decay of N(6)-methyladenosine-modified RNA**. *Cell Res* 2017, **27**(3):315-328.
76. Zhao Boxuan Simen, Wang Xiao, Beadell Alana V, Lu Zhike, Shi Hailing, Kuuspalu Adam, Ho Robert K, He Chuan: **m(6)A-dependent maternal mRNA clearance facilitates zebrafish maternal-to-zygotic transition**. *Nature* 2017, **542**(7642):475-478.
77. Roundtree Ian A, Luo Guan-Zheng, Zhang Zijie, Wang Xiao, Zhou Tao, Cui Yiquang, Sha Jiahao, Huang Xingxu, Guerrero Laura, Xie Phil *et al*: **YTHDC1 mediates nuclear export of N(6)-methyladenosine methylated mRNAs**. *Elife* 2017, **6**.
78. Xiao Wen, Adhikari Samir, Dahal Ujwal, Chen Yu-Sheng, Hao Ya-Juan, Sun Bao-Fa, Sun Hui-Ying, Li Ang, Ping Xiao-Li, Lai Wei-Yi *et al*: **Nuclear m6A Reader YTHDC1 Regulates mRNA Splicing**. *Molecular Cell* 2016, **61**(4):507-519.
79. Yang Ying, Sun Bao-Fa, Xiao Wen, Yang Xin, Sun Hui-Ying, Zhao Yong-Liang, Yang Yun-Gui: **Dynamic m6A modification and its emerging regulatory role in mRNA splicing**. *Science Bulletin* 2015, **60**(1):21-32.
80. Coles Charlotte H, Bradke Frank: **Microtubule self-organization via protein-RNA network crosstalk**. *Cell* 2014, **158**(2):245-247.

81. Shuman Stewart: **What messenger RNA capping tells us about eukaryotic evolution**. *Nat Rev Mol Cell Bio* 2002, **3**(8):619-625.
82. Chen Jixiang, Sun Yaocheng, Xu Xiao, Wang Dawei, He Junbo, Zhou Hailang, Lu Ying, Zeng Jian, Du Fengyi, Gong Aihua *et al*: **YTH domain family 2 orchestrates epithelial-mesenchymal transition/proliferation dichotomy in pancreatic cancer cells**. *Cell Cycle* 2017, **16**(23):2259-2271.
83. Li Jiangfeng, Meng Shuai, Xu Mingjie, Wang Song, He Liujia, Xu Xin, Wang Xiao, Xie Liping: **Downregulation of N(6)-methyladenosine binding YTHDF2 protein mediated by miR-493-3p suppresses prostate cancer by elevating N(6)-methyladenosine levels**. *Oncotarget* 2018, **9**(3):3752-3764.
84. Sabatinos Sarah A, Forsburg Susan L: **Managing Single-Stranded DNA during Replication Stress in Fission Yeast**. *Biomolecules* 2015, **5**(3):2123-2139.
85. Trzeciak Andrzej R, Mohanty Joy G, Jacob Kimberly D, Barnes Janice, Ejiogu Ngozi, Lohani Althaf, Zonderman Alan B, Rifkind Joseph M, Evans Michele K: **Oxidative damage to DNA and single strand break repair capacity: relationship to other measures of oxidative stress in a population cohort**. *Mutat Res* 2012, **736**(1-2):93-103.
86. Thomas Marshall P, Liu Xing, Whangbo Jennifer, McCrossan Geoffrey, Sanborn Keri B, Basar Emre, Walch Michael, Lieberman Judy: **Apoptosis Triggers Specific, Rapid, and Global mRNA Decay with 3' Uridylated Intermediates Degraded by DIS3L2**. *Cell Rep* 2015, **11**(7):1079-1089.
87. Tokmakov Alexander A, Iguchi Sho, Iwasaki Tetsushi, Fukami Yasuo, Sato Ken-Ichi: **Global decay of mRNA is a hallmark of apoptosis in aging Xenopus eggs**. *RNA biology* 2017, **14**(3):339-346.
88. Zhou Jun, Rode Kara A, Qian Shu-Bing: **m(6)A: A novel hallmark of translation**. *Cell Cycle* 2016, **15**(3):309-310.
89. Jones Joshua D, Monroe Jeremy, Koutmou Kristin S: **A molecular-level perspective on the frequency, distribution, and consequences of messenger RNA modifications**. *Wiley Interdiscip Rev RNA* 2020, **11**(4):e1586.
90. Boccaletto Pietro, Machnicka Magdalena A, Purta Elzbieta, Piatkowski Pawel, Baginski Blazej, Wirecki Tomasz K, de Crecy-Lagard Valerie, Ross Robert, Limbach Patrick A, Kotter Annika *et al*: **MODOMICS: a database of RNA modification pathways. 2017 update**. *Nucleic Acids Res* 2018, **46**(D1):D303-D307.
91. Engel Mareen, Eggert Carola, Kaplick Paul M, Eder Matthias, Roh Simone, Tietze Lisa, Namendorf Christian, Arloth Janine, Weber Peter, Rex-Haffner Monika *et al*: **The Role of m(6)A/m-RNA Methylation in Stress Response Regulation**. *Neuron* 2018, **99**(2):389-403 e389.
92. Xiang Yang, Laurent Benoit, Hsu Chih-Hung, Nachtergaele Sigrid, Lu Zhike, Sheng Wanqiang, Xu Chuanyun, Chen Hao, Ouyang Jian, Wang Siqing *et al*: **RNA m(6)A methylation regulates the ultraviolet-induced DNA damage response**. *Nature* 2017, **543**(7646):573-576.
93. Liu Nian, Dai Qing, Zheng Guanqun, He Chuan, Parisien Marc, Pan Tao: **N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions**. *Nature* 2015, **518**(7540):560-564.
94. Feng Jianxing, Liu Tao, Qin Bo, Zhang Yong, Liu Xiaole Shirley: **Identifying ChIP-seq enrichment using MACS**. *Nat Protoc* 2012, **7**(9):1728-1740.

95. Meng Jia, Lu Zhiliang, Liu Hui, Zhang Lin, Zhang Shaowu, Chen Yidong, Rao Manjeet K, Huang Yufei: **A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package**. *Methods* 2014, **69**(3):274-281.
96. Liu Lian, Zhang Shao-Wu, Huang Yufei, Meng Jia: **QNB: differential RNA methylation analysis for count-based small-sample sequencing data with a quad-negative binomial model**. *BMC Bioinformatics* 2017, **18**(1):387.
97. Li Jianwei, Huang Yan, Cui Qinghua, Zhou Yuan: **m6Acorr: an online tool for the correction and comparison of m(6)A methylation profiles**. *BMC Bioinformatics* 2020, **21**(1):31.
98. Liu Shun, Zhu Allen, He Chuan, Chen Mengjie: **REPIC: a database for exploring the N(6)-methyladenosine methylome**. *Genome Biol* 2020, **21**(1):100.
99. Liu Zhen, Zhang Jianzhi: **Most m6A RNA Modifications in Protein-Coding Regions Are Evolutionarily Unconserved and Likely Nonfunctional**. *Mol Biol Evol* 2018, **35**(3):666-675.
100. Wu Xiangyu, Wei Zhen, Chen Kunqi, Zhang Qing, Su Jionglong, Liu Hui, Zhang Lin, Meng Jia: **m6Acomet: large-scale functional prediction of individual m(6)A RNA methylation sites from an RNA co-methylation network**. *BMC Bioinformatics* 2019, **20**(1):223.
101. Speir Matthew L, Zweig Ann S, Rosenbloom Kate R, Raney Brian J, Paten Benedict, Nejad Parisa, Lee Brian T, Learned Katrina, Karolchik Donna, Hinrichs Angie S *et al*: **The UCSC Genome Browser database: 2016 update**. *Nucleic Acids Res* 2016, **44**(D1):D717-725.
102. Drosten Christian, Gunther Stephan, Preiser Wolfgang, van der Werf Sylvie, Brodt Hans-Reinhard, Becker Stephan, Rabenau Holger, Panning Marcus, Kolesnikova Larissa, Fouchier Ron A M *et al*: **Identification of a novel coronavirus in patients with severe acute respiratory syndrome**. *N Engl J Med* 2003, **348**(20):1967-1976.
103. Zaki Ali M, van Boheemen Sander, Bestebroer Theo M, Osterhaus Albert D M E, Fouchier Ron A M: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia**. *N Engl J Med* 2012, **367**(19):1814-1820.
104. Huang Chaolin, Wang Yeming, Li Xingwang, Ren Lili, Zhao Jianping, Hu Yi, Zhang Li, Fan Guohui, Xu Jiuyang, Gu Xiaoying *et al*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China**. *Lancet* 2020, **395**(10223):497-506.
105. Zhu Na, Zhang Dingyu, Wang Wenling, Li Xingwang, Yang Bo, Song Jingdong, Zhao Xiang, Huang Baoying, Shi Weifeng, Lu Roujian *et al*: **A Novel Coronavirus from Patients with Pneumonia in China, 2019**. *N Engl J Med* 2020, **382**(8):727-733.
106. Lu Mijia, Zhang Zijie, Xue Miaoge, Zhao Boxuan Simen, Harder Olivia, Li Anzhong, Liang Xueya, Gao Thomas Z, Xu Yunsheng, Zhou Jiyong *et al*: **N(6)-methyladenosine modification enables viral RNA to escape recognition by RNA sensor RIG-I**. *Nat Microbiol* 2020, **5**(4):584-598.
107. Zhou Yuan, Zeng Pan, Li Yan-Hui, Zhang Ziding, Cui Qinghua: **SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features**. *Nucleic Acids Res* 2016, **44**(10):e91.

108. Tortorici M Alejandra, Veesler David: **Structural insights into coronavirus entry**. *Advances in virus research* 2019, **105**:93-116.
109. Edgar Robert C: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.
110. Ovcharenko Anna, Rentmeister Andrea: **Emerging approaches for detection of methylation sites in RNA**. *Open Biol* 2018, **8**(9).
111. Roundtree Ian A, Evans Molly E, Pan Tao, He Chuan: **Dynamic RNA Modifications in Gene Expression Regulation**. *Cell* 2017, **169**(7):1187-1200.
112. Novoa Eva Maria, Mason Christopher E, Mattick John S: **Charting the unknown epitranscriptome**. *Nat Rev Mol Cell Biol* 2017, **18**(6):339-340.
113. Li Xiaoyu, Xiong Xushen, Yi Chengqi: **Epitranscriptome sequencing technologies: decoding RNA modifications**. *Nat Methods* 2016, **14**(1):23-31.
114. van Dijk Erwin L, Jaszczyszyn Yan, Naquin Delphine, Thermes Claude: **The Third Revolution in Sequencing Technology**. *Trends Genet* 2018, **34**(9):666-681.
115. Liu Huanle, Begik Oguzhan, Lucas Morghan C, Ramirez Jose Miguel, Mason Christopher E, Wiener David, Schwartz Schraga, Mattick John S, Smith Martin A, Novoa Eva Maria: **Accurate detection of m(6)A RNA modifications in native RNA sequences**. *Nat Commun* 2019, **10**(1):4079.
116. Van Gestel T, Suykens J A K, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J: **Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel Fisher discriminant analysis**. *Neural Comput* 2002, **14**(5):1115-1147.