DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Avenue
Madison, WI 53706

TECHNICAL REPORT NO. 1151

January 12, 2009

# A Hierarchical Semi-Markov Model for Detecting Enrichment with Application to ChIP-Seq Experiments

Pei Fen Kuan
Department of Statistics,
University of Wisconsin, Madison, WI 53706.

Guangjin Pan
Genome Center of Wisconsin, Madison, WI 53706.

James A. Thomson
Morgridge Institute for Research, Madison, WI 53707.
School of Medicine and Public Health,
University of Wisconsin, Madison, WI 53706.

Ron Stewart
Morgridge Institute for Research, Madison, WI 53707.

Sündüz Keleş
Department of Statistics,
Department of Biostatistics and Medical Informatics,
University of Wisconsin, Madison, WI 53706.

# A Hierarchical Semi-Markov Model for Detecting Enrichment with Application to ChIP-Seq Experiments

Pei Fen Kuan, Guangjin Pan, Ron Stewart and Sündüz Keleş

**Abstract**

Chromatin immunoprecipitation followed by direct sequencing (ChIP-Seq) has revolutionalized the experiments in profiling DNA-protein interactions and chromatin remodeling patterns. However, limited statistical tools are available for modeling and analyzing the ChIP-Seq data thoroughly. We carefully study the data generating mechanism of ChIP-Seq data and propose a new model-based approach for detecting enriched regions. Our model is based on a hierarchical mixture model which gives rise to a zero-inflated negative binomial (ZINB), coupled with a hidden semi-Markov model (HSMM) to address the sequencing depth and biases, the inherent spatial data structure and allows for detection of multiple non-overlapping variable size peaks. In particular, we demonstrate that the proposed ZINB accounts for the excess zeroes and over-dispersion in the observed data relative to a Poisson distribution, and this model provides a better fit as the background distribution. We also propose a new meta false discovery rate (FDR) control at peak level as an alternative to the usual heuristic postprocessing of enriched bins identified via bin level FDR control. We show with simulations and case studies that this new procedure allows for the boundaries of peak regions to be declared probabilistically and provides accurate FDR control.

KEYWORDS: Hidden semi-Markov model; Hierarchical mixture model; Zero-inflated Poisson; Zero-inflated negative binomial; False discovery rate; ChIP-Seq; ChIP-chip.

# 1 Introduction

The study of protein-DNA interactions is important in molecular biology to understand its implication in gene regulation. In recent years, significant progress has been made in profiling transcription factor binding sites and histone modifications using chromatin immunoprecipitation (ChIP) techniques (Mikkelsen et al., 2007; Robertson et al., 2007). Such measurements are important for systems-level studies as they provide a global map of candidate gene network input connections. The ChIP experiments are usually followed by a microarray hybridization experiment (ChIP-chip) for quantifying different binding or chromatin modification activities. Although the ChIP-chip experiments have been successfully used to interrogate different genomes, there are some limitations of this technology, especially in studying the mammalian genome (Mikkelsen et al., 2007; Barski et al., 2007). Large number of arrays are required to cover the mammalian genome and current array designs for large genomes usually have lower resolution. The ChIP-chip experiments require large amounts of DNA, thus involve extensive amplifications and could potentially introduce bias. In addition, the probes are typically subjected to cross-hybridization which hinders the study of repeated sequences and allelic variants (Mikkelsen et al., 2007; Wei et al., 2008).

More recently, a new technology has been developed to directly sequence the fragments (ChIP-Seq) and offers whole-genome coverage at a lower cost. While ChIP-Seq technologies are currently evolving, most of the published work in ChIP-Seq are conducted via the Solexa/Illumina platform (Mikkelsen et al., 2007; Barski et al., 2007). This high throughput sequencing technology works by sequencing one end of each fragment ($\sim 25-36$ bps) in the ChIP sample, thereby generate millions of short reads/tags. These tags are then mapped to the reference genome, followed by summarizing the total tag counts within specified genomic windows, i.e., bins and analysis to detect enriched/bound regions, i.e., peaks. Although this technology offers promising results for surveying large genomes at high resolution, there are limited statistical tools to detect enriched regions. In contrast, numerous model-based approaches are available for the analysis of ChIP-chip data (Ji and Wong, 2005; Keleş, 2007; Gottardo et al., 2008). In addition, published statistical methodologies mainly control the false positive or false discovery rates (FDR) at probe/bin level and rely on heuristic post-processing to merge contiguous probes/bins declared to be statistically significant as a peak.

Our goal in this paper is to develop a comprehensive statistical model for detecting enriched regions in ChIP-Seq data via a hierachical semi-Markov model. By studying the underlying data generating process, our modeling framework incorporates the variability in the sequencing depths and other source of biases. We investigate numerous candidate models for ChIP-Seq data and provide the merits/demerits of each model analytically and empirically. One of the major drawbacks with the current ChIP-Seq data analysis is the absence of control experiments in detecting enriched regions to reduce experimental costs. To allow for broader applicability of our modeling framework, we develop a model which allows for (1) identifying enriched regions in one sample ChIP-Seq, (2) identifying enriched regions in the presence of input, and (3) detecting differential enrichments between two samples. We also introduce a new meta approach for controlling FDR at peak level, which allows for the boundaries of binding sites to be declared probabilistically. We demonstrate the advantages

of this new procedure over FDR control at bin level in both simulations and case studies. Although our discussion is dedicated to ChIP-Seq experiments measuring transcription factor binding and histone modifications, the proposed hierarchical semi-Markov model is a general framework that can be applied to other types of data (e.g., ChIP-chip) which exhibit spatial structure, by modifying the observed emission distributions.

# 2 A hierarchical model for ChIP-Seq data

In ChIP-Seq experiments measuring both the transcription factor binding and histone modification, enrichment due to specific binding/modification site is detected by a cluster of bins mapping in the vicinity of the binding sites on the genome. This spatial data structure is an important characteristic of ChIP-Seq (likewise ChIP-Chip) data, of which we will account for via a hidden semi-Markov model (HSMM) in Section 3. We first investigate the distribution of the observed ChIP-Seq tags mapped to specified genomic windows/bins. The total mappable reads/tags sequenced from an experiment is also known as the sequencing depth of the experiment.

In a typical ChIP-Seq experiment, the probability that a bin is sequenced is affected by numerous factors. The most important determinant is the latent state of the bin, reflecting whether or not fragments mapping to that bin are enriched. Fragments mapping to enriched bins are over represented in the sample and almost surely being sequenced, resulting in high tag counts. On the other hand, a non-enriched bin has a small probability of being sequenced, since the DNA fragments corresponding to these regions are under represented/absent in the sample. The sequencing affinity of a bin is also influenced by non-specific immunoprecipitation and amplification biases, both of which are related to the underlying sequence composition of the DNA fragments. This consideration gives rise to bin specific distributions that account for the non-uniform genomic background as shown in Zhang et al. (2008). They attributed the substantial variations in tag distribution on the genomic background to preferential sequencing specific to the sequencing platform or protocol.

An important factor that is usually ignored in modeling the background/non-enriched distribution is the sequencing depth (total mappable reads) of the experiment that affects the overall genome coverage. That is, bins exhibiting zero tag counts are a consequence of non-enrichment and non-coverage due to insufficient sequencing depth, especially for large genomes. In Sections 2.1 and 2.2, we provide detailed formulation of our modeling framework for the observed ChIP-Seq data that accounts for (1) bin specific distributions and (2) sequencing depth of the experiment for one sample and two sample problems separately.

## 2.1 One sample problem

A natural choice to model the observed tag counts is a Poisson distribution. However, insufficient sequencing depth results in ChIP-Seq data having excess bins with zero counts compared to a Poisson distribution. Both the bin specific distribution and excess zeroes result

in over-dispersion relative to a Poisson model for the observed tag counts. To motivate this idea, we consider a subset of the data set measuring SMAD2/3 binding activities in embryonic stem cells (ES) from the Thomson Lab, University of Wisconsin-Madison and the Morgridge Institute for Research, Madison, WI. This subset consists of tag counts summarized at bins of size 100 bps generated from 3 lanes on the same Illumina-Solexa machine in a single run, labeled as replicates 1 to 3. These 3 replicates have comparable sequencing depths ($2.18M$, $2.22M$, $2.33M$) and equal concentration ($3pM$) of sample materials loaded to the machine. Figure 1 plots the mean and variance of the tag counts for each bin for Chromosomes 19 and 21, respectively. Since the bin specific means and variances are approximately equal and the mean varies for different bins, this supports the bin specific Poisson distribution to characterize the tag counts. In addition, a substantial proportion of the bins has zero counts across all the 3 replicates, which suggest the use of an indicator variable to model the excess zeroes.

Let $Y_j$ denote the observed tag counts for bin $j$ (e.g., summarization based on tag shifting by MACS (Zhang et al., 2008)), and $Z_j$ be the unobserved random variable specifying if bin $j$ comes from enriched ($Z_j = 1$) or non-enriched ($Z_j = 0$) distribution. Let $D_Y$ be the sequencing depth for the treatment sample. Based on the observations above, we consider several candidate models for the non enriched bins ($Z_j = 0$) to address (1) over-dispersion, (2) excess zeroes, and (3) bin specific distributions:

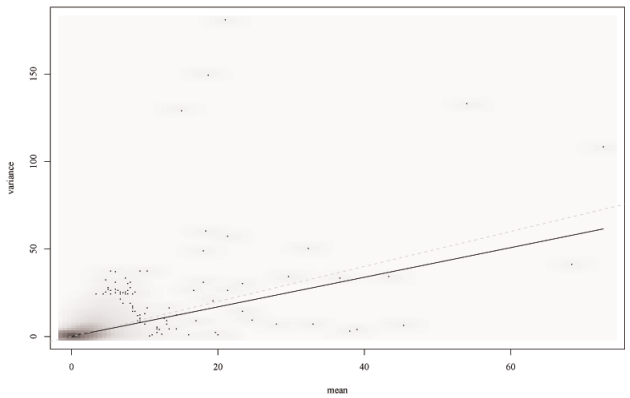
1. Model 1 (Poisson)
   $Y_j = N_j(D_Y)$, where $N_j(D_Y) \sim Po(\lambda D_Y)$.

2. Model 2 (Zero Inflated Poisson, ZIP)
   $Y_j = N_j(D_Y)I(B_j(D_Y) = 1)$ where $N_j(D_Y) \sim Po(\lambda(D_Y))$ and $B_j(D_Y) \sim Be(p(D_Y))$.

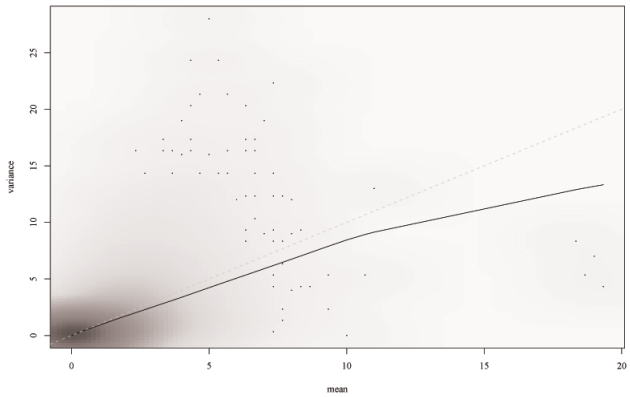
3. Model 3 (Negative Binomial)
   $Y_j = N_j(D_Y)$ where $N_j(D_Y) \sim Po(\lambda_j(D_Y))$ and $\lambda_j(D_Y) \sim \mathcal{G}a(a_0, b)$.

4. Model 4 (Zero Inflated Negative Binomial, ZINB)
   $Y_j = N_j(D_Y)I(B_j(D_Y) = 1)$ where $N_j(D_Y) \sim Po(\lambda_j(D_Y))$, $\lambda_j(D_Y) \sim \mathcal{G}a(a_0, b)$, $B_j(D_Y) \sim Be(p_j(D_Y))$ and $p_j(D_Y) \sim \mathcal{B}eta(1, \beta(D_Y))$.

Models 1 and 3 have been considered in Robertson et al. (2007) and Ji et al. (2008), respectively. MACS (Zhang et al., 2008) considered a variant of Model 3 with $\lambda_j$ estimated from $\max(\lambda_{BG}, \lambda_{5k}, \lambda_{10k})$ which could potentially result in higher false negatives if a peak spans a large region, such as in histone modifications. Here $N_j(D_Y)$ measures non-specific sequencing which is usually attributed to sequence and amplification bias. Non-specific sequencing could result in tags mapping to multiple genomic regions and such tags are usually omitted in summarizing total tag counts in each bin. $B_j(D_Y)$ indicates if bin $j$ is sequenced and it depends on the sequencing depths. Without loss of generality, we assume that $D_Y = 1M$. Model 1 assumes equivalence of mean and variance whereas Model 2 accounts for over-dispersion.




(a) Chromosome 19




(b) Chromosome 21

Figure 1: *Mean/variance relationship by bin for Chromosomes 19 and 21.* The mean and variance are computed using 3 technical replicates measuring ES SMAD 2/3 binding sites. The black solid line is the lowess fit.

Under Model 2,

$$\begin{aligned} E(Y_j|Z_j=0) &= p\lambda, \\ \mathrm{Var}(Y_j|Z_j=0) &= p\lambda[1+\lambda(1-p)], \\ \Rightarrow \frac{\mathrm{Var}(Y_j|Z_j=0)}{E(Y_j|Z_j=0)} &\geq 1. \end{aligned}$$

Therefore, the presence of excess zeroes results in an over-dispersion relative to a Poisson model, known as a zero inflated Poisson (ZIP) model (Lambert, 1992). Both Model 1 and Model 2 assume common bin distribution. Model 4 is analogous to Model 2 but it allows for bin specific distributions, with $p$ and $\lambda$ replaced by $p_j$ and $\lambda_j$ respectively. The priors governing bin specific distributions in Model 4 are based on the following justifications. In a ChIP-Seq experiment, the tags counts over different lanes/runs are usually pooled to increase the sequencing depths instead of treating them as individual replicates, unless these replicates exhibit other sources of variations (e.g., different machines, concentration, run day). This suggests that one typically has a single observation to estimate $\lambda_j$ and $p_j$. Therefore, we introduce conjugate priors which allow for information sharing across bins,

$$p_j \sim \mathcal{B}eta(1,\beta) \quad , \quad \lambda_j \sim \mathcal{G}a(a_0,b),$$

Then, we have

$$\begin{aligned} P(B_j=z) &= P(\text{Bernoulli}(1/(1+\beta))=z) \text{ for } z\in\{0,1\}, \\ P(N_j=y) &= P(\mathcal{NB}(a_0,b)=y). \end{aligned}$$

The marginal density for the observed counts for a non-enriched ($Z=0$) bin is given by:

$$\begin{aligned} P(Y_j=0|Z_j=0) &= P(B_j=0)+P(B_j=1,N_j=0) \\ &= \frac{\beta}{1+\beta}+\frac{1}{1+\beta}P(\mathcal{NB}(a_0,b)=0) \\ &= \frac{\beta}{1+\beta}+\frac{1}{1+\beta}\left(\frac{b}{b+1}\right)^{a_0}, \\ P(Y_j=y|Z_j=0) &= P(B_j=1)P(N_j=y) \\ &= \frac{1}{1+\beta}\frac{\Gamma(y+a_0)b^{a_0}}{\Gamma(a_0)(b+1)^{y+a_0}y!} \text{ for } y>0. \end{aligned}$$

When the sequencing depth is sufficient with $\beta=0$, the model reduces to a negative binomial distribution with parameters $a_0$ and $b$ (Model 3). Therefore the proposed model offers greater flexibility than a regular negative binomial model. Both Model 3 and Model 4 are derived in a hierarchical setting to account for bin specific biases.

To assess the goodness of fit of Models 1-4, we fit the each model on the tag counts sequenced from naked DNA (histone-free DNA), which is a control sample and does not contain any enrichment. Therefore, the variability and excess zeroes in the observed tag

counts can be attributed to sequencing biases and sequencing depth. In addition, we also evaluate these models in characterizing the background/non-enriched distribution on a ChIP sample from a publicly available ChIP-Seq data measuring Ezh2 binding (Ku et al., 2008) using the bottom 99% (an estimate of the non-enriched bins) of the data. The unknown parameters of Models 1 and 3 are obtained via maximum likelihood estimation (MLE) or methods of moments estimation (MME). Although MLE and MME estimation for Models 1 and 3 are straight forward, Models 2 and 3 require an EM algorithm to handle unobserved data.

Figure 2 plots the distribution of the actual data against the simulated data of each model using the estimated parameters and the BIC score of each model is displayed on the main title of the corresponding plot. Model 4 appears to fit the data best (lowest BIC score) in both the naked DNA control sample and the ChIP sample. An interesting extension of the proposed hierarchical modeling framework would be to model $\log \lambda_j = X\beta$ and $\log\left(\frac{p_j}{1-p_j}\right) = M\alpha$, where $X$ and $M$ are the covariate matrices (e.g., functions of the sequence compositions) which explain the sequencing biases. We are currently investigating the characteristics (e.g., function of nucleotides) contributing to sequencing biases from naked DNA control experiment.

Given that the background (non-enriched) distribution is best characterized by a ZINB, we next consider analogous model for the enriched bins $Z_j = 1$ to account for bin specific distribution via a hierarchical model. Since the DNA fragments for enriched regions are highly represented in the ChIP sample, the probability of an enriched bin being sequenced can be assumed to be 1. The potential models are:

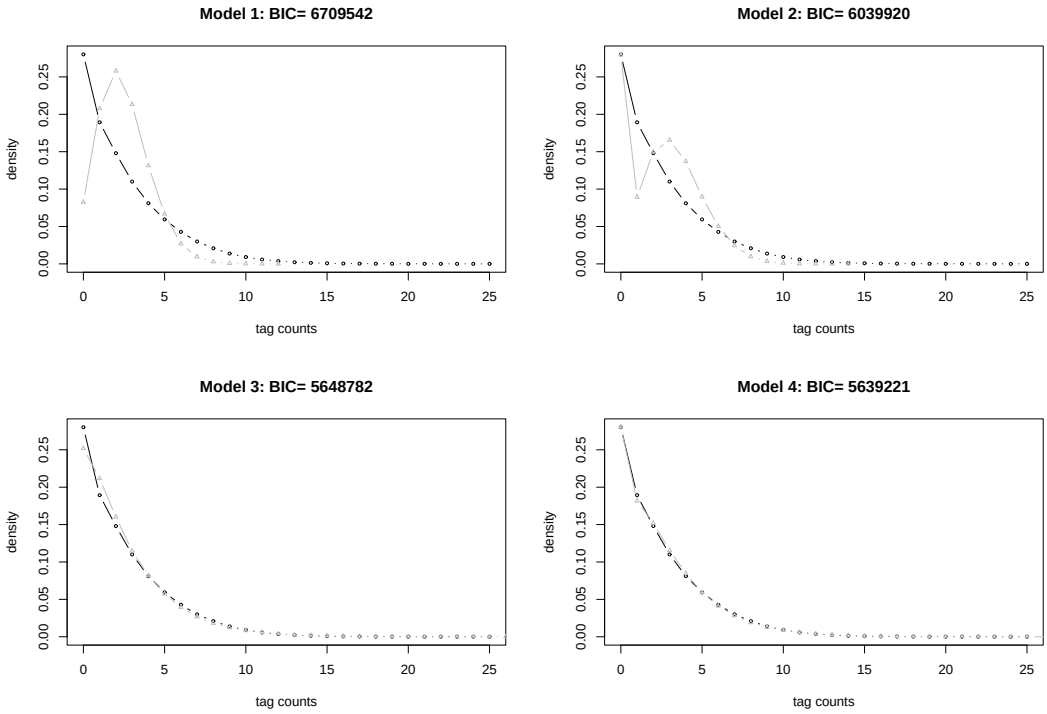
1. Model 1a
   $Y_j = S_j$ where $S_j \sim Po(\gamma_j)$, $\gamma_j \sim \mathcal{G}a(a_2, b_2)$, under the constraint $E(Y|Z = 1) \geq E(Y|Z = 0)$.

2. Model 2a
   $Y_j = N_j + S_j$ where $N_j \sim Po(\lambda_j)$, $\lambda_j \sim \mathcal{G}a(a_0, b)$, $S_j \sim Po(\gamma_j)$, $\gamma_j \sim \mathcal{G}a(a_1, b)$.

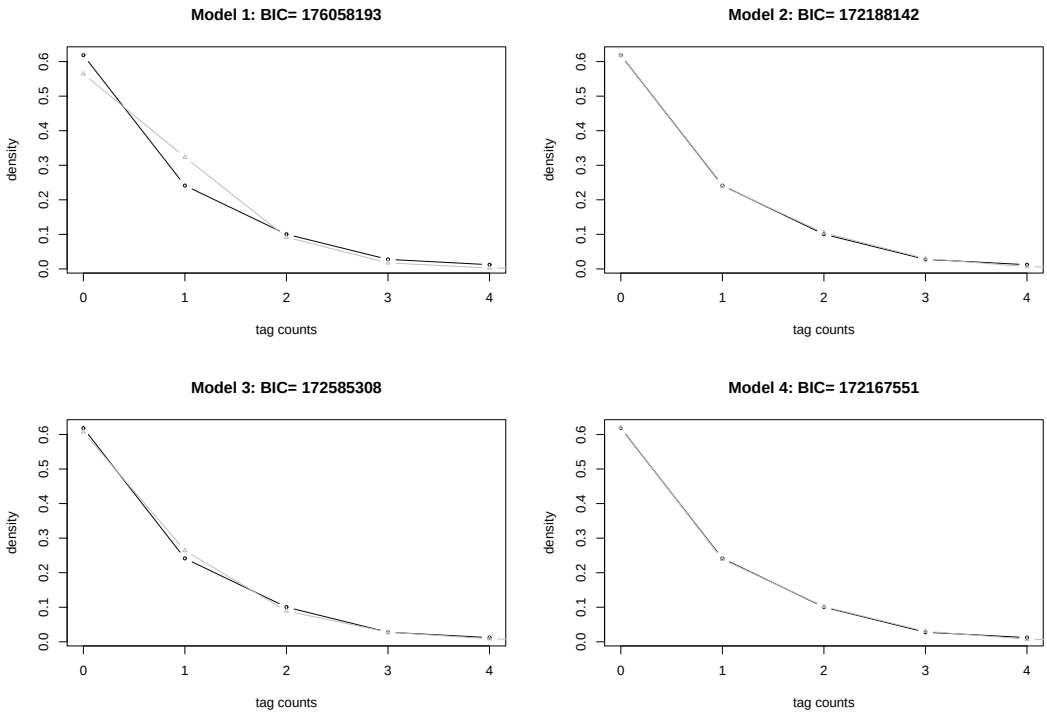
Both models assume bin specific distribution and variable enrichment levels. Model 2a is nested in Model 1a with the restriction

$$\frac{\text{Var}(Y|Z = 0)}{E(Y|Z = 0)} = \frac{\text{Var}(Y|Z = 1)}{E(Y|Z = 1)}.$$

Although Model 1a appears to offer more flexibility, it does not guarantee that $P(Y_j = y|Z_j = 0) \leq P(Y_j = y|Z_j = 1)$, $\forall y \geq y^*$, where $y^*$ is a sufficiently large tag count number. On the other hand, Model 2a explicitly assumes that the tag counts for an enriched bin is contributed by non-specific sequencing bias ($N_j$) and the actual level of enrichment ($S_j$), and guarantees that $P(Y_j = y|Z_j = 0) \leq P(Y_j = y|Z_j = 1)$, $\forall y \geq y^*$. Therefore, we choose Model 2a to characterize the distribution of an enriched bin. Under this model, $\lambda_j + \gamma_j \sim \mathcal{G}a(a_0 + a_1, b)$ and $N_j + S_j \sim Po(\lambda_j + \gamma_j)$. The marginal density for the observed




(a) Naked DNA




(b) Ezh2 ChIP-Seq from Ku et al. (2008)

Figure 2: *Goodness of fit of Models 1-4.* Black lines are the density of the actual data. Gray lines are the density for simulated data from each model. The BIC score for each model is given in the header.

counts in an enriched ($Z = 1$) bin is given by:

$$\begin{aligned} P(Y_j = y|Z_j = 1) &= P(N_j + S_j = y) \\ &= \frac{\Gamma(y + a_0 + a_1)b^{a_0+a_1}}{\Gamma(a_0 + a_1)(b + 1)^{y+a_0+a_1}y!} \text{ for } y \geq 0. \end{aligned}$$

We derive an EM algorithm for fitting this hierarchical mixture distribution characterized by Model 4 for non-enriched bins and Model 2a for enriched bins. The details of the algorithm are provided in Appendix A.1.

## 2.2 Two sample problem

Two sample problem can arise in ChIP-Seq experiments in two different ways. The first is comparison of a chromatin-immunoprecipitated (ChIPed) sample with a control sample. Valouev et al. (2008) observed an under-representation of coverage in AT-rich regions of the genome in their data. They attributed the inefficient sequencing in these genomic regions to the lower melting temperature and showed that such sequencing biases were reduced by normalizing against data from a control experiment. A second reason for two sample comparison is that a relative comparison between two samples to detect differential enrichments could reduce the sequencing biases. We introduce a modeling framework for inferring (1) enriched regions relative to a control experiment (e.g., total genomic DNA) or (2) comparing differential enrichment between two ChIP samples. We first consider case (1) and let $(Y_j, X_j)$ be the observed sample 1 (treatment) and sample 2 (control) tag counts for bin $j$. Similarly, we define $Z_j$ to be the unobserved random variable specifying the underlying latent state of bin $j$. In one sample ChIP-Seq, the ZINB model that arises from a hierarchical setting appears to fit the observed data well. Thus, we consider possible extensions of this model to account for both the bin specific bias and excess zeroes due to insufficient sequencing depth within the two sample context. Let $D_X$ and $D_Y$ denote the sequencing depths of control and treatment experiments, respectively. Most of the current approaches in the analysis of two sample ChIP-Seq apply linear scaling to the observed tag counts to normalize for the difference in sequencing depths. This is undesirable since the distribution of the scaled tag counts is different from the original distribution under the Poisson assumption. More formally, if $Y = Po(\lambda)$, then $cY$ is no longer distributed as Poisson since $E(cY) \neq \text{Var}(cY)$. Another popular strategy is to randomly sample $D_X$ counts from $Y$ (assuming $D_Y > D_X$). This is again undesirable, since the non-uniform background (Zhang et al., 2008) indicates that random sampling is inappropriate. Moreover, using only a fraction of the original data results in some information loss. Thus, we propose a model that appropriately incorporates the sequencing depths of the two samples.

We introduce Bernoulli random variables $B_{j1}$ and $B_{j2}$ to denote if bin $j$ is sequenced under control and treatment experiments, respectively. These random variables will again be utilized to account for the excess zeroes in the observed data. For ease of exposition, we first assume $B_{j1} = B_{j2} = 1$. Let $\lambda_{j1}$ and $\lambda_{j2}$ denote the bin specific latent mean tag counts for $X_j$ and $Y_j$. We assume that $X_j$ and $Y_j$ are random samples from $p_X(.|\lambda_{j1}) = Po(\lambda_{j1}D_X)$ and $p_Y(.|\lambda_{j2}) = Po(\lambda_{j2}D_Y)$ respectively, and

$$\lambda_{j1} = \lambda_{j2} \text{ if } Z_j = 0,$$
$$\lambda_{j1} < \lambda_{j2} \text{ if } Z_j = 1.$$

As in Newton et al. (2004) and Keleş (2007), we assume the latent mean counts $(\lambda_{j1}, \lambda_{j2})$ to be a random pair from an unknown bivariate distribution $f$, which is taken to be a mixture over the two hypotheses of interest:

$$f(\lambda_{j1}, \lambda_{j2}) \;=\; P(Z_j = 0) f_0(\lambda_{j1}, \lambda_{j2}) + P(Z_j = 1) f_1(\lambda_{j1}, \lambda_{j2}),$$

where the densities $f_0$ and $f_1$ describe the fluctuations of the means within each hypothesis. The joint distribution of $\lambda_{j1}$ and $\lambda_{j2}$ is related to a one-dimensional base distribution $\pi$ so that the unknown components are estimable. In addition, we observe that the tag counts for the control and treatment sample in the real ChIP-Seq data (see case studies) exhibit significant correlation. An advantage of this hierarchical mixture modeling approach is that it automatically incorporates the correlation between $X$ and $Y$ via $\lambda_{j1}$ and $\lambda_{j2}$ based on the following data generating process:

1. Draw $Z_j \sim Be(p_0)$.
2. If $Z_j = 0$, draw $\lambda_{j1}$ from $\pi$ and $X_j \sim Po(\lambda_{j1} D_X)$ and $Y_j \sim Po(\lambda_{j2} D_Y)$.
3. If $Z_j = 1$, draw $\theta_{j1}$, $\theta_{j2}$ from $\pi$. Set $\lambda_{j1} = \min(\theta_{j1}, \theta_{j2})$ and $\lambda_{j2} = \max(\theta_{j1}, \theta_{j2})$. Draw $X_j \sim Po(\lambda_{j1} D_X)$ and $Y_j \sim Po(\lambda_{j2} D_Y)$.

We will now consider two different modeling approaches for the observed tag counts to identify enriched regions in $Y$ relative to $X$. The first approach is to model the bivariate distribution of $(Y_j, X_j)$ jointly via a mixture model. According to the data generation process described above, the mixture distribution $f$ and $\pi$ are related as follows:

$$f_0(\lambda_{j1}, \lambda_{j2}) = \pi(\lambda_{j1}) \quad \text{and} \quad f_1(\lambda_{j1}, \lambda_{j2}) = 2\pi(\lambda_{j1})\pi(\lambda_{j2}) I[\lambda_{j1} < \lambda_{j2}].$$

We take $\pi = \mathcal{G}a(a, 1/b)$ because of the conjugacy property of Poisson-Gamma which makes the computations analytically tractable. Given the hierarchical modeling framework, the marginal density of the data can be derived as follows. For notation brevity, we drop the subscript $j$. Then

$$P(X, Y) \;=\; P(Z = 0) P(X, Y | Z = 0) + P(Z = 1) P(X, Y | Z = 1),$$

and

$$g_0^{(1)} \;=\; P(X, Y | Z = 0) = \binom{X + Y + a - 1}{X, Y, a - 1} \left( \frac{b}{b + D_X + D_Y} \right)^a \frac{D_X^X D_Y^Y}{(b + D_X + D_Y)^{X+Y}},$$
$$g_1^{(1)} \;=\; P(X, Y | Z = 1) = 2\mathcal{NB}_X(a, b/D_X)\mathcal{NB}_Y(a, b/D_Y) P\left( B < \frac{b + D_X}{2b + D_X + D_Y} \right).$$

where $B \sim \mathcal{B}eta(X + a, Y + a)$. The details on this derivation are given in Appendix A.2.1.

An alternative to modeling the joint distribution of $(X, Y)$ is to model the distribution of $Y$ conditioned on $X + Y$ and $Z$, since $X$ and $Y$ are distributed as Poisson conditional on $Z$. Under the same data generating mechanism described above, we have

$$\begin{aligned} g_0^{(2)} &= P(Y|X+Y, Z=0) = \left( \begin{array}{c} X+Y \\ Y \end{array} \right) \left( \frac{D_Y}{D_X + D_Y} \right)^Y \left( \frac{D_X}{D_X + D_Y} \right)^X, \\ g_1^{(2)} &= P(Y|X+Y, Z=1) = \left( \begin{array}{c} X+Y \\ Y \end{array} \right) \frac{\int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{Y+a-1}(1-v)^{X+a-1}}{\left[1+\frac{b}{D_Y}v+\frac{b}{D_X}(1-v)\right]^{X+Y+2a}} dv}{\int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{a-1}(1-v)^{a-1}}{\left[1+\frac{b}{D_Y}v+\frac{b}{D_X}(1-v)\right])^{X+Y+2a}} dv}, \end{aligned}$$

as given in Appendix A.2.2. We investigate the power of these two proposed formulations for two sample ChIP-Seq in discriminating $Z = 1$ from $Z = 0$. In both models, $g_1^{(k)}/g_0^{(k)}$ is an increasing function of $Y$ for a fixed value of $X$, which is desirable since it is easier to discriminate enriched from non-enriched bin as the difference between $X$ and $Y$ increases. Next, we define the following quantities for fixed values of $X = x$, $a$, and $b$ :

$$\begin{aligned} y^{*(k)} &= \text{argmin}_Y \{g_1^{(k)}/g_0^{(k)} > 1\}, \\ R(x)^{(k)} &= y^{*(k)}/x. \end{aligned}$$

The quantity $R(x)^{(k)}$ can be interpreted as an analog of the minimum fold change in microarray data analysis such that the probability of the observed tag counts under an enriched bin is greater than that of a non-enriched bin. Figure 3 provides examples on the behavior of $R(x)^{(k)}$ as a function of $x$ for two arbitrary chosen values of $a$. In general, $R(x)^{(1)}$ exhibits increasing trend with $x$, while $R(x)^{(2)}$ exhibits decreasing trend with $x$ for fixed $a$ and $b$. In other words, for $k = 1$ where we model the joint distribution of $(X, Y)$, for larger $X$, $Y$ has to be a few times larger for a bin to be called enriched. On the other hand, for $k = 2$, the reverse holds, which is perhaps more desirable if $X$ is the genomic DNA input. We reason this as follows based on the observation that the profile for control and treatment sample in two sample ChIP-Seq data are highly correlated:

1. Genomic/chromatin DNA which is commonly used as control input in ChIP experiments differ from the treatment sample in that no antibody is added to immunoprecipitate the DNA fragments bound by DNA proteins. However, because of the cross-linking of protein to DNA, regions tightly bound by proteins are less likely to be sheared, compared to unbound regions. Therefore, DNA fragments corresponding to these regions are more abundant than fragments that are randomly sheared as observed in the ChIP-Seq data.

2. For a region with large tag counts in the control experiment, if the corresponding region in the treatment sample has higher counts, this suggests some degree of enrichment in the treatment sample and vice versa for regions with zero or small tag counts.

On the other hand, the first formulation may be more appropriate in cases in which high tag counts in the control sample are due to technical bias instead of the underlying chromatin

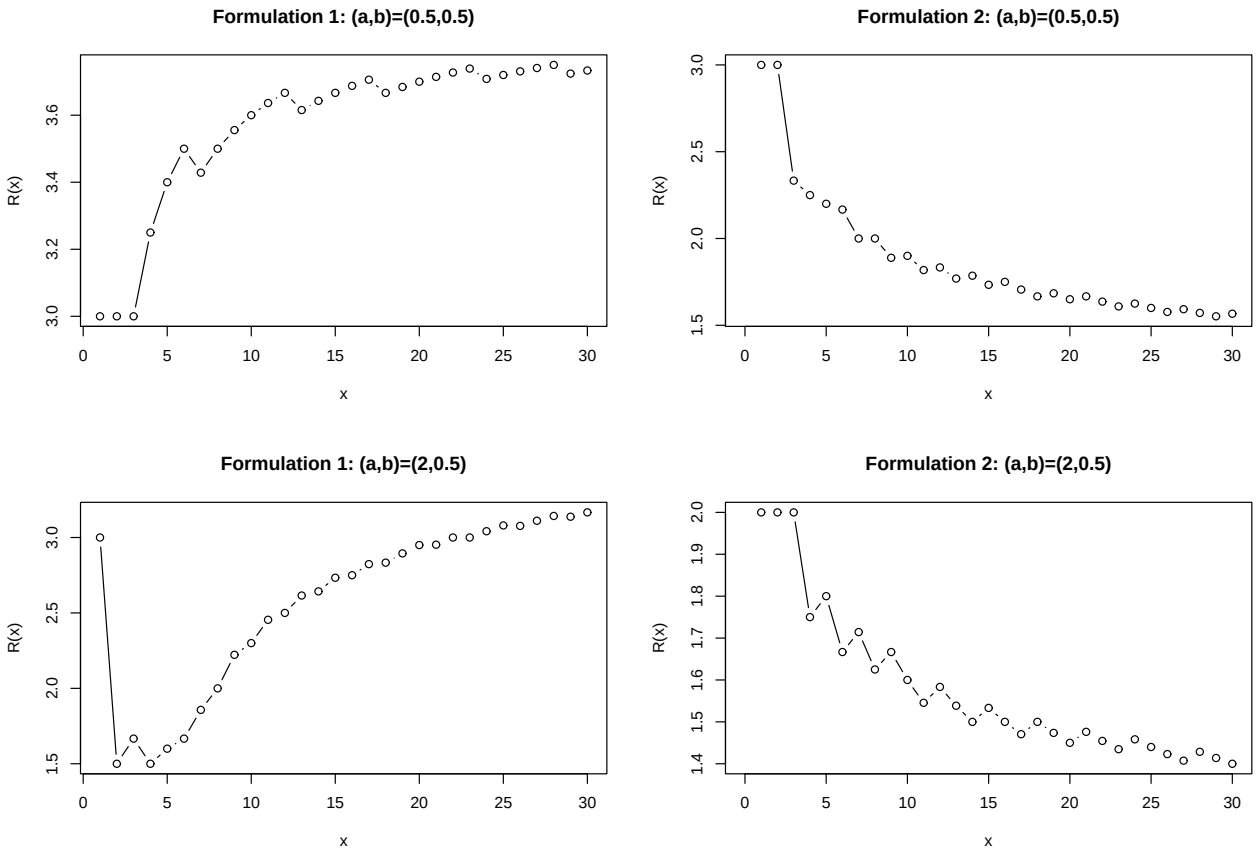



Figure 3: *Power comparisons of Formulation 1 (bivariate mixture modeling) and Formulation 2 (conditional mixture modeling).* We plot $R(x)^{(k)}$ against $x$ for two arbitrary values of $a$ for Formulation 1 ($k = 1$) and Formulation 2 ($k = 2$).

structure. In such cases, a much higher tag counts in the treatment sample is required to confidently declare a bin with high tag counts in control as enriched.

Next, we discuss the case in which $B_{j1}$ or $B_{j2} = 0$. Under the bivariate framework in Formulation 1, we consider the following complete data generating mechanism:

1. Draw $Z_j \sim Be(p_0)$.
2. If $Z_j = 0$, draw $p_{j1} \sim h_1$ and $B_{j1} \sim Be(p_{j1})$ and $\lambda_{j1}$ from $\pi$.
   (a) If $B_{j1} = 0$, set $Y_j = X_j = 0$.
   (b) If $B_{j1} = 1$, draw $p_{j2} \sim h_2$ and $B_{j2} \sim Be(p_{j2})$.
      i. If $B_{j2} = 0$, set $Y_j = 0$ and draw $X_j \sim Po(\lambda_{j1} D_X)$.
      ii. If $B_{j2} = 1$, set $\lambda_{j2} = \lambda_{j1}$ and draw $X_j \sim Po(\lambda_{j1} D_X)$ and $Y_j \sim Po(\lambda_{j2} D_Y)$.
3. If $Z_j = 1$, set $B_{j2} = 1$. Draw $p_{j1} \sim h_1$ and $B_{j1} \sim Be(p_{j1})$. Draw $\theta_{j1}$, $\theta_{j2}$ from $\pi$. Set $\lambda_{j1} = \min(\theta_{j1}, \theta_{j2})$ and $\lambda_{j2} = \max(\theta_{j1}, \theta_{j2})$.
   (a) If $B_{j1} = 0$, set $X_j = 0$ and draw $Y_j \sim Po(\theta_{j1} D_Y)$.
   (b) If $B_{j1} = 1$, draw $X_j \sim Po(\lambda_{j1} D_X)$ and $Y_j \sim Po(\lambda_{j2} D_Y)$.

We take $h_1 = \mathcal{B}eta(1, \beta_1)$ and $h_2 = \mathcal{B}eta(1, \beta_2)$ for the conjugacy properties of Poisson-Gamma and Bernoulli-Beta which makes the computations analytically tractable. Then

$$P(X, Y) = P(Z = 0)P(X, Y|Z = 0) + P(Z = 1)P(X, Y|Z = 1)$$

and

$$\begin{aligned} P(X, Y|Z = 0) &= I(X = 0)I(Y = 0)\frac{\beta_1}{1 + \beta_1} \\ &\quad + I(Y = 0)\mathcal{NB}_X(a, b/D_X)\frac{\beta_2}{(1 + \beta_1)(1 + \beta_2)} \\ &\quad + \binom{X + Y + a - 1}{X, Y, a - 1}\left(\frac{b}{b + D_X + D_Y}\right)^a \frac{D_X^X D_Y^Y}{(b + D_X + D_Y)^{X+Y}} \frac{1}{(1 + \beta_1)(1 + \beta_2)}, \end{aligned}$$

$$\begin{aligned} P(X, Y|Z = 1) &= I(X = 0)\mathcal{NB}_Y(a, b/D_Y)\frac{\beta_1}{1 + \beta_1} \\ &\quad + 2\mathcal{NB}_X(a, b/D_X)\mathcal{NB}_Y(a, b/D_Y)P\left(B < \frac{b + D_X}{2b + D_X + D_Y}\right)\frac{1}{1 + \beta_1}, \end{aligned}$$

where $B \sim \mathcal{B}eta(X + a, Y + a)$.

On the other hand, the conditional distribution of $Y$ given $X + Y$ does not have a closed form if we model $Y$ and $X$ as zero inflated Poisson. Therefore, we consider an alternative strategy. Ideally, if the control sample is the total genomic DNA, the number of tags in each bin is approximately equal to the number of DNA copy sequenced ($\geq 1$). We assume

that $X_j = 0$ is attributed to non-coverage due to insufficient sequencing depths and it does not contain information about the enrichment level of bin $j$. Therefore, the above model $(g_0^{(2)}, g_1^{(2)})$ is defined for $X_j \geq 1$, and for $X_j = 0$ we model the observed tag counts for the treatment sample $Y$ as in one sample ChIP-Seq.

If the interest is in comparing treatment 1 to treatment 2, the corresponding bin specific hypotheses of interest for bin $j$ are

$$\begin{array}{c} \lambda_{j1} = \lambda_{j2} \text{ if } Z_j = 0 \text{ (Non enriched)}, \\ \lambda_{j1} < \lambda_{j2} \text{ if } Z_j = 1 \text{ (Enriched in treatment 2)}, \\ \lambda_{j1} > \lambda_{j2} \text{ if } Z_j = 2 \text{ (Enriched in treatment 1)}, \end{array}$$

and the latent mean variables are distributed as

$$\begin{aligned} f(\lambda_{j1}, \lambda_{j2}) &= P(Z_j = 0)\pi(\lambda_{j1}) + 2P(Z_j = 1)\pi(\lambda_{j1})\pi(\lambda_{j2})I[\lambda_{j1} < \lambda_{j2}] \\ &\quad +2P(Z_j = 2)\pi(\lambda_{j1})\pi(\lambda_{j2})I[\lambda_{j1} > \lambda_{j2}]. \end{aligned}$$

The marginal distribution for $P(X, Y|Z = 1)$ under formulation 1 is similar to above, whereas

$$\begin{aligned} P(X, Y|Z = 0) &= I(X = 0)I(Y = 0)\left(1 - \frac{1}{(1+\beta_1)(1+\beta_2)}\right) \\ &\quad +\binom{X+Y+a-1}{X, Y, a-1}\left(\frac{b}{b+D_X+D_Y}\right)^a \frac{D_X^X D_Y^Y}{(b+D_X+D_Y)^{X+Y}} \frac{1}{(1+\beta_1)(1+\beta_2)}, \end{aligned}$$

and

$$\begin{aligned} P(X, Y|Z = 2) &= I(Y = 0)\mathcal{NB}_X(a, b/D_X)\frac{\beta_2}{1+\beta_2} \\ &\quad +2\mathcal{NB}_X(a, b/D_X)\mathcal{NB}_Y(a, b/D_Y)P\left(B > \frac{b+D_X}{2b+D_X+D_Y}\right)\frac{1}{1+\beta_2}. \end{aligned}$$

The conditional distributions for $g_0 = P(Y|X + Y, Z = 0)$ and $g_1 = P(Y|X + Y, Z = 1)$ under formulation 2 are similar to above, whereas

$$g_2 = P(Y|X+Y, Z=2) = \binom{X+Y}{Y} \frac{\int_0^{\frac{D_Y}{D_X+D_Y}} \frac{v^{Y+a-1}(1-v)^{X+a-1}}{\left[1+\frac{b}{D_Y}v+\frac{b}{D_X}(1-v)\right]^{X+Y+2a}} dv}{\int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{a-1}(1-v)^{a-1}}{\left[1+\frac{b}{D_Y}v+\frac{b}{D_X}(1-v)\right])^{X+Y+2a}} dv}.$$

# 3 A hidden semi-Markov model for spatial structure

As discussed earlier, an important characteristic of ChIP-Seq experiments is the spatial data structure, in which an enriched region is represented by a cluster of bins mapping in the vicinity of the binding site on the genome. We consider an automated algorithm that incorporates the distribution of the peak sizes in inferring bound regions. As we will illustrate

below, our proposed framework allows for an arbitrary number of non-overlapping peaks of variable lengths in each contiguous genomic region to be declared probabilistically. This bypasses the adhoc postprocessing procedure to combine contiguous bins in reporting final list of bound regions (Ji et al., 2008).

Although our model is formulated in a hierarchical manner, the existence of analytic marginal distributions allows us to easily recast the underlying spatial data structure as a hidden semi-Markov process. In a hidden semi-Markov model (HSMM), explicit duration distributions are introduced for each latent/hidden states. The peak size distribution $\rho$ specifies the duration distribution for $Z = 1$. On the other hand, the duration distribution for $Z = 0$ (non enriched region) is taken to be $W \sim Geo(1 - p_0) = p_0^{w-1}(1 - p_0)$, where $p_0$ is interpreted as the the probability of self transition to state $Z = 0$. Let $O_j = (X_j, Y_j)$ and $O_1^L = (O_1, ..., O_L)$ denote the observed data. The quantities needed to specify the HSMM are the initial distribution $\pi$, transition probabilities $a_{mn} = P(Z_j = n|Z_{j-1} = m)$ and the emission distributions of the observations $b_z(O_j)$, where $b_z(O_j) = P(Y_j|Z_j = z)$ for one sample problem and $b_z(O_j) = P(X_j, Y_j|Z_j = z)$ (bivariate mixture) or $P(Y_j|X_j + Y_j, Z_j = z)$ (conditional mixture) for two sample problem. Since self-transitions are prohibited in HSMM, in the case of comparing mixture of two hypotheses ($Z = 0, Z = 1$), the underlying data structure consists of segments of non-enriched regions alternating with enriched regions. To motivate the HSMM in detecting multiple enriched regions, we consider the following data generating process:

1. Set $j = 1$. Draw $Z_1$ from $\pi_z$.
   (a) If $Z_1 = 0$, draw a duration $w$ from $d_0 = Geo(1 - p_0)$ and set $Z_1, ..., Z_{1+w-1} = 0$, otherwise draw $w$ from $d_1 = \rho$ and set $Z_1, ..., Z_{1+w-1} = 1$.
   (b) Draw $O_k \sim b_z(.)$ for $k = 1, .., 1 + w - 1$.
   (c) Set $j = 1 + w$.
2. While $j \leq L$, draw $w$ from $d_{1-Z_{j-1}}$ and set $Z_j, ..., Z_{\min(j+w-1,L)} = 1 - Z_{j-1}$.
   (a) Draw $O_k \sim b_z(.)$ for $k = j, .., \min(j + w - 1, L)$.
   (b) Set $j = j + w$.

where $b_z(.)$ is the marginal distribution, e.g. $b_1(.) = P(Y|X + Y, Z = 1)$ in the conditional mixture modeling framework. The semi-Markov model offers a flexible framework to capture binding regions of variable lengths which is specified by $\rho$. We will discuss the choice of $\rho$ below.

We provide a motivating example of using a HSMM in a simulated ChIP-Seq data in Figure 4. Each vertical bar corresponds to tag count for a bin. True enriched regions are between bins 316 and 320 and between bins 442 and 448. We computed $P(Z_j = z|O_1^L)$ for each bin. Table 1 lists the tag counts for a few selected bins based on Figure 4.

Although bins 299 and 411 have higher tag counts than bins 316 and 446, the HSMM is able to distinguish the true states of these bins by utilizing the underlying spatial structure

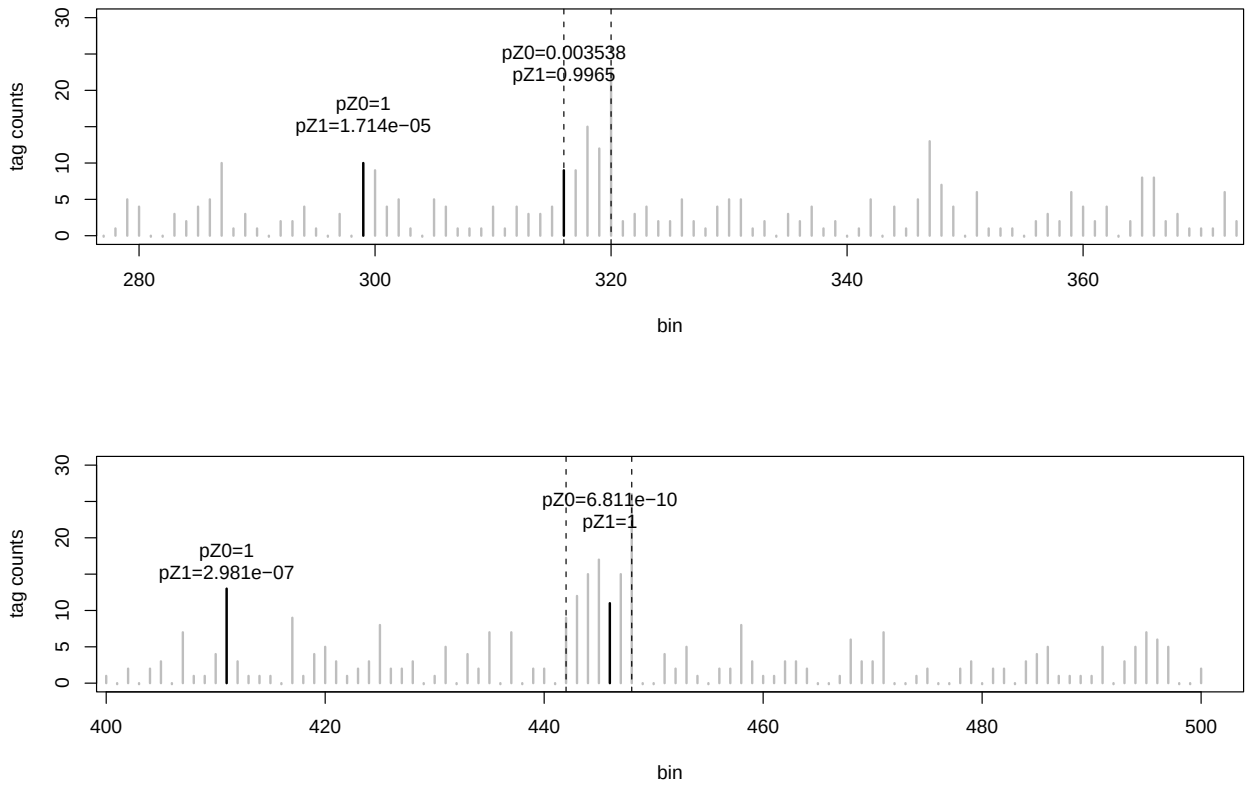



Figure 4: *Illustration of the effect of spatial structure.* Dotted lines indicate the boundaries of enriched regions.

Table 1: *Posterior probabilities for selected bins from Figure 4.*

| Bin | Tag count | True Z | $P(Z=0\|O_1^L)$ | $P(Z=1\|O_1^L)$ |
|---|---|---|---|---|
| 299 | 10 | 0 | 0.9901 | $9.867 \times 10^{-3}$ |
| 300 | 9 | 0 | 0.9901 | $9.869 \times 10^{-3}$ |
| 316 | 9 | 1 | $3.087 \times 10^{-3}$ | 0.9969 |
| 317 | 9 | 1 | $3.471 \times 10^{-5}$ | $\sim 1$ |
| 318 | 15 | 1 | $1.648 \times 10^{-8}$ | $\sim 1$ |
| 319 | 12 | 1 | $8.699 \times 10^{-9}$ | $\sim 1$ |
| 320 | 23 | 1 | $5.490 \times 10^{-6}$ | $\sim 1$ |
| 411 | 13 | 0 | 0.9997 | $3.392 \times 10^{-4}$ |
| 417 | 9 | 0 | $\sim 1$ | $4.011 \times 10^{-5}$ |
| 442 | 9 | 1 | $7.236 \times 10^{-2}$ | 0.9276 |
| 443 | 12 | 1 | $7.996 \times 10^{-4}$ | 0.9992 |
| 444 | 15 | 1 | $6.312 \times 10^{-9}$ | $\sim 1$ |
| 445 | 17 | 1 | $3.933 \times 10^{-14}$ | $\sim 1$ |
| 446 | 11 | 1 | $3.995 \times 10^{-12}$ | $\sim 1$ |
| 447 | 15 | 1 | $7.718 \times 10^{-9}$ | $\sim 1$ |
| 448 | 26 | 1 | $4.504 \times 10^{-6}$ | $\sim 1$ |

as indicated by the posterior probabilities $P(Z = z|O_1^L)$. In addition, $P(Z = 0|O_1^L)$ is lower for a bin that is in the center of an enriched region compared to a bin near the boundary of an enriched region, although both are in an enriched region. This is desirable since it is less likely to commit a mistake in declaring bins that are near the center of an enriched region compared to those at the boundaries.

Fitting a HSMM is challenging and more difficult than a regular hidden Markov model, since the powerful Baum-Welch algorithm (Rabiner, 1989) is not readily applicable. The Baum-Welch forward/backward algorithm involves multiplication of a large number of probabilities, thus generating underflowing errors. In a regular hidden Markov model, numerical underflow can be avoided via ad-hoc scaling the forward and backward variables. However, the analog of this scaling procedure is not available for HSMMs. Fortunately, a new procedure was derived recently by Guedon (2003) in recent years that is immune to numerical underflow and does not require ad-hoc scaling procedures. We adapt the derivation of Guedon (2003) in our model fitting strategy. The unknown parameters in the HSMM and the marginal distributions are estimated via the EM algorithm, coupled with the dynamic programming strategy to estimate the location of multiple peaks in each region, which is presented in the next section. Alternative strategies for mapping multiple peaks per region include several heuristic methods in multiple motif finding (Bailey and Elkan, 1995; Keleş et al., 2003). However, the dynamic programming in HSMM is more advantageous because it does not rely on any heuristic strategies to infer multiple instances of peak regions. In

addition, the by-products of the E-step in our proposed model allow for control of false positives or false discoveries at peak level, which will be described in Section 3.2.

*Choice of peak size distribution* $\rho$
The peak size distribution which usually ranges from 500 to 1000 bps for transcription factor binding can be estimated from the agarose gel image. Alternatively, it could be estimated via a cross-validation approach. In either case, the distribution can be approximated by a non-parametric discrete distribution over the range of binding lengths and we refer the readers to Keleş et al. (2006) for details on the estimation procedures. On the other hand, genomic regions undergoing histone modifications cover a larger range of sizes. An example of the distribution of peak sizes in H3K4me and H3K27me in human embryonic stem cells is given in Figure 2(A) of Pan et al. (2007). This suggests that the distribution can be approximated by a shifted geometric distribution, $w \sim p(1-p)^{w-C}$ for $w \geq C$. $C$ is usually the minimum size of a histone modified region. To access the goodness of fit with such peak distribution, we downloaded the annotated histone modification regions from the Canada's Michael Smith Genome Sciences Centre website at `http://www.bcgsc.ca/` and plotted the distribution of actual peak sizes against simulated peak sizes in Figure 5. The peak sizes were simulated from a shifted geometric distributions with $C = 200$ and $p = (0.0017, 0.00135, 0.0015, 0.0025, 0.0026, 0.002)$. Figure 5 illustrates that this shifted geometric distribution is sufficient to approximate the lengths of histone modified regions. Note that when the duration distributions for $Z = 0$ and $Z = 1$ are a geometric and shifted geometric at $C$ , the HSMM is equivalent to a regular hidden Markov model architecture given in Figure 6.

## 3.1 Model fitting via EM algorithm and dynamic programming

Apart from the unobserved $Z_j = z \in \{0, 1\}$ which specifies the hidden state of bin $j$, we introduce two additional latent variables $(T_j, V_j)$, where $T_j = z$ denote the event "state $z$ starts at bin $j$" and $V_j = z$ denote the event "state $z$ ends at bin $j$". Let $\theta = (\pi_z, d_z, b_z)$ denote the unknown parameters in the model. Here $d_0(w) = p_0^{w-1}(1-p_0)$ and $d_1 = \rho$. The unknown parameters in the marginal distributions $b_z$ are $(a_0, a_1, b)$ in one sample problem and $(a, b)$ in two sample problem. Given the latent variables $(Z, T, V)$ and $\theta$, the complete data likelihood is given by

$$\begin{aligned} P(O_1^L, Z_1^L, T_1^L, V_1^L|\theta) &= \left[\prod_{z=0}^{1} \pi_z^{I(T_1=z)}\right] \times \left[\prod_{z=0}^{1}\prod_{j=0}^{L-1}\prod_{w\geq 1} d_z(w)^{I(T_{j+1}=z, V_{j+u}=z)}\right] \\ &\quad \times \left[\prod_{z=0}^{1}\prod_{j=1}^{L} b_z(O_j)^{I(Z_j=z)}\right] \end{aligned}$$

where $O_1^L = (O_1, ..., O_L)$.

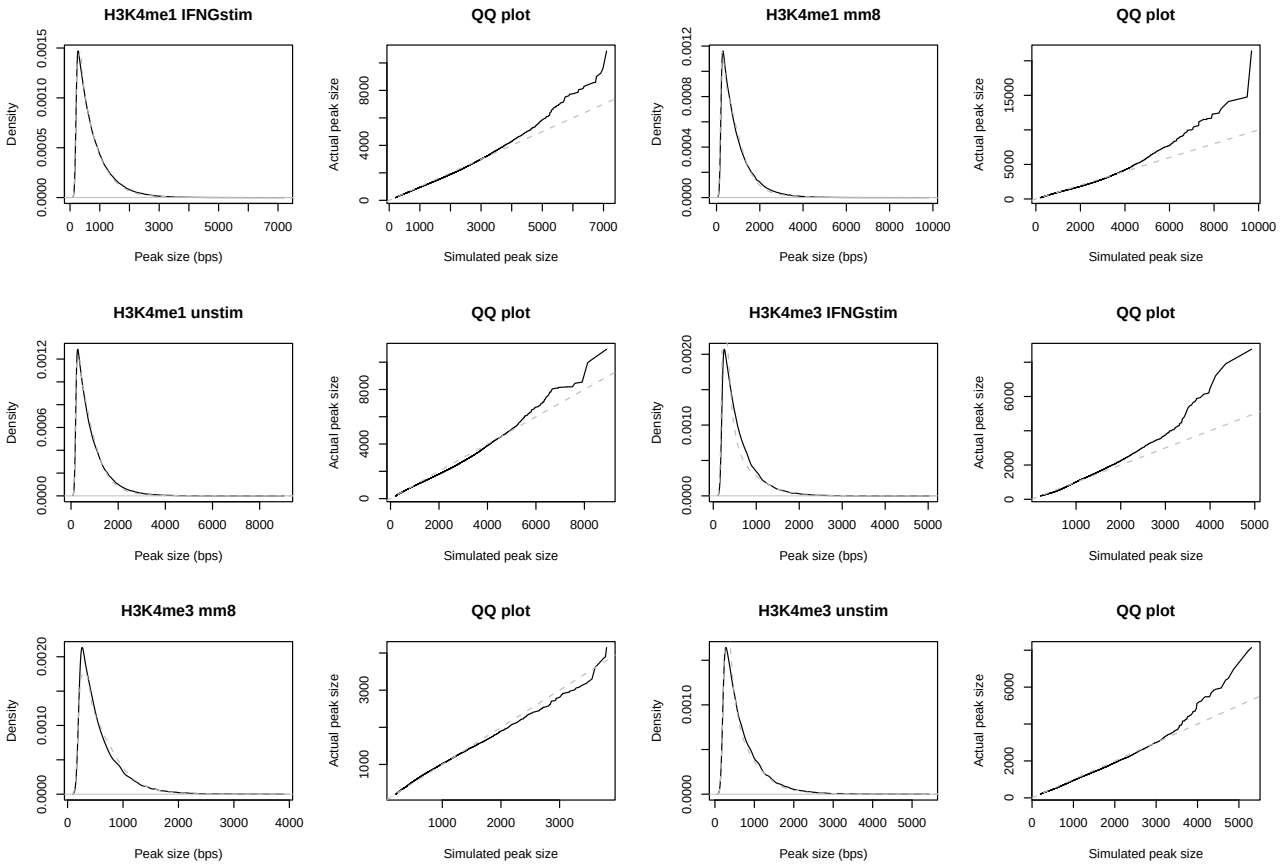



Figure 5: *Peak size distributions for histone modifications.* The density and quantile-to-quantile plots of simulated peak sizes against observed peak sizes. The data are simulated from $Geo(p)+200$, where $p = (0.0017, 0.00135, 0.0015, 0.0025, 0.0026, 0.002)$ for the 6 histone modifications. The black and gray line in the density plots correspond to simulated and actual data, respectively.

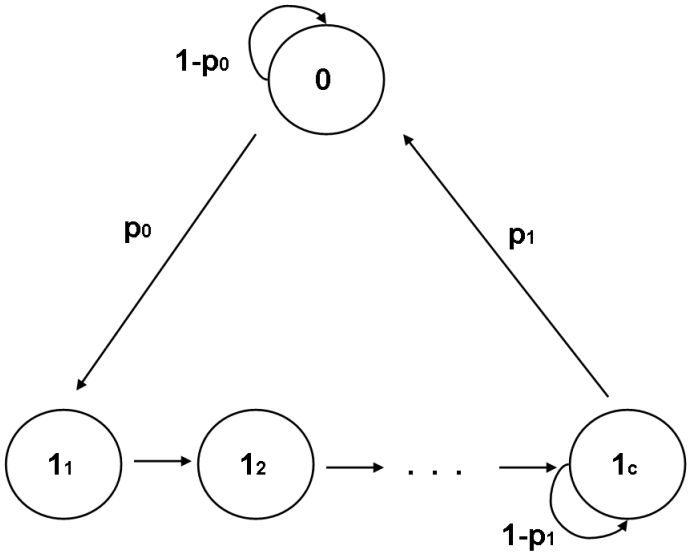



Figure 6: *Equivalent regular HMM representation.* A hidden semi-Markov model in which all the duration distributions are geometric/shifted geometric is equivalent to a regular HMM with enlarged state space.

The E-step in the EM algorithm includes computation of the following quantities:

$$P(T_1 = z|O_1^L, \theta),$$
$$P(T_{j+1} = z, V_{j+u} = z|O_1^L, \theta),$$
$$P(Z_j = z|O_1^L, \theta).$$

Direct calculations of the quantities above is computationally prohibitive. We utilized the dynamic programming scheme for HSMM by Guedon (2003) that is computationally efficient and immune to numerical underflow problems through a normalizing factor $N_j = P(O_j|O_1^{j-1})$. The key quantities in the algorithm are

$$\begin{aligned} F_j(z) &= P(V_j = z|O_1^j) \text{ (forward variable)}, \\ L1_j(z) &= P(V_j = z|O_1^L), \\ L_j(z) &= P(Z_j = z|O_1^L) \text{ (backward variable)}, \end{aligned}$$

which are computed recursively. The derivation tailored for two hidden states ($Z = 0, Z = 1$) are given in Appendix A.5. The M-step involves re-estimation of $\theta$ given the E-step variables. To reduce computation time, we assume that the peak size distribution $\rho$ has been estimated and fixed. However, the M-step can be extended to incorporate the re-estimation of $\rho$, i.e., a discrete non-parametric distribution in the case of transcription factor binding or a shifted geometric distribution in the case of histone modifications.

## 3.2 Inference

Comparisons of enriched regions from multiple experiments are meaningful if the peak set for each experiment is declared under a pre-specified error control. Most of the available tools for ChIP-chip and ChIP-Seq data control the FDR at probe/bin level, despite the interest in inferring a set of bins which constitutes a peak/enriched region instead of individual probes/bins (Ji et al., 2008). Although MACS (Zhang et al., 2008) proposed a version of peak level FDR control based on sample swap, their definition of empirical FDR could be violated in some cases (e.g., # control peaks > # ChIP peaks $\Rightarrow$ FDR> 1). The sample swap approach is also not applicable in two sample comparison of differential enrichments. On the other hand, for bin level FDR control, reporting a peak set is usually carried out as a heuristic postprocessing to merge contiguous bins declared to be statistically significant and requires the user to pre-specify the maximum allowable bins below the threshold and the minimum number of bins within a peak region. To bypass this ad hoc postprocessing approach, we propose a meta FDR approach for controlling FDR at peak level. We will now discuss several useful posterior probabilities that are byproducts of the E-step of the EM algorithm and can be utilized for error control. A quantity of interest for inferring the most probable start and end of an enriched region is $P(T_j = 1, V_k = 1|O_1^L, \theta)$, which is the posterior probability of bins $j$ and $k$ defining the boundary of an enriched region and can be used to rank candidate enriched regions. The boundaries of enriched regions could also be decoded via the Viterbi algorithm (Rabiner, 1989) to determine the most likely sequence of hidden states generating the observed data. In a HSMM, the Viterbi decoding automatically generates a set of non overlapping enriched regions that maximizes the likelihood function of the observed sequence of tag counts. Let $\mathcal{P}_V = \{\hat{p}\}$ be the list of enriched regions identified via the Viterbi algorithm, where $\hat{p} = (\hat{j}, \hat{k})$ are the start and end positions of an inferred enriched region. Define $\beta_{j,k}$ to represent the posterior probability of region covered by bins $j, .., k$ being a false peak. The choice of $\beta_{j,k}$ is discussed below. Consider the goal of identifying a list of enriched regions that is as large as possible while bounding the FDR by $\alpha$. We propose the following strategy for identifying the most probable enriched regions while controlling FDR at level $\alpha$. This strategy can be considered as a modified version of the *direct posterior probability* approach of Newton et al. (2004).

1. Initialize:
   (a) List of enriched regions: $\mathcal{P} = \emptyset$.
   (b) Candidate start positions: $\mathcal{J} = \{1, ..., L - \min(\mathcal{W}) + 1\}$.
   (c) Candidate end positions given a start position $j$: $\mathcal{V}|j \in \mathcal{J} = \{j + \min(\mathcal{W}) - 1, ..., j + \min(L - j + 1, \max(\mathcal{W})) - 1\}$. Here, $\min(\mathcal{W})$ and $\max(\mathcal{W})$ are the minimum and maximum peak sizes, respectively.
   (d) Actual FDR: $\hat{\alpha} = 0$.
2. Compute actual FDR:
   Define $\hat{\alpha} = \sum_{(j,k)\in\mathcal{P}_V} \beta_{j,k}/|\mathcal{P}_V|$, where $|\mathcal{P}_V|$ is the cardinality of $\mathcal{P}_V$ from the Viterbi algorithm. If $\hat{\alpha} \geq \alpha$, go to step 3. Else go to step 4.

3. Bound actual FDR:
   Sort $\beta_{j,k}(1) \leq \beta_{j,k}(2) \leq ... \leq \beta_{j,k}(|\mathcal{P}_V|)$. Let $n \in \{1, ..., |\mathcal{P}_V|\}$ be the largest value such that $\sum_{r=1}^{n} \beta_{j,k}(r)/n \leq \alpha$. Update $\mathcal{P} = \{\hat{p}(1), ..., \hat{p}(n)\}$, where $\hat{p}(r)$ corresponds to the start and end coordinate in $\beta_{j,k}(r)$.

4. Pre-select Viterbi identified regions as enriched:
   Update $\mathcal{P} = \mathcal{P}_{\mathcal{V}}$, $\mathcal{J} = \mathcal{J} \backslash \{\hat{p} \in \mathcal{P}_{\mathcal{V}}\}$ and $\mathcal{V}|j \in \mathcal{J} \backslash \{\hat{p} \in \mathcal{P}_{\mathcal{V}}\}$. Go to step 5.

5. Update the set of enriched regions until the desired FDR level is reached:
   While $\hat{\alpha} \leq \alpha$:

   (a) Let $(\hat{i}, \hat{j}, \hat{w}) = \text{argmax}_{i,j \in \mathcal{J}, k \in \mathcal{V}} P(T_j = 1, V_k = 1 | O_1^L, \theta)$ and $\hat{p} = (\hat{j}, \hat{k})$ be the start and end position of the inferred enriched region.

   (b) Update $\mathcal{P} = \mathcal{P} \bigcup \{\hat{p}, \hat{j} - \max(\min(\mathcal{W}), \max(\mathcal{W})/2) + 1, ..., \hat{j} - 1\}$.

   (c) Update $\mathcal{J} = \mathcal{J} \backslash \{\hat{p}\}$ and $\mathcal{V}|j \in \mathcal{J} \backslash \{\hat{p}\}$.

   (d) Update $\hat{\alpha} = \sum_{(j,k) \in \mathcal{P}} \beta_{j,k}/|\mathcal{P}|$, where $|\mathcal{P}|$ is the cardinality of $\mathcal{P}$.

The procedure described above allows for meta FDR control at peak level by utilizing the byproducts of the EM algorithm, an added advantage of the proposed hierarchical semi-Markov framework. Since the Viterbi algorithm outputs the most probable candidate enriched regions that maximizes the observed likelihood, we first utilize this decoding to get an initial set of enriched regions. If the empirical FDR $\hat{\alpha}$ of this set is larger than $\alpha$, we remove some candidate enriched regions in Step 3. On the other hand, if $\hat{\alpha} \leq \alpha$, the set $\mathcal{P}$ is expanded by including additional candidate enriched regions in Steps 4 and 5. We use $P(T_j = 1, V_k = 1 | O_1^L, \theta)$ in Step 5(a) to guide the selection of the most probable boundary of an enriched region, whereas Step 5(c) is to avoid inferring two highly overlapping regions as separate enriched regions. There are several choices for defining $\beta_{j,k}$ (the posterior probability of region covered by bins $j, .., k$ being a false peak):

1. $1 - P(T_j = 1, V_k = 1 | O_1^L, \theta)$
2. $1 - \sum_{t=j}^{k} P(Z_t = 1 | O_1^L, \theta)/(k - j + 1)$
3. $1 - \max_{t \in \{j, ..., k\}} P(Z_t = 1 | O_1^L, \theta)$

If (1) is chosen as the definition of $\beta_{j,k}$, a false discovery will be declaring the boundary of an enriched region wrongly. On the other hand, (2) and (3) can be interpreted as the average and maximum significance level of declaring region covering bins $j$ to $k$ as enriched region. We investigate the performance of these choices in extensive simulation studies.

# 4 Simulation studies

## 4.1 Choice of $\beta_{j,k}$

We consider a simple simulation setup with $L = 2000$ and $p_0 = 0.98$. In addition, we assume a discrete peak size distribution $\rho = P(\mathcal{W}) = (1, 2, 3, 4, 3, 2, 1)/16$ over the range $3 \leq \mathcal{W} \leq 9$

and sufficient sequencing depth. The unknown state $Z_j$ for each bin is simulated according to a HSMM while the emission distribution is simulated from a one sample hierarchical model with $\lambda_{j0} \sim \mathcal{G}a(2, c/(1-c))$ and $\lambda_{j1} \sim \mathcal{G}a(2+a_1, c/(1-c))$, where $c \sim U(0.4, 0.5)$. We consider $a_1 = (8, 13, 18, 23)$ which corresponds to signal to noise ratio (SNR=$\sqrt{(2+a_1)/2}$) of $(2.2, 2.7, 3.2, 3.5)$. An example of simulated data is given in Figure 7.

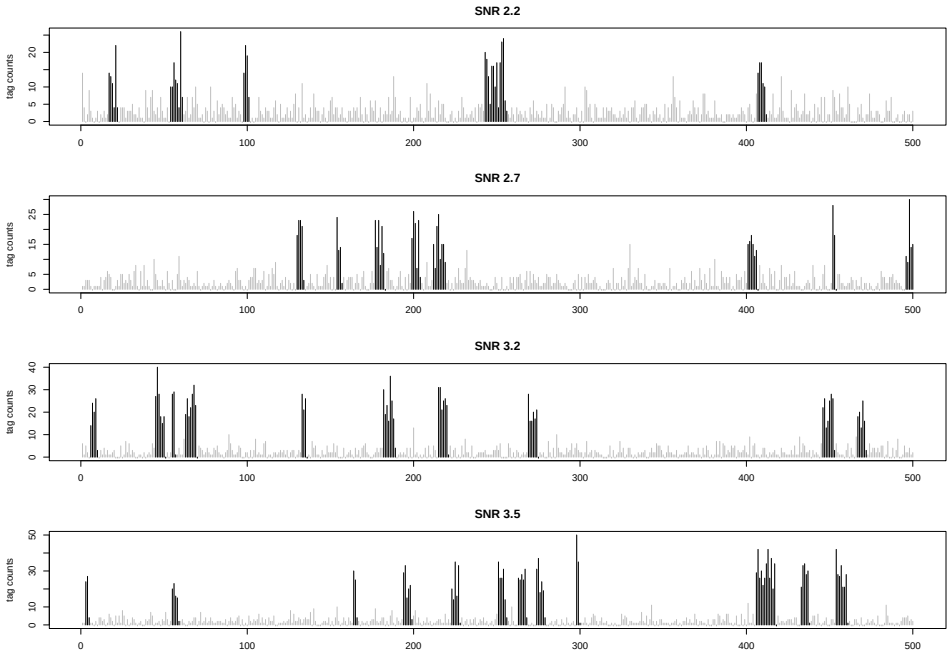



Figure 7: *An illustrative example of simulated data for various SNR.* Black and gray bars denote enriched and non enriched bins, respectively.

We evaluate the FDR control using the proposed procedure in Section 3.2 for the three choices of $\beta_{j,k}$. At various nominal FDR levels $\alpha$, a set of peaks is obtained according to Section 3.2. A peak is considered a true discovery if both the start and end position are within a small margin (2 bins) of the set of known true peaks. Figure 8 plots the empirical FDR against the nominal FDR for the three choices of $\beta_{j,k}$ from 50 simulations. We also included the bin level empirical FDR control for the p-values computed from the null distribution $NB(2, c/(1-c))$ and adjusted according to Benjamini and Hochberg (1995). In all four cases, bin level FDR tends to declare more false positives because it does not utilize the spatial structure of the enriched regions. For low SNR, using (1) as the definition of $\beta_{j,k}$ appears to be more conservative compared to (2) and (3). At nominal FDR$\leq$ 0.05, the set of peaks identified by (1) does not contain any false peaks, thereby have zero empirical FDR value. It is not surprising that (1) is the most conservative among the three since a false discovery is committed if the boundaries of a peak is declared wrongly. (3) is comparable to (2), but slightly too liberal at small nominal FDR levels for low SNR. It is interesting to observe that as the SNR increases, all the three choices of $\beta_{j,k}$ provide accurate FDR

control. At high SNR, the posterior probabilities $P(T_j = 1, V_k = 1|O_1^L, \theta)$ are able to locate the boundaries of enriched regions accurately. Based on the simulation results, (2) appears to be the best choice for defining $\beta_{j,k}$ in the proposed meta FDR control at peak level.

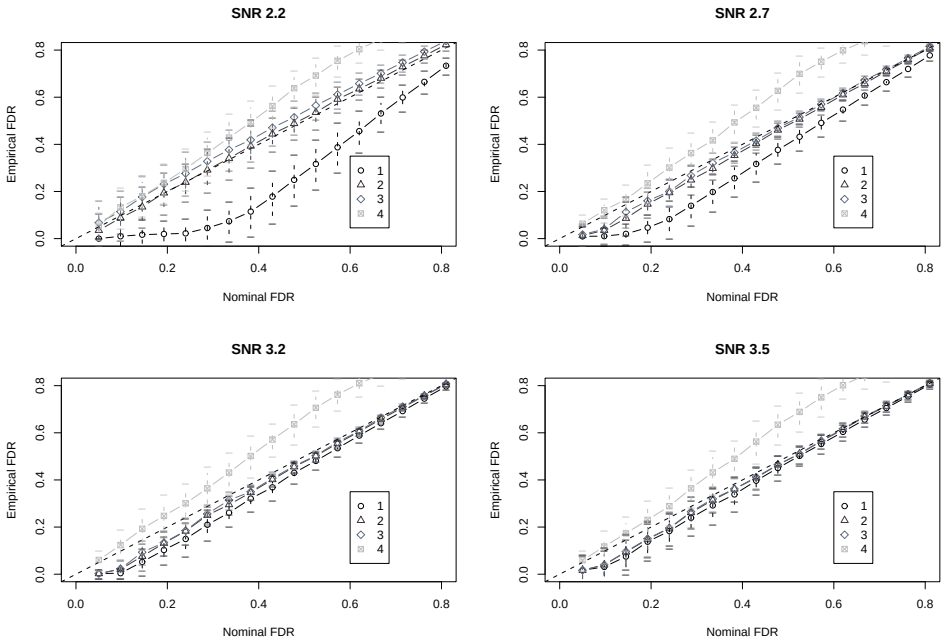



Figure 8: *Empirical versus nominal FDR for various choices of* $\beta_{j,k}$. The different choices of $\beta_{j,k}$ are (1) $1 - P(T_j = 1, V_k = 1|O_1^L, \theta)$, (2) $1 - \sum_{t=j}^{k} P(Z_t = 1|O_1^L, \theta)/(k - j + 1)$, (3) $1 - \max_{t\in\{j,\ldots,k\}} P(Z_t = 1|O_1^L, \theta)$ and (4) bin level FDR. Vertical bars are the corresponding standard errors over 50 simulations.

We also evaluate the accuracy of the Viterbi algorithm in detecting the boundaries of true enriched regions. The sensitivity is defined as the fraction of true enriched regions that is within $m$ bins of the peak regions from the Viterbi decoding. As shown in Figure 9, this algorithm is able to identify all the peak regions accurately by allowing one bin margin of error. The number of peaks detected by the Viterbi algorithm is approximately equal to the number of true peaks indicating that it has an extremely low false positive rate, i.e., no additional false peaks is detected. This provides evidence for pre-selecting Viterbi identified regions as enriched in Step (4) of the proposed procedure in Section 3.2.

## 4.2 Simulations in two sample problem

Direct maximum likelihood estimation for the unknown parameters in two sample problem requires intensive numerical optimization which could result in unstable estimates as shown in Appendix A.3.2. Therefore, we propose a simpler approximate re-estimation for two sample problem as given in Appendix A.3.1 and evaluate the accuracy of the estimates via simulation

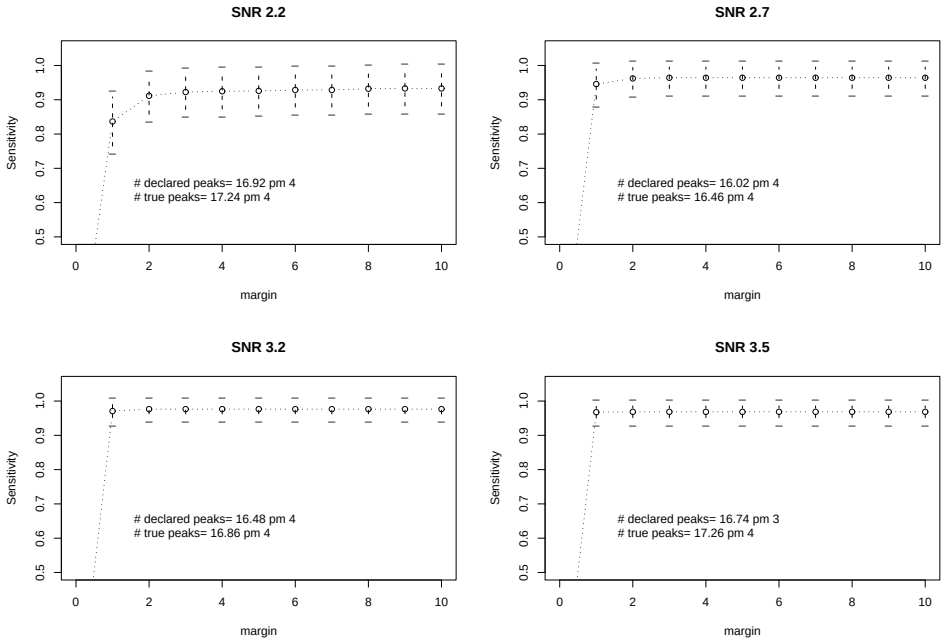



Figure 9: *Sensitivity of the Viterbi decoding in identifying the boundaries of true enriched regions at various tolerance/margin of errors.* Vertical bars are the corresponding standard errors over 50 simulations.

studies. The data is simulated according to Section 2.2 with $\omega_1 = 1/(\beta_1 + 1) \sim U(0.5, 1)$, $\omega_2 = 1/(\beta_2 + 1) \sim U(0.5, 1)$, $\pi \sim \mathcal{G}a(a, c/(1-c))$, where $a \sim U(0.5, 10)$ and $c \sim U(0.4, 0.9)$. Figure 10 plots the estimated values against the simulated true values for $\omega_1, \omega_2, a$ and $b$ for 20 simulated data. As evident from this figure, the proposed re-estimation procedure provides relatively good estimates for the unknown parameters in two sample problem.

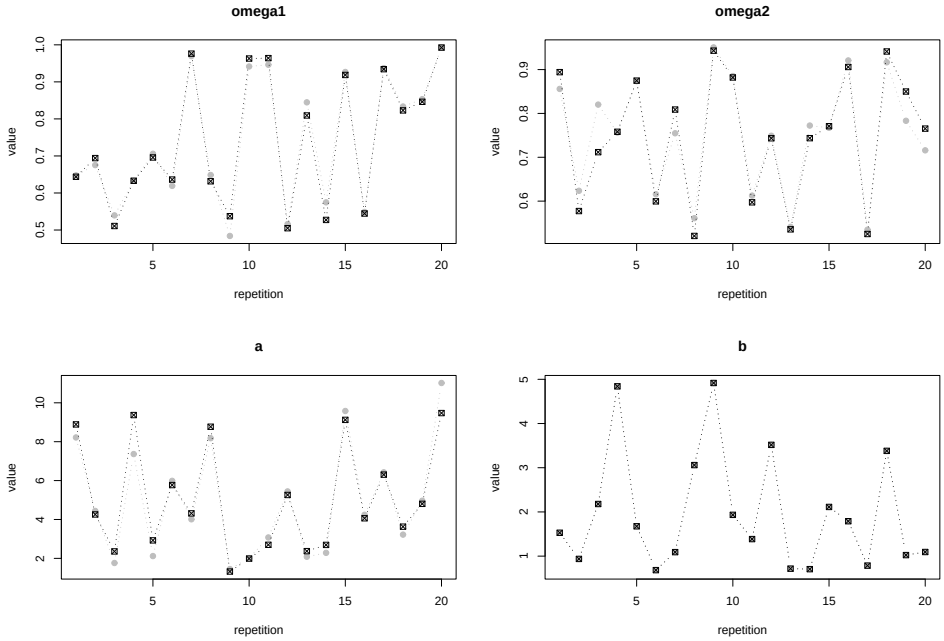



Figure 10: *Estimated versus true parameters.* Each panel plot the estimated versus true values for the four emission distribution parameters in two sample problem for 20 simulated data. Black lines/points are the true values. Gray lines/points are the estimated values.

Next, we evaluate the proposed meta FDR control procedure on two sample problem. For bin level FDR control, we calculate the p-values from $Bin(X+Y, D_Y/(D_X+D_Y))$. The results over 50 simulations are summarized in the left panel of Figure 11. Bin level FDR control has the worst performance in two sample problem since it does not account for the spatial structure of enriched regions. On the other hand, the empirical FDR from proposed meta FDR control with $\beta_{j,k} = 1 - \sum_{t=j}^{k} P(Z_t = 1|O_1^L, \theta)/(k-j+1)$ is the closest to nominal FDR. We also evaluate the accuracy of the Viterbi decoding in detecting boundaries of simulated enriched regions in two sample problem. The right panel of Figure 11 summarizes the sensitivities from 50 simulations. The average number of enriched regions from the Viterbi decoding (17.2) is close to the average number of true enriched regions (17.88), which again indicates a very low false positive rate. Most of the enriched regions from Viterbi decoding are within $m = 2$ bins of the true enriched regions.

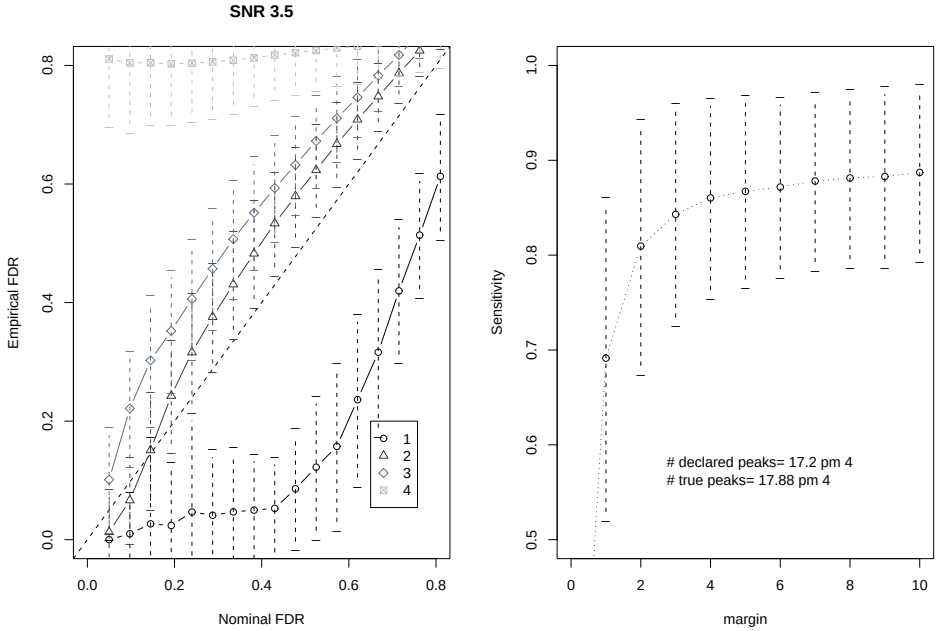



Figure 11: *Simulation results for conditional mixture emission in two sample problem.* Left panel plots the empirical versus nominal FDR for various choices of $\beta_{j,k}$, in which (1) $1 - P(T_j = 1, V_k = 1|O_1^L, \theta)$, (2) $1 - \sum_{t=j}^{k} P(Z_t = 1|O_1^L, \theta)/(k-j+1)$, (3) $1 - \max_{t\in\{j,\ldots,k\}} P(Z_t = 1|O_1^L, \theta)$ and (4) bin level FDR. Right panel plots the sensitivity of the Viterbi decoding in identifying the boundaries of true enriched regions at various tolerance/margin of errors. Vertical bars are the corresponding standard errors over 50 simulations.

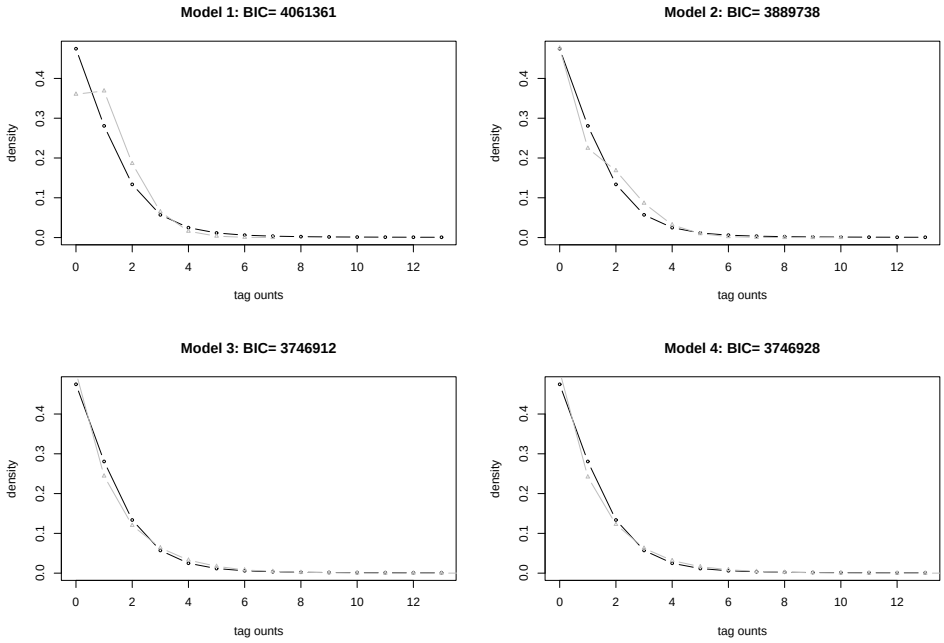
# 5 Case studies

TGFb superfamily plays an important role in regulating self renewal and differentiation potential of embryonic stem (ES) cells and lineage choices at gastrulation in embryogenesis (Tam and Loebel, 2007). The growth factors of the TGFb superfamily consists of two branches, namely NODAL and BMP. Interplay between these two branches determines the fate of ES cells, i.e., maintaining or exiting pluripotency. In particular, NODAL signaling helps maintain pluripotency while BMP signaling triggers differentiation. Upon binding to the receptors, NODAL branch signaling catalyzes phosphorylations on transcription factors SMAD2/3, while the signals from BMP branch phosphorylate transcription factors SMAD1/5/8 (Ross and Hill, 2008). It is therefore crucial to understand the mechanisms governing the two TGFb signaling pathways. ChIP-Seq experiments were conducted at the Thomson Lab, University of Wisconsin-Madison and the Morgridge Institute for Research, Madison, WI to map in vivo binding regions of SMAD2/3, SMAD4 and SMAD1/5/8 under untreated and BMP4 ES cells treated for six hours. The data were generated from the Illumina/Solexa sequencer.

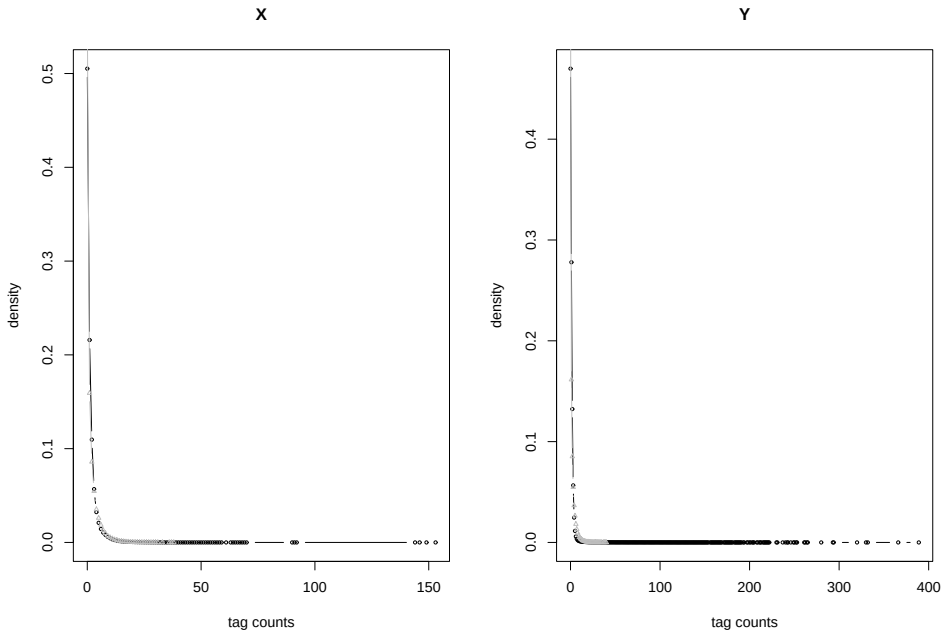
We illustrate the proposed hierarchical semi-Markov model in a ChIP-Seq experiment measuring transcription factor SMAD1/5/8 binding on BMP4 cells treated for six hours. Locating the binding sites of this transcription factor an important step to elucidate how BMP signaling initiates differentiation in ES cells. Our analysis is conducted using a bin size of 100 bps on Chromosome 10. The peak size distribution which ranges from 200 to 2200 bps is determined empirically by a preliminary one sample bin level analysis without the spatial structure. The corresponding control experiment is the genomic/chromatin DNA input from BMP4 cells treated for six hours.

We analyse the data using both the one sample hierarchical mixture model (without the genomic DNA input) and the two sample conditional hierarchical mixture model. For computational efficiency, the parameters in the emission and duration distributions are initialized and fixed according to Appendices A.1, A.3.1, and A.4. In one sample analysis, the Bernoulli random variable in ZINB (Model 4) converges to 1, which reduces the model to a regular negative binomial model as shown in Figure 12(a). Figure 13 illustrates the annotation from both analyses on selected regions at FDR=0.05. A total of 2445 and 1274 enriched regions is obtained from one sample and two sample conditional hierarchical mixture model, respectively. Among the 1274 enriched regions identified from two sample analysis, 95.7% of the regions overlap with the enriched regions identified from one sample analysis. In addition, 88.5% of the 1274 regions is an exact subset of the larger peak set from one sample analysis, i.e., the peak boundaries from two sample analysis fall within the peaks from one sample analysis. This indicates that two sample conditional hierarchical model is able to refine the boundaries of identified enriched regions as evident from Subfigures 13(a)-13(c), 13(f). Subfigures 13(d) and 13(e) further demonstrate the advantage of using the genomic DNA input in two sample analysis in removing non-specific enriched regions.

Two genes of interests on Chromosome 10 are GATA3 and NODAL. GATA3 is an early trophoblast associated gene which is expressed at very low level but is significantly induced upon BMP signaling. On the other hand, NODAL is highly expressed in ES state but is




(a) One sample goodness of fit




(b) Two sample goodness of fit

Figure 12: *Goodness of fit for BMP4 SMAD 1/5/8 analysis.* Top panel is the goodness of fit of Models 1 to 4 in one sample analysis. Bottom panel is the goodness of fit for two sample conditional mixture model, where `X` is the genomic DNA input and `Y` is the ChIP sample.

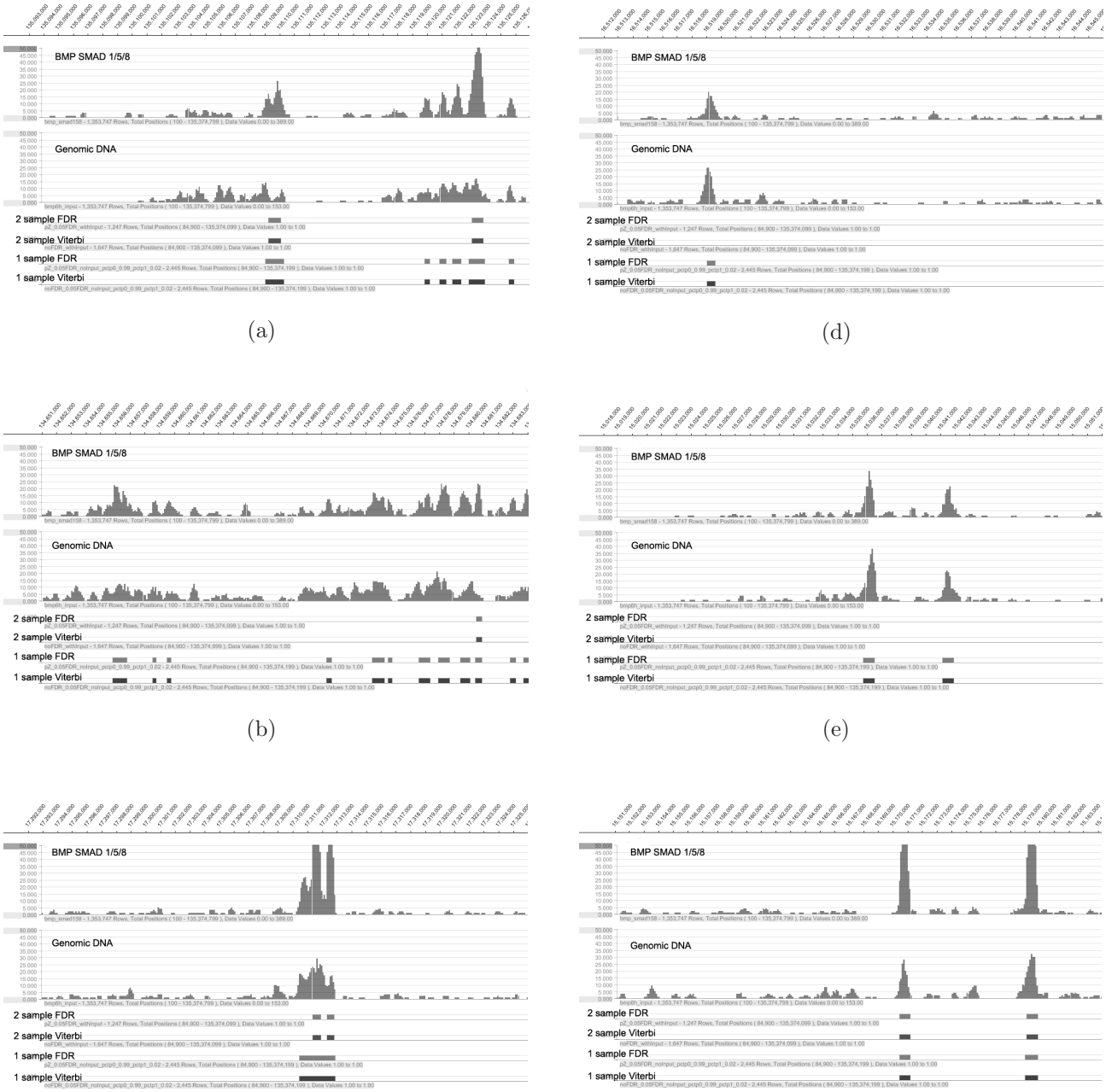



Figure 13: *Example of identified enriched regions.* Track 1 and 2 are the observed tag counts for each bin in treatment (BMP4 SMAD1/5/8) and control (Genomic DNA input) sample, respectively. Track 3 and 5 are the annotations by applying the FDR control procedure in Section 3.2 for two sample conditional mixture and one sample mixture (ignoring the control sample), respectively. Track 4 and 6 are the corresponding Viterbi identified enriched regions without FDR control for two sample conditional mixture and one sample mixture, respectively.

significantly suppressed upon differentiation, and has been reported as a direct target of NODAL signaling (Besser, 2004). The binding pattern of SMAD 1/5/8 at the promoter regions of GATA3 and NODAL are given in Figure 14. We further map the 1274 identified enriched regions to the promoter and UCSC gene regions in Table 2. More than 70% of the identified enriched regions are located within -10000 bps of a transcription start site (TSS) plus gene regions in Chromosome 10. To validate the specificity of the identified regions enriched in SMAD1/5/8 binding in BMP4 treated cells, we examine the corresponding binding pattern in untreated ES cells. In untreated ES cells, BMP signaling is inactive and this is reflected by the decrease in binding activities of SMAD1/5/8. For each of the peak regions, we compute the average ratio $R_E$ of the emission distribution under enrichment ($Z = 1$) against non enrichment ($Z = 0$), i.e.,

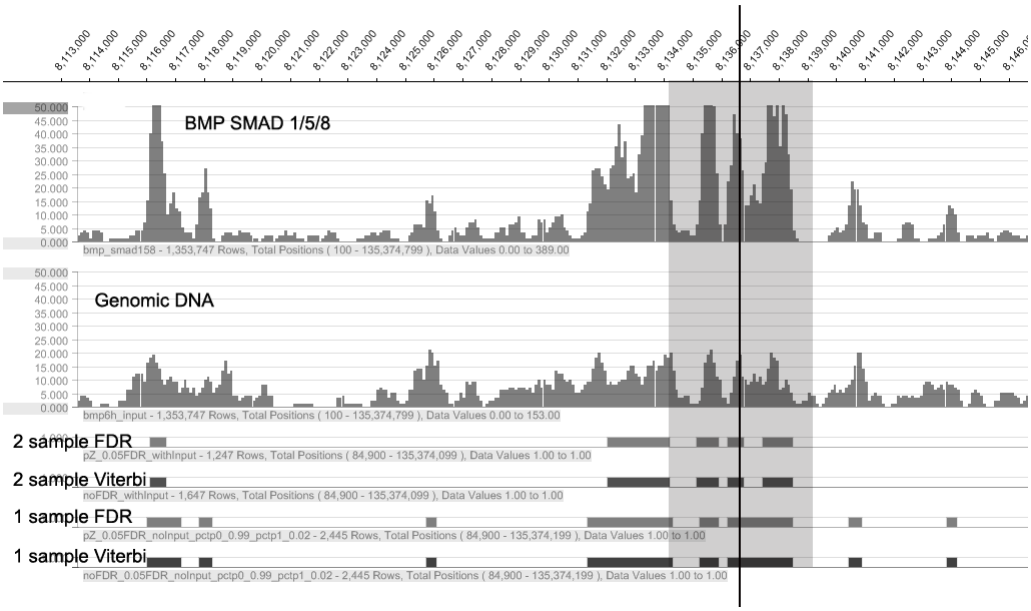
$$R_E = \frac{\sum_{i \in E} P(Y_i | X_i + Y_i, Z_i = 1)}{\sum_{i \in E} P(Y_i | X_i + Y_i, Z_i = 0)},$$

where $E$ is the set of bins in a peak region. We randomly draw 1274 non peak regions and computed $R_{NE}$, where $NE$ is the set of bins in a randomly drawn non peak region and this process is repeated 50 times. Large values of $R_E$ or $R_{NE}$ indicate reduction in binding between BMP4 treated and untreated ES cells. As evident from Figure 15, the peak regions show significant decrease in binding from BMP4 treated to ES cell compared to non peak regions.

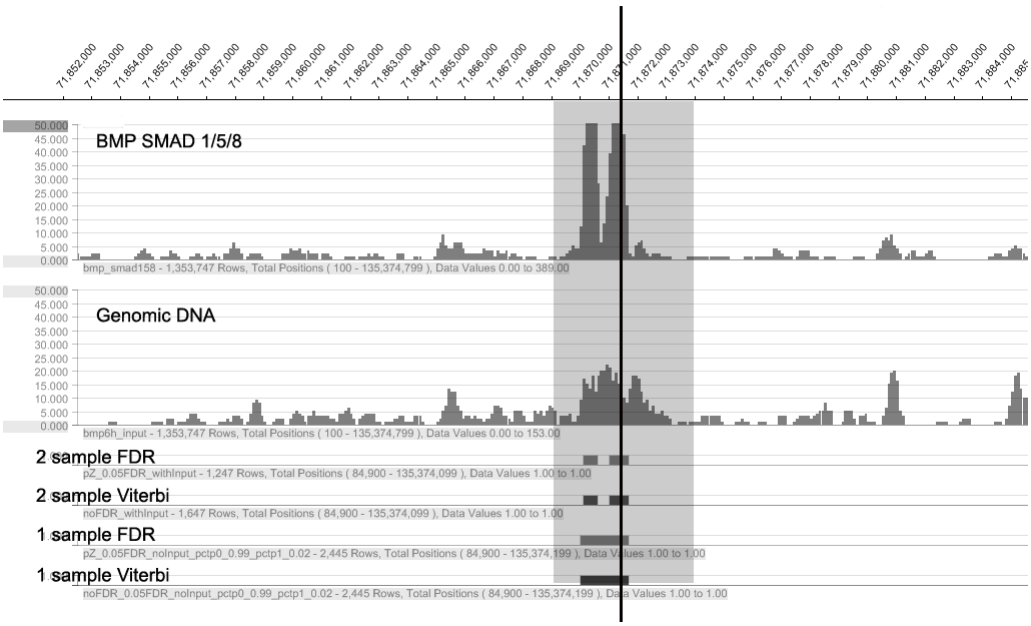
Table 2: *Percentage of enriched regions in promoter and gene regions.*

| Promoter | Percentage mapped |
|---|---|
| $\pm$2500 | 0.407 |
| $\pm$5000 | 0.464 |
| $\pm$10000 | 0.526 |
| $\pm$25000 | 0.658 |
| $\pm$50000 | 0.766 |
| $\pm$100000 | 0.868 |

| $x$-bps upstream + gene region | Percentage mapped |
|---|---|
| $x = 0$ | 0.415 |
| $x = 2500$ | 0.678 |
| $x = 5000$ | 0.694 |
| $x = 10000$ | 0.715 |




(a) GATA3




(b) NODAL

Figure 14: *BMP4 SMAD1/5/8 binding at the promoter regions of GATA3 and NODAL.* The rectangular boxes highlight ±2500-bps TSS of GATA3 and NODAL.

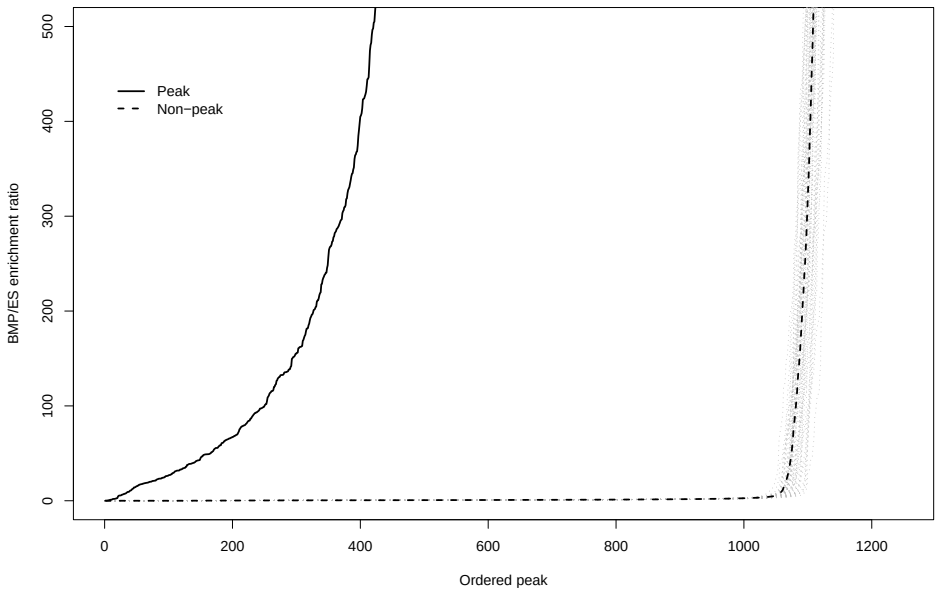



Figure 15: *Comparison of enrichment level for SMAD 1/5/8 in BMP4 treated against ES cell in peak and non-peak regions.* Gray lines are the 50 randomly drawn non-peak regions.

# 6 Discussion

The introduction of next generation sequencing instruments in recent years has enabled whole-genome regulatory DNA-protein binding interactions (ChIP-Seq) to be elucidated at lower costs and is becoming a popular alternative to the tiling array (ChIP-chip) experiments. Although this technology offers promising results for surveying large genomes at high resolution, limited statistical tools are available to analyze the ChIP-Seq data. Current models for the background distribution of ChIP-Seq data include the regular Poisson and negative binomial distribution. In this paper, we carefully studied the data generating process of ChIP-Seq data and introduced zero-inflated Poisson (ZIP) and negative binomial (ZINB) models to account for the excess zeroes in the observed tag counts. In particular, we demonstrated that the more flexible ZINB for modeling the background distribution fits the observed ChIP-Seq data better. The proposed hierarchical modeling offers a general framework that incorporates bin specific distribution and sequencing biases, and allows for information sharing across bins. Although our current hierarchical model implementation is based on conjugate priors, our proposed hierarchical framework is extendable to include additional covariates contributing to non-specific biases.

We also proposed a hierarchical mixture model for the two sample problem for inferring enriched regions relative to a control experiment or detecting differential enrichment between two treatment samples/libraries. The available tools for two sample ChIP-Seq data analysis

usually normalize the sequencing depth between the two samples to the same number by linear scaling. However, we showed that this is undesirable if the underlying distribution of the tag counts is indeed Poisson or negative binomial. Instead of a linear scaling, the sequencing depth is included as a parameter in our two sample hierarchical model. We introduced (1) bivariate mixture model and (2) conditional mixture model, and investigated the power of the two formulations in discriminating enriched from non-enriched distribution. Our power analysis suggested that two sample conditional mixture model is more suitable if the goal is to detect enriched regions relative to a genomic DNA input.

Most of the available tools for ChIP-Seq data analysis control the FDR at the bin level, despite the inherent spatial structure in the observed data and the interest in inferring individual peaks instead of individual bins. Reporting a list of enriched regions is usually carried out as a heuristic postprocessing step to merge consecutive bins declared to be enriched as well as removing small peaks, which affects the actual FDR level. We proposed a model that incorporates the spatial structure in ChIP-Seq data via a hidden semi-Markov model (HSMM). This allows for automatic detection of multiple non overlapping variable size peaks. We also introduced a new meta approach for controlling FDR at peak level by utilizing the byproducts of the EM algorithm and demonstrated that this approach provides accurate FDR control in extensive simulation studies. Since optimizing model parameters in the HSMM is computationally intensive, we proposed methods to pre-estimate the unknown parameters and showed that this procedure provides good estimates in simulation and case studies. By pre-estimating and fixing the unknown parameters, only one forward/backward recursion is needed and this offers a reasonable computational time to analyze massive amounts of ChIP-Seq data. Source codes for fitting the hierarchical semi-Markov model are available upon request. (An `R` package will be made publicly available soon).

# Acknowledgements

This research has been supported in part by a PhRMA Foundation Research Starter Grant in Informatics (P.K. and S.K.), the NIH grant HG003747 (P.K. and S.K.), the NSF grant DMS004597 (P.K. and S.K.) and the Morgridge Institute for Research support for Computation and Informatics in Biology and Medicine (P.K). The authors thank Michael A. Newton for discussions on ChIP-Seq data.

# A Appendix

## A.1 Re-estimation for one sample problem

Let $B_j, Z_j$ be the latent variables and $B_j \sim$ Bernoulli($\omega$) where $\omega = 1/(\beta+1)$. The E-step of the $k$ iteration involves calculating

$$\begin{aligned}
P(B_j = z, Z_j = 0|Y)^{(k)} &= \frac{P(Y_j|B_j = z, Z_j = 0)\omega^{(k)z}(1-\omega^{(k)})^{1-z}\pi_0^{(k)}}{P(Y_j)}, \\
\text{where } P(Y_j) &= I(Y_j = 0)(1-\omega^{(k)})\pi_0^{(k)} + NB_{Y_j}(a_0, b)\omega^{(k)}\pi_0^{(k)} \\
&\quad + NB_{Y_j}(a_0 + a_1, b)(1-\pi_0^{(k)}), \\
\omega^{(k)} &= \frac{\sum_{j=1}^N P(B_j = 1, Z_j = 0|Y)^{(k-1)}}{\sum_{j=1}^N P(Z_j = 0|Y)^{(k-1)}}, \\
\pi_0^{(k)} &= \frac{\sum_{j=1}^N P(Z_j = 0|Y)^{(k-1)}}{N}.
\end{aligned}$$

For the M-step, we consider MME for re-estimation. Although $b$ is a common parameter for both enriched and non-enriched distribution, we use the non-enriched bins to re-estimate $b$ since they are the majority, and to simplify calculation.

$$\begin{aligned}
a_0 = \frac{\mu_0^2}{\sigma_0^2 - \mu_0}, \qquad b = \frac{\mu_0}{\sigma_0^2 - \mu_0}, \qquad a_1 = \frac{\mu_1^2}{\sigma_1^2 - \mu_1} - a_0 \\
\text{where } \mu_0 = \frac{\sum_{j=1}^N Y_j P(B_j = 1, Z_j = 0|Y)}{\sum_{j=1}^N P(B_j = 1, Z_j = 0|Y)}, \qquad \sigma_0^2 = \frac{\sum_{j=1}^N (Y_j - \mu_0)^2 P(B_j = 1, Z_j = 0|Y)}{\sum_{j=1}^N P(B_j = 1, Z_j = 0|Y)}, \\
\mu_0 = \frac{\sum_{j=1}^N Y_j P(Z_j = 1|Y)}{\sum_{j=1}^N P(Z_j = 1|Y)}, \qquad \sigma_1^2 = \frac{\sum_{j=1}^N (Y_j - \mu_1)^2 P(Z_j = 1|Y)}{\sum_{j=1}^N P(Z_j = 1|Y)}.
\end{aligned}$$

$P(Z_j = z|Y)$ are byproducts of the hidden semi-Markov model.

## A.2 Marginal distributions for mixture model

### A.2.1 Bivariate mixture model

$$P(X, Y|Z = 0, B_1 = 0) = I(X = 0)I(Y = 0),$$

$$
\begin{aligned}
P(X,Y|Z=0,B_1=1,B_2=0) &= I(Y=0)\int P(X|\lambda_1)\pi(\lambda_1)d\lambda_1 \\
&= I(Y=0)\int_0^\infty \frac{\exp(-\lambda_1 D_X)(\lambda_1 D_X)^X}{X!}\frac{b^a\lambda_1^{a-1}\exp(-b\lambda_1)}{\Gamma(a)}d\lambda_1 \\
&= I(Y=0)\frac{\Gamma(X+a)b^a D_X^X}{\Gamma(a)(b+D_X)^{X+a}X!} \\
&= I(Y=0)\mathcal{NB}_X(a,b/D_X),
\end{aligned}
$$

$$
\begin{aligned}
P(X,Y|Z=0,B_1=1,B_2=1) &= \int\int P(X|\lambda_1)P(Y|\lambda_2)f_0(\lambda_1,\lambda_2)\lambda_1 d\lambda_2 \\
&= \int_0^\infty \frac{\exp(-\lambda_1(D_X+D_Y))(\lambda_1 D_X)^X(\lambda_1 D_Y)^Y}{X!Y!}\frac{b^a\lambda_1^{a-1}\exp(-b\lambda_1)}{\Gamma(a)}d\lambda_1 \\
&= \frac{\Gamma(X+Y+a)b^a D_X^X D_Y^Y}{(b+D_X+D_Y)^{X+Y+a}\Gamma(a)X!Y!} \\
&= \binom{X+Y+a-1}{X,Y,a-1}\left(\frac{b}{b+D_X+D_Y}\right)^a\frac{D_X^X D_Y^Y}{(b+D_X+D_Y)^{X+Y}},
\end{aligned}
$$

and

$$
\begin{aligned}
P(X,Y|Z=0) &= P(X,Y|Z=0,B_1=0)P(B_1=0) \\
&\quad +P(X,Y|Z=0,B_1=1,B_2=0)P(B_1=1,B_2=0) \\
&\quad +P(X,Y|Z=0,B_1=1,B_2=1)P(B_1=1,B_2=1),
\end{aligned}
$$

where

$$
\begin{aligned}
P(B_1,B_2) &= \int\int p_1 p_2 g_1(p_1)g_2(p_2)dp_1 dp_2 \\
&= Be\left(\frac{1}{1+\beta_1}\right)Be\left(\frac{1}{1+\beta_2}\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P(X,Y|Z=0) &= I(X=0)I(Y=0)\frac{\beta_1}{1+\beta_1} \\
&\quad +I(Y=0)\mathcal{NB}_X(a,b/D_X)\frac{\beta_2}{(1+\beta_1)(1+\beta_2)} \\
&\quad +\binom{X+Y+a-1}{X,Y,a-1}\left(\frac{b}{b+D_X+D_Y}\right)^a\frac{D_X^X D_Y^Y}{(b+D_X+D_Y)^{X+Y}}\frac{1}{(1+\beta_1)(1+\beta_2)}.
\end{aligned}
$$

Now

$$\begin{aligned}
P(X,Y|Z=1,B_1=0) &= I(X=0)\int P(Y|\theta_1)\pi(\theta_1)d\theta_1 \\
&= I(X=0)\int_0^\infty \frac{\exp(-\theta_1 D_Y)(\theta_1 D_Y)^Y}{Y!}\frac{b^a\theta_1^{a-1}\exp(-b\theta_1)}{\Gamma(a)}d\theta_1 \\
&= I(X=0)\frac{\Gamma(Y+a)b^a D_Y^Y}{\Gamma(a)(b+D_Y)^{Y+a}Y!} \\
&= I(X=0)\mathcal{NB}_Y(a,b/D_Y),
\end{aligned}$$

and

$$\begin{aligned}
P(X,Y|Z=1,B_1=1) &= \int\int P(X|\lambda_1)P(Y|\lambda_2)f_1(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2 \\
&= \int\int P(X|\lambda_1)P(Y|\lambda_2)2\pi(\lambda_1)\pi(\lambda_2)I[\lambda_1<\lambda_2]d\lambda_1 d\lambda_2 \\
&= \int_0^\infty 2P(Y|\lambda_2)\pi(\lambda_2)I(\lambda_2)d\lambda_2,
\end{aligned}$$

where

$$\begin{aligned}
I(\lambda_2) &= \int_0^{\lambda_2} P(X|\lambda_1)\pi(\lambda_1)d\lambda_1 \\
&= \int_0^{\lambda_2} \frac{\exp(-\lambda_1 D_X)(\lambda_1 D_X)^X}{X!}\frac{b^a\lambda_1^{a-1}\exp(-b\lambda_1)}{\Gamma(a)}d\lambda_1 \\
&= \frac{\Gamma(X+a)b^a D_X^X}{\Gamma(a)(b+D_X)^{X+a}X!}\int_0^{\lambda_2}\frac{\exp(-\lambda_1(b+D_X))\lambda_1^{X+a-1}(b+D_X)^{X+a}}{\Gamma(X+a)}d\lambda_1 \\
&= \mathcal{NB}_X(a,b/D_X)\int_0^{\lambda_2} P(\psi_1)d\psi_1,
\end{aligned}$$

and $\psi_1 \sim \mathcal{G}a(X+a, 1/(b+D_X))$. Plugging in $P(X,Y|Z=1)$,

$$\begin{aligned}
& P(X,Y|Z=1,B_1=1) \\
= \; & 2\mathcal{NB}_X(a,b/D_X)\int_0^\infty\int_0^{\lambda_2}\frac{\exp(-\lambda_2 D_Y)(\lambda_2 D_Y)^Y}{Y!}\frac{b^a\lambda_2^{a-1}\exp(-b\lambda_2)}{\Gamma(a)}P(\psi_1)d\psi_1 d\lambda_2 \\
= \; & 2\mathcal{NB}_X(a,b/D_X)\mathcal{NB}_Y(a,b/D_Y)\int_0^\infty\int_{\psi_1}^\infty\frac{\exp(-\lambda_2(b+D_Y))\lambda_2^{Y+a-1}(b+D_Y)^{Y+a}}{\Gamma(Y+a)}d\lambda_2 P(\psi_1)d\psi_1 \\
= \; & 2\mathcal{NB}_X(a,b/D_X)\mathcal{NB}_Y(a,b/D_Y)\int_0^\infty\int_{\psi_1}^\infty P(\psi_2)P(\psi_1)d\psi_2 d\psi_1,
\end{aligned}$$

where $\psi_2 \sim \mathcal{G}a(Y+a, 1/(b+D_Y))$ and $\psi_1 \perp \psi_2$. Let $\omega_1 = (b+D_X)\psi_1$ and $\omega_2 = (b+D_Y)\psi_2$. Thus, $\omega_1 \sim \mathcal{G}a(X+a,1)$, $\omega_2 \sim \mathcal{G}a(Y+a,1)$ and $B = \omega_1/(\omega_1+\omega_2) \sim \mathcal{B}eta(X+a, Y+a)$

and

$$\begin{aligned} P(X,Y|Z=1,B_1=1) &= 2\mathcal{NB}_X(a,b/D_X)\mathcal{NB}_Y(a,b/D_Y)P(\psi_1<\psi_2) \\ &= 2\mathcal{NB}_X(a,b/D_X)\mathcal{NB}_Y(a,b/D_Y)P\left(B<\frac{b+D_X}{2b+D_X+D_Y}\right). \end{aligned}$$

Hence

$$\begin{aligned} P(X,Y|Z=1) &= P(X,Y|Z=1,B_1=0)P(B_1=0)+P(X,Y|Z=1,B_1=1)P(B_1=1) \\ &= I(X=0)\mathcal{NB}_Y(a,b/D_Y)\frac{\beta_1}{1+\beta_1} \\ &\quad +2\mathcal{NB}_X(a,b/D_X)\mathcal{NB}_Y(a,b/D_Y)P\left(B<\frac{b+D_X}{2b+D_X+D_Y}\right)\frac{1}{1+\beta_1}. \end{aligned}$$

### A.2.2 Conditional mixture model

Under $Z=0$, $\lambda_1=\lambda_2$ and $X\sim Po(\lambda_1 D_X)$, $Y\sim Po(\lambda_1 D_Y)$. Thus, $Y|X+Y,Z=0\sim Bin\left(X+Y,\frac{D_Y}{D_X+D_Y}\right)$. On the other hand for $Z=1$, first we obtain

$$f_1(\lambda_1,\lambda_2) = \frac{2\exp(-b\lambda_1)\lambda_1^{a-1}b^a}{\Gamma(a)}\frac{\exp(-b\lambda_2)\lambda_2^{a-1}b^a}{\Gamma(a)}I(\lambda_1<\lambda_1).$$

Let $\omega_1=\lambda_1 D_X$, $\omega_2=\lambda_2 D_Y$. Then

$$\begin{aligned} f_2(\omega_1,\omega_2) &= f_1\left(\frac{\omega_1}{D_X},\frac{\omega_2}{D_Y}\right)\frac{1}{D_X D_Y} \\ &= \frac{2\exp(-b\omega_1/D_X)\omega_1^{a-1}b^a}{\Gamma(a)D_X^a}\frac{\exp(-b\omega_2/D_Y)\omega_2^{a-1}b^a}{\Gamma(a)D_Y^a}I(\omega_1 D_Y<\omega_2 D_X). \end{aligned}$$

Now let

$$\begin{aligned} u &= \omega_1+\omega_2, \\ v &= \frac{\omega_2}{\omega_1+\omega_2}, \\ \Rightarrow \omega_1=u(1-v), &\quad \omega_2=uv. \end{aligned}$$

The Jacobian is $u$.

$$\begin{aligned} f_3(u,v) &= \frac{2\exp(-\frac{b}{D_X}u(1-v))u^{a-1}(1-v)^{a-1}b^a}{\Gamma(a)D_X^a}\frac{\exp(-\frac{b}{D_Y}uv)u^{a-1}v^{a-1}b^a}{\Gamma(a)D_Y^a}I(u(1-v)D_Y<uvD_X)u \\ &= \frac{2b^{2a}\exp(-u[\frac{b}{D_Y}v+\frac{b}{D_X}(1-v)])u^{2a-1}v^{a-1}(1-v)^{a-1}}{D_X^a D_Y^a\Gamma(a)^2}I(v>\frac{D_Y}{D_X+D_Y}). \end{aligned}$$

Back to deriving $P(Y|X+Y,Z=1)$:

$$\begin{aligned}
& P(Y|X+Y,Z=1) \\
= & \int\int P(Y|X+Y,\lambda_1,\lambda_2) f(\lambda_1,\lambda_2|X+Y) d\lambda_1 d\lambda_2 \\
= & \int\int \left( \begin{array}{c} X+Y \\ Y \end{array} \right) \left( \frac{\lambda_2 D_Y}{\lambda_1 D_X + \lambda_2 D_Y} \right)^Y \left( \frac{\lambda_1 D_X}{\lambda_1 D_X + \lambda_2 D_Y} \right)^X f(\lambda_1,\lambda_2|X+Y) d\lambda_1 d\lambda_2 \\
= & \int\int \left( \begin{array}{c} X+Y \\ Y \end{array} \right) v^Y (1-v)^X f(u,v|X+Y) dv du.
\end{aligned}$$

Now

$$f(u,v|X+Y) \;=\; f(X+Y|u) f(u,v) / f(X+Y),$$

since $X+Y \sim Po(u)$.

$$\begin{aligned}
& f(X+Y) \\
= & \int_0^\infty \int_{\frac{D_Y}{D_X+D_Y}}^1 f(X+Y|u) f(u,v) dv du \\
= & \int_0^\infty \int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{\exp(-u) u^{X+Y}}{(X+Y)!} \frac{2b^{2a} \exp(-u[\frac{b}{D_Y} v + \frac{b}{D_X}(1-v)]) u^{2a-1} v^{a-1} (1-v)^{a-1}}{D_X^a D_Y^a \Gamma(a)^2} dv du \\
= & \frac{2b^{2a}}{(X+Y)! D_X^a D_Y^a \Gamma(a)^2} \int_0^\infty \int_{\frac{D_Y}{D_X+D_Y}}^1 \exp\left( -u \left[ 1 + \frac{b}{D_Y} v + \frac{b}{D_X}(1-v) \right] \right) u^{X+Y+2a-1} \\
& v^{a-1} (1-v)^{a-1} dv du \\
= & C_1 \int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{a-1}(1-v)^{a-1} \Gamma(X+Y+2a)}{\left[ 1 + \frac{b}{D_Y} v + \frac{b}{D_X}(1-v) \right]^{X+Y+2a}} dv,
\end{aligned}$$

where $C_1 = \frac{2b^{2a}}{(X+Y)! D_X^a D_Y^a \Gamma(a)^2}$.

Next

$$\begin{aligned}
& \int\int \left( \begin{array}{c} X+Y \\ Y \end{array} \right) v^Y (1-v)^X f(X+Y|u) f(u,v) dv du \\
= & \left( \begin{array}{c} X+Y \\ Y \end{array} \right) C_1 \int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{Y+a-1}(1-v)^{X+a-1} \Gamma(X+Y+2a)}{\left[ 1 + \frac{b}{D_Y} v + \frac{b}{D_X}(1-v) \right]^{X+Y+2a}} dv.
\end{aligned}$$

Thus

$$P(Y|X+Y,Z=1) \;=\; \left( \begin{array}{c} X+Y \\ Y \end{array} \right) \frac{\int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{Y+a-1}(1-v)^{X+a-1}}{\left[ 1+\frac{b}{D_Y} v + \frac{b}{D_X}(1-v) \right]^{X+Y+2a}} dv}{\int_{\frac{D_Y}{D_X+D_Y}}^1 \frac{v^{a-1}(1-v)^{a-1}}{\left[ 1+\frac{b}{D_Y} v + \frac{b}{D_X}(1-v) \right])^{X+Y+2a}} dv}.$$

## A.3 Re-estimation for two sample problem

### A.3.1 Simplified re-estimation for two sample problem

We consider the following approximate re-estimation procedure for two sample problem. We use the data from $X$ to estimate $a$ and $b$ via EM algorithm with latent variable $B_{j1} \sim$ Bernoulli($\omega_1$) where $\omega_1 = 1/(\beta_1 + 1)$. The E-step of the $k$ iteration involves calculating

$$\begin{aligned} P(B_{j1} = 1|X)^{(k)} &= \frac{P(X_j|B_{j1} = 1)\omega_1^{(k)}}{P(X_j)}, \\ \text{where } P(X_j) &= I(X_j = 0)(1 - \omega_1^{(k)}) + NB_{X_j}(a, b/D_X)\omega_1^{(k)}, \\ \omega_1^{(k)} &= \sum_{j=1}^{N} P(B_{j1} = 1|X)^{(k-1)}/N. \end{aligned}$$

For the M-step, we consider MME for re-estimation of $a$ and $b$.

$$a = \frac{\mu^2}{\sigma^2 - \mu}, \qquad b = \frac{\mu D_X}{\sigma^2 - \mu},$$
$$\text{where } \mu = \frac{\sum_{j=1}^{N} X_j P(B_{j1} = 1|X)}{\sum_{j=1}^{N} P(B_{j1} = 1|X)}, \qquad \sigma^2 = \frac{\sum_{j=1}^{N} (X_j - \mu)^2 P(B_{j1} = 1|X)}{\sum_{j=1}^{N} P(B_{j1} = 1|X)}.$$

For given $a$ and $b$, we then estimate $\tilde{\omega}_2$ from

$$\begin{aligned} P(\tilde{B}_{j2} = 1|Y)^{(k)} &= \frac{P(Y_j|\tilde{B}_{j2} = 1)\tilde{\omega}_2^{(k)}}{P(Y_j)}, \\ \text{where } P(Y_j) &= I(Y_j = 0)(1 - \tilde{\omega}_2^{(k)}) + NB_{Y_j}(a, b/D_Y)\tilde{\omega}_2^{(k)}, \\ \tilde{\omega}_2^{(k)} &= \sum_{j=1}^{N} P(\tilde{B}_{j2} = 1|Y)^{(k-1)}/N. \end{aligned}$$

From the data generating process in Section 2.2, $\omega_2 = \tilde{\omega}_2\pi_0/\omega_1$, where $\pi_0 = \sum_{j=1}^{N} P(Z_j = 0|X, Y)$ is from the hidden semi-Markov model.

### A.3.2 Complicated direct re-estimation for two sample problem

Let $B_{j1}, B_{j2}, Z_j$ be the latent variables, $B_{j1} \sim$ Bernoulli($\omega_1$) and $B_{j2} \sim$ Bernoulli($\omega_2$) where $\omega_1 = 1/(\beta_1 + 1)$ and $\omega_2 = 1/(\beta_2 + 1)$. The E-step of the $k$ iteration involves calculating

$$P(B_{j1} = 0, Z_j = 0|Y)^{(k)}, P(B_{j1} = 1, B_{j2} = 0, Z_j = 0|Y)^{(k)},$$
$$P(B_{j1} = 1, B_{j2} = 1, Z_j = 0|Y)^{(k)}, P(B_{j1} = 0, Z_j = 1|Y)^{(k)}, P(B_{j1} = 1, Z_j = 1|Y)^{(k)}.$$

$$
\begin{aligned}
\omega_1^{(k)} &= \frac{\sum_{j=1}^N P(B_{j1}=1, B_{j2}=0, Z_j=0|Y)^{(k-1)}}{N} \\
&\quad + \frac{\sum_{j=1}^N P(B_{j1}=1, B_{j2}=1, Z_j=0|Y)^{(k-1)}}{N} + \frac{\sum_{j=1}^N P(B_{j1}=1, Z_j=1|Y)^{(k-1)}}{N}, \\
\omega_2^{(k)} &= \frac{\sum_{j=1}^N P(B_{j1}=1, B_{j2}=1, Z_j=0|Y)^{(k-1)}}{\sum_{j=1}^N P(B_{j1}=1, B_{j2}=0, Z_j=0|Y)^{(k-1)} + P(B_{j1}=1, B_{j2}=1, Z_j=0|Y)^{(k-1)}}.
\end{aligned}
$$

In the M-step, for fixed $a$ the partial derivative of the expected complete log likelihood with respect to $b$ is given by

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \sum_{j=1}^N P(B_{j1}=1, B_{j2}=0, Z_j=0|Y)\left(\frac{a}{b} - \frac{X_j+a}{b+D_X}\right) \\
&\quad + \sum_{j=1}^N P(B_{j1}=1, B_{j2}=1, Z_j=0|Y)\left(\frac{a}{b} - \frac{X_j+Y_j+a}{b+D_X+D_Y}\right) \\
&\quad + \sum_{j=1}^N P(B_{j1}=0, Z_j=1|Y)\left(\frac{a}{b} - \frac{Y_j+a}{b+D_Y}\right) \\
&\quad + \sum_{j=1}^N P(B_{j1}=1, Z_j=1|Y)\left(\frac{2a}{b} - \frac{X_j+a}{b+D_X} - \frac{Y_j+a}{b+D_Y} + \frac{\partial}{\partial b}\log P\left(B < \frac{b+D_X}{2b+D_X+D_Y}\right)\right),
\end{aligned}
$$

where $B \sim \mathcal{B}eta(X+a, Y+a)$. Now

$$
\frac{\partial}{\partial b}\log P\left(B < \frac{b+D_X}{2b+D_X+D_Y}\right) = \frac{(b+D_X)^{X+a-1}(b+D_Y)^{Y+a}}{(2b+D_X+D_Y)^{X+Y+2a} P\left(B < \frac{b+D_X}{2b+D_X+D_Y}\right)}.
$$

We find the root of the partial derivative $\frac{\partial L}{\partial b}$ for a fixed $a$, and then use `optim` function to find the value of $a$ that maximizes the expected complete log likelihood.

## A.4 Initialization of $p_0$

Let $q$ be an estimate of the percentage of enriched region. Let $m$ be the number of distinct peaks and $E(P)$ be the expected size of a peak. Then

$$
\begin{aligned}
\frac{mE(P)}{L} &= q, \\
\Rightarrow m &= \frac{qL}{E(P)},
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{m+1}{1-p_0} + mE(P) &= L, \\
\Rightarrow p_0 &= 1 - \frac{m+1}{L - mE(P)}.
\end{aligned}
$$

## A.5 Dynamic programming and EM algorithm

We introduce the following notations:

$$\begin{aligned} O_j(i) &:= (X_j(i), Y_j(i)) \text{ (two samples) or } Y_j(i) \text{ (one sample)}, \\ O_k(i), ..., O_r(i) &:= O_k^r(i), \\ b_z(O_j(i)) &:= P(O_j(i)|Z_j(i) = z). \end{aligned}$$

The latent variables consist of $(T_j(i), V_j(i))$, where $T_j(i) = z$ denote 'state $z$ starts at bin $j$' and $V_j(i) = z$ denote 'state $z$ ends at bin $j$'. For notation brevity, we drop the subscript $i$ in the following equations. The key quantities for the new algorithm proposed by Guedon (2003) are:

$$\begin{aligned} F_j(z) &= P(V_j = z|O_1^j) \text{ forward variable}, \\ L1_j(z) &= P(V_j = z|O_1^L), \\ L_j(z) &= P(Z_j = z|O_1^L) \text{ backward variable}. \end{aligned}$$

Define the normalizing factor $N_j$:

$$N_j = P(O_j|O_1^{j-1}).$$

Then

$$\begin{aligned} P(O_1^j) &= \frac{P(O_1^j)}{P(O_1^{j-1})} \frac{P(O_1^{j-1})}{P(O_1^{j-2})} ... \frac{P(O_1^2)}{P(O_1)} P(O_1) \\ &= \prod_{s=1}^{j} N_s, \end{aligned}$$

where $N_1 = P(O_1) = \sum_{z=0}^{1} \pi_z b_z(O_1)$. Let $d_z(w)$ be the duration density at state $z$, where $d_z(w) > 0$ $w = m_z, ..., M_z$.

### A.5.1 Forward recursion

*Initialization:*
For $j = 1$ and $z = 0, 1$:

$$\begin{aligned} F_1(z) &= P(V_1 = z|O_1) \\ &= P(T_1 = z, V_1 = z|O_1) \\ &= \pi_z d_z(1) \frac{b_z(O_1)}{N_1}. \end{aligned}$$

*Induction:*
For $j = 2, ..., L-1$ and $z = 0, 1$:

$$\begin{aligned} F_j(z) &= P(V_j = z|O_1^j) \\ &= P(T_1 = z, V_j = z|O_1^j) + \sum_{k=2}^{j} P(T_k = z, V_j = z|O_1^j) \\ &= \pi_z d_z(j) \prod_{s=1}^{j} \frac{b_z(O_s)}{N_s} + \sum_{k=2}^{j} \left\{ \prod_{s=k}^{j} \frac{b_z(O_s)}{N_s} \right\} d_z(j-k+1) F_{k-1}(1-z), \end{aligned}$$

since

$$\begin{aligned} P(T_1 = z, V_j = z|O_1^j) &= \frac{P(O_1^j|T_1 = z, V_j = z)P(V_j = z|T_1 = z)P(T_1 = z)}{P(O_1^j)} \\ &= \pi_z d_z(j) \frac{\prod_{s=1}^{j} b_z(O_s)}{P(O_1^j)} \\ &= \pi_z d_z(j) \prod_{s=1}^{j} \frac{b_z(O_s)}{N_s}, \end{aligned}$$

and

$$\begin{aligned} P(T_k = z, V_j = z|O_1^j) &= \frac{P(O_1^{k-1}, T_k = z, O_k^j, V_j = z)}{P(O_1^j)} \\ &= \frac{P(O_k^j|T_k = z, V_j = z)P(V_j = z|T_k = z)P(T_k = z|O_1^{k-1})P(O_1^{k-1})}{P(O_1^j)} \\ &= \frac{\prod_{s=k}^{j} b_z(O_s) d_z(j-k+1) P(V_{k-1} = 1-z|O_1^{k-1}) \prod_{s=1}^{k-1} N_s}{\prod_{s=1}^{j} N_s} \\ &= \prod_{s=k}^{j} \frac{b_z(O_s)}{N_s} d_z(j-k+1) F_{k-1}(1-z). \end{aligned}$$

*Termination:*
For $j = L$ and $z = 0, 1$:

$$\begin{aligned} F_L(z) &= P(Z_L = z|O_1^L) \\ &= P(T_1 = z, Z_L = z|O_1^L) + \sum_{k=2}^{j} P(T_k = z, Z_L = z|O_1^L) \\ &= \pi_z D_z(L) \prod_{s=1}^{L} \frac{b_z(O_s)}{N_s} + \sum_{k=2}^{L} \left\{ \prod_{s=k}^{L} \frac{b_z(O_s)}{N_s} \right\} D_z(L-k+1) F_{k-1}(1-z), \end{aligned}$$

where $D_z(L) = \sum_{j \geq L} d_z(j)$.

The normalizing factor $N_j$ is directly obtained during the forward recursion. For $j = 1, ..., L$:

$$\begin{aligned} N_j &= P(O_j|O_1^{j-1}) \\ &= \sum_{z=0}^{1} P(Z_j = z, O_j|O_1^{j-1}) \\ &= \sum_{z=0}^{1} \left[ P(T_1 = z, Z_j = z, O_j|O_1^{j-1}) + \sum_{k=2}^{j} P(T_k = z, Z_j = z, O_j|O_1^{j-1}) \right] \\ &= \sum_{z=0}^{1} \left[ b_z(O_j)\pi_z D_z(j) \prod_{s=1}^{j-1} \frac{b_z(O_s)}{N_s} + \sum_{k=2}^{j} b_z(O_j) \left\{ \prod_{s=k}^{j-1} \frac{b_z(O_s)}{N_s} \right\} D_z(j-k+1) F_{k-1}(1-z) \right]. \end{aligned}$$

### A.5.2 Backward recursion

*Initialization:*
For $j = L$ and $z = 0, 1$:

$$\begin{aligned} L_L(z) &= P(Z_L = z|O_1^L) \\ &= F_L(z). \end{aligned}$$

*Induction:*
For $j = L-1, ..., 1$ and $z = 0, 1$:

$$\begin{aligned} & L1_j(z) \\ = \; & P(V_j = z|O_1^L) \\ = \; & P(V_j = z, Z_L = 1-z|O_1^L) + \sum_{k=j+1}^{L-1} P(V_j = z, V_k = 1-z|O_1^L) \\ = \; & \prod_{s=j+1}^{L} \frac{b_{1-z}(O_s)}{N_s} D_{1-z}(L-j) F_j(z) + \sum_{k=j+1}^{L-1} \left[ \frac{L1_k(1-z)}{F_k(1-z)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j) F_j(z) \right] \\ = \; & \left[ \prod_{s=j+1}^{L} \frac{b_{1-z}(O_s)}{N_s} D_{1-z}(L-j) + \sum_{k=j+1}^{L-1} \left[ \frac{L1_k(1-z)}{F_k(1-z)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j) \right] \right] F_j(z), \end{aligned}$$

since

$$\begin{aligned}
& P(V_j = z, Z_L = 1-z|O_1^L) \\
= & \frac{P(O_{j+1}^L|T_{j+1}=1-z, Z_L=1-z)P(Z_L=1-z|T_{j+1}=1-z)P(V_j=z|O_1^j)P(O_1^j)}{P(O_1^L)} \\
= & \prod_{s=j+1}^{L} \frac{b_{1-z}(O_s)}{N_s} D_{1-z}(L-j)F_j(z),
\end{aligned}$$

and

$$\begin{aligned}
& P(V_j = z, V_k = 1-z|O_1^L) \\
= & \frac{1}{P(O_1^L)} \times P(O_{k+1}^L|T_{k+1}=z) \times P(O_{j+1}^k|T_{j+1}=1-z, V_k=1-z) \\
\times & P(V_k=1-z|T_{j+1}=1-z)P(V_j=z|O_1^j) \times P(O_1^j) \\
= & \frac{P(O_{k+1}^L|T_{k+1}=z)P(O_1^k)}{P(O_1^L)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j)F_j(z) \\
= & \frac{P(O_{k+1}^L, V_k=1-z)}{P(O_{k+1}^L, V_k=1-z)} \frac{P(O_{k+1}^L|V_k=1-z)P(O_1^k)}{P(O_1^L)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j)F_j(z) \\
= & \frac{P(O_1^k, V_k=1-z)}{P(O_1^k, V_k=1-z)} \frac{P(O_{k+1}^L|V_k=1-z)P(O_1^k)}{P(O_1^L)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j)F_j(z) \\
= & \frac{P(O_1^L, V_k=1-z)}{P(O_1^L)} \frac{P(O_1^k)}{P(O_1^k, V_k=1-z)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j)F_j(z) \\
= & \frac{L1_k(1-z)}{F_k(1-z)} \left\{ \prod_{s=j+1}^{k} \frac{b_{1-z}(O_s)}{N_s} \right\} d_{1-z}(k-j)F_j(z).
\end{aligned}$$

Thus

$$\begin{aligned}
L_j(z) &= P(Z_j = z|O_1^L) \\
&= P(Z_j = z, Z_{j+1} = 1-z|O_1^L) + P(Z_j = z, Z_{j+1} = z|O_1^L) \\
&= P(Z_j = z, Z_{j+1} = 1-z|O_1^L) + P(Z_{j+1} = z|O_1^L) - P(Z_j = 1-z, Z_{j+1} = z|O_1^L) \\
&= P(V_j = z|O_1^L) + P(Z_{j+1} = z|O_1^L) - P(T_{j+1} = z|O_1^L) \\
&= L1_j(z) + L_{j+1}(z) - L1_j(1-z).
\end{aligned}$$

Define the following auxiliary variables:

$$\begin{aligned} G^u_{j+1}(z) &= \frac{P(O^L_{j+1}, V_{j+u} = z | T_{j+1} = z)}{P(O^L_{j+1}|O^j_1)} \\ &= \frac{L1_{j+u}(z)}{F_{j+u}(z)} \left\{ \prod_{v=0}^{u-1} \frac{b_z(O_{j+u-v})}{N_{j+u-v}} \right\} d_z(u) \text{ for } u = 1, ..., L-j-1 \\ G^{L-j}_{j+1}(z) &= \frac{P(O^L_{j+1}, Z_L = z | T_{j+1} = z)}{P(O^L_{j+1}|O^j_1)} \\ &= \left\{ \prod_{v=0}^{L-j-1} \frac{b_z(O_{L-v})}{N_{L-v}} \right\} D_z(L-j). \end{aligned}$$

Then

$$\begin{aligned} G_{j+1}(z) &= \frac{P(O^L_{j+1}|T_{j+1} = z)}{P(O^L_{j+1}|O^j_1)} \\ &= \sum_{u=1}^{L-j} G^u_{j+1}(z), \end{aligned}$$

and

$$\begin{aligned} L1_j(z) &= \sum_{u=1}^{L-j} G^u_{j+1}(1-z) F_j(z) \\ &= G_{j+1}(1-z) F_j(z). \end{aligned}$$

### A.5.3 E-step

Expected complete log likelihood is given by

$$\begin{aligned} \mathcal{L}_{EC} = E[\log P(O^L_1, Z^L_1, T^L_1, V^L_1 | \theta)] &= \sum_{z=0}^{1} P(T_1 = z | O^L_1, \theta) \log \pi_z \\ &+ \sum_{z=0}^{1} \sum_{j=0}^{L-1} \sum_{u \geq 1} P(T_{j+1} = z, V_{j+u} = z | O^L_1, \theta) \log d_z(u) \\ &+ \sum_{z=0}^{1} \sum_{j=1}^{L} P(Z_j = z | O^L_1, \theta) \log b_z(O_j), \end{aligned}$$

where

$$P(T_1 = z | O^L_1, \theta) = L_1(z).$$

For $j = 1, ..., L-1$ and $u = 1, ..., L-j-1$ :

$$\begin{aligned} P(T_{j+1} = z, V_{j+u} = z|O_1^L, \theta) &= \frac{P(T_{j+1} = z, V_{j+u} = z, O_1^L|\theta)}{P(O_1^L|\theta)} \\ &= \frac{P(O_{j+1}^L, V_{j+u} = z|T_{j+1} = z, \theta)P(T_{j+1} = z|O_1^j, \theta)P(O_1^j, \theta)}{P(O_1^L|\theta)} \\ &= \frac{P(O_{j+1}^L, V_{j+u} = z|T_{j+1} = z, \theta)P(V_j = 1-z|O_1^j, \theta)}{P(O_{j+1}^L|O_1^j, \theta)} \\ &= G_{j+1}^u(z)F_j(1-z). \end{aligned}$$

For $j = 1, ..., L-1$ and $u \geq L-j$ :

$$\begin{aligned} & P(T_{j+1} = z, V_{j+u} = z|O_1^L, \theta) \\ = & \frac{P(T_{j+1} = z, V_{j+u} = z, O_1^L|\theta)}{P(O_1^L|\theta)} \\ = & \frac{P(O_{j+1}^L|T_{j+1} = z, V_{j+u} = z, \theta)P(V_{j+u} = z|T_{j+1} = z, \theta)P(V_j = 1-z|O_1^j, \theta)}{P(O_{j+1}^L|O_1^j, \theta)} \\ = & \left\{ \prod_{v=0}^{L-j-1} \frac{b_z(O_{L-v})}{N_{L-v}} \right\} d_z(u)F_j(1-z). \end{aligned}$$

For $j = 0$ and $u = 1, ..., L-1$:

$$\begin{aligned} & P(T_1 = z, V_u = z|O_1^L, \theta) \\ = & \frac{P(T_1 = z, V_u = z, O_1^L|\theta)}{P(O_1^L|\theta)} \\ = & \frac{P(O_{u+1}^L|V_u = z, \theta)P(V_u = z|T_1 = z, \theta)P(O_1^u|T_1 = z, V_u = z, \theta)P(T_1 = z|\theta)}{P(O_1^L|\theta)} \\ = & \frac{P(O_1^u, V_u = z|\theta)}{P(O_1^u, V_u = z|\theta)} \frac{P(O_{u+1}^L|V_u = z, \theta)P(O_1^u|\theta)}{P(O_1^L|\theta)} \left\{ \prod_{v=0}^{u-1} \frac{b_z(O_{u-v})}{N_{u-v}} \right\} d_z(u)\pi_z \\ = & \frac{P(O_1^L, V_u = z|\theta)P(O_1^u|\theta)}{P(O_1^u, V_u = z|\theta)P(O_1^L|\theta)} \left\{ \prod_{v=0}^{u-1} \frac{b_z(O_{u-v})}{N_{u-v}} \right\} d_z(u)\pi_z \\ = & \frac{P(V_u = z|O_1^L, \theta)}{P(V_u = z|O_1^u, \theta)} \left\{ \prod_{v=0}^{u-1} \frac{b_z(O_{u-v})}{N_{u-v}} \right\} d_z(u)\pi_z \\ = & \frac{L1_u(z)}{F_u(z)} \left\{ \prod_{v=0}^{u-1} \frac{b_z(O_{u-v})}{N_{u-v}} \right\} d_z(u)\pi_z. \end{aligned}$$

For $j = 0$ and $u \geq L$:

$$P(T_1 = z, V_u = z|O_1^L, \theta) = \left\{ \prod_{v=0}^{L-1} \frac{b_z(O_{L-v})}{N_{L-v}} \right\} d_z(u)\pi_z.$$

The quantities $P(T_{j+1} = 1, V_{j+u} = 1|O_1^L, \theta)$ for $j = 0, ..., L-1$ and $u = m_1, ..., M_1$ will be used to infer the most probable boundaries of enriched regions. Finally

$$P(Z_j = z|O_1^L, \theta) \quad = \quad L_j(z).$$

### A.5.4 M-step

Maximizing the expected complete likelihood,

$$\begin{aligned} &\max \mathcal{L}_{EC} \\ &s.t \sum_{Z=0}^{1} \pi_z = 1 \end{aligned}$$

yields

$$\begin{aligned} \hat{\pi}_z &= P(T_1 = z|O_1^L, \theta) \\ \hat{p}_0 &= \frac{\sum_{j=0}^{L-1} \sum_{u \geq 1} P(T_{j+1} = 0, V_{j+u} = 0|O_1^L, \theta)(u-1)}{\sum_{j=0}^{L-1} \sum_{u \geq 1} P(T_{j+1} = 0, V_{j+u} = 0|O_1^L, \theta)u} \\ &= 1 - \frac{\sum_{j=0}^{L-1} \sum_{u \geq 1} P(T_{j+1} = 0, V_{j+u} = 0|O_1^L, \theta)}{\sum_{j=0}^{L-1} \sum_{u \geq 1} P(T_{j+1} = 0, V_{j+u} = 0|O_1^L, \theta)u}, \end{aligned}$$

where

$$\begin{aligned} \sum_{j=0}^{L-1} \sum_{u \geq 1} P(T_{j+1} = 0, V_{j+u} = 0|O_1^L, \theta) &= \sum_{j=1}^{L-1} P(V_j = 1|O_1^L, \theta) + P(T_1 = 0|O_1^L, \theta) \\ &= \sum_{j=1}^{L-1} L1_j(1) + L_1(0), \end{aligned}$$

and

$$\begin{aligned}
&\sum_{j=0}^{L-1}\sum_{u\geq 1} P(T_{j+1}=0, V_{j+u}=0|O_1^L,\theta)u \\
=\;& \sum_{j=0}^{L-1}\sum_{r\geq j+1} P(T_{j+1}=0, V_r=0|O_1^L,\theta)(r-j) \\
=\;& \sum_{j=0}^{L-1}\sum_{r\geq j+1} rP(T_{j+1}=0, V_r=0|O_1^L,\theta) - \sum_{j=0}^{L-1} jP(T_{j+1}=0|O_1^L,\theta) \\
=\;& \sum_{r\geq 1} r\sum_{j=0}^{r-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) - \sum_{j=1}^{L-1} jP(V_j=1|O_1^L,\theta) \\
=\;& \sum_{r\geq 1} rP(V_r=0|O_1^L,\theta) - \sum_{j=1}^{L-1} jL1_j(1) \\
=\;& \sum_{r=1}^{L-1} rL1_r(0) + \sum_{r\geq L} r\sum_{j=0}^{r-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) - \sum_{j=1}^{L-1} jL1_j(1).
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{r\geq L} r\sum_{j=0}^{r-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) \\
=\;& \sum_{r\geq L} rP(T_1=0, V_r=0|O_1^L,\theta) + \sum_{r\geq L} r\sum_{j=1}^{r-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) \\
=\;& \left[\frac{1}{1-p_0^{old}} - \sum_{r=1}^{L-1} rd_0(r)\right]\left[\pi_0\left\{\prod_{v=0}^{L-1}\frac{b_0(O_{L-v})}{N_{L-v}}\right\} + \sum_{j=1}^{L-1}\left\{\prod_{v=0}^{L-j-1}\frac{b_z(O_{L-v})}{N_{L-v}}\right\}\frac{F_j(1)}{(p_0^{old})^j}\right] \\
=\;& \left[\frac{1}{1-p_0^{old}} - \sum_{r=1}^{L-1} rd_0(r)\right]\mathtt{denom}_{j=0}^{L-1},
\end{aligned}$$

and

$$\begin{aligned}
\sum_{r\geq L} rP(T_1=0, V_r=0|O_1^L,\theta) &= \sum_{r\geq L}\pi_0 d_0(r) r\left\{\prod_{v=0}^{L-1}\frac{b_0(O_{L-v})}{N_{L-v}}\right\} \\
&= \pi_0\left\{\prod_{v=0}^{L-1}\frac{b_0(O_{L-v})}{N_{L-v}}\right\}\left[\frac{1}{1-p_0^{old}} - \sum_{r=1}^{L-1} rd_0(r)\right],
\end{aligned}$$

$$
\begin{aligned}
& \sum_{r\geq L} r \sum_{j=1}^{r-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) \\
= \; & L\sum_{j=1}^{L-1} P(T_{j+1}=0, V_L=0|O_1^L,\theta) + \sum_{r\geq L+1} r \sum_{j=1}^{L-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) \\
= \; & \sum_{r\geq L} r \sum_{j=1}^{L-1} \left\{ \prod_{v=0}^{L-j-1} \frac{b_z(O_{L-v})}{N_{L-v}} \right\} d_0(r-j) F_j(1) \\
= \; & \sum_{j=1}^{L-1} \left\{ \prod_{v=0}^{L-j-1} \frac{b_z(O_{L-v})}{N_{L-v}} \right\} \frac{F_j(1)}{(p_0^{old})^j} \sum_{r\geq L} r d_0(r) \\
= \; & \sum_{j=1}^{L-1} \left\{ \prod_{v=0}^{L-j-1} \frac{b_z(O_{L-v})}{N_{L-v}} \right\} \frac{F_j(1)}{(p_0^{old})^j} \left[ \frac{1}{1-p_0^{old}} - \sum_{r=1}^{L-1} r d_0(r) \right].
\end{aligned}
$$

`denom` is computed recursively. However, when $L$ is large,

$$
\begin{aligned}
\sum_{r\geq L} r \sum_{j=0}^{r-1} P(T_{j+1}=0, V_r=0|O_1^L,\theta) &= \left[ \frac{1}{1-p_0^{old}} - \sum_{r=1}^{L-1} r d_0(r) \right] \texttt{denom}_{j=0}^{L-1} \\
&\approx 0,
\end{aligned}
$$

since $\sum_{r=1}^{L-1} r d_0(r) \approx \frac{1}{1-p_0^{old}}$. Thus,

$$
\hat{p}_0 \approx 1 - \frac{\sum_{j=1}^{L-1} L1_j(1) + L_1(0)}{\sum_{r=1}^{L-1} r[L1_r(0) - L1_r(1)]}.
$$

### A.5.5 Viterbi algorithm for hidden semi-Markov model

Define

$$
\delta_j(z) = \max_{Z_1,\ldots,Z_{j-1}} \log P(O_1^j, V_j = z|\theta).
$$

For $j=1$ and $z=0,1$:

$$
\delta_j(z) = \log b_z(O_1) + \log d_z(1) + \log \pi_z.
$$

For $j=2,\ldots,L-1$ and $z=0,1$:

$$
\begin{aligned}
& \delta_j(z) \\
= \; & \log b_z(O_j) + \max[\max_{1\leq u\leq j-1} \left[ \left\{ \sum_{v=1}^{u-1} \log b_z(O_{j-v}) \right\} + \log d_z(u) + \delta_{j-u}(1-z) \right], \\
& \left\{ \sum_{v=1}^{j-1} \log b_z(O_{j-v}) \right\} + \log d_z(j) + \log \pi_z].
\end{aligned}
$$

For $j = L$ and $z = 0, 1$:

$$\begin{aligned}
& \delta_j(z) \\
= \; & \log b_z(O_L) + \max[\max_{1 \le u \le L-1} \left[ \left\{ \sum_{v=1}^{u-1} \log b_z(O_{L-v}) \right\} + \log D_z(u) + \delta_{L-u}(1-z) \right], \\
& \left\{ \sum_{v=1}^{L-1} \log b_z(O_{L-v}) \right\} + \log D_z(L) + \log \pi_z].
\end{aligned}$$

The likelihood optimal state sequence associated with the observations $O_1^L$ is $\exp[\max_z\{\delta_z(L)\}]$. For backtracking purposes, define

$$\begin{aligned}
& \psi_j(z) \\
= \; & \text{argmax}[\log b_z(O_j) + \max[\max_{1 \le u \le j-1} \left[ \left\{ \sum_{v=1}^{u-1} \log b_z(O_{j-v}) \right\} + \log d_z(u) + \delta_{j-u}(1-z) \right], \\
& \left\{ \sum_{v=1}^{j-1} \log b_z(O_{j-v}) \right\} + \log d_z(j) + \log \pi_z]].
\end{aligned}$$

# References


Bailey, T. and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using em. *Machine Learning* **21,** 51–80.

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,** 823–837.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57,** 289–300.

Besser, D. (2004). Expression of nodal, lefty-a, and lefty-b in undifferentiated human embryonic stem cells requires activation of smad2/3. *J. Biol. Chem.* **279,** 45076–45084.

Gottardo, R., Li, W., Johnson, W., and Liu, X. (2008). A flexible and powerful bayesian hierarchical model for chip-chip experiments. *Biometrics* **64,** 468–478.

Guedon, Y. (2003). Estimating hidden semi-markov chains from discrete sequences. *Journal of Computational and Graphical Statistics* **12,** 604–639.

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

Ji, H. and Wong, W. (2005). Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21,** 3629–3636.

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

Keleş, S., van der Laan, M., Dudoit, S., Xing, B., and Eisen, M. (2003). Supervised detection of regulatory motifs in dna sequences. *Statistical Applications in Genetics and Molecular Biology* **2,**.

Keleş, S., van der Lann, M., Dudoit, S., and Cawley, S. (2006). Multiple testing methods for ChIP-chip high density oligonucleotide array data. *Journal of Computational Biology* **13,** 579–613.

Ku, M., Koche, R., Rheinbay, E., Mendenhall, E., Endoh, M., Mikkelsen, T., Presser, A., Nusbaum, C., Xie, X., Chi, A., Adli, M., Kasif, S., Ptaszek, L., Cowan, C., Lander, E., Koseki, H., and Bernstein, B. (2008). Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. *PLoS Genetics* **4,**.

Lambert, D. (1992). Zero-inflated poisson regression models with an application to defects in manufacturing. *Technometrics* **34,** 1–14.

Mikkelsen, T., Ku, M., Jaffe, D., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E., and Bernstein, B. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 653–560.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5,** 155–176.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G., Stewart, R., and Thomson, J. (2007). Whole-genome analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in human embryonic stem cells. *Cell Stem Cell* **1,** 299–312.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77,** 257–286.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* .

Ross, S. and Hill, C. (2008). How the smads regulate transcription. *Int J Biochem Cell Biol* **40,** 383–408.

Tam, P. and Loebel, D. (2007). Gene function in mouse embryogenesis: get set for gastrulation. *Nature Reviews Genetics* **8,** 368–381.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J., Costa, G., McKernan, K., Sidow, A., Fire, A., and Johnson, S. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research* **18,** 1051–1063.

Wei, H., Kuan, P., Tian, S., Yang, C., Nie, J., Sengupta, S., Ruotti, V., Jonsdottir, G., Keles, Ş., Thomson, J., and Stewart, R. (2008). A study of the relationships between oligonucleotide properties and hybridization signal intensities from nimblegen microarray datasets. *Nucleic Acids Research* **36,** 2926–2938.

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

Zhang, Z., Rozowsky, J., Snyder, M., Chang, J., and Gerstein, M. (2008). Modeling chip sequencing in silico with applications. *PLoS Computational Biology* **4,** e1000158.