Bioinformatics, 32, 2016, i378–i385
doi: 10.1093/bioinformatics/btw281
ISMB 2016

OXFORD

A novel algorithm for calling mRNA m^{6}A peaks by modeling biological variances in MeRIP-seq data

Xiaodong Cui[1], Jia Meng[2], Shaowu Zhang[3], Yidong Chen[4,5] and Yufei Huang[1,5,*]

[1]Department of Electrical and Computer Engineering, University of Texas at San Antonio, TX 78249, USA, [2]Department of Biological Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China, [3]College of Automation, Northwestern Polytechnical University, Xi'an 710072, China, [4]Greehey Children's Cancer Research Institute and [5]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, TX 78229, USA

*To whom correspondence should be addressed.

Abstract

Motivation: *N*^{6}-methyl-adenosine (m^{6}A) is the most prevalent mRNA methylation but precise prediction of its mRNA location is important for understanding its function. A recent sequencing technology, known as Methylated RNA Immunoprecipitation Sequencing technology (MeRIP-seq), has been developed for transcriptome-wide profiling of m^{6}A. We previously developed a peak calling algorithm called exomePeak. However, exomePeak over-simplifies data characteristics and ignores the reads' variances among replicates or reads dependency across a site region. To further improve the performance, new model is needed to address these important issues of MeRIP-seq data.

Results: We propose a novel, graphical model-based peak calling method, MeTPeak, for transcriptome-wide detection of m^{6}A sites from MeRIP-seq data. MeTPeak explicitly models read count of an m^{6}A site and introduces a hierarchical layer of Beta variables to capture the variances and a Hidden Markov model to characterize the reads dependency across a site. In addition, we developed a constrained Newton's method and designed a log-barrier function to compute analytically intractable, positively constrained Beta parameters. We applied our algorithm to simulated and real biological datasets and demonstrated significant improvement in detection performance and robustness over exomePeak. Prediction results on publicly available MeRIP-seq datasets are also validated and shown to be able to recapitulate the known patterns of m^{6}A, further validating the improved performance of MeTPeak.

Availability and implementation: The package 'MeTPeak' is implemented in R and C++, and additional details are available at https://github.com/compgenomics/MeTPeak

Contact: yufei.huang@utsa.edu or xdchoi@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Compared with DNA methylation, which is a well-established and extensively studied epigenetic phenomenon, m^{6}A methylation in mRNA is still largely an uncharted territory. Recent surge of interest comes mainly as a result of two groundbreaking studies (Dominissini *et al.*, 2012; Meyer *et al.*, 2012) that show high abundancy of m^{6}A in >25% of transcripts in human and mouse cells. A number of subsequent studies have provided significant insights into the mechanism as well as the functions of mRNA m^{6}A methylation. It is known now that the m^{6}A methyltransferase complex, or the m^{6}A writer, consists of Wilms' tumor 1-associating protein (WTAP), methyltransferase-like 3 (METTL3) and methyltransferase-like 14 (METTL14) (Liu *et al.*, 2014; Nilsen, 2014; Wang *et al.*, 2014b). Different from DNA methylation, m^{6}A methylation is reversible and

can be removed by m^{6}A demethylase, or m^{6}A erasers, such as the fat mass- and obesity-associated protein (FTO) (Fu *et al.*, 2014; Meyer *et al.*, 2012). In addition, the YTH protein domain functions as an m^{6}A reader that recruits the binding of the m^{6}A methyltransferase to mRNA (Wang *et al.*, 2014a). The mRNA m^{6}A methylation has been shown to play a role in mRNA stability, processing and translational efficiency (Alarcon *et al.*, 2015; Niu *et al.*, 2013). These recent advances notwithstanding, the fact of wide-spread m^{6}A in transcriptome suggests that m^{6}A is potential involved in many other unknown biological processes.

These recent breakthroughs of m^{6}A are largely attributed to the development of the Methylated RNA Immunoprecipitation Sequencing technology, or MeRIP-seq (also known as 'm^{6}A-seq'), which is designed to profile transcriptome-wide m^{6}A sites. MeRIP-seq is technically a synthesis of two well-established methods: chromatin immunoprecipitation sequencing (Kidder *et al.*, 2011) and RNA sequencing (RNA-seq) (Garber *et al.*, 2011). In MeRIP-seq, mRNAs are first fragmented into approximately 100-nucleotide-long fragments. A large portion of the fragments are immunoprecipitated by anti-m^{6}A antibody and subsequently measured by high-throughput sequencing to form Immunoprecipitation Sequencing (IP) samples. In addition, input samples, which are used to measure the background mRNA abundance for the IP experiments, are generated by sequencing the un-immunoprecipitated mRNA fragments. To predict m^{6}A sites, the IP and input reads are aligned to the transcriptome and the reads enrichment of IP over the combined reads in IP and input are assessed.

To meet the computational needs for MeRIP-seq-based m^{6}A peak detection, we previously developed the exomePeak R package (Meng *et al.*, 2014). Basically, exomePeak calculates the overall methylation degree by dividing the IP reads by the sum of IP and input reads. Given the hypothesis that all the IP reads at a particular location follow the same Binomial distribution, parameterized by the overall methylation degree, exomePeak computes a C test (Przyborowski and Wilenski, 1940) to determine the methylation sites. Although exomePeak could achieve fairly robust m^{6}A detection, there are still two major limitations. First, exomePeak does not model the reads variance within transcripts and across replicates. Usually, highly fluctuating read enrichments can be observed around an m^{6}A site and in MeRIP-seq replicates. However, exomePeak assumes that the site regions share the same reads enrichment. Second, exomePeak ignores the dependency of reads enrichment along the mRNA transcripts and therefore could miss the true peaks with low enrichment or erroneously predict the noisy outlier as true peaks. To address these issues, we introduce in this article a new algorithm, MeTPeak. MeTPeak deploys a hierarchical beta-binomial model to model the variance of reads enrichment and a Hidden Markov Model (HMM) to account for the dependency of neighboring enrichment. MeTPeak is an open-source R package, where core heavy computation part of the algorithm is written in C++.

2 Methods

2.1 Graphical representation of MeTPeak

MeTPeak detects m^{6}A peaks on each gene separately, where a particular gene is first divided into N mutually connected bins whose length is equal to the sequencing fragment L, MeTPeak models the reads dependency of the continuous bins by an HMM and the reads count in each bin by the mixture of Beta-binomial distribution. Now

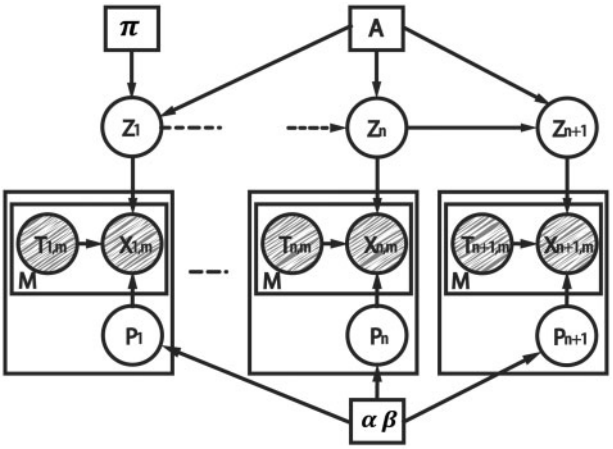

Fig. 1. Graphical Model of MeTPeak for a gene. N is the total number of bins, M is the number of replicates, $X_{n,m}$ is the observed IP reads and $T_{N,m}$ is the total number of observed reads in the m_{th} replicate at the n_{th} bin. P is the methylation degree, Z is a latent variable that denotes the methylation status, indicating whether this is a peak or not. A, π, α, β are the unknown parameters. (See detailed explanation in main text.)

assume that there are M pairs of IP and input replicate samples. For bin n, let X_{mn} denote the reads count in the m^{th} IP sample and Y_{mn} the counts in the m^{th} input samples. We assume that the read counts follow the Poisson distribution as

$$\begin{aligned} P(X_{mn}) &= Pois(S_{IP,m}\lambda_{IP,m}) \\ P(Y_{mn}) &= Pois(S_{ctrl,m}\lambda_{ctrl,m}) \end{aligned} \tag{1}$$

where $S_{IP,m}$ and $S_{ctrl,m}$ are the total reads (sequencing depth) in the m^{th} IP and the m^{th} input samples, respectively, and $\lambda_{IP,m}$ and $\lambda_{ctrl,m}$ are the respective normalized Poisson rates. Let $Z_n \in [1,2]$ denote the unknown hidden methylation status, with 1 representing methylated and 2 otherwise. As shown in Fig. 1, given (1), the conditional probability of observing X_{mn} given the total counts $T_{mn} = X_{mn} + Y_{mn}$ follows the binomial distribution

$$P(X_{mn}|p_n, Z_n, T_{mn}) = \prod_{k=1}^{2}\left(\binom{T_{mn}}{X_{mn}} p_n{}^{X_{mn}}(1-p_n)^{Y_{mn}}\right)^{I(Z_n=k)} \tag{2}$$

where p_n represents methylation degree at the n_{th} bin, which is proportional to reads count in the IP over the total counts at bin n, $I(\cdot)$ is the indicator function. In addition, to model the read counts variances, p_n is further assumed to follow *a priori* the Beta distribution,

$$P(p_n|Z_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{k=1}^{2}\left(\frac{\Gamma(\alpha_k)\Gamma(\beta_k)}{\Gamma(\alpha_k+\beta_k)} p_n{}^{\alpha_k-1}(1-p_n)^{\beta_k-1}\right)^{I(Z_n=k)} \tag{3}$$

where $\boldsymbol{\alpha} = [\alpha_1\alpha_2]^{\mathrm{T}}$, $\boldsymbol{\beta} = [\beta_1, \beta_2]^{\mathrm{T}}$ are the unknown parameters. After integrating p_n from (2) and (3), X_{mn} follows the Beta-binomial distribution

$$P(X_{mn}|Z_n;\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{k=1}^{2}\left(C \bullet \frac{\Gamma(X_{mn}+\alpha_k)\Gamma(Y_{mn}+\beta_k)}{\Gamma(T_{mn}+\alpha_k+\beta_k)} \frac{\Gamma(\alpha_k+\beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}\right)^{I(Z_n=k)} \tag{4}$$

where $C = \Gamma(T_{mn}+1)/(\Gamma(X_{mn}+1)\cdot\Gamma(Y_{mn}+1))$ is the normalization constant. It becomes clear from (4) that the parameters α and β are shared for all bins across the replicates and they function to model

the variance of reads in M replicates. As a result, the joint log-likelihood function of all replicates can be expressed as

$$l = \log\sum_{\mathbf{Z}} P(\mathbf{X},\mathbf{Z}|\Theta) = \log\sum_{\mathbf{Z}} P(Z_1|\pi)\prod_{n=2}^{N}P(Z_n|Z_{n-1},A)\prod_{n=1}^{N}P(\mathbf{X_n}|Z_n,\boldsymbol{\alpha},\boldsymbol{\beta})$$
$$= \log\left(\sum_{\mathbf{Z}}\left[\begin{array}{l}\prod_{k=1}^{2}\pi_k^{I(z_1=k)}\prod_{n=2}^{N}\prod_{i=1}^{2}\prod_{j=1}^{2}A_{ij}^{I(z_{n-1}=i,z_n=j)} \\ \cdot\prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{2}\left(C\cdot\frac{\Gamma(X_{mn}+\alpha_k)\Gamma(Y_{mn}+\beta_k)}{\Gamma(T_{mn}+\alpha_k+\beta_k)}\frac{\Gamma(\alpha_k+\beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}\right)^{I(z_n=k)}\end{array}\right]\right) \quad (5)$$

where
$\boldsymbol{\Theta} = [\boldsymbol{\alpha},\boldsymbol{\beta},\mathrm{A},\pi]$, $\boldsymbol{Z} = [\mathrm{Z}_1,\mathrm{Z}_2,\ldots,\mathrm{Z}_\mathrm{N}]^\mathrm{T}$, $\boldsymbol{X} = [\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ldots,\boldsymbol{X}_\mathbf{N}]^\mathrm{T}$ and $\boldsymbol{X_n} = [X_{1n},X_{2n},\ldots,X_{Mn}\ ]^T$.

2.2 Sites detection and parameter inference

Identifying methylation sites requires inferring the hidden methylation status variable $\boldsymbol{Z}$ and the model parameters $\boldsymbol{\Theta}$. However, given the non-linearity of the likelihood function (5), no closed-form solution for the maximum likelihood estimators can be obtained. Alternatively, we turn to the expectation maximization (EM) framework for a numerical solution (Lindstrom and Bates, 1988). Furthermore, as the parameters for Beta-binomial distribution $[\boldsymbol{\alpha},\boldsymbol{\beta}]$ are constrained to be positive, we device a barrier function and propose a constrained maximization procedure to achieve that goal.

The EM algorithm estimates $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ in an iterative fashion, where each iteration consists of an E step and an M step. In the E step, the posterior distributions of $\boldsymbol{Z}$ given an initial estimate or the previously computed parameters $\boldsymbol{\Theta^{old}}$ is calculated. To do so, the joint marginal $P(Z_n,\boldsymbol{X})$ is first computed through the forward–backward recursion algorithm (Supplementary Equation (9–11)), from which the conditional posterior distributions $P(Z_n|\boldsymbol{X},\boldsymbol{\Theta^{old}})$ and $P\left(Z_{n-1},Z_n\boldsymbol{X},\boldsymbol{\Theta^{old}}\right)$ can be subsequently obtained.

In the M step, a new estimate of $\boldsymbol{\Theta^{new}}$ is obtained through maximizing $Q(\boldsymbol{\Theta},\boldsymbol{\Theta^{old}})$, which is a lower bound of l in (5) and has the expression

$$Q(\Theta,\Theta^{old}) = \sum_{\mathbf{Z}} q(\mathbf{Z})\log(P(Z_1|\pi)\prod_{n=2}^{N}P(Z_n|Z_{n-1},A)\prod_{n=1}^{N}P(\mathbf{X_n}|Z_n,\boldsymbol{\alpha},\boldsymbol{\beta})) \quad (6)$$

where $q(\boldsymbol{Z})$ is a shorthand expression of $P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta^{old}})$. Among $\boldsymbol{\Theta}$, closed-form expressions can be obtained for $[A,\pi]$ and the detailed derivations are provided in Supplementary (S14–S17). However, because the derivative of $Q(\Theta,\Theta^{old})$ w.r.t. $[\boldsymbol{\alpha},\boldsymbol{\beta}]$ has no closed form and $[\boldsymbol{\alpha},\boldsymbol{\beta}]$are constrained to be positive, an estimate of $[\boldsymbol{\alpha},\boldsymbol{\beta}]$ cannot be computed directly. To address this difficulty, we propose a constrained Newton's method, whose optimization objective is formulated as

$$\begin{aligned}&\text{minimize}\quad -s\\ &\text{s.t.}\quad F\Delta \le h\end{aligned} \quad (7)$$

where s is an equivalent form $Q(\boldsymbol{\Theta},\boldsymbol{\Theta^{old}})$ containing only the terms w.r.t $[\boldsymbol{\alpha},\boldsymbol{\beta}]$, $\Delta = \left[\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T\right]^T$ and F and h are the coefficients for the inequality constrains. To constrain $\Delta > 0$ as desired, h is set to zero and F is set to a diagonal matrix with -1s as the diagonal elements; however, this general inequality constrain formulation allows inclusion of any linear constrains such as an upper limit on $[\boldsymbol{\alpha},\boldsymbol{\beta}]$. The basic idea to solve (7) is to use a logarithmic barrier to make the inequality constrain implicit, i.e.

$$\text{minimize}\quad -s - \frac{1}{t}\sum_{i=1}^{m}\log(h_i - f_i^T\Delta) \quad (8)$$

where $t > 0$ is a parameter that sets the accuracy of estimation, where the estimation becomes more accurate as the t increases, m is the number of inequality constrains, and $f_1^T,.,\ f_m^T$ are the rows of F. Minimization of (9) can be carried out for each component separately and the parameters Δ_k for the kth component for $k \in [1,2]$ is calculated iteratively as

$$\Delta_k{}^{new} = \Delta_k{}^{old} - H_k{}^{-1}J_k \quad (9)$$

where J_k and H_k are the gradient and Hessian of (9) w.r.t. $\Delta_k = [\alpha_k,\beta_k]^T$. Specifically, the gradient has the form

$$J_k = -t\bullet\left[\begin{array}{l}\sum_{n=1}^{N}q(Z_{nk})\left[\begin{array}{l}\Phi(\alpha_k+\beta_k)\cdot M - \sum_{m=1}^{M}\Phi(T_{mn}+\alpha_k+\beta_k)\\ -\Phi(\alpha_k)\cdot M + \sum_{m=1}^{M}\Phi(X_{mn}+\alpha_k)\end{array}\right]\\ \sum_{n=1}^{N}q(Z_{nk})\left[\begin{array}{l}\Phi(\alpha_k+\beta_k)\cdot M - \sum_{m=1}^{M}\Phi(T_{mn}+\alpha_k+\beta_k)\\ -\Phi(\beta_k)\cdot M + \sum_{m=1}^{M}\Phi(Y_{mn}+\beta_k)\end{array}\right]\end{array}\right] - F_k{}^{\mathrm{T}}d_k, \quad (10)$$

where F_k and h_k are elements in F and h that correspond to Δ_k, $q(Z_{nk}) = P(Z_n = k|\boldsymbol{X})$, which is the posterior probability of $Z_n = k$, and $\Phi(\cdot) = \log\Gamma(\cdot)$, and $d_k = F_k\Delta_k - h_k$. The Hessian can be expressed as

$$H_k = -t\cdot S_k + F_k{}^{\mathrm{T}}diag(d_k)F_k, \quad (11)$$

where S_k is the Hessian for s in (10) with respect to Δ_k, whose detailed form can be found in Supplementary Equation (23–24).

Taken together, the proposed MeTPeak algorithm can be summarized as follows:

For every gene in the list
  Initialization: choose an initial random $\boldsymbol{\Theta}$
  Repeat: until the increment of likelihood in (5) $\le 10^{-5}$
    E step: use the previous estimate $\boldsymbol{\Theta_{m-1}}$ to update the posterior probability $q(Z_n)$ in (6)
    M step: optimize the parameters $\boldsymbol{\Theta_m}$ through maximizing $Q(\boldsymbol{\Theta},\boldsymbol{\Theta^{old}})$ in (6)
      Compute parameters $[A,\pi]$ according to Supplementary Equation (14–17)
      Initialize $t = 1$ in (8)
      Repeat: until $t \ge 10^6$
        Compute gradient and Hessian according to (10-11)
        Compute Δ by Newton's method in (9)
        Increase t
Report peak according to the threshold FDR (See in Supplementary)

3 Results

3.1 Performance evaluation by simulation

We started by first evaluating the performance of the proposed constrained Newton's method for estimating the parameters $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ by examining the goodness-of-fit to the mixture of Beta-binomials distribution. The results showed that MeTPeak can accurately fit the mixture under a variety of variance conditions (Supplementary Fig. S1). Then, we assessed the performance of MeTPeak using the simulated MeRIP-seq dataset where the true methylation sites were known according to the proposed generative model described in Section 2.1. The reads for each gene were simulated independently, where the count for each bin was generated according to the HMM-based Beta-binomial model. To obtain proper parameters for α and β that mimic the methylation sites, we examined the genes with methylation degree p_n and fitted the Beta distributions from a real MeRIP-seq case-control study as shown in Supplementary Fig. S2. The estimated parameters for the site region was $[\alpha_1, \beta_1] = [6, 2]$, with the mean of methylation degree $p_{mean} = 0.75$, the parameters for non-peak region were set to $[\alpha_2, \beta_2] = [1, 5]$, and the proportion weight for the peak versus non-peak region was set to $w = [0.3, 0.7]$. For each of the following experiments, MeRIP-seq reads of 1000 genes with methylation sites were simulated according to the above-defined parameters. Detection results of both MeTPeak and exomePeak were obtained and to evaluate the performance, areas under the receiver operating characteristic curves (AUCs) were obtained for both methods. Unless otherwise specified, two replicates were simulated for each experiment.

3.1.1 MeTPeak is robust against data variance

We first investigated the detection robustness of MeTPeak by considering different variances of data because in the real MeRIP-seq datasets, the reads variation in peak regions across these replicates can vary dramatically. Particularly, a number of beta distributions of peak regions with variances designed to range from 0.014 to an extreme high level 0.134, corresponding to $[\alpha, \beta] = [0.3, 0.1]$, were simulated to generate reads, where $p_{mean} = 0.75$ was held for each case. The AUC curves in Fig. 2 show that MeTPeak was highly robust against the change of variance and achieved close to >95% AUC even at the highest variance. In contrast, the performance of exomePeak clearly degraded with variance and at the highest variance level, it dropped >20% in AUC and more than four times than MeTPeak. This is because exomePeak assumed that the methylation degree from all the sites are the same, which obviously violates the real case of MeRIP-seq data.

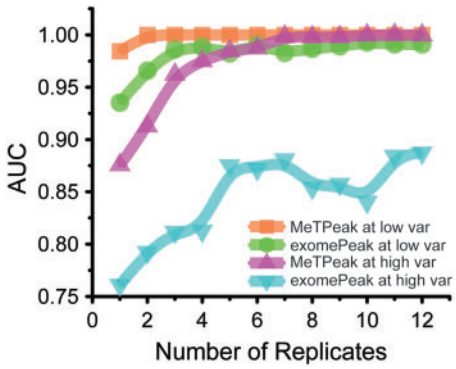

Fig. 2. MeTPeak achieves higher AUCs and is robust against the increase of variance

3.1.2 MeTPeak is robust for small replicates

We next evaluated the influence of MeRIP-seq replicates on the detection performance. Usually in the MeRIP-seq experiments, it is hard to get more than four biological replicates, and robustness under the condition for small replicates is useful for the peak-calling algorithms. Simulated data with the number of replicates varying from 1 to 12 were generated under two variances conditions and results are shown in Fig. 3. AUCs of both methods increased with the number of replicates. However, MeTPeak was robust and achieved about 5% and 10% more AUC than exomePeak even with one replicate at low and high data variance, respectively. In addition, even under the condition of high data variance, with more than four replicates, MeTPeak could excel the AUC of exomePeak in low data variance. In contrast, exomePeak can barely reach that of MeTPeak at high variance despite the increase of replicates. This improvement of MeTPeak is mainly owing to careful modeling of reads variance and replicates.

3.1.3 MeTPeak is robust against data outlier

We next considered a specific case in which MeRIP-seq data contain reads outlier, a scenario where counts in some replicates were significantly larger than those in the remaining replicates. In such case, because exomePeak is likely to produce erroneous predictions because its decision is based on the average counts of all the replicates, and the outlier reads could significantly skew the true distribution of the reads, leading to erroneous predictions. To evaluate the impact of reads outlier on MeTPeak performance, we simulated a case with four replicates and three samples were generated with parameter $[\alpha, \beta] = [6, 2]$ for the methylation peaks, whereas in the fourth, outlier sample, peaks were simulated from different locations and reads counts in peak regions were made to artificially vary from 1 to 15 times of the averaged counts of peak regions in the other three replicates. As shown in Fig. 4, the performance of exomePeak drops significantly with the increase of outlier counts. In contrast, MeTPeak performance is highly robust (~95% of AUC) against different levels of outlier.

3.1.4 MeTPeak is sensitive to lowly enriched peaks

Furthermore, we examined the detection sensitivity, especially for lowly enriched peaks. We simulated the data with varying methylation degree from a low methylation degree with $p_{mean} = 0.33$, to a

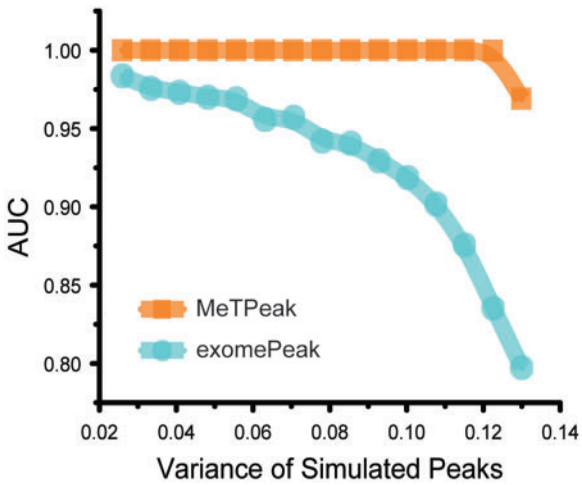

Fig. 3. MeTPeak achieves good performance under high variance with small number of replicates

highly enriched with $p_{mean} = 0.89$. The result is depicted in Fig. 5 and it demonstrates that at all tested methylation degrees, MeTPeak outperformed exomePeak for detecting lowly enriched peaks. Even at an extremely low methylation degree of $p_{mean} = 0.33$, a scenario where the read counts in IP samples are half of that in input samples, MeTPeak still can achieve >70% AUC, about 5% higher than exomePeak. This improvement by MeTPeak in detecting lowly enriched peaks demonstrates the advantage of modeling the dependency of reads.

3.2 Performance evaluation using real MeRIP-seq datasets

To further validated MeTPeak, we applied it to real publically available MeRIP-seq datasets. The first two MeRIP-seq datasets are from the study of the function of FTO, where FTO is a demethylase that functions to remove m^{6}A methylation. In this study, FTO was knocked out from mouse midbrain cells, and the MeRIP-seq datasets before (wildtype or WT-FTO) and after knock-out (KO-FTO) were obtained (Meyer *et al.*, 2012). We should expect an increase of m^{6}A methylation after FTO knockout. The other four datasets come from the experiments that studied the roles of the components in human methyltransferase complex, namely METTL3, METTL14 and WTAP. The datasets include MeRIP-seq samples from the wild-type HeLa cells (WT-HeLa) and HeLa cells after knocking out each of the three proteins (KO-METTL3, KO-METTL14 and KO-WTAP) (Liu *et al.*, 2014). In this case, we should expect a decrease of m^{6}A methylation in knockout cells. The datasets were downloaded from Gene Expression Omnibus under the accession number GSE47217 for KO-FTO experiment and GSE46705 for the others. The dataset of WT-FTO (KO-FTO) includes three replicates for each of WT-FTO IP samples (KO-FTO IP) and WT-FTO input samples (KO-FTO input). Each dataset of KO-METTL3, KO-METTL14 and KO-WTAP has two IP and input replicates and WT-HeLa has four replicates, respectively. For all the samples, reads were aligned to the corresponding transcriptome (UCSC mm9 or UCSC hg19) by TopHat2 (Kim *et al.*, 2013). After quality control (QC > 30), the depth of the samples was normalized to the geometric mean.

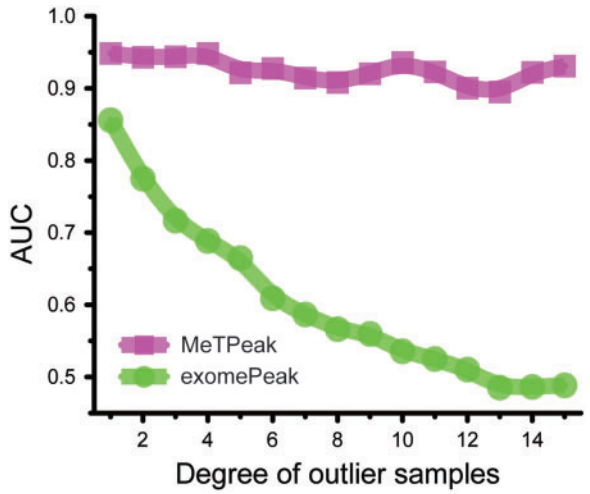

Fig. 4. MeTPeak is robust against biased reads

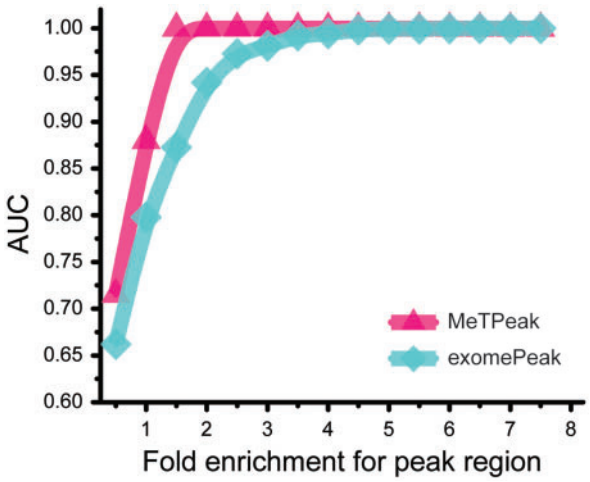

Fig. 5. MeTPeak is sensitive to lowly enriched peaks

We assessed the prediction specificity of MeTPeak and exomePeak by evaluating their false positives (FPs), i.e. the percentage of falsely detected peaks among all un-methylated peaks. To calculate false-positive rates, we need to know the true unmethylation peaks, which however were not readily available. To circumvent this difficulty, we created artificial datasets, where for each dataset, we replaced input (or IP) samples by a subset of IP (or input) replicate samples and treat them as pseudo input (or IP) samples for a particular condition WT (or KO). Because the IP and input samples in each artificial dataset are replicate samples from the same IP (or input), there should not be any reads enrichment and thus any detected peaks should be FPs. Eight artificial datasets were generated as WT-FTO IP versus WT-FTO IP, WT-FTO input versus WT-FTO input, KO-FTO IP versus KO-FTO IP, KO-FTO input versus KO-FTO input. Each dataset can be manipulated to generate two pseudo datasets by the strategy of using IP versus IP and input versus input (i.e. WT-FTO IP versus WT-FTO IP, WT-FTO input versus WT-FTO input). Therefore, a method has higher specificity if it detects less number of m^{6}A sites across all these artificial datasets at the same FDR threshold.

To evaluate the specificity at different detection thresholds, a series of FDR thresholds (0.00001–0.1) was evaluated for both MeTPeak and exomePeak. At a particular FDR, a ratio of FPs (RFP) was computed as number of FPs by MeTPeak divided by that by exomePeak. MeTPeak will have higher specificity if RFP < 1. Fig. 6 shows the RFPs for the artificial datasets from WT- (KO-) FTO and WT- (KO-) METTL14 datasets; the results of the rest datasets can be found in Supplementary Fig. S3. We can see RFPs are always less than one for all FDR thresholds in all 12 tested datasets. On average, MeTPeak has three times less FPs than exomePeak at any given FDR threshold. For exomePeak, even at FDR threshold of 0.00001, it still reported about more than 10 times more FPs than MeTPeak did. These results suggest that MeTPeak has a consistently higher specificity than exomePeak.

3.3 Prediction results on real MeRIP-seq datasets

After comprehensive evaluation of MeTPeak performance on both simulation and real datasets, we analyzed the MeTPeak predictions on those six MeRIP-seq datasets to investigate context-specific and common characteristics of m^{6}A methylation sites. The results for WT(KO)-FTO and WT(KO)-HeLa(METTL14) are shown in Fig. 7 (see Supplementary Fig. S4 for results for KO-METTL3 and KO-WTAP). On average, 13 172 m^{6}A sites were detected by MeTPeak (FDR < 0.05), whereas the number of detected peaks by exomePeak is 16 691. Although in all datasets, exomePeak reports more peaks

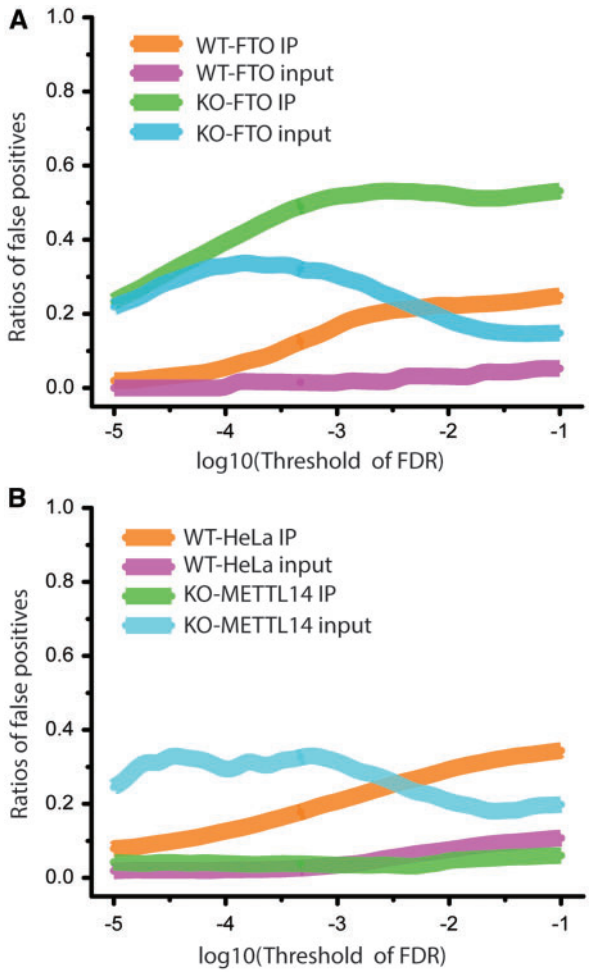

Fig. 6. Ratios of FPs between MeTPeak and exomePeak (MeTPeak over exomePeak) versus log10 FDR threshold. A and B show the results of two different artificial datasets. As can be seen, MeTPeak commits far less FPs than exomePeak

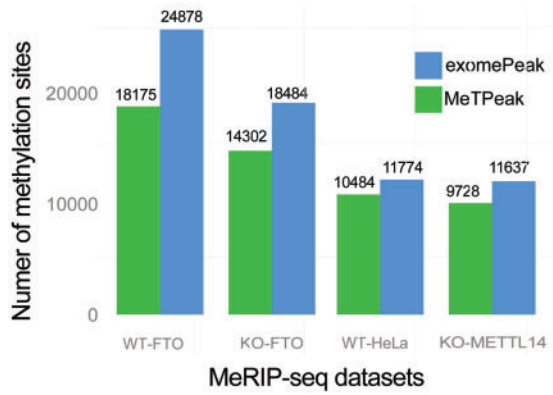

Fig. 7. Prediction results on real MeRIP-seq datasets by MeTPeak

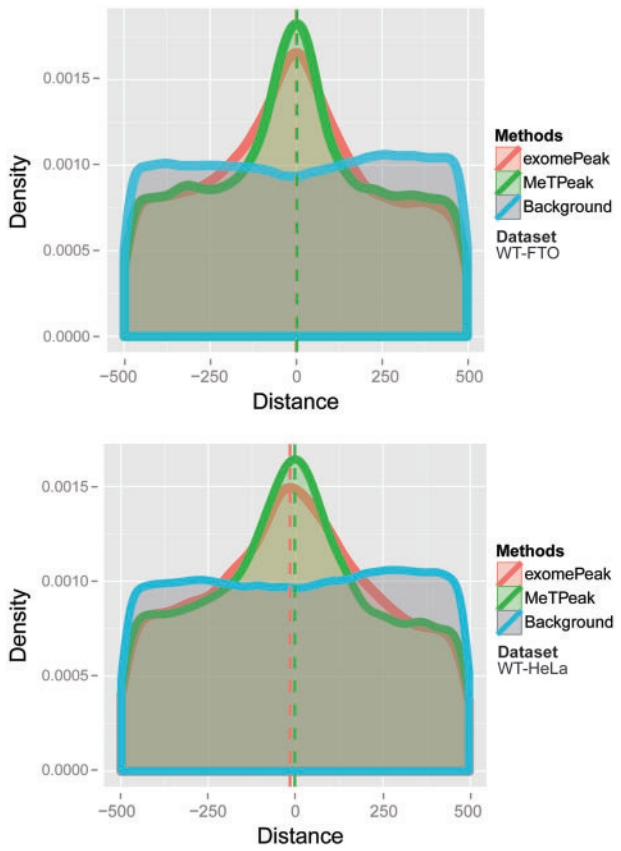

Fig. 8. Comparison of distributions of the distance of motif 'GGAC' to the center of the detected m^{6}A sites between MeTPeak and exomePeak. (A) Distribution of distance of the motif to the center on WT-FTO dataset. (B) Distribution of distance of the motif to the center on WT-HeLa dataset

than MeTPeak, exomePeak commits far more FPs as demonstrated in Section 3.2. As a result, MeTPeak is likely to produce more reliable m^{6}A sites. Besides, we carefully examined the methylation sites in IGV2.1 (Thorvaldsdóttir *et al.*, 2013). We found that some obvious methylation sites were missed by exomePeak (see gene Tgfbr3 in Fig. S5), while a lot more unlikely methylated loci were reported as methylation sites, as the read count was extremely small (see gene Ccdc60 in Fig. S5). It takes MeTPeak about 3∼4 hours to predict m^{6}A sites for the datasets with three replicates on an ordinary desktop (see detailed runtime analysis in Supplementary).

To further validate the methylation sites identified by MeTPeak, we searched the m^{6}A motif in the sequences of detected peaks. It has been shown in (Liu *et al.*, 2014) that 'GGAC' is the most significant consensus sequence. Therefore, we extracted the sequences of the top 5000 most enriched peaks ranked according to FDR for each algorithms using BEDTools (Quinlan and Hall, 2010) and searched 'GGAC' motif in each one of these sequences, where the distance from the location of the identified motif to the center of the peak was computed. The distributions of the distance for WT-FTO and WT-HeLa are plotted in Fig. 8 (See Supplementary S6 for other datasets). We can see that the distribution of MeTPeak is peakier, indicating that 'GGAC' motif is closer to the center in the MeTPeak detected peaks, demonstrating again higher accuracy in the detected methylation site. Next, we investigated the m^{6}A fold enrichment, defined as IP reads count divided by input reads count in the peak region. The distributions of fold enrichment of the identified peaks are plotted in Fig. 9 and Supplementary Fig. S7. The figures show that the distribution of fold enrichment for MeTPeak is significantly higher than that for exomePeak (Welch *t*-test, $P < 2.2e\text{-}16$). The results show that MeTPeak is robust to the reads variance and favors peaks with higher fold enrichment than exomePeak, demonstrating the advantages of modeling variances for the read counts.

We next examined the shared m^{6}A peaks between WT and KO conditions in these four MeRIP-seq datasets to evaluate the

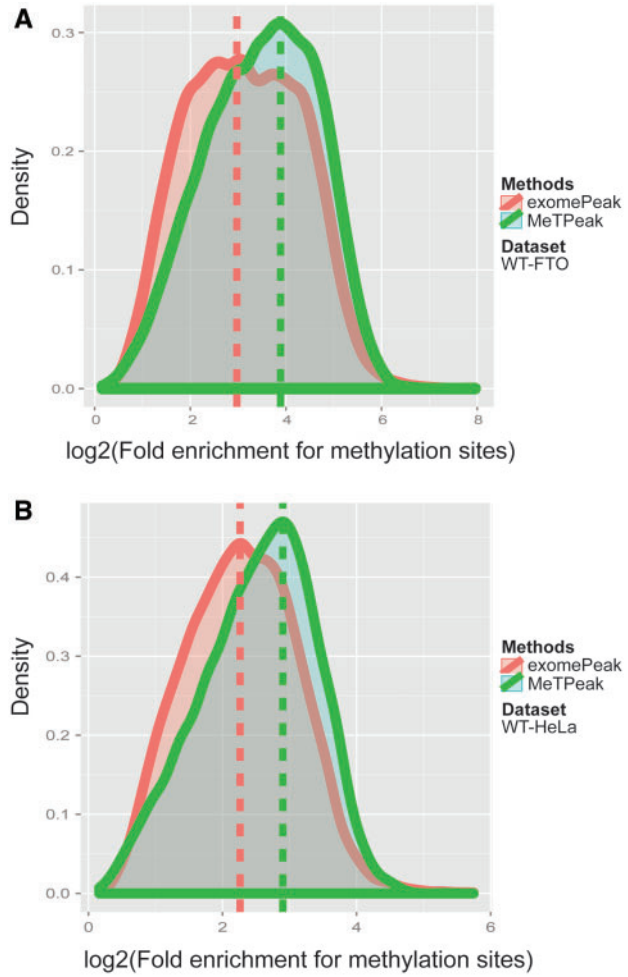

Fig. 9. Comparison of distributions from fold enrichment of identified methylation sites between MeTPeak and exomePeak. (A) Distribution of fold enrichment of methylation sites on WT-FTO dataset. (B) Distribution of fold enrichment of methylation sites on WT-HeLa dataset

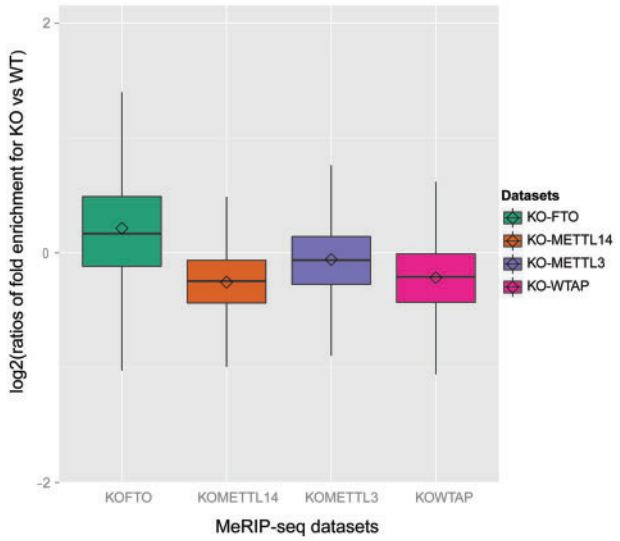

Fig. 10. Boxplot of the differential fold enrichment change of the shared m^{6}A sites between treated condition and wild-type condition

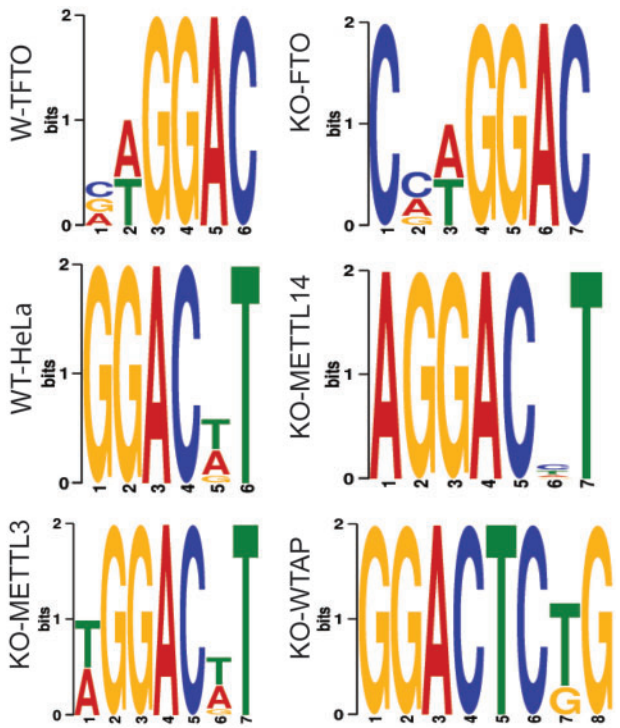

Fig. 11. Motifs detected by DREME on the MeRIP-seq datasets

enrichment change of the common methylation sites. There are 11 534 common m^{6}A peaks reported by MeTPeak in WT-FTO and KO-FTO datasets; for KO-METTL14, KO-METTL3 and KO-WTAP datasets, we found 6170, 8256 and 6067 m^{6}A peaks, respectively. Later, we extracted the fold enrichments for the shared m^{6}A peaks and differential ratios were computed by using fold enrichment in the treated condition divided by that in the WT condition. The boxplots of differential ratios for each group of MeRIP-seq datasets are shown in Fig. 10. In KO-FTO dataset, both the median and mean are above zero, meaning that methylation sites detected in KO-FTO have higher fold enrichment than that in WT-FTO, which is consistent with the fact that knocking out FTO would lead to increased methylation because FTO is an m^{6}A demethylase. In contrast, in the other datasets, KO-METTL14/-METTL3/-WTAP leads to less methylation in the treated conditions, which is again consistent with the fact that METTL4, METTTL3 and WTAP form the m^{6}A methyltransferase complex. Interestingly, KO-METTL14 and KO-WTAP introduce more significant reduction of methylation than KO-METTL3. Indeed, it has been reported that both METTL14 and WTAP have a larger impact on m^{6}A level than METTL3 (Liu *et al.*, 2014).

We also assessed the sequence motif of the detected m^{6}A peaks. For each MeRIP-seq dataset, DREME (Bailey, 2011) was used for each dataset separately and the results are shown in Fig. 11. The enriched motifs from these datasets all contain 'GGAC' motif. Then, we examined the distributions of m^{6}A peaks. As shown in Fig. 12, in all cases, the distributions in mRNAs are consistent with the previously reported general distribution of m^{6}A (Meyer *et al.*, 2012), where a clear enrichment can be seen around the stop codon. In addition, the distributions on lncRNA are also examined in Fig. 12. This consistent pattern of m^{6}A distribution in mRNA and lncRNA across different cell lines and species suggests that the m^{6}A methylation is likely regulated by a similar mechanism. Note that the distributions of KO-M14 and KO-WTAP are somewhat different from

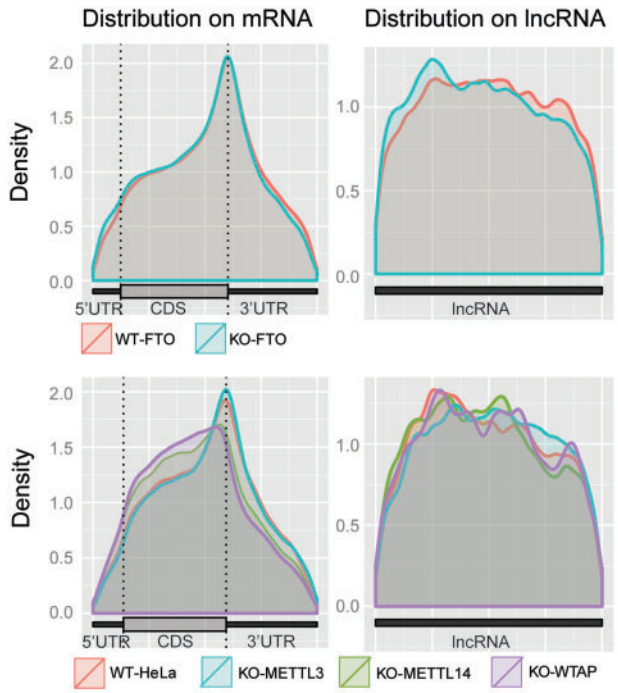

Fig. 12. Distribution of m^{6}A sites in mRNAs and lncRNAsAQ1

the rest. This could suggest that M14 and WTAP may engage in carrying out similar functions.

4 Conclusion

We presented MeTPeak, a novel graphical model-based peak calling method for detecting m^{6}A sites from MeRIP-seq data. We formulated the m^{6}A peak calling problem as an inference of the mixture of binomial distributions by introducing a hierarchical layer of Beta distribution to model the variances among m^{6}A sites in replicates. We also used HMM model to capture the reads dependency among the bins. To find the solution to the inference problem, we proposed a constrained Newton's method for estimating the Beta parameters and developed a fast inference mothed for estimating the m^{6}A peaks. The method is suitable for processing MeRIP-seq data with replicates and high variance, which are common cases in MeRIP-seq data. We have compared MeTPeak with exomePeak both on simulated and real MeRIP-seq datasets and the results showed that MeTPeak is highly robust against the data variances and outliers and making far less FPs than exomePeak. Finally, prediction of MeTPeak on real MeRIP-seq datasets have suggested that it precisely recapitulates the motif and distribution of m^{6}A sites, as well as correctly predicting the methylation differences among these methyltransferases, demonstrating the validity of MeTPeak.

Acknowledgments

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

Funding

National Institutes of Health (NIH-NCIP30CA54174, 5 U54 CA113001 to Y.C. and R01GM113245 to Y.H.); National Science Foundation (CCF-1246073 to Y.H.); The William and Ella Medical Research Foundation grant, Thrive Well Foundation and The Max and Minnie Tomerlin Voelcker Fund to M.K.R.; Natural Science Foundation of China (61473232) to S.Z.

Conflict of Interest: none declared.

References

Alarcon,C.R. *et al.* (2015) N6-methyladenosine marks primary microRNAs for processing. *Nature*, **519**, 482–485.

Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

Dominissini,D. *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.

Fu,Y. *et al.* (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.*, **15**, 293–306.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Kidder,B.L. *et al.* (2011) ChIP-Seq: technical considerations for obtaining high-quality data. *Nat. Immunol.*, **12**, 918–922.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Lindstrom,M.J. and Bates,D.M. (1988) Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Stat. Assoc.*, **83**, 1014–1022.

Liu,J. *et al.* (2014) A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.*, **10**, 93–95.

Meng,J. *et al.* (2014) A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods*, **69**, 274–281.

Meyer,K.D. *et al.* (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, **149**, 1635–1646.

Nilsen,T.W. (2014) Internal mRNA methylation finally finds functions. *Science*, **343**, 1207–1208.

Niu,Y. *et al.* (2013) N 6-Methyl-adenosine (m 6 A) in RNA: an old modification with a novel epigenetic function. *Genomics Proteomics Bioinformatics*, **11**, 8–17.

Przyborowski,J. and Wilenski,H. (1940) Homogeneity of results in testing samples from Poisson series: With an application to testing clover seed for dodder. *Biometrika*, **31**, 313–323.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, **14**, 178–192.

Wang,X. *et al.* (2014a) N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, **505**, 117–120.

Wang,Y. *et al.* (2014b) N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.*, **16**, 191–198.