Joint modelling of ChIP-seq data via a Markov random field model

Y. Bao[1], V. Vinciotti[1,*], E. Wit[2] and P. 't Hoen[3,4]

[1]*School of Information Systems, Computing and Mathematics, Brunel University, UK*

[2]*Institute of Mathematics and Computing Science, University of Groningen, The Netherlands*

[3]*Department of Human Genetics, Leiden University Medical Center, The Netherlands*

[4]*Netherlands Bioinformatics Centre, The Netherlands*

Abstract: Chromatin ImmunoPrecipitation-sequencing (ChIP-seq) experiments have now become routine in biology for the detection of protein binding sites. In this paper, we present a Markov random field model for the joint analysis of multiple ChIP-seq experiments. The proposed model naturally accounts for spatial dependencies in the data, by assuming first order Markov dependence, and for the large proportion of zero counts, by using zero-inflated mixture distributions. In contrast to all other available implementations, the model allows for the joint modelling of multiple experiments, by incorporating key aspects of the experimental design. In particular, the model uses the information about replicates and about the different antibodies used in the experiments. An extensive simulation study shows a lower false non-discovery rate for the proposed method, compared to existing methods, at the same false discovery rate. Finally, we present an analysis on real data for the detection of histone modifications of two chromatin modifiers from eight ChIP-seq experiments, including technical replicates with different IP efficiencies.

1 Introduction

ChIP-sequencing, also known as ChIP-seq, is a well-known biological technique to detect protein-DNA interactions, DNA methylation, and histone modifications in vivo. ChIP-seq combines Chromatin ImmunoPrecipitation (ChIP) with massively parallel DNA sequencing to identify all DNA binding sites of a transcription factor or genomic regions with certain histone modification marks. The final data produced by the experiment provide the number of DNA fragments in the sample aligned to each location of the genome. From this, the aim of the statistical analysis is to distinguish the truly enriched regions along the genome from the background noise. Whereas conventional transcription factors, that bind directly to the DNA, show sharp peaks at the regions of enrichment, chromatin modifiers tend to have much broader regions of enrichment and do not follow a peak-like pattern. The latter cannot

be captured by standard peak-calling algorithms and require more sophisticated statistical models. This is the focus of the present paper.

As regions of the genome are either bound by the protein in question or not, it is quite natural to analyse such data using a mixture model. Here the observed counts are assumed to come either from a signal or from a background distribution. A number of methods have adopted this approach, with some differences in the distribution chosen for the mixture. Spyrou et al. (2009), in their BayesPeak R package, adopt a Negative Binomial (NB) mixture model, with a NB distribution used both for signal and background. Kuan et al. (2011), in their MOSAiCS package, adopt a more flexible NB mixture model, where an offset is included in the signal distribution and this distribution itself is taken as a mixture of NBs. Spyrou et al. (2009) show evidence that a NB mixture model outperforms a Poisson mixture model, such as the one used by iSeq (Mo, 2012). Qin et al. (2010), in their HPeak implementation, suggest to use a zero-inflated Poisson model for the background and a generalized Poisson distribution for the signal. Zero-inflated distributions have been used successfully also for modelling other types of sequencing data (e.g. Dhavala et al., 2010; Van De Wiel et al., 2013). In this paper we consider the more flexible framework by allowing a zero-inflated Poisson or NB distribution for the background and a Poisson or NB distribution for the signal component.

Another feature of ChIP-seq data on histone modifications is the spatial dependency of counts for neighbouring windows along the genome. This is mainly the result of a common pre-processing step, whereby the genome is divided into bins of some ad-hoc length. It is quite common to consider fixed-width windows, although dynamic approaches have also been considered (Mo, 2012). The sum of counts within each window is subsequently considered for the analysis. As a result of this, true regions of the genome that are bound by the protein in question could be easily found to cross two or more pre-processed bins. This issue has been addressed in the literature with the use of Hidden Markov Models (HMMs) (Spyrou et al., 2009; Mo, 2012; Qin et al., 2010).

With few exceptions, the methods developed so far are limited to the analysis of single experiments, with the optional addition of a control experiment. When technical replicates or biological replicates are available, the standard procedure is to perform the peak calling on each individual data set and then combine the results by retaining the common regions. This process has inherent statistical problems, as pointed out by Bardet et al. (2012) and Bao et al. (2013). Despite the recognition of the need for biological replicates for ChIP-seq analyses (Tuteja et al., 2009) and despite the fact that several normalization methods have been proposed for multiple ChIP-seq experiments (Bardet et al., 2012; Shao et al., 2012), very few methods have been developed that combine technical and biological replicates at the modelling stage, so that the variability in the data can be properly accounted for. Zeng et al. (2013) extend MOSAiCS (Kuan et al., 2011), by developing a mixture model for multiple ChIP-seq datasets: individual models are used to analyse counts for each experiment and a final model is considered to govern the relationship of enrichment among different samples. Bao et al. (2013) build mixture models for multiple experiments, where replicates are modelled jointly by an assumption of a shared latent binding profile. They show how such a joint modelling approach leads to a more accurate detection of enriched and differentially

bound regions and how it allows to account for the different IP efficiencies of individual experiments. The latter has probably been the main reason why joint modelling approaches of ChIP-seq data have rarely been considered so far.

In this paper, we combine all the aspects described above into a single model, by proposing a one-dimensional Markov random field model for the analysis of multiple ChIP-seq data. Our model can be viewed as a hidden Markov model where the initial distribution is a stationary distribution. As such, we follow the existing literature on the use of hidden Markov models for ChIP-seq data in order to account for the spatial dependencies in the data (Spyrou et al., 2009; Mo, 2012). In contrast to the existing HMM-based methods, we propose a joint statistical model for ChIP-seq data, under general experimental designs. In particular, we discuss the case of technical and biological replicates as well as the case of different antibodies and/or IP efficiencies associated to each experiment. The remainder of this paper is organized as follows. In Section 2, we describe the Markov random field model and its Bayesian Markov Chain Monte Carlo (MCMC) implementation. In Section 3, we perform a simulation study to compare our method with two existing HMM-based methods, BayesPeak and iSeq, as well as with the joint mixture model of Bao et al. (2013) and that implemented in jMOSAiCS (Zeng et al., 2013). A real data analysis on eight experiments for the detection of histone modifications of two proteins, CBP and p300, is given in Section 4, where we also compare our results with two widely used methods for ChIP-seq analyses, MACS (Zhang et al., 2008) and CisGenome (Ji et al., 2008). In Section 5, we conclude with a brief discussion.

2 Methods

2.1 A joint latent mixture model and its limitations

The data generated by ChIP-sequencing experiments report the number of aligned DNA fragments in the sample for each position along the genome. Due to noise and the size of the genome, it is common to summarise the raw counts by dividing the genome into consecutive windows, or bins. Since the majority of the genome is expected not to be enriched, we would generally expect that some bins are enriched regions, with a lot of tags, and most other bins are not enriched, containing only few tags. This scenario is well suited to a mixture model framework.

Let M be the total number of bins and Y_{mcar} the counts in the mth bin, $m = 1, 2, \cdots, M$, under condition c, antibody a and replicate r. In the ChIP-seq context, the condition c stands for a particular protein and/or a particular time point, and $r = 1, \ldots, R_{ca}$ is the number of replicates for antibody a under condition c, with $a = 1, \ldots, A$. It is well known how a different level of ChIP efficiency is associated to different antibodies and how different IP efficiencies have been observed also for technical replicates (Bao et al., 2013). The current setup allows to account for these effects in the joint statistical modelling of multiple ChIP-seq experiments, under a variety of common experimental designs. The counts Y_{mcar} are either from a background population (non-enriched region) or a from a signal population (enriched region). Let X_{mc} be the unobserved random variable specifying if the mth bin is enriched ($X_{mc} = 1$) or non-enriched ($X_{mc} = 0$) under condition c. Clearly, this latent state does not

depend on ChIP efficiencies. As in Bao et al. (2013), we define a joint mixture model for Y_{mcar} as follows:

$$Y_{mcar} \sim p_c f(y|\theta^S_{car}) + (1-p_c) f(y|\theta^B_{car}),$$

where $p_c = P(X_{mc} = 1)$ is the mixture proportion of the signal component and $f(y, \theta^S_{car})$ and $f(y, \theta^B_{car})$ are the signal and background densities for condition c, antibody a and replicate r, respectively. Using this model, the regions are detected as enriched or not by controlling the False Discovery Rate (FDR).

Since we divide the genome arbitrarily in fixed-size windows, it is possible that a region in a certain chromatin state crosses two or more bins. As a consequence of this, it is reasonable to expect spatial dependencies in the data. Figure 1 (left) plots the bin counts Y_m for 200bp fixed windows in a region of the genome, for one ChIP-seq experiment. On the right, the plot shows the posterior probability of enrichment, using the latent (non-Markovian) mixture model described above. This plot clearly shows regions of consecutive enriched bins.
FIGURE 1 ABOUT HERE.

In this paper, we propose a natural extension of the mixture model in (1) by allowing first-order Markov dependencies. This is described in the next section.

2.2 A one-dimensional Markov random field model

The number of reads Y_{mcar} in bin m, under condition c, antibody a and replicate r, is either drawn from a signal or a background distribution. The first issue is the choice of the mixture distribution. Together with the general expectation that a large part of the genome is not bound by the protein in question, unmapped genome regions and insufficient sequencing depth, i.e. an insufficient total number of reads, give rise to an excess of zeroes in the observed data. This forms part of the background noise and gives us the motivation to use a zero-inflated distribution to model the background. As the data is in forms of counts, it is natural to consider either a Zero-Inflated Poisson (ZIP) or a Zero-Inflated Negative Binomial (ZINB) distribution. That is, conditional on the latent state X_{mc},

$$\begin{aligned} Y_{mcar}|X_{mc} = 0 &\sim \text{ZIP}(\pi_{car}, \lambda_{0car}) \text{ or } \text{ZINB}(\pi_{car}, \mu_{0car}, \phi_{0car}), \\ Y_{mcar}|X_{mc} = 1 &\sim \text{Poisson}(\lambda_{1car}) \text{ or } \text{NB}(\mu_{1car}, \phi_{1car}), \end{aligned}$$

where the probability density function of the zero-inflated Negative Binomial is given by:

$$ZINB(y|\pi, \mu, \phi) = \begin{cases} (1-\pi) + \pi\Big(\dfrac{\phi}{\mu+\phi}\Big)^{\phi} & \text{if } y = 0, \\ \dfrac{\Gamma(y+\phi)}{\Gamma(\phi)\Gamma(y+1)}\Big(\dfrac{\mu}{\mu+\phi}\Big)^{y}\Big(\dfrac{\phi}{\mu+\phi}\Big)^{\phi} & \text{if } y > 0, \end{cases} \quad (1)$$

and similarly for a zero-inflated Poisson (Qin et al., 2010).

A zero-inflated model can be seen as a mixture model of a Poisson/NB distribution and a zero mass distribution, which represents the extra zeroes in the background regions that

a standard Poisson/NB distribution cannot account for. If we introduce an inner latent variable Z_{mcar} and $P(Z_{mcar}=1|X_{mc}=0)=\pi_{car}$, then conditional on X_{mc}, we have

$Y_{mcar}|X_{mc}=0, Z_{mcar}=0 \sim 1(y=0)$, $Y_{mcar}|X_{mc}=0, Z_{mcar}=1 \sim \text{Poisson}(\lambda_{0car})$ or $\text{NB}(\mu_{0car}, \phi_{0car})$,
$Y_{mcar}|X_{mc}=1 \sim \text{Poisson}(\lambda_{1car})$ or $\text{NB}(\mu_{1car}, \phi_{1car})$.

Note that the parameters of the background and signal components vary for each replicate, in order to account for the different IP efficiencies of individual experiments.

The latent variable, X_{mc}, representing the binding profile under condition c, is assumed to satisfy one dimensional Markov properties, that is,

$$P(X_{mc}=i|X_{-mc}) = P(X_{mc}=i|X_{m-1,c}, X_{m+1,c}), i \in \{0,1\}, \tag{2}$$

where $X_{-mc} = \{X_{1c}, \ldots, X_{m-1,c}, X_{m+1,c}, \ldots, X_{Mc}\}$. By imposing a first-order Markov assumption on the latent variable, the model class becomes richer, since a nearest neighbour latent Markov model can induce long-range conditional dependencies on the observed data. The Markov assumption leads to the classical factorization of the joint density

$$P(X_{1c}, \ldots, X_{Mc}) \;=\; f_{0c}(X_{1c}) \prod_{m=1}^{M-1} q_{X_{mc}, X_{m+1,c}} \tag{3}$$

in terms of the initial state distribution f_{0c} and transition probabilities $q_{i,j,c} = P(X_{m+1,c} = j|X_{mc}=i), i,j \in \{0,1\}$. Unlike the model above, in this paper we use a more natural representation of the joint density of the latent states for a one-dimensional Markov random field model, namely:

$$P(X_{1c}, \ldots, X_{Mc}) \;=\; \frac{\prod_{m=1}^{M-1} P(X_{mc}, X_{m+1,c})}{\prod_{m=2}^{M-1} P(X_{mc})} \tag{4}$$

where $P(X_{mc}, X_{m+1,c})$ is the joint probability of X_{mc} and $X_{m+1,c}$ and $P(X_{mc})$ is the marginal probability of X_{mc}. In particular, we have $P(X_{mc}) = \sum_{x_{m+1,c}} P(X_{mc}; X_{m+1,c} = x_{m+1,c})$.

When the X_{mc} are binary variables, as in our case, we can further re-write the model (4) as

$$P(X_{1c}, \ldots, X_{Mc}) = {\delta_1}^{I(X_{1c}=1)} {\delta_0}^{I(X_{1c}=0)} \Big(\frac{\delta_{1,1,c}}{\delta_{1c}}\Big)^{n_{1,1,c}} \Big(\frac{\delta_{1,0,c}}{\delta_{1c}}\Big)^{n_{1,0,c}} \Big(\frac{\delta_{0,1,c}}{\delta_{0c}}\Big)^{n_{0,1,c}} \Big(\frac{\delta_{0,0,c}}{\delta_{0c}}\Big)^{n_{0,0,c}},$$

where

$n_{i,j,c} = \#\{X_{mc}=i, X_{m+1,c}=j\}$, $\delta_{i,j,c} = P(X_{mc}=i, X_{m+1,c}=j), i,j \in \{0,1\}, m = 1, \ldots, M-1$,
$\delta_{1c} = P(X_{mc}=1) = \delta_{1,1,c} + \delta_{1,0,c}$, $\delta_{0c} = P(X_{mc}=0) = 1 - \delta_{1c}$, $\delta_{0,1,c} = \delta_{1,0,c} = \dfrac{1-\delta_{1,1,c}-\delta_{0,0,c}}{2}$.

One can show that this model satisfies (2), that is the model is a one-dimensional Markov random field model. And if we notice that the transition probabilities satisfy $q_{i,j,c} = \delta_{i,j,c}/\delta_{ic}$, the model can be further written in terms of the transition probabilities $q_{i,j,c}$ as follows

$$P(X_{1c}, \ldots, X_{Mc}) = \Big(\frac{q_{0,1,c}}{q_{0,1,c} + q_{1,0,c}}\Big)^{I(X_{1c}=1)} \Big(\frac{q_{1,0,c}}{q_{0,1,c} + q_{1,0,c}}\Big)^{I(X_{1c}=0)} q_{1,1,c}^{n_{1,1,c}} q_{1,0,c}^{n_{1,0,c}} q_{0,1,c}^{n_{0,1,c}} q_{0,0,c}^{n_{0,0,c}}. \quad (5)$$

The most attractive property of this model is that the initial state distribution under (4) is the stationary distribution. This is different from BayesPeak (Spyrou et al., 2009), where an equal mass probability for the states is taken as the initial state distribution. Note also that the Ising model of Mo (2012) has one parameter less than the model presented here: this corresponds to assuming that $q_{1,1,c} + q_{0,1,c} = 1$, which is an unnecessary assumption and it is not normally satisfied by the data. More details about the comparison between the presented Markov random field model, a classical hidden Markov model and an Ising model are provided in the supplementary material.

2.3 Parameter Estimation

For simplicity, in this section we consider one condition and assume that the same antibody is used for all replicates under this condition, which is often the case in practice. A similar derivation applies to the more general case. Furthermore, we define $\tilde{q}_1 = q_{1,1}$ and $\tilde{q}_0 = q_{0,1}$ for the probabilities that the current state of a bin is 1 (enriched) given that the state of the left bin is 1 and 0, respectively. We denote with R the number of replicates under the current condition. Assuming a ZINB-NB mixture model (zero-inflated NB for the background and NB for the signal), we aim to estimate the parameters $\Theta = (\tilde{q}_0, \tilde{q}_1, \pi_1, \cdots, \pi_R, \mu_{01}, \cdots, \mu_{0R}, \phi_{01}, \cdots, \phi_{0R}, \mu_{11}, \cdots, \mu_{1R}, \phi_{11}, \cdots, \phi_{1R})$. The joint likelihood of this model given the latent states, $\mathbf{X}$, the inner variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_R$ and data $\mathbf{Y}_1, \ldots, \mathbf{Y}_R$, is given by

$$\begin{aligned}
P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}|\Theta) &= P(\mathbf{X}|\Theta) P(\mathbf{Z}|\mathbf{X}=0, \Theta) P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \Theta) \\
&\propto \Big(\frac{\tilde{q}_0}{\tilde{q}_0 + 1 - \tilde{q}_1}\Big)^{I(X_1=1)} \Big(\frac{1-\tilde{q}_1}{\tilde{q}_0 + 1 - \tilde{q}_1}\Big)^{I(X_1=0)} \tilde{q}_1^{n_{1,1}} (1-\tilde{q}_1)^{n_{1,0}} \tilde{q}_0^{n_{0,1}} (1-\tilde{q}_0)^{n_{0,0}} \\
&\times \prod_{r=1}^{R} \pi_r^{\Sigma_m I(X_m=0, Z_{mr}=1)} \times (1-\pi_r)^{\Sigma_m I(X_m=0, Z_{mr}=0)} \quad (6) \\
&\times \prod_{r=1}^{R} \prod_{m=1}^{M} \Big[\frac{\Gamma(y_{mr} + \phi_{0r})}{\Gamma(\phi_{0r})\Gamma(y_{mr}+1)} \Big(\frac{\mu_{0r}}{\mu_{0r} + \phi_{0r}}\Big)^{y_{mr}} \Big(\frac{\phi_{0r}}{\mu_{0r} + \phi_{0r}}\Big)^{\phi_{0r}}\Big]^{I[X_m=0, Z_{mr}=1]} \\
&\times \prod_{r=1}^{R} \prod_{m=1}^{M} \Big[\frac{\Gamma(y_{mr} + \phi_{1r})}{\Gamma(\phi_{1r})\Gamma(y_{mr}+1)} \Big(\frac{\mu_{1r}}{\mu_{1r} + \phi_{1r}}\Big)^{y_{mr}} \Big(\frac{\phi_{1r}}{\mu_{1r} + \phi_{1r}}\Big)^{\phi_{1r}}\Big]^{I[X_m=1]}.
\end{aligned}$$

Here we assume that technical and biological replicates share the same binding profiles, i.e. that the latent states X are common between replicates. This results in the joint probabilities $P(X_m, X_{m+1})$ in equation (4) being equal for all replicates, and consequently, the transition probabilities $\tilde{q}_0$ and $\tilde{q}_1$ in equation (5) are also equal across replicates. A similar derivation

applies for a ZIP-Poisson mixture model for the estimation of the parameters $\Theta = (\tilde{q}_0, \tilde{q}_1, \pi_1, \ldots, \pi_R, \lambda_{01}, \ldots, \lambda_{0R}, \lambda_{11}, \ldots, \lambda_{1R})$.

We use a Bayesian framework, together with a Metropolis-within-Gibbs procedure, to estimate the model parameters and states. In particular, using a direct Gibbs method, we draw each X_m, for $m = 1, \ldots, M$, from its full conditional distribution

$$P(X_m = i | X_{-m}, \mathbf{Y}_1, \ldots, \mathbf{Y}_R, \Theta) \propto q_{X_{m-1},i} q_{i,X_{m+1}} \prod_{r=1}^{R} P_i(Y_{mr}|\Theta)$$

where $P_i(Y_{mr}|\Theta) = P(Y_{mr}|X_m = i, \Theta)$ and the normalising constant is the sum over all possible values of i. Given $X_m = 0$, the inner latent variable Z_{mr} is drawn from its full conditional distribution

$$P(Z_{mr} = i | X_m = 0, Y_{mr} = y_{mr}, \Theta) \propto P(y_{mr}|X_m = 0, Z_{mr} = i, \Theta) P(Z_{mr} = i | X_m = 0).$$

More details about the prior and posterior distributions are given in the supplementary material.

2.4 Assuming the same number of binding sites across conditions

The method above can be used in the presence of technical and biological replicates. Whereas technical and biological replicates share the same binding profile $\mathbf{X}$, different proteins will generally have a different binding profile. Under certain situations, e.g. when comparing the binding profiles across two conditions or between highly similar transcription factors, we can assume that the total number of binding sites is the same. Bao et al. (2013) show how this assumption can be included in a mixture modelling framework. In this paper, we show how the same assumption can be included also in the proposed Markov random field mixture model.

In particular, if $\mathbf{X}_1$ and $\mathbf{X}_2$ are the binding profiles of conditions 1 and 2, respectively (e.g. protein 1 and protein 2), we can include the a priori biological knowledge in the joint model that the two conditions have the same number of binding sites, i.e. $P(X_{m1} = 1) = P(X_{m2} = 1)$ for any region m. This constraint is satisfied under an assumption of equal transition probabilities for the two conditions. However, this is quite a strong assumption and it is difficult to know this beforehand, unless we are in the presence of technical replicates. If we note that the stationary distribution $P(X = 1) = \dfrac{\tilde{q}_0}{\tilde{q}_0 + 1 - \tilde{q}_1} = \dfrac{1}{1 + (1 - \tilde{q}_1)/\tilde{q}_0}$, we can see that if $\dfrac{1 - \tilde{q}_{11}}{\tilde{q}_{01}} = \dfrac{1 - \tilde{q}_{12}}{\tilde{q}_{02}}$ then $P(X_1 = 1) = P(X_2 = 1)$. Here $\tilde{q}_{01}$, $\tilde{q}_{11}$ and $\tilde{q}_{02}$, $\tilde{q}_{12}$ denote the transition probabilities corresponding to the binding profiles $\mathbf{X}_1$ and $\mathbf{X}_2$ of the two conditions, respectively. This shows that a weaker condition on the transition probabilities is necessary to achieve equal probabilities of enrichment.

In general, let $s = \dfrac{1 - \tilde{q}_{1c}}{\tilde{q}_{0c}}$ for protein c and assume that s is common for all proteins c, with $c = 1, \ldots, C$, that is the different proteins have the same proportion of binding sites. If R_c is the number of replicates for protein c, then the joint likelihood given the latent states

$\mathbf{X}_1, \ldots, \mathbf{X}_C, \mathbf{Z}_{11}, \ldots, \mathbf{Z}_{1R_1}, \ldots, \mathbf{Z}_{C1}, \ldots, \mathbf{Z}_{CR_C}$, and the data $\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1R_1}, \ldots, \mathbf{Y}_{C1}, \ldots, \mathbf{Y}_{CR_C}$, is given by:

$$\begin{aligned}
P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}|\Theta) &= \prod_{c=1}^{C} \Big(\frac{1}{1+s}\Big)^{I(X_{1c}=1)} \Big(1-\frac{1}{1+s}\Big)^{I(X_{1c}=0)} (1-s\tilde{q}_{0c})^{n^c_{1,1}} (s\tilde{q}_{0c})^{n^c_{1,0}} \tilde{q}_{0c}^{n^c_{0,1}} (1-\tilde{q}_{0c})^{n^c_{0,0}} \\
&\times \prod_{r=1}^{R_c} \pi_{cr}^{\Sigma_m I(X_{mc}=0, Z_{mcr}=1)} \times (1-\pi_{cr})^{\Sigma_m I(X_{mc}=0, Z_{mcr}=0)} \\
&\times \prod_{r=1}^{R_c} \prod_{m=1}^{M} \Big[\frac{\Gamma(y_{mcr}+\phi_{0cr})}{\Gamma(\phi_{0cr})\Gamma(y_{mcr}+1)} \Big(\frac{\mu_{0cr}}{\mu_{0cr}+\phi_{0cr}}\Big)^{y_{mcr}} \Big(\frac{\phi_{0cr}}{\mu_{0cr}+\phi_{0cr}}\Big)^{\phi_{0cr}}\Big]^{I[X_{mc}=0, Z_{mcr}=1]} \\
&\times \prod_{r=1}^{R_c} \prod_{m=1}^{M} \Big[\frac{\Gamma(y_{mcr}+\phi_{1cr})}{\Gamma(\phi_{1cr})\Gamma(y_{mcr}+1)} \Big(\frac{\mu_{1cr}}{\mu_{1cr}+\phi_{1cr}}\Big)^{y_{mcr}} \Big(\frac{\phi_{1cr}}{\mu_{1cr}+\phi_{1cr}}\Big)^{\phi_{1cr}}\Big]^{I[X_{mc}=1]}.
\end{aligned}$$

This is used in a Metropolis-within-Gibbs procedure similar to the one described in the previous section (more details in the supplementary material).

2.5 Identification of enriched regions and differentially bound regions

In this section, we show how the statistical model described above is used to detect the regions in the genome that are bound by a protein of interest. Let $\mathbf{X_c^{(1)}}, \ldots, \mathbf{X_c^{(N)}}$ be N Gibbs draws of the joint states $\mathbf{X_c}$ under condition c, where $\mathbf{X_c^{(k)}} = (X_{1c}^{(k)}, \ldots, X_{Mc}^{(k)})$. Under the proposed random field model, a natural estimate of the posterior probability that the mth bin is enriched is given by $\hat{P}(X_{mc}=1|\mathbf{Y}) = \sum_{k=1}^{N} I(X_{mc}^{(k)}=1)$ (Scott, 2002). To decide whether a bin is enriched or not, we set a threshold on these probabilities based on the false discovery rate. If D is set of declared enriched regions corresponding to a particular cut-off on the posterior probabilities, then the estimated false discovery rate for this cut-off is given by $\widehat{FDR} = \dfrac{\sum_{m\in D} \hat{P}(X_{mc}=0|\mathbf{Y})}{|D|}$.

When data are available for more than one protein, the interest is also on finding the regions that are bound only by one of the proteins. Following Bao et al. (2013), we define a probability of differential binding by

$$P(X_{m1} \neq X_{m2}|\mathbf{Y}) = P(X_{m1}=0|\mathbf{Y}_1)P(X_{m2}=1|\mathbf{Y}_2) + P(X_{m1}=1|\mathbf{Y}_1)P(X_{m2}=0|\mathbf{Y}_2)$$

where $P(X_{mc}=0|\mathbf{Y}_c) = P(X_{mc}=0|\mathbf{Y}_{c1}, \ldots, \mathbf{Y}_{cR_c})$ is the posterior probability that the mth bin is enriched for protein c, estimated by joint model from all the data on protein c. In this way, all technical replicates under the same condition are considered in the estimation of the posterior probabilities, returning a more robust set of differentially bound regions.

3 Simulation study

In this section, we perform an extensive simulation study where we compare our Markov Random Field model (MRF) with four competitive methods: iSeq (Mo 2012), BayesPeak (Spyrou et al. 2009) and the mixture model approaches of Bao et al. (2013) (here denoted as Mixture) and Zeng et al. (2013) (jMOSAiCS). For a number of different scenarios, we generate the data for $M = 10000$ regions and we repeat the simulation for 100 times. For all methods, we identify the enriched regions by controlling the False Discovery Rate (FDR) at 0.05. We then compute the False Non-discovery Rate (FNDR), that is the fraction of all the non-discovered regions that were actually enriched. Finally, we report the p-values of a t-test for the null hypothesis that the FNDR of our method is greater or equal than the FNDR of each of the other methods.

In the first four scenarios, we compare our method with the other HMM-based methods, namely iSeq (Mo 2012) and BayesPeak (Spyrou et al. 2009). In the first scenario, we simulate data from a mixture model with a ZINB background distribution and a NB signal distribution. We set the parameters of these distributions using the values estimated by a MRF model on two of our real ChIP-seq datasets. We choose the experiments on the basis of their ChIP efficiency. In particular, we consider the case of a not very efficient experiment (CBPT0) and the one of a more efficient experiment (p300T302). In terms of the mixture distribution, the more efficient experiment corresponds to background and signal distributions that are better separated. Since neither iSeq nor BayesPeak can deal with multiple experiments, we perform these comparisons on single experiments. The results are given in Table 1 in scenario 1. BayesPeak is in general inferior to both iSeq and MRF. Between iSeq and MRF, there is no significant difference for the less efficient experiment, whereas MRF is superior to iSeq for the more efficient experiment. In general, we find that the use of zero-inflated distributions is particularly suited to the case of efficient experiments, where there is combination of a large number of zeroes and a relatively large number of high counts. A mixture of Poisson distributions, which is implemented in iSeq, cannot capture this situation very well. We further extend this simulation to scenarios where some assumptions are shared between MRF and iSeq. Firstly, we generate data from a mixture of Poisson distributions and $\tilde{q}_1 + \tilde{q}_0 = 1$. These are the two main assumptions imposed by the Ising model implemented in iSeq. The results are given in Table 1 in scenario 2. In this case, as expected, there is no difference between iSeq and MRF, whereas BayesPeak is still inferior to both. Secondly, we consider the case of a Poisson mixture, as in iSeq, but we relax the assumption of $\tilde{q}_1 + \tilde{q}_0 = 1$ (Table 1, scenario 3). In both cases, the MRF method is superior to iSeq, although the difference is not so large. Finally, in the fourth scenario (Table 1, scenario 4), we generated data which satisfies the constraint $\tilde{q}_1 + \tilde{q}_0 = 1$, but which does not follow a Poisson mixture distribution. In particular, we use a ZINB-NB mixture distribution. This is the case where the MRF method performs much better than either of the two other methods. In general, the first four scenarios in Table 1 show how iSeq and MRF perform equally well when the data is generated from a Poisson mixture distribution and $\tilde{q}_1 + \tilde{q}_0 = 1$, but MRF is superior to both iSeq and BayesPeak when either of these two conditions is not satisfied. TABLE 1 ABOUT HERE.

In the last two scenarios, we compare the MRF model with our previously developed

mixture model for multiple experiments (Bao et al., 2013) and the recently developed jMOSAiCS (Zeng et al., 2013), which also does not account for spatial dependencies. As the jMOSAiCS implementation requires a control sample, we use a sample of 10000 bins from the IgGrabbit control sample provided by Wang et al. (2009). For a fairer comparison, we extend the model of Bao et al. (2013) to include zero-inflated distributions for the background. In Table 1 in scenario 5, the data are generated from a MRF model, using the parameter values estimated from two real datasets. As expected, the MRF model performs better than the mixture models in this case, as it accounts for the Markov dependencies. In the final scenario (Table 1, scenario 6), we generated data without Markov properties, that is, the latent state X follows a Bernoulli distribution. In this case, the MRF and our mixture model give the same results, but the MRF model is still superior to jMOSAiCS.

From all scenarios considered, one can conclude that the proposed MRF model performs as well as the other methods under similar conditions, but it outperforms the other models under more general mixture distributions and modelling assumptions.

4 Real data analysis

In this section, we use the new model on real ChIP-seq data on two proteins, p300 and CBP (CREB-binding protein). P300 and CBP are two histone acetyltransferaces: they are transcriptional co-activators whose regulatory mechanisms are not fully understood, but are thought to be crucial for a number of biological functions. As chromatin modifiers, they do not bind directly to the DNA and thus they generally show broad binding profiles. We analyze ChIP-seq data from six experiments, three for CBP and three for p300 (Ramos et al., 2010). For each protein, one experiment is conducted at time point 0 (T0) and two technical replicates are performed after 30 minutes (T301 and T302). We also use CBP and p300 ChIP-seq data from an earlier independent study (Wang et al., 2009), where CBP and p300 binding was evaluated in resting cells, using a different cell line and different antibodies. The data are further described in Bao et al. (2013), which includes a discussion of the effect of the different IP efficiencies on the data. This is the case also for the technical replicates, with one replicate having a higher IP efficiency than the other. We divide the whole genome into 200 base pair windows and summarise the raw counts for each window by the number of tags whose first position is in the window. The window length was chosen as it matches the fragment size used in the ChIP-seq experiment. Furthermore, we exclude from the analysis genomic regions that have been found to exhibit anomalous or unstructured read counts from the analysis (Hoffman et al., 2012).

First of all, we have compared the fit of a NB mixture model, where a NB distribution is chosen both for the background and signal, against a ZINB-NB mixture, where a zero-inflated NB is considered for the background and a NB distribution for the signal. In general, we find that the BIC values are lower for the ZINB-NB mixture than for the NB mixture, suggesting a better fit for the zero-inflated mixture (supplementary material). In the following, we will therefore use zero-inflated distributions, distinguishing it to iSeq (Mo, 2012) and BayesPeak (Spyrou et al., 2009), which use Poisson and NB mixtures, respectively.

Within the eight data sets that we analyzed, CBPT0, p300T0, WangCBP and Wangp300,

are single experiments, i.e. with no replicates. We therefore compare the proposed MRF model with iSeq and BayesPeak on these four experiments. We also include in the comparison two widely used methods for analysing ChIP-seq data, namely MACS (Zhang et al., 2008) and CisGenome (Ji et al., 2008), which do not account for spatial dependencies. For simplicity we provide the results only for chromosome 21. Figure 2 shows the Venn diagrams of the detected regions for two representative experiments (CBPT0 and WangCBP). As MACS does not provide FDR control, we use the suggested default p-value cut-off of 10^{-5}. For the other methods, we use a 5% FDR cutoff, although we found the empirical FDR values returned by CisGenome to be rather unreliable. The results show that MRF detects more regions than any of the other four methods. Furthermore, MRF tends to agree more with iSeq and BayesPeak, than with the other two methods. This is probably due to the fact that these three methods all account for spatial dependencies. Finally, Figure 2 shows how the overlap between MRF and iSeq and the overlap between MRF and BayesPeak are both larger than the overlap between iSeq and BayesPeak. The results for the other datasets are provided in the supplementary material. FIGURE 2 ABOUT HERE.

CBP and p300 both have largely overlapping roles in transcriptional activation. We use `ChromHMM` (Ernst and Kellis, 2010) to explore whether the regions identified by MRF are likely functional in transcription activation and whether different chromatin features are enriched in the regions identified by the different methods. Figure 3 shows the results of `ChromHMM` using a 4-state hidden Markov model on the enrichment profile given by MRF, BayePeak and iSeq, each at a 5% FDR, for two representative experiments (CBPT0 and p300T0). The left plots give the emission probabilities for the different analyses, that is the probability of the observed enrichment given each of the four possible states. These plots show how, for all analyses, three of the four states explain most of the enrichment pattern in the identified lists. The right plots give the relative fold enrichment for several annotations. These plots show how these three states are represented by a similar enrichment of features for the three methods, mainly CpGisland and RefSeq Transcription Start Sites (RefSeqTSS), suggesting that the additional regions detected by MRF are likely to be genuine binding events. FIGURE 3 ABOUT HERE.

For replicated experiments (CBPT301, CBPT302, p300T301, p300T302), we compare the proposed MRF model for multiple experiments, with the previously developed mixture model (Bao et al., 2013) and the recently developed jMOSAiCS (Zeng et al., 2013). Table 2 reports the number of enriched regions at a 5% FDR on chromosome 21. For the MRF and Mixture methods, we also include the number of regions differentially bound between p300 and CBP. The results show that by including the assumption of Markov properties, the number of enriched regions detected is larger than when the Markov property is not considered. This is to be expected since regions with a relatively small number of counts but with neighbouring enriched regions may not be detected by the mixture model but would be detected by the MRF model. As before, the `ChromHMM` validation shows a similar enrichment pattern in the identified lists, with predominantly TSS and CpGIsland features (supplementary material). Overall, these results lead us to the conclusion that by taking into account Markov properties while combining replicates many more regions are found at the same FDR, and that these regions are of the same nature as those found by latent mixture models. TABLE 2 ABOUT HERE.

Similar analyses can be performed under the assumption that p300 and CBP have the same number of binding sites, using the method discussed in section 2.4. The results of these analyses are provided in the supplementary material.

5 Conclusion

In this paper, we have proposed a one-dimensional Markov random field model for the analysis of ChIP-seq data. Our model can be viewed as a hidden Markov model where the initial distribution is the stationary distribution. As such, we follow the literature on existing HMM-based models, such as BayesPeak (Spyrou et al., 2009) and iSeq (Mo, 2012). Similarly to these models, we capture the spatial dependencies of local bins by an assumption of first-order Markov dependence. Differently from these methods, we propose a joint model for multiple ChIP-seq experiments under general experimental designs, such as replicated experiments, experiments using different antibodies and two-sample analyses. Furthermore, similarly to a previously developed mixture model (Bao et al., 2013), we show how a priori knowledge of the same number of binding sites for different proteins can also be added to the model, in order to better account for the different ChIP efficiencies of individual experiments. Finally, we advocate the use of zero-inflated background distributions, as these better account for the large number of zeroes in the data.

In an extensive simulation study, we have shown that the proposed Markov random field model performs at least as well as competitive existing methods under similar conditions, but that it outperforms the other models under more general mixture distributions and modelling assumptions. Finally, we present an analysis on real data for the detection of histone modifications of two transcriptional activators from eight ChIP-sequencing experiments, including technical replicates with different IP efficiencies. Future work will extend the current methodology to account for sequencing biases and for the possibility of more than two mixture components.

6 Software

The method is available from CRAN in the R package `enRich`. The current parallel implementation of the ZINB-NB model takes about 46mins on one chromosome with 10000 MCMC iterations, using a 64-bit machine with two 2.39GHz processors. At the moment, the running time is higher than that of the competitive methods, mainly due to the use of the Metropolis-within-Gibbs sampling procedure.

7 Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

This work was supported by the BBSRC [BB/H017275/1 to Y.B.]; the European Commission 7th Framework Program GEUVADIS [project nr. 261123 to P.'t H.]; and the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. *Conflict of Interest*: None declared.

References

Bao, Y., V. Vinciotti, E. Wit, and P. 't Hoen (2013). Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC Bioinformatics 14*, 169.

Bardet, A., Q. He, J. Zeitlinger, and A. Stark (2012). A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols 7*(1), 45–61.

Dhavala, S., S. Datta, B. Mallick, R. Carroll, S. Khare, S. Lawhon, and L. Adams (2010). Bayesian modeling of MPSS data: Gene expression analysis of bovine salmonella infection. *Journal of the American Statistical Association 105*(491), 956–967.

Ernst, J. and M. Kellis (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology 28*(8), 817–825.

Hoffman, M., J. Ernst, S. Wilder, A. Kundaje, R. Harris, M. Libbrecht, B. Giardine, P. Ellenbogen, J. Bilmes, E. Birney, R. Hardison, I. Dunham, M. Kellis, and W. Noble (2012). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research 41*(2), 827–841.

Ji, H., H. Jiang, W. Ma, D. Johnson, R. Myers, and W. Wong (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology 26*(11), 1293–1300.

Kuan, P., D. Chung, G. Pan, J. Thomson, R. Stewart, and S. Keles (2011). A statistical framework for the analysis of ChIP-Seq data. *Journal of the American Statistical Association 106*(495), 891–903.

Mo, Q. (2012). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics 13*(1), 113–128.

Qin, Z., J. Yu, J. Shen, C. Maher, M. Hu, S. Kalyana-Sundaram, Y. J., and A. Chinnaiyan (2010). HPeak: an HMM-based algorithm for defining read-enriched regions in chip-seq data. *BMC Bioinformatics 11*(369).

Ramos, Y., M. Hestand, M. Verlaan, E. Krabbendam, Y. Ariyurek, H. van Dam, G. van Ommen, J. den Dunnen, A. Zantema, and P. 't Hoen (2010). Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. *Nucleic Acids Research 38*(16), 5396–5408.

Scott, S. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association 97*(457), 337–351.

Shao, Z., Y. Zhang, G. Yuan, S. Orkin, and D. Waxman (2012). MAnorm: a robust model for quantitative comparision of ChIP-Seq data sets. *Genome Biology 13*(3), R16.

Spyrou, C., R. Stark, A. Lynch, and S. Tavare (2009). BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics 10*(1), 299.

Tuteja, G., P. White, J. Schug, and K. Kaestner (2009). Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Research 37*(17), e113.

Van De Wiel, M., G. Leday, L. Pardo, H. Rue, A. Van Der Vaart, and W. Van Wieringen (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors.

Wang, Z., C. Zang, K. Cui, D. Schones, A. Barski, W. Peng, and K. Zhao (2009). Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell 138*, 1019–1031.

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

Zhang, Y., T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nussbaum, R. Myers, M. Brown, and W. Li (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology 201*(1), R137.

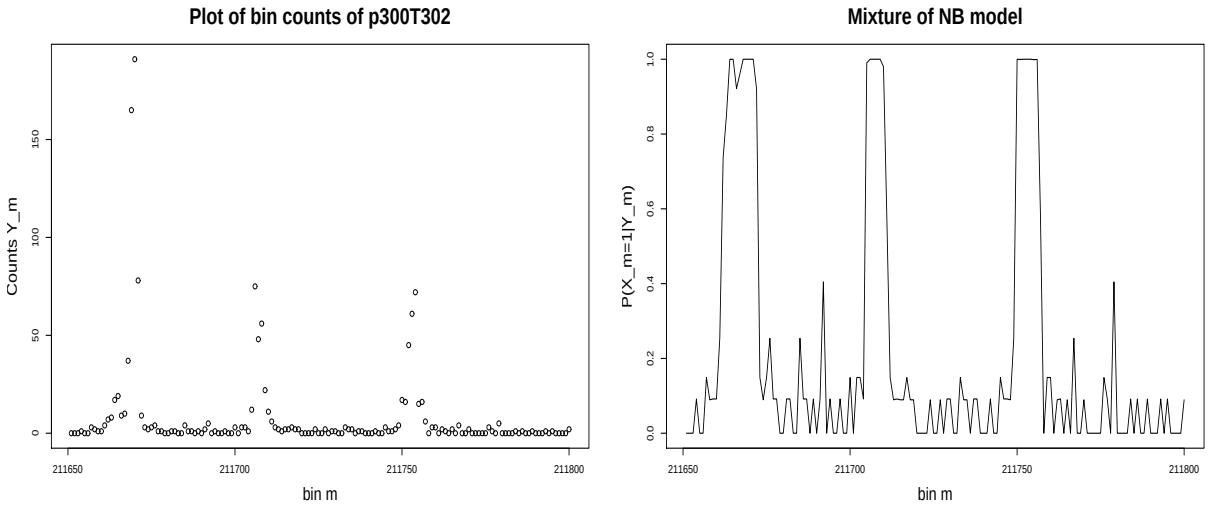

Figure 1: Plots of bin counts Y_m versus bin m (left) for bins of size 200bp on a region of chromosome 1 for the p300T302 experiment, and $P(X_m = 1|Y_m, \hat{\theta})$ versus bin m (right) under mixture of NB model.

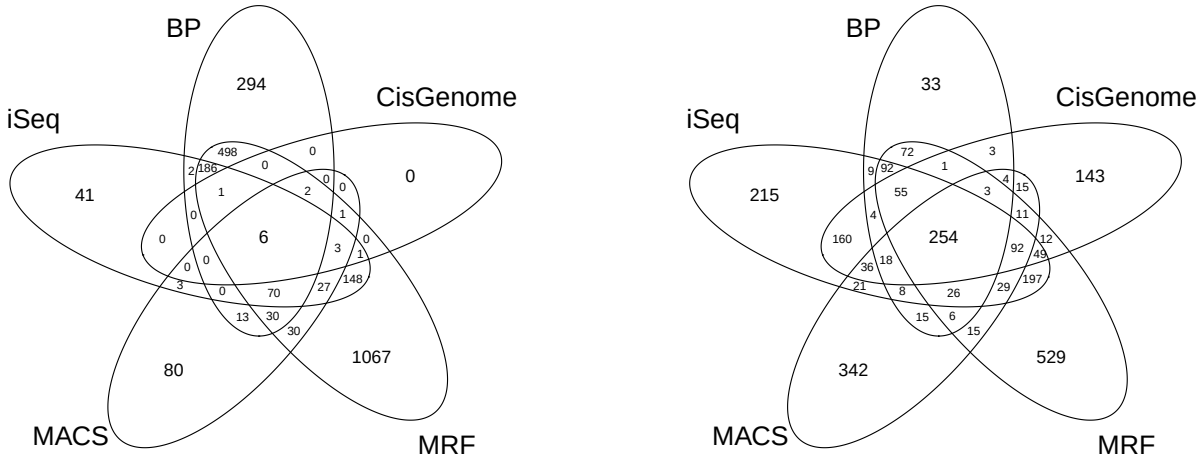

Figure 2: Venn diagrams of the number of enriched regions detected by MRF, iSeq, BayesPeak (BP), CisGenome and MACS, at the 5% FDR, for the CBPT0 and WangCBP experiments.

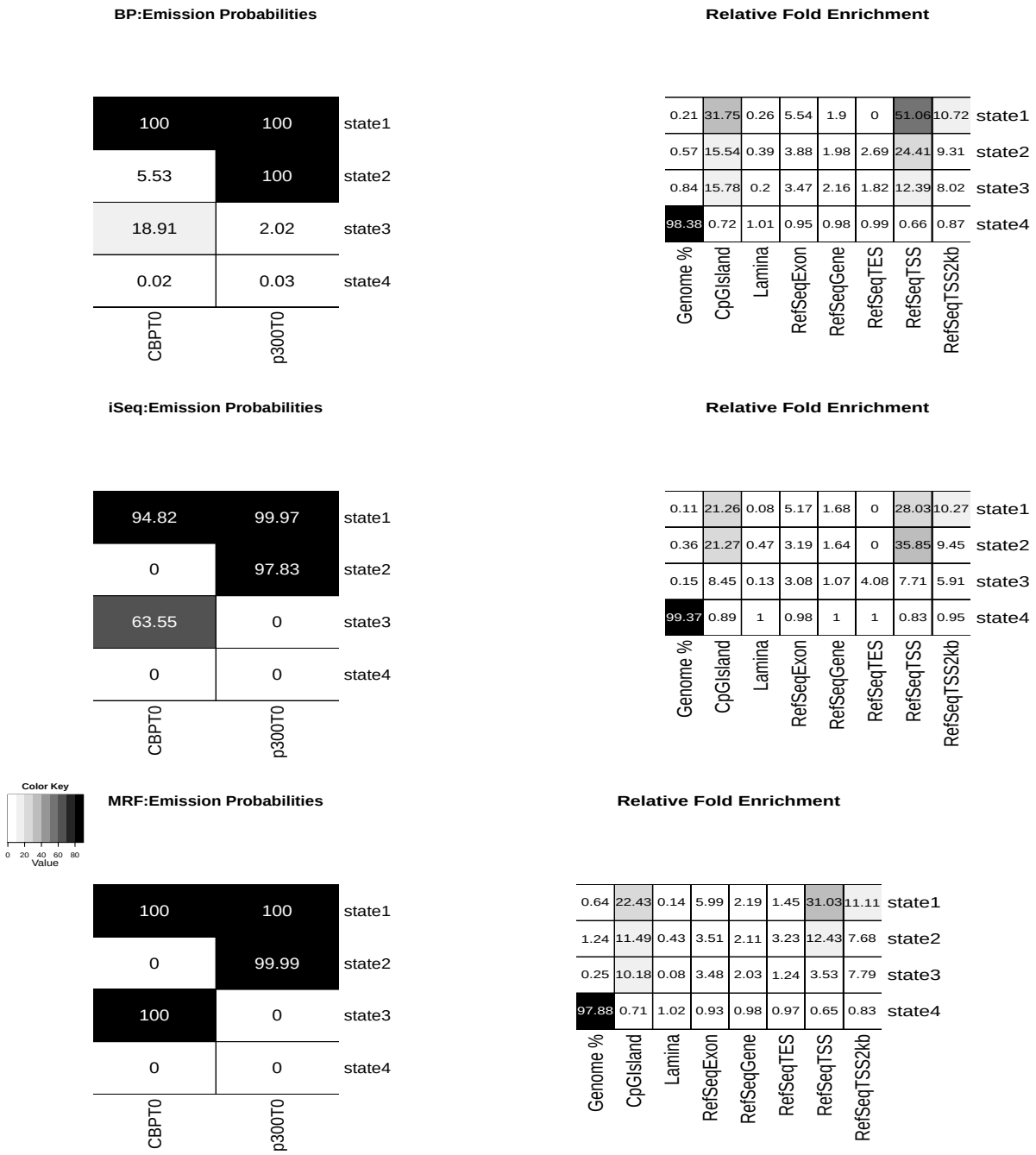

Figure 3: Validation of the enriched bins detected by BayesPeak (BP), iSeq and MRF for CBPT0, p300T0. We use a 4-state ChromHMM. The left plots show heatmaps of the probabilities (in %) that the detected bins are enriched given each identified chromatin-state. The right plots show the relative percentage of the genome represented by each chromatin state (column 1) and the relative fold enrichment for several types of annotation (columns 2-8).

Table 1: Simulated count data is generated for $M = 10000$ regions under different scenarios, with parameter values given in brackets for signal (S) and background (B) distributions. The table reports the average FNDR over 100 iterations, at a controlled FDR of 5%, for MRF, iSeq, BayesPeak, Mixture and jMOSAiCS. The p-values show whether the MRF model has a significantly lower FNDR than each of the other methods.

	Less Efficient Experiment		More Efficient Experiment	
	Scenario 1: ZINB-NB mixture with $\tilde{q}_1 + \tilde{q}_0 \neq 1$ (as MRF).			
	S: NB(1.38,2.07); B: ZINB(0.66, 0.33, 2.01)		S: NB(6.95,0.89); B: ZINB(0.53, 0.36, 0.88)	
	$(\tilde{q}_0, \tilde{q}_1)$=(0.002,0.940)		$(\tilde{q}_0, \tilde{q}_1)$=(0.003,0.866)	
	FNDR	p-value	FNDR	p-value
MRF	0.0090	-	0.0020	-
iSeq	0.0086	0.7778	0.0052	$< 2.2e-16$
BayesPeak	0.0292	$< 2.2e-16$	0.0088	$< 2.2e-16$
	Scenario 2: Poisson-Poisson mixture with $\tilde{q}_1 + \tilde{q}_0 = 1$ (as iSeq).			
	S: Poisson(1.5); B: Poisson(0.5)		S: Poisson(9.0); B: Poisson(0.5)	
	$\tilde{q}_1 = 1 - \tilde{q}_0 = 0.98$		$\tilde{q}_1 = 1 - \tilde{q}_0 = 0.98$	
	FNDR	p-value	FNDR	p-value
MRF	0.0606	-	3.79e-06	-
iSeq	0.0586	0.7661	1.04e-05	0.1073
BayesPeak	0.4547	$< 2.2e-16$	0.2707	$< 2.2e-16$
	Scenario 3: Poisson-Poisson mixture with $\tilde{q}_1 + \tilde{q}_0 \neq 1$.			
	S: Poisson(3.0); B: Poisson(0.5)		S: Poisson(6.0); B: Poisson(0.2)	
	$(\tilde{q}_0, \tilde{q}_1) = (0.02, 0.5)$		$(\tilde{q}_0, \tilde{q}_1) = (0.02, 0.5)$	
	FNDR	p-value	FNDR	p-value
MRF	0.0225	-	0.0011	-
iSeq	0.0287	$< 2.2e-16$	0.0016	$1.737e-12$
BayesPeak	0.0299	$< 2.2e-16$	0.0200	$< 2.2e-16$
	Scenario 4: ZINB-NB mixture with $\tilde{q}_1 + \tilde{q}_0 = 1$.			
	S: NB(3.0, 1.0); B: ZINB(0.5, 0.5,0.5)		S: NB(6.0, 1.0); B: ZINB(0.5, 0.5,0.5)	
	$(\tilde{q}_0, \tilde{q}_1) = (0.02, 0.98)$		$(\tilde{q}_0, \tilde{q}_1) = (0.02, 0.98)$	
	FNDR	p-value	FNDR	p-value
MRF	0.0168	-	0.0039	-
iSeq	0.2903	$< 2.2e-16$	0.1874	$< 2.2e-16$
BayesPeak	0.4100	$< 2.2e-16$	0.4310	$< 2.2e-16$
	Scenario 5: multiple experiments and Markov property.			
	Rep 1 S: NB(2.74, 1.55); B: ZINB(0.63, 0.43, 2.32)		Rep 1 S: NB(3.80, 1.14); B: ZINB(0.66, 0.40, 3.01)	
	Rep 2 S: NB(5.99, 0.96); B: ZINB(0.48, 0.48, 1.25)		Rep 2 S: NB(7.40, 0.96); B: ZINB(0.49, 0.40, 1.06)	
	$(\tilde{q}_0, \tilde{q}_1) = (0.003, 0.839)$		$(\tilde{q}_0, \tilde{q}_1) = (0.003, 0.830)$	
	FNDR	p-value	FNDR	p-value
MRF	0.0011	-	0.0008	-
Mixture	0.0072	$< 2.2e-16$	0.0057	$< 2.2e-16$
jMOSAiCS	0.0104	$< 2.2e-16$	0.0081	$< 2.2e-16$
	Scenario 6: multiple experiments and no Markov property.			
	Rep 1 S: NB(2.74, 1.55); B: ZINB(0.63, 0.43, 2.32)		Rep 1 S: NB(3.80, 1.14); B: ZINB(0.66, 0.40, 3.01)	
	Rep 2 S: NB(5.99, 0.96); B: ZINB(0.48, 0.48, 1.25)		Rep 2 S: NB(7.40, 0.96); B: ZINB(0.49, 0.40, 1.06)	
	$p(X = 1) = 0.017$		$p(X = 1) = 0.020$	
	FNDR	p-value	FNDR	p-value
MRF	0.0073	-	0.0058	-
Mixture	0.0073	0.5001	0.0058	0.6738
jMOSAiCS	0.0099	$< 2.2e-16$	0.0080	$< 2.2e-16$

Table 2: Number of enriched and differentially bound regions identified by MRF, a ZINB-NB mixture model (Mixture) and jMOSAiCS, using technical replicates of CBP and p300 at time 30. The last row reports the number of regions identified by all three methods.

	Enriched regions		Differentially bound regions	
Method	CBPT30	p300T30	only CBP	only p300
MRF	2977	3970	69	347
Mixture	981	1848	29	395
jMOSAiCS	1231	1970	-	-
Overlap	899	1672	-	-