Signal Detection and Biological Feature Extraction for High-throughput Data of N6-methyladenosine (m6A)

N6-メチルアデノシン (m6A) の
ハイスループットデータの信号検出と
生物学的特徴抽出

October 2020

Yiqian ZHANG
張 怡倩

Signal Detection and Biological Feature Extraction for High-throughput Data of N6-methyladenosine (m6A)

N6-メチルアデノシン (m6A) の
ハイスループットデータの信号検出と
生物学的特徴抽出

October 2020

Waseda University

Graduate School of Advanced Science and Engineering

Department of Electrical Engineering and Bioscience

Research on Bioinformatics

Yiqian ZHANG

張 怡倩

ABSTRACT

RNA modification is biochemical modifications of RNA and plays their roles in gene regulation at post-transcription level. So far, more than 150 types of RNA modifications have been discovered with N6-methyladenosine (m6A) as being one of the most abundant types found in nature. m6A is featured with its preferential location near 3' untranslated regions (3' UTR) and its nearby sequences mostly conforming to a certain motif, i.e., DRACH (where D = A, G or U; R = A or G; H= A, C or U) in the mammalian genome.

There are two main kinds of high-throughput sequencing technologies for mapping transcriptome-wide m6A. One is named as MeRIP-Seq(Methylated RNA immunoprecipitation sequencing, also known as m6A-seq) that was developed in 2012. The other is named as miCLIP-Seq that was developed in 2015. MeRIP-Seq detects genomic regions containing m6A sites and is economical, therefore more popular for biologists while miCLIP-Seq detects positions of m6A sites at single-base resolution.

m6A participates in essential RNA activities including alternative splicing, export, translation, and decay. During these biological processes, m6A exerts its function through interaction with several RNA binding proteins (RBPs) that can be considered as m6A-associated RBPs. There are three main types of m6A-associated RPBs, i.e. writer, eraser, and reader. m6A writer is methyltransferase including METTL3, METTL14, WTAP, RBM15/15B; m6 eraser is demethyltransferase including FTO, ALKBH5; m6A reader is proteins that can recognize m6A including YTH domain-containing proteins (YTHDF1/2/3), EIF3, FMR1. m6A writers and erasers can be considered as m6A regulators which directly regulate m6A while m6A readers can be considered as m6A effectors which participate in m6A regulatory network. These m6A-associated RBPs cooperate with each other to facilitate both temporal and spatial regulation. On the other hand, given m6A's essential roles in gene regulation, it has been found that dysfunction of m6A-associated RBPs is related to cancer progression. Therefore, the study of m6A and m6A-associated RBPs enables us to develop a better understanding of gene regulation mechanism and leads to potential therapeutic opportunities.

For m6A data analysis, I would like to be focused on two main challenges that researchers are interested in, one is the signal detection and the other is the biological feature extraction.

For the signal detection, it means to apply statistical models to detect genomic regions with m6A from MeRIP-Seq data. Commonly used tools for m6A signal detection either require a long time or do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling. In addition, with more attention on the treatment experiments (perturbation of methyltransferases) of MeRIP-Seq, biologists need intuitive evaluation on the treatment effect from comparison. Therefore, efficient and user-friendly software that can solve these tasks must be developed.

For the biological feature extraction, it means to identify biological features surrounding m6A-containing sequences using machine learning. Some tools built prediction models using random forest (RF) or support vector machine (SVM) algorithm with existing knowledge as feature input like a combination of k-mers and chemical properties, however these features are not easily interpretable, therefore it needs a prediction model for extracting meaningful biological information such as RNA binding proteins that can assist biologists in studying regulation mechanism of m6A. In that case, a deep learning model equipped with a motif (sequences recognized by certain proteins) detector is a good choice. In addition, although motifs learned from a deep learning model can help identify potential m6A-associated RBPs, the sequence-based feature has its limitation because not all the motifs of RBPs are available and sequences cannot reflect actual protein binding. Therefore, it also needs to integrate with other kinds of data like RBPs' binding data to extract biological information beyond sequence level.

The purpose of this thesis is to develop efficient and user-friendly software for detection of m6A signal and extract biological features surrounding m6A by applying statistical models and state-of-the-art machine learning methods. The main contributions of this thesis can be summarized into three aspects. 1) I developed DeepM6ASeq, a deep learning framework, to predict m6A-containing sequences and characterize biological features surrounding m6A sites. DeepM6ASeq is a sequence-based predictor that showed competitive performance of prediction, learned known m6A readers and a newly recognized one, FMR1, and also helped to visualize locations of m6A sites with a saliency map. 2) I developed MoAIMS (model-based analysis and inference of MeRIP-Seq), an efficient and easy-to-use software for analysis of MeRIP-Seq. For detection of m6A regions, MoAIMS achieves excellent speed and competitive performance compared with other tools, and provides user-friendly outputs for downstream analysis. MoAIMS also provides intuitive evaluation on treatment effects for MeRIP-Seq treatment datasets. 3) I designed an

integrative computational framework for the identification of m6A-associated RBPs from reproducible m6A regions. The framework is composed of an enrichment analysis and a classification model. Utilizing the RBPs' binding data, the framework is able to identify known m6A-associated RBPs and also found some potential m6A-associated RBPs like RBM3 for mouse. Besides, it also helps infer interaction between m6A and m6A-associated RBPs including actions of reading and repelling beyond sequence level.

The thesis is composed of five chapters, which are demonstrated briefly in the followings,

[**Chapter 1**] **Introduction** presents the research background of m6A, one of the most abundant RNA modification, including its essential roles in gene regulation with m6A-associated RBPs, its detection methods of high-throughput sequencing technologies, overview of algorithms for m6A data analysis. Next, research objectives are proposed and the contribution of the thesis is described. Finally, the organization of the thesis is demonstrated.

[**Chapter 2**] **DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning** presents the work on m6A biological features extraction at sequence level. Existing tools can predict m6A at single-base resolution, however, the features they used for prediction such as k-mers and chemical properties are less interpretable. It needs a model to extract meaningful biological information like RNA binding proteins in the m6A-containing sequences in order to provide more insights for the biologists, therefore I implemented a deep learning framework, named DeepM6ASeq, to predict m6A-containing sequences and characterize surrounding biological features based on miCLIP-Seq data, which detects m6A sites at single-base resolution. DeepM6ASeq showed better performance as compared to other machine learning classifiers. Moreover, an independent test on MeRIP-Seq data, which identifies m6A-containing genomic regions, revealed that our model is competitive in predicting m6A-containing sequences. DeepM6ASeq utilized the convolutional neural network (CNN) layer as a motif detector and identified known m6A readers. Notably, DeepM6ASeq also identifies a newly recognized m6A reader: FMR1. Besides, I found that a saliency map in the deep learning model could be utilized to visualize locations of m6A sites.

[**Chapter 3**] **MoAIMS: efficient software for detection of enriched regions of MeRIP-Seq** presents the work on m6A signal detection of MeRIP-Seq(Methylated RNA

immunoprecipitation sequencing). MeRIP-Seq is an economical and popular high-throughput sequencing method for studying m6A. The signal detection is a main challenge for data analysis of MeRIP-Seq, however current tools either require a long time or do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling. On the other hand, with more attention on the treatment experiments of MeRIP-Seq, biologists need intuitive evaluation on the treatment effect from comparison. Therefore, efficient and user-friendly software that can solve these tasks must be developed. I developed a software named "model-based analysis and inference of MeRIP-Seq (MoAIMS)" to detect enriched regions of MeRIP-Seq and infer signal proportion based on a mixture negative-binomial model. MoAIMS is designed for transcriptome immunoprecipitation sequencing experiments; therefore, it is compatible with different RNA sequencing protocols. MoAIMS offers excellent processing speed (nearly ten times faster) and competitive performance on motif occurrence in the m6A regions and the overlapping percentage with miCLIP-Seq data when compared with other tools. Furthermore, signal proportion inferred from MoAIMS showed a decreasing trend for m6A treatment dataset (perturbation of m6A methyltransferases) compared with m6A wild-type dataset, which is consistent with experimental observations, suggesting that the signal proportion can be used as an intuitive indicator of treatment effect.

[**Chapter 4**] **Identification of m6A-associated RNA binding proteins using an integrative computational framework** presents the work on m6A biological features extraction at protein-binding level. Existing tools for extracting m6A biological features are sequence-based ones which are limited because not all the motifs of RBPs are available and sequences cannot reflect actual protein binding, therefore in this study I designed an integrative computational framework to extract m6A biological features, i.e. m6A-associated RBPs, utilizing RBP's binding data. I identified reproducible m6A regions from independent studies in certain cell lines and then utilized RBPs' binding data of the same cell line to identify m6A-associated RBPs. The computational framework is composed of an enrichment analysis and a classification model. The enrichment analysis identified known m6A-associated RBPs including YTH domain-containing proteins; it also identified a potential m6A-associated RBP, RBM3, for mouse. I observed a significant correlation for the identified m6A-associated RBPs at the protein expression level rather than the gene expression. In addition, I built a Random Forest classification model for the reproducible m6A regions using the information of RBPs' binding. The RBP-based predictor not only demonstrated competitive performance compared with sequence-based ones and but also

helped identify m6A-repelled RBP. These results suggested that the framework enabled us to infer interaction between m6A and m6A-associated RBPs beyond sequence level when utilizing RBPs' binding data.

[**Chapter 5**] **General conclusions and future work** presents general conclusion for the thesis and future work. The future work has main three parts, including model improvement for the signal detection, algorithms and data worth investigating for biological feature extraction and application in other RNA modifications.

In conclusion, the key contribution of the thesis is to extract meaningful biological features like m6A-associated RBPs by applying mathematical models in the analysis of m6A high-throughput data so that it could help biologists develop more insights into regulation mechanism of m6A.

ACKNOWLEDGMENTS

First, I would like to thank my supervisor Prof. Michiaki Hamada for the strong support and freedom he gave me during my Ph.D. study. His rigorous attitude toward research and his commitment deeply impacted me and guided me to be a professional researcher. Second, I would like to thank Prof. Masato Inoue , Prof. Noboru Murata, and Prof. Martin Frith, who were the members of my thesis committee. They helped me improve this thesis with many useful comments. Third, I would like to thank Otsuka Toshimi Scholarship for their financial support that enabled me to be committed into research. Next, I would like to thank the members of Hamada Lab, Dr. Tsukasa Fukunaga and Dr. Chao Zeng. They gave me helpful advices on my research. Besides, I would like to thank my friends for their care in this special year. Last, I would like to thank my parents for their love and encouragement.

Contents

Chapter 1

Introduction

1.1 Research Background

1.1.1 m6A's essential roles in gene regulation

RNA modification is biochemical modifications of RNA and plays their roles in gene regulation at post-transcription level, which is also known as epitranscriptome [1]. So far, more than 150 types of RNA modifications have been discovered [2]. Among them, N6-methyladenosine (m6A) [3], firstly reported in 1970 [4], is one of the most abundant types found in various species, including human, mouse, and yeast [5–7]. m6A is featured with its preferential location near 3' untranslated regions (3' UTR) and its nearby sequences mostly conforming to a certain motif, i.e., DRACH (where D = A, G or U; R = A or G; H= A, C or U) in the mammalian genome [8].

Several RNA binding proteins (RBPs) have been found to be associated with regulation of m6A modification, which can be considered as m6A-associated RBPs. There are three main kinds of m6A-associated RBPs, i.e. writer, eraser, and reader [9]. m6A writer is methyltransferase responsible for the formation of m6A. METTL3 is the first writer discovered in the 1980s [10]. Then, other writers including partners of METTL3 were found such as METTL14 [11], WTAP [12], RBM15/15B [13]. m6 eraser is demethyltransferase that facilitates removing the methylation. The identified m6A erasers are FTO [14] and ALKBH5 [15]. m6A reader is proteins that can recognize m6A and influence activities of

targeted RNA. A representative m6A reader group is YTH proteins (YTHDF1/2/3, YTHDC1/2) [6,16]. YTH proteins have a YTH domain that forms a hydrophobic pocket to promote the binding of m6A [17]. Other readers were also discovered including EIF3 and its subunit [18], FMR1 [19]. m6A writers and erasers can be considered as m6A regulators which directly regulate m6A while m6A readers can be considered as m6A effectors which participate in m6A regulatory network. These m6A-associated RBPs cooperate with each other to facilitate both temporal and spatial regulation where writers work in the nucleus to introduce the m6A modification which is then recognized by various readers in the nucleus and cytoplasm, which can influence activities of their target RNAs as shown in the Figure 1.1.

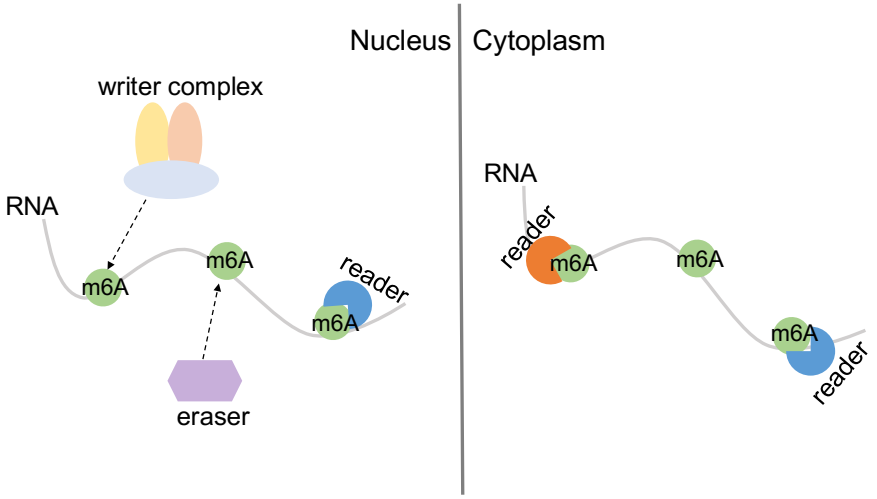


Fig. 1.1 Illustration of m6A modification and m6A-associated RBPs. In the nucleus, m6A modification is a dynamic process where m6A can be installed by writer complex or removed by erasers. In both the nucleus and cytoplasm, m6A can be recognized by readers.

m6A and m6A-associated RBPs have influence on several essential RNA activities. First, because m6A is preferentially located near 3' UTR, which is the region of mRNA behind translation termination codon, some m6A-associated RBPs participate in RNA activities related to regulatory roles of 3' UTR like translation, export, and decay [18]. For example, the reader YTHDC1 is related to export of methylated mRNA from the nucleus to the cytoplasm out of the observation that knock-down of YTHDC1 leaded to the accumulation of transcripts in the nucleus and the depletion within the cytoplasm [20]. The reader YTHDF3 is reported to affect methylated mRNA decay with the observation that knockdown of YTHDF3 resulted in decreased translation efficiency and increased cellular m6A level [21]. Besides, m6A is also related to alternative splicing [22,23]. A study found that the reader YTHDC1, which works

as a recruiter for splicing factors, impacted massive alternative splicing defects in mouse oocytes when it was lost [23].

Given the essential roles of m6A and m6A-associated RBPs in gene regulation, their roles in cancer are getting attention. The writer METTL3 was early noticed because of its overexpression in acute myeloid leukemia (AML). It was found that m6A promotes the translation of oncogenes like c-MYC, BCL2, and PTEN in the human acute myeloid leukemia MOLM-13 cell line [24]. Knock-down of METTL3 induced differentiation and failure to establish leukaemia in immuno-deficient mice [25]. Because of necessity of METTL3 in the maintain the leukaemic state, it is identified as a potential therapeutic target for AML. In addition, in human hepatocellular carcinoma (HCC), METTL3 is also significantly up-regulated and METTL3-mediated m6A modification repressed the expression of SOCS2 (suppressor of cytokine signaling 2) in HCC. Knockout of METTL3 showed remarkable suppression of HCC tumorigenicity and lung metastasis in mice [26]. Apart from METTL3, a study found that the reader YTHDF2 silenced in HCC cells can provoke inflammation, vascular reconstruction, and metastatic progression [27]. Furthermore, m6A and the reader YTHDF1 have been reported to control anti-tumor immunity. YTHDF1 deficient mice had enhanced therapeutic efficacy of PD-1 checkpoint blockade which suggested YTHDF1's potential in anti-cancer immunotherapy [28]. Therefore, the study of m6A and m6A-associated RBPs enables us to develop a better understanding of gene regulation mechanism and leads to potential therapeutic opportunities.

1.1.2 Transcriptome-wide mapping of m6A

With the rapid development of high-throughput sequencing technologies, there are two methods for mapping transcriptome-wide m6A. One is named as MeRIP-Seq (Methylated RNA immunoprecipitation sequencing, also known as m6A-seq) and the other is miCLIP-Seq (m6A individual-nucleotide-resolution cross-linking and immunoprecipitation). The m6A data used in this thesis is from these two kinds of high-throughput sequencing technologies. The experimental procedures are shown in Figure 1.2 and Figure 1.3.

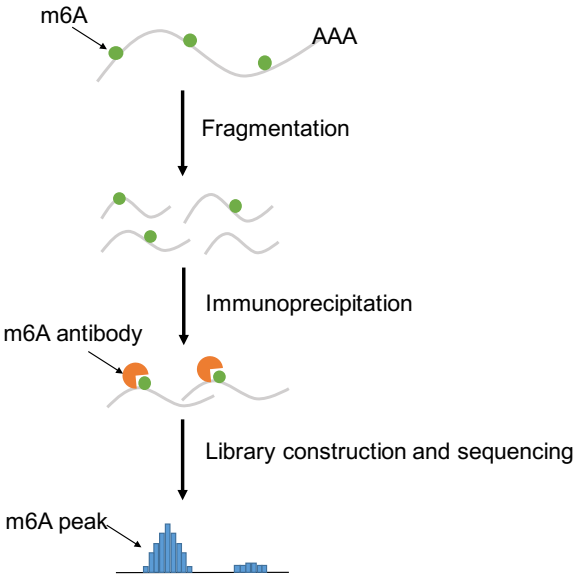


Fig. 1.2 Schematics of experimental procedures for MeRIP-Seq. In MeRIP-Seq, fragmented RNA is recognized by antibody specific to m6A and then captured for high-throughput sequencing. Peaks in sequencing reads signal indicate the regions with m6A.

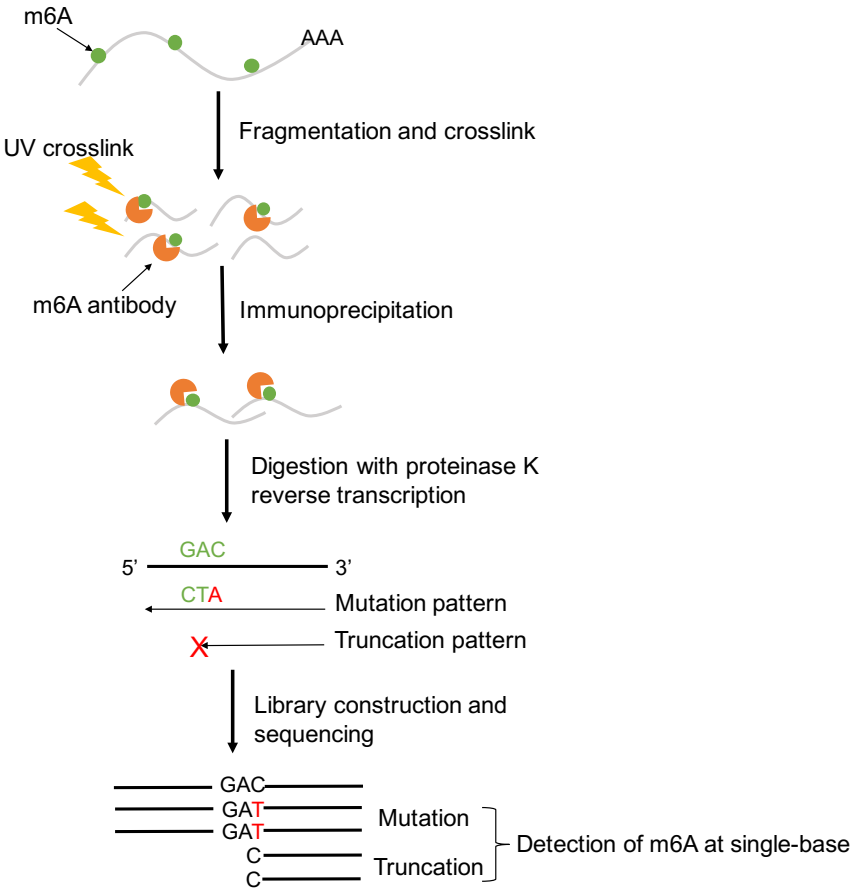


Fig. 1.3 Schematics of experimental procedures for miCLIP. In miCLIP, m6A antibody is firstly cross-linked to methylated RNA, which is then followed by immunoprecipitation. Digestion with proteinase K and reverse transcription detect truncations or mutations signal close to m6A sites so as to identify precise locations of m6A.

MeRIP-Seq was developed independently by two research groups in 2012 [5,6]. In MeRIP-Seq, an antibody specific to m6A is used to immunoprecipitate fragmented RNAs, then targeted RNAs are subjected to sequencing. Three years later in 2015, miCLIP aimed to detect m6A at single-base resolution appeared [8,29]. miCLIP-Seq is a UV-based sequencing method in that m6A antibody is firstly cross-linked to methylated RNA, which is then followed by immunoprecipitation. miCLIP utilized the truncations or mutations signal close to m6A sites during the process of digestion with proteinase K and reverse transcription to locate precise m6A site. MeRIP-Seq can detect genomic regions containing m6A sites at a resolution of 100-200 bp which is lower than miCLIP-Seq that calls m6A sites at single-base resolution, however, because of simpler experimental procedures and the economical feature, MeRIP-Seq is still more popular in the m6A research.

1.1.3 Overview of algorithms for m6A data analysis

For m6A data analysis, researchers are faced with several key issues to be answered, and I decided to be focused on the two main challenges, one is the signal detection and the other is the biological feature extraction. The algorithms for these two challenges are discussed in the following sections.

Signal detection

For the signal detection, it means to apply statistical models to detect genomic regions with m6A from MeRIP-Seq data. Efficient analysis fo MeRIP-Seq data helps biologists study which genes are regulated by m6A. In high-throughput sequencing technologies like MeRIP-Seq, data is in the form of tag counts, therefore one common statistical model is Poisson distribution. Poisson distribution assumes that the discrete variable, i.e. tag count, follows the probability distribution $Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ with a parameter mean λ. Commonly used computational tools for MeRIP-Seq such as MACS [30], exomePeak [31], and MeTPeak [32] assume the Poisson distribution for tag counts, however, negative binomial(NB) model distribution is considered as a better choice. NB distribution follows the probability distribution $Pr(X = k) = \frac{\Gamma(k+r)}{k!\Gamma(r)} p^r (1-p)^k$ with two parameters r and p, representing size and probability. Because NB distribution has an extra parameter to model for variance while Poisson distribution assuming equal mean and variance, NB distribution is more suitable for RNA sequencing data with higher variance than mean.

To model MeRIP-Seq data better and improve three points in the existing tools, they are: 1) require a long time for analysis; 2) do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling; 3) cannot provide intuitive evaluation on treatment effect of treatment experiments (perturbation of methyltransferases) for biologists, it needs to develop efficient and user-friendly software that can solve these tasks.

Biological feature extraction

For the biological feature extraction, it means to identify biological features surrounding m6A-containing sequences using machine learning, which can assist biologists in studying

regulation mechanism of m6A like which RPBs interact with m6A and influence RNA activities. miCLIP-Seq is useful data for prediction of m6A because it provides precise locations of m6A sites. Traditional machine learning classifiers like Random forest (RF) and support vector machine (SVM) have been applied by existing tools for the prediction [33, 34]. RF is an ensemble model consisting of many decision trees. It makes prediction using bootstrapping, random selection of features, and average votes [35]. For SVM, the basic idea is to find a decision boundary that can separate data into different classes with the maximum margins between them [36]. One drawback of traditional classifiers is that they require existing knowledge as feature input. Features like combinations of k-mers and chemical properties have been used for the prediction of m6A but they are not easily interpretable [33, 34], therefore it needs a prediction model that can extract meaningful biological information like which RBPs appear near m6A sites. In that case, a deep learning model equipped with a motif (sequences recognized by certain proteins) detector is a good choice [37].

In recent years, deep learning models are getting unprecedented attention in the research field owing to its excellent performance with convolutional neural network (CNN) [38] and recurrent neural network (RNN) [39] as being popular models. The CNN model is widely used in the computer vision [40]. Figure 1.4 shows an example of how a CNN model recognizes a picture of a dog. The advantage of CNN model is that the neural network itself can learn how to detect the important features specific to the corresponding classes. With this strength, CNN models have been applied to solve recognition of biological sequence motifs in the computational biology [37, 41, 42] in the way of utilizing the CNN layer for learning position specific scoring matrix(PSSM) [43]. The RNN model including its variation long short-term memory(LSTM) [44] is popular in speech recognition because it is a model designed to utilize sequential information of input data with cyclic connections [43]. Combining CNN with LSTM can enable us to identify biological sequence motifs and capture latent sequential structures, therefore I would like to use deep learning models to extract biological information such as m6A-associated RBPs from m6A miCLIP-Seq data.

In addition, although motifs learned from a deep learning model can help identify potential m6A-associated RBPs, there is limitation for their utility because not all the motifs of RBPs are available, and sequences cannot reflect actual protein binding. Therefore, it also needs to integrate with other kinds of data like RBPs' binding data to extract biological information beyond sequence level.

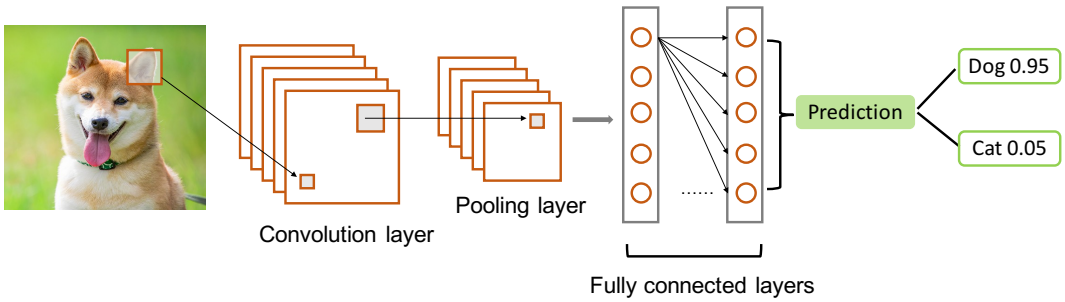


Fig. 1.4 A simplified CNN model for recognition of a dog. The network is composed of a convolution layer, a pooling layer, and fully connected layers. The convolution layer contains multiple filters similar to response of neurons. This layer enables the model to automatically learn the important features specific to the corresponding classes. The pooling layer is used for reducing redundant features. The fully connected layer is to take the results of the previous layers for final classification.

1.2 Research Objectives

The purpose of this thesis is to develop efficient and user-friendly software for detection of m6A signal and extract biological features surrounding m6A by using statistical models and state-of-the-art machine learning algorithms. To achieve these goals, the main contributions of this thesis are summarized as following,

- I developed DeepM6ASeq, a deep learning framework, to predict m6A-containing sequences and characterize biological features surrounding m6A sites. DeepM6ASeq is a sequence-based predictor that showed competitive performance of prediction, learned known m6A readers and a newly recognized one, FMR1, and also helped to visualize locations of m6A sites with a saliency map.

- I developed MoAIMS(model-based analysis and inference of MeRIP-Seq), an efficient and easy-to-use software for analysis of MeRIP-Seq. For detection of m6A regions, MoAIMS achieves excellent speed and competitive performance compared with other tools, and provides user-friendly outputs for downstream analysis. MoAIMS also provides intuitive evaluation on treatment effects for MeRIP-Seq treatment datasets (perturbation of m6A methyltransferases).

- I designed an integrative computational framework for the identification of m6A-associated RBPs from reproducible m6A regions. The framework is composed of an enrichment analysis and a classification model. Utilizing the RBPs' binding data,

the framework is able to identify known m6A-associated RBPs and also found some potential m6A-associated RBPs such as RBM3 for mouse. Besides, it also helps infer interaction between m6A and m6A-associated RBPs like actions of reading and repelling beyond sequence level.

1.3 Dissertation Organization

The thesis is composed of five chapters and the relationship between chapters is summarized in the Figure 1.5.

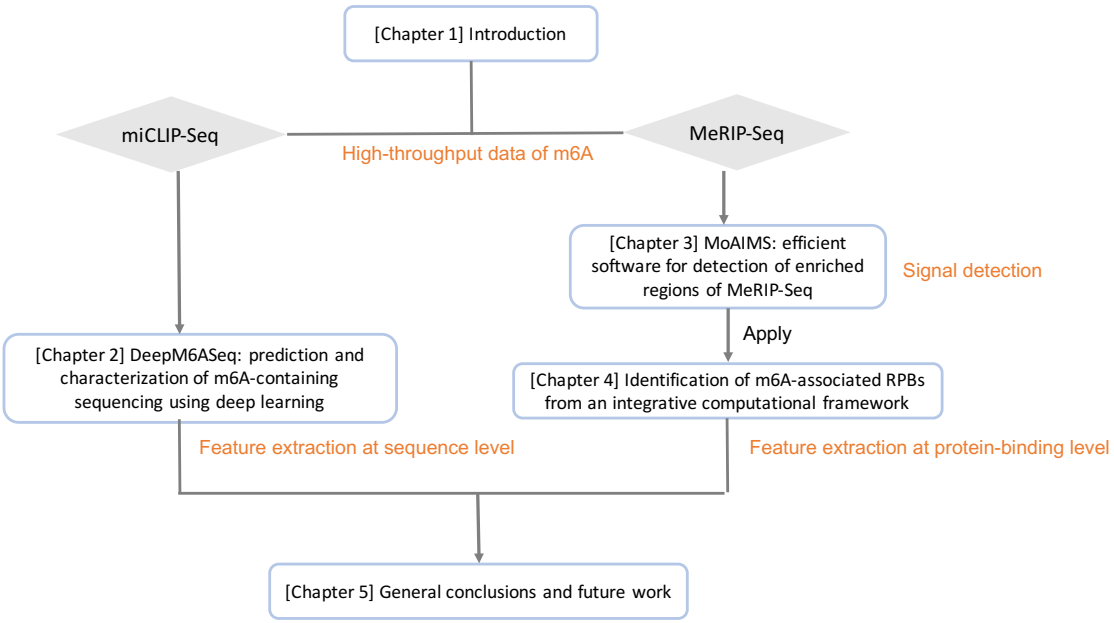


Fig. 1.5 Illustration of the thesis organization.

The content of the chapters are demonstrated briefly in the followings,

[**Chapter 1**] **Introduction** presents the research background of m6A, one of the most abundant RNA modification, including its essential roles in gene regulation with m6A-associated RBPs, its detection methods of high-throughput sequencing technologies, overview of algorithms for m6A data analysis. Next, research objectives are proposed and the contribution of the thesis is described. Finally, the organization of the thesis is demonstrated.

[**Chapter 2**] **DeepM6ASeq: prediction and characterization of m6A-containing**

sequences using deep learning presents the work on m6A biological features extraction at sequence level. Existing tools can predict m6A at single-base resolution, however, the features they used for prediction such as k-mers and chemical properties are less interpretable. It needs a model to extract meaningful biological information like RNA binding proteins in the m6A-containing sequences in order to provide more insights for the biologists, therefore I implemented a deep learning framework, named DeepM6ASeq, to predict m6A-containing sequences and characterize surrounding biological features based on miCLIP-Seq data, which detects m6A sites at single-base resolution. DeepM6ASeq showed better performance as compared to other machine learning classifiers. Moreover, an independent test on MeRIP-Seq data, which identifies m6A-containing genomic regions, revealed that DeepM6ASeq is competitive in predicting m6A-containing sequences. DeepM6ASeq utilized the convolutional neural network(CNN) layer as a motif detector and identified known m6A readers. Notably, DeepM6ASeq also identifies a newly recognized m6A reader: FMR1. Besides, I found that a saliency map in the deep learning model could be utilized to visualize locations of m6A sites.

[**Chapter 3**] **MoAIMS: efficient software for detection of enriched regions of MeRIP-Seq** presents the work on m6A signal detection of MeRIP-Seq(Methylated RNA immunoprecipitation sequencing). MeRIP-Seq is an economical and popular high-throughput sequencing method for studying m6A. The signal detection is a main challenge for data analysis of MeRIP-Seq, however current tools either require a long time or do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling. On the other hand, with more attention on the treatment experiments of MeRIP-Seq, biologists need intuitive evaluation on the treatment effect from comparison. Therefore, efficient and user-friendly software that can solve these tasks must be developed. I developed a software named "model-based analysis and inference of MeRIP-Seq (MoAIMS)" to detect enriched regions of MeRIP-Seq and infer signal proportion based on a mixture negative-binomial model. MoAIMS is designed for transcriptome immunoprecipitation sequencing experiments; therefore, it is compatible with different RNA sequencing protocols. MoAIMS offers excellent processing speed (nearly ten times faster) and competitive performance on motif occurrence in the m6A regions and the overlapping percentage with miCLIP-Seq data when compared with other tools. Furthermore, signal proportion inferred from MoAIMS showed a decreasing trend for m6A treatment dataset (perturbation of m6A methyltransferases) compared with m6A wild-type

dataset, which is consistent with experimental observations, suggesting that the signal proportion can be used as an intuitive indicator of treatment effect.

[**Chapter 4**] **Identification of m6A-associated RBPs with an integrative computational framework** presents the work on m6A biological features extraction at protein-binding level. Existing tools for extracting m6A biological features are sequence-based ones which are limited because not all the motifs of RBPs are available and sequences cannot reflect actual protein binding, therefore in this study I designed an integrative computational framework to extract m6A biological features, i.e. m6A-associated RBPs, utilizing RBP's binding data. I identified reproducible m6A regions from independent studies in certain cell lines and then utilized RBPs' binding data of the same cell line to identify m6A-associated RBPs. The computational framework is composed of an enrichment analysis and a classification model. The enrichment analysis identified known m6A-associated RBPs including YTH domain-containing proteins; it also identified a potential m6A-associated RBP, RBM3, for mouse. I observed a significant correlation for the identified m6A-associated RBPs at the protein expression level rather than the gene expression. In addition, I built a Random Forest classification model for the reproducible m6A regions using the information of RBPs' binding. The RBP-based predictor not only demonstrated competitive performance compared with sequence-based ones and but also helped identify m6A-repelled RBP. These results suggested that the framework enabled us to infer interaction between m6A and m6A-associated RBPs beyond sequence level when utilizing RBPs' binding data.

[**Chapter 5**] **General conclusions and future work** presents general conclusion for the thesis and future work. The future work has main three aspects, including model improvement for the signal detection, algorithms and data worth investigating for biological feature extraction, and application in other RNA modifications.

In summary, in Chapter 2, DeepM6A-Seq was developed to extract biological features at sequence level using miCLIP-Seq data; In Chapter 3, MoAIMS was developed to detect m6A signal from MeRIP-Seq data; In Chapter 4, An integrative computational framework was designed to extract biological features at protein-binding level in reproducible m6A regions which were identified by applying MoAIMS.

Chapter 2

DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning

2.1 Abstract

N6-methyladensine (m6A) is a common and abundant RNA methylation modification found in various species. As a type of post-transcriptional methylation, m6A plays an important role in diverse RNA activities such as alternative splicing, an interplay with microRNAs, and translation efficiency. Although existing tools can predict m6A at single-base resolution, it is still challenging to extract the biological information surrounding m6A sites.

I implemented a deep learning framework, named DeepM6ASeq, to predict m6A-containing sequences and characterize surrounding biological features based on miCLIP-Seq data, which detects m6A sites at single-base resolution. DeepM6ASeq showed better performance as compared to other machine learning classifiers. Moreover, an independent test on m6A-Seq data, which identifies m6A-containing genomic regions, revealed that DeepM6ASeq is competitive in predicting m6A-containing sequences. The learned motifs from DeepM6ASeq correspond to known m6A readers. Notably, DeepM6ASeq also identifies a newly recognized m6A reader: FMR1. Besides, I found that a saliency map in the deep learning model could be utilized to visualize locations of m6A

*Chapter 2 is adapted from the publication [45]

sites.

In conclusion, I developed a deep-learning-based framework to predict and characterize m6A-containing sequences and hope to help investigators to gain more insights for m6A research. The source code is available at https://github.com/rreybeyb/DeepM6ASeq

2.2 Introduction

More than 100 types of RNA modification have been discovered in eukaryotic RNAs [2]; among them, N6-methyladenosine (m6A) is a common and abundant RNA modification type found in various species, such as human, mouse, and yeast [5–7]. m6A is preferentially located near 3' untranslated regions (3' UTR) and its nearby sequences mostly conform to certain motifs, i.e., DRACH (where D = A, G or U; R = A or G; H= A, C or U) in the mammalian genome [8] and RAC in the yeast genome [46]. m6A is involved in diverse RNA activities including alternative splicing [47], an interplay with microRNAs [48] and translation efficiency [49]. In addition, m6A has been linked with cancer progression. It is reported that METTL3 and METTL4, which are both m6A-forming enzymes, have an impact on differentiation and apoptosis of human myeloid leukemia cell lines [24, 50].

m6A can be detected in a high-throughput manner owing to the rapid development of high-throughput sequencing technologies. m6A-Seq and Methylated RNA immunoprecipitation sequencing (MeRIP-Seq) [5, 6] are the main sequencing methods for detection of genomic regions with m6A sites via antibody capturing. Recently, m6A individual-nucleotide-resolution cross-linking and immunoprecipitation (miCLIP-Seq) enables detection of m6A at single-base resolution [8, 29]. Several bioinformatics tools have been developed to predict m6A sites in different species, e.g., m6Apred [51] and iRNA-Methyl [52] for the yeast genome, SRAMP [33] for the mammalian genome. These tools mainly apply existing knowledge as feature input such as a combination of k-mers and chemical properties to build models using random forest (RF) or support vector machine (SVM) algorithm. Although these tools can predict single-base m6A, the biological information surrounding m6As is still limited; this situation poses a challenge for researchers. Therefore, here I implemented a deep-learning-based framework, named DeepM6ASeq, to predict m6A-containing sequences and characterize biological features surrounding m6A. In recent years, deep learning became an state-of-the-art technology and

is now employed more and more in the field of biology [37, 42, 53]. The strength of deep learning is not only in its better prediction power (in comparison with traditional machine learning classifiers), but also its ability to recognize motifs in genomic sequences. Because miCLIP-Seq data revealed precise locations of m6A sites, I explored on such data by utilizing convolutional neural network (CNN) layer as a motif detector to characterize biological features surrounding m6A, then capturing m6A's positional preference out of the deep learning model I built. In addition, I made use of a saliency map to visualize locations of m6A sites in the sequences. The development of DeepM6ASeq, model performance and analysis of biological information will be discussed in details in the following sections.

2.3 Methods

2.3.1 Datasets

The miCLIP-Seq dataset

Given that miCLIP-Seq data can pinpoint m6A sites at single-base resolution, these data provide us with ideal conditions to study sequences surrounding m6A sites. I collected miCLIP-Seq data from human, mouse, and zebrafish [8, 29, 54]. Human and mouse data are from the same source as SRAMP, which included five cell line and tissue types, that is A549(adenocarcinomic human alveolar basal epithelial cells), CD8T(cytotoxic T cells), HEK293(human embryonic kidney 293 cells), brain and liver. For zebrafish, the data consisted of two biological replicates from embryonic stem cells.

For positive samples, I defined sequences with the window size of 101 bp containing m6A sites. First, all m6A sites were mapped to the longest transcripts of genes using the ENSEMBL database (release 91, http://www.ensembl.org/). Then, I randomly located m6A sites in the fixed-size windows and extracted the surrounding sequences with length up to 101 bp (if m6A sites are near a terminus of a transcript, I sliced 101-bp-size windows from the terminus). To avoid sample redundancy (because m6A sites are reported to cluster together [5]), before randomly locating I merged m6A sites within 50 bp and chose the centered one among the merged sites. Because zebrafish data consisted of two replicates, I chose common sites as positive samples.

For negative samples, I used nearby windows (with the same fixed window size) not containing any m6A sites. The nearby negative controls are from the windows 100 bp upstream or downstream the positive windows; these windows are generated by a stride of 10 bp and 100 steps. I chose the closest one for each positive sample. (If there were two closest ones on both sides of a positive sample, I randomly picked one of the two.) In rare cases, there were no control windows nearby because m6A sites are mapped to very short transcripts. Nevertheless, the ratio of positive to negative samples was approximately 1:1. For each species, I split the dataset into an 80% part (as training data) and a 20% part (as independent test data). The dataset information is listed in Table 2.1.

Table 2.1 A summary of dataset size

	Training	Independent test
Human	49050	12611
Mouse	37716	9401
Zebrafish	22108	5651

The m6A-Seq dataset

To test DeepM6ASeq on real peaks data, I used m6A-Seq data from the HepG2 (human liver cancer) cell line and human brain (two different cell types from those used in the model) from Dominissini's study [6] and processed this dataset according to their protocol [55]. For positive samples, I retrieved the top 1000 positive peaks detected by MACS [30] with the highest fold enrichment and the false discovery rate (FDR) ≤ 0.05. I extracted sequences of 101 bp around the peak summits and overlapped these regions with peaks from MeT-DB database [56] (The MeT-DB peak score greater than 6 was required, which is the median score for human data.) to obtain reliable m6A-containing sequences. As negative samples, I used negative peaks detected by MACS (MACS identifies negative peaks by swapping immunoprecipitation samples and control samples) and split each peak into bins with a size of 101 bp (because HepG2 has limited negative peaks, I used a sliding window with a step of 20 bp when splitting peaks for data augmentation). I chose bins overlapping with exon regions and not overlapping with peaks from MeT-DB database. To evaluate the generalization of DeepM6ASeq and to conduct a fair comparison with SRAMP, I used CD-HIT [57] to remove test sequence redundancy with the training data of both DeepM6ASeq and SRAMP at an

80% similarity threshold, which is the lowest threshold provided by CD-HIT. Besides, I kept only sequences with DRACH motifs because SRAMP scans only A sites with DRACH motifs in given sequences. Finally, I got 663 positive samples and 413 negative samples in total.

2.3.2 Models

The development of deep learning models

The sequences were one-hot encoded as inputs with the padding of half filter size on each side, that is, A, C, G, U, and N were encoded as (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1), and (0,0,0,0) respectively. The main structure of the deep learning model consists of two layers of CNN [40], one bidirectional long short-term memory (BLSTM) layer [44], and one fully connected (FC) layer as presented in Figure 2.1. The first convolution layer works as a motif detector, while the second convolution layer captures higher-level features. The BLSTM layer is useful to get sequential-order information embedded in the sequences.

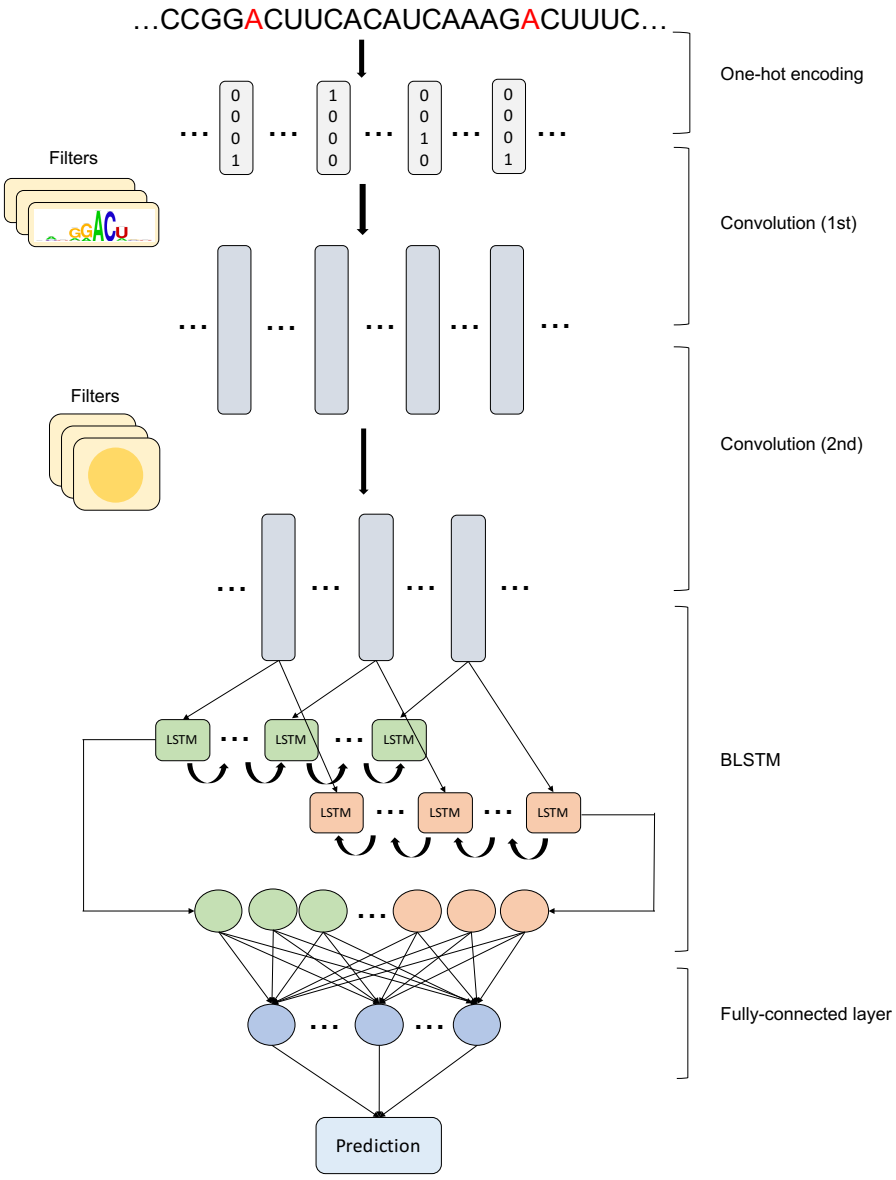


Fig. 2.1 A graphic illustration of DeepM6ASeq model structure. The genome sequence (A in red represents an m6A site) is first one-hot encoded as input, then the input is sequentially fed into two layers of CNN in order. The first CNN layer functions as a motif detector while the second CNN layer captures features of a higher level. After the CNN layers is one BLSTM layer to capture sequential order. The output units of the BLSTM layer are followed by the fully connected layer, and finally the model outputs the prediction result.

During the process of model construction, I chose the filter sizes of 10 and 5, the filter numbers of 256 and 128 for each convolution layer. The activation function for CNN layers is rectified linear unit (ReLU) , tanh for the BLSTM layer, and sigmoid activation after the FC layer to obtain prediction output. Additionally, I applied batch normalization and dropout [58] after each convolutional procedure to accelerate training and avoid overfitting separately. I used binary cross entropy as a loss function to measure the difference between the target and the predicted output and Adam as an optimization algorithm. The deep learning framework is implemented using Pytorch (https://pytorch.org).

There are three phases during the process of model building. First, I performed five-fold cross-validation on training data for optimization of hyperparameters. In this phase, I used the grid-search strategy for optimization of hyperparameters. The details of tuning parameters are given in Table 2.2. Then, I used 1/8 of training data, which equals to 10% of the whole dataset, as validation data and fed the best parameters from the previous phase to the training phase. In the last phase, I applied the model to the independent dataset. I selected a batch size of 256, 50 maximum epochs and an early stopping strategy of patience to 5 in the first two phases.

Table 2.2 Optimization of hyperparameters of DeepM6ASeq

Hyper-parameter	Choices
Maxpooling	0,2
Dropout ratio	0.25,0.5
Output units of BLSTM layer	$32 \times 2, 64 \times 2$
Neurons of FC layer*	32,64
Learning rate	0.01,0.001

*Number of neurons corresponds to 1/2 output units of BLSTM layer

Conversion of filters to motifs

I employed the method from previous papers [37, 42] to convert filters to motifs in position weight matrix (PWM) format. For each input sequence, the subsequence with the filter length that responds to the corresponding filter maximally is extracted in a one-hot encoded matrix, which is then multiplied by the responding score from ReLU in the first CNN layer

as follows

$$M_{l,4}^{(k)} = \sum_{i=1}^{n} \alpha_i^{(k)} X_{l,4}^{(i)} \tag{2.1}$$

where X is the subsequence matrix, α is the responding score, l represents the filter length, k denotes the motif detector, and n is the number of input sequences. The cumulative matrix of these subsequences forms a PWM, each element of which is then normalized as described below

$$m_{p,q} = \frac{m_{p,q}}{\sum_{q=1}^{4} m_{p,q}} \tag{2.2}$$

where m stands for each element in M, and p and q are the row number and column number respectively.

The saliency map

A saliency map is used to determine which nucleotide makes the most contribution to the prediction score for a class (S_c). I calculated the saliency map according to the method described by Lanchantin *et al.* [59]. First, the class score could be approximated with a liner function by computing the first-order Taylor expansion:

$$S_c(X) \approx w(X)^T X + b. \tag{2.3}$$

Then, for a given sequence X in one-hot encoding, the saliency score S was obtained by a point-wise multiplication of the absolute value of a derivative of $S_c(X)$ and its one-hot encoding formally expressed as

$$w(X) = \frac{\partial S_c}{\partial X} \tag{2.4}$$

and

$$S(X) = |w(X)| * X \tag{2.5}$$

Derivation of other classifiers

I built models of RF, Logistic Regression (LR), and SVM on mammalian dataset using sklearn (http://scikit-learn.org). For RF and LR, the feature inputs were normalized counts

of kmers of 1-5. For SVM, the feature inputs were commonly used 4-mer for saving training time. I applied the grid-search strategy on hyperparameter optimization for each classifier and chose the parameters with the best performance. The parameters used in the grid-search were listed in the Table 2.3.

Table 2.3 Hyper-parameters optimization of other classifiers

Classifiers	Hyper-parameters optimization
RF	tree number: 300,400,500,600,700,800,900
LR	regularization strength: 0.001, 0.01, 0.1, 1, 10, 100, 1000; penalty: L1, L2
SVM	C: 1,10,100; gamma: 0.01,0.001; kernel: linear, polynomial, rbf

Evaluation metrics

To measure performance of the models, I calculated accuracy, sensitivity, specificity, the F1-score, and the Matthews correlation coefficient (MCC) as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.6}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{2.7}$$

$$Specificity = \frac{TN}{TN+FP} \tag{2.8}$$

$$F1{-}score = \frac{2TP}{2TP+FP+FN} \tag{2.9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{2.10}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Additionally, I plotted Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves and calculated the areas under the curves, which are denoted by AUROC and AUPR, respectively.

2.4 Results

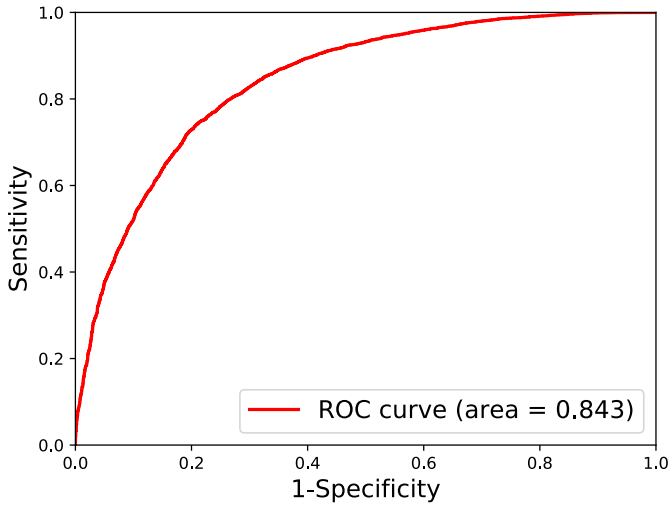
2.4.1 Prediction of m6A-containing sequences

Model training and hyperparameter optimization

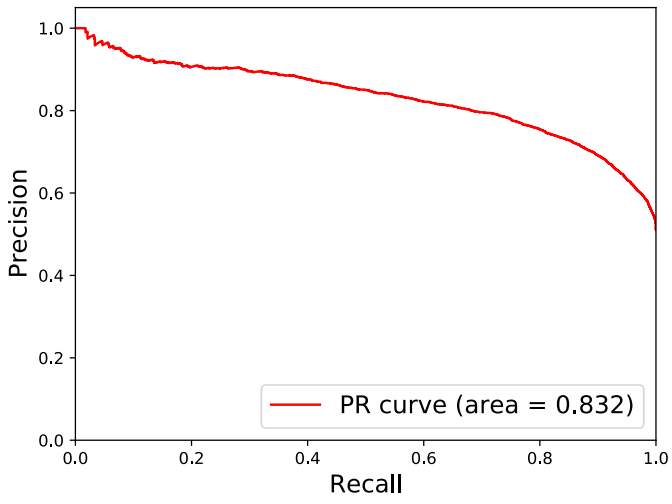
I used the mammalian dataset that consists of both human and mouse miCLIP-seq data, for optimizing the hyperparameters during the development of the model. The details of the model development are described in the Materials and Methods section. In brief, I built a deep-learning-based model that mainly consists of two CNN layers, one BLSTM layer, and one FC layer, to predict whether a sequence contains m6A sites. During hyperparameter optimization, the grid-search strategy was applied to find the best parameter combination of maxpooling size, dropout rate, learning rate, units of the BLSTM layer, and the FC layer. The metrics of mean performance for different parameters settings are shown in the Table 2.4. I found that no maxpooling, a higher dropout rate, and a more complicated model structure contribute to the improvement of performance. Then, I chose the best parameter setting to train the model on the mammalian validation dataset and got AUROC = 0.843 and AUPR = 0.832 for validation as illustrated in the Figure 2.2.

Table 2.4 Metrics of mean performance for hyper-parameters tuning

	Maxpooling	Dropout	Layers	Learning rate	AUROC	AUPR	Accuracy	MCC	F1-score
Model1	0	0.25	BLSTM:32; FC:32	0.001	0.846	0.833	0.768	0.538	0.775
Model2	0	0.25	BLSTM:32; FC:32	0.01	0.847	0.834	0.770	0.541	0.775
Model3	0	0.25	BLSTM:64; FC:64	0.001	0.847	0.835	0.771	0.543	0.777
Model4	0	0.25	BLSTM:64; FC:64	0.01	0.846	0.833	0.769	0.540	0.777
Model5	0	0.5	BLSTM:32; FC:32	0.001	0.848	0.835	0.772	0.544	0.776
Model6	0	0.5	BLSTM:32; FC:32	0.01	0.847	0.834	0.770	0.542	0.778
Model7	0	0.5	BLSTM:64; FC:64	0.001	**0.850**	**0.838**	0.772	**0.546**	0.777
Model8	0	0.5	BLSTM:64; FC:64	0.01	0.849	0.836	**0.773**	**0.546**	**0.780**
Model9	2	0.25	BLSTM:32; FC:32	0.001	0.844	0.830	0.767	0.536	0.776
Model10	2	0.25	BLSTM:32; FC:32	0.01	0.843	0.830	0.768	0.536	0.774
Model11	2	0.25	BLSTM:64; FC:64	0.001	0.844	0.832	0.767	0.535	0.775
Model12	2	0.25	BLSTM:64; FC:64	0.01	0.844	0.831	0.767	0.535	0.772
Model13	2	0.5	BLSTM:32; FC:32	0.001	0.848	0.836	0.771	0.543	0.777
Model14	2	0.5	BLSTM:32; FC:32	0.01	0.846	0.832	0.770	0.541	0.776
Model15	2	0.5	BLSTM:64; FC:64	0.001	0.849	0.836	0.771	0.542	0.776
Model16	2	0.5	BLSTM:64; FC:64	0.01	0.847	0.834	0.770	0.541	0.774



(a)

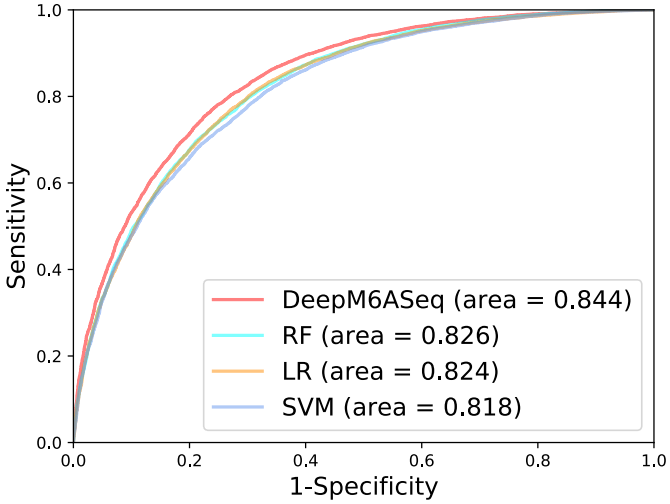


(b)

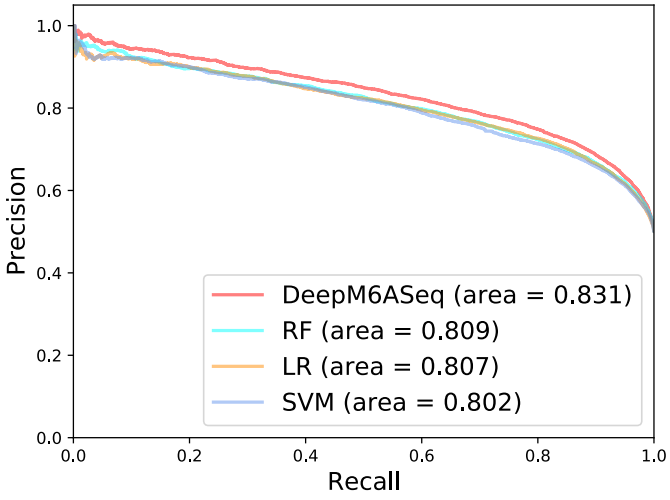
Fig. 2.2 Performance of the mammalian model on the validation dataset. The performance is shown by (a) plot of ROC and (b) plot of precision-recall curve.

The comparison of DeepM6ASeq with other classifiers

I evaluated the mammalian model on the mammalian independent dataset and compared the model with other classifiers, including LR, RF, and SVM. The hyperparameter optimization was performed too for each of these traditional classifiers which as presented in Table 2.3. DeepM6ASeq showed improved performance, with AUROC = 0.844 and AUPR = 0.831 (Figure 2.3). The performance metrics are listed in the Table 2.5 in which DeepM6ASeq ranks first in terms of all the evaluation metrics. To check the statistical significance of the improved performance, I applied the t-test on ROC values from five-fold cross-validation results between DeepM6ASeq and other three classifiers. The mean and standard deviation of ROC values were 0.8504 ± 0.0025, 0.8304 ± 0.0030, 0.8298 ± 0.0031, 0.8258 ± 0.0037 for DeepM6ASeq, RF, LR, and SVM respectively. All the t-test yielded p-value less than 4.5*10e-6, which is indicative of DeepM6ASeq's superiority. Besides, I also tested the mammalian model on an unbalanced mammalian dataset, consisting of the closest nearby windows without any m6A sites on both sides of the positive samples; this arrangement results in the ratio of positives to negatives nearly 1:2. The performance metrics of the mammalian model on the unbalanced independent dataset are compiled in the Table 2.6: DeepM6ASeq showed the stable performance on the unbalanced dataset and still outperformed the other classifiers. The deep learning model has its strengths: it does not require existing knowledge as input and extracts the features automatically, whereas traditional classifiers need predefined features. Additionally, DeepM6ASeq also takes into account the sequential-order information by applying the BLSTM layer. In summary, the results indicate that DeepM6ASeq performs better than the other three algorithms with only sequence-based feature input.



(a)



(b)

Fig. 2.3 The comparison of DeepM6ASeq with other classifiers, including random forest (RF), logistic regression (LR), and support vector machine (SVM), on the mammalian independent dataset. The performance is presented as (a) a plot of ROC and (b) a graph of precision-recall curves.

Table 2.5 Performance metrics for comparison of DeepM6ASeq with other classifiers on the mammalian independent dataset

	Accuracy	F1-score	AUROC	AUPR	MCC
DeepM6ASeq	**0.764**	**0.762**	**0.844**	**0.831**	**0.528**
Random Forest	0.747	0.756	0.826	0.809	0.494
Logistic Regression	0.743	0.736	0.824	0.807	0.487
Support Vector Machine	0.736	0.732	0.818	0.802	0.472

The highest value for each accuracy measure is highlighted in bold.

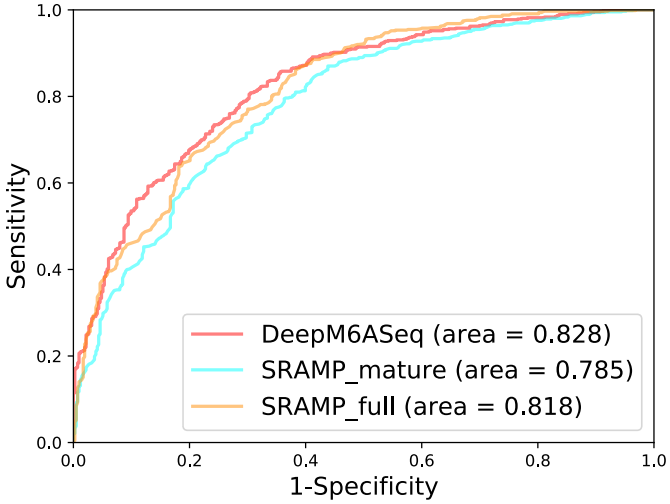
Table 2.6 Performance metrics for comparison of DeepM6ASeq with other classifiers on the mammalian unbalanced independent dataset

	Accuracy	F1-score	AUROC	AUPR	MCC
DeepM6ASeq	**0.763**	**0.687**	**0.841**	**0.725**	**0.505**
Random Forest	0.732	0.667	0.822	0.692	0.466
Logistic Regression	0.750	0.662	0.821	0.694	0.469
Support Vector Machine	0.740	0.654	0.815	0.687	0.453

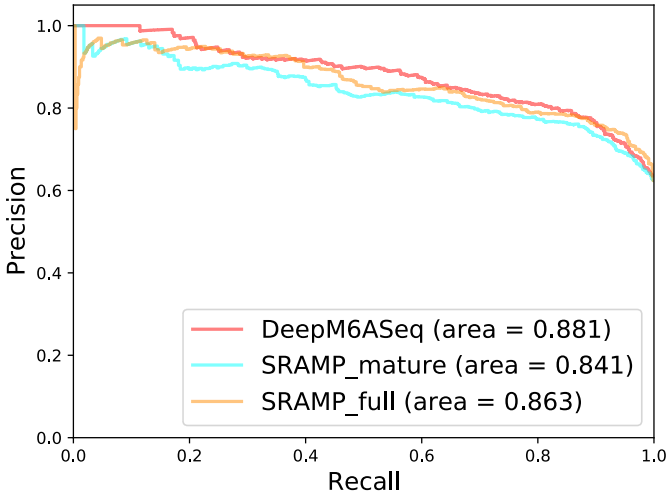
The highest value for each accuracy measure is highlighted in bold.

DeepM6ASeq performance on m6A-Seq data

Given that independent test samples are generated by a stochastic process, I wondered how the model performs on the real m6A-Seq peak data. I retrieved m6A-Seq peak data from HepG2 cell line and human brain (see the Materials and Methods section) and compared the performance of the mammalian model with that of SRAMP, which is also a sequence-based predictor built for the mammalian genome. Both the full mode and mature mode of SRAMP were compared, where the full mode is for whole transcripts and the mature mode for cDNA sequences. I used SRAMP's highest score among all the scores for predicted A sites as the prediction score for a given sequence. DeepM6ASeq showed better performance in terms of AUROC and AUPR as presented in Figure 2.4, and I list performance metrics in Table 2.7. The results indicate that DeepM6ASeq is competitive in predicting m6A-containing sequences.



(a)



(b)

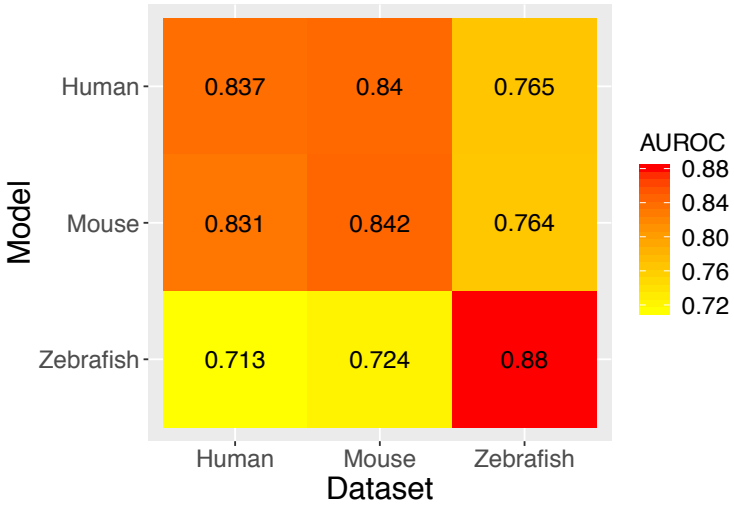
Fig. 2.4 Comparison of DeepM6ASeq with SRAMP in full mode and mature mode (the full mode for whole-transcript sequences and the mature model for cDNA sequences) on the m6A-Seq dataset. The performance is shown as (a) a plot of ROC and (b) a graph of precision-recall curves.

Table 2.7 Performance metrics for comparison of DeepM6ASeq with SRAMP on the m6A-Seq dataset

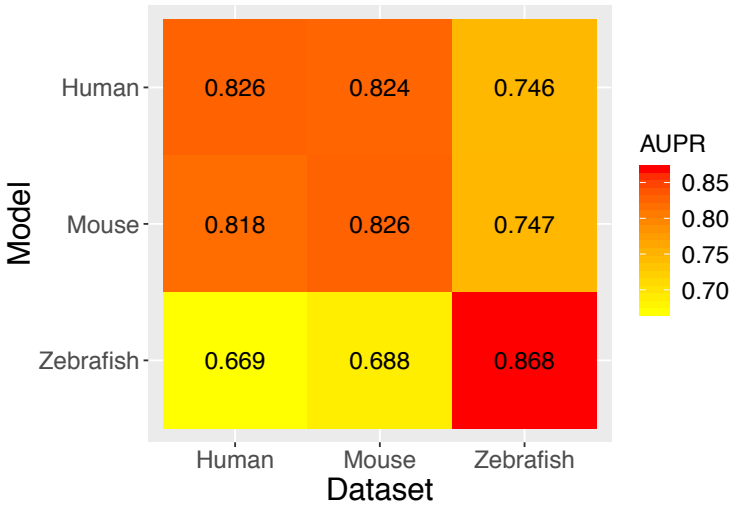
	Accuracy	F1-score	AUROC	AUPR	MCC
DeepM6ASeq	**0.763**	0.808	**0.828**	**0.881**	**0.499**
SRAMP-Mature	0.732	0.787	0.785	0.841	0.428
SRAMP-Full	0.762	**0.824**	0.818	0.863	0.483

Cross-species performance

I built models for human, mouse, and zebrafish separately. The cross-species performance is illustrated in Figure 2.5. As expected, the cross-species prediction was stable between human and mouse; however, there was a gap in the prediction of the mouse and human dataset by the zebrafish model and vice versa. Because the zebrafish dataset is from one cell line, it is possible that models from other species have limitations in terms of generalization due to the cell-line specificity.



(a)



(b)

Fig. 2.5 Cross-species performance. The rows of heatmaps represent a model type and the columns indicate a dataset type. Values in the heatmaps are (a) AUROC (b) AUPR.

2.4.2 Biological information on sequences surrounding m6A sites

Learned motifs for each species

The first CNN layer of the deep learning model is a motif detector, thus I wondered what biological information could be captured by models for different species. The filters of the first CNN layer are converted to the motifs in the ways described in refs. [37, 42], in which I extracted the subsequences with the filter length that respond to the filters maximally from positive training sequences and converted these subsequences to PWMs.These learned motifs were aligned to known motifs using TOMTOM [60]. Under the threshold of E-value=0.05, were 18, 21, 15 out of 256 convolutional filters (7%, 8%, and 6%) corresponding to known motifs for human, mouse, and zebrafish respectively. As depicted in Figure 2.6, among the most significant motifs (E-value $\leq$ 0.01), I found Rbmx (also known as HNRNPG) in both the human and mouse model, which is a known m6A reader [61]. Interestingly, the human predictor detects FMR1, which is a recently discovered m6A reader [19]. FMR1 has been detected in the mouse predictor, albeit not so significant as that in the human predictor (E-value = 0.013). In the zebrafish predictor, the most significant motif was LIN28A, which is one of the core pluripotency regulators. Because the zebrafish data came from embryotic cell line, this outcome is consistent with m6A's role in controlling cell fate development [62]. The results above suggest that DeepM6ASeq could capture meaningful biological information surrounding m6A sites which is also consistent with biological experiments. Furthermore, I used RSAT [63] for clustering motifs and got 161, 158, and 177 clusters separately for human, mouse, and zebrafish (Figure 2.7). The detailed information on motifs and clusters information can be found at https://github.com/rreybeyb/DeepM6ASeq.

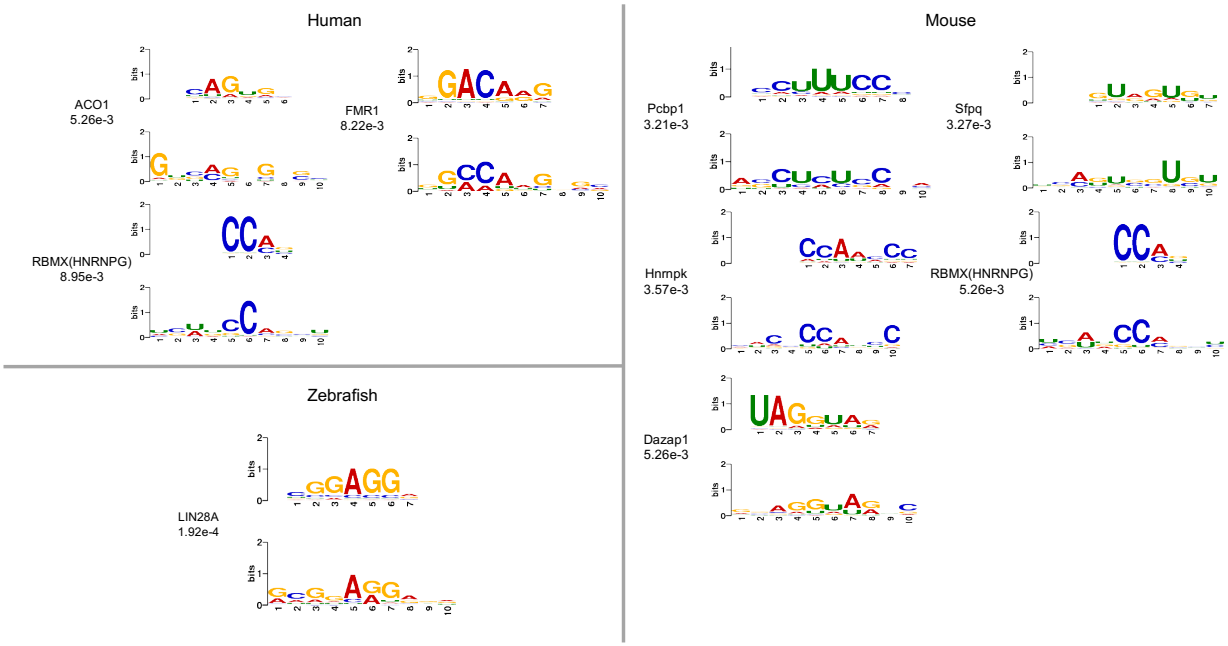
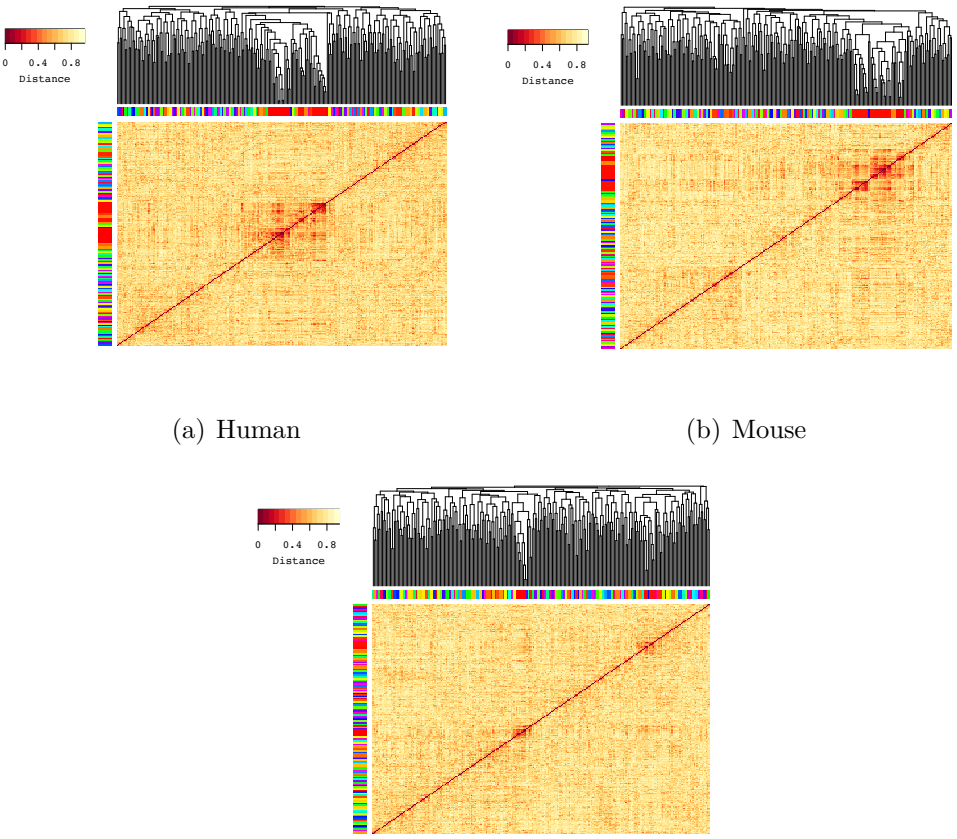


Fig. 2.6 Significant learned motifs (E-value < 0.01) in human, mouse, and zebrafish. The learned motifs from the first CNN layer of each species model are aligned with known motifs by means of TOMTOM. For each aligned result, the upper panel is the known motif, while the bottom panel is the learned motif. The names of known motifs and the significant scores (E-value) are shown on the side.



(a) Human

(b) Mouse

(c) Zebrafish

Fig. 2.7 The clusters of learned motifs from RSAT for (a) Human, (b) Mouse and (c) Zebrafish. The blocks of different colors along the heatmaps represent different clusters.

Location preference for m6A-containing sequences

m6A is characterized by enrichment near 3' UTR of transcripts, thus I wanted to know if DeepM6ASeq could capture such location information. I performed the position analysis in a way without prior knowledge in which I split the transcripts of the independent test dataset into bins of 101-bp size, get bins with confident prediction scores and check if these bins have location preference with regard to the transcript structure. I established three confidence categories (moderate, high, and very high) for prediction scores, which corresponds to 90%, 95%, and 99% specificity respectively in the validation datasets (see Table 2.8).

Table 2.8 Prediction scores at different confidence thresholds for species models

	Moderate*	High*	Very high*
Mammalian	0.725	0.818	0.929
Human	0.772	0.841	0.920
Mouse	0.724	0.813	0.890
Zebrafish	0.715	0.820	0.90

*Moderate, High, and Very high correspond to 90%, 95%, and 99% specificity respectively.

First, I computed the percentage of potential m6A-containing bins with scores above moderate confidence in the bins of the whole transcripts including introns, all exons, and last exons. The result indicated that these potential m6A-containing bins are not enriched in the last exons. This finding suggests that sequences with a potential to contain m6A sites are widely distributed along the exons of transcripts (Figure 2.8).

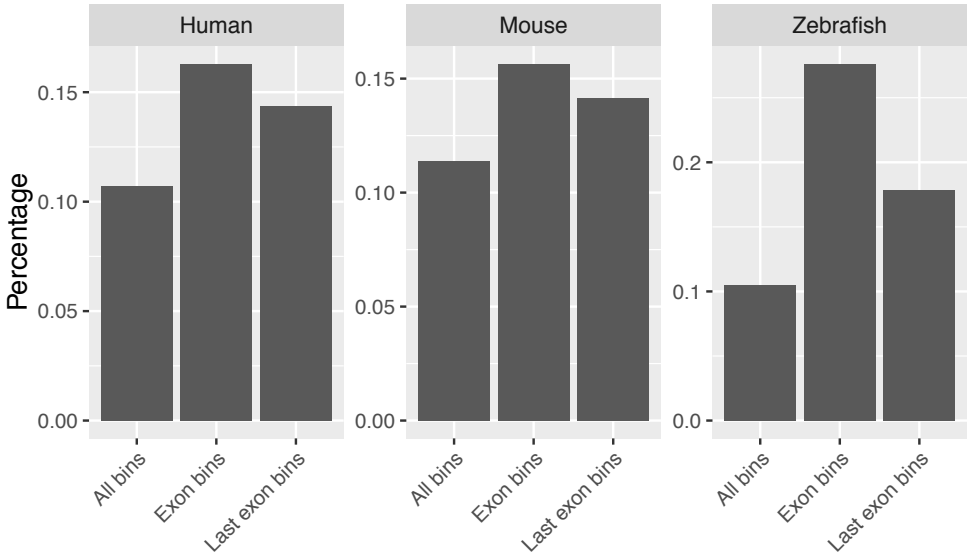


Fig. 2.8 A comparison of percentages of potential bins in different categories for human, mouse, and zebrafish. The X-axis represents different categories, including all bins, exon bins, and last exon bins. Potential bins are the ones with confidence above the moderate threshold.

Then, I checked the relative position of bins of moderate-to-very high confidence in the last exons toward 3' UTR. I profiled the relative distances from the center of these bins to the start of 3' UTR as shown in Figure 2.9. (The distance was normalized to the length of 3' UTR.) The relative distance less than -2 is not shown in the figure because some values are huge owing to the small size of 3' UTR, and because such bins account for less than 3% in the mammal and 7% in the zebrafish. This finding suggests that predicted potential m6A-containing bins were enriched near the start of 3' UTR as the confidence level increased. This result is consistent with the known m6A location bias.

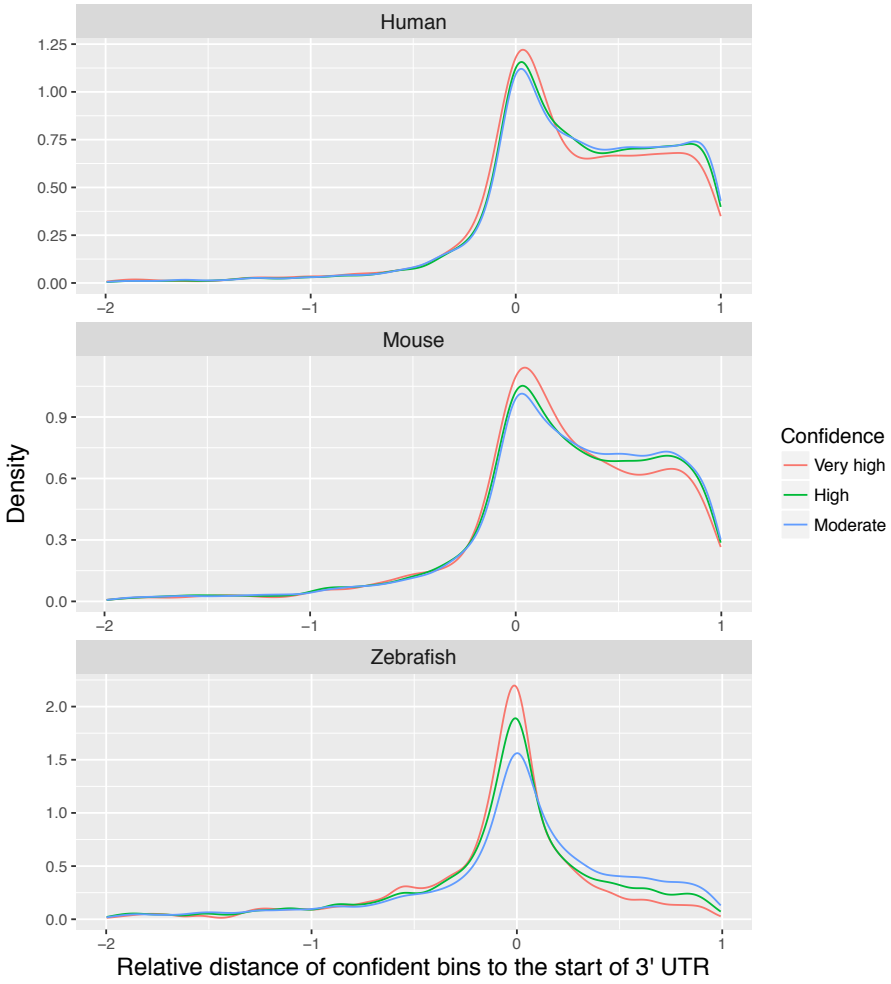


Fig. 2.9 Position profiles of potential m6A-containing bins with a size of 101bp in the last exons for human, mouse, and zebrafish. The X-axis represents the relative distances from m6A-containing bins in the last exons to 3‘ UTR, which is the distance from bins’ center to the start of 3’ UTR normalized to the length of the 3’ UTR. Different colors of lines represent confidence levels from moderate to very high, which corresponds to 90%, 95%, and 99% specificity respectively.

In summary, the location analysis indicates that sequences with a potential to contain m6A sites are widely distributed along the exons of transcripts, in particular, the potential m6A-containing sequences in the last exons are preferentially located near the start of 3' UTR.

The saliency map for visualizing m6A sites

A saliency map is commonly used in computation version for showing each pixels' unique quality. In the context of a genome sequence, a saliency map can measure the nucleotide importance which can have an impact on the prediction scores. Given that I had precise m6A locations from miCLIP-Seq data, I was curious whether locations of m6A sites could be uncovered by way of a saliency map. I obtained saliency maps for potential m6A-containing sequences in the independent datasets with prediction scores with higher-than-moderate confidence via the method described by Lanchantin *et al.* [59], which, in briefly, performs point-wise multiplication of the absolute derivative of the input sequences from back-propagation and their one-hot encoding.

First, I checked the distribution of the types of the most salient nucleotides in the sequences. I extracted the nucleotides with the highest saliency score for each sequence and plotted the distribution. As shown in Figure 2.10, nucleotide type A accounted for the majority among all the most salient nucleotides. For those most salient nucleotides rather than A, I plotted the distribution of the distance from these non-A nucleotides to the closest mapped miCLIP m6A sites as depicted in Figure 2.11, in which the majority of these most salient non-A nucleotides are located near mapped miCLIP m6A sites.

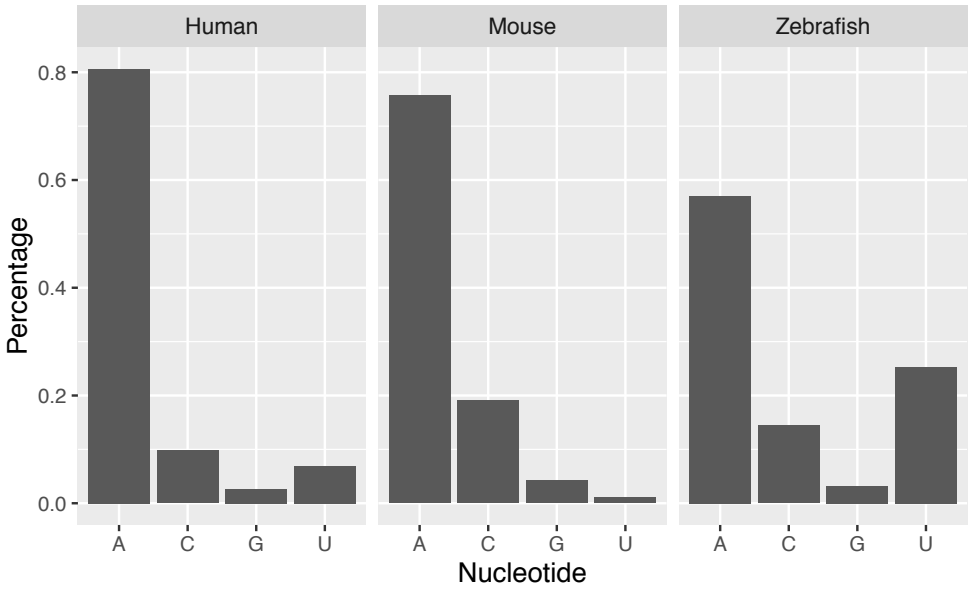


Fig. 2.10 The distribution of nucleotide types of the most salient nucleotides in the independent test sequences with confidence above the moderate threshold for human, mouse, and zebrafish. The X-axis represents the nucleotide types A, C, G, and U.

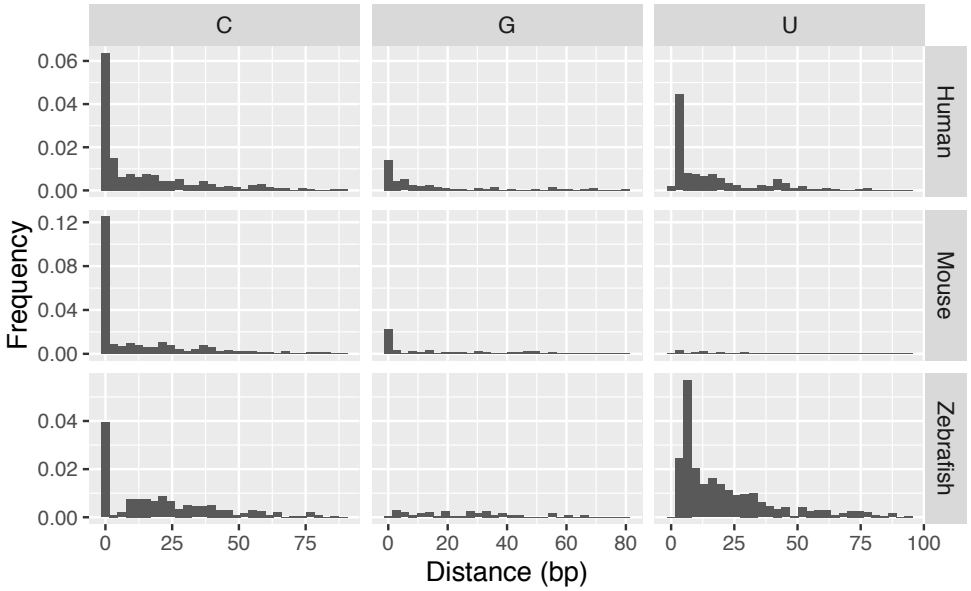


Fig. 2.11 The distribution of distances from the most salient non-A nucleotides in the independent test sequences with confidence above the moderate threshold to mapped miCLIP m6A in human, mouse, and zebrafish. The X-axis denotes the distance.

After that, I wondered how many of these most salient A*s* are overlapped with known m6A sites. It is revealed that nearly 40%–50% of these A*s* belong to known m6A sites from miCLIP-data as shown in Figure 2.12. Besides, some of non-miCLIP m6A could be mapped to the predicted m6A sites in the Met-DB single-base m6A database. Although in zebrafish, the most salient A*s* overlapping neither with miCLIP-Seq data nor Met-DB are more than those in human and mouse, actually, over 30% of these A*s* belongs to the miCLIP m6A sites of one of the replicate zebrafish samples.

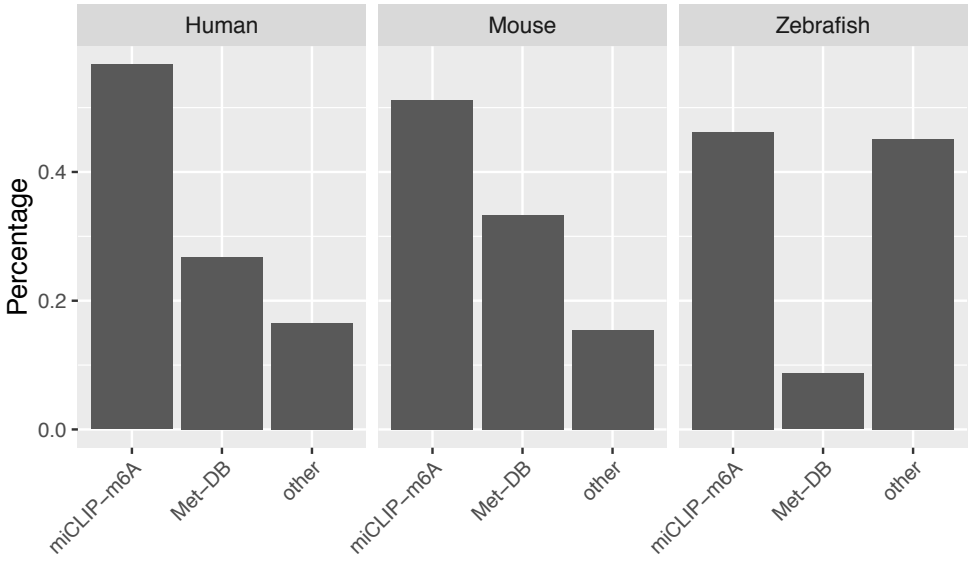


Fig. 2.12 The distribution of the most salient A*s* (in the independent test sequences with confidence above the moderate threshold) by different categories, including miCLIP-Seq m6A, Met-DB predicted m6A and the other in human, mouse, and zebrafish.

Even though most salient nucleotides are overlapped with known miCLIP m6A sites to some extent, I wonder if these known miCLIP m6A sites have higher saliency scores as compared to the other A*s* in the sequences. Thus, I evaluated the ranking percentile of the saliency scores for known miCLIP-Seq m6A sites in the sequences. I found that most of miCLIP m6A sites ranked ahead as shown in Figure 2.13. I also provide examples of visualization of saliency maps as illustrated in Figure 2.14, in which obvious red bands for A*s* are consistent with mapped miCLIP-Seq m6A sites. In the saliency map example for mouse, even though one miCLIP-Seq m6A was missing, I found that this m6A site conforms to a non-DRACH motif and is located between two more significant m6A sites. All the above results indicate that a saliency map could serve as an efficient tool to visualize locations of m6A sites.

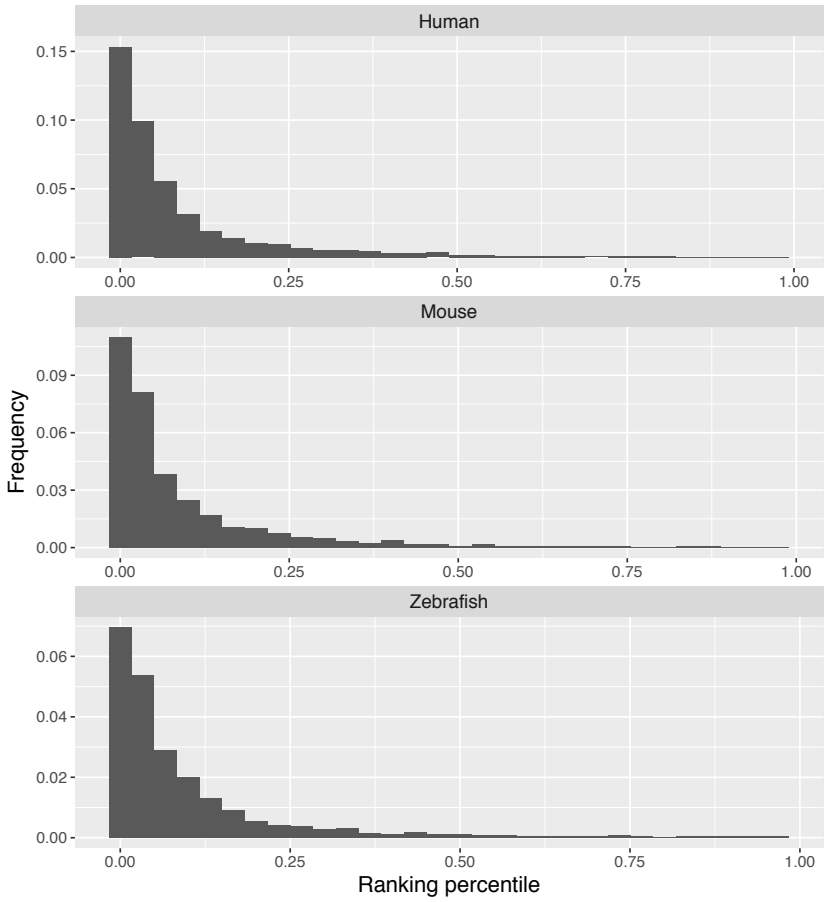


Fig. 2.13 The distribution of ranking percentiles of saliency scores of miCLIP-Seq m6A sites in human, mouse, and zebrafish. The X-axis is the ranking percentile of saliency scores of miCLIP-Seq m6As among those of all the As in the independent test sequences with confidence above a moderate threshold.

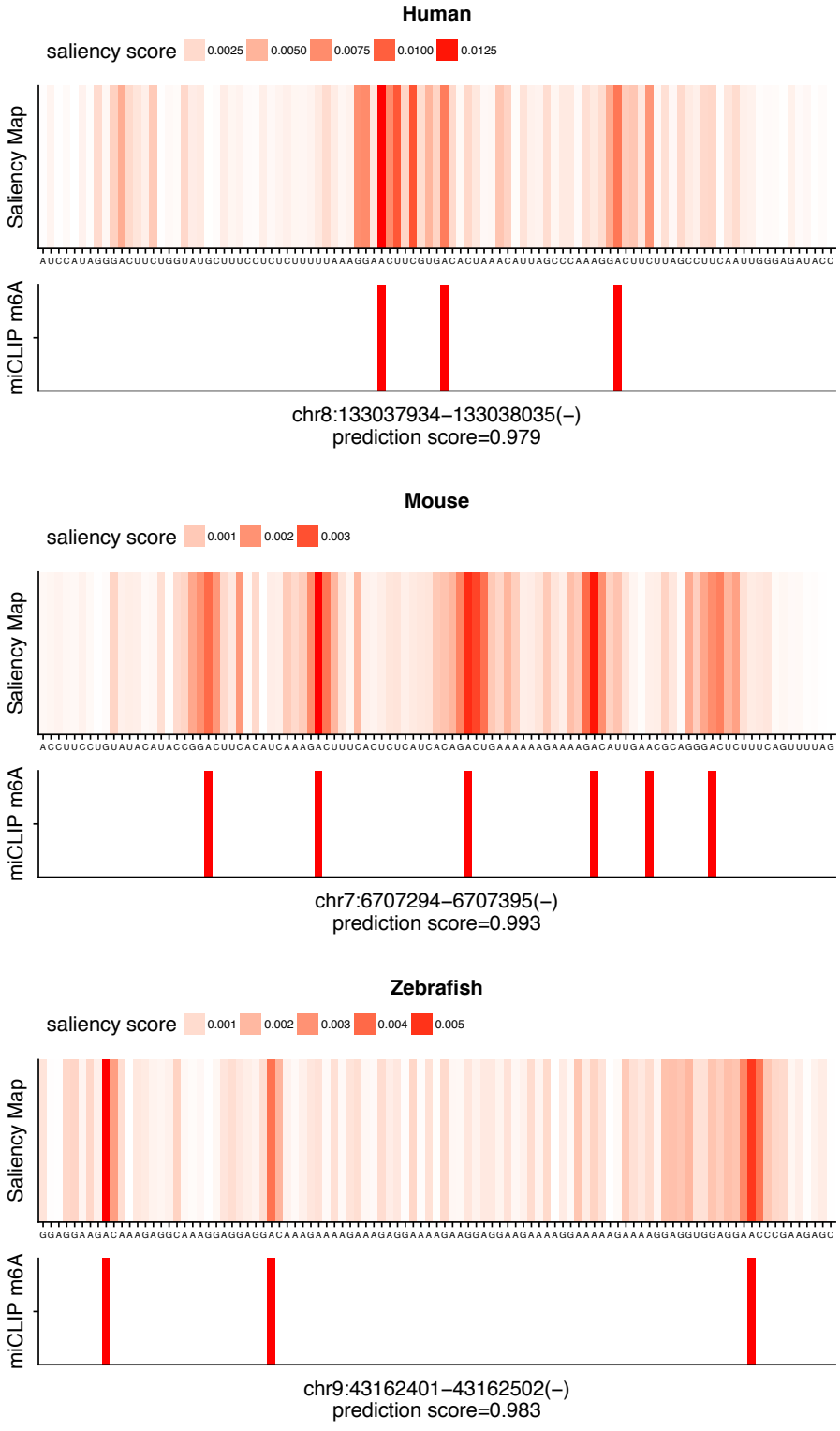


Fig. 2.14 Examples of saliency maps in human, mouse, and zebrafish. For each species, the upper panel presents saliency scores of each nucleotide in the sequence and the bottom panel reveals the locations of mapped miCLIP-m6A sites. The position information and the prediction scores for the sequences are listed at the bottom.

2.5 Discussion and Conclusion

I propose DeepM6ASeq as a framework useful for identifying m6A-containing sequences. Nonetheless, I have some thoughts about the future research. First, although the zebrafish model has higher predictive power, biological information extracted from this model is limited probably due to the single source of the cell type. I expect additional miCLIP-Seq data to become available for zebrafish in the future to improve the current model and provide more biological information. Second, because the second CNN layer detects the combination of motifs at a higher level, it would be interesting to explore what the deep learning model could detect in this layer. An alternative approach is to apply word-embedding, a strategy widely used in the natural language processing. In this way, input sequences can be converted to words and then a deep learning model can be built to discern some patterns among the sequence words. The word-embedding strategy has been utilized for identifying chromatin accessibility [64]. Finally, to characterize biological features surrounding m6A sites in some way without prior knowledge, I employed all the m6A sites rather than being limited to m6A sites with DRACH motifs. I believe that deep leaning method may also exert its power for predicting single-base m6A sites with DRACH motifs, in particular combined with other features such as secondary structure and conservation score.

In conclusion, I developed DeepM6ASeq, a model based on deep learning framework, to predict m6A-containing sequences and characterize biological features surrounding m6A sites. DeepM6ASeq showed better performance as compared to other machine learning classifiers and is competitive at predicting m6A-containing sequences. In addition, DeepM6ASeq can recognize the position preference of sequences harboring m6A sites. All these data corroborate the effectiveness of DeepM6ASeq. Furthermore, taking advantage of function of motif detectors and saliency maps in the deep learning model, DeepM6ASeq learned a newly recognized m6A reader, FMR1 and helped to visualize mapped and potential m6A sites. I hope that DeepM6ASeq will provide more insights for m6A research.

Chapter 3

MoAIMS: efficient software for detection of enriched regions of MeRIP-Seq

3.1 Abstract

Methylated RNA immunoprecipitation sequencing (MeRIP-Seq) is a popular sequencing method for studying RNA modifications and, in particular, for N6-methyladenosine (m6A), the most abundant RNA methylation modification found in various species. The detection of enriched regions is a main challenge of MeRIP-Seq analysis, however current tools either require a long time or do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling. On the other hand, with more attention on the treatment experiments of MeRIP-Seq, biologists need intuitive evaluation on the treatment effect from comparison. Therefore, efficient and user-friendly software that can solve these tasks must be developed.

I developed a software named "model-based analysis and inference of MeRIP-Seq (MoAIMS)" to detect enriched regions of MeRIP-Seq and infer signal proportion based on a mixture negative-binomial model. MoAIMS is designed for transcriptome immunoprecipitation sequencing experiments; therefore, it is compatible with different RNA sequencing protocols. MoAIMS offers excellent processing speed and competitive

*Chapter 3 is adapted from the publication [65]

performance when compared with other tools. When MoAIMS is applied to studies of m6A, the detected enriched regions contain known biological features of m6A. Furthermore, signal proportion inferred from MoAIMS for m6A treatment datasets (perturbation of m6A methyltransferases) showed a decreasing trend that is consistent with experimental observations, suggesting that the signal proportion can be used as an intuitive indicator of treatment effect.

In conclusion, MoAIMS is efficient and easy-to-use software implemented in R. MoAIMS can not only detect enriched regions of MeRIP-Seq efficiently but also provide intuitive evaluation on treatment effect for MeRIP-Seq treatment datasets.

3.2 Introduction

RNA modification refers to biochemical modifications of RNAs that are involved in functional regulations such as translation efficiency and mRNA stability without a change in the RNA sequence, which is also known as epitranscriptome [1]. Over 100 types of RNA modifications have been reported [2]. Among them, researchers have recently focused on certain abundant modifications such as N6-methyladenosine (m6A) [3], N1-methyladenosine (m1A) [66], and 5-methylcytidine (m5C) [67].

With the fast growth of next-generation sequencing (NGS), scientists can study RNA modifications at a whole-transcriptome scale. Methylated RNA immunoprecipitation sequencing (MeRIP-Seq) is a type of NGS technology for studying RNA modifications and is particularly widely used to detect m6A, a modification found in various species including human, mouse, and zebrafish [5, 6, 54]. In MeRIP-Seq, an antibody specific to a certain type of RNA modification (such as m6A or m1A) is used to immunoprecipitate RNA; it is similar to another popular sequencing technology, i.e., ChIP-Seq (Chromatin immunoprecipitation sequencing) [68], which is used in studies of transcription factor binding. However, based on the inherent features of DNA and RNA, there is some difference between MeRIP-Seq and ChIP-Seq data. First, the distribution of ChIP-Seq read counts is relatively uniform while that of MeRIP-Seq is more variable owing to transcript abundance so that MeRIP-Seq requires an input RNA-Seq sample as a control. Second, high duplication rate is often observed in RNA sequencing data due to highly expressed genes, which must be considered in preprocessing MeRIP-Seq data. Third, because RNA

sequencing can store strand information, which provides more accurate transcriptome profiling by strand-specific protocols [69], strand information must be well utilized when analyzing MeRIP-Seq data.

Commonly used tools for identifying enriched regions of MeRIP-Seq include MACS [30], exomePeak [31], and MeTPeak [32]. MACS, which is a popular software in ChIP-Seq analysis, assumes the Poisson distribution for read counts. Applying MACS in MeRIP-Seq analysis requires the genome size to be set [55]; furthermore, because no gene information is considered, the enriched regions contain ambiguous annotations. exomePeak and MeTPeak are both exome-based peak callers that also assuming the Poisson distribution for read counts, and MeTPeak is developed based on exomePeak by integrating a hidden Markov Model (HMM). Although these two tools are exome-based, they do not process strand-specific and paired-end cases and are time consuming. Besides, with more attention on the treatment experiments of MeRIP-Seq, these tools cannot satisfy the need for intuitive evaluation on the treatment effect from the comparison.

To facilitate the analysis of MeRIP-Seq, I developed "model-based analysis and inference of MeRIP-Seq (MoAIMS)," which is efficient and user-friendly software designed for transcriptome immunoprecipitation sequencing. MoAIMS can detect enriched regions and infer the signal proportion of MeRIP-Seq based on a mixture negative-binomial(NB) model. It is compatible with different RNA sequencing protocols including paired/single-end and non-strand/strand-specific sequencing. The results demonstrated the excellent processing speed (it only takes several minutes to finish analysis of one dataset) and competitive performance of MoAIMS compared with other tools. When MoAIMS is applied to studies of m6A, the detected enriched regions contain known biological features of m6A. Furthermore, MoAIMS can provide an intuitive indicator of treatment effect for treatment experiments. The signal proportion inferred from MoAIMS for m6A treatment datasets (perturbation of m6A methyltransferases) showed a decreasing trend, consistent with experimental observations. Finally, functional analysis on the m6A perturbation datasets reveals the interplay between m6A and histone modification. In conclusion, I developed efficient and user-friendly software for MeRIP-seq analysis.

3.3 Implementation

A MeRIP-Seq dataset consists of one immunoprecipitation (IP) sample and one input sample (used as control). MoAIMS takes aligned IP and input bams as input. Aligned bams are generated from pre-processing as shown in the workflow of MeRIP-Seq analysis (Figure 3.1). In the pre-processing, reads are aligned to a target genome by transcriptome-based aligners such as STAR [70], Tophat [71], and HISAT [72]. Only uniquely mapped reads are kept. Then, reads are sorted and marked for duplication using PicardTools [73] or samtools [74]. Given the RNA sequencing protocol (single-end or paired-end, strand-specific or not) and a target genome annotation (in GTF format), MoAIMS is ready for analysis. Typically, MoAIMS requires several minutes to complete the analysis of one MeRIP-Seq dataset. The primary outputs of MoAIMS contain enriched regions (in BED12 format), goodness of fitting (GOF) plot (Figure 3.2), and a summary table of the fitted models (Table 3.1). The source code and the user's manual are available at https://github.com/rreybeyb/MoAIMS

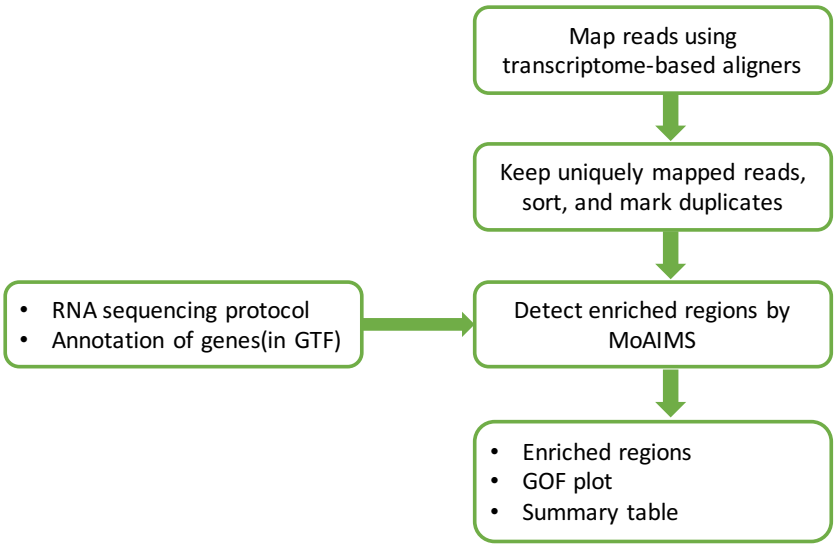
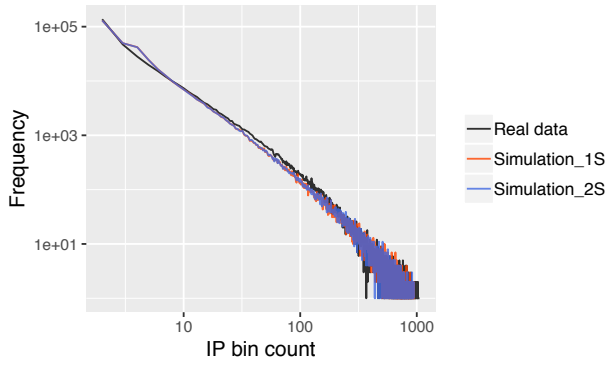
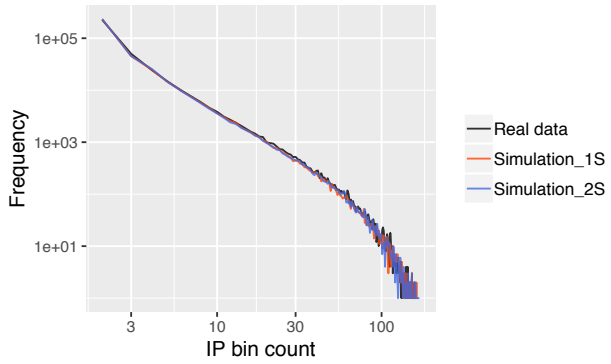


Fig. 3.1 Workflow of MeRIP-Seq analysis using MoAIMS. Reads are pre-processed through alignment, sort (by coordinates), and mark-duplication. Given the RNA sequencing protocol and annotation of genes in GTF format, MoAIMS is ready for analysis. The primary outputs include detected enriched regions (in BED12 format), goodness of fitting (GOF) plots, and a model summary table.



(a) Human example



(b) Mouse example

Fig. 3.2 Examples of goodness of fitting (GOF) plots for a human and a mouse dataset. X-axis is bin count and Y-axis is frequency. Real data, simulation data of 1S (one-signal) mode, and simulation data of 2S (two-signal) mode, are plotted in black, red, and blue lines, respectively.

Table 3.1 An example of the model summary table

Dataset	π_s	BIC_1S	BIC_2S	optim_k	optim_reg
WT_rep1	0.138	1679168	1678590	2	rlm
WT_rep2	0.11	1212063	1212005	2	rlm

The columns represent dataset names, signal proportion, BIC values for 1S (one-signal) mode, BIC values for 2S (two-signal) mode, optimized k, and optimized regression methods.

In the analysis performed by MoAIMS, it firstly obtains transcriptome bins by concatenating all exons for the expressed genes. Then, it uses featureCounts for counting reads in the bins. Subsequently, it models the distribution of the bin counts by a mixture NB distribution and detects the enriched regions. The details are described as follows.

3.3.1 Read counts of bins

Counting reads in bins was performed for the transcriptome of expressed genes because unexpressed genes provide little information for signal detection. The default threshold for expressed genes is 0.5 TPM(transcripts per million). All exons for the expressed genes were concatenated and split into bins with size 200 bp(default setting). Subsequently, featureCounts [75] was used for counting reads in the bins. The parameters used in featureCounts include the following: requireBothEndsMapped=TRUE (for paired-end sequencing), read2pos=5, ignoreDup=T, allowMultiOverlap=T.

3.3.2 Model construction

A negative-binomial mixture model

MoAIMS implements and extends the statistical framework proposed by MOSAiCS [76], which is used to detect ChIP-Seq enriched regions and cannot be directly applied to MeRIP-Seq data because it is designed for processing DNA Sequencing and models the bin counts on the whole-genome scale. The statistical framework uses the negative-binomial to model the distribution of background reads distribution (bin counts) in an individual sample. It assumes that the observed bin counts of an IP sample follows a mixture negative-binomial

model composed of a background component and a signal component that are unobserved. Let Z represent the components, where $Z \in \{0,1\}$ (0 for the background component and 1 for the signal component) and Y_j is the observed read count of the jth bin; therefore, the mixture model can be written as Equation(3.1),

$$P(Y_j) = (1-\pi_s)P(Y_j|Z_j = 0, \Theta_B) + \pi_s P(Y_j|Z_j = 1, \Theta_s), \tag{3.1}$$

of which π_s is the *signal proportion*($\pi_s \in [0,1]$), i.e. the proportion of bins from the signal component with s representing signal, equal to $P(Z_j = 1)$; $(1-\pi_s)$ is equal to $P(Z_j = 0)$; Θ_B and Θ_s are parameters of background and signal distribution respectively.

When the bin is from the background component, the read count follows the distribution $NB(a, \frac{a}{a+\mu_j})$, with a the size parameter and $\frac{a}{a+\mu_j}$ the probability parameter of the NB distribution. When the bin is from the signal component, the read count is represented as $Y_j = N_j + S_j + k$, where N_j is the count from a non-specific background following $NB(a, \frac{a}{a+\mu_j})$, S_j is the count from an actual enrichment, and k is the minimal read count required for the signal component. Thus, the distribution of the signal component is a convolution of negative binomials. There are two modes for the distribution of S_j, named one-signal (1S) mode and two-signal (2S) mode. In the 1S mode, S_j follows $NB(b, \frac{c}{c+1})$ ($c = \frac{b}{\mu}$, μ is the mean). Details of the distributions are provided in the Appendix A. In the 2S mode, S_j follows a mixture NB distribution, i.e. $\pi_{s1} NB(b_1, \frac{c_1}{c_1+1}) + (1-\pi_{s1}) NB(b_2, \frac{c_2}{c_2+1})$, with π_{s1} ($\pi_{s1} \in [0,1]$) representing the first signal proportion. MoAIMS implements 2S mode from MOSAiCS. The difference between 1S mode and 2S is that 1S mode assumes a single NB distribution for the actual enrichment while 2S model assumes a mixture NB distribution that considers the complexity of the signal component.

Parameters estimation

The parameters of NB to be estimated in the model are represented as $\Theta = \{\Theta_B, \Theta_{s1}, \Theta_{s2}\}$, where $\Theta_B = (a, \mu_j)$ for the background component, $\Theta_{s1} = (b, c)$ for the signal component in 1S mode and $\Theta_{s2} = (b_1, c_1, b_2, c_2)$ for the signal component in 2S mode.

First, the parameters of the background component, $\Theta_B = (a, \mu_j)$ are estimated. μ_j is the expected bin counts of the jth bin and estimated by regression using the input bin count data. A simple illustrative figure for the regression process is shown in Figure 3.3. The detailed explanation is described as follows.

Gene1:IP	1	2	5	1	1
Gene1:input	3	3	3	3	3

Gene2:IP	2	2	3	2
Gene2:input	6	6	6	6

Steps

- Group IP count(Y) according to input count(X), x_i is the group value:
  $\mathcal{S}_1 = \{Y = 1, 2, 5, 1, 1 | X = x_1 = 3\}$
  $\mathcal{S}_2 = \{Y = 2, 2, 3, 2 | X = x_2 = 6\}$
- Get median Y value of each group
  $\mu_1 = E(\mathcal{S}_1) = 1$
  $\mu_2 = E(\mathcal{S}_2) = 2$
- To estimate the mean background count of each bin from IP sample, regression is performed using x_i as the predictor variable and $\mu_i$as the response variable

Fig. 3.3 A simple illustration on estimating background means. Tables list the IP and input bin counts for Gene1 and Gene2 as examples. Cells with numbers represent bins. The following steps show how to estimate the mean background count of each bin from in sample in brief.

Each IP bin count Y_j has a corresponding input bin count X_j. For the bins from the background component, it is assumed that $\{Y_j\}(j = 1, 2, ..., T)$ with the same input bin count from the same distribution; thus, $\{Y_j\}$ are grouped by the input bin count to $\mathcal{S}_i = \{Y_j | X_j = x_i\}$ (x_i is the group value equal to available and unique bin count value, i.e. 0,1,2,..., in input sample and i is the group index) in the transcriptome-wide. Figure 3.3 gives an example of how to estimate background means. For $Y_j \in \mathcal{S}_i$, it follows that $NB(a, \frac{a}{a+\mu_i})$. In another word, the expected bin counts μ_j of Y_j is decided by the group $\mathcal{S}_i$ which Y_j belongs to. Subsequently, regression is performed with x_i as the predictor variable and μ_i(euqal to $E(\mathcal{S}_i)$, the median value of $Y_j \in \mathcal{S}_i$. Median is a robust and efficient estimator because there are both background and signal bins in an IP sample.) as the response variable. For the regression method, MOSAiCS uses the weighted robust fitting of linear model (RLM) [77] with the function $\log(\mu_i) = \beta_0 + \beta_1 \log(x_i)$, of which β_0 and β_1 are the coefficients. However, in some cases of RNA sequencing, I found that the generalized additive model (GAM) [78] can provide better fitting as shown in Figure 3.4. GAM is a flexible model that uses a sum of unspecified smooth functions $\sum_{s=1}^{G} f_s(v_s)$ to replace the linear form $\sum_{s=1}^{G} \beta_s v_s$ in the generalized linear model where v is predictor variable and G is the number of predictor

variables. Here, I used only one predictor variable, that is, the input bin count. Therefore, when using GAM, μ_i can be estimated by $\log(\mu_i) = \beta_0 + f(\log(x_i)|\boldsymbol{\beta})$, where f is represented using smoothing splines and $\boldsymbol{\beta}$ is a vector of coefficients for the spline term with length of 9 as default. I implemented GAM using R package mgcv [79] and set the restricted maximum likelihood [80] as the method for estimating the smoothing parameters. To optimize the model, MoAIMS implements both RLM and GAM and subsequently uses that with a lower BIC (Bayesian Information Criterion) [81]. BIC scores were calculated in the general method by $r\ln(T) - 2\ln(\hat{L})$, where r is the number of parameters, T the number of bins, and $\hat{L}$ the maximum likelihood.

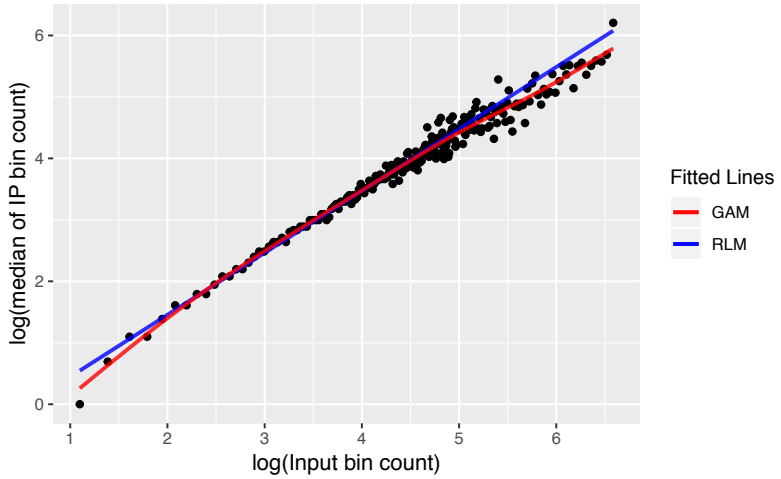


Fig. 3.4 Comparison of generalized additive model (GAM) and robust fitting of linear model (RLM) regression for estimating the distribution means of the background component. X-axis is the available and unique read count for input sample and Y-axis is the median read count of the IP bins of which corresponding input bins have the same read count. Both are log transformed.

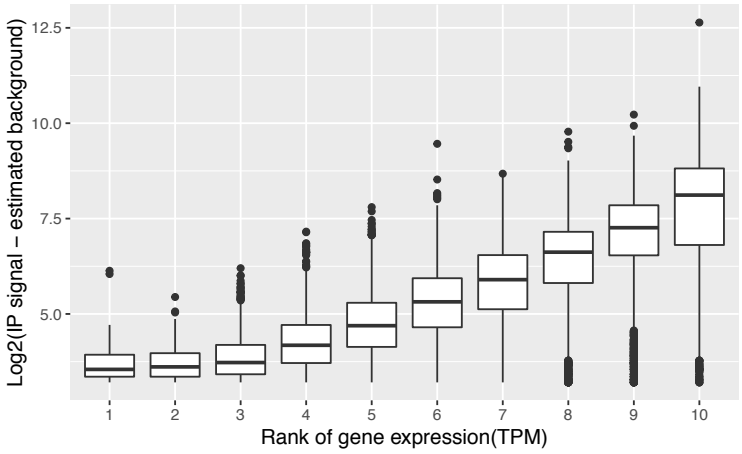
The size parameter a is a weighted value estimated by $\hat{a} = \sum_i n_i \hat{a}_i / \sum_i n_i$, where $\hat{a}_i = [E(\mathcal{S}_i)]^2/[Var(\mathcal{S}_i) - E(\mathcal{S}_i)]$ (i is the group index; the expectation is calculated using median value; the variation is calculated using the median absolute deviation) and n_i is the number of bins.

Then, after estimating the parameters of the background component, the parameters of the signal component in 1S mode, $\Theta_{s1} = (b, c)$, and π_s are estimated using expectation maximization (EM) algorithm [82]. For the parameters b and c, the method of moments is used as MOSAiCS. For π_s, it is estimated in the maximization step with optimized k value rather than based on a pre-defined k value in MOSAiCS. The details of modified EM process

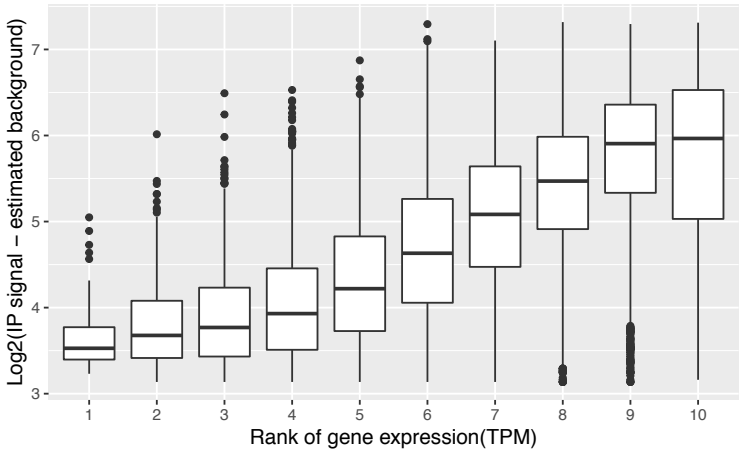
for 1S mode are provided in the Appendix A. Finally, the parameters of the signal component in 2S mode, $\Theta_{s2} = (b_1, c_1, b_2, c_2)$, and π_{s1} are estimated in the way of MOSAiCS.

Model design for MeRIP-Seq analysis

I developed MoAIMS based on the statistical framework proposed by MOSAiCS and made some modification and extension in order that MoAIMS is more suitable for MeRIP-Seq analysis. This framework applies the negative-binomial distribution that is commonly used to deal with the overdispersion, and uses input read counts as covariates to estimate background means in the IP sample, which is an efficient way that helps provide information like transcript abundance and regional bias. In addition, to check if signal detection is influenced for genes with higher expression, I plotted the residuals between IP signal and estimated background corresponding to the gene expression. As Figure 3.5 shows, IP signal increases as the gene expression increases, which indicates that the model can detect enrichment for highly-expressed genes.



(a) Human shGFP_rep1



(b) Mouse WT_rep1

Fig. 3.5 Diagnostic plots of a human dataset shGFP_rep1 and a mouse dataset WT_rep1 for the residuals between IP signal and estimated background corresponding to the gene expression. X-axis is the rank of gene expression(TPM) from lowest to highest. Y-axis is the residuals between IP signal and estimated background in log2 scale.

The modification and extension involved three aspects. First, I used log-transformation in estimating the background means instead of power-transformation in MOSAiCS. Log-transformation is often used for covariates with skewed distribution in the regression analysis [83]. It can simplify the parameter tuning required in power transformation. Second, I set k, the minimum count in the signal regions, flexible instead of pre-defined in MOSAiCS. Because k may depend on the library size and signal-to-background ratio of the experiments [84] , I set k flexible and optimized in the model fitting. With the optimized k, the signal proportion (π_s) was estimated by EM rather than based on a pre-defined k value in MOSAiCS. Third, in addition to the RLM used by MOSAiCS in estimating background means, I applied GAM for regression to obtain better fitting for some cases of RNA sequencing data, as shown in Figure 3.4. An example of summary table of the fitted models is shown as Table 3.1 that provides signal proportion, BIC values for 1S (one-signal) mode, BIC values for 2S (two-signal) mode, optimized k, and optimized regression methods.

3.3.3 Detection of enriched regions

The enriched regions were decided under the threshold of the false discovery rate (FDR), which was calculated as in [84, 85]. In this study, false discovery means a genomic region that is claimed to be significant when it is not. For a set $\mathcal{M}$ of m enriched regions that satisfies a defined cut-off (default is 0.05), the estimated FDR is equal to $(1/m)\Sigma_{j\in\mathcal{M}}P(Z=0|Y_j)$, where $P(Z=0|Y_j)$ is equal to $\frac{(1-\hat{\pi}_s)\hat{p}_{0,j}}{(1-\hat{\pi}_s)\hat{p}_{0,j}+\hat{\pi}_s\hat{p}_{1,j}}$ for the 1S mode and $\frac{(1-\hat{\pi}_s)\hat{p}_{0,j}}{(1-\hat{\pi}_s)\hat{p}_{0,j}+\hat{\pi}_s[\hat{\pi}_{s1}\hat{p}_{1,j}+(1-\hat{\pi}_{s1})\hat{p}_{1,j}]}$ for the 2S mode with $\hat{p}_{0,j}$ and $\hat{p}_{1,j}$ as the post probability for the jth bin from the background component and the signal component respectively. Finally, the enriched regions were merged and output in the BED12 format with the highest bin count of merged regions as the score, which can be used as a filter to obtain higher confident signal region candidates.

3.3.4 Goodness of fitting (GOF)

To display the goodness of fitting (GOF), the simulations is performed using the estimated parameters. For the simulation of the 1S mode, m background bins and n signal bins were randomly sampled according to π_s, where $m+n=T$. The background read count of T bins were generated from the background distribution $NB(a, \frac{a}{a+\mu_j})(j=1,...,T)$. Subsequently,

for n signal bins, the read count was composed of the background read count, the count sampled from the signal distribution $NB(b, \frac{c}{c+1})$, and the minimal count k. For the simulation of the 2S mode, m background bins, $n1$ first-signal bins, and $n2$ second-signal bins were randomly sampled according to π_s and π_{s1}, where $m + n1 + n2 = T$. The background read count of T bins were generated from the background distribution $NB(a, \frac{a}{a+\mu_j})(j = 1, ..., T)$. Subsequently, for the signal bins, the read count was composed of the background read count, the count sampled from the corresponding signal distribution $NB(b_1, \frac{c_1}{c_1+1})$ or $NB(b_2, \frac{c_2}{c_2+1})$, and the minimal count k. Figure 3.2 gives examples of GOF plot.

3.4 Results

3.4.1 Comparison with other tools

Detection of m6A-enriched regions

I performed analysis on two m6A MeRIP-Seq studies. One is from mouse embryonic stem cell [62] that uses the single-end and strand-specific sequencing protocol. The mouse datasets include the wild type and knock-out of Mettl3 (an m6A methyltransferase), of which each has two biological replicates. The other is from human A549(adenocarcinomic human alveolar basal epithelial cells) cell line [86] that uses the paired-end and strand-specific sequencing protocol. The human datasets contain negative control (shGFP) and perturbation of three types of m6A methyltransferases including Mettl14, Mettl3, and WTAP, of which each has two replicates. Table 3.2 summarized the information of datasets. Raw fastq files were retrieved from Gene Expression Omnibus [87] with accession numbers GSE52662 and GSE54365. Reads were aligned to human (hg19) and mouse (mm10) genome using STAR (version 2.6.0c, default setting) [70] with annotation files of GENCODE (human release19 and mouse release M19) [88]. Only uniquely mapped reads were kept. The sorted (by coordinates) and duplication-marked bam files were generated by Picard (version 2.18.1) and subsequently used as input for MoAIMS.

Table 3.2 Information of MeRIP-Seq datasets

Name	Species	Type	Replicates
WT	Mouse	Wild	2
KO_Mettl3	Mouse	Treated	2
shGFP	Human	Negative control	2
shWTAP	Human	Treated	2
shMettl3	Human	Treated	2
shMettl14_1	Human	Treated	2
shMettl14_3	Human	Treated	2

Three commonly-used tools for comparison are MACS(version MACS2), exomePeak(v2.13.2), and MeTPeak(v1.0.0). Duplication-removed bam files were used as input for the three tools. For MACS, I specified parameters "–nomodel –extsize=100 –keep-dup=all -g 286,000,000 (for human)/221,000,000 (for mouse)'. 'I kept the peaks called by MACS overlapped with exonic regions for comparison. For exomePeak and MeTPeak, I used the default setting.

First, I compared the m6A-enriched regions called by MoAIMS with MACS, exomePeak, and MeTPeak. I verified to what extent the enriched regions called by the four tools agree with each other using BEDTools [89]. To obtain higher confident regions, I chose the enriched regions (FDR$\leq$0.05) called by MoAIMS with score $\geq$10. Table 3.3 shows the results for the mouse wild-type datasets. Each cell of the table represents the percentage of enriched regions of tools in the columns detected by tools in the rows; the number in bracket is the number of enriched regions called by each tool. It is indicated that enriched regions of MoAIMS are overlapped more with MACS and exomePeak. Additionally, MeTPeak called relatively less peaks and, in some cases, could miss enriched regions, as shown in Figure 3.6.

Table 3.3 Consistency of enriched regions called by MoAIMS, MACS, exomePeak, and MeTPeak

Mouse WT_rep1	MoAIMS	MACS	exomePeak	MeTPeak
MoAIMS	100(11869)	72.9	64.8	71.0
MACS	81.3	100(14681)	69.2	80.9
exomePeak	96.4	96.6	100(19698)	91.5
MeTPeak	44.0	48.4	37.8	100(9049)

Mouse WT_rep2	MoAIMS	MACS	exomePeak	MeTPeak
MoAIMS	100(9411)	81.3	61.7	69.7
MACS	88.2	100(11161)	66.8	70.8
exomePeak	97.8	98.1	100(16190)	85.5
MeTPeak	61.9	61.9	49.7	100(10133)

WT_rep1 and WT_rep2 are two replicates of wild-type mouse datasets. Each cell is shown in percentage(%) and the number in bracket is the number of enriched regions.

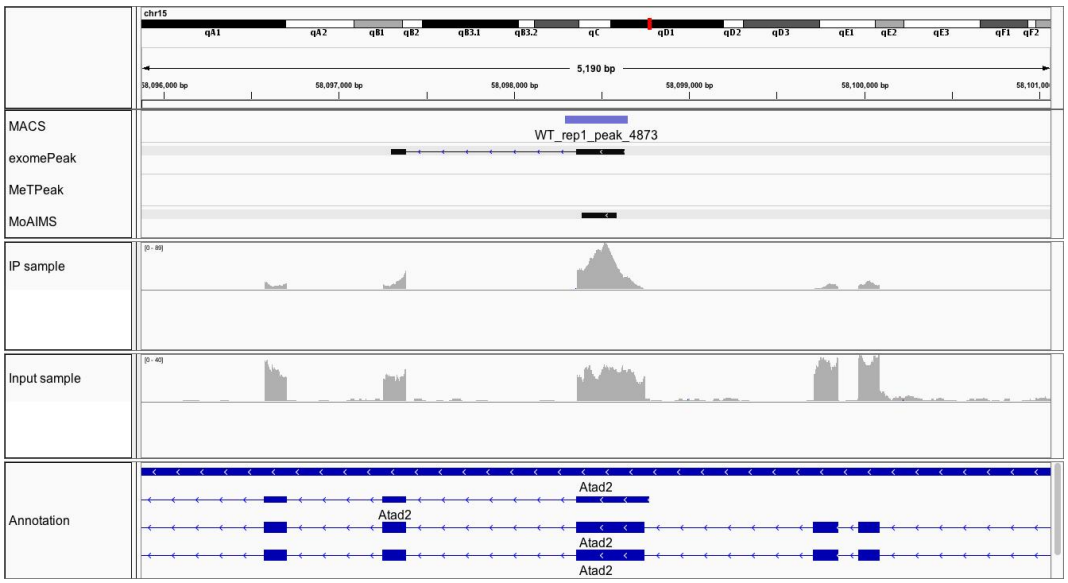
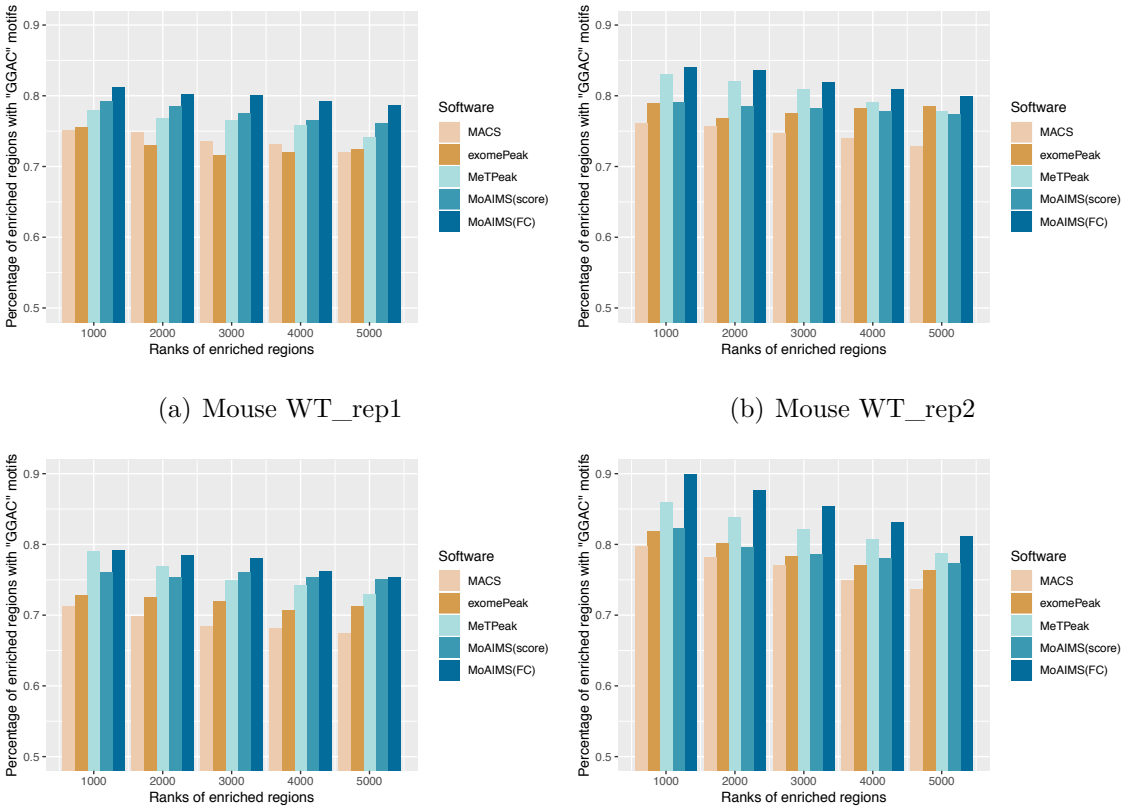


Fig. 3.6 Example of an enriched region missed by MeTPeak. The plot is generated using IGV showing the enriched region called by MACS, exomePeak, MeTPeak, and MoAIMS in the first four tracks. The following tracks are the coverage for IP and input sample and the genome annotation.

Subsequently, I verified the occurrence of the core m6A motif 'GGAC' [90] in the top-5000 enriched regions. The ranking scheme for MACS, exomePeak, and MeTPeak is fold change. For MoAIMS, the ranking scheme of fold change(FC) and score are both used for comparison. Sequences of length 200 bp were extracted around the summits of the enriched regions. For MACS, I used the summits it provided; for MoAIMS, exomePeak, and MeTPeak, the summits were defined as the positions with the highest read coverage. Because I had the strand-specific sequencing data, I only counted the motifs that occurred in the expressed genes with coverages (for MACS, only motifs with coverages were counted). Figure 3.7 compares the percentage of motif occurrence in the decreasing peak ranks for two wild-type mouse datasets and two human negative control datasets. The results indicated that MoAIMS achieved comparable performance to the other three tools.



(a) Mouse WT_rep1

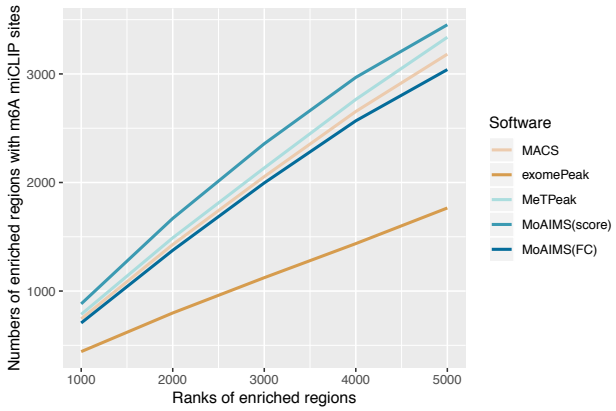
(b) Mouse WT_rep2

(c) Human shGFP_rep1

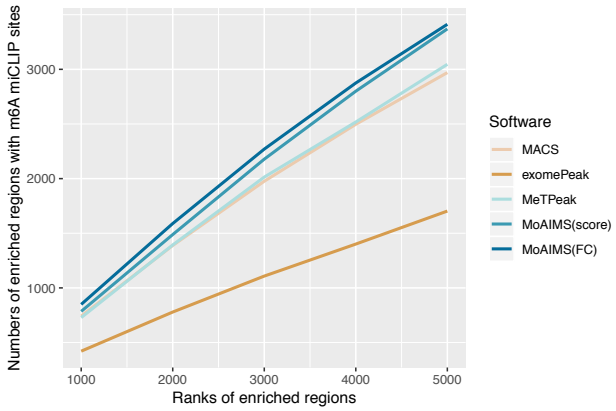
(d) Human shGFP_rep2

Fig. 3.7 Comparison of 'GGAC' motif occurrence for MACS, exomePeak, MeTPeak, and MoAIMS for datasets: (a) WT_rep1 and (b) WT_rep2 (two replicates of mouse wild type), (c) shGFP_rep1 and (d) shGFP_rep2 (two replicates of human negative control). X-axis is the decreasing rank of the enriched regions from the top 1000 to top 5000. The ranking scheme for MACS, exomePeak, and MeTPeak is fold change. For MoAIMS, the ranking scheme of fold change (FC) and score are both used for comparison. Y-axis is the percentage of motif occurrence.

Next, I was interested to know to what extent the m6A miCLIP sites agree with the MeRIP-Seq enriched regions. I collected miCLIP-Seq data of human A549 cell line from [29], which maps m6A sites at single-base resolution.I counted the number of regions containing miCLIP sites in the top-5000 enriched regions detected by the four tools (The ranking scheme is the same as that for counting motif occurrence). MoAIMS with score ranking has the most number of regions with m6A miCLIP sites in the decreasing peak ranks in Figure 3.8 (a) while has the second most in Figure 3.8 (b). To determine whether the number was affected by the length of the enriched regions, I compared the length of the top-5000 enriched regions between the tools, as shown in Table 3.4. The result shows that compared with MeTPeak, which ranks second with regard to consistency with miLCIP sites, MoAIMS can detect more regions with m6A miCLIP sites under the similar resolution.



(a) Human shGFP_rep1



(b) Human shGFP_rep2

Fig. 3.8 Comparison of top enriched regions with m6A miCLIP sites called by MACS, exomePeak, MeTPeak, and MoAIMS for two human negative control datasets. X-axis is the decreasing rank of the enriched regions from the top 1000 to top 5000. The ranking scheme for MACS, exomePeak, and MeTPeak is fold change. For MoAIMS, the ranking scheme of fold change (FC) and score are both used for comparison. Y-axis is the number of enriched regions with m6A miCLIP sites.

Table 3.4 Length comparison of top-5000 enriched regions

Dataset	MoAIMS(score)	MoAIMS(FC)	MeTPeak	exomePeak	MACS
shGFP_rep1	400(400)	400(473)	399(458)	297(386)	244(297)
shGFP_rep2	400(549)	600(653)	300(365)	300(400)	221(271)

shGFP_rep1 andshGFP_rep2 are two replicates of the human negative control datasets. Each cell represents the median length, and the number in bracket is the mean length. The ranking scheme for MACS, exomePeak, and MeTPeak is fold change. For MoAIMS, the ranking scheme of fold change(FC) and score are both used for comparison.

Features of MoAIMS

MoAIMS is efficient software with appealing features, as shown in Table 3.5. Thus, I performed comparison analysis with regard to those features. First, because MoAIMS is compatible with general RNA sequencing protocols in counting reads, I investigated how the methods of counting reads affected the detection of enriched regions for pair-end RNA sequencing. The comparison was conducted for the human shGFP (negative control) datasets among exome-based callers: MoAIMS, exomePeak, and MeTPeak. Table 3.6 lists the number of enriched regions detected by these three tools using pair-end reads and first-in-pair reads, separately. The result indicates that exomePeak and MeTPeak differ in the method of counting paired-end reads, while the difference is limited for MoAIMS.

Table 3.5 Features of MoAIMS compared with other tools

Features	MoAIMS	exomePeak	MeTPeak	MACS
Exome-based	Y	Y	Y	N
Strand-specific/Paired-end	Y	N	N	N
Time-consuming	N	Y	Y	N
Inference of signal proportion	Y	N	N	N
Visualization of model fitting	Y	N	N	N
Output in BED12 format	Y	Y	Y	N
Support for differential methylation analysis	N	Y	N	N

Table 3.6 Comparison of methods of counting reads in bins for pair-end sequencing

	MoAIMS		exomePeak		MeTPeak	
Dataset	Pair-end	First-read-in-pair	Pair-end	First-read-in-pair	Pair-end	First-read-in-pair
shGFP_rep1	14137	15319	24009	17573	14478	9213
shGFP_rep2	21603	24300	26741	18418	13610	7401

shGFP_rep1 and shGFP_rep2 are two human negative control datasets. Pair-end means using both reads in pair-end sequencing as input, while first-read-in-pair means using only the first read in pair-end sequencing. Each cell shows the number of enriched regions.

Next, MoAIMS is a strand-aware caller; thus, it can avoid calling ambiguous regions that are overlapped with other regions on different strands. Figure 3.9 shows an example of how MoAIMS called strand-specific enriched regions. As shown in the figure, a protein-coding gene Mtmr10 and an antisense gene RP23-84M17.2 are partially overlapped. The coverage track in red (colored by strand) indicates the signal in Mtmr10, not the antisense gene. For this case, exomePeak and MeTPeak have callings on both genes, but MoAIMS can avoid the ambiguous callings.

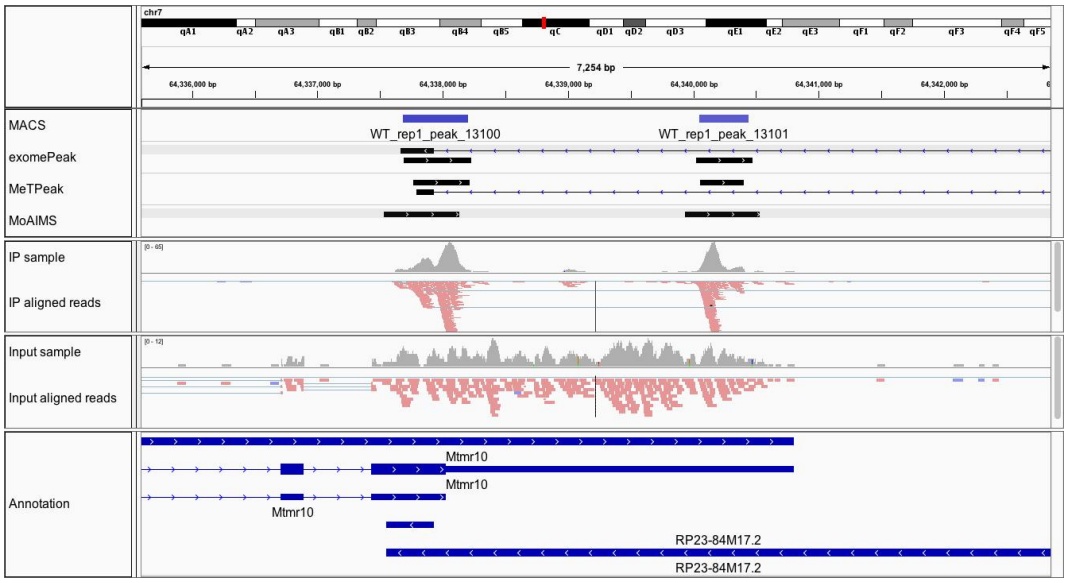


Fig. 3.9 Example of detection of strand-specific enriched regions. The plot is generated using IGV [91], showing the enriched region called by MACS, exomePeak, MeTPeak, and MoAIMS in the first four tracks. The following tracks are coverage and aligned reads (strand orientation is colored) for the IP and input sample, respectively, and the genome annotation.

Finally, MoAIMS offers excellent processing speed compared with exome-based callers exomePeak and MeTPeak, which require approximately 2 hours to analyze one dataset (MeTPeak needs even more time because it applies HMM). Table 3.7 lists the time cost for a human and a mouse dataset, indicating that MoAIMS is competitive as it only requires several minutes and can yield comparable performance.

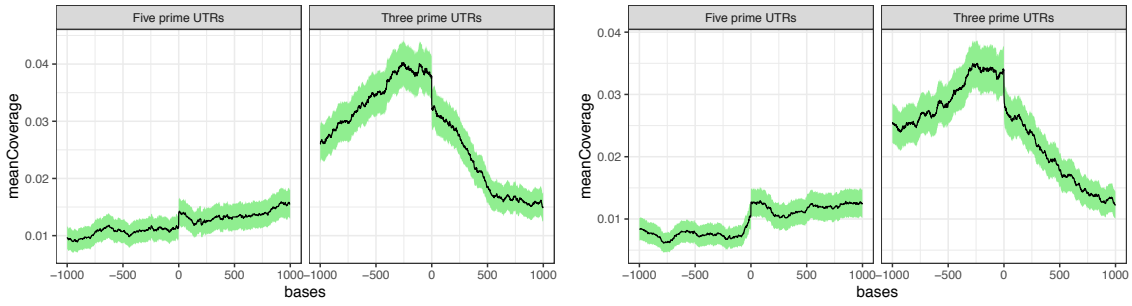
Table 3.7 Performance on the time cost

Dataset	MoAIMS	exomePeak	MeTPeak
Human shGFP_rep1	14.1	141.0	176.4
Mouse WT_rep1	10.6	110.4	143.4

shGFP_rep1 is one human negative control dataset. WT_rep1 is one wild-type mouse dataset. The units of time is minute.

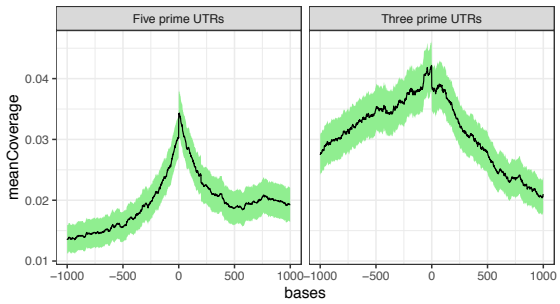
3.4.2 Application on feature and functional analysis of m6A

m6A is characterized by its location preference close to three prime untranslated regions (3' UTRs); thus, I verified the position preference of the enriched regions (with score $\geq$10) called by MoAIMS. For the wild-type mouse datasets, as shown in Figure 3.10 (a) and (b), the enriched regions exhibit location bias near 3' UTRs, which is consistent with the results of the original study [62]. For the human negative control datasets, I observed that enriched regions appeared near 5' UTRs, as shown in Figures 3.10 (c) and (d), which agrees with the findings of the original study [86] regarding methylated m6A at transcription start sites.

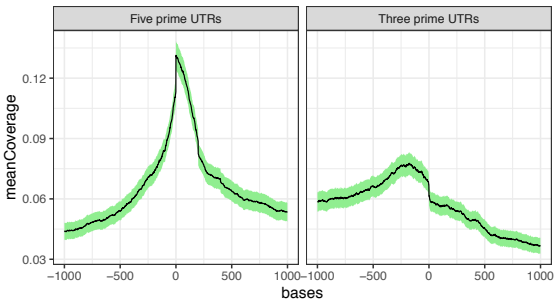


(a) Mouse WT_rep1

(b) Mouse WT_rep2



(c) Human shGFP_rep1



(d) Human shGFP_rep2

Fig. 3.10 Position profile of m6A-enriched regions for the following datasets: (a) WT_rep1 and (b) WT_rep2 (two replicates of mouse wild type), (c) shGFP_rep1, and (d) shGFP_rep2 (two replicates of human negative control). X-axis is the relative position coordinates and Y-axis is the mean coverage of the enriched regions. The plot is generated using RCAS [92].

Because MoAIMS infers the signal proportion from the mixture NB model, I assumed that this value can reflect the treatment effect; for example, the knocking-down/out of methyltransferases (such as WTAP, METTL3, or METTL14) can cause decreased signal proportion. For the mouse datasets, as shown in Figure 3.11, Mettl3 knock-out exhibits a clear decreasing trend for signal proportion, which agrees with the findings of a recent study [93] that include a discussion on the m6A methyltransferase treatment experiments and the effect of treatment in this dataset. For the human datasets, as shown in Figure 3.12, WTAP shows a relatively clear effect after perturbation, while Mettl3 and Mettl14 shows less effect. This trend is consistent with the original study [86], in which the authors observed the necessity of WTAP for m6A methylation, while perturbation of Mettl3 and Mettl14 exhibited milder effects in decreasing methylation level. These results suggest that the signal proportion can be used as an intuitive indicator of the m6A treatment effect, which can facilitate biologists' evaluation on the treatment experiments.

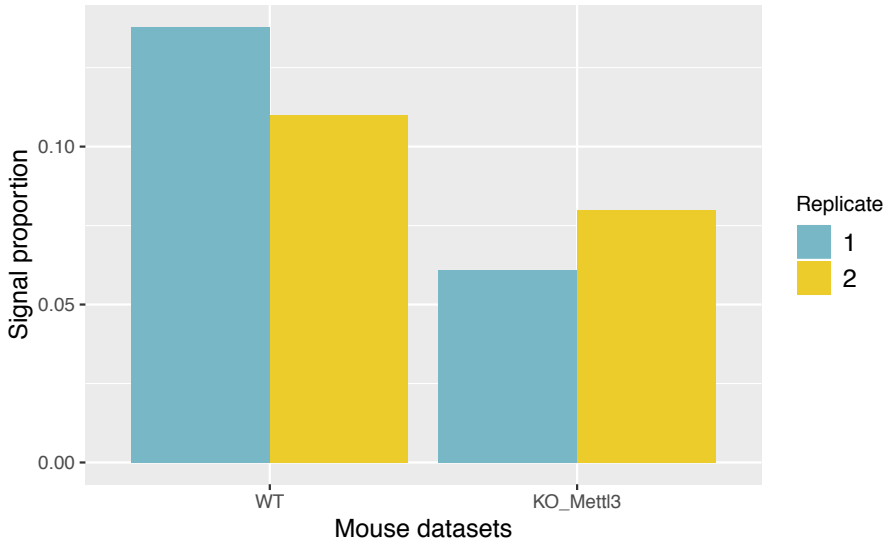


Fig. 3.11 Signal proportion for m6A treatment experiments. X-axis represents MeRIP-Seq datasets, i.e. mouse wild type (WT) and knock-out of METT13 (KO_Mettl3) with blue for replicate 1 and yellow for replicate 2. Y-axis represents the signal proportion.

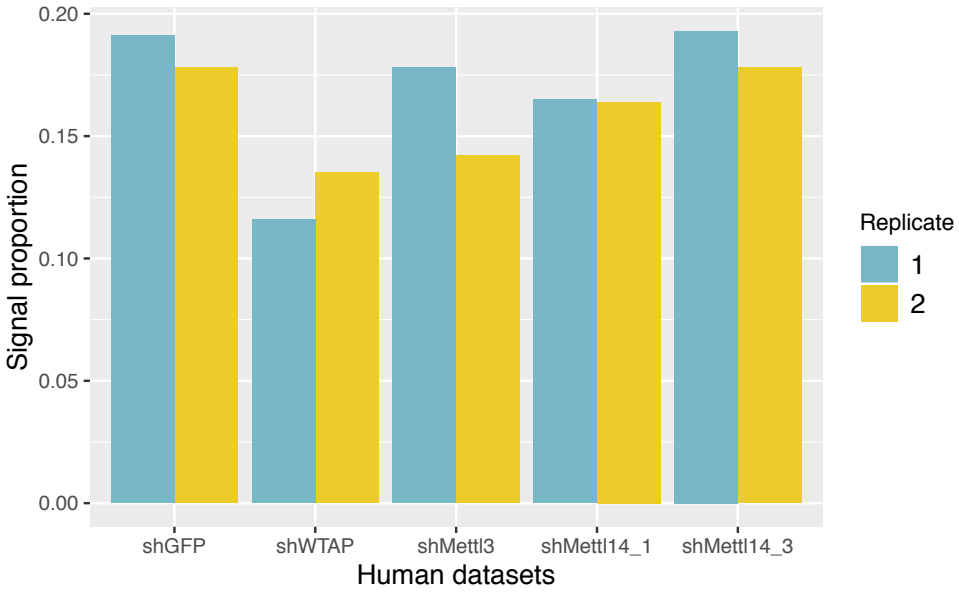
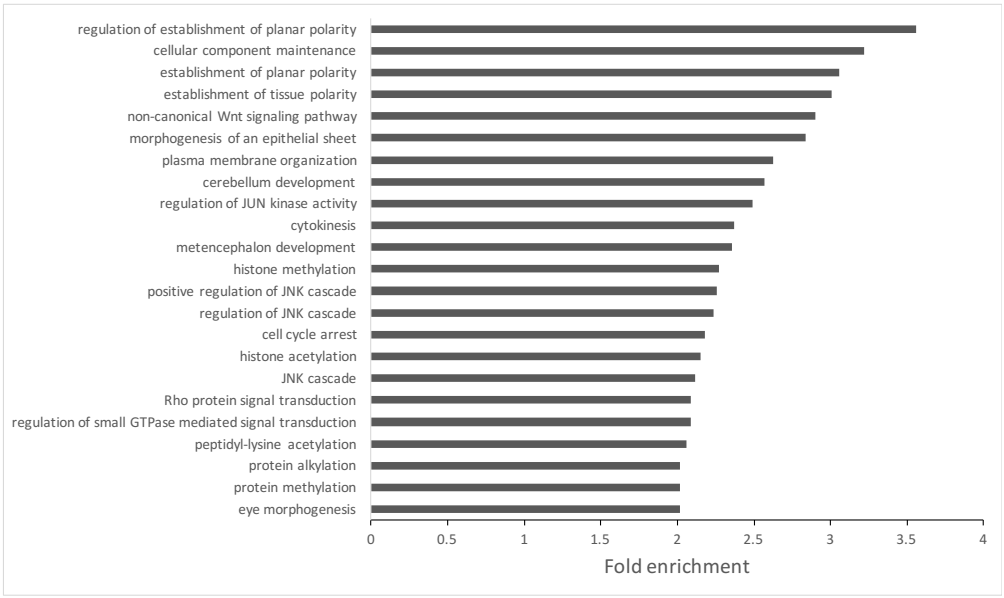
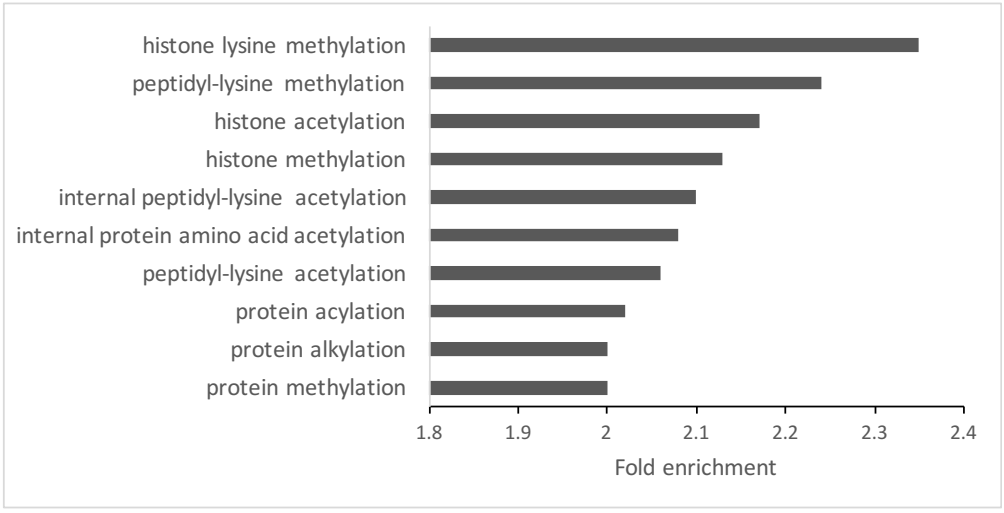


Fig. 3.12 Signal proportion for m6A treatment experiments. X-axis represents MeRIP-Seq datasets, i.e. human negative control(shGFP) and perturbation of WTAP, METTL3, METTL14_1, and METTL14_3 with blue for replicate 1 and yellow for replicate 2. Y-axis represents the signal proportion.

Finally, I conducted a functional analysis on the genes affected by the perturbation of methyltransferases. I performed gene ontology (GO) analysis by RCAS [92] on genes with lost m6A-enriched regions. The loss of m6A-enriched regions is defined as a state from being detected in all the replicates of the wild type to being undetected in all the replicates of the treated type. The GO results of enriched biological process (BP) terms are shown in Figure 3.13. For the mouse datasets of the wild type and Mettl3 knock-out, the enriched BP terms are related to planar polarity and polarity as shown in Figure 3.13 (a), thus suggesting that the loss of m6A affects the development of embryo cells. For the human datasets of negative control and WTAP perturbation, the enriched BP terms are related to histone methylation and acetylation as shown in Figure 3.13 (b), which also appeared in the term list for mouse. This observation agrees with that of [94] regarding m6A's function in destabilizing transcripts that encode histone modification enzymes.



(a) Mouse KO_Mettl3 vs WT



(b) Human shWTAP vs. shGFP

Fig. 3.13 Enriched biological process (BP) term for genes impacted by perturbation of m6A methyltransferase for (a) KO_Mettl3 vs WT (Mouse) and (b) shWTAP vs. shGFP (Human) . The threshold of the adjusted p-value for the terms are set as 0.05.

3.5 Discussion and Conclusion

MoAIMS is an efficient and user-friendly software for the analysis of MeRIP-Seq, which applies a statistical framework with negative-binomial model for signal detection. The NB model is commonly used in the differential expression analysis to model the distribution of counts for a gene across replicates [95, 96]. Besides, NB is also used to model the background read distribution in an individual sample in the analysis of ChIP-Seq [84, 97] and CLIP-Seq [98–100]. A main difference between ChIP-Seq and CLIP-Seq is that transcript abundance has a large impact on CLIP-Seq, which also exists in MeRIP-Seq. Nevertheless, it has been found that NB model can appropriately account for the overdispersion and incorporating input sample can improve the background estimation for CLIP-Seq [100]. This inspired the development of MoAIMS. Though there is no guarantee that NB can work well on all the MeRIP-Seq datasets, a goodness of fitting plot generated by MoAIMS can help check the model fitting.

There are some thoughts for improvements. First, MoAIMS currently supports only the analysis of single samples. For replicate samples, although enriched regions common in all the replicates can be easily extracted using MoAIMS, a joint statistical model can be developed as an alternative that considers the variance among replicates. Next, apart from the NB distribution, other statistical distributions are worth being tested owing to the wide diversity of RNA sequencing data. For example, Poisson-Tweedie is a more general family of count data distributions that can fit RNA sequencing data under situations of heavy tail or zero inflation [101]. Last but not least, because MoAIMS can provide user-friendly outputs for downstream analysis, it is feasible to integrate MeRIP-Seq datasets with other biological data for a comprehensive functional analysis, especially for MeRIP-Seq-treatment experiments.

I developed MoAIMS, which is an efficient and user-friendly software for analysis of MeRIP-Seq. MoAIMS is compatible with general RNA sequencing protocols, achieves excellent speed and competitive performance, and provides user-friendly outputs for downstream analysis. When MoAIMS was applied to studies of m6A, m6A's known biological features and its interplay with histone modification was revealed. Furthermore, the signal proportion inferred from MoAIMS can be used as an intuitive indicator of treatment effect. I hope that MoAIMS would facilitate MeRIP-Seq analysis and provide more insights into studies of RNA modification.

Chapter 4

Identification of m6A-associated RNA binding proteins using an integrative computational framework

4.1 Abstract

N6-methyladenosine (m6A) is one of the most abundant RNA modifications found in various species. Several wet lab studies have identified some RNA binding proteins (RBPs) that is associated with m6A regulation. The objective of this study was to identify potential m6A-associated RBPs using an integrative computational framework.

I identified reproducible m6A regions from independent studies in certain cell lines and then utilized RBPs' binding data of the same cell line to identify m6A-associated RBPs. The computational framework was composed of an enrichment analysis and a classification model. The enrichment analysis identified known m6A-associated RBPs including YTH domain-containing proteins; it also identified a potential m6A-associated RBP, RBM3, for mouse. I observed a significant correlation for the identified m6A-associated RBPs at the protein expression level rather than the gene expression. In addition, I built a Random Forest classification model for the reproducible m6A regions using RBPs' binding data. The RBP-based predictor demonstrated not only competitive performance when compared with sequence-based predictions but also helped to identify m6A-repelled RBP. These

results suggested that this framework allowed us to infer interaction between m6A and m6A-associated RBPs beyond sequence level when utilizing RBPs' binding data.

I designed an integrative computational framework for the identification of known and potential m6A-associated RBPs. I hope the analysis will provide more insights on the studies of m6A and RNA modification.

4.2 Introduction

In recent years, RNA modification has emerged as a mode of post-transcriptional gene regulation and has been gaining increasing attention from researchers around the globe. More than 100 types of post-transcriptional modification have been discovered, with N6-methyladenosine (m6A) as being one of the most abundant RNA modification [67]. m6A is featured with the DRACH motif(where D=A,G or U;R=A or G;H=A,C or U) and is preferentially located near 3' untranslated regions (3' UTR) [8]. It has been reported that m6A participates in essential RNA activities including alternative splicing, export, translation, and decay [18].

m6A exerts its function through interaction with several RNA binding proteins referred to as m6A-associated RBPs. There are three main kinds of m6A-associated RBPs, they are writer, eraser, and reader. m6A writers are methyltransferases like METTL3, METTL14, WTAP, RBM15/15B, while m6 erasers are demethyltransferases like FTO, ALKBH5 and m6A readers are the proteins that can recognize m6A like the YTH domain-containing proteins (YTHDF1/2/3), EIF3 [18], FMR1 [19]. m6A writers and erasers can be considered as m6A regulators which directly regulate m6A while m6A readers can be considered as m6A effectors which participate in m6A regulatory network. These m6A-associated RBPs cooperate with each other to facilitate both temporal and spatial regulation where writers work in the nucleus to introduce the m6A modification which is then recognized by various readers in the nucleus and cytoplasm, which can influence activities of their target RNAs.

There are some computational methods that can be used to identify m6A-associated RBPs. One such method is to build a prediction model based on deep learning and then extract the sequence features [45, 102]. However, not all RBP motifs are available and sequences cannot reflect actual binding, thus limiting their utility in the identification of

m6A-associated RBPs. Another group developed an analysis framework to identify cell-specific trans regulators of m6A [103]. Because there exists the considerable variation among MeRIP-Seq datasets (about 30 to 60% between studies, even in the same cell type) [104], I decided to focus on the use of reproducible m6A regions in order to identify m6A-associated RBPs.

Here, I aimed to identify m6A-associated RBPs from reproducible m6A regions. I developed an integrative computational framework composed of an enrichment analysis and a classification model. The enrichment analysis allows us to identify RBPs enriched in the m6A regions. I was able to identify not only the known m6A-associated RBPs like YTH domain-containing proteins, but also a potential m6A-associated RBP, RBM3, for mouse. I went on to evaluate the correlation of these identified m6A-associated RBPs with some known m6A regulators/effectors and compared these to other RBPs. I observed a significant correlation in the protein expression level rather than the gene expression, which suggested that the m6A-associated RBPs cooperate at the protein-level in regulating the process of modification. In addition, I built a Random Forest classification model for the reproducible m6A regions using RBPs' binding data in an effort to understand how RBPs contribute to the profiling of m6A regions. This RBP-based predictor demonstrated competitive performance when compared with sequence-based methods. Furthermore, the feature importance inferred from this model can be used to help identify m6A-repelled RBP. These results suggested that this framework could enable researchers to infer interaction between m6A and m6A-associated RBPs beyond sequence level when utilizing RBPs' binding data. I hope that the analysis could be used to provide more meaningful insight in future m6A and RNA modification studies.

4.3 Materials and methods

4.3.1 MeRIP-Seq data collection and processing

I collected the raw MeRIP-Seq FASTA files from four independent studies using the human HEK293T(Human embryonic kidney 293T cells) cell line from European Nucleotide Archive with accession numbers SRP090687 [105], SRP039397 [86], SRP007335 [5], and SRP162223. I also collected the MeRIP-Seq data from four independent studies using mouse embryonic

fibroblasts(MEF) with accession numbers SRP039402 [86], SRP048596 [106], SRP115436 [107], and SRP061617 [108].

To detect m6A regions from MeRIP-Seq data, I used MoAIMS [65], an efficient software I developed based on a statistical framework of a mixture negative-binomial distribution. After quality control using FastQC [109] and adapter-trimming with Cutadpat [110], I processed and analyzed the MeRIP-Seq using the method described in the MoAIMS paper. MoAIMS was performed using the default parameters except that I set sep_bin_info=F when analyzing studies with replicates. MoAIMS split genes to bins for signal detection, therefore the output is m6A regions with a size of 200bp as default. After that, I identified reproducible m6A regions that were called in at least 60% of the replicates in any one study and further in at least three studies.

4.3.2 The enrichment analysis

I retrieved the binding site data of RBPs from the POSTAR2 database [111] and identified RBPs enriched in the reproducible m6A regions. A permutation test was adopted to assess the significance of the RBP binding in the m6A regions. The rest of regions in genes with m6A was used as control and then sampled 1000 times. I kept the ratio of the number of bins in exons to the number of bins spanning exons the same for both m6A and control regions to avoid the regions' position being a confounding factor. For each RBP, I calculated the enrichment ratio using the Equation (4.1) where N_t is the number of m6A regions with the RBP and $E(N_c)$ is the average number of control regions with the RBP from 1000 times of sampling. Then, a p-value was calculated as the proportion of N_c which were greater than or equal to N_t. After that, multiple testing was performed using Benjamini & Hochberg [112].

$$R = \frac{N_t}{E(N_c)} \tag{4.1}$$

4.3.3 The classification model

I built a Random Forest (RF) classifier to evaluate how much RBPs contribute in discriminating reproducible m6A regions. I used the human m6A regions with RBPs' binding as the positive data (13978 in total) and generated 10 sets of control data from the control regions which were set to be an equal data size. I kept the ratio of the number of

bins in exons to the number of bins spanning exons the same in both m6A and control regions. The binding information (1 for binding, 0 for non-binding) of RBPs was used as the input features. The data was divided into training and test groups at a ratio of 80:20. I implemented the RF classifier using the R package caret [113] and randomForest [114] with 5-fold cross validations and 'mtry' (the tuning parameters) as 8 (nearly the square root of the number of features). I used the accuracy to measure the performance of the models as shown in the Equation (4.2) where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.2}$$

4.4 Results

4.4.1 Identification of m6A-associated RBPs enriched in reproducible m6A regions

Because of the considerable variation in the m6A datasets [104], I collected the data from nine samples of human HEK293T cell line from four independent studies and six samples of mouse MEF cell line from four independent studies to generate the *reproducible* m6A regions used in this study. The details of the detection of these m6A regions is provided in the Methods section. I required that the m6A regions were called in at least 60% of replicates within each study and that they appeared in at least three of the total studies analyzed for them to be qualified as reproducible. Under these criteria, I finally obtained 14803 reproducible m6A regions for the HEK293T cell line and 5576 reproducible m6A regions for the MEF cell line.

To identify the RBPs enriched in these m6A regions, 71 RBPs for HEK293T/HEK293 and nine RBPs for MEF were retrieved from the POSTAR2 database. For each RBP, I calculated an enrichment score and assessed its significance using a permutation test as described in Methods section. When setting the threshold for the enrichment ratio to $\geq$1.3 and FDR (false discovery rate) adjusted p-value to $\leq$0.05, I obtained the enriched RBPs listed in Table 4.1. For HEK293T, I identified several known m6A readers including YTH family proteins, FMR1, EIF3, and m6A writers RBM15/15B [18]. For MEF, I found a common RBP, CPSF6, which is enriched for both human and mouse. CPSF6 has been reported to interact with VIRMA, which mediates preferential m6A methylation in the 3' UTR and

near stop codon and is associated with alternative polyadenylation (APA) in human [115]. In addition, I noticed that RBM3 was highly enriched in m6A regions of MEF. RBM3 is an important regulator of circadian gene expression by controlling APA [116]. This suggests that RBM3 could participate in m6A regulation which is associated with APA. The full list of enrichment ratios for each of the RBPs (including raw p-values and FDR adjusted p-values) is provided in Table B.1 and B.2 of Appendix B. Besides, for each enriched RBP (overlap with more than 100 m6A regions), I also listed the RBPs that more than 60% of the enriched RPB is overlapped with for HEK293T in Table B.3 of Appendix B. As expected, YTHDF1 and DDX3X were shown to have the highest overlapping percentage as they have a considerable overlap with m6A regions.

Table 4.1 RNA binding proteins (RBPs) enriched in reproducible m6A regions

HEK293T	Enrichment ratios*	# m6A regions with RBPs	p-value**	FDR adjusted p-value
YTHDF2	3.90	6964	<0.001	<0.003
RBM15	2.73	3534	<0.001	<0.003
YTHDF3	2.70	52	<0.001	<0.003
YTHDF1	2.49	9196	<0.001	<0.003
RBM15B	2.32	6375	<0.001	<0.003
YTHDC1	2.15	7224	<0.001	<0.003
EIF3D	1.88	593	<0.001	<0.003
NOP58	1.74	159	<0.001	<0.003
HNRNPH1	1.57	47	0.002	0.006
NUDT21	1.48	5201	<0.001	<0.003
FMR1	1.46	4443	<0.001	<0.003
DDX3X	1.44	9470	<0.001	<0.003
EIF3A	1.39	293	<0.001	<0.003
CPSF6	1.34	3593	<0.001	<0.003
CPSF7	1.31	4413	<0.001	<0.003

MEF	Enrichment ratio*	# m6A regions with RBPs	p-value**	FDR adjusted p-value
RBM3	5.81	485	<0.001	<0.001
CREBBP	2.47	24	<0.001	<0.001
SRSF2	2.24	793	<0.001	<0.001
SRSF1	2.13	467	<0.001	<0.001
CPSF6	2.07	94	<0.001	<0.001
CIRBP	1.76	401	<0.001	<0.001

* RBPs are ranked by their enrichment ratios. ** P-values were calculated from 1000 times of permutation. When p-value is zero, it is shown in the table as < 0.001 becasue it is possible that the p-value is actually less than 0.001 if times of permutation were increased.

The RBPs in Table 4.1 are considered as m6A-associated RBPs, therefore I wondered how these identified m6A-associated RBPs are correlated with known m6A regulators/effectors when compared with other RPBs at both the transcription and the protein expression level. I performed a correlation analysis for all the human RBPs. To do the correlation analysis at the transcription level, I downloaded Illumina Body Map (HBM) [117–119] from ArrayExpress [120] with the accession number E-MTAB-513, which provides gene expression data for 16 human tissues. For the correlation analysis at the protein level, I downloaded mass spectrometry data from Human Proteome Map (HPM) [121] for 30 human tissues/cell lines. I checked some known m6A regulators/effectors including YTHDF2, RBM15, EIF3D which ranked at the top of Table 4.1 and METTL3 of which binding data is not available but is a well-known m6A writer, and compared their correlation with the identified m6A-associated RBPs (15 in total) or with the rest of RBPs (56 in total). Correlation was calculated using the Spearman's correlation coefficient. I observed a similar trend in all the investigated known m6A regulators/effectors which showed that the identified m6A-associated RBPs are more correlated with them at the protein-level than the transcription level (Figure 4.1). Because the protein data included more tissues/cell lines than the transcription data, I chose to compare a subset of 17 adult tissues to check the correlation values for avoiding any biased introduced by different dataset sizes. The higher correlation at the protein level was still observed in this subset evaluation as shown in Figure 4.1. This observation supports the hypothesis that m6A-associated RBPs are more likely to cooperate at the protein level. Then, I went on to confirm to if these higher correlation values are the result of protein-protein interactions. To do this I retrieved the protein-protein interaction data from STRING [122]. The interaction scores do not show significant difference between m6A-associated RBPs and other RBPs except for METTL3 (Figure 4.2), which suggests that the higher correlation at the protein-level is only marginally related to protein-protein interaction. This indicates that the regulation of the m6A modification is a dynamic process involving both temporal and spatial interactions between the m6A-associated RBPs.

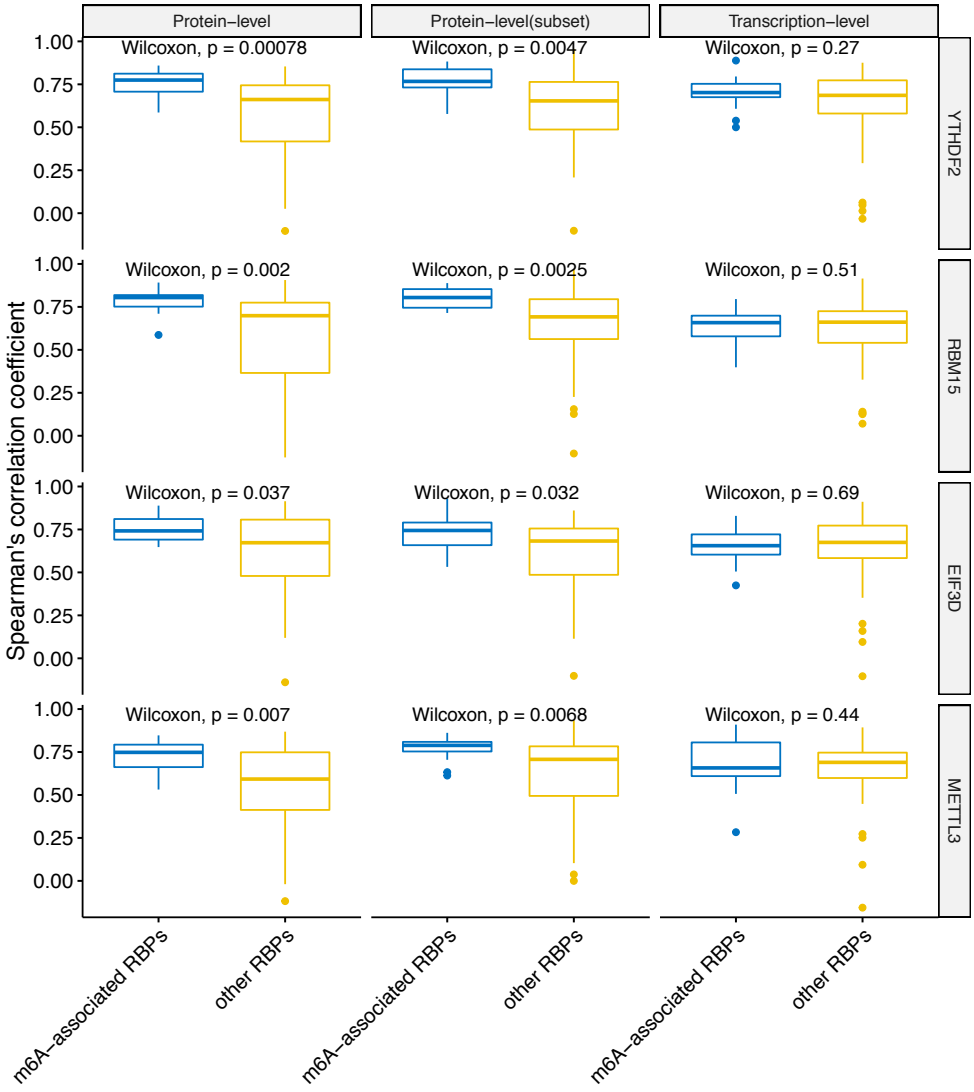


Fig. 4.1 Comparison of the correlation values for known m6A regulators/effectors (YTHDF2, RBM15, EIF3D, and METTL3) with identified m6A-associated RBPs (15 in total) or other RBPs. The boxplot shows the distribution of the Spearman's correlation coefficient between known m6A regulators/effectors and identified m6A-associated RBPs/other RBPs at the protein- and transcript-level (the subset protein-level results describe the correlation coefficients calculated from a subset of the protein data which included only 17 adult tissues). Significance was evaluated using a one-sided Wilcoxon test.

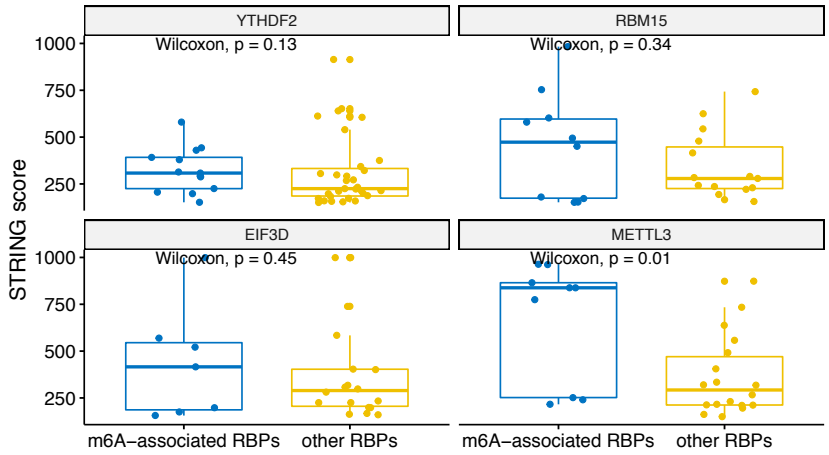


Fig. 4.2 Comparison of protein-protein interactions between known m6A regulators/effectors (YTHDF2, RBM15, EIF3D, and METTL3) and identified m6A-associated RBPs (15 in total)/other RBPs. The boxplot shows the distribution of the interaction scores between known m6A regulators/effectors and identified m6A-associated RBPs/other RBPs. Significance was evaluated using a one-sided Wilcoxon test.

4.4.2 Identification of m6A-associated RBPs contributing to the classification of m6A regions

Although I identified RBPs enriched in the reproducible m6A regions, I wanted to develop a more comprehensive understanding of how the RBPs contribute to the profile of the m6A regions. To do this, I performed a further analysis on the human RBPs. First, I investigated the overall profile of the binding information of RBPs (0 for non-binding and 1 for binding) in the reproducible m6A regions. I calculated the pairwise distance between RBPs using cosine similarity and performed clustering (Figure 4.3). The result of the clustering analysis demonstrated the co-occurrence of YTH family proteins and RBM15B which all ranked in the top of the enrichment analysis. Then, I built a Random Forest classifier which incorporated the binding information for each of the RBPs as features. The details of models are described in the Methods section. The classifier achieved an average accuracy of 0.736 and AUROC (Area Under Receiver Operating Characteristic) of 0.788 as shown in Figure 4.4. I also compared the RBP-based classifier with two sequence-based predictors SRAMP [33] in mature mRNA mode and DeepM6ASeq which showed an accuracy of 0.660 and 0.686, respectively (Figure 4.4). I plotted top10 most important

features as shown in Figure 4.5 and found that apart from the known m6A-associated RPBs such as the readers YTHDF1/2, YTHDC1, the writers RBM15/15B that have been shown to be enriched in the m6A regions, ELAVL1, which is reported to have action of being repelled by m6A [123], also contributed to the classification of m6A regions to some extent. The repelling action of m6A against ELAVL1 is consistent with the enrichment results, which show that its enrichment ratio is 0.816. In summary, the RBP-based classifier not only demonstrated competitive performance in the prediction of reproducible m6A regions but also helped to infer interaction between m6A and m6A-associated RBPs beyond sequence level when combined with the results of the enrichment analysis.

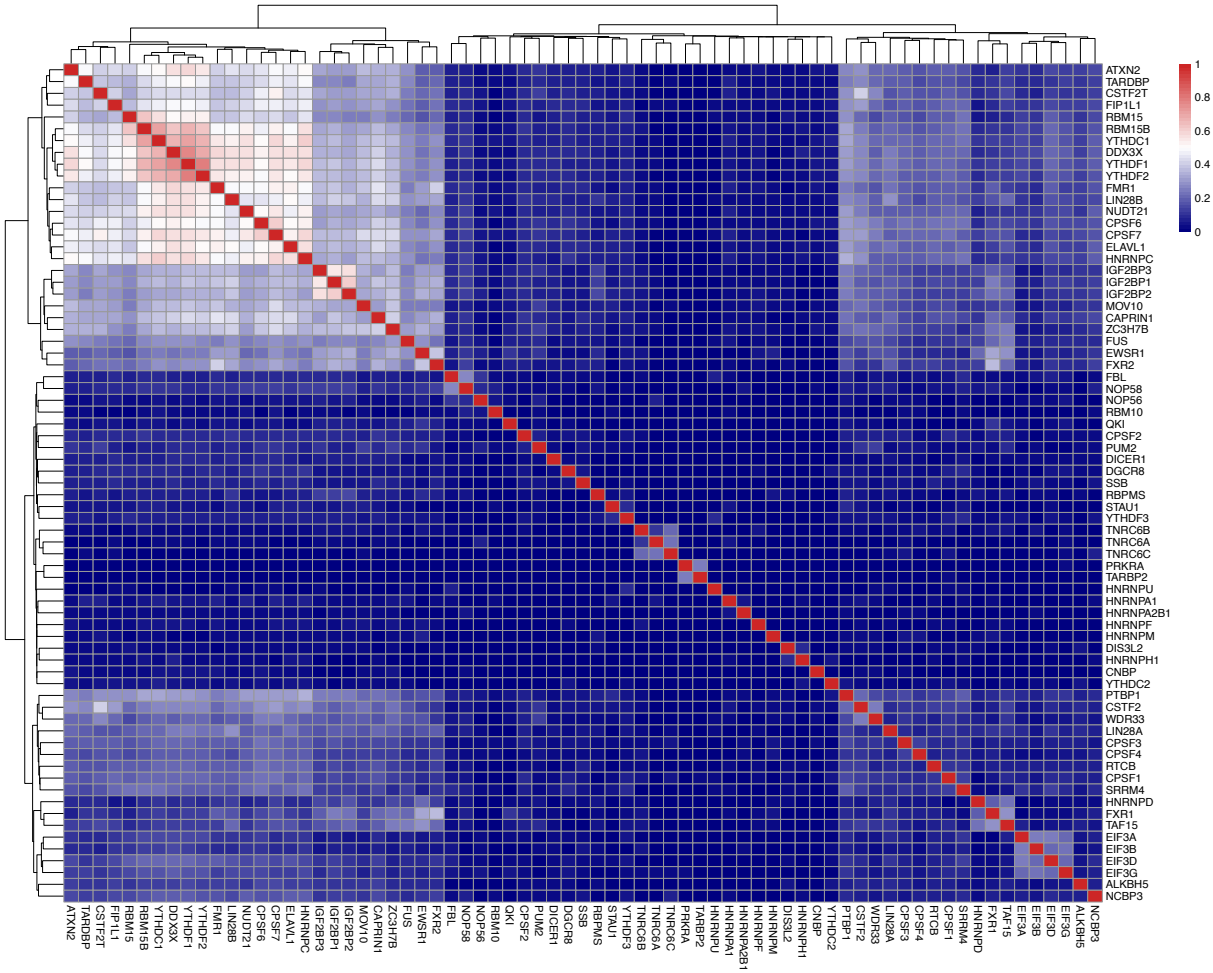


Fig. 4.3 Clustering of RNA binding proteins (RBPs) in the m6A regions of HEK293T cells. X-axis and Y-axis represent the names of the RBPs. The color scale indicates the cosine similarity between the RBPs.

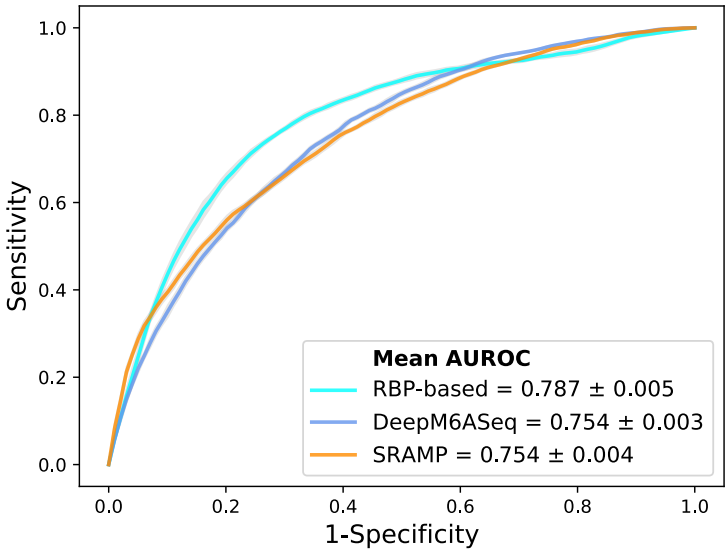


Fig. 4.4 Comparison of AUROC between the RBPs (RNA binding proteins)-based predictor, DeepM6ASeq, and SRAMP in mature mRNA mode for the classification of HEK293T m6A regions. The plot represents average ROC from ten times of sampling control regions for each predictor.

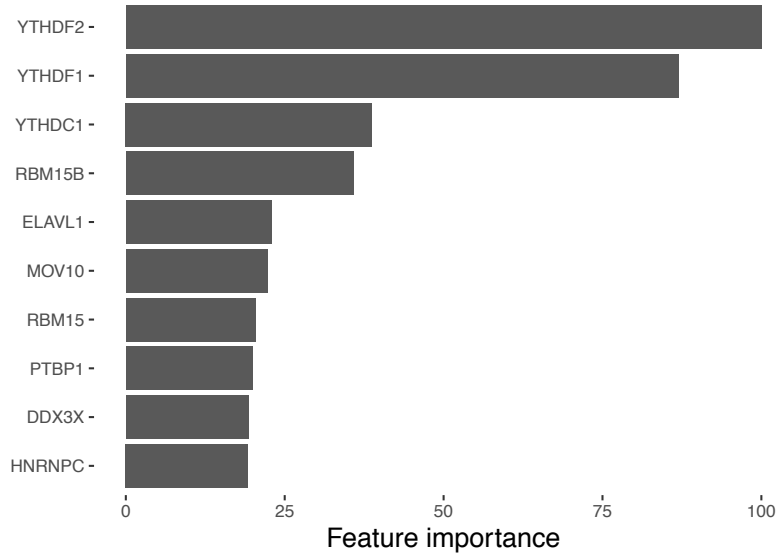


Fig. 4.5 Top 10 RNA binding proteins (RBPs) identified from the classification of the HEK293T reproducible m6A regions. The bar graph shows the top 10 RBPs extracted from the classifier for the m6A regions. X-axis represents the name of RBPs and Y-axis represents the average importance score from ten times of sampling control regions.

4.5 Discussion and Conclusion

This computational framework enabled us to identify potential m6A-associated RBPs and infer interaction between m6A and m6A-associated RBPs. Some studies have reported the connection of dysfunction of m6A-associated RBPs to disease. Fox example, YTHDF2 silenced in human hepatocellular carcinoma(HCC) cells can provoke inflammation, vascular reconstruction and metastatic progression [27]. Therefore, it is expected that the study of m6A and m6A-associated RBPs can lead to a better understanding of gene regulation mechanism and potential therapeutic opportunities.

This analysis serves as a first step, and future analyses may include some improvements and expansions. First, this framework was designed and tested on a limited number of cell types and organisms. With the increasing amount of data available for m6A and RBPs in more cell lines and tissues, this framework could be tested on much larger datasets and may provide valuable insights into the m6A regulatory network. In addition, this framework could be applied to other RNA modifications such as N1-methyladenosine (m1A) [66] and 5-

methylcytidine(m5C) [67], which have also been identified as critical RNA modification. Such analyses could help improve experimental design in wet lab applications and help researchers narrow their focus. Third, apart from RBPs, other genomic features like transcription factors and histone modification are worth inspecting for studying the m6A regulation networks at multiple layers. These applications highlight the future utility of this framework and its value in the current research climate.

I designed an integrative computational framework for the identification of m6A-associated RBPs in reproducible m6A regions. This computational framework is composed of an enrichment analysis and a classification model. Using the enrichment analysis, I was able to identify known m6A-associated RBPs and several potential m6A-associated RBPs including RBM3 from mouse. These identified m6A-associated RBPs show a significant degree of correlation at their protein level, although this is not seen in their transcriptional profile, which suggests that these m6A-associated RBPs cooperate at the protein-level to regulate the process of modification. In addition, I built a classification model for m6A regions using a Random Forest algorithm that uses RBPs' binding information as its input features. The RBP-based predictor not only demonstrated comparable performance to sequence-based predictions but also helped infer interaction between m6A and m6A-associated RBPs like actions of reading and repelling beyond sequence level. I hope that this analysis framework can assist biologists in their study of RNA modifications.

Chapter 5

General conclusions and future work

5.1 Conclusions

To solve the key issues of signal detection and biological feature extraction in the analysis of high-throughput m6A data, I present two softwares and one analysis framework in this thesis.

In Chapter 2, I developed DeepM6ASeq, a deep learning framework, to predict m6A-containing sequences and characterize biological features surrounding m6A sites at sequence level. DeepM6ASeq showed competitive performance of prediction, learned known m6A readers and a newly recognized one, FMR1, and also helped to visualize positions of m6A sites.

In Chapter 3, I developed MoAIMS, an efficient and easy-to-use software for analysis of MeRIP-Seq. MoAIMS achieves excellent speed and competitive performance in detection of m6A regions, and provides user-friendly outputs for downstream analysis. MoAIMS also provide intuitive evaluation on treatment effect for MeRIP-Seq treatment datasets.

In Chapter 4, I designed an integrative computational framework for the identification of m6A-associated RBPs in the reproducible m6A regions. Utilizing RBP's binding data, the framework is able to identify known m6A-associated RBPs and also found some potential ones such as RBM3 for mouse. Besides, it also helps infer interaction between m6A and m6A-associated RBPs like actions of reading and repelling beyond sequence level.

In conclusion, this thesis presents state-of-the-art prediction models and statistical methods for the systematic analysis of high-throughput m6A data and it is expected that the thesis can provide more insights for the research on m6A and assist biologists in studying the regulation mechanism of m6A.

5.2 Future Work

About the future work, there are three main aspects. First, in the aspect of signal detection, HMM(hidden Markov model) [124] can be applied in the future which is a statistical model of sequential data. HMM takes into consideration the dependency between bins and works well for smaller bin size. Besides, it needs to develop a joint model that considers the variance among sample replicates. Furthermore, it needs to apply other statistical distributions like Poisson-Tweedie [101] in order to deal with the wide diversity of RNA sequencing data.

Second, in the aspect of biological feature extraction, several algorithms and high-throughput data requires further studies. At sequence level, it needs to develop models that are more competitive and interpretable. Models in the field of natural language processing(NLP) are inspiring for the analysis of biological sequence data. Especially, the state-of-the-art model 'BERT'(Bidirectional Encoder Representations from Transformers) [125] is a promising one. BERT is a pre-training language model based on Transformer model [126] and can learn information from both left and right sides of sequences. If an appropriate pre-trained model is used to learn the language of transcriptome, it is possible to improve the prediction power. With regard to the interpretability of models, at sequence level, word-embedding [127] is a feasible strategy. With word-embedding, the features are k-mer sequences as motif, and the deep learning model can learn the combinations of motifs, which makes the model more interpretable.

At the feature level beyond sequences, other genomic features like transcription factors and histone modification are worth inspecting for studying the m6A regulation networks at multiple layers. Besides, it needs to pay attention to the dependency of features in the interpretation of prediction models, although feature importance extracted from the random forest model built on the protein-binding data in my study did revealed some known and important biological features. For the possible dependency between features like multicollinearity, some nonparametric classifiers like SVM(support vector machine) with

non-linear kernel are less sensitive, and feature selection(for example, the leave-one-out procedure can be used to check the change on prediction) needs to be applied to get important features.

The last aspect is about the application of the systematic analysis framework, which is presented in the thesis including prediction and feature extraction, on the study of other RNA modifications such as N1-methyladenosine (m1A) [66] and 5-methylcytidine(m5C) [67]. I hope my study to facilitate biologists' experiment design in the future.

Bibliography

[1] F. Morena, C. Argentati, M. Bazzucchi, C. Emiliani, and S. Martino, "Above the Epitranscriptome: RNA Modifications and Stem Cell Identity," Genes (Basel), vol.9, no.7, Jun 2018.

[2] I.A. Roundtree, M.E. Evans, T. Pan, and C. He, "Dynamic RNA Modifications in Gene Expression Regulation," Cell, vol.169, no.7, pp.1187–1200, Jun 2017.

[3] T. Pan, "N6-methyl-adenosine modification in messenger and long non-coding RNA," Trends Biochem. Sci., vol.38, no.4, pp.204–209, Apr 2013.

[4] J.M. Adams and S. Cory, "Modified nucleosides and bizarre 5'-termini in mouse myeloma mRNA," Nature, vol.255, no.5503, pp.28–33, May 1975.

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

[6] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, and G. Rechavi, "Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq," Nature, vol.485, no.7397, pp.201–206, Apr 2012.

[7] M.J. Clancy, M.E. Shambaugh, C.S. Timpte, and J.A. Bokar, "Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene," Nucleic Acids Res., vol.30, no.20, pp.4509–4518, Oct 2002.

[8] B. Linder, A.V. Grozhik, A.O. Olarerin-George, C. Meydan, C.E. Mason, and S.R. Jaffrey, "Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome," Nat. Methods, vol.12, no.8, pp.767–772, Aug 2015.

[9] R. Wu, D. Jiang, Y. Wang, and X. Wang, "N (6)-methyladenosine (m(6)a) methylation in mrna with a dynamic and reversible epigenetic modification," Molecular biotechnology, vol.58, no.7, pp.450–459, July 2016.

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

[11] J. Liu, Y. Yue, D. Han, X. Wang, Y. Fu, L. Zhang, G. Jia, M. Yu, Z. Lu, X. Deng, Q. Dai, W. Chen, and C. He, "A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation," Nat. Chem. Biol., vol.10, no.2, pp.93–95, Feb 2014.

[12] S.D. Agarwala, H.G. Blitzblau, A. Hochwagen, and G.R. Fink, "RNA methylation by the MIS complex regulates a cell fate decision in yeast," PLoS Genet., vol.8, no.6, p.e1002732, 2012.

[13] D.P. Patil, C.K. Chen, B.F. Pickering, A. Chow, C. Jackson, M. Guttman, and S.R. Jaffrey, "m(6)A RNA methylation promotes XIST-mediated transcriptional repression," Nature, vol.537, no.7620, pp.369–373, 09 2016.

[14] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y.G. Yang, and C. He, "N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO," Nat. Chem. Biol., vol.7, no.12, pp.885–887, Oct 2011.

[15] G. Zheng, J.A. Dahl, Y. Niu, P. Fedorcsak, C.M. Huang, C.J. Li, C.B. Vågbø, Y. Shi, W.L. Wang, S.H. Song, Z. Lu, R.P. Bosmans, Q. Dai, Y.J. Hao, X. Yang, W.M. Zhao, W.M. Tong, X.J. Wang, F. Bogdan, K. Furu, Y. Fu, G. Jia, X. Zhao, J. Liu, H.E. Krokan, A. Klungland, Y.G. Yang, and C. He, "ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility," Mol. Cell, vol.49, no.1, pp.18–29, Jan 2013.

[16] D. Theler, C. Dominguez, M. Blatter, J. Boudet, and F.H. Allain, "Solution structure of the YTH domain in complex with N6-methyladenosine RNA: a reader of methylated RNA," Nucleic Acids Res., vol.42, no.22, pp.13911–13919, Dec 2014.

[17] S. Berlivet, J. Scutenaire, J.M. Deragon, and C. Bousquet-Antonelli, "Readers of the m6A epitranscriptomic code," Biochim Biophys Acta Gene Regul Mech, vol.1862, no.3, pp.329–342, 03 2019.

[18] Y. Lee, J. Choe, O.H. Park, and Y.K. Kim, "Molecular Mechanisms Driving mRNA Degradation by m6A Modification," Trends Genet., vol.36, no.3, pp.177–188, Mar 2020.

[19] R.R. Edupuganti, S. Geiger, R.G.H. Lindeboom, H. Shi, P.J. Hsu, Z. Lu, S.Y. Wang, M.P.A. Baltissen, P.W.T.C. Jansen, M. Rossa, M. Muller, H.G. Stunnenberg, C. He, T. Carell, and M. Vermeulen, "N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis," Nat. Struct. Mol. Biol., vol.24, no.10, pp.870–878, Oct 2017.

[20] I.A. Roundtree, G.Z. Luo, Z. Zhang, X. Wang, T. Zhou, Y. Cui, J. Sha, X. Huang, L. Guerrero, P. Xie, E. He, B. Shen, and C. He, "YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs," Elife, vol.6, 10 2017.

[21] H. Shi, X. Wang, Z. Lu, B.S. Zhao, H. Ma, P.J. Hsu, C. Liu, and C. He, "YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA," Cell Res., vol.27, no.3, pp.315–328, Mar 2017.

[22] W. Xiao, S. Adhikari, U. Dahal, Y.S. Chen, Y.J. Hao, B.F. Sun, H.Y. Sun, A. Li, X.L. Ping, W.Y. Lai, X. Wang, H.L. Ma, C.M. Huang, Y. Yang, N. Huang, G.B. Jiang, H.L. Wang, Q. Zhou, X.J. Wang, Y.L. Zhao, and Y.G. Yang, "Nuclear m(6)a reader ythdc1 regulates mrna splicing," Molecular cell, vol.61, no.4, pp.507–519, February 2016.

[23] S.D. Kasowitz, J. Ma, S.J. Anderson, N.A. Leu, Y. Xu, B.D. Gregory, R.M. Schultz, and P.J. Wang, "Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development," PLoS Genet., vol.14, no.5, p.e1007412, 05 2018.

[24] L.P. Vu, B.F. Pickering, Y. Cheng, S. Zaccara, D. Nguyen, G. Minuesa, T. Chou, A. Chow, Y. Saletore, M. MacKay, J. Schulman, C. Famulare, M. Patel, V.M. Klimek, F.E. Garrett-Bakelman, A. Melnick, M. Carroll, C.E. Mason, S.R. Jaffrey, and M.G.

Kharas, "The N6-methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells," Nat. Med., vol.23, no.11, pp.1369–1376, Nov 2017.

[25] I. Barbieri, K. Tzelepis, L. Pandolfini, J. Shi, G. Millán-Zambrano, S.C. Robson, D. Aspris, V. Migliori, A.J. Bannister, N. Han, E. De Braekeleer, H. Ponstingl, A. Hendrick, C.R. Vakoc, G.S. Vassiliou, and T. Kouzarides, "Promoter-bound METTL3 maintains myeloid leukaemia by m6A-dependent translation control," Nature, vol.552, no.7683, pp.126–131, 12 2017.

[26] M. Chen, L. Wei, C.T. Law, F.H. Tsang, J. Shen, C.L. Cheng, L.H. Tsang, D.W. Ho, D.K. Chiu, J.M. Lee, C.C. Wong, I.O. Ng, and C.M. Wong, "RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2," Hepatology, vol.67, no.6, pp.2254–2270, 06 2018.

[27] J. Hou, H. Zhang, J. Liu, Z. Zhao, J. Wang, Z. Lu, B. Hu, J. Zhou, Z. Zhao, M. Feng, H. Zhang, B. Shen, X. Huang, B. Sun, C. He, and Q. Xia, "YTHDF2 reduction fuels inflammation and vascular abnormalization in hepatocellular carcinoma," Mol. Cancer, vol.18, no.1, p.163, 11 2019.

[28] D. Han, J. Liu, C. Chen, L. Dong, Y. Liu, R. Chang, X. Huang, Y. Liu, J. Wang, U. Dougherty, M.B. Bissonnette, B. Shen, R.R. Weichselbaum, M.M. Xu, and C. He, "Anti-tumour immunity controlled through mrna m6a methylation and ythdf1 in dendritic cells," Nature, vol.566, no.7743, pp.270–274, February 2019.

[29] S. Ke, E.A. Alemu, C. Mertens, E.C. Gantman, J.J. Fak, A. Mele, B. Haripal, I. Zucker-Scharff, M.J. Moore, C.Y. Park, C.B. Vågbø, A. Kusśnierczyk, A. Klungland, J.E. Darnell, and R.B. Darnell, "A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation," Genes Dev., vol.29, no.19, pp.2037–2053, Oct 2015.

[30] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, and X.S. Liu, "Model-based analysis of ChIP-Seq (MACS)," Genome Biol., vol.9, no.9, p.R137, 2008.

[31] J. Meng, X. Cui, M.K. Rao, Y. Chen, and Y. Huang, "Exome-based analysis for RNA epigenome sequencing data," Bioinformatics, vol.29, no.12, pp.1565–1567, Jun 2013.

[32] X. Cui, J. Meng, S. Zhang, Y. Chen, and Y. Huang, "A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data," Bioinformatics, vol.32, no.12, pp.i378–i385, 06 2016.

[33] Y. Zhou, P. Zeng, Y.H. Li, Z. Zhang, and Q. Cui, "SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features," Nucleic Acids Res., vol.44, no.10, p.e91, 06 2016.

[34] W. Chen, P. Xing, and Q. Zou, "Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines," Sci Rep, vol.7, p.40242, 01 2017.

[35] L. Breiman, "Random forests," Machine Learning, vol.45, no.1, pp.5–32, 2001.

[36] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," Neurocomputing, vol.55, no.1–2, pp.169 – 186, 2003.

[37] B. Alipanahi, A. Delong, M.T. Weirauch, and B.J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," Nat. Biotechnol., vol.33, no.8, pp.831–838, Aug 2015.

[38] Y. LeCun, B.E. Boser, J.S. Denker, D. Henderson, R.E. Howard, W.E. Hubbard, and L.D. Jackel, "Handwritten digit recognition with a back-propagation network," in Advances in Neural Information Processing Systems 2, ed. D.S. Touretzky, pp.396–404, Morgan-Kaufmann, 1990.

[39] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors," Nature, vol.323, no.6088, pp.533–536, Oct. 1986.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol.521, no.7553, pp.436–444, May 2015.

[41] D.R. Kelley, J. Snoek, and J.L. Rinn, "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks," Genome Res., vol.26, no.7, pp.990–999, 07 2016.

[42] D. Quang and X. Xie, "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences," Nucleic Acids Res., vol.44, no.11, p.e107, 06 2016.

[43] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," Brief. Bioinformatics, vol.18, no.5, pp.851–869, 09 2017.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol.9, no.8, pp.1735–1780, Nov. 1997.

[45] Y. Zhang and M. Hamada, "DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning," BMC Bioinformatics, vol.19, no.Suppl 19, p.524, Dec 2018.

[46] S. Schwartz, S.D. Agarwala, M.R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T.S. Mikkelsen, R. Satija, G. Ruvkun, S.A. Carr, E.S. Lander, G.R. Fink, and A. Regev, "High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis," Cell, vol.155, no.6, pp.1409–1421, Dec 2013.

[47] I.A. Roundtree and C. He, "Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing," Trends Genet., vol.32, no.6, pp.320–321, 06 2016.

[48] T. Chen, Y.J. Hao, Y. Zhang, M.M. Li, M. Wang, W. Han, Y. Wu, Y. Lv, J. Hao, L. Wang, A. Li, Y. Yang, K.X. Jin, X. Zhao, Y. Li, X.L. Ping, W.Y. Lai, L.G. Wu, G. Jiang, H.L. Wang, L. Sang, X.J. Wang, Y.G. Yang, and Q. Zhou, "m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency," Cell Stem Cell, vol.16, no.3, pp.289–301, Mar 2015.

[49] X. Wang, B.S. Zhao, I.A. Roundtree, Z. Lu, D. Han, H. Ma, X. Weng, K. Chen, H. Shi, and C. He, "N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency," Cell, vol.161, no.6, pp.1388–1399, Jun 2015.

[50] H. Weng, H. Huang, H. Wu, X. Qin, B.S. Zhao, L. Dong, H. Shi, J. Skibbe, C. Shen, C. Hu, Y. Sheng, Y. Wang, M. Wunderlich, B. Zhang, L.C. Dore, R. Su, X. Deng, K. Ferchen, C. Li, M. Sun, Z. Lu, X. Jiang, G. Marcucci, J.C. Mulloy, J. Yang, Z. Qian, M. Wei, C. He, and J. Chen, "METTL14 Inhibits Hematopoietic Stem/Progenitor Differentiation and Promotes Leukemogenesis via mRNA m6A Modification," Cell Stem Cell, vol.22, no.2, pp.191–205, Feb 2018.

[51] W. Chen, H. Tran, Z. Liang, H. Lin, and L. Zhang, "Identification and analysis of the N(6)-methyladenosine in the Saccharomyces cerevisiae transcriptome," Sci Rep, vol.5, p.13859, Sep 2015.

[52] W. Chen, P. Feng, H. Ding, H. Lin, and K.C. Chou, "iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition," Anal. Biochem., vol.490, pp.26–33, Dec 2015.

[53] J. Zhou and O.G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," Nat. Methods, vol.12, no.10, pp.931–934, Oct 2015.

[54] C. Zhang, Y. Chen, B. Sun, L. Wang, Y. Yang, D. Ma, J. Lv, J. Heng, Y. Ding, Y. Xue, X. Lu, W. Xiao, Y.G. Yang, and F. Liu, "m6A modulates haematopoietic stem and progenitor cell specification," Nature, vol.549, no.7671, pp.273–276, 09 2017.

[55] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, and G. Rechavi, "Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing," Nat Protoc, vol.8, no.1, pp.176–189, Jan 2013.

[56] H. Liu, H. Wang, Z. Wei, S. Zhang, G. Hua, S.W. Zhang, L. Zhang, S.J. Gao, J. Meng, X. Chen, and Y. Huang, "MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome," Nucleic Acids Res., vol.46, no.D1, pp.D281–D287, Jan 2018.

[57] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," Bioinformatics, vol.28, no.23, pp.3150–3152, Dec 2012.

[58] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol.15, no.1, pp.1929–1958, Jan. 2014.

[59] J. Lanchantin, R. Singh, B. Wang, and Y. Qi, "DEEP MOTIF DASHBOARD: VISUALIZING AND UNDERSTANDING GENOMIC SEQUENCES USING DEEP NEURAL NETWORKS," Pac Symp Biocomput, vol.22, pp.254–265, 2017.

[60] S. Gupta, J.A. Stamatoyannopoulos, T.L. Bailey, and W.S. Noble, "Quantifying similarity between motifs," Genome Biol., vol.8, no.2, p.R24, 2007.

[61] N. Liu, K.I. Zhou, M. Parisien, Q. Dai, L. Diatchenko, and T. Pan, "N6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein," Nucleic Acids Res., vol.45, no.10, pp.6051–6063, Jun 2017.

[62] P.J. Batista, B. Molinie, J. Wang, K. Qu, J. Zhang, L. Li, D.M. Bouley, E. Lujan, B. Haddad, K. Daneshvar, A.C. Carter, R.A. Flynn, C. Zhou, K.S. Lim, P. Dedon, M. Wernig, A.C. Mullen, Y. Xing, C.C. Giallourakis, and H.Y. Chang, "m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells," Cell Stem Cell, vol.15, no.6, pp.707–719, Dec 2014.

[63] J.A. Castro-Mondragon, S. Jaeger, D. Thieffry, M. Thomas-Chollier, and J. van Helden, "RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections," Nucleic Acids Res., vol.45, no.13, p.e119, Jul 2017.

[64] X. Min, W. Zeng, N. Chen, T. Chen, and R. Jiang, "Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding," Bioinformatics, vol.33, no.14, pp.i92–i101, Jul 2017.

[65] Y. Zhang and M. Hamada, "MoAIMS: efficient software for detection of enriched regions of MeRIP-Seq," BMC Bioinformatics, vol.21, no.1, p.103, Mar 2020.

[66] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M.S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, W.C. Clark, G. Zheng, T. Pan, O. Solomon, E. Eyal, V. Hershkovitz, D. Han, L.C. Dore, N. Amariglio, G. Rechavi, and C. He, "The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA," Nature, vol.530, no.7591, pp.441–446, Feb 2016.

[67] T. Amort, D. Rieder, A. Wille, D. Khokhlova-Cubberley, C. Riml, L. Trixl, X.Y. Jia, R. Micura, and A. Lusser, "Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain," Genome Biol., vol.18, no.1, p.1, 01 2017.

[68] D.S. Johnson, A. Mortazavi, R.M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," Science, vol.316, no.5830, pp.1497–1502, Jun 2007.

[69] J.D. Mills, Y. Kawahara, and M. Janitz, "Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling," Curr. Genomics, vol.14, no.3, pp.173–181, May 2013.

[70] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," Bioinformatics, vol.29, no.1, pp.15–21, Jan 2013.

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

[72] D. Kim, B. Langmead, and S.L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," Nat. Methods, vol.12, no.4, pp.357–360, Apr 2015.

[73] Broad Institute, "Picard tools." http://broadinstitute.github.io/picard/, (Accessed: 2018/02/21; version 2.17.8).

[74] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics, vol.25, no.14, pp.1754–1760, Jul 2009.

[75] Y. Liao, G.K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," Bioinformatics, vol.30, no.7, pp.923–930, Apr 2014.

[76] P.F. Kuan, D. Chung, G. Pan, J.A. Thomson, R. Stewart, and S. Keles, "A Statistical Framework for the Analysis of ChIP-Seq Data," J Am Stat Assoc, vol.106, no.495, pp.891–903, 2011.

[77] W.N. Venables and B.D. Ripley, Modern Applied Statistics with S, Springer Publishing Company, Incorporated, 2010.

[78] T. Hastie and R. Tibshirani, "Generalized additive models," Statistical Science, vol.1, pp.297–310, 1986.

[79] S.N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," Journal of the Royal Statistical Society (B), vol.73, no.1, pp.3–36, 2011.

[80] G. Wahba, "A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem," The Annals of Statistics, vol.13, 12 1985.

[81] E. Wit, E.v.d. Heuvel, and J.W. Romeijn, "All models are wrong...: an introduction to model uncertainty," Statistica Neerlandica, vol.66, no.3, pp.217–236, 2012.

[82] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, vol.39, no.1, pp.1–38, 1977.

[83] I. Zwiener, B. Frisch, and H. Binder, "Transforming RNA-Seq data to improve the performance of prognostic gene signatures," PLoS ONE, vol.9, no.1, p.e85150, 2014.

[84] Y. Bao, V. Vinciotti, E. Wit, and P.A. 't Hoen, "Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data," BMC Bioinformatics, vol.14, p.169, May 2013.

[85] P. Broet and S. Richardson, "Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model," Bioinformatics, vol.22, no.8, pp.911–918, Apr 2006.

[86] S. Schwartz, M.R. Mumbach, M. Jovanovic, T. Wang, K. Maciag, G.G. Bushkin, P. Mertins, D. Ter-Ovanesyan, N. Habib, D. Cacchiarelli, N.E. Sanjana, E. Freinkman, M.E. Pacold, R. Satija, T.S. Mikkelsen, N. Hacohen, F. Zhang, S.A. Carr, E.S. Lander, and A. Regev, "Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites," Cell Rep, vol.8, no.1, pp.284–296, Jul 2014.

[87] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–update," Nucleic Acids Res., vol.41, no.Database issue, pp.D991–995, Jan 2013.

[88] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T.J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project," Genome Res., vol.22, no.9, pp.1760–1774, Sep 2012.

[89] A.R. Quinlan and I.M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," Bioinformatics, vol.26, no.6, pp.841–842, Mar 2010.

[90] Y. Zeng, S. Wang, S. Gao, F. Soares, M. Ahmed, H. Guo, M. Wang, J.T. Hua, J. Guan, M.F. Moran, M.S. Tsao, and H.H. He, "Refined RIP-seq protocol for epitranscriptome analysis with low input materials," PLoS Biol., vol.16, no.9, p.e2006092, 09 2018.

[91] H. Thorvaldsdottir, J.T. Robinson, and J.P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," Brief. Bioinformatics, vol.14, no.2, pp.178–192, Mar 2013.

[92] B. Uyar, D. Yusuf, R. Wurmus, N. Rajewsky, U. Ohler, and A. Akalin, "RCAS: an RNA centric annotation system for transcriptome-wide regions of interest," Nucleic Acids Res., vol.45, no.10, p.e91, Jun 2017.

[93] A.B. McIntyre, N.S. Gokhale, L. Cerchietti, S.R. Jaffrey, S.M. Horner, and C.E. Mason, "Limits in the detection of m6a changes using merip/m6a-seq," bioRxiv, 2019.

[94] Y. Wang, Y. Li, M. Yue, J. Wang, S. Kumar, R.J. Wechsler-Reya, Z. Zhang, Y. Ogawa, M. Kellis, G. Duester, and J.C. Zhao, "N6-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications," Nat. Neurosci., vol.21, no.2, pp.195–206, 02 2018.

[95] M.I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," Genome Biol., vol.15, no.12, p.550, 2014.

[96] D.J. McCarthy, Y. Chen, and G.K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," Nucleic Acids Res., vol.40, no.10, pp.4288–4297, May 2012.

[97] H. Ji, H. Jiang, W. Ma, and W.H. Wong, "Using CisGenome to analyze ChIP-chip and ChIP-seq data," Curr Protoc Bioinformatics, vol.Chapter 2, p.Unit2.13, Mar 2011.

[98] Y. Li, D.Y. Zhao, J.F. Greenblatt, and Z. Zhang, "RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments," Nucleic Acids Res., vol.41, no.8, p.e94, Apr 2013.

[99] P.J. Uren, E. Bahrami-Samani, S.C. Burns, M. Qiao, F.V. Karginov, E. Hodges, G.J. Hannon, J.R. Sanford, L.O. Penalva, and A.D. Smith, "Site identification in high-throughput RNA-protein interaction data," Bioinformatics, vol.28, no.23, pp.3013–3020, Dec 2012.

[100] X. Chen, D. Chung, G. Stefani, F. Slack, and H. Zhao, "Statistical issues in binding site identification through clip-seq," Statistics and Its Interface, vol.8, pp.419–436, 01 2015.

[101] M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, and J.R. Gonzalez, "A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments," BMC Bioinformatics, vol.14, p.254, Aug 2013.

[102] J. Wang and L. Wang, "Deep analysis of RNA N6-adenosine methylation (m6A) patterns in human cells," NAR Genomics and Bioinformatics, vol.2, no.1, 02 2020. lqaa007.

[103] S. An, W. Huang, X. Huang, Y. Cun, W. Cheng, X. Sun, Z. Ren, Y. Chen, W. Chen, and J. Wang, "Integrative network analysis identifies cell-specific trans regulators of m6A," Nucleic Acids Res., vol.48, no.4, pp.1715–1729, 02 2020.

[104] A.B.R. McIntyre, N.S. Gokhale, L. Cerchietti, S.R. Jaffrey, S.M. Horner, and C.E. Mason, "Limits in the detection of m6A changes using MeRIP/m6A-seq," Sci Rep, vol.10, no.1, p.6590, Apr 2020.

[105] G. Lichinchi, B.S. Zhao, Y. Wu, Z. Lu, Y. Qin, C. He, and T.M. Rana, "Dynamics of Human and Viral RNA Methylation during Zika Virus Infection," Cell Host Microbe, vol.20, no.5, pp.666–673, Nov 2016.

[106] S. Geula, S. Moshitch-Moshkovitz, D. Dominissini, A.A. Mansour, N. Kol, M. Salmon-Divon, V. Hershkovitz, E. Peer, N. Mor, Y.S. Manor, M.S. Ben-Haim, E. Eyal, S. Yunger, Y. Pinto, D.A. Jaitin, S. Viukov, Y. Rais, V. Krupalnik, E. Chomsky, M. Zerbib, I. Maza, Y. Rechavi, R. Massarwa, S. Hanna, I. Amit, E.Y. Levanon, N. Amariglio, N. Stern-Ginossar, N. Novershtern, G. Rechavi, and J.H. Hanna, "Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation," Science, vol.347, no.6225, pp.1002–1006, Feb 2015.

[107] J. Zhou, J. Wan, X.E. Shu, Y. Mao, X.M. Liu, X. Yuan, X. Zhang, M.E. Hess, J.C. Brüning, and S.B. Qian, "N6-Methyladenosine Guides mRNA Alternative Translation during Integrated Stress Response," Mol. Cell, vol.69, no.4, pp.636–647, 02 2018.

[108] J. Zhou, J. Wan, X. Gao, X. Zhang, S.R. Jaffrey, and S.B. Qian, "Dynamic m(6)A mRNA methylation directs translational control of heat shock response," Nature, vol.526, no.7574, pp.591–594, Oct 2015.

[109] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett, "FastQC." Babraham Institute, Jan. 2012.

[110] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," EMBnet.journal, vol.17, no.1, pp.10–12, 2011.

[111] Y. Zhu, G. Xu, Y.T. Yang, Z. Xu, X. Chen, B. Shi, D. Xie, Z.J. Lu, and P. Wang, "POSTAR2: deciphering the post-transcriptional regulatory logics," Nucleic Acids Res., vol.47, no.D1, pp.D203–D211, Jan 2019.

[112] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," Journal of the Royal Statistical Society Series B (Methodological), vol.57, no.1, pp.289–300, 1995.

[113] M. Kuhn, "Building predictive models in r using the caret package," Journal of Statistical Software, Articles, vol.28, no.5, pp.1–26, 2008.

[114] A. Liaw and M. Wiener, "Classification and regression by randomforest," R News, vol.2, no.3, pp.18–22, 2002.

[115] Y. Yue, J. Liu, X. Cui, J. Cao, G. Luo, Z. Zhang, T. Cheng, M. Gao, X. Shu, H. Ma, F. Wang, X. Wang, B. Shen, Y. Wang, X. Feng, C. He, and J. Liu, "VIRMA mediates preferential m6A mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation," Cell Discov, vol.4, p.10, 2018.

[116] Y. Liu, W. Hu, Y. Murakawa, J. Yin, G. Wang, M. Landthaler, and J. Yan, "Cold-induced RNA-binding proteins regulate circadian gene expression by controlling alternative polyadenylation," Sci Rep, vol.3, p.2054, 2013.

[117] Y.W. Asmann, B.M. Necela, K.R. Kalari, A. Hossain, T.R. Baker, J.M. Carr, C. Davis, J.E. Getz, G. Hostetter, X. Li, S.A. McLaughlin, D.C. Radisky, G.P. Schroth, H.E. Cunliffe, E.A. Perez, and E.A. Thompson, "Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer," Cancer Res., vol.72, no.8, pp.1921–1928, Apr 2012.

[118] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D.G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J.B. Brown, L. Lipovich, J.M. Gonzalez, M. Thomas, C.A. Davis, R. Shiekhattar, T.R. Gingeras, T.J. Hubbard, C. Notredame, J. Harrow, and R. Guigó, "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression," Genome Res., vol.22, no.9, pp.1775–1789, Sep 2012.

[119] N.L. Barbosa-Morais, M. Irimia, Q. Pan, H.Y. Xiong, S. Gueroussov, L.J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C.M. Misquitta-Ali, M.D. Wilson, P.M. Kim, D.T. Odom, B.J. Frey, and B.J. Blencowe, "The evolutionary landscape of alternative splicing in vertebrate species," Science, vol.338, no.6114, pp.1587–1593, Dec 2012.

[120] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N.A. Fonseca, R. Petryszak, I. Papatheodorou, U. Sarkans, and A. Brazma, "ArrayExpress update - from bulk to single-cell expression data," Nucleic Acids Res., vol.47, no.D1, pp.D711–D715, Jan 2019.

[121] M.S. Kim, S.M. Pinto, D. Getnet, R.S. Nirujogi, S.S. Manda, R. Chaerkady, A.K. Madugundu, D.S. Kelkar, R. Isserlin, S. Jain, J.K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N.A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L.D. Selvan, A.H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S.K. Sreenivasamurthy, A. Marimuthu, G.J. Sathe, S. Chavan, K.K. Datta, Y. Subbannayya, A. Sahu, S.D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K.R. Murthy, N. Syed, R. Goel, A.A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.C. Huang, J. Zhong, X. Wu, P.G. Shaw, D. Freed, M.S. Zahari, K.K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C.J. Mitchell, S.K. Shankar, P. Satishchandra, J.T. Schroeder, R. Sirdeshmukh, A. Maitra, S.D. Leach, C.G. Drake, M.K. Halushka, T.S. Prasad, R.H. Hruban, C.L. Kerr, G.D. Bader, C.A. Iacobuzio-Donahue, H. Gowda, and A. Pandey, "A draft map of the human proteome," Nature, vol.509, no.7502, pp.575–581, May 2014.

[122] C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork, "STRING: known and predicted protein-protein

associations, integrated and transferred across organisms," Nucleic Acids Res., vol.33, no.Database issue, pp.D433–437, Jan 2005.

[123] Y. Wang, Y. Li, J.I. Toth, M.D. Petroski, Z. Zhang, and J.C. Zhao, "N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells," Nat. Cell Biol., vol.16, no.2, pp.191–198, Feb 2014.

[124] L. Rabiner and B. Juang, "An introduction to hidden Markov models," IEEE ASSP Magazine, vol.3, no.1, pp.4–16, 1986.

[125] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. cite arxiv:1810.04805Comment: 13 pages.

[126] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. cite arxiv:1706.03762Comment: 15 pages, 5 figures.

[127] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation.," vol.14, pp.1532–1543, 2014.

Appendix A.

Chapter 3 Supplementary Materials

Supplementary Text

Negative-Binomial distribution

A negative binomial distribution is defined as $NB(r,p)$ with two parameters r and p, representing size (a shape parameter) and probability, respectively. The density function is

$$P(Y=y)=\frac{\Gamma(y+r)}{y!\Gamma(r)}p^r(1-p)^y \tag{S1.1}$$

with $y=0,1,2,...,r>0$, and $0<p\leq 1$.

As p can be represented by $p=\frac{r}{r+\mu}$, where μ is the mean, the density function can be written as

$$P(Y=y)=\frac{\Gamma(y+r)}{y!\Gamma(r)}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y. \tag{S1.2}$$

Implementation and extension of 1S mode of MOSAiCS[1]

MoAIMS implements and extends the statistical framework proposed by MOSAiCS. The followings provide details of modified 1S mode of MOSAiCS.

A MeRIP-Seq dataset consists of one (IP) sample and one input sample. It is assumed that the observed bin counts of an IP sample follow a mixture (NB) model composed of

a background component and a signal component that are unobserved. Let Z represent the components, where $Z \in \{0,1\}$(0 for the background component and 1 for the signal component), and Y_j is the observed read count of the jth bin; therefore, the mixture model can be written as the following equation,

$$P(Y_j) = (1-\pi_s)P(Y_j|Z_j = 0, \Theta_B) + \pi_s P(Y_j|Z_j = 1, \Theta_s), \tag{S2}$$

where π_s is the *signal proportion*($\pi_s \in [0,1]$), equal to $P(Z_j = 1)$, and $(1-\pi_s)$ is equal to $P(Z_j = 0)$; Θ_B and Θ_s are parameters of background and signal distribution respectively.

When the bin is from the background component, the read count follows the distribution $NB(a, \frac{a}{a+\mu_j})$ which can be written as

$$P(Y_j = y|Z_j = 0) = \frac{\Gamma(y+a)}{y!\Gamma(a)} \left(\frac{a}{a+\mu_j}\right)^a \left(\frac{\mu_j}{a+\mu_j}\right)^y. \tag{S3}$$

When the bin is from the signal component, the read count can be represented as $Y_j = N_j + S_j + k$, where N_j is the count from a non-specific background following $NB(a, \frac{a}{a+\mu_j})$ as defined in Equation(S3), S_j is the count from an actual enrichment following $NB(b, \frac{c}{c+1})$($c = \frac{b}{\mu}$,μ is the mean), and k is the minimal read count required for the signal component. Thus, the distribution of the signal component is a convolution of negative binomials. The convolution of two discrete distributions is defined as $P(X = X_1 + X_2) = P(X_1) * P(X_2) = \sum_{n=0}^{x} P_1(n)P_2(x-n)$, when X_1 and X_2 are two random variables with distributions $P_1(X_1)$ and $P_2(X_2)$, respectively; therefore, the distribution of the signal component can be written as

$$\begin{aligned} &P(Y_j = y-k|Z_j = 1) \\ &\quad = P(S_j) * P(N_j) \\ &\quad = \sum_{q=0}^{y-k} \left[\frac{\Gamma(y-k-q+b)}{(y-k-q)!\Gamma(b)} \left(\frac{c}{c+1}\right)^b \left(\frac{1}{c+1}\right)^{(y-k-q)}\right] \left[\frac{\Gamma(q+a)}{q!\Gamma(a)} \left(\frac{a}{a+\mu_j}\right)^a \left(\frac{\mu_j}{a+\mu_j}\right)^q\right]. \end{aligned} \tag{S4}$$

For estimating the parameters, a and μ_j of the background component are estimated by regression using the input bin counts, while b and c of the signal component and π_s are estimated by expectation maximization(EM).

Each IP bin count Y_j has a corresponding input bin count X_j. For the bins from the background component, it is assumed that $\{Y_j\}(j = 1,2,...,T)$ with the same input bin count

from the same distribution; thus, $\{Y_j\}$ are grouped by the input bin count to $\mathcal{S}_i = \{Y_j | X_j = x_i\}$($x_i$ is the group value equal to available and unique input bin count value(0,1,2,...) and i is the group index). For $Y_j \in \mathcal{S}_i$, it follows that $NB(a, \frac{a}{a+\mu_i})$. Let μ_i be $E(\mathcal{S}_i)$ (the median value of $Y_j \in \mathcal{S}_i$); then, the regression is fitted through RLM or GAM as Equations (S5) and (S6),

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_i), \tag{S5}$$

where β_0 and β_1 are coefficients, and

$$\log(\mu_i) = \beta_0 + f(\log(x_i)|\boldsymbol{\beta}), \tag{S6}$$

where f is represented using thin plate regression splines and $\boldsymbol{\beta}$ is a vector of coefficients for the spline term with length of 9 as default. The regression method is optimized in MoAIMS based on the BIC value.

a is estimated by $\hat{a} = \sum_i n_i \hat{a}_i / \sum_i n_i$, where $\hat{a}_i = [E(\mathcal{S}_i)]^2 / [Var(\mathcal{S}_i) - E(\mathcal{S}_i)]$ (the expectation is calculated using the median value; the variation is calculated using the median absolute deviation) and n_i is the number of bins.

The estimations of π_s, b, and c using EM are shown as follows with the initiation values for π_s, b, and c set empirically to 0.02, 0.2, and 2, respectively. Because this algorithm employed various approximated estimations for efficient calculation, EM cannot ensure the monotonic increase of likelihood. I set the initiation value of the signal proportion low enough so that it is expected to be closer to the real value after each iteration.

The complete data likelihood can be written as Equations (S7),(S8), where T is the number of bins and $I(Z)$ is the indicator function,

$$L = \Pi_{j=1}^{T} [(1-\pi_s) P(Y_j | Z_j = 0, \Theta_B)]^{I(Z_j=0)} + [\pi_s P(Y_j | Z_j = 1, \Theta_s)]^{I(Z_j=1)}, \tag{S7}$$

$$\begin{aligned} \log L = \sum_{j=1}^{T} [& I(Z_j = 0)(\log(1-\pi_s) + \log P(Y_j | Z_j = 0, \Theta_B)) \\ & + I(Z_j = 1)(\log \pi_s) + \log P(Y_j | Z_j = 1, \Theta_s))]. \end{aligned} \tag{S8}$$

The expected complete data likelihood is

$$\begin{aligned} Q = \sum_{j=1}^{T} [& P(Z_j = 0 | Y_j)(\log(1-\pi_s) + \log P(Y_j | Z_j = 0, \Theta_B)) \\ & + P(Z_j = 1 | Y_j)(\log \pi_s + \log P(Y_j | Z_j = 1, \Theta_s))]. \end{aligned} \tag{S9}$$

E-step:

$$\begin{aligned} z_{1,j}^{(t)} &= P(Z_j = 1|Y_j) \\ &= \frac{\pi_s^{(t)} P(Y_j|Z_j = 1, \Theta_s^{(t)})}{(1-\pi_s^{(t)}) P(Y_j|Z_j = 0, \Theta_B^{(t)}) + \pi_s^{(t)} P(Y_j|Z_j = 1, \Theta_s^{(t)})}, \end{aligned} \tag{S10}$$

$$\begin{aligned} z_{0,j}^{(t)} &= P(Z_j = 0|Y_j) \\ &= 1 - z_{1,j}^{(t)}. \end{aligned} \tag{S11}$$

M-step:

For the parameter π_s, to maximize the expected log likelihood with respect to π_s, I obtained

$$\frac{\partial Q}{\partial \pi_s} = \frac{\sum_{j=1}^{T} P(Z_j = 0|Y_j)}{1-\pi_s} - \frac{\sum_{j=1}^{T} P(Z_j = 1|Y_j)}{\pi_s} = 0. \tag{S12}$$

Solving Equation (S12), I obtained

$$\pi_s^{(t+1)} = \frac{1}{T} \sum_{j=1}^{T} z_{1,j}^{(t)}. \tag{S13}$$

For the parameters b and c, the method of moments is used by utilizing

$$\begin{aligned} Var(S_j) &= Var(Y_j|Z_j = 1) - Var(N_j) \\ &= E(S_j) + \frac{E(S_j)^2}{b^{(t)}}, \end{aligned} \tag{S14}$$

$$c^{(t)} = \frac{b^{(t)}}{E(S_j)}. \tag{S15}$$

Solving Equations (S14) and (S15), I obtained

$$b^{(t+1)} = \frac{E(S_j)^2}{Var(Y_j|Z_j = 1) - Var(N_j) - E(S_j)}, \tag{S16}$$

$$c^{(t+1)} = \frac{E(S_j)}{Var(Y_j|Z_j = 1) - Var(N_j) - E(S_j)}, \tag{S17}$$

where $E(S_j) = E(Y_j|Z_j = 1) - E(N_j) - k$. I calculate $E(Y_j|Z_j = 1), Var(Y_j|Z_j = 1), E(N_j)$ and $Var(N_j)$ by,

$$
\begin{aligned}
E(Y_j|Z_j = 1) &= \frac{\sum_{j=1}^{T} z_{1,j}^{(t)} Y_j}{\sum_{j=1}^{T} z_{1,j}^{(t)}}, \\
Var(Y_j|Z_j = 1) &= \frac{\sum_{j=1}^{T} z_{1,j}^{(t)} [Y_j - E(Y_j|Z_j = 1)]^2}{\sum_{j=1}^{T} z_{1,j}^{(t)}}, \\
E(N_j) &= \hat{\mu_0} = \frac{\sum_{j=1}^{T} \hat{\mu}_j}{T}, \\
Var(N_j) &= \hat{\mu_0}(1 + \hat{\mu_0}/\hat{a}),
\end{aligned}
$$

in that $E(Y_j|Z_j = 1)$ and $Var(Y_j|Z_j = 1)$ are the weighted mean and variance, respectively; $E(N_j)$ and $Var(N_j)$ are calculated using the method of moments with $\hat{\mu}_j$ and $\hat{a}$ estimated from the previous steps, of which $\hat{\mu}_j$ is equal to $\frac{\sum_{j=1}^{T} \exp[\hat{\beta}_0 + \hat{\beta}_1 \log(x_j)]}{T}$ using RLM or $\frac{\sum_{j=1}^{T} \exp[\hat{\beta}_0 + f(\log(x_j)|\hat{\boldsymbol{\beta}})]}{T}$ using GAM.

References

[1] Kuan, P.F., Chung, D., Pan, G., Thomson, J.A., Stewart, R., Keles, S.: A Statistical Framework for the Analysis of ChIP-Seq Data. J Am Stat Assoc 106(495), 891–903 (2011)

Appendix B.

Chapter 4 Supplementary Materials

Supplementary Tables

Table B..1 Enrichment ratios for RNA binding proteins (RBPs) associated with the reproducible m6A regions identified in the HEK293T datasets

HEK293T	Enrichment ratio	# m6A regions with RBP	p-value*	FDR adjusted p-value
ALKBH5	1.09	265	0.065	0.171
ATXN2	1.04	5378	<0.001	<0.003
CAPRIN1	1.17	2331	<0.001	<0.003
CNBP	1.34	32	0.052	0.142
CPSF1	0.96	558	0.850	1.000
CPSF2	1.09	124	0.157	0.384
CPSF3	0.93	680	0.988	1.000
CPSF4	0.89	506	1.000	1.000
CPSF6	1.34	3593	<0.001	<0.003
CPSF7	1.31	4413	<0.001	<0.003
CSTF2	0.68	1104	1.000	1.000
CSTF2T	0.86	2738	1.000	1.000
DDX3X	1.44	9470	<0.001	<0.003
DGCR8	1.07	67	0.273	0.606
DICER1	0.94	92	0.773	1.000
DIS3L2	0.46	11	1.000	1.000
EIF3A	1.39	293	<0.001	<0.003
EIF3B	1.10	255	0.073	0.185
EIF3D	1.88	593	<0.001	<0.003

Continued on next page

Table B..1 – *Continued from previous page*

HEK293T	Enrichment ratio	# m6A regions with RBP	p-value*	FDR adjusted p-value
EIF3G	1.14	453	0.001	0.003
ELAVL1	0.82	4594	1.000	1.000
EWSR1	0.75	853	1.000	1.000
FBL	1.22	88	0.026	0.074
FIP1L1	1.16	3419	<0.001	<0.003
FMR1	1.46	4443	<0.001	<0.003
FUS	0.79	1179	1.000	1.000
FXR1	0.95	178	0.795	1.000
FXR2	1.23	966	<0.001	<0.003
HNRNPA1	0.60	53	1.000	1.000
HNRNPA2B1	0.67	11	0.948	1.000
HNRNPC	0.95	5404	1.000	1.000
HNRNPD	0.29	123	1.000	1.000
HNRNPF	0.95	35	0.661	1.000
HNRNPH1	1.57	47	0.002	0.006
HNRNPM	0.34	7	1.000	1.000
HNRNPU	0.31	2	0.996	1.000
IGF2BP1	1.07	1847	<0.001	<0.003
IGF2BP2	0.91	1583	1.000	1.000
IGF2BP3	0.77	1796	1.000	1.000
LIN28A	1.01	791	0.426	0.864
LIN28B	1.06	4997	<0.001	<0.003
MOV10	0.60	2001	1.000	1.000
NCBP3	1.29	529	<0.001	<0.003
NOP56	0.82	20	0.860	1.000
NOP58	1.74	159	<0.001	<0.003
NUDT21	1.48	5201	<0.001	<0.003
PRKRA	1.65	5	0.176	0.417
PTBP1	0.61	1397	1.000	1.000
PUM2	0.41	108	1.000	1.000
QKI	0.48	32	1.000	1.000
RBM10	1.20	17	0.251	0.575
RBM15	2.73	3534	<0.001	<0.003
RBM15B	2.32	6375	<0.001	<0.003
RBPMS	0.68	87	0.999	1.000
RTCB	1.17	619	<0.001	<0.003
SRRM4	0.95	490	0.908	1.000
SSB	1.06	73	0.319	0.686

Continued on next page

Table B..1 – *Continued from previous page*

HEK293T	Enrichment ratio	# m6A regions with RBP	p-value*	FDR adjusted p-value
STAU1	0.23	43	1.000	1.000
TAF15	0.76	239	1.000	1.000
TARBP2	0.66	3	0.848	1.000
TARDBP	0.90	4194	1.000	1.000
TNRC6A	1.05	9	0.478	0.943
TNRC6B	1.18	11	0.338	0.706
TNRC6C	0.38	2	0.981	1.000
WDR33	0.58	650	1.000	1.000
YTHDC1	2.15	7224	<0.001	<0.003
YTHDC2	0.96	36	0.636	1.000
YTHDF1	2.49	9196	<0.001	<0.003
YTHDF2	3.90	6964	<0.001	<0.003
YTHDF3	2.70	52	<0.001	<0.003
ZC3H7B	0.62	1889	1.000	1.000

* P-values were calculated from 1000 times of permutation. When p-value is zero, it is shown in the table as < 0.001 becasue it is possible that the p-value is actually less than 0.001 if times of permutation were increased.

Table B..2 Enrichment ratios for RNA binding proteins (RBPs) in the reproducible m6A regions identified in MEF studies

MEF	Enrichment ratio	# m6A regions with RBP	p-value*	FDR adjusted p-value
CIRBP	1.76	401	<0.001	<0.001
CPSF6	2.07	94	<0.001	<0.001
CREBBP	2.47	24	<0.001	<0.001
MBNL1	2.16	8	0.040	0.045
MBNL2	1.55	5	0.216	0.216
MBNL3	NA	1	NA	NA
RBM3	5.81	485	<0.001	<0.001
SRSF1	2.13	467	<0.001	<0.001
SRSF2	2.24	793	<0.001	<0.001

* P-values were calculated from 1000 times of permutation. When p-value is zero, it is shown in the table as < 0.001 becasue it is possible that the p-value is actually less than 0.001 if times of permutation were increased.

Table B..3 Overlapped RNA binding proteins (RBPs) for proteins enriched in the m6A-HEK293T dataset

RBPs enriched in m6A	RBPs with overlapping ratio more than 60%*
YTHDF2	YTHDF1(89.7%), DDX3X(80.5%), YTHDC1(68.5%), RBM15B(60.6%)
RBM15	YTHDF1(86.8%), RBM15B(83.6%), DDX3X(82.8%), YTHDC1(81.7%), YTHDF2(74.6%)
YTHDF1	DDX3X(77.4%), YTHDF2(67.9%), YTHDC1(63.0%)
RBM15B	YTHDF1(80.4%), DDX3X(76.9%), YTHDC1(75.7%), YTHDF2(66.2%)
YTHDC1	YTHDF1(80.2%), DDX3X(76.3%), RBM15B(66.8%), YTHDF2(66.0%)
EIF3D	DDX3X(87.0%), YTHDF1(83.1%), YTHDC1(72.0%), YTHDF2(68.3%), RBM15B(65.6%)
NOP58	DDX3X(81.1%), YTHDC1(77.4%), YTHDF1(75.5%), RBM15B(74.2%), CPSF7(73.0%), NUDT21(71.1%), HNRNPC(70.4%), LIN28B(66.7%), YTHDF2(63.9%), FMR1(62.3%), CPSF6(60.4%)
NUDT21	DDX3X(76.7%), YTHDF1(75.7%), YTHDC1(67.7%), YTHDF2(60.5%)
FMR1	DDX3X(87.2%), YTHDF1(83.1%), YTHDF2(69.5%), YTHDC1 (69.1%)
DDX3X	YTHDF1(75.1%)
EIF3A	DDX3X(85.0%), YTHDF1(80.9%), YTHDC1(70.3%), YTHDF2(68.9%), RBM15B(68.6%)
CPSF6	DDX3X(84.5%), YTHDF1(82.0%), YTHDC1(72.6%), YTHDF2(68.7%), CPSF7(67.2%), HNRNPC(65.0%), RBM15B(63.9%), NUDT21(61.9%)
CPSF7	DDX3X(83.3%), YTHDF1(82.4%), YTHDC1(75.0%), YTHDF2(68.5%), HNRNPC(64.4%), RBM15B(63.4%), NUDT21(60.5%)

*Numbers in brackets indicate the percentage of overlap and are listed in decreasing order.

List of Academic Achievements

Category	Description
Journal	[1] **Yiqian Zhang** and Michiaki Hamada, DeepM6ASeq: Prediction and Characterization of m6A-containing Sequences using Deep Learning, BMC Bioinformatics. 2018 Dec 31;19(Suppl 19):524. [2] **Yiqian Zhang** and Michiaki Hamada, MoAIMS: Efficient Software for Detection of Enriched Regions of MeRIP-Seq, BMC Bioinformatics. 2020 Mar 14;21(1):103.
International Conferences	[1] **Yiqian Zhang** and Michiaki Hamada, DeepM6ASeq: Prediction and Characterization of m6A-containing Sequences using Deep Learning, Genome Informatics Workshop (GIW), Dec. 03-05, 2018, Kunming, China (Talk).
Domestic Conference	[1] **Yiqian Zhang** and Michiaki Hamada, Prediction of mRNA m6A sites in the mammalian genome, 第6回生命医薬情報学連合大会 (IIBMP), Sep. 27-29, 2017, 札幌 (Poster). [2] **Yiqian Zhang** and Michiaki Hamada, Prediction and Characterization of m6A-contained Sequences using Deep Neural Network, バイオ情報学研究会(IPSJ-BIO), Jun. 13-15, 2018, 沖縄 (Talk). [3] **Yiqian Zhang** and Michiaki Hamada, DeepM6ASeq: Prediction and Characterization of m6A-contained Sequences using Deep Neural Networks, 第 20回日本RNA学会年会, Jul. 09-11, 2018, 大阪 (Poster). [4] **Yiqian Zhang** and Michiaki Hamada, Developing an efficient software for MeRIP-Seq signal detection， 第 3 回 Tokyo Bioinformatics Meeting研究会，Aug. 29, 2019, 東京 (Talk). [5] **Yiqian Zhang** and Michiaki Hamada, Developing efficient software for detection of enriched regions of MeRIP-Seq, 第42回日本分子生物学会年会, Dec. 03-06, 2019, 福岡 (Poster).