PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods

Song-Yao Zhang[1,2], Shao-Wu Zhang[1]*, Xiao-Nan Fan[1], Jia Meng[3], Yidong Chen[4], Shou-Jiang Gao[5], Yufei Huang[2,4]*

1 Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, China, **2** Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, Texas, United States of America, **3** Department of Biological Sciences, HRINU, SUERI, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China, **4** Department of Epidemiology and Biostatistics, University of Texas Health San Antonio, San Antonio, Texas, United States of America, **5** UPMC Hillman Cancer Center and Department of Microbiology and Molecular Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

* zhangsw@nwpu.edu.cn (SWZ); yufei.huang@utsa.edu (YH)

Citation: Zhang S-Y, Zhang S-W, Fan X-N, Meng J, Chen Y, Gao S-J, et al. (2019) Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. PLoS Comput Biol 15(1): e1006663. https://doi.org/10.1371/journal.pcbi.1006663

Editor: Andrey Rzhetsky, University of Chicago, UNITED STATES

Received: June 1, 2018

Accepted: November 21, 2018

Published: January 2, 2019

Data Availability Statement: The HEK293 miCLIP data were deposited in NCBI's Gene Expression Omnibus (GEO) under accession number GSE63753; the HEK293 MeRIP-Seq data were deposited in NCBI's Gene Expression Omnibus (GEO) under accession number GSE29714; the MOLM13 miCLIP data were deposited in NCBI's Gene Expression Omnibus (GEO) under accession number GSE98623; the MOLM13 MeRIP-Seq data were deposited in NCBI's Gene Expression Omnibus (GEO) under accession number GSE94613; all the human MeRIP-Seq Data comes

Abstract

N6-methyladenosine (m^6A) is the most abundant methylation, existing in >25% of human mRNAs. Exciting recent discoveries indicate the close involvement of m^6A in regulating many different aspects of mRNA metabolism and diseases like cancer. However, our current knowledge about how m^6A levels are controlled and whether and how regulation of m^6A levels of a specific gene can play a role in cancer and other diseases is mostly elusive. We propose in this paper a computational scheme for predicting m^6A-regulated genes and m^6A-associated disease, which includes Deep-m^6A, the first model for detecting condition-specific m^6A sites from MeRIP-Seq data with a single base resolution using deep learning and Hot-m^6A, a new network-based pipeline that prioritizes functional significant m^6A genes and its associated diseases using the Protein-Protein Interaction (PPI) and gene-disease heterogeneous networks. We applied Deep-m^6A and this pipeline to 75 MeRIP-seq human samples, which produced a compact set of 709 functionally significant m^6A-regulated genes and nine functionally enriched subnetworks. The functional enrichment analysis of these genes and networks reveal that m^6A targets key genes of many critical biological processes including transcription, cell organization and transport, and cell proliferation and cancer-related pathways such as Wnt pathway. The m^6A-associated disease analysis prioritized five significantly associated diseases including leukemia and renal cell carcinoma. These results demonstrate the power of our proposed computational scheme and provide new leads for understanding m^6A regulatory functions and its roles in diseases.

Author summary

The goal of this work is to identify functional significant m^6A-regulated genes and m^6A-associated diseases from analyzing an extensive collection of MeRIP-seq data. To achieve

from MeT-DB2 database (http://www.xjtlu.edu.cn/metdb2); the phenotype-gene relationship downloaded from OMIM database (https://www.omim.org/)

Funding: This work was supported by the National Natural Science Foundation of China (http://www.nsfc.gov.cn/ 61473232, 61873202, 31671373 and 91430111) awarded to SWZ and JM; and the National Institutes of Health (https://www.nih.gov/ R01GM113245) awarded to YH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

this, we first developed Deep-m^{6}A, a CNN model for single-base m^{6}A prediction. To our knowledge, this is the first condition-specific single-base m^{6}A site prediction model that combines mRNA sequence feature and MeRIP-Seq data. The 10-fold cross-validation and test on an independent dataset show that Deep-m^{6}A outperformed two sequence-based models. We applied Deep-m^{6}A followed by network-based analysis using HotNet2 and RWRH to 75 human MeRIP-Seq samples from various cells and tissue under different conditions to globally detect m^{6}A-regulated genes and further predict m^{6}A mediated functions and associated diseases. This is also to our knowledge the first attempt to predict m^{6}A functions and associated diseases using only computational methods in a global manner on a large number of human MeRIP-Seq samples. The predicted functions and diseases show considerable consistent with those reported in the literature, which demonstrated the power of our proposed pipeline to predict potential m^{6}A mediated functions and associated diseases.

Introduction

N^6-methyl-adenosine (m^{6}A) methylation is a paradigm-shifting research filled with exciting discoveries. Recent research has shown that m^{6}A exists in > 25% of mRNAs in mammalian cells [1, 2] and forms an important regulatory circuitry that controls many aspects of RNA metabolism [3–10]. Evidence of m^{6}A's involvement in cancer and other diseases [11–21] and its role in regulating viral life cycle [22–25] are also accumulating. However, our current knowledge about how m^{6}A levels are regulated and whether and how regulation of m^{6}A levels of a specific gene can play a role in cancer and other diseases is largely elusive.

The purpose of this study is to conduct a comprehensive prediction of m^{6}A mediated functions and associated diseases through global analysis of m^{6}A regulated genes using 75 human methylated RNA immunoprecipitation sequencing (MeRIP-seq) [1, 2] samples curated by MeT-DB2 [26]. To this end, prediction of context-specific m^{6}A sites is an essential first step. Several informatics tools have been developed to predict condition independent m^{6}A sites from RNA sequences [27–35] or condition-specific m^{6}A peaks in MeRIP-seq data [36, 37]. Chen et al. [35] proposed the first sequence-based model iRNA-Methyl to predict m^{6}A site using features extracted from RNA sequences. Subsequently, Zhou et al. developed SRAMP [29] to improve the performance of predicting single-base m^{6}A sites using three kinds of features extracted from pri- and mature RNA sequences. Since then, a line of sequence-based algorithms [38–40], all based on different handcrafted features extracted from RNA sequences have been developed. As an alternative to handcrafted RNA sequence features, Wei et al. [33] proposed DeepM6APred to extract features automatically using a deep belief network (DBN). However, because RNA sequences are independent of any study conditions, none of these sequence-based models can predict condition-specific m^{6}A sites. Because m^{6}A has been shown to play different regulatory roles in different cell conditions and disease types, its methylation status is highly dynamic in nature, being either methylated or demethylated depending on the biological contexts. Also, m^{6}A status can be experimentally manipulated through silencing or overexpressing key m^{6}A related proteins, a commonly used approach for studying m^{6}A functions. Therefore, predicting condition-specific single-base m^{6}A site is an essential task in m^{6}A research. Currently, algorithms including exomePeak [41] and MeTPeak [42] are proposed to predict context-specific m^{6}A peaks from MeRIP-seq data. However, MeRIP-Seq has a limited resolution of ~100bp and its large biological and technical variations often result in

high false positive rates in the predicted peaks. No existing algorithm can predict condition-specific m^6A site at a single-base resolution.

In addition to a lack of single-base, context-specific site prediction algorithms, computational prediction of m^6A functions has not been adequately addressed. In our previous work [43], we developed m^6A-Driver, a network-based approach to identify m^6A driven genes with significant functions under a specific context. However, m^6A-Driver has several limitations. First, m^6A-Driver to identify m^6A driven genes in two different conditions, e.g., gene knock-down vs. normal. Therefore, m^6A-Driver cannot be used for the intended global analysis that includes samples from multiple conditions. Second, because of the small sample size, m^6A-Driver tends to identify a large number of significant genes, which makes it difficult to prioritize these genes. In this work, we propose to develop a new approach to address these limitations of the site or peak detection algorithms and m^6A-Driver so that a global prediction of m^6A mediated functions and associated diseases across samples from multiple conditions can be reliably performed.

To improve the resolution and accuracy of condition-specific m^6A site prediction from MeRIP-seq, we propose a novel Convolutional Neural Network (CNN) [44] based method named Deep-m^6A to detect m^6A sites from MeRIP-Seq peak regions with a single base resolution. This model integrates mRNA sequence information with MeRIP-seq data and trained on single-base m^6A sites identified by the miCLIP [45], the state-of-the-art high throughput technology for single-base detection of m^6A sites. Deep-m^6A makes it possible to identify single-base m^6A sites from the 75 human MeRIP-seq samples with higher accuracy and resolution, providing us a chance to investigate the relationship of m^6A methylation with gene expression and diseases in a global manner. To this end, we propose Hot-m^6A, a new network-based pipeline. We hypothesize that m^6A regulates biological processes and pathways by regulating a set of functionally interacted genes through regulating their mRNA stability. Therefore, the expression of a gene regulated by an m^6A site is likely to be correlated with the m^6A level across multiple conditions and this regulation also influences its neighboring genes in the PPI network. The stronger an m^6A regulates a gene, the more significant the m^6A-expression correlation and the higher the degree of influence on its neighbors. Our goal is to identify m^6A regulated genes, whose expressions are correlated with their m^6A methylations across 75 samples and which are closely interacting with other m^6A genes in a PPI network. To achieve this goal, we adopt the HotNet2 [46] algorithm. The basic idea of Hot-m^6A is to diffuse "heat", i.e., correlation of gene expression and m^6A level in this study, of genes in a network and select the significant genes, where the "heats" they diffused to each other are high. Therefore, the m^6A regulated genes identified by HotNet2 will in general have relatively high expression-methylation correlation and closely interacted with each other in the PPI network. However, HotNet2 still can also identify m^6A regulated genes with lower "heat" but closely interacted with "hot" genes. After m^6A regulated genes are identified, our pipeline prioritizes the m^6A associated diseases by applying the random walk with restart on heterogeneous network (RWRH) method [47] on a gene-disease heterogeneous network.

We applied Deep-m^6A and Hot-m^6A to the 75 MeRIP-seq samples, which produced a compact set of 709 functional significant m^6A-regulated genes and nine functional enriched sub-networks. The functional enrichment analysis of these genes and networks reveal that m^6A targets key genes of many critical biological processes including transcription, cell organization and transport, cell proliferation and cancer-related pathways such as the Wnt pathway. The m^6A-associated disease analysis prioritized five significantly associated diseases including leukemia and renal cell carcinoma.

Result

Overview of the method

The flowchart of our pipeline is illustrated in Fig 1. We intend to perform a global analysis of existing human m^6A site in different samples to uncover m^6A regulated functions and associated disease. To achieve this, we first set out to determine the single-base m^6A sites. To this end, we developed a novel CNN model, called Deep-m^6A, which takes both sequence feature and MeRIP-Seq IP reads count as an input to predict context specific single-base m^6A sites from MeRIP-Seq data. We then applied Deep-m^6A to predict single-base m^6A sites for all 75 human MeRIP-Seq samples extracted from MeT-DB V2.0 [26]. Then, we extracted the m^6A sites appeared in at least 12 samples according to a test based on Fisher's z transformation [48] and calculated the revised Fisher's z-transformation [49] of the Pearson correlation of methylation degree and gene expression level across all their appeared samples (see Methods and material for details). Genes that contain at least one m^6A site that occurred in more than 12 samples and whose methylation degree and gene expression level are significantly correlated were defined as candidate m^6A regulated genes. We selected the largest absolute z-transformed

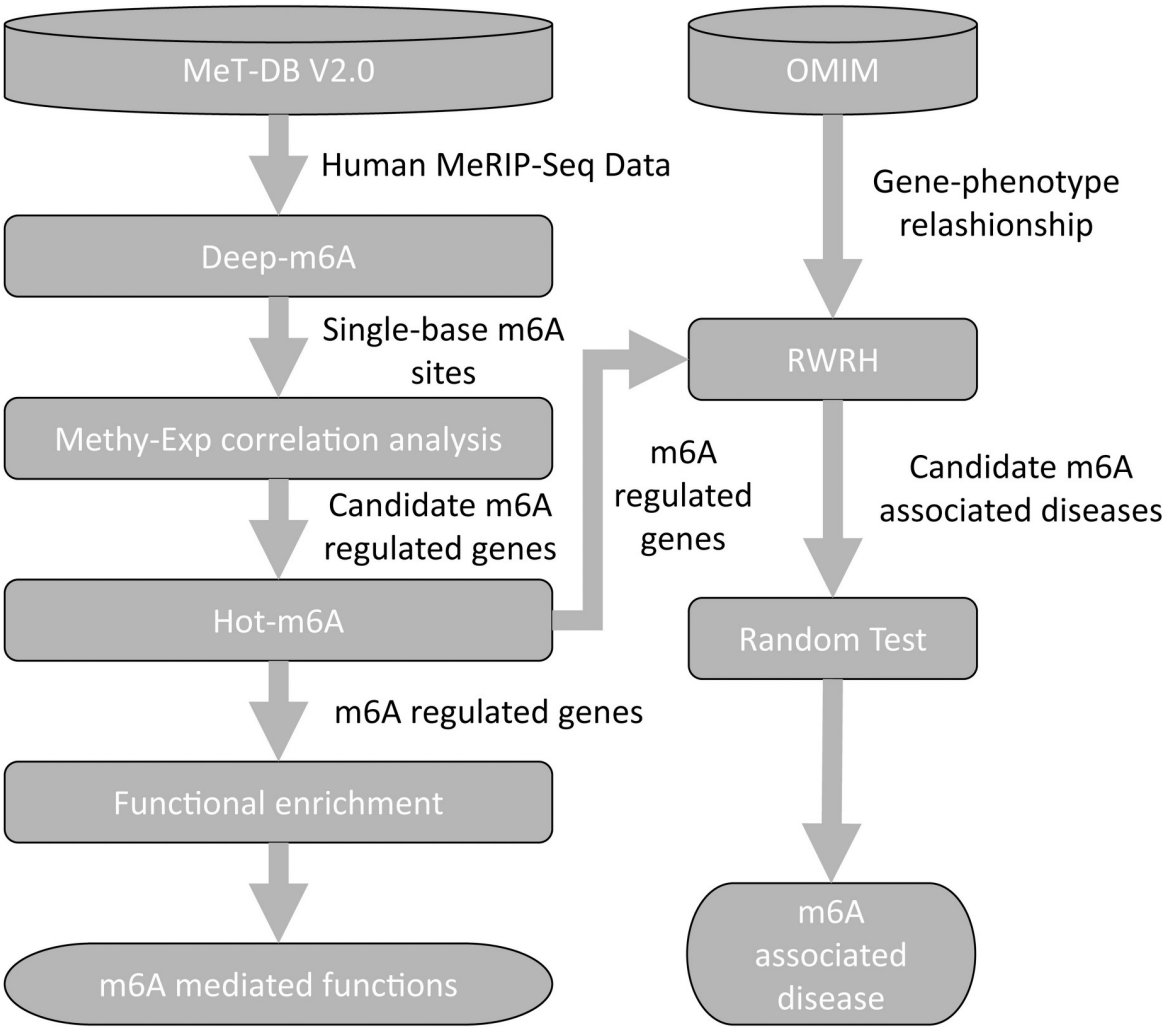

Fig 1. Flowchart of our proposed prediction pipeline.

https://doi.org/10.1371/journal.pcbi.1006663.g001

correlation from all m^{6}A sites in a gene to denote the methylation-expression correlation of the gene. To further study the m^{6}A regulated functions and select functional m^{6}A-regulated genes, we developed Hot-m^{6}A and applied it to 4 different PPI networks, taking the absolute z-transformed correlations for all candidate m^{6}A-regulated genes as the heat vector, to identify significant PPI subnetworks that are regulated by m^{6}A more than expected by chance. The 4 PPI networks includes BioGRID [50], HINT+HI2012 [51, 52], MultiNet [53], and iRefIndex [54]. The last three networks are used in the HotNet2 paper. The edges that are identified by all the four networks were extracted to form significant consensus subnetworks, which were then extended to include edges identified in 3, 2, and 1network, respectively (See Methods for the detail). All the genes in the extended subnetworks are defined as significant m^{6}A-regulated genes and genes reported in the significant consensus networks are defined as consensus m^{6}A-regulated genes. Finally, we took all these significant m^{6}A-regulated genes as gene seeds and their correlated diseases according the OMIM as disease seeds and applied RWRH on a heterogeneous gene-disease network. We also built four heterogeneous gene-disease networks corresponding to the four PPI networks. For each heterogeneous gene-disease network, the top 10 ranked diseases were selected as candidate m^{6}A-associated disease. A random network test was applied to calculate an empirical p-value for each candidate m^{6}A-associated disease. The candidate disease with a p-value < 0.05 was selected as significant candidate m^{6}A-associated disease. The significant candidate m^{6}A-associated disease that identified by all the 4 networks were finally identified as m^{6}A-associated disease.

Deep-m^{6}A achieves higher precision and sensitivity in 10-fold cross validation

We first evaluated the performance of Deep-m^{6}A using the training data from the HEK293 cell lines and compared its performance with Deep-m^{6}A-S. As described in Methods, Deep-m^{6}A-S and Deep-m^{6}A use the same CNN structure but different input features. Deep-m^{6}A-S takes 101-nt long sequences centered at a DRACH motif as input, whereas Deep-m^{6}A takes both sequence and the corresponding IP reads count as input. As detailed in the "Dataset" section, this training data include 4,742 positive samples that are CITS miCLIP m^{6}A sites in the MeRIP-seq peak regions and also centered at DRACH motif, and 33,718 negative samples that are also centered at a DRACH motif in the MeRIP-seq peak regions but are $>$ 50-nt away from the positive samples and not CIMS miCLIP sites or single-base m^{6}A sites reported in other experiments [55]. Before model comparison, we investigated the difference between MeRIP-seq IP reads coverage in the sequence regions of the positive and negative samples. As shown in Fig 2(A), the average reads count of positive samples are higher and more centered at the DRACH m^{6}A motif than those of the negative samples.

The CNN architecture of Deep-m^{6}A and Deep-m^{6}A-S includes one convolutional layer followed by a max-pooling layer, a fully connected layer and an output layer. We used 10-fold cross validation to evaluate the performance of Deep-m^{6}A and Deep-m^{6}A-S and took one of the 10 CVs to optimize the hyperparameters. To reduce the impact of significant imbalance between the positive and negative samples, we split the negative samples into 7 subsets, each of which has the equal size to positive samples and for each CV, we trained 7 models for each balanced pair of positive/negative samples and used the average predicted score of the 7 models as the final predicted probability of a test DRACH motif to be a single base m^{6}A methylation site. The CNN architecture was determined by a grid search method [56], where the searched parameters are the kernel size: 4x3, 4x4 and 4x5; # of kernels: 16, 32 and 48; the max-pooling size: 1x3, 1x4 and 1x5; # of the nodes of the fully connected layer: 8, 12, 16 and 32; the fully connected layer dropout rate: 0.2, 0.25, 0.3, and 0.5; and the softmax output layer dropout rate:

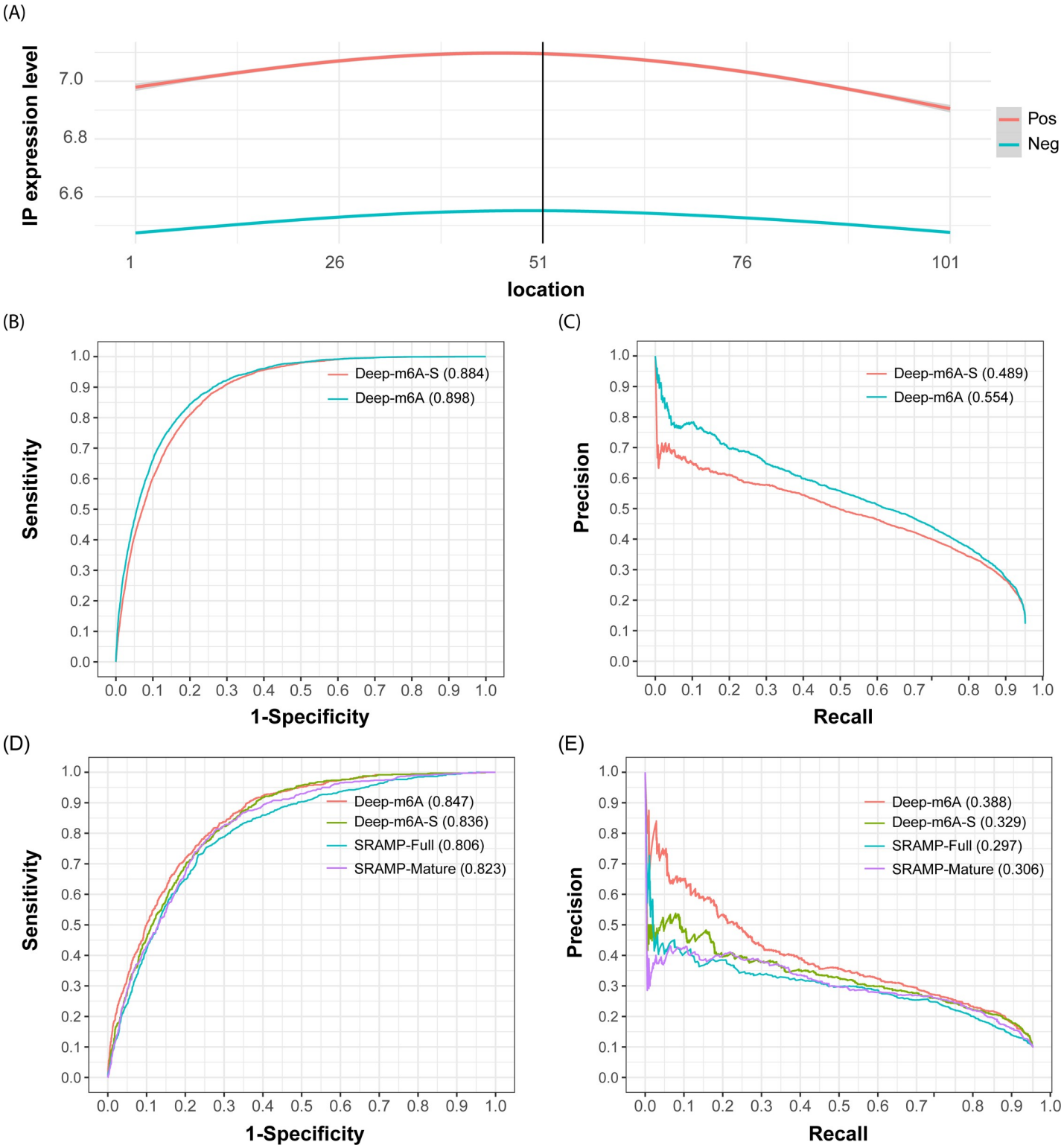

Fig 2. Performance of Deep-m^{6}A. (A) shows the IP reads count coverage in the 101-nt positive and negative training samples. The x-axis denotes the relative location of the nucleotides in the sample sequences, where the 51st position is the center A of the DRACH motif and the 1st position is the 5' end of the sequence. The IP expression level represents the IP reads input RC_{norm}. (B) and (C) are the ROC and PR curves of Deep-m^{6}A-S and Deep-m^{6}A obtained from the 10-fold CV on the HEK293 training data. (D) and (E) are performances of Deep-m^{6}A, Deep-m^{6}A-S, SRAMP-Mature and SRAMP-Full model on the independent MOLM13 data. The number after the method names are corresponding area under the curve (AUC).

https://doi.org/10.1371/journal.pcbi.1006663.g002

Table 1. Optimized hyperparameters of Deep-m^{6}A and Deep-m^{6}A-S.

Hyperparameters	Kernel size	# Filters	Max-pooling size	Dropout rate1	Node of dense	Dropout rate2
Value	4x5	32	1 x 4	0.25	12	0.25

* "Kernel size" is size of CNN kernel, "# Filters" is number of CNN filters, "Dropout rate1" is the dropout rate after max-pooling, "Node of dense" is the number of nodes in the fully connected layer after convolution layers, and "Dropout rate2" is the dropout rate after the fully connected layer.

https://doi.org/10.1371/journal.pcbi.1006663.t001

0.2, 0.25, 0.3 and 0.5. Table 1 showed the optimized hyperparameters by grid search. The categorical cross entropy loss function and Adadelta [57] optimizer were adopted in the training.

Fig 2(B) and 2(C) shows the receiver operating characteristic (ROC) curves and the precision-recall (PR) curves of the CV test. We have trained seven CNN models to balance the scale of the positive and negative training samples and the predicted probability of a site is an average predicted probably of these seven models. We can see Deep-m^{6}A achieved a higher area under the ROC curve (AUC = 0.898) and area under the PR curve (PRAUC = 0.554) than Deep-m^{6}A-S model (AUC = 0.884, PRAUC = 0.489). The AUC and PRAUC of Deep-m^{6}A are 1.4% and 5.6% higher than Deep-m^{6}A-S, respectively. There is especially a higher improvement in precision. This suggests that including IP reads help lower the false positive rate, especially in the higher ranked predictions. Having a higher precision is particularly important for subsequent biological validation and functional study because in practice attention is most likely given to a limited top ranked predictions.

Deep-m^{6}A outperforms sequence-based predictions on an independent dataset

To further validate the performance of Deep-m^{6}A, we applied it along with Deep-m^{6}A-S on the independent MOLM13 dataset and compared their performance with SRAMP [29]. As described in the "Dataset" section, this dataset includes 726 positive and 6,577 negative samples. Deep-m^{6}A and Deep-m^{6}A-S were trained on HEK293 dataset as described in the previous section. For SRAMP, we downloaded the SRAMP tool from their webserver (http://www.cuilab.cn/sramp/) and applied it to the 101-nt RNA (without intron) and DNA (with intron) sequences of each sample. The RNA sequences are used by the SRAMP mature RNA model and the DNA sequences are used by SRAMP full DNA model. When feeding these sequences into the SRAMP tools, they output probabilities of all DRACH motifs contained in the sequence. We only picked the results for the centered motif of each sample as the prediction probabilities for SRAMP. Fig 2(D) and 2(E) shows the ROC curve and PR curve results of these models. First, among the three sequence-based models, Deep-m^{6}A-S reports about 1% improvement in AUC and 2% in PRAUC over the two SRAMP models. This suggests that the CNN model can capture additional discriminate sequence features than SRAMP. Second, Deep-m^{6}A outperforms all three sequence-based prediction models in both AUC and PRAUC. The improvement in precision is even more pronounced (~6% over Deep-m^{6}A-S). This once again speaks the benefit of including IP reads in the prediction.

Identification of candidate m^{6}A-regulated genes from all *human* MeRIP-Seq data

The relatively large number of human MeRIP-Seq data of different cells and tissues under different conditions curated by MeT-DB2 gives us a chance to investigate the relationship of m^{6}A methylation with gene expression and diseases in a global manner. As the first step of this global investigation, we applied Deep-m^{6}A to the 75 human MeRIP-seq samples to identify

candidate m^{6}A-regulated genes. A positive site was predicted when the prediction probability calculated by Deep-m^{6}A is greater than 0.907; this threshold is chosen because at this threshold the 10-fold CV test in training can achieve a precision of 0.7. This resulted in 23,456 single-base m^{6}A sites in all 75 samples. Fig 3(A) shows the frequency of a site being predicted in the 75 samples. As shown, ~23% of these sites are sample specific, i.e. they only appear in one sample. In contrast, ~30% sites appear in more than 12 data samples.

To conduct a global analysis across all the 75 samples, we extracted sites that appear in more than 12 samples. The distributions of these sites on mRNA as well as all predicted m^{6}A sites were tally using Guitar R package [58] and shown in Fig 3(B). As illustrated, for all the predicted sites, the distribution tends to be enriched in 3' UTR and 5' UTR, whereas the sites appeared in more than 12 samples are more enriched around stop codon and in 3' UTR. Because 3'UTR contain binding sites of miRNA and RNA binding proteins such as HuR that are known to post-transcriptionally regulate gene expression, this distribution may indicate that the sites that appeared in > = 12 samples may potentially be involved in regulating gene expression. Next, we selected the genes which harbor m^{6}A sites that appear in > = 12 samples

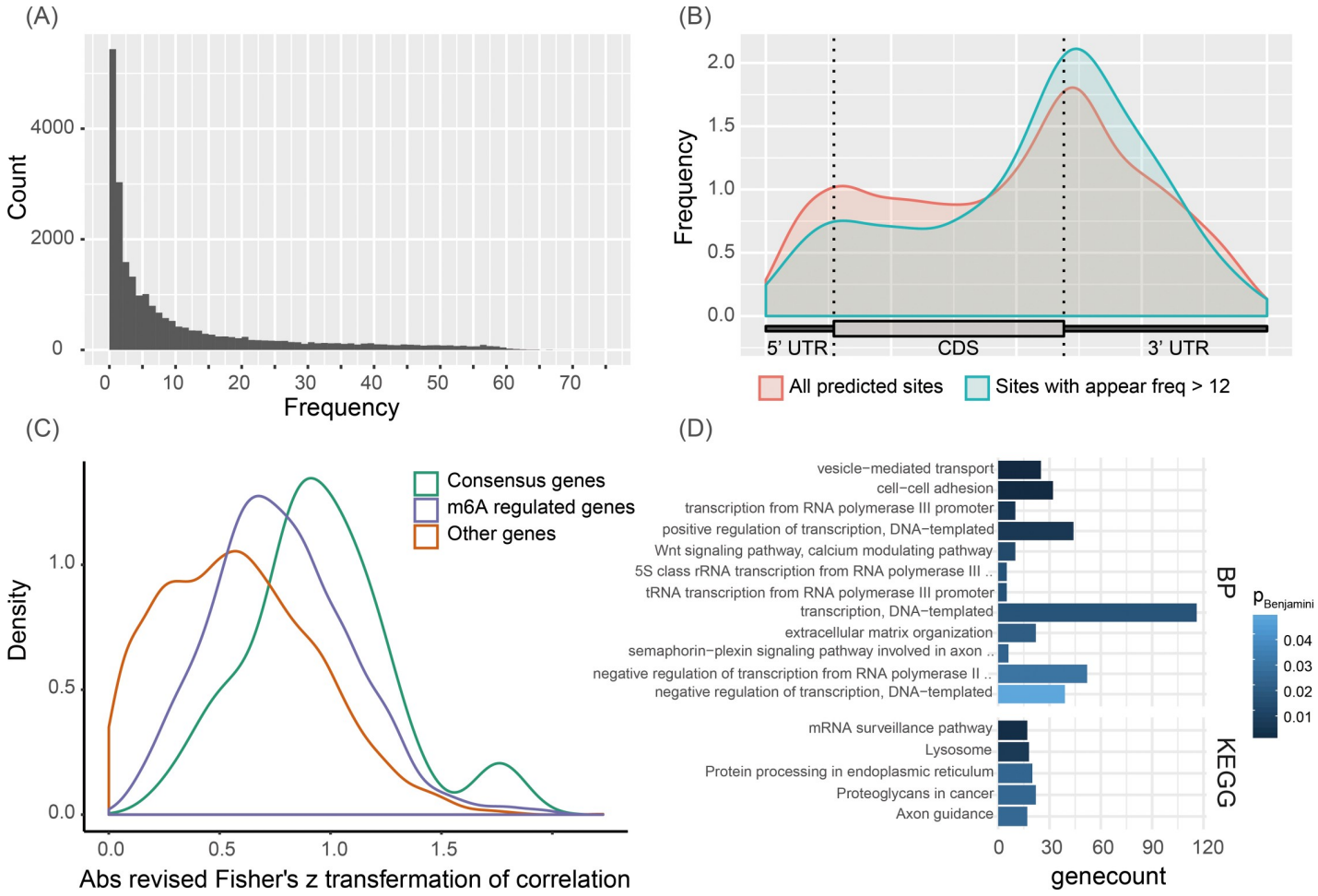

Fig 3. Analysis of predicted m^{6}A sites in 75 human samples. (A) Distribution of the number of occurrence of a predicted single-base m^{6}A site in the 75 samples. (B) Distributions of all predicted m^{6}A sites and sites appeared in more than 12 samples in a meta-mRNA. (C) Distributions of the absolute revised Fisher's z transformed Pearson correlations of the 49 consensus m^{6}A-regulated genes, 709 m^{6}A-regulated genes identified by HotNet2, and other remaining candidate genes. (D) GO BP and KEGG pathway enrichment result for all 709 m^{6}A-regulated genes. Gene count means the number of genes involved in the corresponding terms and $P_{Benjamini}$ is the adjusted FDR of the enrichment p-value.

https://doi.org/10.1371/journal.pcbi.1006663.g003

as candidate m^{6}A-regulated genes. In total, 3,670 m^{6}A-regulated genes that contain totally 7,090 m^{6}A sites were extracted. To study the relationship of their expression and m^{6}A methylation, we calculated the Pearson's correlation of methylation degree and gene expression level for all the candidate m^{6}A-regulated genes and then converted it to a revised Fisher's z score as described in "Methods and material" section.

Functional significant m^{6}A-regulated genes identified by Hot-m^{6}A are involved in important functional pathways

We then set out to identify functional significant m^{6}A-regulated genes, which are defined as candidate genes whose expression levels are influenced by the m^{6}A methylation more than expected by chance and which are closely interacting with each other in PPI networks. We applied Hot-m^{6}A (See Methods and material for detailed algorithms) to the candidate m^{6}A-regulated genes, taking their absolute revised Fisher's z score as the heat vector and performing heat diffusion in the PPI network to identify functional interacted genes with relative high correlation. 49 consensus m^{6}A-regulated genes were identified across all four PPI networks using Hot-m^{6}A and 709 m^{6}A-regulated genes were identified in the extended network (S1 Fig; see Methods and material for detail). Most of the 709 m^{6}A-regulated genes (681 or ~96%) are in the biggest subnetwork. We then compared the distributions of the absolute revised correlation z scores of the 49 consensus m^{6}A-regulated genes, 709 m^{6}A-regulated genes, and remaining non-significant candidate genes (Fig 3(C)). We can see that most of the m^{6}A-regulated genes have larger correlation than other nonsignificant candidate genes and the consensus m^{6}A-regulated genes tend to the largest correlations among the three groups. This result speaks for the ability of our pipeline to identify likely m6A regulated genes. Notice there are some m^{6}A-regulated genes with low correlations; they are identified by Hot-m^{6}A because they interacted tightly with "hot" genes with high correlation in the PPI network. To further detect the function module from these 709 m^{6}A-regulated genes, we removed edges that are identified only in one PPI network, and obtained 22 subnetworks including 113 genes in total. We further isolated the 9 largest functional significant subnetworks that have at least 3 genes (S2 Fig).

We further examined the functional enrichment for all the 709 significant genes using DAVID [59] (Fig 3(D)). Among the twelve enriched GO BPs ($P_{Benjamini} < 0.05$), seven are directly related to transcription including transcription from RNA polymerase III promoter (10 genes, $P_{Benjamini} = 4.78\times10^{-3}$), positive regulation of DNA-templated transcription (44 genes, $P_{Benjamini} = 5.79\times10^{-3}$), transcription, DNA-templated (116 genes, $P_{Benjamini} = 1.67\times10^{-2}$), tRNA transcription from RNA polymerase III promoter (5 genes, $P_{Benjamini} = 1.62\times10^{-2}$), 5S class rRNA transcription from RNA polymerase III type 1 promoter (5 genes, $P_{Benjamini} = 1.62\times10^{-2}$), negative regulation of transcription from RNA polymerase II promoter (52 genes, $P_{Benjamini} = 2.92\times10^{-2}$) and negative regulation of transcription, DNA-templated (39 genes, $P_{Benjamini} = 4.82\times10^{-2}$). Indeed, the m^{6}A is shown to involve in every stage of RNA metabolism including transcription. It is shown to regulate mRNA stability and speculated to also regulate mRNA splicing [60]. Also, we also see that m^{6}A genes could regulate cell organization and transport as they are enriched in vesicle-mediated transport (25 genes, $P_{Benjamini} = 3.83\times10^{-5}$), cell–cell adhesion (32 genes, $P_{Benjamini} = 3.61\times10^{-4}$) and extracellular matrix organization (22 genes, $P_{Benjamini} = 2.04\times10^{-2}$) are also enriched. We also noticed that that the Wnt signaling pathway is enriched (10 genes, $P_{Benjamini} = 1.26\times10^{-2}$). Wnt is an important pathway widely involved in cancer and cell development [61–63]. The potential involvement of m^{6}A in cancer is reinforced by the enriched proteoglycans in cancer KEEG pathway (22 genes, $P_{Benjamini} = 2.43\times10^{-2}$). This is not surprising as there is increasing evidence demonstrating its regulatory roles in different cancer [20, 64–70]. Another enriched KEGG pathways are Protein

processing in the endoplasmic reticulum (20 genes, $P_{Benjamini} = 2.37\times10^{-2}$) and Lysosome pathway (18 genes, $P_{Benjamini} = 6.31\times10^{-3}$), both of which are protein processing pathways. These predictions are corroborated by our knowledge of the m^{6}A's role in regulating translational efficiency [71, 72].

We next performed GO and KEGG pathway enrichment analysis to the nine largest significant subnetworks (S3 and S4 Figs). These enriched BP and pathways present reasonably clear and consistent interpretations of these subnetworks. Subnetwork A is mostly involved in protein processing (6 genes enriched in Protein processing in endoplasmic reticulum, 6 genes enriched in ER-associated ubiquitin-dependent protein catabolic process, 3 genes enriched in retrograde protein transport, ER to cytosol and 4 gene enriched in protein stabilization) and potentially regulate mRNA stability via SMG9, UPF1, SMG8 and SMG1 genes enriched in nuclear-transcribed mRNA catabolic process, nonsense-mediated decay BP term. Subnetwork B is predicted to mostly involve in Notch signaling pathway via NOTCH2, MAML1 and MAML3 genes. Subnetwork C is closely related to cell motility, proliferation and survival through enrichment of m^{6}A-regulated genes including VEGFB, PDGFB, VEGFA, COL5A1, NRP1, SLIT2 and SPARC in pathways like Focal adhesion ($P_{Benjamini} = 1.75\times10^{-2}$), positive regulation of endothelial cell proliferation ($P_{Benjamini} = 1.56\times10^{-3}$), cell migration involved in sprouting angiogenesis ($P_{Benjamini} = 1.26\times10^{-3}$) and positive regulation of endothelial cell migration ($P_{Benjamini} = 1.26\times10^{-3}$). It is also involved in neuronal development related pathways including Axon guidance (NRP1, PLXNA1, SEMA3F and SLIT2 genes with $P_{Benjamini} = 1.30\times10^{-3}$), branchiomotor neuron axon guidance (NRP1, PLXNA1 and SEMA3F genes with $P_{Benjamini} = 8.54\times10^{-3}$), semaphorin-plexin signaling pathway involved in axon guidance (NRP1, PLXNA1 and SEMA3F genes with $P_{Benjamini} = 1.19\times10^{-3}$) and axon extension involved in axon guidance (NRP1, SEMA3F and SLIT2 genes with $P_{Benjamini} = 1.19\times10^{-3}$). The function of subnetwork D is mostly related to transcription including regulation of transcription from RNA polymerase II promoter (MED26, MED18, MED9, MED11 and MED21 genes with $P_{Benjamini} = 4.16\times10^{-4}$), transcription, DNA-templated (MED29, MED26, POLR2I, MED9, MED11 and MED21 genes with $P_{Benjamini} = 4.92\times10^{-3}$) and transcription initiation from RNA polymerase II promoter (MED26, POLR2I and MED13 genes with $P_{Benjamini} = 1.48\times10^{-2}$). Subnetwork E's function is defined through mostly the enrichment Wnt signaling pathway (5 genes enriched including FZD8, DKK1, LRP6, FZD5, LRP5 with $P_{Benjamini} = 6.75\times10^{-6}$). For subnetwork F, 3 m^{6}A-regulated genes are STK11, WDR6 and STRADA and they are involved in cell cycle and cell proliferation related pathways including cell cycle arrest, mTOR signaling pathway and AMPK signaling pathway with $P_{Benjamini} < 0.05$. Subnetwork G serves a role in antigen processing (3 genes, TAP2, HLA-E, TAPBP enriched in Antigen processing and presentation, antigen processing and presentation of peptide antigen via MHC class I and antigen processing and presentation of endogenous peptide antigen via MHC class I with $P_{Benjamini} < 0.05$). Subnetwork H includes ITGAV, ITGB5 and CYR61 as m^{6}A regulated genes, which are involved in cell adhesion and subnetwork I are involved in MAPK signaling pathway via gene MAP3K4, GADD45B and MAP2K7. Taken together, these results suggest that m^{6}A-regulated genes are enriched in significant biological process and pathways that can influence RNA transcription, cell motility, cell proliferation, cell survival and cell death, and therefore are likely involved in caner and cancer related pathways. Moreover, m^{6}A regulated genes tend to be in the upstream of these enriched pathways (S5 Fig).

Prioritized m^{6}A-associated diseases

We next investigated the association of m^{6}A with diseases. We first searched the OMIM database for diseases related to m^{6}A-regulated genes, where we found 308 phenotypes associated with 177 m^{6}A-regulated genes according to OMIM phenotype-disease relationship record (S1 File). Most

of the selected phenotypes are associated with only one m^{6}A-regulated gene. To prioritize these diseases to identify significant m^{6}A-associated diseases, we mapped the 709 m^{6}A-regulated genes to each of the four PPI networks and the 308 phenotypes to the disease network and constructed four gene-disease networks. We then applied RWRH to each of the four networks (See "Methods and material" for the detailed algorithm), taking the regulated genes and their correlated diseases as seed nodes. Then, we selected the top 10 ranked diseases as candidate m^{6}A-associated diseases and calculated an empirical *p*-value to assess if a disease association is selected by chance. Five significant m^{6}A-associated diseases with p-value < 0.05 and reported by all the four networks were finally identified as the m^{6}A-associated diseases (Table 2).

As shown in Table 2, acute myeloid leukemia is prioritized as the most significant m^{6}A associated disease, where m^{6}A-regulated genes NSD1, PICALM, and ABL2 are OMIM-annotated disease genes, suggesting that they may be potential m^{6}A-associated biomarkers in leukemia. Several lines of evidence have shown the direct involvement of m^{6}A in regulating leukemia [68–70]. The second significant disease is type 2 diabetes mellitus. Yang et al. has established this association and reported that glucose is involved in the dynamic regulation of m^{6}A in patients with type 2 diabetes [73]. Our prediction identified the m^{6}A genes HNF1B, HNF1A, WFS1, and IRS2 as the potential disease genes, which could provide a new clue to study the role of m^{6}A in type 2 diabetes. Also, m^{6}A is predicted to have a role in renal cell carcinoma. This is also corroborated by Xiao et al. [74], which reported that the m^{6}A methyltransferase METTL3 acted as a tumor suppressor in renal cell carcinoma and publications in [75, 76], which identified YTHDF2, an m^{6}A binding protein, to be involved in renal cell carcinoma. Taken together, we found existing evidence to support 3 out of 5 predicted association diseases. These results demonstrate the power of the proposed network-based analysis and the RWRH algorithm in identifying m^{6}A associated diseases.

Discussion

The accumulation of a large number of MeRIP-Seq samples from different cells and tissues under different conditions gives us a chance to analyze m^{6}A-regulated genes and m^{6}A-associated functions in a global manner. However, existing informatics tools for predicting m^{6}A sites from MeRIP-seq are hampered by high false positive rates and low resolutions. To address this issue, we developed Deep-m^{6}A, the first CNN model that predicts single-base m^{6}A sites in MeRIP-Seq peak regions by integrating mRNA sequence features with MeRIP-Seq IP reads. Test results from 10-fold CV on training HEK293 data and an independent OMLM13 dataset showed that Deep-m^{6}A outperformed sequence-based algorithms including Deep-m^{6}A-S and SRAMP in both precision and sensitivity. Although the miCLIP technology has been proposed to profile transcriptome-wide m^{6}A at a single-base resolution, its adoption is still very limited because of its more complex protocol. Therefore, MeRIP-seq will continue to serve as the go-to high throughput technology for global m^{6}A profiling in the near future. Give the ability of Deep-m^{6}A to provide MeRIP-seq a single-base detection resolution, we expect Deep-m^{6}A to

Table 2. Predicted m^{6}A-associated diseases.

Disease	OMIM annotated m^{6}A genes	Exp-m^{6}A corr (# samples)
acute myeloid leukemia	NSD1, PICALM, ABL2	0.76 (52), 0.77 (28), 0.89 (27)
type 2 diabetes mellitus	HNF1B, HNF1A, WFS1, IRS2	0.73 (14), 0.89 (13), 0.51 (14), 0.60 (23)
microphthalmia	BCOR, SOX2	0.67 (13), 0.34 (26)
MHC class I deficiency	TAPBP, TAP2	0.81 (24), 0.46 (36)
renal cell carcinoma	HNF1A, OGG1, VHL, HNF1B	0.89 (13), 0.41 (14), 0.83 (16), 0.73 (14)

https://doi.org/10.1371/journal.pcbi.1006663.t002

be an important tool in m^{6}A research. Currently, Deep-m^{6}A is trained to detect sites that reside on DRACH motifs. Even though this requirement of containing motifs helps reduce the false positive predictions, it also sacrifices the prediction sensitivity and will inevitably miss the positive sites that do not have any motifs. Therefore, further improvement of Deep-m^{6}A in the future to be able to detect sites without motifs will provide additional value for Deep-m^{6}A.

In our scheme, to further reduce the false positive rate and prioritize functional significant m^{6}A-regulated genes, we examined the correlation between expression and m^{6}A methylation of m^{6}A-regulated genes and assessed their function significance using Hot-m^{6}A implemented on the PPI network. We applied Deep-m^{6}A and Hot-m^{6}A to 75 human MeRIP-seq data, which resulted in a compact collection of 709 m^{6}A-regualted genes and several interacting sub-networks of m^{6}A-regualted genes. Functional analysis revealed that these genes are mainly involved in transcription and Wnt pathway (Fig 2(D)). Wnt pathway is one of the key cascades regulating cell migration and cell development, and is also tightly associated with cancer such as glioblastoma (GBM). Even though direct involvement of m^{6}A in Wnt pathway has not been reported, YTHDF2, an m^{6}A reader, has been shown to suppress cancer cell migration by inhibiting EMT in an m^{6}A dependent manner [77]. Also, ALKBH5, an m^{6}A demethylase, is reported to promote GBM tumorigenesis by stabilizing nascent FOXM1 transcripts through mediating its m^{6}A levels [78] and FOXM1 has also been show to control Wnt target gene expression in GBM. It is highly likely that there are alternative pathways such as Wnt, by which m^{6}A regulates cell migration and tumorigenesis. Functional enrichment of subnetworks also presents a highly consistent interpretations of their functions (S3 and S4 Figs). This suggests that the m^{6}A-regulated genes in the subnetworks are likely to involve in the same processes and pathways and thus share similar functions. In terms of their enriched processes and functions, we observed again important pathways such as focal adhesion, mTOR signaling pathway and AMPK signaling pathway, which are also known to regulate cell cycle, and cell migration, and are critical in cancer.

Finally, the network-based disease association analysis on the m^{6}A-regulated genes reported 5 significant associated diseases, where 3 of them have been corroborated by the existing publications. Our predictions could provide clues for the mechanisms, by which m^{6}A regulating these diseases. While there is no published evidence to support the other two associated diseases, our prediction points to potentially new disease associations of m^{6}A.

Last but not the least, our proposed methods have several issues that need to be further addressed in the future. First, because of the lack of miCLIP data, the model has only been trained on data from HEK293 cell line; this may not be enough to capture the features of all different kinds of true positive m^{6}A sites. Second, the scale of phenotype-gene relationships in the OMIM database are relative small, which might not be able to capture all potential significant disease-gene correlations. A potential solution is to integrate other disease-gene annotation information like DisGeNET [79] to make the network more complete. Finally, we only analyzed the common m^{6}A sites across many samples and their influence on gene expression. However, some of these context specific m^{6}A sites appear only in a unique sample and they could be important for understanding m^{6}A functions under this specific condition. Developing algorithms to detect these unique genes would be another future work.

Methods and materials

Datasets

The positive single-base m^{6}A sites for our training data for Deep-m^{6}A and Deep-m^{6}A-S were obtained from [45], which developed m^{6}A individual-nucleotide-resolution cross-linking and immunoprecipitation (miCLIP) technology. This paper includes two alternative technologies

including cross-linking–induced mutation sites (CIMSs) and cross-linking–induced truncation sites (CITSs) for human embryonic kidney (HEK293) cells using total cellular RNA. The CIMS miCLIP uses Abcam as antibody and C→T transitions as feature, whereas the CITS miCLIP uses SySy as antibody and truncations as feature. We chose CITS miCLIP generated single base m^6A sites as true positive m^6A sites because the corresponding MeRIP-Seq data of the HEK293 cell line that we used for training were also generated with SySy as antibody [77]. Another reason is that 74% (4847/6543) of CITS sites can be mapped to peaks generated by MeRIP-Seq data, whereas this ratio is only 55% (5202/9536) for CIMS sites. The MeRIP-Seq peaks were detected using the exomePeak R package [41] from 2 replicates of MeRIP-Seq samples [77]. A consistent peak that appears in both replicates and also contains at least one CITS miCLIP site was determined as a single-base m6A-containing MeRIP-Seq peak. The mRNA sequences of these peaks (without introns) were subsequently extracted and any sites in the sequences that contain DRACH motifs were defined as candidate m^6A sites. A candidate m^6A site was then extended to 101 nt centered at the "A" of the DRACH motif to capture the sequence and reads count features around the motif. The candidate sites that also are CITS miCLIP m^6A sites were determined as the positive samples (4,742 in total) and other candidate sites that are at least 50-nt away from a positive "A" and are not any CIMS miCLIP sites or single base m^6A sites reported in other experiments [55] were defined as the negative samples (33,718 in total).

The independent miCLIP test data were obtained from [69], which has 3 miCLIP replicates for the leukemia cell line MOLM13, each of which contains 8,113, 2,886 and 2,050 miCLIP sites, respectively. There were in total 11,746 miCLIP m^6A sites after combining all the three replicates, whereas only 136 sites were common across all the three replicates. Here, we considered miCLIP m^6A sites that appeared in at least two replicates as true positive m^6A sites (1,147 in total). The corresponding MeRIP-Seq test data were obtained from [80], which included 2 MeRIP-Seq replicates for the MOLM13 cell line. One of the 2 replicates has very low sequencing depth, so we took the combined peaks of these two replicates reported by exomePeak as MeRIP-seq peaks. Similar as the training data, the MeRIP-Seq peaks that contain at least one miCLIP m^6A sites were determined as single-base m^6A-containing MeRIP-Seq peaks and the sites with DRACH motifs in these regions were defined as candidate single-base m^6A sites. The candidate m^6A sites that are also miCLIP m^6A sites were determined as positive samples (726 in total) and other candidate sites that are >50-nt away from any positive sites and are not miCLIP sites from any of the three miCLIP replicates were determined as negative samples (6,577 in total).

The human MeRIP-Seq data were downloaded from MeT-DB V2.0 [26], which include 75 human samples from different human cell lines and tissues under different conditions, including 9 cell lines (A549, Dendritic cells, embryonic stem cell, Endoderm, HEK293T, HeLa, HepG2, neural progenitor cells, OKMS inducible fibroblasts and U2OS) and brain tissue samples under different conditions. The reference PPI networks were built based on BioGRID (release 3.4.128) [50], HINT+HI2012 [51, 52], MultiNet [53], and iRefIndex [54]. After removing the isolated proteins and self-interaction proteins, we established a PPI network with a total of 16,062 proteins and 152,676 interactions. The last three PPI networks were downloaded from http://compbio.cs.brown.edu/pancancer/hotnet2/. The HINT+HI2012 network contains 9,858 genes and 40,704 edges; the iRefIndex network contains 12,128 genes and 91,808 edges; and the Multinet network contains 14,398 genes and 109,569 edges. Diseases ontology terms were collected from the Disease ontology [81]. The "doSim" function of R package "DOSE" [82] was used to calculate the semantic similarity between two DO terms, which are used as the edge weight of the disease network. The disease gene relationship information was extracted from the OMIM database [83]. The OMIM ID was mapped to DO ID so that we can use OMIM gene–phenotype relationship to connect the PPI network and DO

disease network to construct a gene-disease network. Four gene-disease heterogeneous networks were built for each of the PPI networks.

Deep-m^{6}A and Deep-m^{6}A-S for single-base m^{6}A prediction

We developed 2 CNN models for single-base m^{6}A prediction, one called Deep-m^{6}A and the other was Deep-m^{6}A-S. These 2 models use the same CNN structure and hyperparameters but with different inputs. For Deep-m^{6}A-S, the input is the OneHot encoded 101nt RNA sequences centered at the "A" of a DRACH motif. OneHot encoding translates the A, U, C, G characters into a binary vector of (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1), respectively. Therefore, the input of Deep-m^{6}A-S becomes a 4 x 101 matrix, M_s. On the contrary, Deep-m^{6}A takes both RNA sequences and the features of MeRIP-Seq IP reads count at each nucleotide of the RNA sequence. However, the input for Deep-m^{6}A is similar to OneHot encoded M_s for Deep-m^{6}A-S but with 1s replaced by the IP reads count feature at that nucleotide. The IP reads count features were calculated by the same approach as exomePeak, i.e.,

$$RC_{norm} = \ln(RC/RC_{total} * 10^8) \quad (1)$$

where RC_{norm} is a 101-dimension vector of normalized reads count, RC is another 101-dimension vector of raw reads count, and RC_{total} is the total number of reads for the MeRIP-Seq IP sample. Because there are two replicates in each of the HEK293 and MOLM13 datasets, we calculated the average RC_{norm} of the two replicates as input IP reads count. The Deep-m^{6}A input is then calculated as

$$M_{sr} = M_s D_{RC_{norm}} \quad (2)$$

where $D_{RC_{norm}}$ is a 101x101 diagonal matrix with diagonal entries being those of RC_{norm}.

Deep-m^{6}A and Deep-m^{6}A-S use CNN [44] to capture the non-linear features of input sequences and IP reads count. The adopted CNN architecture consists of a convolutional, a max-pooling and two fully connected layers (Fig 4). The convolutional layer outputs the pointwise product between the input matrix (M_{sr} for Deep-m^{6}A and M_s for Deep-m^{6}A-S) and filters, which is followed by a rectified linear (ReLU) activation. Then, a max-pooling layer, which selects the maximum value over a window, is applied to reduce the dimensionality, which is followed by a dropout operation to reduce the complexity of the model. Finally, a fully connected dense layer is added followed by another ReLU and dropout to combine all the features learned by each filter. The output of the dense layer is passed on to the softmax function to generate the probability of the input sample to be an m^{6}A site.

Because the positive and negative sample sizes are imbalanced, we split the negative samples into seven subsets of equal size and trained 7 CNN models, each on a set that paired positive samples with a subset of negative samples. Seven models are trained as a result. For any prediction, the averaged prediction probability of the seven models is taken as the final predicted probability for an input sample. We implemented the CNN models using the keras R package. We refer Deep-m^{6}A and Deep-m^{6}A-S as the combination of the seven models. Input data files and Deep-m^{6}A source code and the models are made publicly available on GitHub (https://github.com/NWPU-903PR/Deepm6A.git). The inputs of the Deep-m^{6}A function are a MeRIP-Seq IP sample bam file and a bed file that annotates the peaks identified by exomePeak R package from MeRIP-Seq data. The output is an excel file, where each row contains information about a predicted single-base m^{6}A site extracted from an exomePeak peak region and each column denotes the chromatin, the chromatin start, the chromatin end, Entrez gene ID, the predicted probability, the strand and the motif at this location, respectively.

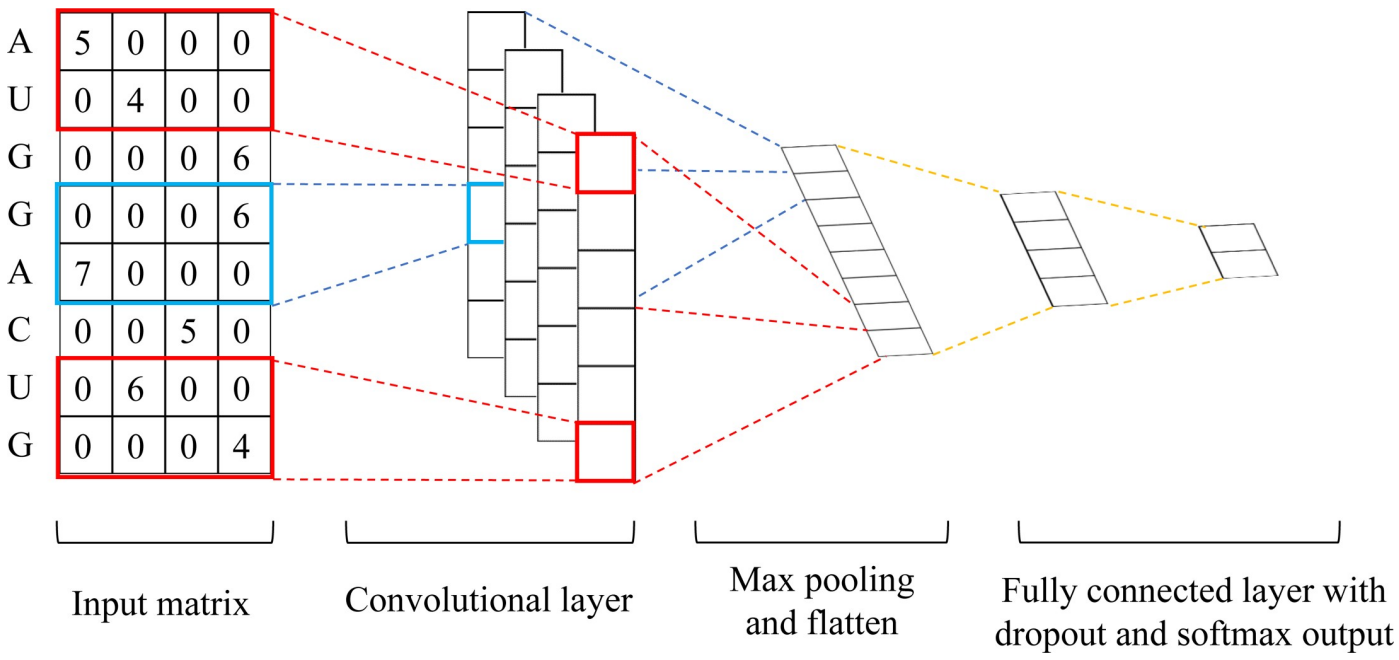

Fig 4. The architecture of the proposed CNN model for Deep-m^{6}A. The input matrix is M_{sr} for Deep-m^{6}A and M_s for Deep-m^{6}A-S.

https://doi.org/10.1371/journal.pcbi.1006663.g004

Prediction of single base m^{6}A peak for all *human* MeRIP-Seq data

We applied Deep-m^{6}A to predict single-base m^{6}A sites in 75 human MeRIP-Seq samples from MeT-DB2. First, we applied exomePeak to predict m^{6}A peaks in each MeRIP-Seq sample and then searched for DRACH motifs in the peak regions. "A"s in these motifs were treated as candidate single-base m^{6}A sites and the 101nt RNA sequences centered at these "A"s and the corresponding IP reads counts were extracted to construct the input matrix M_{sr}. For each candidate site, we separately applied Deep-m^{6}A trained on the training data to calculate the probability of it to be a m^{6}A site. We predict a positive site when the output probability is greater than 0.907; this threshold is chosen because at this threshold the 10-fold CV test in training can achieve a precision of 0.7. In this way, we detected in total 23,456 single-base m^{6}A sites.

Hot-m^{6}A for identification of m^{6}A-regulated genes

We define m^{6}A-regulated genes as genes whose expression level is influenced by its m^{6}A methylation level more than expected by chance. We propose a new network based algorithm, Hot-m^{6}A, to identify m^{6}A-regulated genes. Hot-m^{6}A first determines that genes whose expression levels are significantly influenced by m^{6}A by assessing the correlation between the methylation level and the corresponding expression. If the expression levels of a gene change together with its m^{6}A levels across more samples, then there will be a higher chance that the gene expression is influenced by m^{6}A. To obtain a statistically meaningful correlation, we need to have the same m^{6}A sites appearing in many samples. In this study, we estimated the sample size needed to detect a correlation of 0.8 with a significance level of $\alpha = 0.05$. Using a two-sided test based on the Fisher's z transformation, at $\alpha = 0.05$ with power 90%, the required sample size is approximately 12 (i.e., n = 12) [48]. Thus, we select single-base m^{6}A sites that appear in at least 12 samples and calculated the Pearson's correlations of the methylation degree and gene

expression level across all occurred samples. The methylation degree was calculated as:

$$Meth_{level} = \ln(mean(e^{RC_{norm}})/G_{FPKM}) \quad (3)$$

where RC_{norm} is defined in (1), G_{FPKM} is the FPKM of gene harbored the corresponding m^{6}A site, calculated from the MeRIP-seq input sample by Cufflinks [84]. The expression level is denoted by the *ln* scale gene FPKM. The correlation coefficient is then transformed into a z score using a revised Fisher's z-transformation (which is known as Hotelling's (1953) second-order transformations) [49] that considers the influence of sample size *n*, which is expressed as

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad (4)$$

$$z^* = z - \frac{3z+r}{4n} \quad (5)$$

where *r* is the correlation, *n* is the sample size, and *z* is Fisher's transformation. We used z^* of each gene as its correlation for the subsequent analysis in the pipeline. Genes containing this kind of m^{6}A sites are treated as candidate m^{6}A-regulated genes. For a candidate gene that has multiple m^{6}A sites, the z^* score with the largest absolute value among all its harbored m^{6}A sites is selected as the correlation degree for that gene and the absolute z^* score are used to denote the degree of regulation.

To further select m^{6}A regulated genes more than expected by chance, Hot-m^{6}A adopts the HotNet2 algorithm [85], which was originally developed to identify significant mutation genes for cancer. HotNet2 takes the gene mutation frequency or mutation score as input heat vector and applies a heat diffusion method to identify subnetworks of a genome-scale interaction network that is mutated more than expected by chance. The advantage of HotNet2 is that it not only detects significant mutated genes with high mutation frequency but also can identify significant mutation genes with relatively low mutation frequency but interact closely with other significant genes. In our case, we also want to identify significant m^{6}A-regulated genes that not only have high m^{6}A-expression correlations but also cooperate with each other functionally in a functional network. To this end, Hot-m^{6}A takes the absolute m^{6}A-expression correlation coefficients z^* scores of candidate m^{6}A regulated genes as the input heat vector and uses the four PPI networks as the reference functional network, which includes BioGRID, which has been shown to contain the significant interactions between m^{6}A methylated genes [43], and the other three PPI networks that were used in the HotNet2 paper.

Hot-m^{6}A has two parameters β and δ; β is the fraction of the heat that a node in the network retains for itself at each time step and δ is the threshold, which determines whether there is an edge between 2 nodes in the final subnetwork. β can influence the amount of heat that a gene shares with its neighbors and is determined by the topology of the PPI network. For the three networks used in the HotNet2 paper, we selected β as 0.4 for HINT+HI2012, 0.45 for iRefIndex, and 0.5 for Multinet, which are reported values in [46]. For the BioGRID network, we performed the same analysis as the HotNet2 paper did to determine the value of β. For each different β from {0.05, 0.1, 0.15, . . ., 0.95}, we analyzed the inflection point at which the heat kept in the direct interacting neighbors of a gene drops and selected β as the one with the largest inflection point. As shown in S6 Fig, β is selected as 0.5 for the BioGRID PPI network. On the other hand, δ influences the scale of the subnetwork that we generated. To automatically determine it, for each PPI network, we firstly generated 100 random PPI networks [86]. We used the same "heat" vector at each run and select δ as the minimum one where all strongly connected subnetwork components identified by Hot-m^{6}A have a size less than and equal to a

threshold L_{max}. For each L_{max} = 5, 10, 15, 20, we reported the median of the 100 δ_{min}s for the 100 random networks. For each run, we selected the smallest δ with the most significant ($P < 0.05$) subnetwork sizes k as described in HotNet2 [46]. The P-value, which denotes the significance of subnetwork size k is computed for the statistic X_k, the number of subnetworks of size $\geq k$ reported by Hot-m^{6}A. To compute an empirical distribution of X_k for computing the P-value, we permuted the heat scores, i.e., the z transformation of correlation, among the genes in the original PPI network for 1000 times and then applied HotNet2 to the network using the permutated heat scores. In the end, δ was selected as 0.00769 for BioGRID, 0.0148 for HINT+HI2012, 0.00998 for iRefIndex, and 0.00777 for Multinet.

We applied Hot-m^{6}A to each of the four PPI networks and pooled all the genes and edges reported in at least one network to form a candidate m^{6}A regulated gene network, G. We then assigned the number of the PPI networks in which an edge exists as the weight of this edge. Next, we initialized the consensus subnetworks C as the connected components of G with edges of the weight = 4, i.e., the edges that are reported in all 4 PPI networks. Then we extended each consensus subnetwork S ∈ C by adding edges of weight = 3, and afterward further extended them by adding edges of weight = 2 and then 1. Finally, we defined the genes in the final extended networks as the m^{6}A regulated genes and the genes in the initial consensus subnetworks as the consensus m^{6}A regulated genes. To further detect the functional modules of m^{6}A regulated genes, we removed edges with weight = 1 in the m^{6}A regulated subnetworks to generate functional significant subnetworks.

Prioritize m^{6}A-associated diseases

To further reveal the association between m^{6}A and disease, we adopted the RWRH (random walk with restart on heterogeneous network) [47] algorithm to prioritize the candidate m^{6}A regulated diseases. RWRH is a well-known heterogeneous network-based algorithm to infer the gene-phenotype relationship. The heterogeneous network contains three parts: gene-gene interaction network, phenotype (disease) network, and gene-phenotype relationship. Similarly, we also built four heterogeneous networks based on each of the four PPI networks. The disease network was generated from the semantic similarities between two DO terms of Disease Ontology database and gene-phenotype relationships were extracted from OMIM. Specifically, the OMIM phenotype ID was mapped to DO disease ID to match the disease network with the gene-phenotype relationship. RWRH performs a random walk with restart from certain gene and disease seed nodes in the heterogeneous network. After several steps of random walk, the probability of each node, that the seed genes and seed diseases will walk to, become steady and is returned as vectors p_g^∞ and p_d^∞. $p_{g_i}^\infty$ denotes the probability that the seed genes and seed diseases will walk to gene i and $p_{d_j}^\infty$ is the probability that the seed genes and seed diseases will walk to disease j. In another world, RWRH can prioritize how closely a gene or a disease is correlated with the seed genes and seed diseases.

To study the correlation between m^{6}A and diseases, we first mapped all the 709 significant m^{6}A-regulated genes obtained from HotNet2 to OMIM gene-phenotype relationship, which resulted in 308 phenotypes regulated by 177 m^{6}A genes. Most of these phenotypes contain only one disease gene (The details of the m^{6}A gene-disease relationships are included in S1 File). Next, to obtain candidate m^{6}A-associated diseases, we mapped these 308 m^{6}A gene-correlated OMIM phenotypes to DO ID, which resulted in 90 m^{6}A gene-associated DO diseases.

For each of the four heterogeneous networks, these 90 diseases were taken as disease seeds for RWRH, whereas all the 709 significant m^{6}A-regulated genes were mapped to the PPI network to serve as gene seeds for RWRH. Finally, we applied RWRH using these genes and disease as seed nodes to prioritize all the DO diseases in the heterogeneous networks. Fig 5

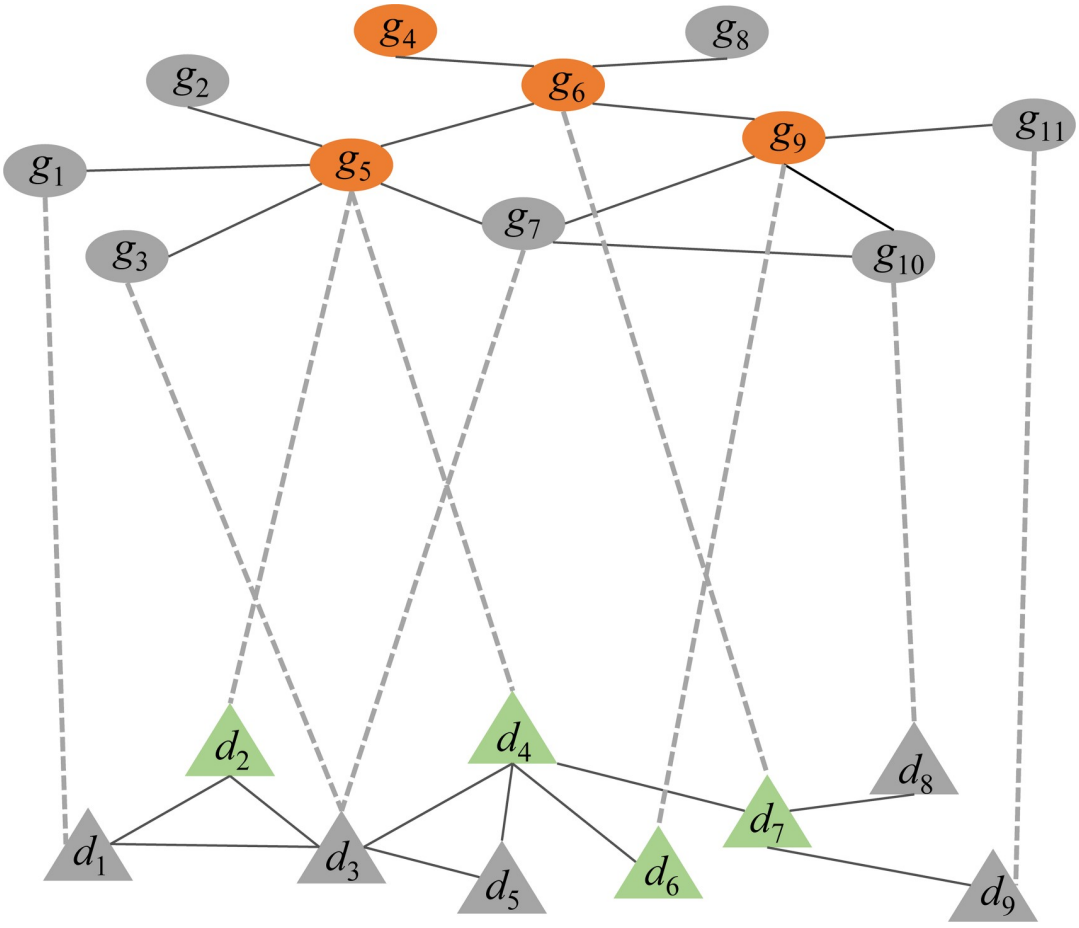

Fig 5. Gene-disease heterogeneous network. The top network is gene-gene interaction network, the bottom network is disease-disease similarity network and they are connected by gene-disease relationship (dashed grey lines). The orange gene nodes denoted the m^{6}A regulated genes and the green disease nodes are m^{6}A regulated genes correlated diseases.

https://doi.org/10.1371/journal.pcbi.1006663.g005

illustrates the heterogeneous network in this study. As shown, there is a chance that d_3 may be ranked as significant m^{6}A-associated diseases even though there is no m^{6}A-regulated gene directly connected with it. The reason is that the genes (i.e., g_3, g_7) that closely interacted with m^{6}A regulated genes (i.e., g_5) are connected to d_3 and m^{6}A regulated genes (i.e., g_5) connected with disease (i.e., d_2, d_4) also closely interacted with d_3. RWRH can efficiently capture this significant network topology and prioritized d_3 as significant, which showed the power of this study to prioritize the potential m^{6}A-associated disease with no prior knowledge that correlated with m^{6}A-regulated genes.

The top 10 ranked diseases based on the RWRH output probability p_d^∞ were selected as candidate m^{6}A-associated diseases. To ensure these top diseases were really influenced by m^{6}A-regulated genes rather than by chance, we implemented a random test to calculate an empirical p-value to assess if the disease is randomly selected. For this test, 100 random PPI networks that only keep the degree distribution of original network were generated using the method in [86]. Then, for each random PPI network, we connected it with the disease network using the disease-gene relationship and deleted the relationship between m^{6}A-regulated genes and their corresponding diseases to generate a corresponding random heterogeneous network whose gene interaction relationship is random and contains no prior knowledge of the relationship

between m^{6}A regulated genes and any disease. After that, we applied RWRH in each of the 100 random heterogeneous networks. Finally, for each candidate m^{6}A-associated disease d_j (j = 1, 2, ⋯, 10), the empirical p-value was calculated as: $p = \frac{\{\pi(d_j)\}}{100}$, where $\pi(d_j)$ is a random heterogeneous network in which d_j is found as the top 10 candidate disease after RWRH. The empirical p-value is calculated as the probability that a candidate m^{6}A-associated disease is selected by random. The candidate m^{6}A-associated diseases that with $p < 0.05$ were selected as significant candidate m^{6}A-associated diseases. The significant candidate m^{6}A-associated diseases that reported by all the four heterogeneous networks are determined as significant m^{6}A-associated diseases.

Supporting information

S1 Fig. Consensus m^{6}A-regulated subnetworks and m^{6}A-regulated subnetworks. The color of the node denotes the heat i.e. expression-methylation correlation degree it has, red denotes higher and blue denotes lower. The solid orange edge is edge that identified in all 4 PPI networks, the dashed edge is edge that identified in at least 3 networks, the dash dot edge is in 2 networks and the dotted is 1.
(TIF)

S2 Fig. Functional significant m^{6}A-regulated subnetworks. (A)-(I) denotes the 9 functional significant subnetworks of m6A-regulated genes. The color of the node denotes the heat i.e. expression-methylation correlation degree it has, red denotes higher and blue denotes lower. The solid edge is edge that identified in all 4 PPI networks, the dashed edge is edge that identified in at least 3 networks and the dotted edge is in 2 networks.
(TIF)

S3 Fig. Significantly enriched GO BP terms for each of the 9 largest significant subnetworks. There is no significant BP for subnetwork (B). Gene count means number of genes involved in the corresponding BP terms and $P_{Benjamini}$ is the adjusted FDR of enriched p-value. All the involved BP terms have a $P_{Benjamini} < 0.05$ and we only list the top 5 if the enriched terms are more than 5.
(TIF)

S4 Fig. Significantly enriched KEGG pathways for the 9 largest significant subnetworks. There is no significant KEGG pathways for network (D). Gene count means number of genes involved in the corresponding pathways and $P_{Benjamini}$ is the adjusted FDR of enriched p-value. All the involved pathways have a $P_{Benjamini} < 0.05$.
(TIF)

S5 Fig. Parts of the enriched KEGG pathways of the 9 significant subnetworks. (A) is WNT signaling pathway. (B) is mTOR signaling pathway. (C) is Focal adhesion. (D) is Notch signaling pathway. The genes marked with red star are m^{6}A regulated genes in the subnetworks. As is shown, m^{6}A regulated genes tend to be in the upstream of these enriched pathways.
(TIF)

S6 Fig. Selection of β of the BioGRID network. Figures (A)-(G) represent the distributions of number of nodes with influence larger than a cutoff for β from 0.6 to 0.3 for gene TP53, which has a high betweenness centrality. The x-axis of each distribution represents θ, a cutoff of influence. The y-axis represents the number of nodes in the interaction network with influence larger than θ. Red dotted circles denote the number of all nodes with an influence larger than different θs, green dotted circles denote that of the level-one nodes, and blue dotted circles denote that of the level-two nodes. The red vertical lines in all the distributions represent the

location of the inflection point in level one for the β we chose for different interaction networks, $\beta = 0.5$ for BioGRID. (H) depicts the inflection point θ as a function of β ranging from 0.05 to 0.95.
(TIF)

S1 File. OMIM diseases related to m^{6}A-regulated genes.
(XLSX)

Acknowledgments

We thank the computational support from UTSA's HPC cluster Shamu, operated by the Office of Information Technology.

Author Contributions

Conceptualization: Song-Yao Zhang, Yufei Huang.

Data curation: Song-Yao Zhang, Xiao-Nan Fan, Jia Meng, Shou-Jiang Gao.

Formal analysis: Song-Yao Zhang.

Funding acquisition: Song-Yao Zhang, Jia Meng, Yufei Huang.

Investigation: Song-Yao Zhang.

Methodology: Song-Yao Zhang.

Project administration: Shao-Wu Zhang, Yufei Huang.

Resources: Song-Yao Zhang.

Software: Song-Yao Zhang, Jia Meng, Shou-Jiang Gao.

Supervision: Shao-Wu Zhang, Yidong Chen, Yufei Huang.

Validation: Song-Yao Zhang, Xiao-Nan Fan, Shou-Jiang Gao.

Visualization: Song-Yao Zhang.

Writing – original draft: Song-Yao Zhang, Yufei Huang.

Writing – review & editing: Yufei Huang.

References

1. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell. 2012; 149(7):1635–46. Epub 2012/05/23. S0092-8674(12)00536-3 [pii] https://doi.org/10.1016/j.cell.2012.05.003 PMID: 22608085; PubMed Central PMCID: PMC3383396.
2. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012; 485 (7397):201–6. Epub 2012/05/12. https://doi.org/10.1038/nature11112 PMID: 22575960.
3. Ping XL, Sun BF, Wang L, Xiao W, Yang X, Wang WJ, et al. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. Cell research. 2014; 24(2):177–89. https://doi.org/10.1038/cr.2014.3 PMID: 24407421; PubMed Central PMCID: PMC3915904.
4. Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. Nat Cell Biol. 2014; 16(2):191–8. https://doi.org/10.1038/ncb2902 PMID: 24394384; PubMed Central PMCID: PMCPMC4640932.
5. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016; 32(12):1832–9. https://doi.org/10.1093/bioinformatics/btw074 PMID: 26873929

6. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, et al. N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. Cell. 2015; 161(6):1388–99. https://doi.org/10.1016/j.cell.2015.05.014 PMID: 26046440; PubMed Central PMCID: PMCPMC4825696.
7. Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian SB. Dynamic m(6)A mRNA methylation directs translational control of heat shock response. Nature. 2015; 526(7574):591–4. https://doi.org/10.1038/nature15377 PMID: 26458103.
8. Fustin J-M, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, et al. RNA-Methylation-Dependent RNA Processing Controls the Speed of the Circadian Clock. Cell. 2013; 155(4):793–806. http://dx.doi.org/10.1016/j.cell.2013.10.026. PMID: 24209618
9. Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang CM, Li CJ, et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol Cell. 2013; 49(1):18–29. https://doi.org/10.1016/j.molcel.2012.10.015 PMID: 23177736; PubMed Central PMCID: PMCPMC3646334.
10. Slobodin B, Han R, Calderone V, Vrielink JA, Loayza-Puch F, Elkon R, et al. Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. Cell. 2017; 169(2):326–37 e12. https://doi.org/10.1016/j.cell.2017.03.031 PMID: 28388414; PubMed Central PMCID: PMCPMC5388891.
11. Ratti A, Fallini C, Colombrita C, Pascale A, Laforenza U, Quattrone A, et al. Post-transcriptional regulation of neuro-oncological ventral antigen 1 by the neuronal RNA-binding proteins ELAV. Journal of Biological Chemistry. 2008; 283(12):7531–41. https://doi.org/10.1074/jbc.M706082200 PMID: 18218628
12. Oka Y, Tsuboi A, Elisseeva O, Udaka K, Sugiyama H. WT1 as a novel target antigen for cancer immunotherapy. Current cancer drug targets. 2002; 2(1):45–54. PMID: 12188920
13. Loeb DM, Evron E, Patel CB, Sharma PM, Niranjan B, Buluwela L, et al. Wilms' tumor suppressor gene (WT1) is expressed in primary breast tumors despite tumor-specific promoter methylation. Cancer research. 2001; 61(3):921–5. PMID: 11221883
14. Bansal H, Yihua Q, Iyer SP, Ganapathy S, Proia DA, Penalva LO, et al. WTAP is a novel oncogenic protein in acute myeloid leukemia. Leukemia. 2014; 28(5):1171–4. https://doi.org/10.1038/leu.2014.16 PMID: 24413322; PubMed Central PMCID: PMCPMC4369791.
15. Long J, Zhang B, Signorello LB, Cai Q, Deming-Halverson S, Shrubsole MJ, et al. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. PLoS One. 2013; 8(4):e58350. https://doi.org/10.1371/journal.pone.0058350 PMID: 23593120
16. Kaklamani V, Yi N, Sadim M, Siziopikou K, Zhang K, Xu Y, et al. The role of the fat mass and obesity associated gene (FTO) in breast cancer risk. BMC medical genetics. 2011; 12(1):52.
17. Franchini M, Liumbruno GM. ABO blood group: old dogma, new perspectives. Clinical Chemistry and Laboratory Medicine. 2013; 51(8):1545–53. https://doi.org/10.1515/cclm-2013-0168 PMID: 23648637
18. Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, et al. FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N6-Methyladenosine RNA Demethylase. Cancer cell. 2017; 31(1):127–41. https://doi.org/10.1016/j.ccell.2016.11.017 PMID: 28017614.
19. Keith B, Simon MC. Hypoxia-inducible factors, stem cells, and cancer. Cell. 2007; 129(3):465–72. https://doi.org/10.1016/j.cell.2007.04.019 PMID: 17482542
20. Zhang C, Samanta D, Lu H, Bullen JW, Zhang H, Chen I, et al. Hypoxia induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m6A-demethylation of NANOG mRNA. Proceedings of the National Academy of Sciences. 2016; 113(14):E2047–E56.
21. Lin S, Choe J, Du P, Triboulet R, Gregory RI. The m 6 A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. Molecular cell. 2016.
22. Lichinchi G, Zhao BS, Wu Y, Lu Z, Qin Y, He C, et al. Dynamics of Human and Viral RNA Methylation during Zika Virus Infection. Cell Host Microbe. 2016; 20(5):666–73. https://doi.org/10.1016/j.chom.2016.10.002 PMID: 27773536; PubMed Central PMCID: PMCPMC5155635.
23. Kennedy EM, Bogerd HP, Kornepati AV, Kang D, Ghoshal D, Marshall JB, et al. Posttranscriptional m(6)A Editing of HIV-1 mRNAs Enhances Viral Gene Expression. Cell Host Microbe. 2016; 19(5):675–85. https://doi.org/10.1016/j.chom.2016.04.002 PMID: 27117054; PubMed Central PMCID: PMCPMC4867121.
24. Tirumuru N, Zhao BS, Lu W, Lu Z, He C, Wu L. N(6)-methyladenosine of HIV-1 RNA regulates viral infection and HIV-1 Gag protein expression. Elife. 2016; 5. https://doi.org/10.7554/eLife.15528 PMID: 27371828; PubMed Central PMCID: PMCPMC4961459.
25. Tan B, Liu H, Zhang S, da Silva SR, Zhang L, Meng J, et al. Viral and cellular N 6-methyladenosine and N 6, 2′-O-dimethyladenosine epitranscriptomes in the KSHV life cycle. Nature microbiology. 2018; 3(1):108. https://doi.org/10.1038/s41564-017-0056-8 PMID: 29109479

26. Liu H, Wang H, Wei Z, Zhang S, Hua G, Zhang S-W, et al. MeT-DB V2. 0: elucidating context-specific functions of N 6-methyl-adenosine methyltranscriptome. Nucleic acids research. 2017; 46(D1):D281–D7.
27. Liu Z, Xiao X, Yu D-J, Jia J, Qiu W-R, Chou K-C. pRNAm-PC: Predicting N 6-methyladenosine sites in RNA sequences via physical–chemical properties. Analytical biochemistry. 2016; 497:60–7. https://doi.org/10.1016/j.ab.2015.12.017 PMID: 26748145
28. Chen W, Feng P, Ding H, Lin H, Chou K-C. iRNA-Methyl: Identifying N 6-methyladenosine sites using pseudo nucleotide composition. Analytical biochemistry. 2015; 490:26–33. https://doi.org/10.1016/j.ab.2015.08.021 PMID: 26314792
29. Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic acids research. 2016; 44(10):e91–e. https://doi.org/10.1093/nar/gkw104 PMID: 26896799
30. Xiang S, Liu K, Yan Z, Zhang Y, Sun Z. RNAMethPre: A Web Server for the Prediction and Query of mRNA m 6 A Sites. PloS one. 2016; 11(10):e0162707. https://doi.org/10.1371/journal.pone.0162707 PMID: 27723837
31. Xing P, Su R, Guo F, Wei L. Identifying N6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. Scientific Reports. 2017; 7.
32. Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N6-methyladenosine sites. Journal of Biomolecular Structure and Dynamics. 2017; 35(3):683–7. https://doi.org/10.1080/07391102.2016.1157761 PMID: 26912125
33. Wei L, Su R, Wang B, Li X, Zou Q, Gao X. Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. Neurocomputing. 2018.
34. Jiang S, Xie Y, He Z, Zhang Y, Zhao Y, Chen L, et al. m6ASNP: a tool for annotating genetic variants by m6A function. GigaScience. 2018; 7(5):giy035.
35. Chen W, Feng P, Ding H, Lin H, Chou K-C. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Analytical biochemistry. 2015; 490:26–33. https://doi.org/10.1016/j.ab.2015.08.021 PMID: 26314792
36. Meng J, Lu Z, Liu H, Zhang L, Zhang S, Chen Y, et al. A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. Methods. 2014; 69(3):274–81. https://doi.org/10.1016/j.ymeth.2014.06.008 PMID: 24979058; PubMed Central PMCID: PMCPMC4194139.
37. Cui X, Zhang L, Meng J, Rao M, Chen Y, Huang Y. MeTDiff: a Novel Differential RNA Methylation Analysis for MeRIP-Seq Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2015.
38. Xiang S, Yan Z, Liu K, Zhang Y, Sun Z. AthMethPre: A web server for the prediction and query of mRNA m 6 A sites in Arabidopsis thaliana. Molecular BioSystems. 2016; 12(11):3333–7. https://doi.org/10.1039/c6mb00536e PMID: 27550167
39. Chen W, Xing P, Zou Q. Detecting N 6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. Scientific reports. 2017; 7:40242. https://doi.org/10.1038/srep40242 PMID: 28079126
40. Xing P, Su R, Guo F, Wei L. Identifying N 6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. Scientific reports. 2017; 7:46757. https://doi.org/10.1038/srep46757 PMID: 28440291
41. Meng J, Cui X, Rao MK, Chen Y, Huang Y. Exome-based analysis for RNA epigenome sequencing data. Bioinformatics. 2013; 29(12):1565–7. https://doi.org/10.1093/bioinformatics/btt171 PMID: 23589649
42. Cui X, Meng J, Zhang S, Chen Y, Huang Y. A novel algorithm for calling mRNA m 6 A peaks by modeling biological variances in MeRIP-seq data. Bioinformatics. 2016; 32(12):i378–i85. https://doi.org/10.1093/bioinformatics/btw281 PMID: 27307641
43. Zhang S-Y, Zhang S-W, Liu L, Meng J, Huang Y. m6A-Driver: identifying context-specific mRNA m6A methylation-driven gene interaction networks. PLoS computational biology. 2016; 12(12):e1005287. https://doi.org/10.1371/journal.pcbi.1005287 PMID: 28027310
44. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11):2278–324.
45. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nature methods. 2015; 12(8):767. https://doi.org/10.1038/nmeth.3453 PMID: 26121403
46. Leiserson MD, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics. 2015; 47(2):106. https://doi.org/10.1038/ng.3168 PMID: 25501392

47. Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010; 26(9):1219–24. https://doi.org/10.1093/bioinformatics/btq108 PMID: 20215462

48. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled clinical trials. 1981; 2(2):93–113. PMID: 7273794

49. Hotelling H. New light on the correlation coefficient and its transforms. Journal of the Royal Statistical Society Series B (Methodological). 1953; 15(2):193–232.

50. Chatr-aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2014:gku1204.

51. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC systems biology. 2012; 6(1):92.

52. Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, et al. Next-generation sequencing to generate interactome datasets. Nature methods. 2011; 8(6):478. https://doi.org/10.1038/nmeth.1597 PMID: 21516116

53. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS computational biology. 2013; 9(3):e1002886. https://doi.org/10.1371/journal.pcbi.1002886 PMID: 23505346

54. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC bioinformatics. 2008; 9(1):405.

55. Ke S, Alemu EA, Mertens C, Gantman EC, Fak JJ, Mele A, et al. A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. Genes & development. 2015; 29(19):2037–53.

56. Pan X, Shen H-B. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics. 2018.

57. Zeiler MD. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:12125701. 2012.

58. Cui X, Wei Z, Zhang L, Liu H, Sun L, Zhang S-W, et al. Guitar: an R/bioconductor package for gene annotation guided Transcriptomic analysis of RNA-related genomic features. BioMed Research International. 2016; 2016.

59. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome biology. 2003; 4(9):R60.

60. Zhao X, Yang Y, Sun B-F, Shi Y, Yang X, Xiao W, et al. FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell research. 2014; 24(12):1403. https://doi.org/10.1038/cr.2014.151 PMID: 25412662

61. Polakis P. Wnt signaling and cancer. Genes & development. 2000; 14(15):1837–51.

62. Reya T, Clevers H. Wnt signalling in stem cells and cancer. Nature. 2005; 434(7035):843. https://doi.org/10.1038/nature03319 PMID: 15829953

63. Bienz M, Clevers H. Linking colorectal cancer to Wnt signaling. Cell. 2000; 103(2):311–20. PMID: 11057903

64. Jaffrey SR, Kharas MG. Emerging links between m 6 A and misregulated mRNA methylation in cancer. Genome medicine. 2017; 9(1):2. https://doi.org/10.1186/s13073-016-0395-8 PMID: 28081722

65. Chen M, Wei L, Law CT, Tsang FHC, Shen J, Cheng CLH, et al. RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2. Hepatology. 2018; 67(6):2254–70. https://doi.org/10.1002/hep.29683 PMID: 29171881

66. Esteller M, Pandolfi PP. The epitranscriptome of noncoding RNAs in cancer. Cancer discovery. 2017.

67. Wang X, Li Z, Kong B, Song C, Cong J, Hou J, et al. Reduced m6A mRNA methylation is correlated with the progression of human cervical cancer. Oncotarget. 2017; 8(58):98918. https://doi.org/10.18632/oncotarget.22041 PMID: 29228737

68. Bansal H, Yihua Q, Iyer SP, Ganapathy S, Proia D, Penalva L, et al. WTAP is a novel oncogenic protein in acute myeloid leukemia. Leukemia. 2014; 28(5):1171. https://doi.org/10.1038/leu.2014.16 PMID: 24413322

69. Vu LP, Pickering BF, Cheng Y, Zaccara S, Nguyen D, Minuesa G, et al. The N 6-methyladenosine (m 6 A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. Nature medicine. 2017; 23(11):1369. https://doi.org/10.1038/nm.4416 PMID: 28920958

70. Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, et al. FTO plays an oncogenic role in acute myeloid leukemia as a N 6-methyladenosine RNA demethylase. Cancer Cell. 2017; 31(1):127–41. https://doi.org/10.1016/j.ccell.2016.11.017 PMID: 28017614

71. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, et al. N 6-methyladenosine modulates messenger RNA translation efficiency. Cell. 2015; 161(6):1388–99. https://doi.org/10.1016/j.cell.2015.05.014 PMID: 26046440

72. Lin S, Choe J, Du P, Triboulet R, Gregory RI. The m 6 A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. Molecular cell. 2016; 62(3):335–45. https://doi.org/10.1016/j.molcel.2016.03.021 PMID: 27117702

73. Yang Y, Shen F, Huang W, Qin S, Huang J-T, Sergi C, et al. Glucose is involved in the dynamic regulation of m6A in patients with type 2 diabetes. The Journal of clinical endocrinology and metabolism. 2018.

74. Li X, Tang J, Huang W, Wang F, Li P, Qin C, et al. The M6A methyltransferase METTL3: acting as a tumor suppressor in renal cell carcinoma. Oncotarget. 2017; 8(56):96103. https://doi.org/10.18632/oncotarget.21726 PMID: 29221190

75. Liu N, Pan T. RNA epigenetics. Translational Research. 2015; 165(1):28–35. https://doi.org/10.1016/j.trsl.2014.04.003 PMID: 24768686

76. Cardelli M, Marchegiani F, Cavallone L, Olivieri F, Giovagnetti S, Mugianesi E, et al. A polymorphism of the YTHDF2 gene (1p35) located in an Alu-rich genomic domain is associated with human longevity. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2006; 61(6):547–56.

77. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell. 2012; 149(7):1635–46. https://doi.org/10.1016/j.cell.2012.05.003 PMID: 22608085

78. Zhang S, Zhao BS, Zhou A, Lin K, Zheng S, Lu Z, 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. e6. https://doi.org/10.1016/j.ccell.2017.02.013 PMID: 28344040

79. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015; 2015.

80. Barbieri I, Tzelepis K, Pandolfini L, Shi J, Millán-Zambrano G, Robson SC, et al. Promoter-bound METTL3 maintains myeloid leukaemia by m 6 A-dependent translation control. Nature. 2017; 552(7683):126. https://doi.org/10.1038/nature24678 PMID: 29186125

81. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012; 40(Database issue):D940–6. Epub 2011/11/15. https://doi.org/10.1093/nar/gkr972 PMID: 22080554; PubMed Central PMCID: PMCPMC3245088.

82. Yu GC, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015; 31(4):608–9. https://doi.org/10.1093/bioinformatics/btu684 WOS:000350059600025. PMID: 25677125

83. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2005; 33(suppl_1):D514–D7.

84. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012; 7(3):562. https://doi.org/10.1038/nprot.2012.016 PMID: 22383036

85. Leiserson MD, Vandin F, Wu H-T, Dobson JR, Raphael BR. Pan-cancer identification of mutated pathways and protein complexes. AACR; 2014.

86. Milo R, Kashtan N, Itzkovitz S, Newman ME, Alon U. On the uniform generation of random graphs with prescribed degree sequences. arXiv preprint cond-mat/0312028. 2003.