Bayesian Analysis (2013) **8**, Number 4, pp. 759–780

Nonparametric Bayesian Bi-Clustering for Next Generation Sequencing Count Data

Yanxun Xu [*¶], Juhee Lee [†], Yuan Yuan [‡], Riten Mitra [§], Shoudan Liang [¶], Peter Müller [§] and Yuan Ji [‖]

Abstract. Histone modifications (HMs) play important roles in transcription through post-translational modifications. Combinations of HMs, known as chromatin signatures, encode specific messages for gene regulation. We therefore expect that inference on possible clustering of HMs and an annotation of genomic locations on the basis of such clustering can contribute new insights about the functions of regulatory elements and their relationships to combinations of HMs. We propose a nonparametric Bayesian local clustering Poisson model (NoB-LCP) to facilitate posterior inference on two-dimensional clustering of HMs and genomic locations. The NoB-LCP clusters HMs into HM sets and lets each HM set define its own clustering of genomic locations. Furthermore, it probabilistically excludes HMs and genomic locations that are irrelevant to clustering. By doing so, the proposed model effectively identifies important sets of HMs and groups regulatory elements with similar functionality based on HM patterns.

Keywords: ChIP-Seq, Histone modifications, Nonparametric Bayes, Bi-Clustering, Markov chain Monte Carlo

1 Introduction

Histones are proteins that package DNA into structural units called nucleosomes. Through post-translational modifications, histones play key roles in transcription (Bernstein et al. (2002); Roh et al. (2005)), chromosomal segregation (Andersson et al. (2009)), and DNA repair. Combinations of such histone modifications (HMs) are known as the "histone code", which modulates chromatin structure to regulate gene expression. For example, combinations of HMs have been linked to cancer prognosis (Kurdistani (2007)) and clinical decisions (Kurdistani (2011)).

Recently, several HM patterns have been shown to be associated with various classes of regulatory elements, known as chromatin signatures (Bernstein et al. (2006)). For example, distinct and predictive chromatin signatures are used to characterize active promoters and enhancers (Heintzman et al. (2007); Heintzman et al. (2009)). These

[*]Department of Statistics, Rice University, Houston, TX, U.S.A.
[†]Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A.
[‡]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX, U.S.A
[§]Department of Mathematics, University of Texas Austin, Austin, TX, U.S.A. pmueller@math.utexas.edu
[¶]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.
[‖]NorthShore University HealthSystem, Chicago, IL, U.S.A. yji@northshore.org

 DOI:10.1214/13-BA822

results lead us to look for more such patterns. We expect that regulatory elements with similar functionality are likely to share similar patterns of some subset of HMs. We conjecture that annotating genomic location on the basis of such patterns could be a promising step towards deciphering the histone code.

We consider data from ChIP-Seq experiments, which are applications of next generation sequencing (NGS) technology and will be introduced in the next Section. The sequencing data is a matrix of HM counts, with rows representing genomic locations and columns representing HMs. Traditional one-dimensional clustering techniques aim to partition either the HMs or genomic locations. While useful, such clustering methods are often inadequate to identify co-localized HMs that are important factors in deciding functions of genomic regions. In addition, how genomic regions cluster should depend on which subset of HMs we focus on. Different HM sets might partition genomic locations in different ways, which might indicate different cellular or chromatin states.

These considerations lead us to consider two-dimensional clustering. Getz et al. (2000) presented a coupled two-way clustering approach that employs hierarchical clustering to each separate dimension, combining the clustering results along each dimension in a problem-specific manner. Later, Cheng and Church (2000) introduced the concept of biclustering to find biclusters within a data matrix. They proposed a quantitative measure as a guide to search for biclusters in gene expression data. Lazzeroni and Owen (2002) developed the plaid model that describes gene expression data as a sum of biclusters. In their model, each bicluster contains a group of genes expressed similarly within a given set of samples, indicating the presence of a particular biological process. Turner et al. (2005) proposed an improved algorithm for fitting the plaid model. Li et al. (2009) reported an effective and computationally efficient biclustering algorithm, QUBIC, to identify overlapping biclusters by employing a combination of qualitative measures of gene expression data and a combinatorial optimization technique.

We extend these approaches to incorporate two important new features: first, we develop models for discrete count data as opposed to continuous measurements. Second, we introduce full model-based inference that defines a posterior probability model for the random partitions, including a full probabilistic description of the associated uncertainties. Specifically, we propose a nonparametric Bayesian local clustering Poisson model (NoB-LCP) to close this gap in the existing literature. The proposed method builds on Lee et al. (2013a) who developed bi-directional clustering for continuous protein activation data. The proposed NoB-LCP model clusters any two HMs (columns) together if they give rise to the same partition of genomic locations. That is, the partitions of genomic locations (rows) are nested within clusters of HMs, with a separate partition of locations for each HM cluster. This definition of HM clusters based on inducing the same (nested) clustering of genomic locations distinguishes the proposed model from most currently used models, including Bayesian nonparametric approaches, that define clusters based on common parameters in the sampling model. We will refer to the column clusters as "HM clusters" and to the row clusters as "location clusters". Location clusters can be used to define different functional signatures that are characterized by subsets of HMs, while HM clusters suggest unique combinatory patterns that annotate chromatin states. One advantage of nonparametric Bayesian clustering

is that it provides model-based posterior probability models for the random partitions. It entirely avoids the problem of specifying the number of clusters in advance. Another key difference between NoB-LCP and other biclustering methods is that we allow that some HMs and some genomic locations might not meaningfully cluster with the other HMs or locations. In practice, experimental data usually include noisy rows and/or columns that are irrelevant to the scientific problem being addressed. Excluding them significantly increases the power of detecting meaningful signals in the remaining rows and columns.

The paper proceeds as follows. We introduce the motivating application and the data set in Section 2. In Section 3, we present probability models and computational methods for posterior inference. We present a simulation study in Section 4, and in Section 5, we report inference results on the ChIP-Seq data. We conclude with a discussion in Section 6.

2 ChIP-Seq Data

ChIP-Seq integrates chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (Seq) to identify genome-wide expression patterns of DNA-binding proteins. ChIP-Seq data record the counts of sequence tags mapped onto non-overlapping positions that cover the genome. By applying HM-specific antibodies, ChIP-Seq experiments can record the counts of DNA fragments that include a certain HM. And the fragments are mapped to specific locations across the whole genome. A large count of DNA fragments indicates high occurrence of the targeted HM.

We consider a ChIP-Seq experiment for CD4+ T lymphocytes (Barski et al. (2007); Wang et al. (2008)), in which 39 types of HMs, including 18 acetylations, 20 methylations, and a special histone modification H2A.Z, are reported. We focus on genomic locations with at least one enriched HM for meaningful inference and use the peak-calling program SICER (Zang et al. (2009)) to decide enrichment. SICER parameters were set to W_SIZE=200, GAP_SIZE=600, EVALUE=1000, FRAG_SIZE=150. Also any adjacent windows with unchanged SICER calls for the 39 HM counts are merged to create larger regions.

3 Methodology

3.1 Probability Model

The ChIP-Seq data is arranged in an $N \times G$ matrix $\mathbf{Y} = [y_{ig}]$ with each element y_{ig} representing the read count for HM g in genomic location i, $i = 1, 2, \ldots, N$ and $g = 1, 2, \ldots, G$. Here, genomic locations are defined as windows of 200 base pairs. We start the model construction with a random partition of HMs $\{1, \ldots, G\}$ into non-overlapping subsets C_q as $\{1, \ldots, G\} = \bigcup_{q=0}^{Q} C_q$. The unusual indexing starting with $q = 0$ is in anticipation of the upcoming discussion. The number $Q + 1$ of subsets is random itself. It is part of the random partition $\{C_0, \ldots, C_Q\}$. In the following discussion we find

it convenient to index the partition equivalently by cluster membership indicators c_g, $g = 1, \ldots, G$ with $c_g = q$ if $g \in C_q$. Under the proposed model some HMs are singled out as not giving rise to a nested partition of genomic locations. We refer to these HMs as the "idle HMs", and to the remaining ones as "active HMs". We use the special cluster C_0 to combine the idle HMs, i.e., $c_g = 0$ for all idle HMs. Assume that there are $G' < G$ active HMs and $(G - G')$ idle HMs. We propose a zero-enriched Pólya urn (Sivaganesan et al. (2011)) prior for $\boldsymbol{c} = (c_1, c_2, \ldots, c_G)^T$:

$$P(\boldsymbol{c}) = \pi_0^{G'}(1-\pi_0)^{G-G'} \frac{\alpha^Q \prod_{q=1}^{Q} \Gamma(p_q)}{\prod_{g=1}^{G'}(\alpha + g - 1)}, \tag{1}$$

where p_q is the number of HMs in HM set q and α is the total mass parameter of the Pólya urn scheme. Under this model, $c_g = 0$ with probability $(1 - \pi_0)$, i.e., HM g falls into the idle HM set with probability $(1 - \pi_0)$. When c_g is non-zero, HM g is either assigned to an existing active HM set q with probability proportional to p_q, or assigned to a new singleton active HM set with probability proportional to α. We refer to (1) as a nonparametric Bayesian prior model. The Pólya urn is traditionally considered a nonparametric Bayesian model since it can be constructed as the partition that is implied by the ties under i.i.d. sampling from a probability measure with a Dirichlet process prior. See, for example, a recent review by Lee et al. (2013b).

Next, we consider clustering of genomic locations for each of the Q active HM sets. Recall that the partition of locations is nested within HM sets, i.e., we want to allow for a different set of location clusters with respect to each HM cluster. We define $\boldsymbol{r}_q = (r_{q1}, r_{q2}, \ldots, r_{qN})^T$ to be the N cluster labels $r_{qi} \in \{0, \ldots, D_q\}$ that describe the partition of genomic locations corresponding to the q-th HM set. Again we allow for a special cluster $r_{qi} = 0$ of inactive genomic locations that do not meaningfully co-cluster with other loci with respect to the q-th HM set. We assume that $\boldsymbol{r}_q$ includes D_q active location clusters with $r_{qi} = d$ indicating that locus i is assigned to active location cluster d, and $r_{qi} = 0$ indicating that genomic location i is assigned to the idle location cluster. Let $\boldsymbol{r} = (\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_Q^T)^T$. We assume independent zero-enriched Pólya urn priors for each $\boldsymbol{r}_q$ given by

$$P(\boldsymbol{r} \mid \boldsymbol{c}) = \prod_{q=1}^{Q} P(\boldsymbol{r}_q) \quad \text{and} \quad P(\boldsymbol{r}_q) = \pi_1^{m_q}(1-\pi_1)^{N-m_q} \frac{\beta^{D_q} \prod_{d=1}^{D_q} \Gamma(n_{qd})}{\prod_{i=1}^{m_q}(\beta + i - 1)}. \tag{2}$$

Note that Q is random and depends on $\boldsymbol{c}$. In (2), for a given active HM set q, n_{qd} is the number of genomic locations in the active location cluster $d(> 0)$ and $m_q = \sum_{d=1}^{D} n_{qd}$. In addition, β is the total mass parameter of the Pólya urn. The cluster label r_{qi} is allowed to be 0 with probability $(1 - \pi_1)$, characterizing the idle location cluster.

The described prior probability model can be characterized as a partition of HMs and a nested partition of locations, nested within each (active) cluster of HMs. In words, we identify subsets of HMs that are characterized by the fact that genomic locations cluster into the same subsets with respect to all HMs in a HM cluster. These subsets will provide important information on the co-location patterns of HMs and actionable target HMs

for diagnosis and prognosis. In addition, the resulting clusterings of genomic regions can be examined and integrated with other information (e.g., transcription binding sites) to potentially achieve better understanding of gene regulation.

Given $\boldsymbol{c}$ and $\boldsymbol{r}$, we now define a sampling model for the observed counts y_{ig}. Let $\text{Poi}(\theta)$ denote a Poisson distribution with mean θ. We start with a Poisson sampling model for the count data, i.e.,

$$y_{ig} \sim \text{Poi}(\theta_{ig}).$$

The prior probability model for θ_{ig} makes use of the clustering. Let $\text{Ga}(a, b)$ denote a gamma distribution with mean a/b. We define $P(\theta_{ig} \mid \boldsymbol{c}, \boldsymbol{r})$ as follows. Assume $c_g = q$ and $r_{qi} = d$. The model gives meaning to the partition of locations by assuming a shared rate $\widetilde{\theta}_{dg}$ for all locations in the same location cluster, i.e., $\theta_{ig} = \widetilde{\theta}_{dg}$ for all i with $r_{qi} = d$. But HMs in the same HM cluster share the same partition of locations only, not the same rate, i.e., $\theta_{jh} = \widetilde{\theta}_{dh} \neq \widetilde{\theta}_{dg}$ for all (h, j) with $c_h = q$ and $r_{qj} = d$ and $h \neq g$. We assume

$$\widetilde{\theta}_{dg} \overset{\text{iid}}{\sim} \text{Ga}(k_{0g}, \lambda_{0g}),$$

For the idle genomic locations in the active HM sets, i.e., $r_{qi} = 0$ with $q > 0$, we assume a priori $\theta_{ig} \overset{\text{iid}}{\sim} \text{Ga}(k_{1g}, \lambda_{1g})$. For idle HMs, i.e., $c_g = 0$, we assume $\theta_{ig} \overset{\text{iid}}{\sim} \text{Ga}(k_{2g}, \lambda_{2g})$ for all locations i. Note that taking a Poisson sampling model with parameter θ_{ig} and a gamma prior for θ_{ig}, we equivalently constructed a negative binomial sampling model for the count data, which provides additional variabilities to account for potential over dispersion.

Finally, denoting with $\text{Beta}(a, b)$ a beta distribution with parameters (a, b), we assume conditionally conjugate priors

$$\pi_0 \sim \text{Beta}(a_0, b_0), \qquad \pi_1 \sim \text{Beta}(a_1, b_1).$$

The beta hyperprior on π_0 and π_1 is important to allow for inference about the number of active HMs and locations, as it allows adjustment of the priors $p(\boldsymbol{c} \mid \pi_0)$ and $p(\boldsymbol{r}_q \mid \boldsymbol{c}, \pi_1)$ to adapt to the level of noise in the data. See, for example Scott and Berger (2010) for a discussion of this multiplicity correction feature.

Figure 1 is a graphical illustration of the proposed NoB-LCP model. It demonstrates the core idea of how we define local clusters. In Figure 1, we assume that 9 HMs belong to two active HM sets and and an idle HM set, including HMs 5, 8 and 9. In the two active HM sets, cells in off-white are idle genomic locations. The rest of cells marked with the same color in the same column form local clusters of genomic locations (rows). Different colors indicate different values of parameters $\widetilde{\theta}_{dg}$. Within each local cluster, the colors are the same across the genomic locations but different across different HMs. We define an active HM set as the set of HMs that partition the genomic locations in the same way, regardless of the actual values of $\widetilde{\theta}_{dg}$. This highlights the important difference between NoB-LCP and other clustering methods that often assume common values of $\widetilde{\theta}_{dg}$ for items in the same cluster. In other words, in Figure 1, the cells in each local cluster would be marked in the same color across both genomic locations and HMs.

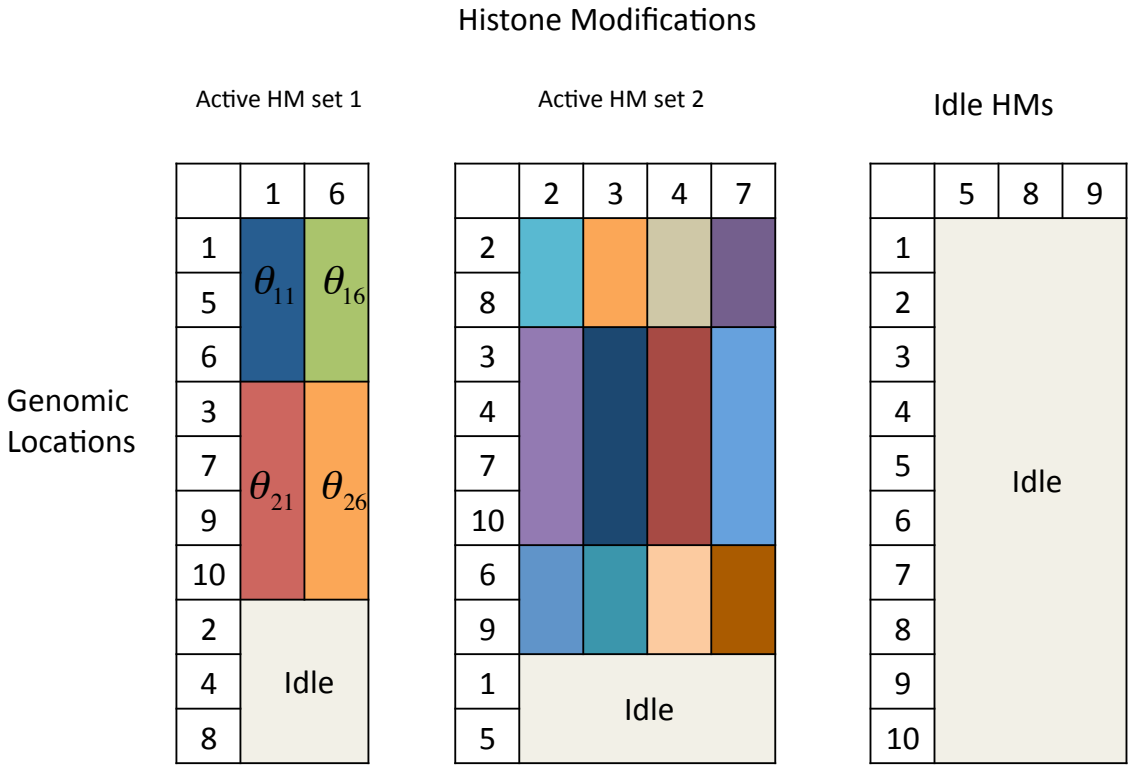

Figure 1: An illustration of the proposed NoB-LCP model with 9 HMs and 10 genomic locations. There are two active HM sets and an idle HM set, including HMs 5, 8, 9. In the two active HM sets, cells in off-white are idle genomic locations. The rest of the cells marked with the same color in the same column form local clusters of genomic locations (rows). Different colors indicate different values of parameters $\widetilde{\theta}_{dg}$.

In summary, the joint model is:

$$P(\mathbf{Y}, \mathbf{c}, \boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{k}, \boldsymbol{\lambda}, \pi_0, \pi_1) = P(\mathbf{Y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{r}, \boldsymbol{k}, \boldsymbol{\lambda}) P(\boldsymbol{r} \mid \boldsymbol{c}) P(\boldsymbol{c}) P(\pi_0) P(\pi_1). \quad (3)$$

3.2 Markov Chain Monte Carlo Simulations

We carry out posterior inference using MCMC simulation. Letting $[x \mid y, z]$ generically denote a transition density that updates an unknown parameter x conditional on currently imputed values for y and z, we propose a Gibbs sampler that iterates over the following sampling steps that draw random values from the transition densities:

$$[\boldsymbol{r} \mid \mathbf{Y}, \boldsymbol{c}, \pi_1], [\boldsymbol{c} \mid \mathbf{Y}, \boldsymbol{r}, \pi_0], [\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{c}, \boldsymbol{r}], [\pi_0 \mid \boldsymbol{c}], [\pi_1 \mid \boldsymbol{c}, \boldsymbol{r}]$$

We start by generating r_{qi}, $q = 1, \ldots, G'$, $i = 1, \ldots, N$, from its full conditional posterior distribution. When resampling r_{qi} and c_g, we marginalize over $\boldsymbol{\theta}$.

Let Q denote the currently imputed number of active HM clusters. A challenge in constructing a valid transition probability arises when $c_g = Q + 1$ is considered, i.e., when we consider placing g into a new, $(Q+1)$-th, singleton HM cluster. The problem is that a proposal $c_g = Q + 1$ gives rise to a new partition $\boldsymbol{r}_{Q+1}$ of locations. We use the pseudo prior mechanism of Carlin and Chib (1995) to construct an MCMC scheme. We introduce a set of auxiliary variables $\widetilde{\boldsymbol{r}}_g = (\tilde{r}_{ig}, i = 1, 2, \ldots, N)$, $g = 1, 2, \ldots, G$, and augment the probability model with a pseudo prior $P(\widetilde{\boldsymbol{r}}_g \mid \pi_1)$. Let $p_{1g}(\boldsymbol{r} \mid \pi_1)$ denote the conditional posterior of the location partition with respect to a singleton HM cluster $\{g\}$. We define $P(\widetilde{\boldsymbol{r}}_g \mid \pi_1) = p_{1g}(\widetilde{\boldsymbol{r}}_g \mid \pi_1)$. Think of $\widetilde{\boldsymbol{r}}_g$ as a potential genomic location partition with respect to a singleton HM set $\{g\}$. In other words, when a new singleton HM set is proposed for c_g, the proposal distribution for the genomic location clusters under this new HM set is determined by imputed value $\widetilde{\boldsymbol{r}}_g$. Lastly we draw $\boldsymbol{\theta}$, π_0 and π_1 whose full conditional posterior distributions are in closed forms. More MCMC technical details are included in the Appendix.

3.3 Posterior Inference

A practical challenge related to posterior inference is the need to summarize a distribution over random partitions. Medvedovic et al. (2004) initially addressed this problem by estimating posterior probabilities that any two HMs are clustered together. They evaluated probabilities $H_{gh} = P(c_g = c_h \mid data)$ of pair-wise co-clustering, and used H as a distance matrix for a (deterministic) hierarchical clustering algorithm. Alternatively, Dahl (2006) proposed a point estimate of a random partition under a Dirichlet process mixture model by reporting a least-squares partition. Specifically, the least-squares clustering $\boldsymbol{c}^{LS}$ is the observed clustering $\boldsymbol{c}$ which minimizes the Frobenius distance (L_2 norm for matrices) between S^c and H, where S^c is an association $G \times G$ matrix whose (g, g') element is an indicator that HM g is clustered with HM g'. We include HMs in the idle HM set by letting $s^c_{g,g'} = 0$ for all g' if $c_g = 0$. Following Dahl (2006), we

propose a least-square summary

$$\boldsymbol{c}^{LS} = \arg\min_{\boldsymbol{c}} \mid\mid S^{\boldsymbol{c}} - H \mid\mid^2,$$

as a point estimate of the clustering of HMs, which minimizes the sum of the squared deviation of association matrix S from the matrix H of the posterior pairwise co-clustering probabilities. Given $\boldsymbol{c}^{LS}$, we compute $\boldsymbol{r}_q^{LS}$, the least square estimate of the clustering for genomic locations, based on the same formulation.

4 Simulation Studies

4.1 Simulation setup

We conducted simulation studies to evaluate the performance of the proposed NoB-LCP model. We compared posterior inference with the simulation truth and with inference under two alternative clustering methods, the plaid model and the QUBIC. Furthermore, to show the importance of zero-enriched Pólya urn priors which allow some HMs or genomic locations to be idle, we performed a sensitivity analysis by using regular Pólya urn priors without zero-enrichment as the prior for the random partitions of HMs and genomic locations. It means that we let $\pi_0 = 1$ and $\pi_1 = 1$ in (1) and (2) respectively.

We simulated a data matrix $\mathbf{Y}$ with $N = 300$ genomic locations and $G = 18$ HMs. We let 13 out of 18 HMs belong to two active HM sets, in which HMs 1-7 belonged to set 1 and HMs 8-13 to set 2. The remaining 5 HMs, HMs 14–18, belonged to the idle HM set. We assumed that the active HM set 1 partitioned the genomic locations into four location clusters including one idle location cluster, i.e., $D_1 = 3$, and that the active HM set 2 partitioned the genomic locations into three location clusters including one idle location cluster, i.e., $D_2 = 2$. We generated location cluster labels, r_{qi}, for each active HM set assuming that a genomic location belonged to one of the location clusters with equal probability. In keeping with the definition of the idle HM set ($q = 0$), we did not generate location clusters with respect to the idle HMs with $c_g = 0$. We fixed $\widetilde{\theta}_{dg}$ for all the active location clusters for each of the 13 HMs residing in the active HM set as listed in Table 1. Finally, denoting with NB(mean $= a$, size $= b$) a negative binomial distribution with mean=a, variance=$a+a^2/b$ and with Unif$(0, 1)$ a Uniform distribution on $(0, 1)$, the remaining θ_{ig} were independently generated from NB(mean $= \mu$, size $= 1$), where $\mu \sim$ Unif$(0, 10)$, including the idle genomic locations in the active HM sets and all the genomic locations in the idle HM set. The NB distribution was chosen to examine the sensitivity of posterior inference with respect to deviations from the assumed Poisson sampling model.

4.2 Simulation Results

The left panel of Figure 2 shows the heatmaps of y_{ig} under the simulation truth. After rearranging the HMs and the genomic locations within each active HM set according to the simulation truth, we can clearly observe the local clustering patterns in the data. In

		HM 1	HM 2	HM 3	HM 4	HM 5	HM 6	HM 7
HM set 1	cluster 1	11	9	7	13	13	9	13
	cluster 2	15	7	15	7	9	7	7
	cluster 3	13	11	9	9	7	11	15

		HM 8	HM 9	HM 10	HM 11	HM 12	HM 13
HM set 2	cluster 1	9	15	9	11	7	9
	cluster 2	11	11	15	9	13	7

Table 1: The true mean counts for active genomic location clusters, $\tilde{\theta}_{dg}$, in the simulated data.

the active HM sets, the idle genomic locations, which are located in the first row block, do not show a noticeable pattern: the colors are more or less randomly scattered. In contrast, the active genomic locations in the columns corresponding to active HM sets show clear patterns and the colors are more homogeneous within each location cluster. In the idle HM set, since the genomic locations do not cluster, the corresponding color mapping exhibits large variability.

We applied the proposed NoB-LCP model to the simulated data. In the MCMC posterior simulation, we initialized the HMs allocation variable $\boldsymbol{c}$ using the clustering result from hierarchical clustering by cutting the dendrogram to achieve two active HM sets and one idle HM set. HMs 2, 4, 5 and 6 belonged to active HM set 1, HMs 8, 11 and 13 belonged to active HM set 2 and the remaining belonged to the idle HM set. The initial values and priors of π_0 and π_1 were set to 0.5 and Beta$(1, 1)$, respectively. We fixed parameters k_{0g} and λ_{0g} by setting the mean of $\tilde{\theta}_{dg}$ equal to g-th column mean of $\mathbf{Y}$ and setting the variance of $\tilde{\theta}_{dg}$ equal to 10. Finally, k_{1g}, λ_{1g}, k_{2g} and λ_{2g} were computed by setting the mean of θ_{ig} equal to g-th column mean of $\mathbf{Y}$ and variance equal to 50. After 10,000 MCMC iterations with 5,000 burn-in, the Markov chains converged and mixed well. We conducted convergence diagnostics using the R package *coda* and found no evidence for convergence problems. Traceplots and empirical autocorrelation plots (not shown) for the imputed parameters indicate a well mixing Markov chain. For example, the empirical autocorrelation of π_0 and π_1 is practically zero beyond lag 2. The simulation was carried out on a MacBook Pro laptop with 2.53 GHz Intel Core and 8GB memory. Computation was completed in 2.5 hours.

The least-squares summary of the posterior on $\boldsymbol{c}$ was $\boldsymbol{c}^{LS} = (1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0)$. Conditional on $\boldsymbol{c}^{LS}$, we further calculated the least-squares estimates of genomic location clusters for active HM sets, $\boldsymbol{r}_q^{LS}$, $q = 1, 2$. Figure 2 right panel shows that the NoB-LCP model correctly detected the two active HM sets in the simulation data: HMs 1-7 belonged to the active HM set 1 and HMs 8-13 belonged to the active HM set 2, the remaining HMs belonged to the idle HM set, consistent with the simulation truth. Tables 2 and 3 show that there are five estimated active genomic

Simulation Truth Posterior Estimate

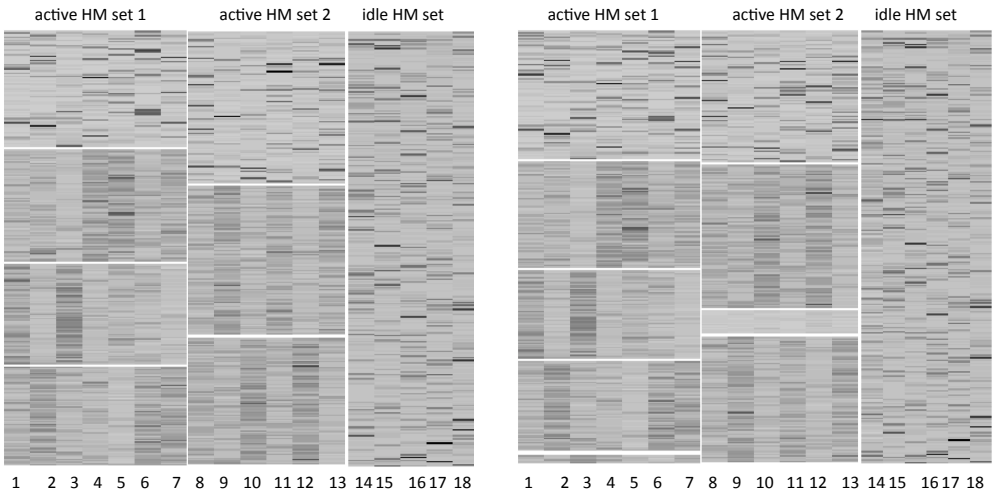

Figure 2: Heatmaps of the HM sets in the simulation truth versus the identified HM sets under the NoB-LCP model. The first row block of each active HM set is the idle genomic location cluster for that HM set. The remaining blocks are active genomic location clusters. The division of genomic location clusters is indicated by white horizontal lines

location clusters and one idle genomic location cluster for HM set 1, where clusters $\{0, 1, 2, 3\}$ dominate and largely overlap with the four true genomic location clusters. And the model identified four active genomic location clusters and one idle genomic location cluster with respect to HM cluster 2, where clusters $\{0, 1, 2\}$ dominated and largely overlapped with the three true genomic location clusters of true HM set 2.

For comparison, the two alternative methods, the plaid model and QUBIC, were applied to the same simulated data. Figure 3 shows the heatmaps of HMs in two biclusters identified by the plaid model. The first bicluster included 18 genomic locations of HMs 8, 9, 10, 11 and 12, all of which belonged to true genomic location cluster $d = 0$ of true

	0	1	2	3	4	5
active HM set 1	89	75	63	64	1	8
active HM set 2	91	101	18	89	1	-

Table 2: The number of genomic locations in each genomic location cluster for active HM sets.

HM set 1

	$\boldsymbol{c}_1^{LS}$					
$\boldsymbol{c}^{TRUE}$	0	1	2	3	4	5
d=0	80	0	0	0	0	0
d=1	3	68	1	1	1	3
d=2	5	0	61	1	0	4
d=3	0	7	1	60	0	1

HM set 2

	$\boldsymbol{c}_2^{LS}$				
$\boldsymbol{c}^{TRUE}$	0	1	2	3	4
d=0	87	0	0	18	0
d=1	3	85	17	0	0
d=2	1	4	84	0	1

Table 3: Comparisons of the location cluster membership estimated by the NoB-LCP model with the true location cluster membership.

HM set 2. The second bicluster included 37 genomic locations of HMs 14 and 15, which belonged to the idle HM set under the simulation truth. The QUBIC method detected 23 biclusters, 17 of which only included one single HM and the other six included two HMs. Figure 4 shows the heatmaps of HMs in the six biclusters with two HMs. Some of those six biclusters included idle HMs such as HMs 14, 16 and 17, and others included either idle genomic locations, active genomic locations, or multiple active location sets. For example, bicluster 1 included 15 genomic locations of HMs 4 and 7, among which 7 belonged to the true genomic location cluster $d = 1$ of true HM set 1, and 5 belonged to the true genomic location cluster $d = 2$ of true HM set 1; bicluster 2 included 21 genomic locations of HMs 8 and 10, among which 17 belonged to the true genomic location cluster $d = 2$ of true HM set 2.

Next we replaced the zero-enriched Pólya urn priors in (1) and (2) with regular Pólya urn priors. And we used the same hyperparameters and initialized the parameters as before, except for $\boldsymbol{c}$. We initialized $\boldsymbol{c}$ by letting HMs 1-13 belong to active HM set 1 and HMs 14-18 belong to active set 2. After 10,000 iterations of MCMC simulation with 5,000 burn-in, the Markov chains converged and mixed well.

The least-squares summary of the posterior on $\boldsymbol{c}$ was $\boldsymbol{c}^{LS} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2)$. Conditional on $\boldsymbol{c}^{LS}$, we further calculated the least-squares estimates of genomic location clusters for active HM sets, $\boldsymbol{r}_q^{LS}$, $q = 1, 2$. Figure 5 shows the heatmaps of two detected active HM sets. Compared to the simulation truth, the model with regular Pólya urn priors failed to differentiate the two active HM sets. In addition, many small and meaningless genomic location clusters nested within two active HM sets can be observed.

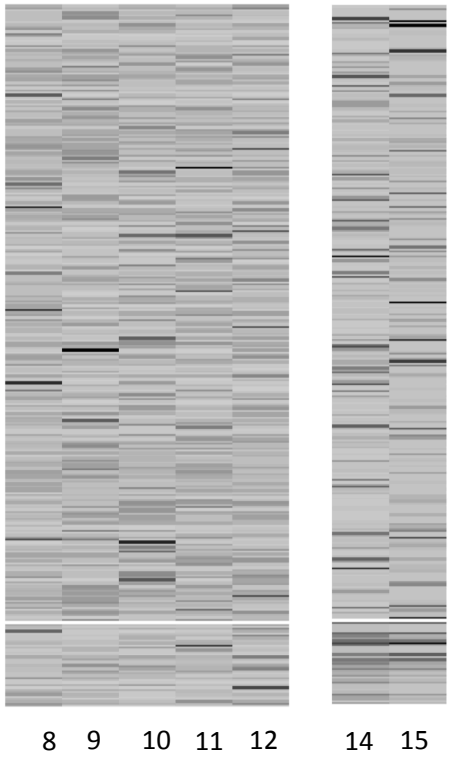

Figure 3: Heatmaps of HMs in two biclusters of the simulated data identified by the plaid model. The division of genomic locations is indicated by white horizontal lines. Below the white line is the detected bi-cluster.

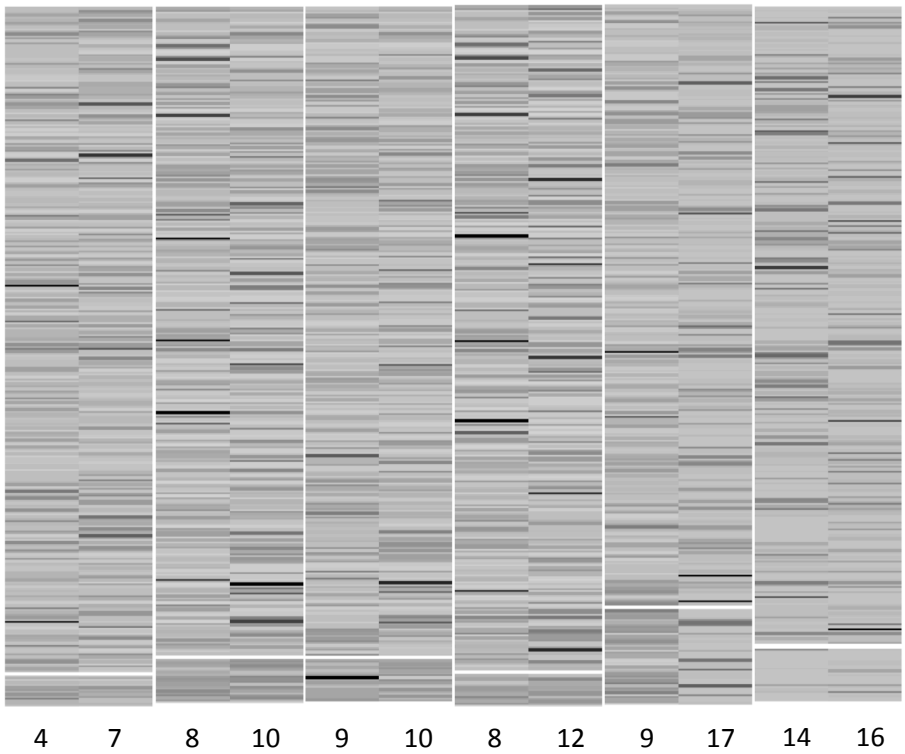

Figure 4: Heatmaps of HMs in six biclusters for the simulated data identified by QUBIC. The division of genomic locations is indicated by white horizontal lines. Below the white line is the detected bi-cluster.

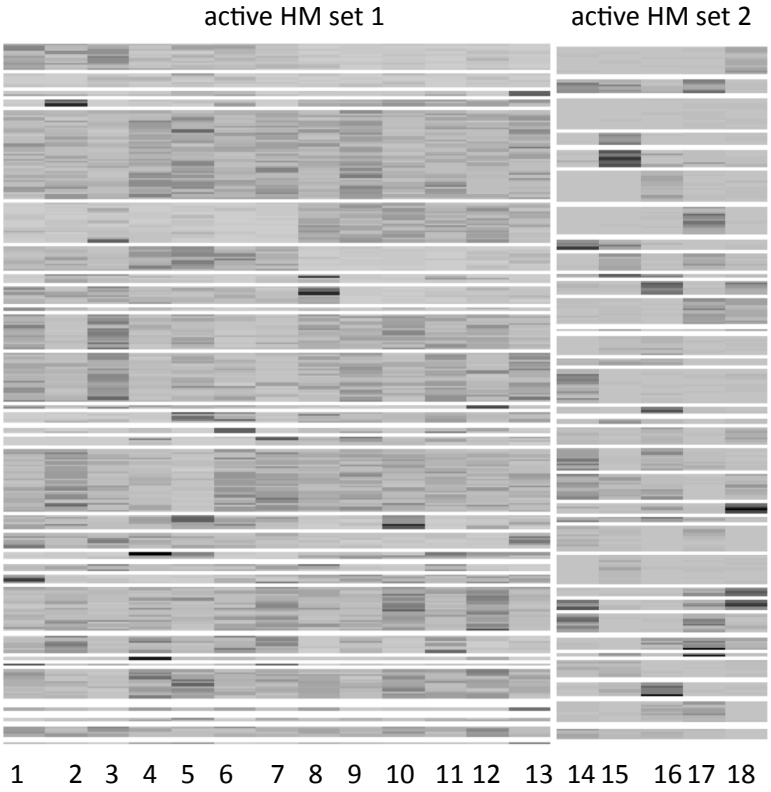

Figure 5: Heatmaps of HMs in two active HM sets for the simulated data identified by the model with regular Pólya urn priors without zero-enrichment. The division of genomic locations is indicated by white horizontal lines.

5 ChIP-Seq Data Analysis

We present local clustering results for the ChIP-Seq data described in Section 2. For demonstration purpose, we apply our NoB-LCP model to clustering of promoters and insulators, both of which are important regulatory elements. Information on the genomic location for promoters was obtained from the UCSC Genome Browser (Fujita et al. (2011)). Read counts were recorded for all genomic locations and all HMs. The insulator information was obtained from the CTCFBSDB (Bao et al. (2008)), a CTCF binding site database to identify insulators.

We consider a small subset of the ChIP-Seq data covering randomly selected 50 genomic locations in promoter regions and 50 genomic locations in insulator regions. The data is a 100×39 matrix with genomic locations as rows and HMs as columns. To fit the NoB-LCP model, $\boldsymbol{c}$ is initialized by the clustering determined by a (deterministic) hierarchical clustering algorithm. We chose parameters k_{0g} and λ_{0g} by fixing the prior variance of $\widetilde{\theta}_{dg}$ at $\mathrm{Var}(\widetilde{\theta}_{dg}) = 10$, and matching the mean of $\tilde{\theta}_{dg}$ with the column means of the data matrix. Similarly, k_{1g}, λ_{1g}, k_{2g} and λ_{2g} are chosen by fixing the prior variance at 50, and matching the prior mean of θ_{ig} with the column means. Finally, π_0 and π_1 are initially set to 0.5 and we used $a_0 = b_0 = a_1 = b_1 = 1$, i.e., uniform hyperpriors. After 10,000 iterations with a 5,000 burn-in for MCMC posterior simulation, we evaluated convergence diagnostics (R package *coda*) and found no evidence for practical convergence problems. The chain mixed well.

We compute the least-squares estimates $\boldsymbol{c}^{LS}$ and $\boldsymbol{r}^{LS}$ to summarize posterior inference. The NoB-LCP model identifies 3 active HM sets, each of which partitions genomic locations differently. Figure 6 shows the heatmaps of all active HM sets. These three sets are candidates of co-localized HMs that relate to gene transcription. In addition, the heatmap shows genomic location clusters nested in each active HM set.

Posterior inference distinguishes different types of regulatory elements and clusters similar types together reasonably well. For example, active HM set 1 includes the following HMs: H4K12ac, H3K79me2, H3K79me3. Genomic location clusters 1 and 5 in active HM set 1 include only promoter regions, in which H4K12ac, H3K79me2 and H3K79me3 clearly show relatively high expression in Figure 6. Our results are consistent with previous findings that H4K12ac counts are elevated in the promoter and transcribed regions of active genes (Wang et al. (2008)), H3K79me2 and H3K79me3 are important histone markers for the prediction of promoter regions (Wang et al. (2009); Weishaupt et al. (2010)). Out of the 12 HMs in active HM set 2, all of them are acetylations; out of the 21 HMs in active HM set 3, 15 of them are methylations. From this fact, we can conjecture that the same types of histone modifications (methylations, acetylations, etc.) are more likely to be clustered together.

In addition, highly correlated HM patterns can be identified by our model. For example, active HM set 2 includes the following HMs: H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H3K18ac, H3K27ac, H3K36ac, H3K4ac, H3K9ac, and H4K91ac, which were reported to have relatively high correlation according to Wang et al. (2008).

For comparison, we applied the plaid model and QUBIC to the same ChIP-Seq data.

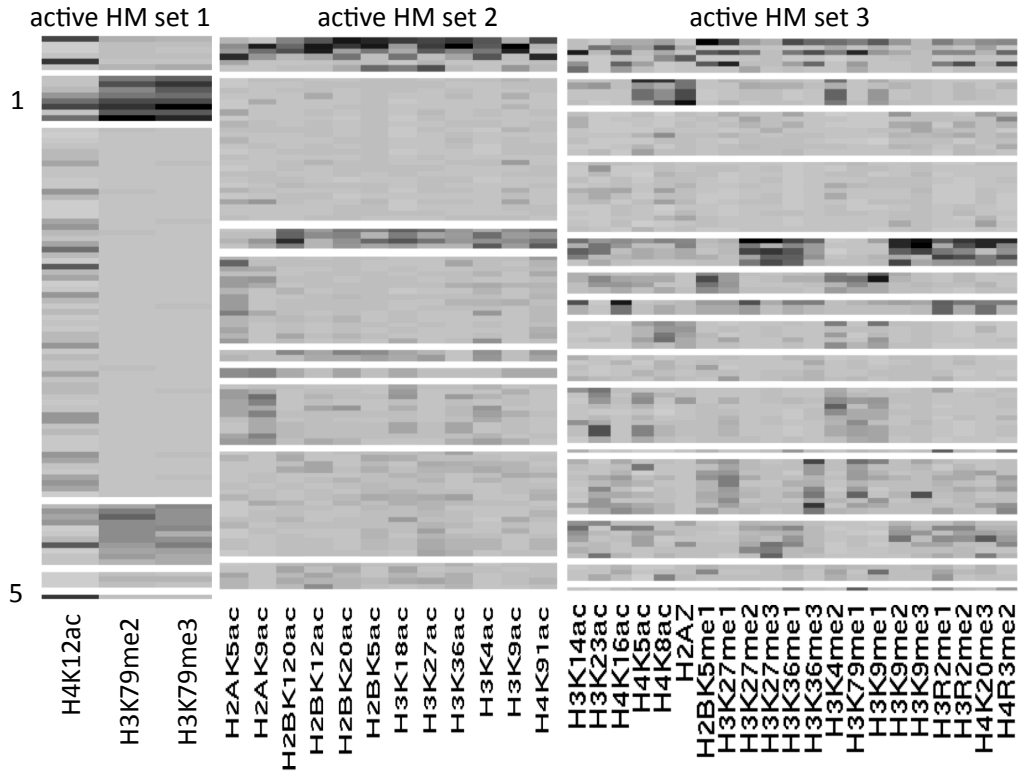

Figure 6: Heatmaps of three active HM sets for ChIP-Seq data. White horizontal lines indicate division of location clusters.

The plaid model did not report any biclusters. QUBIC found 57 biclusters, but none of them provide us clear divisions of regulatory elements. In addition, it is not easy to extract useful information from so many biclusters.

Finally, we used a qq-plot to validate the assumed sampling model. Assuming a Poisson/gamma hierarchical sampling model, we have implicitly defined a negative binomial marginal sampling model. The negative binomial model allows larger variabilities in modeling the counts. We made a qq-plot of the empirical c.d.f. of the observed ChIP-Seq data versus simulated data sampled from the imputed negative binomial distribution. We can see a linear relationship between two quantiles, suggesting that the hierarchical sampling model is well calibrated (Figure not shown).

6 Discussion

We propose a nonparametric Bayesian local clustering Poisson model for a count data matrix. The NoB-LCP model detects local clustering patterns by performing simultaneous clustering on columns and rows of a data matrix. Idle local clusters are introduced to better separate noisy HMs and location from the actual signals in the genomics data. Through simulation studies and the analysis of ChIP-Seq data we demonstrate the effectiveness of our model in grouping regulatory elements with similar functionality based on HMs patterns.

In this paper, we used zero-enriched Pólya urn priors to model random partitions of HMs and genomic locations. Although partitions do not allow overlap between the partitioning subsets in one imputation of the parameters, posterior inference could still report positive (marginal) posterior probability for membership in multiple clusters for the same HM (reporting such probabilities also requires a resolution of the label switching problem). Alternatively, one could use feature allocation models, such as the Indian buffet processes (Griffiths and Ghahramani (2005)) as priors for a random allocation of HMs to subsets, including membership in multiple subsets.

Acknowledgments

Yuan Ji and Peter Müller's research is supported in part by NIH R01 CA132897. Shoudan Liang's research in supported in part by NCI 5 K25 CA123344.

References

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). "Nucleosomes are well positioned in exons and carry characteristic histone modifications." *Genome research*, 19(10): 1732–1741. 759

Bao, L., Zhou, M., and Cui, Y. (2008). "CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators." *Nucleic acids research*, 36(suppl 1): D83–D87. 773

Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). "High-resolution profiling of histone methylations in the human genome." *Cell*, 129(4): 823–837. 761

Bernstein, B., Mikkelsen, T., Xie, X., Kamal, M., Huebert, D., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." *Cell*, 125(2): 315–326. 759

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T., and Schreiber, S. L. (2002). "Methylation of histone H3 Lys 4 in coding regions of active genes." *Proceedings of the National Academy of Sciences*, 99(13): 8695–8700. 759

Carlin, B. and Chib, S. (1995). "Bayesian model choice via Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society. Series B (Methodological)*, 473–484. 765

Cheng, Y. and Church, G. (2000). "Biclustering of expression data." In *Proceedings of the eighth international conference on intelligent systems for molecular biology*, volume 1, 93–103. 760

Dahl, D. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model." In Vannucci, M., Do, K.-A., and Müller, P. (eds.), *Bayesian inference for gene expression and proteomics*, 201–215. Cambridge: Cambridge University Press. 765

Fujita, P., Rhead, B., Zweig, A., Hinrichs, A., Karolchik, D., Cline, M., Goldman, M., Barber, G., Clawson, H., Coelho, A., et al. (2011). "The UCSC genome browser database: update 2011." *Nucleic acids research*, 39(suppl 1): D876–D882. 773

Getz, G., Levine, E., and Domany, E. (2000). "Coupled two-way clustering analysis of gene microarray data." *Proceedings of the National Academy of Sciences*, 97(22): 12079–12084. 760

Griffiths, T. L. and Ghahramani, Z. (2005). "Infinite Latent Feature Models and the Indian Buffet Process." In *In NIPS*, 475–482. MIT Press. 775

Heintzman, N., Hon, G., Hawkins, R., Kheradpour, P., Stark, A., Harp, L., Ye, Z., Lee, L., Stuart, R., Ching, C., et al. (2009). "Histone modifications at human enhancers reflect global cell-type-specific gene expression." *Nature*, 459(7243): 108–112. 759

Heintzman, N., Stuart, R., Hon, G., Fu, Y., Ching, C., Hawkins, R., Barrera, L., Van Calcar, S., Qu, C., Ching, K., et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." *Nature genetics*, 39(3): 311–318. 759

Kurdistani, S. (2007). "Histone modifications as markers of cancer prognosis: a cellular view." *British journal of cancer*, 97(1): 1–5. 759

— (2011). "Histone modifications in cancer biology and prognosis." *Epigenetics and Disease*, 91–106. 759

Lazzeroni, L. and Owen, A. (2002). "Plaid models for gene expression data." *Statistica Sinica*, 12(1): 61–86. 760

Lee, J., Müller, P., Zhu, Y., and Ji, Y. (2013a). "A nonparametric Bayesian model for local clustering with Application to Proteomics." *Journal of the American Statistical Association*, to appear. 760

Lee, J., Quintana, F., Müller, P., and Trippa, L. (2013b). "Defining Predictive Probability Functions for Species Sampling Models." *Statistical Science*, to appear. 762

Li, G., Ma, Q., Tang, H., Paterson, A., and Xu, Y. (2009). "QUBIC: a qualitative biclustering algorithm for analyses of gene expression data." *Nucleic acids research*, 37(15): e101–e101. 760

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004). "Bayesian mixture model based clustering of replicated microarray data." *Bioinformatics*, 20(8): 1222–1232. 765

Roh, T., Cuddapah, S., and Zhao, K. (2005). "Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping." *Genes and development*, 19(5): 542–552. 759

Scott, J. and Berger, J. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. 763

Sivaganesan, S., Laud, P., and Müller, P. (2011). "A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme." *Statistics in Medicine*, 30(4): 312–323. 762

Turner, H., Bailey, T., and Krzanowski, W. (2005). "Improved biclustering of microarray data demonstrated through systematic performance tests." *Computational statistics and data analysis*, 48(2): 235–254. 760

Wang, X., Xuan, Z., Zhao, X., Li, Y., and Zhang, M. (2009). "High-resolution human core-promoter prediction with CoreBoost_HM." *Genome research*, 19(2): 266–275. 773

Wang, Z., Zang, C., Rosenfeld, J., Schones, D., Barski, A., Cuddapah, S., Cui, K., Roh, T., Peng, W., Zhang, M., et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." *Nature genetics*, 40(7): 897–903. 761, 773

Weishaupt, H., Sigvardsson, M., and Attema, J. L. (2010). "Epigenetic chromatin states uniquely define the developmental plasticity of murine hematopoietic stem cells." *Blood*, 115(2): 247–256. 773

Zang, C., Schones, D., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data." *Bioinformatics*, 25(15): 1952–1958. 761

Appendix: MCMC details

Joint pdf

$$\begin{aligned} p(\mathbf{Y}, \boldsymbol{c}, \boldsymbol{r}, \boldsymbol{\theta}, \pi_0, \pi_1) &= p(\pi_0)p(\pi_1)p(\boldsymbol{c})p(\boldsymbol{r} \mid \boldsymbol{c})p(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{r}, \boldsymbol{k}, \boldsymbol{\lambda})p(\mathbf{Y} \mid \boldsymbol{\theta}) \\ &= p(\pi_0)p(\pi_1)p(\boldsymbol{c}) \prod_{q=1}^{Q} [p(\boldsymbol{r}_q) \\ &\quad \times \prod_{d=1}^{D_q} \{ \prod_{g \in V_q} (p(\tilde{\theta}_{dg} \mid k_{0g}, \lambda_{0g}) \prod_{i \in R_{qd}} p(y_{ig} \mid \tilde{\theta}_{dg}) \} \\ &\quad \times \prod_{g \in V_q} \prod_{i \in R_{s0}} (p(\theta_{ig} \mid k_{1g}, \lambda_{1g}) p(y_{ig} \mid \theta_{ig}))] \\ &\quad \times \prod_{g \in V_0} \prod_{i=1}^{N} (p(\theta_{ig} \mid k_{2g}, \lambda_{2g}) p(y_{ig} \mid \theta_{ig})) \end{aligned}$$

where $V_q = \{g \mid c_g = q, g = 1, \ldots, G\}$ is the set of HMs in an HM set q, $q = 0, \ldots, Q$, and $R_{qd} = \{i \mid r_{qi} = d, i = 1, \ldots, N\}$ is the set of genomic locations in genomic location cluster d corresponding to HM set q for $q = 1, \ldots, Q$ and $d = 1, \ldots, D_q$. We include $\boldsymbol{\lambda}$ and $\boldsymbol{k}$ in the conditioning sets to indicate the relevant (fixed) hyperparameters.

The prior probability distributions of $\boldsymbol{c}$ and $\boldsymbol{r}_q$ are a zero-enriched Pólya urn scheme given in Equations (1) and (2) of the main paper.

Full conditional

(i) Update $\boldsymbol{\theta}$

(a) For active HMs ($c_g > 0$) and active genomic locations ($r_{qi} > 0$), $q = 1, \ldots, Q$ and $d = 1, \ldots, D_q$,

$$\tilde{\theta}_{dg} \mid k_{0g}, \lambda_{0g}, \boldsymbol{c}, \boldsymbol{r}_q, \mathbf{y}_g \sim \text{Gamma}(k_{0g} + \sum_{i \in R_{qd}}^{N} y_{ig}, \lambda_{0g} + n_{qd}).$$

(b) For active HMs ($c_g > 0$) and idle genomic locations ($r_{qi} = 0$), $i = 1, \ldots, N$,

$$\theta_{ig} \mid k_{1g}, \lambda_{1g}, \boldsymbol{c}, \boldsymbol{r}_q, \mathbf{y}_g \sim \text{Gamma}(k_{1g} + y_{ig}, \lambda_{1g} + 1).$$

(c) For idle HM set ($c_g = 0$), $i = 1, \ldots, N$,

$$\theta_{ig} \mid k_{2g}, \lambda_{2g}, \boldsymbol{c}, \mathbf{y}_g \sim \text{Gamma}(k_{2g} + y_{ig}, \lambda_{2g} + 1).$$

(ii) Update π_0

$$\pi_0 \mid \boldsymbol{c} \sim \text{Beta}(a_0 + G', b_0 + G - G').$$

(iii) Update π_1

$$\pi_1 \mid \boldsymbol{r} \sim \text{Beta}(a_1 + \sum_q \sum_i I(r_{qi} > 0), b_1 + NQ - \sum_q \sum_i I(r_{qi} > 0)),$$

where I is an indicator function: $I(r_{qi} > 0) = 1$ if $r_{qi} > 0$; $I(r_{qi} > 0) = 0$ if $r_{qi} = 0$.

(iv) Update $\boldsymbol{r}_q$

Update $\boldsymbol{r}_q$ for active HM sets, $q = 1, \ldots, Q$ and $i = 1, \ldots, N$.

Remove $\boldsymbol{\theta}_i^q$, define $m_q^-, \boldsymbol{\theta}^{q-}, D_q^-, \boldsymbol{r}_q^-, n_q^-$ and R_{qd}^-, and integrate with respect to $\boldsymbol{\theta}$. We find

$$p(r_{qi} = d \mid \boldsymbol{c}, \boldsymbol{r}_q^-, \boldsymbol{k}_0, \boldsymbol{\lambda}_0, \mathbf{y})$$

$$\begin{cases} \propto (\pi_1 \frac{n_{qd}^-}{\beta + m_q^-} \prod_{g \in V_q} \frac{1}{y_{ig}!} \prod_{g \in V_q} \frac{\Gamma(k_{0g} + \sum_{l \in R_{qd}^- \bigcup \{i\}} y_{lg})}{\Gamma(k_{0g} + \sum_{l \in R_{qd}^-} y_{lg})} \\ \qquad \times \frac{(\lambda_{0g} + n_{qd}^-)^{k_{0g} + \sum_{l \in R_{qd}^-} y_{lg}}}{(\lambda_{0g} + n_{qd}^- + 1)^{k_{0g} + \sum_{l \in R_{qd}^- \bigcup \{i\}} y_{lg}}}) & \text{if } d = 1, \ldots, D_q^-, \\ \propto \pi_1 \frac{\beta}{\beta + m_q^-} \prod_{g \in V_q} \frac{\Gamma(k_{0g} + y_{ig})}{\Gamma(k_{0g})} \frac{(\lambda_{0g})^{k_{0g}}}{(\lambda_{0g} + 1)^{k_{0g} + y_{ig}} y_{ig}!} & \text{if } d = D^- + 1, \\ \propto (1 - \pi_1) \prod_{g \in V_q} \frac{\Gamma(k_{1g} + y_{ig})}{\Gamma(k_{1g})} \frac{(\lambda_{1g})^{k_{1g}}}{(\lambda_{1g} + 1)^{k_{1g} + y_{ig}} y_{ig}!} & \text{if } d = 0. \end{cases}$$

(v) Update $\boldsymbol{c}$

Remove c_g, and define $G^{'-}, Q^-, p^-, \boldsymbol{c}^-$ and n^-. Sample c_g as follows: $c_g \in \{0, 1, , \ldots, Q^-, (Q^- + 1)\}$. Note that $c_g = 0$ implies becoming idle, $1 \leq c_g \leq Q^-$ joining one of the existing HM sets, and $c_g = Q^- + 1$ starting a new singleton HM set.

$$p(c_g = q \mid \mathbf{y}_g) \propto \begin{cases} p(c_g = 0) p(\mathbf{y}_g \mid c_g = 0) & q = 0 \\ p(c_g = q) p(\mathbf{y}_g \mid \boldsymbol{c}_s) & q = 1, \ldots, Q^- \\ p(c_g = Q^- + 1) \sum_{\boldsymbol{r}} p(\boldsymbol{r} \mid c_g = Q^- + 1) p(\mathbf{y}_g \mid \boldsymbol{r}) & q = Q^- + 1. \end{cases}$$

The marginalization of $\boldsymbol{r}$ is difficult and computationally intensive. To avoid this problem, we consider a pseudo prior $p(\tilde{\boldsymbol{r}_g} \mid \mathbf{y}_g)$ and let $\boldsymbol{r}_{Q^-+1} = \tilde{\boldsymbol{r}_g}$. Finally, after canceling $\prod_{g'=1}^{G} p(\tilde{\boldsymbol{r}_g} \mid \mathbf{y}_g)$, we have the following:

$$p(c_g = q \mid \mathbf{y}_g) \propto \begin{cases} p(c_g = 0) p(\mathbf{y}_g \mid c_g = 0) & q = 0 \\ p(c_g = q) p(\mathbf{y}_g \mid \boldsymbol{c}_s) & q = 1, \ldots, Q^- \\ p(c_g = Q^- + 1) p(\boldsymbol{r}_{Q^-+1}) p(\mathbf{y}_g \mid \boldsymbol{r}_{Q^-+1}) & q = Q^- + 1. \end{cases}$$

For joining an existing cluster, $q = 1, \ldots, Q^-$,

$$p(c_g = q \mid \boldsymbol{c}^-, \text{rest}) \propto \pi_0 \frac{p_q^-}{\alpha + G'^-} \times \prod_{d=1}^{D_q} \left[\frac{\lambda_{0g}^{k_{0g}}}{\Gamma(k_{0g})} \frac{\Gamma(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}{(n_{qd} + \lambda_{0g})^{(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}} \prod_{i \in R_{qd}} \frac{1}{y_{ig}!} \right] \times \prod_{i \in R_{q0}} \frac{\Gamma(k_{1g} + y_{ig})}{\Gamma(k_{1g})} \frac{(\lambda_{1g})^{k_{1g}}}{(\lambda_{1g} + 1)^{k_{1g} + y_{ig}} y_{ig}!},$$

for starting a new (singleton) cluster

$$p(c_g = Q^- + 1 \mid \boldsymbol{c}^-, \text{rest}) \propto \pi_0 \frac{\alpha}{\alpha + G'^-} \pi_1^{m_q} (1 - \pi_1)^{(N - m_q)} \frac{\beta^{D_q} \prod_{d=1}^{D_q} \Gamma(n_{qd})}{\prod_{i=1}^{m_q} (\beta + i - 1)} \times \prod_{d=1}^{D_q} \left[\frac{\lambda_{0g}^{k_{0g}}}{\Gamma(k_{0g})} \frac{\Gamma(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}{(n_{qd} + \lambda_{0g})^{(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}} \prod_{i \in R_{qd}} \frac{1}{y_{ig}!} \right] \times \prod_{i \in R_{q0}}^{N} \frac{\Gamma(k_{1g} + y_{ig})}{\Gamma(k_{1g})} \frac{(\lambda_{1g})^{k_{1g}}}{(\lambda_{1g} + 1)^{k_{1g} + y_{ig}} y_{ig}!},$$

and for joining the inactive cluster

$$p(c_g = 0 \mid \boldsymbol{c}^-, \text{rest}) \propto (1 - \pi_0) \prod_{i=1}^{N} \frac{\Gamma(k_{2g} + y_{ig})}{\Gamma(k_{2g})} \frac{(\lambda_{2g})^{k_{2g}}}{(\lambda_{2g} + 1)^{k_{2g} + y_{ig}} y_{ig}!}.$$