MOSAiCS-HMM: A model-based approach for detecting regions of histone modifications from ChIP-seq data

Dongjun Chung, Qi Zhang, and Sündüz Keleş

Abstract Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) experiments are routinely utilized for studying epigenomics of transcriptional regulation. We review some of the important statistical issues in the analysis of these experiments and extend our previous model for the analysis of ChIP-seq data of transcription factors, named MOSAiCS, with a hidden Markov model architecture (MOSAiCS-HMM). MOSAiCS-HMM provides a model-based approach for modeling read counts in histone modification ChIP-seq experiments and accounts for the spatial dependence in their ChIP-seq profiles. In addition, its `R` package implementation provides many functionality for summarizing these data and generating files that can be directly uploaded to the UCSC genome browser.

1 Introduction

Regulation of gene expression is a multi-faceted process. DNA binding proteins, i.e., transcription factors, and histone modifications are two of the critical mechanisms for regulating gene expression. Transcription factors (TFs) interact with the DNA in a sequence specific or non-specific manner and can act alone or in protein

Dongjun Chung
Department of Biostatistics, Yale School of Public Health, Yale University, 60 College Street, P.O. Box 208034, New Haven, CT 06520-8034. e-mail: dongjun.chung@yale.edu

Qi Zhang
Department of Biostatistics and Medical Informations, School of Public Health and Medicine, University of Wisconsin, 2130C Genetics/Biotechnology Center, 425 Henry Mall, Madison, WI 53706. e-mail: qizhang@stat.wisc.edu

Sündüz Keleş
Departments of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, 2124 Genetics/Biotechnology Center, 425 Henry Mall, Madison, WI 53706. e-mail: keles@stat.wisc.edu

complexes with co-factors. They promote (activate) or block (repress) expression of their specific target genes. In contrast, histones are a specific class of proteins that package DNA. Every 146 base pairs of DNA winds around a histone complex consisting of two of each of the H2A, H2B, H3, and H4 histone proteins, and form the structural unit of DNA called nucleosomes. The H3 and H4 histones have long tails that can be covalently modified at several places. Methylation, acetylation, and phosphorylation are some of the most commonly studied histone modifications and they affect diverse biological processes including gene regulation [30].

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) has become a versatile experimental technique for profiling TF-DNA interactions, histone modifications, chromatin remodeling enzymes, RNA polymerase, and nucleosomes [2, 15]. A typical TF or histone modification ChIP-seq experiment involves isolating regions of the genome interacting with the protein of interest or undergoing the targeted modification. This is accomplished by first cross-linking proteins and associated chromatin in a cell lysate and then shearing DNA to an average of 500 base pair fragments. Then, the DNA fragments associated with the protein of interest are selectively captured by immunoprecipitation with an antibody specific to that protein. In the case of histone modifications, antibodies targeting specific histone proteins with a specific modification are utilized. The associated DNA fragments are then purified and one (single-end sequencing) or both ends (paired-end sequencing) of the captured fragments are sequenced by using a high throughput sequencing platform.

These high throughput *in vivo* biological assays are embraced by large consortia projects such as ENCODE [10] and RoadMap EpiGenomics [4] and have resulted in large volumes of publicly available data. ChIP-seq experiments for transcription factors enable identification of where a protein binds in the genome *in vivo*, whereas experiments targeting histone modifications identify which regions of the genome are undergoing the targeted histone modification. Because both binding of transcription factors and histone modifications play important roles in cell specific gene regulatory programs, their genome-wide mapping is crucial for understanding and diagnosing human diseases.

Characteristics of data from ChIP-seq experiments vary based on what is being profiled (e.g., transcription factor, modified histone, RNA polymerase) and what sequencing parameters (e.g., single-end, paired-end) are being utilized. Illumina platform is by far the most popular choice for ChIP-seq experiments [2, 15, 21, 27]. As a result of sequencing, reads of size 36-100 base pairs (bps) representing one or both ends of immunoprecipitated DNA fragments with varying lengths are obtained. The lengths of DNA fragments are typically kept around 150-300 bps for optimal sequencing by a size selection step in the experimental protocol. ChIP-seq experiments are typically coupled with control experiments which either skip the immunoprecipitation step (Input control) or use a non-specific antibody (IgG control) to measure non-specific protein DNA interactions and characterize background read distribution. Compared to their array-based analogues (ChIP-chip experiments [5, 16]), ChIP-seq provides higher resolution and genomic coverage [38].

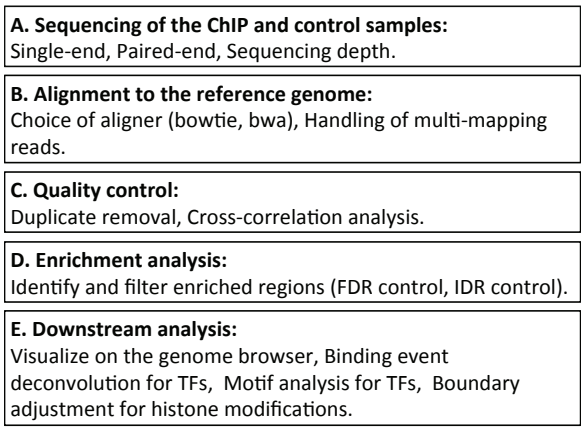

Fig. 1 *Typical work flow of statistical analysis of ChIP-seq experiments.*

Analysis of ChIP-seq data involves multiple steps from quality assessment to downstream analysis for biological interpretation (Figure 1). The main statistical task is, however, identifying regions of the genome that exhibit significantly higher levels of ChIP read counts compared to background read counts. Figure 2 displays ChIP and Input control profiles for such a region from a H3K4me3 experiment in GM12878 cell lines which was generated as part of the ENCODE project [10]. There are a plethora of computational and statistical approaches for analyzing data from ChIP-seq experiments (reviewed in [1]). Most of the well-studied approaches [7, 17, 18, 26, 37] are geared towards ChIP-seq experiments of transcription factors which generate punctuated peaks. In such data, ChIP reads concentrate on the TF-DNA interaction site and have a clear summit. In contrast, ChIP-seq experiments profiling modified histones can result in punctuated, broad (e.g., for H3K27me3, H3K36me3, and H3K9me3), or a mixture of punctuated and broad peaks and show larger variations in the widths of the enriched regions compared to TF ChIP-seq. Methods for analyzing ChIP-seq data of histone modifications either require running methods for punctuated signals in a special ”broad” model [17, 37] or primarily focus on identifying differential histone modifications [23, 28, 32, 34]. Recently, a stochastic Bayesian Change-Point method named BCP [33] has been proposed for the analysis of diffuse histone ChIP-seq data and has been shown to be also effective in analyzing punctuate transcription factor ChIP-seq data.

We have recently developed a model-based, versatile method, named MOSAiCS (Model-based One- and Two-Sample Analysis and Inference for ChIP-seq), for the analysis of ChIP-seq data [18]. MOSAiCS accommodates both one- (in the absence of a control sample) and two-sample analysis of ChIP-seq data. Unlike other popular ChIP-seq methods that consider explicit modeling of data only under the null hypothesis of no enrichment [26, 37], MOSAiCS provides biologically motivated statistical models for reads that arise under both non-enrichment (background)

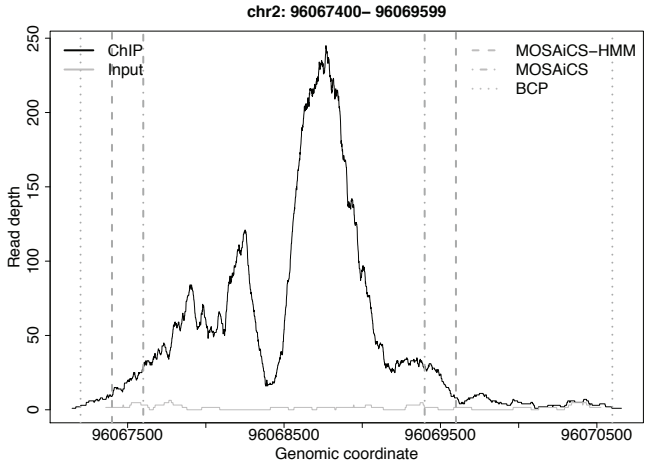

Fig. 2 *H3K4me3 ChIP-seq read profile generated by* `R` *package* `dpeak` *[7].* Black and gray curves depict ChIP and sequencing depth normalized Input read counts for a peak identified by all the three methods. Vertical lines depict the boundaries of the peak as determined by different peak callers.

and enrichment (signal). Furthermore, MOSAiCS builds a parametric background model that takes into account biases such as GC content [8] and mappability [38] that are inherent to ChIP-seq data. MOSAiCS model does not assume punctuated or broad peak structures but instead quantifies whether the ChIP reads show enrichment compared to the background reads for every genomic interval (e.g., bin) of user defined size in the genome. Although such analysis captures most parts of the broad domains, large regions with low but consistent enrichment might be prone to misidentification. In this paper, we extend the MOSAiCS model with a hidden Markov model architecture to allow spatial dependence between adjacent bins and facilitate identification of broad enriched regions in ChIP-seq data. We conclude with a brief discussion of other issues concerning ChIP-seq data analysis (Figure 1).

2 MOSAiCS-HMM Model

2.1 MOSAiCS

We first review the MOSAiCS model [18] that MOSAiCS-HMM builds on. Previous work by others and us have established that next generation sequencing datasets including naked DNA, Input DNA, and ChIP samples are prone to sequencing and other sources of biases [3, 8, 18, 26]. Specifically, observed read counts are affected by local sequence characteristics such as mappability and GC content. In

order to correct these biases and obtain accurate measurements of enrichment signals, we developed MOSAiCS, a flexible mixture model that incorporates various sequence biases in modeling the background read distribution. We implemented the MOSAiCS model as `R` package `mosaics` which is available from *Bioconductor* (http://www.bioconductor.org/) [12]. In this `R` package, the MOSAiCS model is implemented in a computationally efficient way by using `Rcpp` and `parallel R` packages for *C++* implementation and parallel computing, respectively. `mosaics` package also provides various tools for exploratory analysis, model fitting, model selection, and diagnostics for ChIP-seq data analysis with MOSAiCS [31].

In the MOSAiCS model, reference genome is divided into non-overlapping intervals (e.g., bins) of typically 200 bps. We consider ChIP reads in each bin as arising from a mixture of non-enriched and enriched distributions. Let Y_j denote the ChIP read counts in j-th bin. Let M_j and GC_j be the bin-specific mappability and GC content scores. These quantities are defined as functions of base pair mappability and GC scores [6]. For a read length of k and library size of L, let $x_{(i):(i+k-1)}$ denote the $k$mer starting at position i and ending at position $i+k-1$ from $5'$ to $3'$. Let $x^c_{(i):(i-k+1)}$ denote the $k$mer starting at position i and ending in $i-k+1$ in the other strand. Then, the nucleotide-level mappability is defined as:

$$\delta_i = \begin{cases} 1 \text{ if } \quad x_{(i):(i+k-1)} \quad \text{is unique,} \\ 0 \text{ o.w.} \end{cases}$$

Mappability for a position in the reverse strand is similarly defined as:

$$\delta^c_i = \begin{cases} 1 \text{ if } \quad x^c_{(i):(i-k+1)} \quad \text{is unique,} \\ 0 \text{ o.w.,} \end{cases},$$

where $\delta^c_i = \delta_{i-k+1}$. The GC content at the nucleotide level is defined similarly by setting $\delta_i = I(i\text{-th position is a G or C})$. In the MOSAiCS model, bin-level versions of these quantities are utilized to account for the fact that the total number of observed counts at position i could be contributed by forward strand reads that originate between positions $i-L+1$ and i and get extended to L bps or the reverse strand reads that originate between positions i and $i+L-1$ and get extended to L bps. The bin-level mappability/GC content for single-end reads is defined as:

$$\delta^*_i = \frac{1}{2L}\left(\sum_{j=i-L+1}^{i} \delta_j + \sum_{j=i}^{i+L-1} \delta^c_j\right), \tag{1}$$

$$= \frac{1}{2L}\left(\sum_{j=i-L+1}^{i} \delta_j + \sum_{j=i-k+1}^{i+L-k} \delta_j\right). \tag{2}$$

Bin-level mappability and GC content scores for paired-end reads can be computed similarly by taking into account the actual lengths of the fragments that two end reads represent.

When a matching control sample, such as Input control, is available, we further denote the control read counts in j-th bin by X_j. Finally, we denote the indicator of enrichment status of j-th bin as Z_j, where $Z_j = 1$ if j-th bin is enriched, i.e., exhibiting TF binding or histone modification, and $Z_j = 0$ otherwise. We assume that enrichment status of individual bins are independent and is given as follows for $j = 1, 2, \cdots, M$,

$$\Pr(Z_j = 0) = \pi_0, \quad \Pr(Z_j = 1) = 1 - \pi_0. \tag{3}$$

Given these underlying enrichment states for j-th bin, we assume that

$$(Y_j | Z_j = 0) \sim N_j, \quad (Y_j | Z_j = 1) \sim N_j + S_j, \tag{4}$$

where N_j and S_j represent background and signal, respectively. MOSAiCS models reads from the background component with Negative Binomial regression:

$$N_j \sim NegBin(a, a/\mu_j), \tag{5}$$

where we model its mean, μ_j, slightly differently under three different scenarios. The specifications of these models emerged from exploratory analysis of a large collection of ENCODE datasets [18] and other datasets across multiple organisms [14, 22, 29]. The mappability scores contribute the mean model with a log transformation to account for the curvature that is apparent from the mappability versus ChIP read count relationship. Similarly, the piecewise linear B-spline model for the GC-content score enables a flexible way of capturing the GC content versus ChIP read count relationship observed in multiple ChIP-seq datasets. Next, we detail the three mean models and discuss the conditions under which they are appropriate.

- Case 1: In the absence of a control sample:

  $$\log \mu_j = \beta_0 + \beta_M \log_2(M_j + 1) + \beta'_{GC} \mathbf{Sp}(GC_j),$$

  where $\mathbf{Sp}(GC_j)$ is a vector of piecewise linear B-spline basis functions with knots at the first and third quantiles of the GC content. β_{GC} is vector valued and represents all the coefficients in the spline model. Current standard practice for ChIP-seq experiments is to couple each ChIP sample with a Input control sample. However, investigators occasionally generate ChIP samples without control samples especially when choosing among different antibodies for the same factor. This mean model facilitates the analysis of such samples without a control by approximating the background mean read counts using mappability and GC content scores.
- Case 2a: In the presence of a shallowly sequenced control sample:

  $$\begin{aligned} \log \mu_j = \beta_0 + &\left[\beta_M \log_2(M_j + 1) + \beta'_{GC} \mathbf{Sp}(GC_j) \beta_{X1} X_j^d\right] 1\{X_j \le s\} \\ &+ \beta_{X2} X_j^d 1\{X_j > s\}, \end{aligned}$$

where $1\{A\}$ is an indicator function for set A, and s and d are tuning parameters. This model is essentially performing a power transformation with exponent d on the control read counts and incorporating mappability and GC content values for bins with less than or equal to s control read counts. In our previous work [18], we have shown that even in the presence of a control sample, utilizing mappability and GC content values for estimating the background read distribution might improve detection power and eliminate false positives. From a practical point of view, inclusion of mappability and GC content values matters the most when the background read distribution cannot be estimated well just based on the control sample. MOSAiCS framework provides goodness-of-fit plots (Figure 3(a)) which aid in this decision.

- Case 2b: In the presence of an adequately sequenced control sample:

$$\log \mu_j = \beta_0 + \beta_X X_j^d,$$

where d is again the exponent in the power transformation of the control read counts. This model is suitable for cases where the control sample is deeply sequenced and the fit can again be evaluated by the goodness-of-fit plots provided by MOSAiCS. Since its publications, we have applied MOSAiCS to tens to a few hundreds of datasets and observed that $s = 2$ and $d = 0.25$ work well in practice. Therefore, these values are currently the default values in the `mosaics` `R` package.

For the signal component, we consider both a single negative binomial and a mixture of two negative binomial distributions, i.e.,

$$(1) \quad S_j \sim NegBin(b,c) + k,$$
$$(2) \quad S_j \sim p_1 NegBin(b_1,c_1) + (1-p_1) NegBin(b_2,c_2) + k,$$

where k is a constant set to 3 and represents the minimum observable read count in an enriched region. The optimal model for signal component is determined based on Bayesian information criterion (BIC). The parameters in the MOSAiCS model are estimated using a computationally efficient Expectation-Maximization (EM) algorithm described in [18].

After we fit the MOSAiCS model, enriched regions are identified using a direct posterior probability approach [24] for false discovery rate (FDR) control based on the posterior probability that read counts for each bin are generated from the background component. Specifically, we first rank the bins according to increasing values of $\Pr\left(Z_j = 0|\mathbf{Y};\hat{\Theta}\right)$, where $\hat{\Theta}$ denotes the final parameter estimates obtained from the EM algorithm. Let fdr_j denote the sorted $\Pr\left(Z_j = 0|\mathbf{Y};\hat{\Theta}\right)$ values. Then, we increase the cutoff κ until the expected proportion of false discoveries given by

$$\frac{\sum_{j=1}^{M} fdr_j 1\left\{fdr_j \leq \kappa\right\}}{\sum_{j=1}^{M} 1\left\{fdr_j \leq \kappa\right\}},$$

reaches the pre-specified cutoff (α) for false discovery rate. Finally, using this determined cutoff $\hat{\kappa}$, bins satisfying the condition that $\Pr\left(Z_j = 0|\mathbf{Y};\hat{\Theta}\right) \leq \hat{\kappa}$ are reported as enriched regions. FDR control ensures that reported enriched regions achieve a certain level of statistical significance. However, in addition to statistical significance, investigators often would like to require each enriched region to have a minimum number of ChIP reads. Therefore, `R` package `mosaics` allows such a threshold as input. In practice, setting this threshold to a certain percentile (e.g., $0.90-0.99$) of the ChIP read count distribution works well if the control sample is shallowly sequenced (e.g., less than 20 million reads for human samples). In the presence of a deeply sequenced control sample, this threshold can also be set to a depth normalized percentile of the control read count distribution.

2.2 MOSAiCS-HMM

In the MOSAiCS model, enrichment states of adjacent bins are assumed to be independent. This assumption might be mildly violated in practice for the ChIP-seq data of TFs with narrow enrichment profiles that typically span 1-3 bins. However, it is more likely to be invalid for the ChIP-seq data of histone modifications which can easily cover a larger number of bins and might exhibit broad enrichment signals. In the case of broad signals, multiple adjacent bins constitute a wide block-shaped peak and a spatial correlation structure underlies the relation between enrichment status of adjacent bins. Hidden Markov Models (HMMs) provide a graceful way to handle these types of spatial correlations without losing spatial resolution (reviewed in [9] and [25] among many others). This observation motivates our development of the MOSAiCS-HMM framework to account for spatial correlations in ChIP-seq data.

In MOSAiCS-HMM, we assume that enrichment states constitute a Markov chain along each chromosome. Specifically, Eq. (3) of the MOSAiCS model is replaced by

$$\Pr\left(Z_{j+1} = b|Z_j = a\right) \equiv \pi_{ab}, \quad a,b \in \{0,1\}, \quad j = 1,\cdots,M-1 \tag{6}$$

and $\sum_{b=0}^{1} \pi_{ab} = 1$ for $a = 0,1$. Finally, conditional on these underlying enrichment states, ChIP read counts are assumed to follow the read count distributions of the MOSAiCS model, given in Eqs. (4), (5), and (6). This allows effective adjustment of sequence biases in the binding site identification, as shown in [18].

2.3 Parameter Estimation for the MOSAiCS-HMM Model

We estimate the parameters of MOSAiCS-HMM using the Baum-Welch algorithm, which is a special case of the EM algorithm. MOSAiCS-HMM inherits estimates

of the emission distributions from the MOSAiCS fits to the data. Although this is in principle statistically inefficient, the MOSAiCS-HMM goodness-of-fit plots suggest that this procedure results in good fit to the data (Figure 3(b)). More importantly, this approach accelerates fitting of MOSAiCS-HMM significantly because the Baum-Welch algorithm needs to only estimate the transition matrix and state probabilities for the starting bin. We fit the MOSAiCS-HMM model to each chromosome separately because a smooth transition between end of one chromosome and start of another chromosome is not expected. Furthermore, by analyzing each chromosome separately, the fitting of MOSAiCS-HMM can be easily parallelized to decrease computational cost.

Since MOSAiCS-HMM fit utilizes the background and signal distribution estimates of the MOSAiCS fit, the parameters that need to be estimated for each chromosome with the Baum-Welch algorithm are $\Theta = (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}, \pi_{*0}, \pi_{*1})$, where π_{00}, π_{10}, π_{01}, and π_{11} are transition probabilities defined in Eq. (6), and π_{*0} and π_{*1} are probabilities of enrichment states in the first bin of each chromosome, i.e., $\pi_{*0} \equiv \Pr(Z_1 = 0)$, $\pi_{*1} \equiv \Pr(Z_1 = 1)$, and $\pi_{*0} + \pi_{*1} = 1$. Then, the complete data likelihood function is given by

$$L_c = \prod_{k=0}^{1} \pi_{*k}^{Z_{0k}} \prod_{j=1}^{M-1} \prod_{k=0}^{1} \prod_{l=0}^{1} \pi_{kl}^{Z_{jk} Z_{(j+1)l}} \prod_{j=1}^{M} \prod_{l=0}^{1} \left\{ \Pr(Y_j | Z_j = l) \right\}^{Z_{jl}}.$$

Because $\Pr(Y_j | Z_j = l)$ are obtained from the MOSAiCS fit, the MOSAiCS-HMM EM algorithm iterates between the following E- and M-steps until the likelihood or the parameter estimates converge or a fixed number of iterations specified by the user is reached. For the m-th iteration, we have the following and E- and M-steps.

E-step:

We first update the conditional probabilities of the enrichment states $k = 0, 1$ in the first bin of each chromosome as

$$z_{*k}^{(m)} = \Pr\left(Z_1 = k | \mathbf{Y}; \Theta^{(m)} \right) = \frac{\pi_{*k}^{(m)} \Pr(Y_1 | Z_1 = k)}{P(Y_1; \Theta^{(m)})}.$$

The conditional expectation of transition between the states can be computed efficiently using the forward and backward algorithms as follows. In the forward algorithm, we have

$$\begin{aligned} f_{1l}^{(m)} &= \pi_{*l}^{(m)} \Pr(Y_1 | Z_1 = l), \quad l = 0, 1, \\ f_{jl}^{(m)} &= Pr\left(Y_1, Y_2, \cdots, Y_j, Z_j = l; \Theta^{(m)} \right) \\ &= \Pr(Y_j | Z_j = l) \sum_{k=0}^{1} f_{(j-1)k}^{(m)} \pi_{kl}^{(m)}, \quad j =, 2, 3 \cdots, M, \quad l = 0, 1. \end{aligned}$$

In the backward algorithm, we have

$$
\begin{aligned}
b_{Mk}^{(m)} &= 1, \quad , k = 0,1, \\
b_{jk}^{(m)} &= Pr\left(Y_{j+1}, Y_{j+2}, \cdots, Y_M | Z_j = k; \Theta^{(m)}\right) \\
&= \sum_{l=0}^{1} \pi_{kl}^{(m)} \Pr(Y_{j+1} | Z_{j+1} = l) b_{(j+1)l}^{(m)}, \quad j = (M-1), (M-2), \cdots, 1, \quad k = 0,1.
\end{aligned}
$$

Finally, we calculate the conditional probabilities of transition from state $k = 0,1$ to $l = 0,1$ based on the quantities from the forward and backward algorithms as

$$
\begin{aligned}
z_{jkl}^{(m)} &= \Pr\left(Z_j = k, Z_{j+1} = l | \mathbf{Y}; \Theta^{(m)}\right) \\
&= \frac{f_{jk}^{(m)} \pi_{kl}^{(m)} \Pr(Y_{j+1} | Z_{j+1} = l) b_{(j+1)l}^{(m)}}{P(\mathbf{Y}; \Theta^{(m)})}, \quad j = 1, 2, \cdots, (M-1).
\end{aligned}
$$

M step:

For states $k = 0,1$ and $l = 0,1$, we update the transition probabilities as

$$
\pi_{kl}^{(m+1)} = \frac{\sum_{j=1}^{M-1} z_{jkl}^{(m)}}{\sum_{l\prime=0}^{1} \sum_{j=1}^{M-1} z_{jkl\prime}^{(m)}}
$$

and the probabilities of states $k = 0,1$ in the first bin of each chromosome as

$$
\pi_{*k}^{(m+1)} = \frac{z_{*k}^{(m)}}{\sum_{k\prime=0}^{1} z_{*k\prime}^{(m)}}.
$$

We use the scaling procedures provided in [9] to avoid numerical problems in the forward and backward algorithms.

With MOSAiCS-HMM, users can finalize the set of enriched regions by either the Viterbi algorithm or the posterior decoding. If the Viterbi algorithm is used, the most likely sequences of enrichment states are determined across each chromosome. With the posterior decoding approach, enrichment state of each bin is determined using the direct posterior probability approach [24] for FDR control. We next discuss the details of the decoding procedures.

2.3.1 Viterbi Algorithm for the MOSAiCS-HMM Model

The Viterbi algorithm for MOSAiCS-HMM identifies the most likely sequences of enrichment states across each chromosome, i.e.,

$$
\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z}} \Pr\left(\mathbf{Y}, \mathbf{Z}; \hat{\Theta}\right),
$$

where $\hat{\Theta}$ is the final parameter estimates obtained from the EM algorithm. Specifically, the Viterbi algorithm is implemented in the following four steps. First, in the initialization step, for states $l = 0, 1$, we set

$$\begin{aligned} v_{1l} &= \hat{\pi}_{*l} \Pr(Y_1 | Z_1 = l), \\ ptr_{1l} &= 0, \end{aligned}$$

where $\hat{\pi}_{*l}$ is the final estimate for π_{*l}. Second, in the recursion step, from bin $j = 2, 3, \cdots$ to bin M,

$$\begin{aligned} v_{jl} &= \Pr(Y_j | Z_j = l) \max_k \left\{ v_{(j-1)k} \hat{\pi}_{kl} \right\}, \\ ptr_{jl} &= \arg\max_k \left\{ v_{(j-1)k} \hat{\pi}_{kl} \right\}, \end{aligned}$$

where $\hat{\pi}_{kl}$ is the final estimate for π_{kl}. Third, in the termination step, we set

$$\hat{z}_M = \arg\max_k \left\{ v_{Mk} \right\}.$$

Finally, in the trace back step from bin $j = (M-1), (M-2), \cdots$ to bin 1,

$$\hat{z}_j = ptr_{(j+1)\hat{z}_{j+1}},$$

where $\hat{z}_j$ is the estimated state for j-th bin.

2.3.2 Posterior Decoding for MOSAiCS-HMM Model

In the posterior decoding approach, the enrichment state for j-th bin is determined using the direct posterior probability approach of [24] for false discovery rate control based on the following posterior probabilities:

$$\Pr\left(Z_j = k | \mathbf{Y}; \hat{\Theta}\right) = \frac{\hat{f}_{jk} \hat{b}_{jk}}{P(\mathbf{Y}; \hat{\Theta})},$$

where $\hat{\Theta}$ denotes the final parameter estimates obtained from the EM algorithm, and $\hat{f}_{jk}$ and $\hat{b}_{jk}$ are the values from the forward and backward algorithms based on the final parameter estimates. Specifically, we first rank the bins according to increasing values of $\Pr\left(Z_j = 0 | \mathbf{Y}; \hat{\Theta}\right)$ and denote these sorted values with fdr_j. Then, we increase the cutoff κ until the expected proportion of false discoveries given by

$$\frac{\sum_{j=1}^{M} fdr_j 1\left\{fdr_j \leq \kappa\right\}}{\sum_{j=1}^{M} 1\left\{fdr_j \leq \kappa\right\}},$$

reaches the pre-specified false discovery rate α. Finally, using this determined cutoff $\hat{\kappa}$, we report the bins satisfying the condition that $\Pr\left(Z_j = 0 | \mathbf{Y}; \hat{\Theta}\right) \leq \hat{\kappa}$ as enriched regions.

The MOSAiCS-HMM model is now part of the `R` package `mosaics` ($\geq$ 1.6.0).

3 Case Study: H3K4me3 Profiling in GM12878 Cells

We used ChIP-seq data of H3K4me3 in GM12878 cells from the ENCODE project to evaluate performances of MOSAiCS, MOSAiCS-HMM, and BCP. The dataset contained two ChIP replicates with 21.3 and 18.1 million aligned reads each. Each ChIP sample was analyzed with respect to a common Input control sample of 13.4 million aligned reads. We used the default parameter values for BCP and a false discovery rate of 0.05 and a threshold value equal to the 99-th percentile of the bin-level ChIP read counts for MOSAiCS and MOSAiCS-HMM with the posterior decoding approach. Overall, MOSAiCS-HMM fits had better BIC values than the MOSAiCS fits for both replicates (36,273,306 (MOSAiCS) versus 33,169,356 (MOSAiCS-HMM) for replicate 1; 32,568,329 (MOSAiCS) versus 29,652,020 (MOSAiCS-HMM) for replicate 2). The goodness-of-fit plots indicate that both models fit the data adequately (Figure 3(b)).

BCP identified 17664 and 16964 peaks for the two replicates whereas MOSAiCS and MOSAiCS-HMM identified 16438 and 20079 peaks for replicate 1 and 16730 and 20294 peaks for replicate 2, respectively. We allowed MOSAiCS to merge enriched bins that are within 200 bps of each other to facilitate identification of wide enriched regions. We then evaluated the replicate consistency of the methods by ranking and comparing the peaks from the two replicates of each method. For BCP, peak-specific posterior means, which are the only statistical measurements of enrichment reported in the BCP output, were used for ranking whereas for MOSAiCS and MOSAiCS-HMM, maximum signal which denotes the maximum bin-level ChIP read count within the peak region was used. Similar results were obtained when the MOSAiCS and MOSAiCS-HMM peaks were ranked with respect to their maximum posterior probability of enrichment over the bins within the enriched regions. Figure 4(a) depicts the percentage overlap between the peak sets of the two replicates for each method as a function of peak rank. We note that both MOSAiCS and MOSAiCS-HMM provide better ranking of the peaks than BCP. The final overlap percentages of the peak sets of the replicates are comparable between the methods, indicating that MOSAiCS-HMM is identifying more peaks with the same overlap consistency rate. This overlap analysis is based on the original widths of the peaks reported by each method. Figure 4(b) displays the median widths across top $500, 1000, 1500, \cdots, 20000$ peaks for the peak sets of replicate 2 from each method. Similar results are obtained with replicate 1 (data not shown). We observe that BCP peaks are the widest. Despite this, overlap percentages of the ranked BCP peaks are the smallest as illustrated in Figure 4(a). BCP peaks often have long flanking regions lacking enrichment (Figure 2) or a single BCP peak harbours multiple enriched regions separated by long regions lacking enrichment. An example of the latter case is provided in Figure 5, where two enriched regions separated by about 5000 bps are reported as a single peak. We also note that MOSAiCS-HMM actually

provides slightly narrower peaks than MOSAiCS. This indicates that the gain from the HMM architecture cannot simply be attained by merging of enriched bins within close proximity of each other in the MOSAiCS output.

H3K4me3 is a promoter-specific histone modification associated with active transcription; therefore H3K4me3 enrichment is expected at the promoter regions of genes that are transcribed in GM12878. To evaluate biological relevance of peaks identified by each method, we overlapped promoter regions of the expressed genes in GM12878 with each of the peak sets. Expressed genes are defined based on ENCODE2 RNA-seq data from GM12878 by subsetting genes with transcripts per million larger than 20. For each gene, we defined the promoter region as the [-1000, +500] bps interval anchored at the transcription start site. Since wider peaks are expected to provide high overlap by definition, we resized the peaks of each method to 2000 bps by using the midpoint of the peak as the anchoring point. MOSAiCS pipeline reports a summit. Ideally, a summit denoting the location of the highest signal would be a better anchoring point for all the methods; however since BCP only reports intervals of enrichment, using the midpoint as the anchor minimized the summit selection bias between the methods. Table 1 summarizes the total number of promoters that overlap with peak lists of each method and also specifies how many of the promoters are completely within a H3K4me3 peak. We observe that MOSAiCS-HMM peaks overlap with a larger fraction of the active promoters and completely cover the largest number of promoters. When the promoter overlap of the peaks is calculated using the original widths (numbers reported in parentheses in Table 1), a slightly higher number of promoters are overlapping with the BCP peaks; however as depicted in Figure 5, this gain comes at the price of many base pairs that lack any enrichment within the peak regions .

Table 1 *H3K4me3 peak coverage of the promoters of the 5979 expressed genes in GM12878.* The numbers of overlapping promoters are based on the intersection of promoters overlapping with peaks of both replicates. Numbers in parentheses denote the numbers of promoters overlapping with the peaks when the original peak widths are used. The numbers of completely covered promoters are based on the minimum of the number of promoters completely covered by the peaks of each of the replicates.

	BCP	MOSAiCS	MOSAiCS-HMM
# of overlapping promoters	2782 (5484)	4514 (5360)	4745 (5363)
# of completely covered promoters	546	656	704

4 Discussion

We presented an extension of MOSAiCS, named MOSAiCS-HMM, for analyzing ChIP-seq data of histone modifications. MOSAiCS-HMM can analyze ChIP-seq experiments with or without a Input control experiment and provides FDR control.

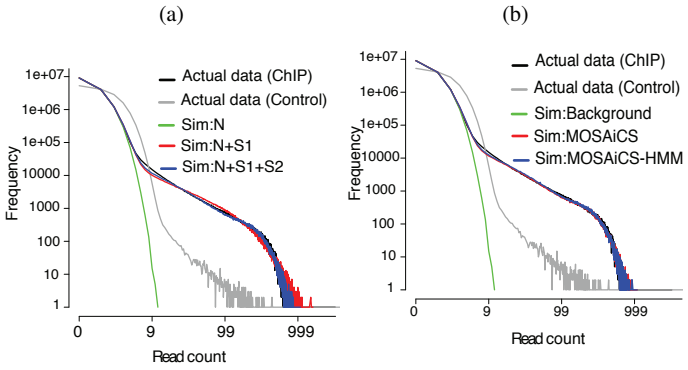

Fig. 3 *Goodness-of-fit (GOF) plots.* (a) MOSAiCS goodness-of-fit plots for replicate 1. Goodness-of-fit plot for replicate 1. Both axes are in the log10 scale. SIM:N: reads simulated from the estimated background read distribution. Sim:N+S1: reads simulated from the MOSAiCS model with one signal component for the ChIP reads. Sim:N+S1+S2: reads simulated from the MOSAiCS model with two signal components for the ChIP reads. Simulated data Sim:N+S1+S2 overlap the actual ChIP data well, indicating good overall fit. (b) MOSAiCS and MOSAiCS-HMM goodness-of-fit plots for replicate 1. Both axes are in the log10 scale. SIM:Background: reads simulated from the estimated background read distribution. Sim: MOSAiCS: reads simulated from the estimated MOSAiCS model with two signal components for the ChIP reads. Sim: MOSAiCS-HMM : reads simulated from the estimated MOSAiCS-HMM model with two signal components for the ChIP reads.

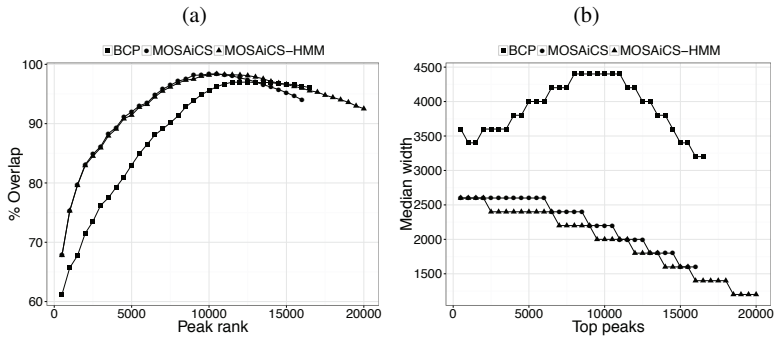

Fig. 4 *Comparison of MOSAiCS, MOSAiCS-HMM, and BCP on H3K4me3 ChIP-seq data from GM12878.* (a) Overlap percentages of enriched regions identified by two independent replicates as a function of the peak rank. (b) Median widths across top $(500, 1000, 1500, \cdots, 20000)$ ranked peaks.

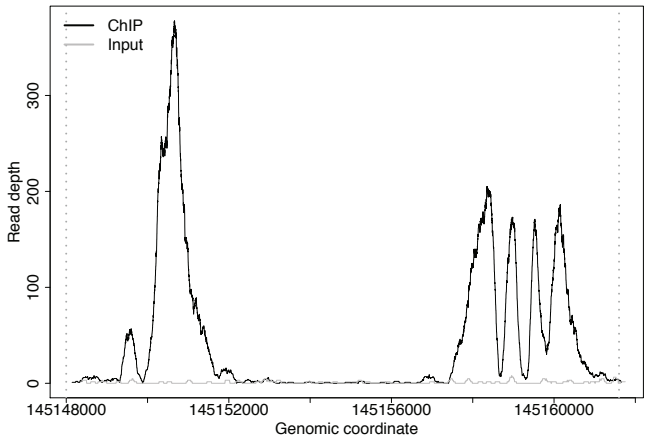

Fig. 5 *Comparison of MOSAiCS, MOSAiCS-HMM, and BCP on H3K4me3 ChIP-seq data from GM12878.* H3K4me3 ChIP-seq read profile generated by the `R` package `dpeak` [7] for a wide BCP peak. Dashed, vertical gray lines mark the boundaries of the BCP peak. Both MOSAiCS and MOSAiCS-HMM identify two peaks within this enriched region.

We conclude by discussing some other key issues related to histone ChIP-seq data, and more generally ChIP-seq data (Figure 1). The commonly used read lengths in ChIP-seq protocols are 50 to 100 bps. This results in about 10-25% of the reads aligning to multiple locations on the reference genome for human and mouse samples. These reads are commonly referred to as *multi-reads* and are typically discarded from the analysis. This leads to missing read data for highly repetitive regions of the genome and such reads are important to recover for studying TFs or histone modifications that interact with repetitive DNA. To address this issue, we developed a ChIP-seq-specific read mapper [6] named CSEM. This mapper utilizes Bowtie [20] alignments of the reads, where multi-reads are retained, and fractionally allocates multi-mapping reads by considering the local read contents of the mapping positions. As a result, it can generate both an alignment file with all the mapping reads and their allocation probabilities and a pseudo alignment file in bed file format where each multi-read is allocated to its most probable mapping location. The latter alignment file is accepted as input by multiple peak callers.

There are multiple quality control procedures developed for ChIP-seq data. Most notable of these is the cross-correlation analysis which is built on calculating cross-correlation between strand-specific genome-wide ChIP read profiles [19]. In ChIP-seq experiments with high signal-to-noise ratios, cross-correlation between the base-pair level forward and reverse strand read counts is expected to attain its maximum value around the average fragment length. A maximum cross-correlation value at a length vastly different than that of the average fragment size indicates potential problems with the ChIP-seq data and requires further attention. ChIP-seq experi-

ments are prone to a wide range of amplification biases. A commonly encountered bias is the extreme amplification of local regions. For such abnormally amplified regions, the same set of nucleotides covering the region appears in the data set hundreds of and even thousands of times. The common practice to alleviate problems due to abnormal amplification effects is the removal of multiple copies of a given read. More specifically, only a single read is allowed to start at each distinct genomic position. This feature is also part of the `mosaics` `R` package. Many ChIP-seq analysis methods provide some level of FDR control. However, the reliability of the FDR control typically depends on how well the assumed model fits the data. An alternative approach, which relies on the consistency of between two independent replicates of the ChIP-seq data, is control of irreproducible discovery rate (IDR). This approach has been widely adapted by the ENCODE project [19] and is shown to stabilize the number of peaks obtained from the same data set by different methods. When the MOSAiCS-HMM GOF plots indicate a lack of fit, IDR provides a robust alternative for choosing the number of peaks in MOSAiCS-HMM.

Once the enriched regions are identified in a ChIP-seq experiment, downstream analysis depends on the specific application. For TFs, especially in compact genomes, an important issue is the deconvolution of closely located binding events. Most of the commonly used ChIP-seq analysis methods [17, 18, 26, 37] are not specifically designed to deconvolve closely located binding; however, the number of methods which can perform such a task is on the increase [7, 13, 36]. In TF ChIP-seq experiments, summits of the peaks (predicted binding locations) are the main parameters of interest. In contrast, for histone ChIP-seq experiments, the boundaries of the enriched regions constitute one of the most important features. Most of the commonly used histone ChIP-seq analysis methods operate by binning the genome into small non-overlapping intervals. As a result, the resulting enriched regions might have inaccurate boundaries and require post trimming and extension procedures. It is often of interest to study multiple histone modifications simultaneously and divide genome into regions exhibiting different combinations of histone modifications [11]. To this end, we developed jMOSAiCS [35], which efficiently analyses multiple TF or histone modification datasets simultaneously and identifies regions showing combinatorial enrichment of the studied factors.

Acknowledgements

We thank Professor Colin Dewey of University of Wisconsin, Madison, for providing us with a RSEM-processed version of ENCODE GM12878 RNA-seq data. This research was supported by National Institutes of Health Grants HG007019 and HG003747 to S.K.

References

[1] Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J.: Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Computational Biology **9**(11), e1003,326 (2013)

[2] Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K.: High-resolution profiling of histone methylations in the human genome. Cell **129**(4), 823–837 (2007)

[3] Benjamini, Y., Speed, T.P.: Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research **40**, e72 (2012)

[4] Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A.: The NIH Roadmap Epigenomics Mapping Consortium. Nature Biotechnology **28**(10), 1045–1048 (2010)

[5] Buck, M.J., Lieb, J.D.: ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments . Genomics **84**, 349–360 (2004)

[6] Chung, D., Kuan, P.F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E.H., Dewey, C., Keleş, S.: Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-seq data. *PLoS Computational Biology* **7**, e1002,111 (2011)

[7] Chung, D., Park, D., Myers, K., Grass, J., Kiley, P., Landick, R., Keleş, S.: dPeak: High resolution identification of transcription factor binding sites from PET and SET ChIP-seq data. PLoS Computational Biology **9**(10), e1003,246 (2013)

[8] Dohm, J., Lottaz, C., Borodina, T., Himmelbauer, H.: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research **36**(16), e105 (2008)

[9] Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge university press (1998)

[10] ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M.: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)

[11] Ernst, J., Kellis, M.: Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology **28**, 817–25 (2010)

[12] Gentleman, R.C., Carey, V.J., Bates, D.M., others: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology **5**, R80 (2004)

[13] Guo, Y., Mahony, S., Gifford, D.K.: High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Computational Biology **8**, e1002,638 (2012)

[14] Jang, S.W., Srinivasan, R., Jones, E.A., Sun, G., Keles, S., Krueger, C., Chang, L.W., Nagarajan, R., Svaren, J.: Locus-wide identification of egr2/krox20 regulatory targets in myelin genes. Journal of Neurochemistry **115**(6), 1409–1420 (2010)

[15] Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B.: Genome-wide mapping of in vivo protein-DNA interactions. Science **316**(5830), 1497–1502 (2007)

[16] Keleş, S.: Mixture modeling for genome-wide localization of transcription factors. Biometrics **63**, 10–21 (2007)

[17] Kharchenko, P.V., Tolstorukov, M., Park, P.J.: Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology **6**, 1351–1359 (2008)

[18] Kuan, P., Chung, D., Pan, G., Thomson, J., Stewart, R., Keleş, S.: A Statistical Framework for the Analysis of ChIP-seq data. Journal of American Statistical Association **106**(459), 891–903 (2011)

[19] Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M.: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research **22**(9), 1813–1831 (2012)

[20] Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology **10**, R25 (2009)

[21] Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S., Bernstein, B.E.: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature **448**, 553–560 (2007)

[22] Myers, K.S., Yan, H., Ong, I.M., Chung, D., Liang, K., Tran, F., Kele, S., Landick, R., Kiley, P.J.: Genome-scale analysis of ¡italic¿escherichia coli¡/italic¿ fnr reveals complex features of transcription factor binding. PLoS Genetics **9**(6), e1003,565 (2013)

[23] Nair, N.U., Sahu, A.D., Bucher, P., Moret, B.M.E.: ChIPnorm: A statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. PLoS ONE **7**(8), e39,573 (2012)

[24] Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics **5**(2), 155–176 (2004)

[25] Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)

[26] Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., Gerstein, M.: PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. Nature Biotechnology **27**(1), 66–75 (2009)

[27] Seo, Y.K., Chong, H.K., Infante, A.M., In, S.S., Xie, X., Osborne, T.F.: Genome-wide analysis of SREBP-1 binding in mouse liver chromatin reveals a preference for promoter proximal binding to a new motif. PNAS **106**(33), 13,765–9 (2009)

[28] Song, Q., Smith, A.D.: Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics **27**(6), 870–1 (2011)

[29] Srinivasan, R., Sun, G., Keles, S., Jones, E.A., Jang, S.W., Krueger, C., Moran, J.J., Svaren, J.: Genome-wide analysis of egr2/sox10 binding in myelinating peripheral nerve. Nucleic Acids Research **40**(14), 6449–6460 (2012)

[30] Strahl, B.D., Allis, C.D.: The language of covalent histone modifications. Nature **403**(6765), 41–45 (2000)

[31] Sun, G., Chung, D., Liang, K., Keleş, S.: Methods in Molecular Biology Series, vol. 1038, chap. Statistical analysis of ChIP-seq data with MOSAiCS, pp. 193–212. Springer (2013)

[32] Taslim, C., Huang, T., Lin, S.: DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models. Bioinformatics **27**(11), 1569–70 (2011)

[33] Xing, H., Mo, Y., Liao, W., Zhang, M.Q.: Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. PLoS Computational Biology **8**(7) (2012)

[34] Xu, H., Wei, C.L., Lin, F., Sung, W.K.: An HMM approach to genome-wide identification of differential histone modification sites form chip-seq data. Bioinformatics **24**(20), 2344–2349 (2008)

[35] Zeng, X., R.Sanalkumar, Bresnick, E.H., Li, H., Chang, Q., Keleş, S.: jMOSAiCS: Joint analysis of multiple ChIP-seq datasets. Genome Biology **14**, R38 (2013)

[36] Zhang, X., Robertson, G., Krzywinski, M., Ning, K., Droit, A., Jones, S., Gottardo, R.: PICS: probabilistic inference for ChIP-seq. Biometrics **67**(1), 151–163 (2011)

[37] Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S.: Model-based analysis of ChIP-Seq (MACS). Genome Biology **9**(9), R137 (2008)

[38] Zhang, Z.D., Rozowsky, J., Snyder, M., Chang, J., Gerstein, M.: Modeling ChIP sequencing in silico with applications. PLoS Computational Biology **4**(8), e1000,158 (2008)