Quantitative Biology 2018, 6(3): 275–286
https://doi.org/10.1007/s40484-018-0149-2

METHODOLOGY ARTICLE

A Bayesian hierarchical model for analyzing methylated RNA immunoprecipitation sequencing data

Minzhe Zhang[1,†], Qiwei Li[1,†] and Yang Xie[1,2,3,*]

[1] Quantitative Biomedical Research Center, Department of Clinical Sciences, U.T. Southwestern Medical Center, Dallas, TX 75390, USA
[2] Department of Bioinformatics, U.T. Southwestern Medical Center, Dallas, TX 75390, USA
[3] Simmons Comprehensive Cancer Center, U.T. Southwestern Medical Center, Dallas, TX 75390, USA
* Correspondence: yang.xie@utsouthwestern.edu

Received November 23, 2017; Revised March 23, 2018; Accepted April 3, 2018

***Background*: The recently emerged technology of methylated RNA immunoprecipitation sequencing (MeRIP-seq) sheds light on the study of RNA epigenetics. This new bioinformatics question calls for effective and robust peaking calling algorithms to detect mRNA methylation sites from MeRIP-seq data.**
***Methods*: We propose a Bayesian hierarchical model to detect methylation sites from MeRIP-seq data. Our modeling approach includes several important characteristics. First, it models the zero-inflated and over-dispersed counts by deploying a zero-inflated negative binomial model. Second, it incorporates a hidden Markov model (HMM) to account for the spatial dependency of neighboring read enrichment. Third, our Bayesian inference allows the proposed model to borrow strength in parameter estimation, which greatly improves the model stability when dealing with MeRIP-seq data with a small number of replicates. We use Markov chain Monte Carlo (MCMC) algorithms to simultaneously infer the model parameters in a *de novo* fashion. The R Shiny demo is available at https://qiwei.shinyapps.io/BaySeqPeak and the R/C ++ code is available at https://github.com/liqiwei2000/BaySeqPeak.**
***Results*: In simulation studies, the proposed method outperformed the competing methods exomePeak and MeTPeak, especially when an excess of zeros were present in the data. In real MeRIP-seq data analysis, the proposed method identified methylation sites that were more consistent with biological knowledge, and had better spatial resolution compared to the other methods.**
***Conclusions*: In this study, we develop a Bayesian hierarchical model to identify methylation peaks in MeRIP-seq data. The proposed method has a competitive edge over existing methods in terms of accuracy, robustness and spatial resolution.**

Keywords: MeRIP-seq data; RNA epigenomics; Bayesian inference; hidden Markov model; zero-inflated negative binomial

Author summary: Methylated RNA immunoprecipatation combined with RNA sequencing (MeRIP-seq), which can be viewed as a marriage of two well-studied techniques: ChIP-seq and RNA-seq, is changing the landscape of RNA epigenomics study at a higher resolution. We propose a Bayesian statistical model to identify the transcriptome methylation sites using MeRIP-seq data. Our approach includes several innovative characteristics by taking into account: (i) the high proportion of zeros in the data due to the insufficient sequencing depth; (ii) the spatial dependence of neighboring read enrichment. Compared to the existing methods, it is shown that our prediction is more consistent with the biological knowledge, and has better accuracy and spatial resolution.

[†] These authors contributed equally to this work.

INTRODUCTION

Epigenetic modifications of DNAs and histones have demonstrated substantial effects on many biological functions, such as cellular differentiation and development [1,2]. Diverse epigenetic modifications of RNAs are also involved in essential regulatory functions in various biological processes [3]. More than 100 chemical modifications of RNA have been found in eukaryotes and some viruses, among which N^6-methyladenosine (m^6A) is the most prevalent in mRNAs and long non-coding RNAs (lncRNAs) [4–6]. Although m^6A modification was first discovered in the 1970s [7–9], not until recently was it found that m^6A modification is a dynamic and reversible process in which adenosine methyltransferases ("writers"), demethylases ("erasers") and m^6A binding proteins ("readers") play distinct roles in methylation, demethylation and recognition of m^6A and conferring downstream effect [10–13]. It has been shown that RNA m^6A methylation contributes to the regulation of a wide range of fundamental biological processes, such as mRNA maturation and degradation, and RNA-protein interaction [14,15]. However, the mechanism that determines methylation sites and the association between m^6A modifications and functional consequences are largely unknown.

In 2012, two independent studies [4,5] developed a new approach called methylated RNA immunoprecipitation sequencing (MeRIP-seq) to map the transcriptome-wide landscape of mRNA m^6A methylome. This technology first fragments total mRNA into short sequences of length approximately 100 nt. Next, the fragmented mRNAs are separated into two parts. One part is used as the control (input) sample, and the other part is subject to immunoprecipitation (IP) by the anti-m^6A antibody to isolate the methylated sequences. Both the control and IP samples are then submitted for high-throughput sequencing. The IP sample contains enriched methylated RNA fragments because unmethylated fragments are washed off in the immunoprecipitation step, while the control sample includes all RNA fragments and thus can be used to adjust methylated transcript abundance with basal gene expression level. Therefore, MeRIP-seq-based m^6A detection can be considered as a peak calling problem, since the read counts of those m^6A sites in the IP samples are enriched compared to those in the control samples. Although the MeRIP-seq experiment measures the transcriptome-wide landscape of mRNA m^6A methylome, it is usually with few replicates. In addition, similar to RNA-seq, MeRIP-seq data usually involve non-negative counts with drastic variation, resulting from transcriptional expression of different genes and isoforms. Finally, strong spatial dependency exists among the read counts of the neighboring sites along mRNA transcripts. All of these pose challenges to MeRIP-seq data analysis, and need to be carefully considered in order to accurately identify the m^6A methylation sites.

Currently, several algorithms have been developed to identify the m^6A sites from MeRIP-seq data. For example, exomePeak [16] models the read counts using a Poisson distribution and conducts a C-test [17] based on the hypothesis that the mean of IP reads at methylated sites is greater than those in the input samples. HEPeak [18] also assumes a Poisson distribution for read count, while introducing a hidden Markov model (HMM) to model the spatial dependency among the reads from neighboring sites. Recently, MeTPeak [19] fits a beta-binomial model to account for highly fluctuating read enrichments across MeRIP-seq replicates. Although each method has its own advantages, the high proportion of zero counts observed in the MeRIP-seq data significantly weakens the stability and performance of the existing methods. In this study, we propose a novel Bayesian method to predict m^6A sites for the MeRIP-seq data. Our model employs a zero-inflated negative binomial model to capture the zero-inflation and over-dispersion observed in sequencing data. It incorporates the spatial dependency of neighboring read enrichments using an HMM model. Furthermore, the Bayesian inference allows our model to borrow strength in parameter estimation, which greatly improves the model stability when dealing with MeRIP-seq data with a small number of replicates. Markov chain Monte Carlo (MCMC) sampling techniques are used to sample the posterior distributions of the model parameters.

The rest of this paper is arranged as follows. In the Section of Hierarchical Model in Methods, we introduce the Bayesian hierarchical modeling framework. In the Section of Details of the MCMC Algorithm in Methods, we present the MCMC algorithm and discuss the resulting posterior inference. In the Section of Simulation in Results, we assess performance of the proposed model on simulated data and carry out comparisons with exomePeak and MeTPeak (HEPeak is no longer accessible). In the Section of Real MeRIP-seq data in Results , we investigate the results of data analysis from a case study.

METHODS

Hierarchical model

Given a MeRIP-seq dataset observed on a set of n samples, we first divide the concatenated exome of mRNA (*i.e.*, the RefSeq gene) of interest into W mutually adjacent bins. Let $\boldsymbol{Y}$ be an n-by-W matrix of read counts defined on the exons of this particular mRNA, with entry $y_{i,w} \in \mathbb{N}$, $i = 1, ..., n$, $w = 1, ..., W$ indicating the number of reads in bin w for sample i. The MeRIP-seq outputs pairs of IP and control samples. We use a binary vector $\boldsymbol{c} = (c_1, ..., c_n)^T$ to

allocate the two different groups of samples, with $c_i = 1$ if sample i is subject to IP, and $c_i = 0$ otherwise.

Modelling counts by a zero-inflated negative binomial model

As most sequencing data are zero-inflated and over-dispersed, we start by modeling the read counts by a zero-inflated negative binomial (ZINB) model. Specifically, for the number of reads in bin w for sample i, we write a mixture model, where one of the kernels is constrained to be degenerate at zero, thereby allowing for zero-inflation:

$$p(y_{i,w}|\pi,\lambda_{i,w},\phi w) = \pi I(y_{i,w}=0) + (1-\pi)\mathrm{NB}(y_{i,w};\lambda_{i,w},\phi w). \quad (1)$$

In model (1), the weight of extra zero counts, *i.e.*, π, is a positive real number less than 1, and $\mathrm{NB}(y;\lambda,\phi)$ denotes a negative binomial distribution for the random variable y, with expectation λ and dispersion $1/\phi$. With this parameterization, the variance of the negative binomial distribution can be written as $\lambda + \lambda^2/\phi$, thereby allowing for over-dispersion. Note that increasing ϕ towards infinity leads the negative binomial to a Poisson distribution with both mean and variance equal to λ. Alternatively, we can write model (1) as $p(y_{i,w}|\eta_{i,w}, \lambda_{i,w}, \phi_w) = I(y_{i,w}=0)^{\eta_{i,w}} \mathrm{NB}(y_{i,w};\lambda_{i,w},\phi_w)^{1-\eta_{i,w}}$, by introducing a binary latent variable $\eta_{i,w}|\pi \sim \mathrm{Bern}(\pi)$, such that if $\eta_{i,w}=1$, then $y_{i,w}=0$, whereas if $\eta_{i,w}=0$, then $y_{i,w} \sim \mathrm{NB}(\lambda_{i,w},\phi_w)$. The Bernoulli prior assumption can be further relaxed by formulating a $\mathrm{Be}(a_\pi, b_\pi)$ hyperprior on π, which leads to a beta-Bernoulli prior for $\eta_{i,w}$ with expectation $a_\pi/(a_\pi + b_\pi)$. A vague prior can be elicited by setting $a_\pi = b_\pi = 1$, which leads to a uniform distribution on π. Lastly, we specify the prior distribution for ϕ_w as $\phi_w \sim \mathrm{Ga}(a_\phi, b_\phi)$. One standard way of setting a weakly informative gamma prior is to choose small values for the two hyperparameters, such as $a_\phi = b_\phi = 0.001$ [20].

Modelling negative binomial mean by a random effect model

Sequencing data is also characterized by high variability in the number of reads between different groups, and even across the samples within the same group. To accommodate this setting, we parameterize the mean parameter of the negative binomial distribution as the multiplicative effect of three random effects for the control and IP samples, respectively,

$$\lambda_{i,w} = \begin{cases} s_i g_w d_{0,w} & \text{if} \quad c_i = 0 \\ s_i g_w d_{1,w} & \text{if} \quad c_i = 1 \end{cases}. \quad (2)$$

We interpret s_i as the size factor for sample i, reflecting the fact that different samples may have been sequenced to different depths, and g_w as the scaling factor for bin w, capturing feature-specific (*i.e.*, bin-specific) levels across all samples. Once the global effects s_i and g_w have been taken into account, the parameters $d_{0,w}$ and $d_{1,w}$ are defined as the relative occurrence rate for the counts in bin w in the control and IP samples, respectively. Note that the Poisson and negative binomial versions of model (1) embedded with (2) have been explicitly used to identify differentially expressed genes between different biological conditions for RNA-seq data [21–27].

The parameterization of the negative binomial mean $\lambda_{i,w}$, as shown in (2), results in an identifiability problem among the three random effects. To avoid this issue, we estimate the size factors s_i and the scaling factor g_w by means of plug-in estimators based on the observed data. For example, in the context of RNA-seq data analyses, a number of methods [21,26,28] fix $\hat{s}_i^{\mathrm{total}} = \sum_{j=1}^{p} x_{i,j} / \sum_{i=1}^{n}\sum_{j=1}^{p} x_{i,j}$, so that $\sum_{i=1}^{n} s_i = 1$, where $x_{i,j}$ is the number of reads mapping to gene j in observation i. Similarly, Anders *et al.* [23] propose $\hat{s}_i^{\mathrm{median}} = m_i / \sum_{i=1}^{n}\sum_{j=1}^{p} x_{i,j}$ where $m_i = \mathrm{median}\left(x_{i,1} / \left(\prod_{i'=1}^{n} x_{i'j} I(x_{i'j} \neq 0) \right)^{1/n}, \ldots, x_{i,p} / \left(\prod_{i'=1}^{n} x_{i'j} I(x_{i'j} \neq 0) \right)^{1/n} \right)$ is the median of the distribution of the ratios of the counts for observation i to their geometric mean. As a further example, Bullard *et al.* [22] propose taking $\hat{s}_i^{\mathrm{quantile}} = q_i / \sum_{i=1}^{n} q_i$, with q_i the 75th percentile of the counts $(x_{i,1}, \ldots, x_{i,p})$ for observation i. Many of the examples further fix $g_i = \sum_{i=1}^{n} x_{i,j}$ [26]. Our estimates on s_i and g_w are similar in spirit. Note that s_i should be computed based on the counts of the whole MeRIP-seq dataset, not just from one single RefSeq gene. For g_w, we estimate $\hat{g}_w = \sum_{i=1}^{n} y_{i,w}$.

The use of the plug-in estimates of these two parameters is convenient, but it has noticeable drawbacks, especially for conducting full Bayesian inference. The plug-in estimators can be regarded as maximum likelihood estimators in multiple stage approaches and somewhat akin to empirical Bayes methods, thus relying on implicit assumptions of exchangeability of the observations, which may not be always justified in practice and can introduce bias in the estimation of posterior uncertainties [29,30]. To address the identifiability issue between s_i and g_w, Li *et al.* [31] developed a novel non-parametric Bayes prior model

with soft constraints on their expected values to normalize the size and scaling factors simultaneously. Integrating this prior model with the proposed model in this paper could be one of our future directions.

Modelling relative occurrence rate by a hidden Markov model

IP samples and control samples can have vastly different total numbers of sequence reads. In general, when a bin is methylated, there are more reads in IP samples than in control samples (*i.e.*, $d_{1,w} > d_{0,w}$); otherwise, there are more reads in control samples than in IP samples (*i.e.*, $d_{0,w} > d_{1,w}$). To model this characteristic, we redefine the parameter space $(d_{0,w}, d_{1,w}, w = 1, ..., W)$ to $(d_{0,w}, \delta_w, w = 1, ..., W)$, where $\delta_w = d_{1,w}/d_{0,w}$ is the fold change in the relative occurrence rate in bin w between the IP and control samples. Thus, given the sample allocation vector c, we rewrite the likelihood for each count as,

$$p(y_{i,w}|\eta_{i,w}, \phi_w, d_{0,w}, \delta_w) = \begin{cases} I(y_{i,w}=0)^{\eta_{i,w}} \, \mathrm{NB}\left(y_{i,w};\, \hat{s}_i \hat{g}_w(d_{0,w}), \phi_w\right)^{1-\eta_{i,w}} & \text{if} \quad c_i = 0 \\ I(y_{i,w}=0)^{\eta_{i,w}} \, \mathrm{NB}\left(y_{i,w};\, \hat{s}_i \hat{g}_w(d_{0,w}\delta_w), \phi_w\right)^{1-\eta_{i,w}} & \text{if} \quad c_i = 1 \end{cases}. \tag{3}$$

For each of the non-negative parameters $d_{0,w}, w = 1, ..., W$, we assume a gamma hyperprior, *i.e.*, $d_{0,w} \sim \mathrm{Ga}(a_d, b_d)$. The choice of $a_d = b_d = 0.001$ would lead to a non-informative distribution as its variance is 1000.

We use an HMM to model $\delta_w, w = 1, ..., W$, which accounts for the variation in the underlying occurrence rate between the methylated and unmethylated bins in the IP samples. The HMM model encourages adjacent bins to be affected by the same process (*i.e.*, methylation/unmethylation). To do this, we first introduce a latent variable z_w to denote the hidden methylation status for bin w, with $z_w = 2$ indicating bin w is methylated and $z_w = 1$ otherwise. Conditional on the hidden state z_w, we assume that δ_w is independent and identically normally distributed, defining the emission distribution of the HMM as $\delta_w | z_w = q \sim N(\mu_q, \sigma_q^2), q \in \{1,2\}$. The parameters μ_q and σ_q^2 represent the mean and variance of the fold changes in the relative occurrence rates for methylated/unmethylated bins. The spatial dependence between the states at adjacent bins is captured by a first order Markov model, which assumes that the probability of being a particular state at bin $w+1$ depends only on the state at its antecedent bin w, *i.e.*, $p(z_w|z_1, ..., z_W) = p(z_w|z_{w-1}) = \psi_{z_w, z_{w-1}}$, with $\psi = [\psi_{q',q}]_{2\times 2}$ forming the matrix of transition probabilities with strictly positive elements. Note that in the proposed model, δ_w only switches according to the status of z_w in the IP samples.

As for the prior specification of the HMM, we assume independent Dirichlet priors across the columns of the transition matrix $\boldsymbol{\psi}$; that is, $\psi_{\cdot,q} \sim Dir(a_1, a_2), q = 1,2$. When both α_1 and α_2 are set to 1, the Dirichlet distribution is uniform over all points in its support. For the priors on $\mu_q, q = 1,2$, we follow Guha *et al.* [32] and assume truncated normal distributions $\mu_q \sim N(m_q, s_q^2) I(\mathrm{low}_{u_q} < u_q, \mathrm{upp}_{u_q})$, $q = 1,2$, where $\mathrm{low}_{\eta_1} = 0, \mathrm{upp}_{\eta_2} = \infty$, and $\mathrm{upp}_{\eta_1} = \mathrm{low}_{\eta_2} = 1$. This setting fully accounts for the fact that the read counts of methylated bins in the IP samples are more enriched than those in the control samples, and *vice versa* for the unmethylated bins. For $\sigma_q^2, q = 1,2$ in the emission distribution, we impose inverse-gamma hyperpriors $\sigma_q^2 \sim IG(a_\sigma, b_\sigma), q = 1,2$. We suggest to set $a_\sigma = 2$ and $b_\sigma = 1$ so as to obtain a flat inverse-gamma distribution.

Model fitting

Our model space consists of $(\pi, H, \phi, d_0, \delta, z, \mu, \sigma^2, \psi)$. The parameters of interest are $z = (z_1, ..., z_w)$, which identify the methylated sites. We first design a Markov chain Monte Carlo (MCMC) algorithm based on Gibbs samplers and Metropolis-Hastings algorithms. Then, the resulting posterior inference for z is discussed.

Details of the MCMC algorithm

We start by writing the likelihood of the model (3) for each read count $y_{i,w}$ given c_i, as,

$$p(y_{i,w}|c_i, \eta_{i,w}, \phi_w, d_{0,w}, \delta_w) = I(y_{i,w}=0)^{\eta_{i,w}} \cdot \mathrm{NB}\left(y_{i,w};\, \hat{s}_i \hat{g}_w(d_{0,w}\delta_w^{c_i}), \phi_w\right)^{1-\eta_{i,w}}.$$

Then, within each iteration, the MCMC updates can be summarized as:

- Update of π: We use a Gibbs sampling step to update π,

$$\pi|H \sim \mathrm{Be}\left(a_\pi + \sum_{i=1}^{n}\sum_{w=1}^{W}\eta_{i,w}, b_\pi + nW - \sum_{i=1}^{n}\sum_{w=1}^{W}\eta_{i,w}\right)$$

- Update of $\boldsymbol{H}$: We update each $\eta_{i,w}, i = 1, ..., n, w = 1, ..., W$ that corresponds to $y_{i,w} = 0$ by sampling from a probability mass distribution, which is proportional to

$$p(\eta_{i,w}|y_{i,w}=0, \pi, \phi_w, d_{0,w}, \delta_w) \propto p(y_{i,w}=0|c_i, \eta_{i,w}, \phi_w, d_{0,w}, \delta_w)\mathrm{Bern}(\eta_{i,w}; \pi)$$

- Update of ϕ: We update each $\phi_w, w = 1, ..., W$, sequentially by using a random walk Metropolis-Hastings algorithm. We first propose a new ϕ_w^* from $\mathrm{Ga}(\phi_w^2/\tau_\phi,$

$\phi_w/\tau_\phi)$ and then accept the proposed value ϕ_w^* with probability min(1, m_{MH}), where

$$m_{MH}=\frac{\prod_{i=1}^{n} p(y_{i,w}|c_i,\eta_{i,w},\ \phi_w^*, d_{0,w},\delta_w)\mathrm{Ga}(\phi_w^*;\ a_\phi,b_\phi)J(\phi_w;\ \phi_w^*)}{\prod_{i=1}^{n} p(y_{i,w}|c_i,\eta_{i,w},\phi_w,d_{0,w},\delta_w)\mathrm{Ga}(\phi_w;a_\phi,b_\phi)J(\phi_w^*;\phi_w)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update.

• Update of d_0: We update each $d_{0,w}, w=1,...,W$ sequentially by using a random walk Metropolis-Hastings algorithm. We first propose a new $d_{0,w}^*$ from $\mathrm{Ga}(d_{0,w}^2/\tau_\phi, d_{0,w}/\tau_d)$ and then accept the proposed value $d_{0,w}^*$ with probability min(1, m_{MH}), where

$$m_{MH}=\frac{\prod_{i=1}^{n} p(y_{i,w}|c_i,\eta_{i,w},\phi_w,d_{0,w}^*,\delta_w)\mathrm{Ga}(d_w^*;a_\phi,b_\phi)J(d_w;d_w^*)}{\prod_{i=1}^{n} p(y_{i,w}|c_i,\eta_{i,w},\phi_w,d_{0,w},\delta_w)\mathrm{Ga}(d_w;a_\phi,b_\phi)J(d_w^*;d_w)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update.

• Update of δ: We update each $\delta_w, w=1,...,W$ sequentially by using a random walk Metropolis-Hastings algorithm. We first propose a new δ_w^* and from $\mathrm{N}(\delta_w,\tau_\delta^2)$ $I(\delta_w^*>0)$ and then accept the proposed value δ_w^* with probability min(1, m_{MH}), where

$$m_{MH}=\frac{\prod\{i:c_i=1\}p(y_{i,w}|c_i,\eta_{i,w},\phi_w,d_{0,w},\delta_w^*)\mathrm{N}(\delta_w^*;\mu_{z_w},\delta_{z_w}^2)J(\delta_w;\delta_w^*)}{\prod\{i:c_i=1\}p(y_{i,w}|c_i,\eta_{i,w},\phi_w,d_{0,w},\delta_w)\mathrm{N}(\delta_w;\mu_{z_w},\delta_{z_w}^2)J(\delta_w^*;\delta_w)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update.

• Update of z: We update each $z_w, w=1,...,W$ sequentially by a stochastic version of the forward-backward algorithm. We first generate z_W from $p(z_W=q|\delta_W,\mu_q,\sigma_q^2,\psi)\propto\mathrm{N}(\delta_W,\mu_q,\sigma_q^2)$ at the beginning of the backward step. Then, the backward step is continued to compute and generate a draw for $z_w, w=W-1,...,1$ from

$$p(z_w=q|\delta_w,\mu_q,\sigma_q^2,\psi)\propto\mathrm{N}(\delta_w;\mu_q,\sigma_q^2)\psi_{z_{w+1},q}.$$

This produces a sample from the joint distribution $p(z|\delta,\mu,\sigma^2,\psi)$.

• Update of μ: We update each $\mu_q, q=1,2$ separately via a Gibbs sampling step,

$$\mu_q|\delta,z,\sigma_q^2\sim\mathrm{N}\left(\frac{m_q/s_q^2+\sum_{w=1}^{W}\delta_w I(z_w=q)/\sigma_q^2}{1/s_q^2+\sum_{w=1}^{W}I(z_w=q)/\sigma_q^2},\ \frac{1}{1/s_q^2+\sum_{w=1}^{W}I(z_w=q)/\sigma_q^2}\right)I(\mathrm{low}_{\mu_q}<\mu_q<\mathrm{upp}_{\mu_q}).$$

• Update of σ^2: We update each $\sigma_q^2, q=1,2$ separately via a Gibbs sampling step,

$$\sigma_q^2|\delta,z,\mu_q\sim\mathrm{IG}\Big(a_\sigma+\sum_{w=1}^{W}I(z_w=q)/2,\ b_\sigma+\sum_{w=1}^{W}(\delta_w-\mu_q)^2I(z_w=q)/2\Big).$$

• Update of ψ: We use a Gibbs sampling step to update each column of the matrix $\psi_{\cdot,q}, q=1,2$,

$$\psi_{\cdot,q}\sim\mathrm{Dir}\Big(a_1+\sum_{w=2}^{W}I(z_{w-1}=1,z_w=q),\ a_2+\sum_{w=2}^{W}I(z_{w-1}=2,z_w=q)\Big).$$

Posterior estimation on the methylation site indicator

For posterior inference, our primary interest lies in the identification of the methylation sites, via the vector z. One way to summarize the posterior distribution of particular parameters is via the *maximum-a-posteriori* (MAP) estimates obtained as

$$\hat{z}^{MAP}=\mathrm{argmax}_{1\leqslant b\leqslant B}\prod_{w=1}^{W}\psi_{z_w^{(b)},z_{w-1}^{(b)}}\mathrm{N}(\delta_w^{(b)};\mu_{z_w^{(b)}}^{(b)},\sigma_{z_w^{(b)}}^{(b)2}),$$

with $b=1,...,B$ indexing the MCMC iterations, after burn-in. Estimation can also be done by thresholding the estimated marginal posterior probabilities of inclusion (PPI) of single bin, obtained as the proportion of MCMC iterations, after burn-in, in which the corresponding z_w is equal to 2. That is,

$$\mathrm{PPI}_w^z=\frac{\sum_{b=1}^{B}(z_w^{(b)}=2|\cdot)}{B}.$$

In choosing the threshold to obtain $\hat{z}_i^{\mathrm{PPI}}=1+I(\mathrm{PPI}_w^z\geqslant c)$,

the value of $c = 0.5$ results in a median model. Alternatively, we can follow the Ref. [33], which guarantees the expected Bayesian false discovery rate (BFDR) to be smaller than a pre-specified threshold.

RESULTS

Simulation

In this study, simulated data with known ground truth were used to validate the performance of the proposed model. The read counts for the control and IP samples were generated using a strategy similar to the HEPeak model by Cui *et al*. [18]. We considered multiple scenarios, of which counts were from four different kernels: Poisson, zero-inflated Poisson (ZIP), negative binomial (NB), and zero-inflated negative binomial (ZINB). The percentage of extra zeros, *i.e.*, π, was set to 0.5 for the ZIP and ZINB kernels. Note that for the real data analyzed in the paper, about one-third of counts were zeros. Each scenario included 100 datasets, each of which had 4 control and 4 IP samples, *i.e.*, $n = 8$, with the number of bins drawn from a uniform distribution, *i.e.*, $W \sim \text{Unif}(100,500)$. Reads of each bin in the control samples were allowed to vary according to the corresponding kernel, where we chose the mean parameter $\lambda_w^{\text{control}} \sim \text{Unif}(5,20)$ and assumed it remained constant for both methylated and unmethylated bins. The means of reads of the methylated and unmethylated bins in the IP samples were set to $\lambda_w^{\text{IP}}|z_i = 2 = k\lambda_w^{\text{control}}$ and $\lambda_w^{\text{IP}}|z_i = 1 \sim \text{Unif}(0, \lambda_w^{\text{control}})$, respectively. Here, k can be seen as the fold change in relative occurrence rates in methylated bins between the IP and control samples. The larger the value of k, the more enriched the read counts of methylated bins in the IP samples than in the control samples. Two different settings of k were considered, $k = 1.5$ and 3. Combined with the four kernels, there were $4 \times 2 = 8$ scenarios in total. The hidden methylation statuses were generated from HMM models with the transition matrix $\psi = \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix}$ at a random initial state. This matrix was based on the estimations using the real data analyzed in Ref. [18].

We implemented our method with different plug-in size factors $\hat{s}_i^{\text{total}}$'s, $\hat{s}_i^{\text{median}}$'s, and $\hat{s}_i^{\text{quantile}}$'s, and called them BaySeqPeak-T, -M, and -Q, respectively. We used the following "non-informative" settings for the priors and hyperpriors. We set the hyperparameters that control the excess of zeros to $Be(a_\pi = 1, b_\pi = 1)$. As for the gamma prior on the over-dispersion parameters $\phi_w, w = 1, ..., W$ and the reference relative occurrence rate $d_{0,w}, w = 1, ..., W$, we set $a_\phi = b_\phi = a_d = b_d = 0.001$, which led to vague gamma priors. For the HMM model, we considered two cases: (i) each column of the transition matrix ψ was independently distributed according to Dir(1,1) and (ii) each column was identical, *i.e.*, $a_{1q} = a_{2q}, q = 1,2$, which corresponds to a special case of a Markov chain, called a Bernoulli scheme. The second case assumes that there was no spatial dependency between adjacent bins and the resulting model was named BaySeqPeak-I, where I stands for independence. Finally, for those reads in the methylated bins, we set $\mu_2 \sim N(m_2 = 0, s_2^2 = 10)I(\mu_2 > 1)$ and $\sigma_2^2 \sim IG(2,1)$, while for those reads in the unmethylated bins, we set $\mu_1 \sim N(m_1 = 0, s_1^2 = 10)I(0 < \mu_1 < 1)$ and $\sigma_1^2 \sim IG(2,1)$. For each simulated dataset, we ran four independent MCMC chains with 10W iterations, discarding the first 50% sweeps as burn-in. Results we report below were obtained by pooling together the MCMC outputs from the four chains. All experiments were implemented in R with an Rcpp package to accelerate computation on a Mac PC with 2.60 GHz CPU and 16 GB memory. In our implementation, one MCMC chain ran about 1 second for data with dimensions $n = 8$ and $W = 100$.

To quantify the accuracy of the identification of methylation sites (via the parameters z) by the proposed model, we calculated the receiver operating characteristic (ROC) curves, and the areas under the curves (AUCs) as the performance metrics. They consider both the true and false positive rates at various threshold settings. The true positive rate is also known as *sensitivity*, which is defined as the number of true methylated bins that are correctly estimated. The false positive rate is equal to 1 minus *specificity*, which is defined as the true unmethylated bins that are correctly estimated. AUC provides a summary of the overall performance across different threshold values. We compared the performance of our proposed method with exomePeak and MeTPeak. Note that they produce threshold *P*-values to control for the false discovery rate (FDR), while our model-based method controls for FDR by generating marginal posterior probabilities of inclusion.

One example of our simulated dataset, which was generated from ZINB kernel and with fold change $k = 3$, is displayed in Figure 1A. It clearly shows the presence of overdispersion as well as excess zeros. Figure 1B plots the marginal posterior probabilities of inclusion of single bins belonging to methylated regions, inferred by the proposed model with total size factor estimates. The red dots indicate the true methylated bins. A threshold of $c = 0.5$ results in a median model that included 28 methylated bins, 26 of which were correctly estimated. Our inference only missed four true methylated bins. We also give the true and the estimated methylated bins by exomePeak, MeTPeak, and BaySeqPeak with different settings at the top. As we can see, the proposed model outperformed the competing methods in this example. Unlike other approaches that provide the confidence level via *P*-values, the posterior distribution of marginal probabilities facil-

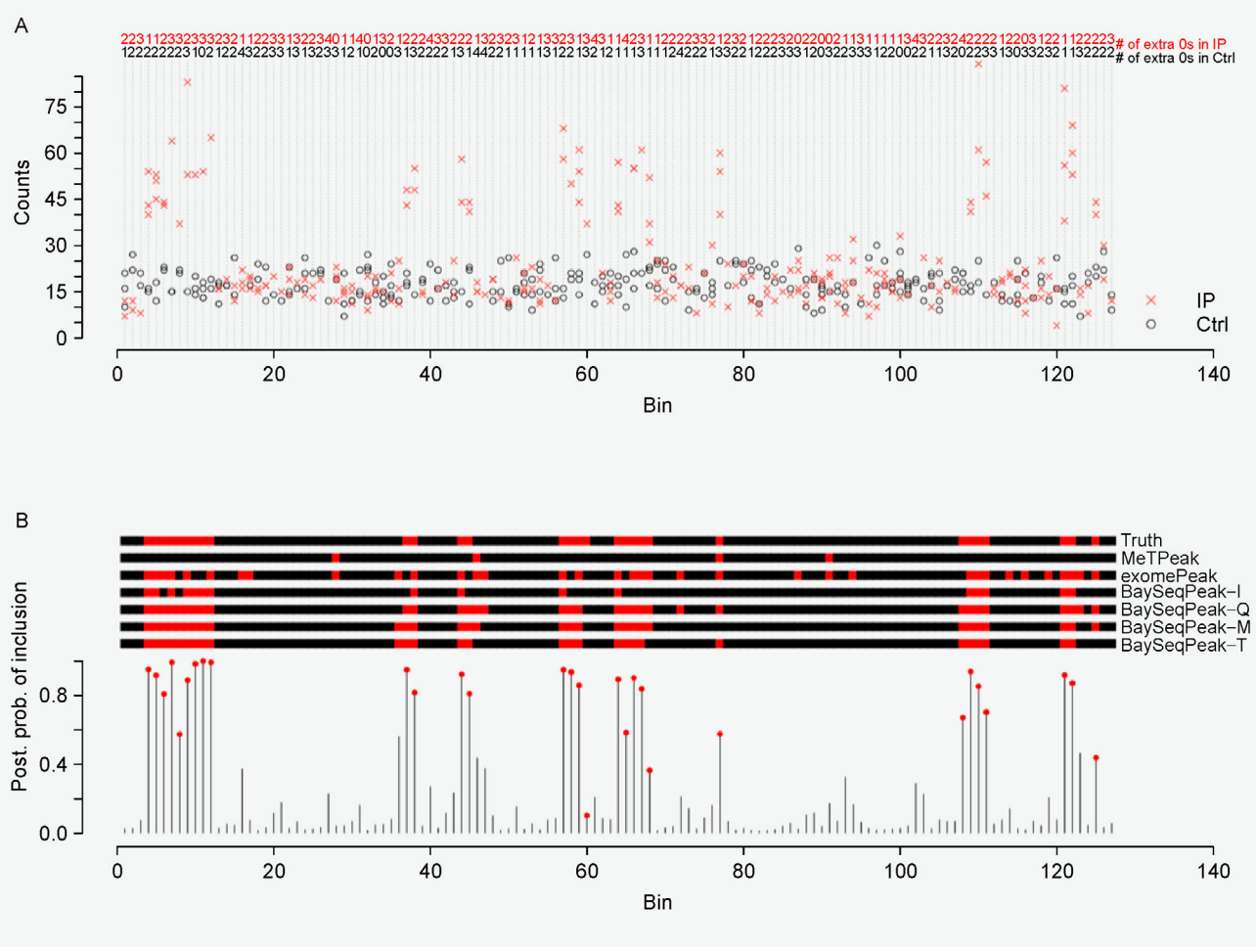

Figure 1. Examples of the model input and output. (A) One simulated data generated from the ZINB kernel and with the fold change $k = 3$. Non-zero counts in the control samples are marked in black circle (o), while non-zero counts in the IP samples are marked in red cross (x). The number of extra zeros for each bin and each sample group is given at the top. (B) The marginal posterior probabilities of inclusion $p(z_w = 2|\cdot)$ inferred by BaySeqPeak-T with the plug-in size factors $\hat{s}_i^{\text{total}}$'s . The red dots indicate the true methylated bins. The true and estimated z by MeTPeak, exomePeak (at a 5% significance level cutoff), and BaySeqPeak (with $c = 0.5$ cutoff) are shown in the top, where the red regions indicate methylated bins.

itates the inference and adds a level of interpretation available only through a Bayesian approach. Figure 2A displays the trace plots of the number of methylated bins $\sum_{w=1}^{W} I(z_w = 2)$ of four independent MCMC chains, which clearly shows that each chain converged and stabilized around the true value 30 in a very short run. We also plotted the trace plots of the transition probabilities A for the first chain in Figure 2B. Furthermore, we used Gelman and Rubin's convergence diagnostics [34] to inspect for signs of convergence of the individual parameters such as A. The statistics were ranging from 1.034 and 1.127, suggesting that the MCMC chains were run for a sufficient number of iterations. We plotted the ROC curves for our method with three different size factor estimates, for different values of the threshold on PPIs, and compared those to the ROC curves obtained with exomePeak and MeTPeak, for different values of the threshold on *P*-values. The ROC curves, averaged over 100 datasets, are shown in Figure 3. The boxplots of AUCs for the 100 simulated datasets for different methods are summarized in Figure 4. Our observations are four-fold. First, our method that makes use of the dependency of consecutive bins achieved the best accuracy under all scenarios. Second, when the signal strength was strong enough (*i.e.*, less variability and zeros), the statistical performances showed no significant differences from each other. However, the increase of variance and number of zeros led to greater disparity between the competing method and ours. Third,

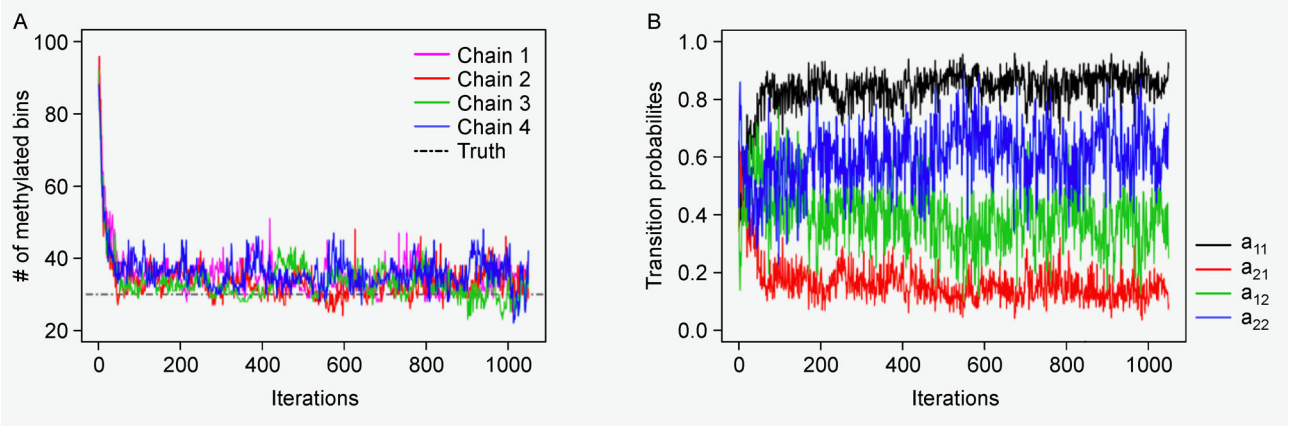

Figure 2. Examples of the MCMC outputs. (A) The trace plots of the number of methylated bins $\sum_{w=1}^{W} I(z_i = 2)$ of four independent chains obtained by BaySeqPeak with the plug-in size factors $\hat{s}_i^{\text{total}}$'s. (B) The trace plots of the transition probabilities a_{11}, a_{21}, a_{12}, and a_{22} of Chain 1.

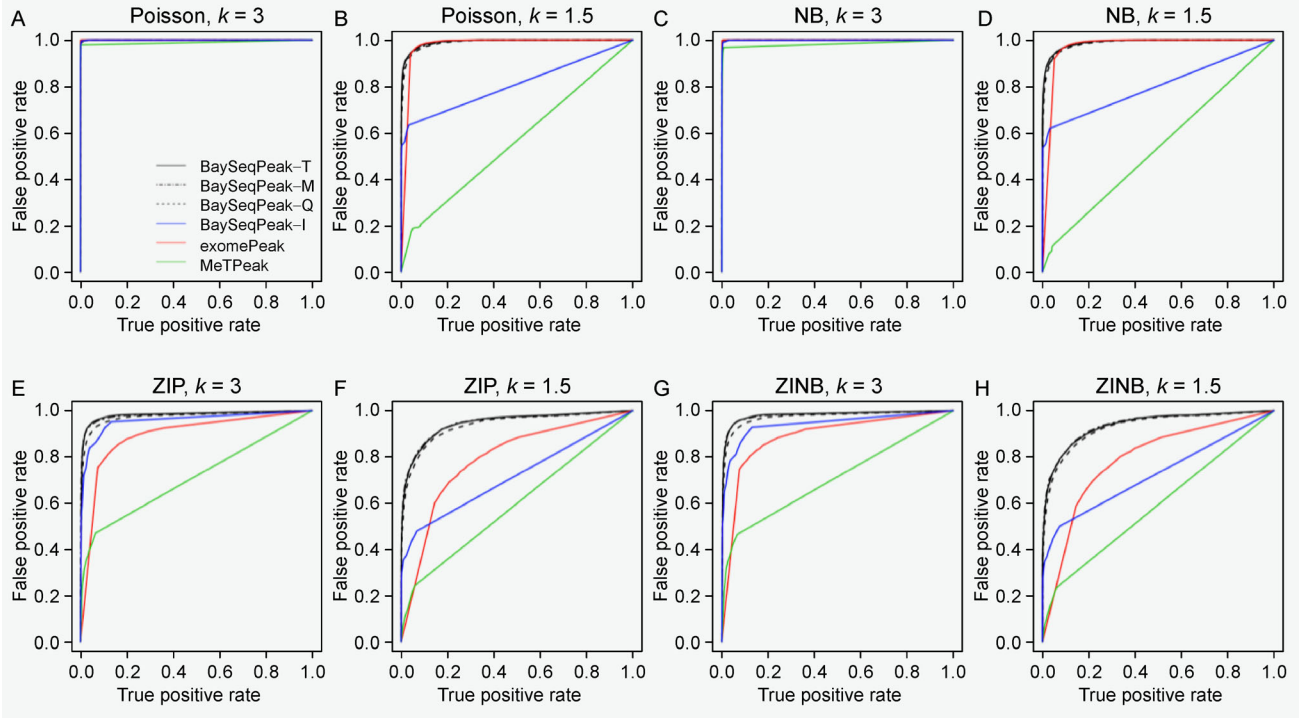

Figure 3. ROC curves produced by different methods. (A–H) The average receiver operating characteristic (ROC) curves for different values of the threshold on PPIs (our method) and on *P*-values (competing methods) for different scenarios. BaySeqPeak-T, -M, -Q stand for our method with different plug-in size factors $\hat{s}_i^{\text{total}}$'s, $\hat{s}_i^{\text{median}}$'s, and $\hat{s}_i^{\text{quantile}}$'s, respectively. BaySeqPeak-I stands for our method with the plug-in size factors $\hat{s}_i^{\text{total}}$'s, and the Bernoulli scheme, a special HMM where no spatial dependency is assumed between adjacent bins.

the proposed model was considerably insensitive to the choice of the three size factor estimates. Fourth, if the hidden methylated states followed a Bernoulli scheme, which ignored the dependency of reads, the model performed poorly, especially when the signal strength was weak, *e.g.*, those scenarios with $k = 1.5$.

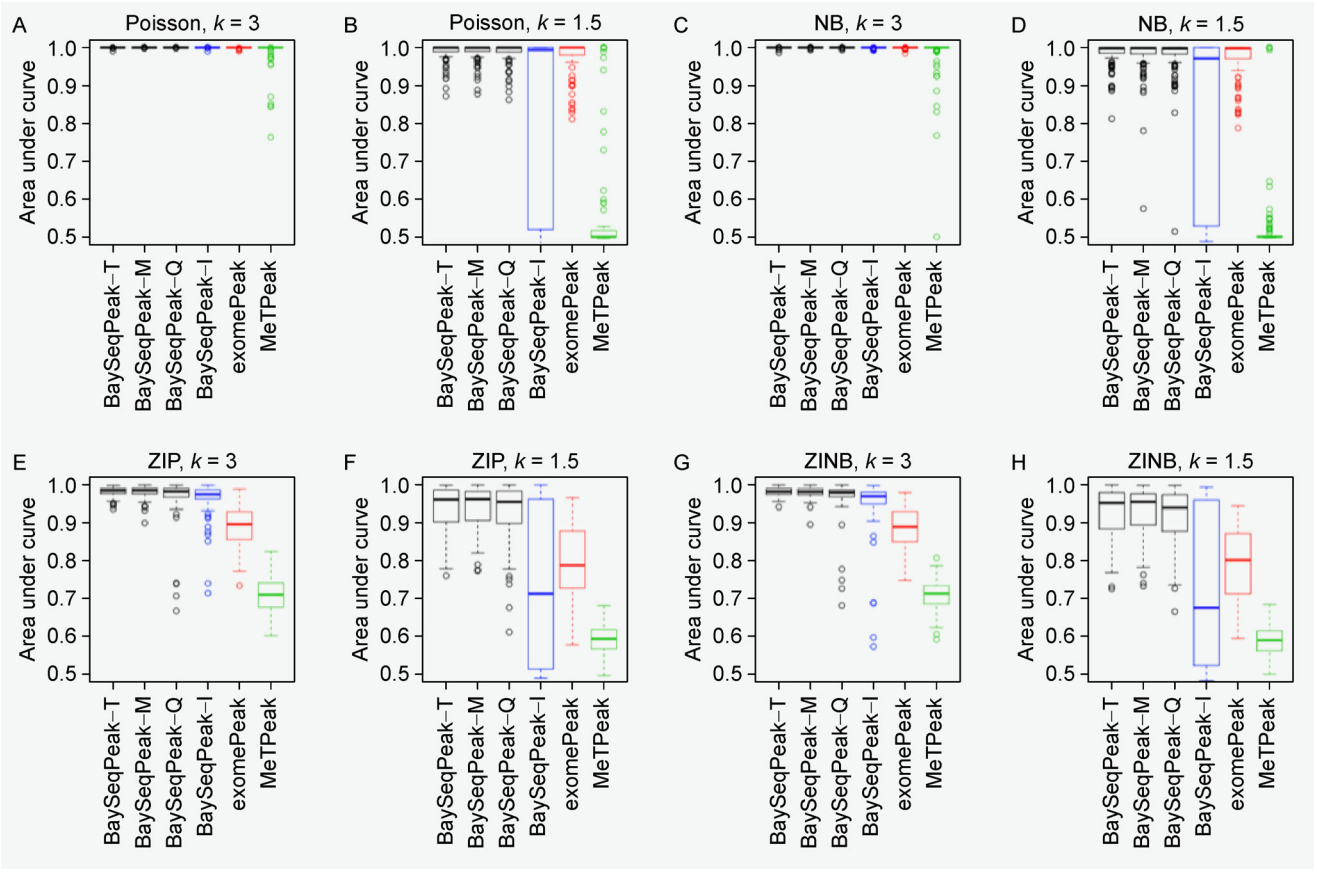

Figure 4. AUCs produced by different methods. (A–H) The boxplot of areas under the curves (AUCs) over 100 datasets for different scenarios, corresponding to Figure 3 (A–H).

Real MeRIP-seq data

To further evaluate the performance of our method in a real biological scenario, we applied both the proposed algorithm and exomePeak to a MeRIP-seq dataset from a study investigating the role of demethylase obesity-associated protein (FTO) in mouse midbrain [35]. MeRIP-seq data was generated in both wild type (WT) and FTO knock-out (FTO-KO) mice, with each condition having 3 IP replicates and 3 input replicates, respectively. The data was downloaded from the Gene Expression Omnibus (GEO) repository [36] (Accession number: GSE47217) and processed following the protocol by Jia *et al*. [37]. The mouse transcriptome (UCSC mm10) was scanned using 200 bp window with sliding step 30 bp (default setting of exomePeak), and reads of each sample in each window were counted. For exomePeak, a C test was performed in each window independently to test the significance of enrichment of reads in IP samples versus input samples. Windows with FDR less than 0.05 were reported as methylated bins. For our methods, a Bayesian hierarchical model either with or without HMM integration was fitted for each transcript, respectively, in order to estimate the posterior probability of inclusion (ppi) of each window within each region. Windows with ppi greater than 0.5 were reported as methylated bins. Finally, the adjacent methylated bins were joined to form a methylation site in the transcriptome.

The mouse mm10 reference transcriptome was divided into a total of 2916,847 bins (gene with all 0 counts were filtered out). exomePeak reported 853,151 bins as significant in WT condition datasets, while our BaySeq-Peak-T and -I reported 664,513 and 512,754 bins (Figure 5A). The overall agreement between the results of our methods and exomePeak indicates that all methods performed reasonably well. However, BaySeqPeak-T reported a smaller number of significant bins, and BaySeqPeak-I reported even less (also observed in the FTO-KO condition). This may due to the failure to detect methylated bins with weak signals that are adjacent to significant regions when HMM is not applied. This trend is also represented at the transcript level. BaySeqPeak-T and -I identified 55,199 and 44,210 methylation sites in WT condition, and 65,110 and 55,829 methylation sites in

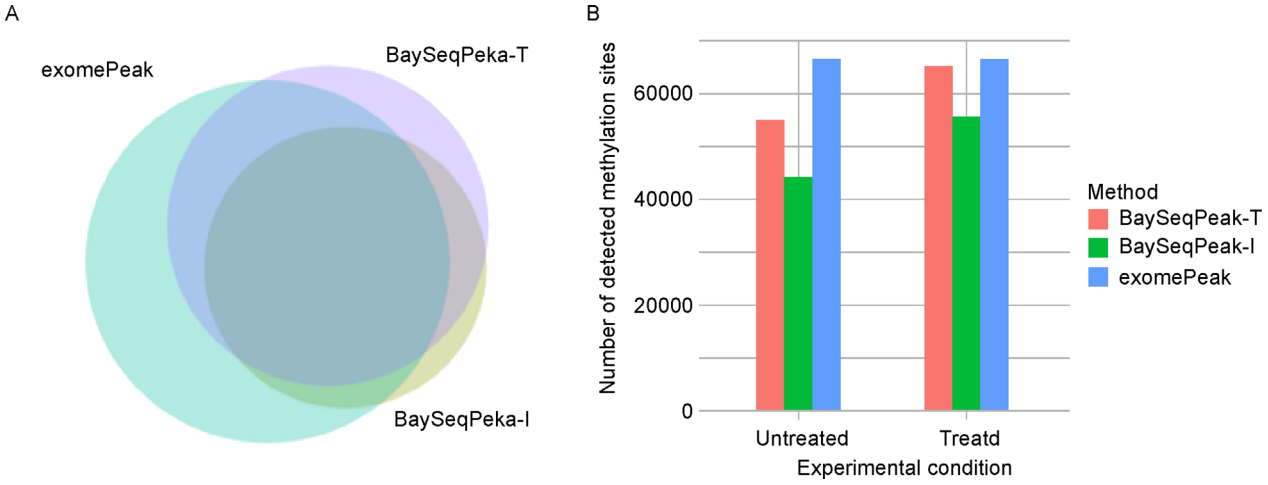

Figure 5. A comparison of the real data results obtained by our method and exomePeak. (A) Venn diagram of total detected methylated windows by BaySeqPeak-T, -I along with exomePeak in WT datasets. The sizes of the circles are proportional to the number of methylated bins found by each method. (B) Total detected methylation sites (adjacent methylated windows are clustered to one methylated site) in WT and FTO knock out datasets by our methods and exomePeak.

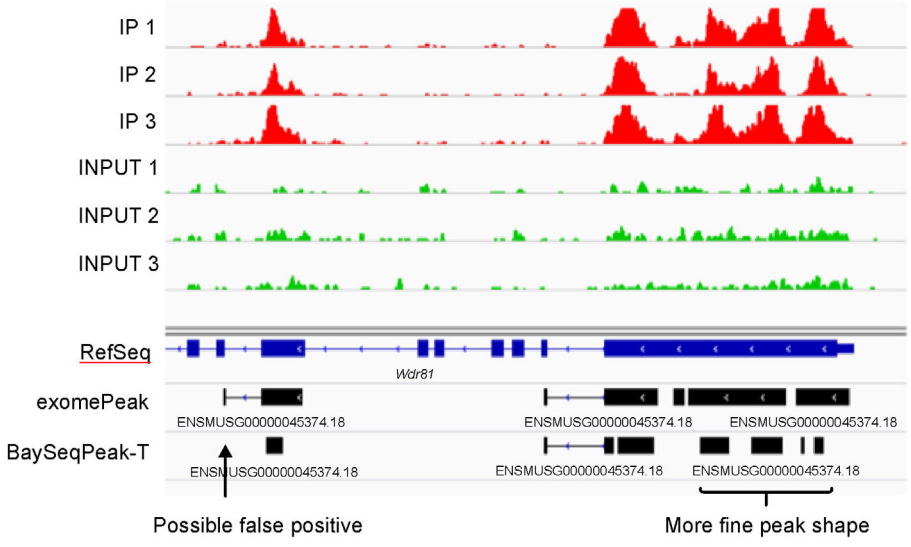

Figure 6. An exemplary methylation region detected by both exomePeak and BaySeqPeak-T shown in the IGV browser.

FTO-KO condition, while exomePeak identified 66,696 and 66,388 methylation sites in two conditions. More methylation sites being found in FTO-KO condition than WT condition by our methods (Figure 5B) is consistent with the fact that FTO is a demethylase, knocking out of which should increase the m^6A peaks. However, the result of exomePeak was contradictory. To investigate the reason why our method detected fewer peaks than exomePeak, we visualize the identified peaks in the IGV browser. Figure 6 shows one example of detected methylation sites in *Wdr81* gene. The figure indicates that there are 4 peaks by comparing IP samples with input samples, and our method correctly detected all 4 peaks. Though exomePeak also reported the corresponding methylated region, it reported 3 additional peaks, which are likely false positive discoveries from the figure. Furthermore, it only claimed the middle 2 peaks to be a single methylated site, but the detailed peak information was lost, and such case happened often. To get a sense of the overall performance methods, we compared the distribution of log fold change

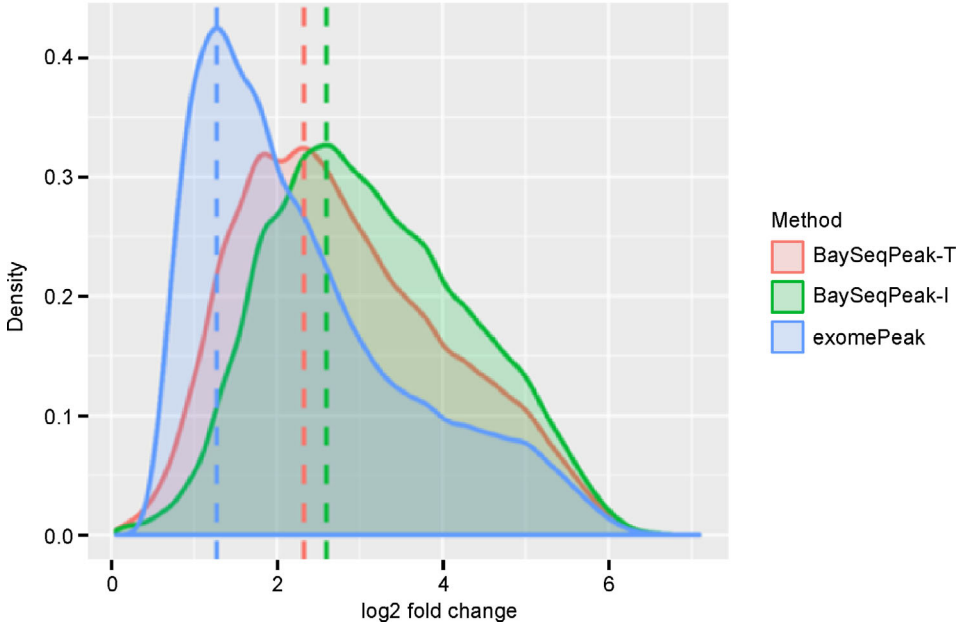

Figure 7. Distribution of log2 fold change in read counts of detected methylated sites by BaySeqPeak-T, -I and exomePeak.

of identified methylation sites (dividing total read counts in IP samples by total read counts in input sample) by each method in Figure 7. The figure shows that the methylation sites identified by BaySeqPeak have larger average fold change than exomePeak, and BaySeqPeak-I has slightly larger average fold change than BaySeqPeak-T, which indicates that the exomePeak may include a lot of false positive discoveries, while BaySeqPeak-I may miss some potential peaks without using spatial information. Overall, our method is more reliable in identifying methylation sites and more capable of detecting peaks in higher resolution, compared with the exomePeak method.

CONCLLUSION

In this paper, a hierarchical Bayesian model is proposed to detect methylation peaks in MeRIP-seq data. By deploying a zero-inflated negative binomial model, our algorithm tackles the zero-inflated and over-dispersed count problem of MeRIP-seq data, which had not been properly solved before. A hidden Markov model (HMM) is also incorporated to model the nearby methylation status dependency. A Markov chain Monte Carlo (MCMC) method was used to draw samples from the posterior distribution and infer the model parameters. The proposed method outperformed the previous package exomePeak under all simulation settings that suffered a high rate of false discovery. A real case study revealed our method to have enough sensitivity to detect the majority of peaks found by exomePeak along with higher specificity against low count data, and to be able to capture the fine structures of peaks that are spatially approximate to each other. Moreover, the proposed method can be easily extended to accommodate study designs that compare multiple factors under an ANOVA setting.

ACKNOWLEDGEMENTS

The authors would like to thank Jessie Norris for helping with proofreading the manuscript. This work was partially supported by the National Institutes of Health (Nos. R01CA172211, P50CA70907, P30CA142543, R01GM-115473, R01GM117597, R15GM113157, and R01CA152301), and the Cancer Prevention and Research Institute of Texas (No. RP120732).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Minzhe Zhang, Qiwei Li and Yang Xie declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Suzuki, M. M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet., 9, 465–476
2. Shi, Y. (2007) Histone lysine demethylases: emerging roles in development, physiology and disease. Nat. Rev. Genet., 8, 829–833
3. Motorin, Y. and Helm, M. (2011) RNA nucleotide methylation. Wiley Interdiscip. Rev. RNA, 2, 611–631

4. Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., *et al.* (2012)Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. Nature 485, 201–206
5. Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E. and Jaffrey, S. R. (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell, 149, 1635–1646
6. Machnicka, M. A., Milanowska, K., Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K. M., *et al.* (2013) MODOMICS: a database of RNA modification pathways – 2013 update. Nucleic Acids Res., 41, D262–D267
7. Desrosiers, R., Friderici, K. and Rottman, F. (1974) Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. Proc. Natl. Acad. Sci. USA, 71, 3971–3975
8. Adams, J. M. and Cory, S. (1975) Modified nucleosides and bizarre 5′-termini in mouse myeloma mRNA. Nature, 255, 28–33
9. Aloni, Y., Dhar, R. and Khoury, G. (1979) Methylation of nuclear simian virus 40 RNAs. J. Virol., 32, 52–60
10. Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., *et al.* (2014) A METTL3-METTL14 complex mediates mammalian nuclear RNA N^6-adenosine methylation. Nat. Chem. Biol., 10, 93–95
11. Ping, X.-L., Sun, B. F., Wang, L., Xiao, W., Yang, X., Wang, W. J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y. S., *et al.* (2014) Mammalian WTAP is a regulatory subunit of the RNA N^6-methyladenosine methyltransferase. Cell Res., 24, 177–189
12. Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y. G., *et al.* (2011) N^6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat. Chem. Biol., 7, 885–887
13. Yue, Y., Liu, J. and He, C. (2015) RNA N^6-methyladenosine methylation in post-transcriptional gene expression regulation. Genes Dev., 29, 1343–1355
14. Meyer, K. D., and Jaffrey, S. R. (2014) The dynamic epitranscriptome: N^6-methyladenosine and gene expression control. Nat. Rev. Mol. Cell Bio., 15, 313–326
15. Cao, G., Li, H-B., Yin, Z., Flavell, R. A. (2016) Recent advances in dynamic m^{6}A RNA modification. Open Biol ., 6,160003
16. Meng, J., Cui, X., Rao, M. K., Chen, Y. and Huang, Y. (2013) Exome-based analysis for RNA epigenome sequencing data. Bioinformatics, 29, 1565–1567
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
18. Cui, X., Meng, J., Rao, M. K., Chen Y. and Huang Y. (2015) HEPeak: an HMM-based exome peak-finding package for RNA epigenome sequencing data. BMC genomics 16(Suppl 4), S2
19. Cui, X., Meng, J., Zhang, S., Chen, Y. and Huang, Y. (2016) A novel algorithm for calling mRNA m^{6}A peaks by modeling biological variances in MeRIP-seq data. Bioinformatics, 32, i378–i385
20. Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal., 1, 515–534
21. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res., 18, 1509–1517
22. Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. BMC Bioinform., 11, 94
23. Anders, S., and Huber W. (2010) Differential expression analysis for sequence count data. Genome Boil., 11, R106
24. Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140
25. Witten, D., Tibshirani, R., Gu, S., Fire, A. and Lui, W. -O. (2010) Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biol., 8, 58
26. Witten, D. M. (2011) Classification and clustering of sequencing data using a Poisson model. Ann. Appl. Stat., 5, 2493–2518
27. Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics, 13, 523–538
28. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621–628
29. Morris, C. N. (1983) Parametric empirical Bayes inference: theory and applications. J. Am. Stat. Assoc., 78, 47–55
30. Gelman, A. (2008) Objections to Bayesian statistics. Bayesian Anal., 3, 445–449
31. Li, Q., Guindani, M., Reich, B. J., Bondell, H. D. and Vannucci, M. (2017) A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. Statistical Analysis and Data Mining: The ASA Data Science Journal, 10, 393–409
32. Guha, S., Li, Y. and Neuberg, D. (2008) Bayesian hidden Markov modeling of array CGH data. J. Am. Stat. Assoc., 103, 485–497
33. Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics, 5, 155–176
34. Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. Stat. Sci., 7, 457–472
35. Hess, M. E., Hess, S., Meyer, K. D., Verhagen, L. A., Koch, L., Brönneke, H. S., Dietrich, M. O., Jordan, S. D., Saletore, Y., Elemento, O., *et al.* (2013) The fat mass and obesity associated gene (*Fto*) regulates activity of the dopaminergic midbrain circuitry. Nat. Neurosci., 16, 1042–1048
36. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE47217
37. Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., Rao, M. K. and Huang, Y. (2014) A protocol for RNA methylation differential analysis with MeRIP-seq data and exomePeak R/Bioconductor package. Methods, 69, 274–281

