Efficient Detection and Classification of Epigenomic Changes Under Multiple Conditions

Pedro L. Baldoni
Naim U. Rashid
Joseph G. Ibrahim
Department of Biostatistics, University of North Carolina at Chapel Hill

December 3, 2019

Abstract

Epigenomics, the study of the human genome and its interactions with proteins and other cellular elements, has become of significant interest in recent years. Such interactions have been shown to regulate essential cellular functions and are associated with multiple complex diseases. Therefore, understanding how these interactions may change across conditions is central in biomedical research. Chromatin immunoprecipitation followed by massively-parallel sequencing (ChIP-seq) is one of several techniques to detect local changes in epigenomic activity (peaks). However, existing methods for differential peak calling are not optimized for the diversity in ChIP-seq signal profiles, are limited to the analysis of two conditions, or cannot classify specific patterns of differential change when multiple patterns exist. To address these limitations, we present a flexible and efficient method for the detection of differential epigenomic activity across multiple conditions. We utilize data from the ENCODE Consortium and show that the presented method, mixNBHMM, exhibits superior performance to current tools and it is among the fastest algorithms available. Our method also allows the classification of combinatorial patterns of differential epigenomic activity, and the characterization of chromatin regulatory states. Supplementary materials, including code to replicate the results, are available as an online supplement. The implemented R package is available for download at https://github.com/plbaldoni/mixNBHMM.

Keywords: Epigenomics; ChIP-seq; Differential Peak Call; Hidden Markov Model; Mixture Model.

1 Introduction

Epigenomics, the study of the human genome and its interactions with proteins and other cellular elements (Fazzari & Greally 2004), has become of significant interest in recent years. Such interactions have been shown to regulate essential cellular functions such as gene expression and DNA packaging (Kim et al. 2018), resulting in downstream phenotypic impact. Therefore, the interrogation of how these interactions may change across conditions, such as cell types or treatments, is of marked interest in biomedical research. For example, several landmark articles have identified specific genomic regions of changing (differential) epigenomic activity between conditions as drivers of cell differentiation (Creyghton et al. 2010), breast cancer progression (Varambally et al. 2002), and a number of human diseases (Portela & Esteller 2010). Within such differential regions, the delineation of specific patterns of change across conditions is also of interest, for example classifying the gain-of- or loss-of-activity in genomic loci due to treatment (Clouaire et al. 2014). Lastly, multiple epigenomic processes may work together to regulate cellular functions (Cui et al. 2009) and the identification of specific combinations of processes acting locally may also be informative, such as for segmenting the genome into regulatory states (Kundaje et al. 2015). Therefore, the development of robust statistical tools to detect genomic regions of differential epigenomic activity, as well as specific patterns of differential activity, would be of great utility in biomedical research.

To quantify local epigenomic activity, a commonly used high-throughput assay is chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq). This assay is often applied to detect genomic regions bound to a specific protein of interest, such as transcription factors (Shukla et al. 2011) or modified histones (Creyghton et al. 2010). ChIP-seq experiments begin with cross-linking DNA and proteins within chromatin structures, which are then fragmented by sonication in a particular sample. DNA fragments bound to the protein of interest are isolated by chromatin immunoprecipitation, which are then sequenced via massively parallel high throughput sequencing to generate short sequencing reads pertaining to the original fragments. These sequences are then mapped onto a reference genome through sequence alignment to determine their likely genomic locations of origin. Genomic coordi-

nates containing a high density of mapped reads, often referred to as "enrichment" regions (peaks), indicate likely locations of protein-DNA interaction sites, and all other regions are referred to as "background" regions. This local read density is often summarized by counting the number of mapped reads falling into non-overlapping windows of fixed length tiling the genome ("window read counts"), forming the basis for downstream modeling approaches. Across multiple conditions, regions exhibiting enrichment in at least one condition, but not across all conditions, indicate the presence of differential epigenomic activity pertaining to the protein-DNA interaction of interest.

To detect differential epigenomic activity in samples from different conditions, many differential peak callers (DPCs) have been proposed (Song & Smith 2011, Stark & Brown 2011, Shen et al. 2013, Chen et al. 2015, Lun & Smyth 2015, Allhoff et al. 2016). However, several challenges affect the ability of existing methods to accurately detect regions of differential activity from the wide range of epigenomic ChIP-seq experiments currently available (Section 2). First, differential regions may be both short or broad in length, causing difficulty for methods that are optimized for a particular type of signal profile (Stark & Brown 2011, Chen et al. 2015). Second, methods that pool experimental replicates together (Song & Smith 2011) often exhibit larger false positive calls compared to methods that jointly model individual replicates from each condition (Steinhauser et al. 2016). Third, the analysis of ChIP-seq data exhibiting broad enrichment regions is often subject to complex trended biases that may vary across the genome, as technical differences in local read enrichment may vary with respect to the overall read abundance in a given region across samples. Therefore, DPCs that solely rely on sample-specific global scaling factors or control subtraction methods (Stark & Brown 2011, Shen et al. 2013, Chen et al. 2015, Allhoff et al. 2016) may be prone to detecting spurious differences due to the lack of non-linear normalization methods (Lun & Smyth 2015). Reflecting these modeling limitations, a recent comprehensive comparison of DPCs demonstrated that current methods tend to detect either a large number of short peaks (low sensitivity) or exhibit a high number of false positive calls (low specificity) in ChIP-seq experiments with broad regions of enrichment (Steinhauser et al. 2016). Furthermore, few methods are also able to test for differential activity across three or more conditions (Lun & Smyth 2015), or can classify specific differential combinatorial patterns.

Altogether, these limitations can impact the drawing of relevant and accurate insights from modern epigenomic studies.

To address these issues, we propose an efficient and flexible statistical method to identify differential regions from epigenomic experiments with diverse signal profiles and collected under common multi-replicate, multi-condition settings. Our method aims to overcome the above limitations of current DPCs with three major features. First, it uses a hidden Markov model (HMM) to account for the diversity in differential enrichment profiles that may result from short and broad epigenomic ChIP-seq data sets. Second, our method captures specific differential combinatorial patterns through a novel finite mixture model emission distribution within the HMM's differential state. Here, each mixture component pertains to a particular differential combinatorial pattern that is formed by the presence or absence of local enrichment across conditions, where a generalized linear model with ANOVA parameterization is used to model the specific differential combinatorial pattern and account for sample-specific normalization factors via GLM offsets. Third, our modeling framework enables the detection and classification of epigenomic changes under three or more conditions, a novelty not yet available in any other DPC algorithm. Through simulations and a case study using ChIP-seq data sets from the Encyclopedia of DNA Elements (ENCODE) Consortium (Landt et al. 2012), we show that our method exhibits superior performance in detecting broad and short differential combinatorial patterns compared to existing DPCs, and performs comparably in terms of computation time. Lastly, we show that our method can be utilized for genomic regulatory state segmentation when comparing multiple types of epigenomic processes from the same cell line, confirming results from known chromatin biology. In all, our proposed approach comprehensively addresses current gaps in the detection of epigenomic changes under multiple conditions and reveals biological insights not possible under existing methods.

2 Data

Histones are proteins that interact and condense DNA in eukaryotic cells into structural units called nucleosomes. Multiple types of enzymatic modifications may be applied to histones, resulting in changes in local DNA packaging and chromatin accessibility mediated

by nucleosomes (Bannister & Kouzarides 2011). In turn, cellular processes such as gene transcription, gene silencing, DNA repair, replication, and recombination are also affected (Bannister & Kouzarides 2011). Proteins that interact with DNA and alter its functional properties are often referred to as "epigenomic marks". For example, the trimethylation of histone H3 at lysines 36 and 27 (H3K36me3 and H3K27me3) are two types of histone modifications that tend to occur in genomic loci containing actively transcribed and repressed genes (Liu et al. 2016), respectively, and exhibit broad enrichment profiles. These marks have been investigated in hematopoiesis and cancer studies, where their absence is often observed in multiple cancer types (Cui et al. 2009, Pfister et al. 2015, Wei et al. 2008). As a result, H3K36me3 and H3K27me3 are considered to be key prognostic indicators in patients with breast, ovarian, and pancreatic cancer. The protein EZH2, a major component of the polycomb complex PRC2 that catalyzes the methylation of H3K27me3 (Margueron & Reinberg 2011), is another example of an epigenomic mark with experimental signal characterized by broad enrichment domains and co-occurs with the activity of H3K27me3.

Using ChIP-seq data pertaining to histone modifications H3K27me3, H3K36me3, and the enhancer EZH2 from the ENCODE Consortium, we find that current DPCs have difficulty in accurately detecting broad regions of differential enrichment between several common cell lines, such as Helas3, Hepg2, Huvec, and H1hesc (Figure 1). In line with previous findings (Steinhauser et al. 2016), we observe that current DPCs tend to detect either overly fragmented differential peaks or call regions exhibiting no difference in experimental signal between conditions as differential (Figure 1A). The low specificity and sensitivity of such methods may impair the biological interpretation of the resulting peak calls in downstream analyses. In addition, nearly all current DPCs restrict their application to the analysis of two experimental conditions. For methods that are tailored for the analysis of three or more conditions, the classification of specific differential combinatorial patterns across conditions (or across various epigenomic processes) is still an open problem. The classification of such patterns would allow researchers to, for example, quantify treatment responses on the epigenomic level (Clouaire et al. 2014), or identify sets of processes working together to regulate local chromatin state. Lastly, we find that the performance of such methods also exhibit low sensitivity and specificity in calling differential regions in broad marks (Figure 1B).

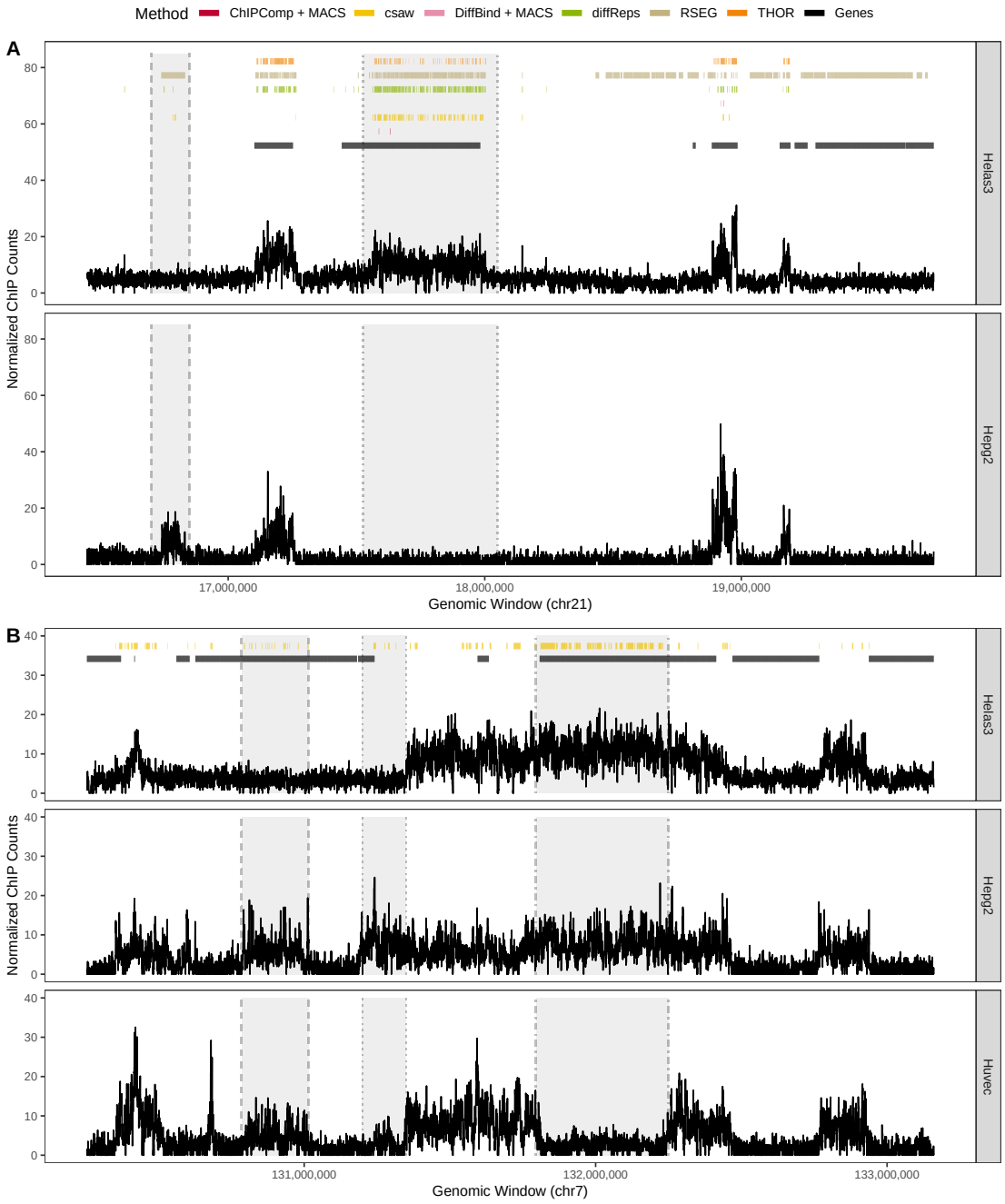

Figure 1: Performance of current DPC methods on calling differential enrichment regions in broad marks under a false discovery rate control of 0.05. (A): Differential peak calls between cell lines Helas3 and Hepg2 for the H3K36me3 histone modification. (B): Differential peak calls between cell lines Helas3, Hepg2, and Huvec for the H3K27me3 histone modification. Shaded regions indicate observed differential enrichment, and each vertical line type bordering each region represents a different combinatorial pattern of enrichment across cell lines. Optimal DPCs would call broad peaks inside shaded regions and no peaks outside them. Csaw is the only DPC able to analyze data from more than two conditions.

In this article, we propose an efficient and flexible tool for the detection and classification of differential epigenomic activity under multiple conditions. We also assess the performance of our proposed model on ChIP-seq experiments characterized by short peaks, namely the histone modifications H3K27ac, H3K4me3, and the transcription factor CTCF (Section 5.2). We show that our model addresses the issues of the current peak callers in broad data, while being flexible for short peaks, and is comparable to the fastest DPCs regarding the computation time. The Online Supplementary Materials A1 and A2 present the data accession codes and the data pre-processing steps, respectively. Code implementing the proposed method and to replicate the presented results are available in the Online Supplementary Materials A3 and A4, respectively.

3 Methods

3.1 Statistical Model

Let Y_{hij} denote the random variable pertaining to the ChIP read count for genomic window j from sample i of condition h, where $j = 1, \ldots, M$, $i = 1, \ldots, n_h$, $h = 1, \ldots, G$, and let y_{hij} be the observed count. Here, n_h is the number of samples in condition h and $N = \sum_{h=1}^{G} n_h$ is the total number of samples across the G conditions. At the jth window, let $\mathbf{y}_{..j} = (y_{11j}, \ldots, y_{Gn_Gj})'$ denote the $N \times 1$ vector of ChIP window read counts across all samples and conditions, and let $\mathbf{y} = (\mathbf{y}'_{..1}, \ldots, \mathbf{y}'_{..M})'$ denote the corresponding $NM \times 1$ vector of window read counts spanning all genomic windows, samples, and conditions.

In our modeling framework, we assume that each window j belongs to one of three possible hidden states: consensus background (state 1), differential (state 2), and consensus enrichment (state 3). Specifically, windows exhibiting low (high) enrichment across all conditions will be modeled by an emission distribution pertaining to the consensus background (consensus enrichment) state. Conversely, windows exhibiting enrichment under at least one condition, but not all conditions, will be modeled by an emission distribution pertaining to the differential state. Assuming that G conditions are of interest, there are $L = 2^G - 2$ possible differential combinatorial patterns of enrichment and background across the G con-

ditions at a given genomic window. The emission distribution pertaining to the differential state of the proposed HMM models all L possible differential combinatorial patterns using a finite mixture model with mixture proportions $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)'$, such that $\sum_{l=1}^{L} \delta_l = 1$ (see Figure B.1 in the Online Supplementary Material B1).

To model transitions between states, we assume a single latent discrete time stationary Markov chain $\mathbf{Z} = \{Z_j\}_{j=1}^{M}$, $Z_j \in \{1, 2, 3\}$, with state-to-state transition probabilities $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{33})'$ and initial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$, such that $\sum_{s=1}^{3} \gamma_{rs} = 1$ and $\sum_{s=1}^{3} \pi_s = 1$ for $r \in \{1, 2, 3\}$. To facilitate the notation, let $f_r(\mathbf{y}_{..j}|\boldsymbol{\psi}_r)$ denote the emission distribution corresponding to the rth hidden state, where $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \boldsymbol{\delta}', \boldsymbol{\psi}')'$ denotes the vector of all model parameters, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \boldsymbol{\psi}_2', \boldsymbol{\psi}_3')'$ denotes each state's set of emission distribution-specific parameters, and $\mathcal{Z}$ denotes the set of 3^M possible state paths of length M. Then, the likelihood function pertaining to the proposed HMM may be written as

$$f(\mathbf{y}|\mathbf{x}; \boldsymbol{\Psi}) = \sum_{\mathbf{Z} \in \mathcal{Z}} \left\{ \prod_{r=1}^{3} \pi_r^{I(Z_1=r)} \times \left(\prod_{j=2}^{M} \prod_{r=1}^{3} \prod_{s=1}^{3} \gamma_{rs}^{I(Z_{j-1}=r, Z_j=s)} \right) \times \right. \tag{1}$$
$$\left. \times \left(\prod_{j=1}^{M} f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right)^{I(Z_j=1)} f_2\left(\mathbf{y}_{..j}|\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\psi}_2\right)^{I(Z_j=2)} f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right)^{I(Z_j=3)} \right) \right\}.$$

Here, $\mathbf{x}$ is a fixed $G \times L$ design matrix enumerating each of the L possible differential combinatorial patterns in terms of the presence or absence of enrichment across each of the G conditions, only in the emission distribution of the differential state.

We assume that read counts pertaining to genomic windows from the consensus background ($r = 1$) and consensus enrichment ($r = 3$) states follow a Negative Binomial (NB) distribution with state-specific parameters $\boldsymbol{\psi}_r = (\mu_{(r,hij)}, \phi_r)'$, with mean $\mu_{(r,hij)}$ and variance $\mu_{(r,hij)}(1 + \mu_{(r,hij)}/\phi_r)$. Assuming independence of read counts across experiments and samples, conditional upon the HMM state, the emission distribution of the consensus background and consensus enrichment states, respectively, can be written as

$$f_r(\mathbf{y}_{..j}|\boldsymbol{\psi}_r) = \prod_{h=1}^{G} \prod_{i=1}^{n_h} \frac{\Gamma(y_{hij} + \phi_r)}{y_{hij}!\Gamma(\phi_r)} \left(\frac{\phi_r}{\mu_{(r,hij)} + \phi_r} \right)^{\phi_r} \left(\frac{\mu_{(r,hij)}}{\mu_{(r,hij)} + \phi_r} \right)^{y_{hij}}, \tag{2}$$

for $r \in \{1, 3\}$ and $y_{hij} \in \{0, 1, 2, \ldots\}$, such that $\log(\mu_{(1,hij)}) = \beta_1 + u_{hij}$, $\log(\phi_1) = \lambda_1$, $\log(\mu_{(3,hij)}) = \beta_1 + \beta_3 + u_{hij}$, and $\log(\phi_3) = \lambda_1 + \lambda_3$. The term u_{hij} is an offset that adjusts for technical artifacts and allows the non-linear normalization of the signal profile across

genomic windows, conditions, and samples (see the Online Supplementary Material B2 for details). When $u_{hij} = 0$, β_1 and λ_1 represent the log-mean and log-dispersion, respectively, of read counts pertaining to consensus background state windows, whereas β_3 and λ_3 represent the difference in log-mean and log-dispersion of read counts from consensus enrichment state windows relative to consensus background state windows.

For windows belonging to the differential state ($r = 2$), we assume that the corresponding read counts are modeled by a L-component finite mixture model with mixture components that follow a Negative Binomial distribution, where each component corresponds to a particular differential combinatorial pattern. To define these patterns, let us consider the sets $S_1, \ldots, S_L$ that delineate the subset of the G conditions that are enriched in each of the L differential combinatorial patterns. For instance, if $G = 3$, the sets $S_1 = \{1\}$, $S_2 = \{2\}$, $S_3 = \{3\}$, $S_4 = \{1, 2\}$, $S_5 = \{1, 3\}$, and $S_6 = \{2, 3\}$ define the six possible differential combinatorial patterns of enrichment and background across three conditions. That is, the set S_1 denotes enrichment in only the first condition and background in all others, whereas the set S_6 denotes enrichment in conditions 2 and 3 and background in condition 1. The presence or absence of enrichment in each of the L sets is encoded into each column of $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, such that $\mathbf{x}_l = (x_{1l}, \ldots, x_{Gl})'$, and $x_{hl} = I(h \in S_l)$ for $l = 1, \ldots, L$ and $h = 1, \ldots, G$. That is, $\mathbf{x}_l$ is the $G \times 1$ vector of binary indicator variables denoting which subset of conditions are enriched in pattern (mixture component) l. A graphical illustration of our proposed mixture model and its design is provided in Figure B.1 in the Online Supplementary Material B1.

Let $\boldsymbol{\psi}_2$ denote the state-specific parameter vector pertaining to the differential state and let $\boldsymbol{\psi}_{(2,l)}$ denote the set of parameters pertaining to the lth mixture component. Assuming independence of read counts across conditions and samples, conditional upon the differential HMM state, the finite mixture model emission distribution can be written as

$$\begin{aligned} f_2(\mathbf{y}_{..j}|\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\psi}_2) &= \sum_{l=1}^{L} \delta_l f_{(2,l)} \left(\mathbf{y}_{..j}|\mathbf{x}_l; \boldsymbol{\psi}_{(2,l)}\right), \\ &= \sum_{l=1}^{L} \delta_l \left[\prod_{h=1}^{G} \prod_{i=1}^{n_h} \frac{\Gamma\left(y_{hij} + \phi_{(2,l,h)}\right)}{y_{hij}!\Gamma\left(\phi_{(2,l,h)}\right)} \left(\frac{\phi_{(2,l,h)}}{\mu_{(2,l,hij)} + \phi_{(2,l,h)}}\right)^{\phi_{(2,l,h)}} \times \right. \\ &\quad \left. \times \left(\frac{\mu_{(2,l,hij)}}{\mu_{(2,l,hij)} + \phi_{(2,l,h)}}\right)^{y_{hij}}\right], \quad y_{hij} \in \{0, 1, 2, \ldots\}, \end{aligned} \tag{3}$$

where $\mu_{(2,l,hij)}$ and $\phi_{(2,l,h)}$ are the mean and dispersion, respectively, pertaining to read counts

originating from window j and sample i in condition h from the lth mixture component. We assume that $\log(\mu_{(2,l,hij)}) = \beta_1 + \beta_3 x_{hl} + u_{hij}$ and $\log(\phi_{(2,l,h)}) = \lambda_1 + \lambda_3 x_{hl}$. That is, in the mixture component l, we utilize the same consensus background (consensus enriched) log-mean and log-dispersion from (2) in all conditions that are specified by $\mathbf{x}_l$ to be background (enriched) in the lth differential combinatorial pattern.

There are several advantages to such a parameterization of our proposed differential emission distribution. For example, this choice of parametrization ensures that windows exhibiting differential enrichment across conditions share means and dispersions that are common between the consensus background and consensus enrichment states, a reasonable assumption that significantly increases computational efficiency. In addition, utilizing a mixture model as the differential state emission distribution avoids the significant computational burden that would come from assuming separate hidden states for each of the L differential combinatorial patterns, particularly as G increases. We evaluate the strength of these assumptions through multiple simulations and a real data benchmarking analysis in Sections 4 and 5.

Two novel features result from our proposed approach that are relevant to the context of differential enrichment detection from ChIP-seq experiments. By using a modified version of the Expectation-Maximization (EM) algorithm to estimate the model parameters, we are able not only to detect differential enrichment regions across multiple conditions, but we can also classify various differential combinatorial patterns of enrichment within broad and short differential enrichment domains. We detail the proposed algorithm for the parameter estimation and the strategies used to accelerate its performance in Section 3.2. We formally evaluate the performance of our model setup and assumptions via comprehensive statistical simulations in Section 4 and in real data benchmarking in Section 5.

3.2 Estimation

To simplify the parameter estimation in (4), we introduce another set of latent variables $\mathbf{W} = (\mathbf{W}_1', \ldots, \mathbf{W}_M')'$, such that $\mathbf{W}_j = (W_{j1}, \ldots, W_{jL})'$ for $j = 1, \ldots, M$. We assume that $\mathbf{W}$ is a sequence of independent random vectors such that $\mathbf{W}_j | (Z_j = 2) \sim \text{Multinomial}(1, \boldsymbol{\delta})$ and $\mathbf{W}_j | (Z_j = r) = 0$ with probability 1 if $r = \{1, 3\}$. Under this setup, one may define

the data generating mechanism when $Z_j = 2$ (differential state) and $W_{jl} = 1$ (l^{th} differential combinatorial pattern) such that read counts pertaining to genomic window j are sampled from $f_{(2,l)}$ given $\boldsymbol{\psi}_{(2,l)}$ and $\mathbf{x}_l$. Let $\mathcal{W}$ denote the set of L^M possible combinations of latent vectors $\mathbf{W}$. Hence, the likelihood function of the observed data (1) can be rewritten as

$$
\begin{aligned}
f(\mathbf{y}|\mathbf{x};\boldsymbol{\Psi}) &= \sum_{\mathbf{Z}\in\mathcal{Z}} \left\{ \left[\sum_{\mathbf{W}\in\mathcal{W}} f(\mathbf{y}|\mathbf{w},\mathbf{z},\mathbf{x};\boldsymbol{\psi}) f(\mathbf{w}|\mathbf{z},\mathbf{x};\boldsymbol{\delta}) \right] f(\mathbf{z}|\mathbf{x};\boldsymbol{\pi},\boldsymbol{\gamma}) \right\}, \qquad (4)\\
&= \sum_{\mathbf{Z}\in\mathcal{Z}} \sum_{\mathbf{W}\in\mathcal{W}} \left\{ \left[\prod_{r=1}^{3} \pi_r^{I(Z_1=r)} \prod_{j=2}^{M} \prod_{r=1}^{3} \prod_{s=1}^{3} \gamma_{rs}^{I(Z_{j-1}=r,Z_j=s)} \right] \times \left[\prod_{j=1}^{M} \left(\prod_{l=1}^{L} \delta_l^{W_{jl}} \right)^{I(Z_j=2)} \right] \times \right.\\
&\quad \left. \times \left[\prod_{j=1}^{M} f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right)^{I(Z_j=1)} \left(\prod_{l=1}^{L} f_{(2,l)}(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)})^{W_{jl}} \right)^{I(Z_j=2)} f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right)^{I(Z_j=3)} \right] \right\},
\end{aligned}
$$

where $f_{(2,l)}(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)})$ is defined as in (3). In the tth step of the EM algorithm, the Q function of the complete data log-likelihood can be written as

$$
\begin{aligned}
Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}\right) &= E_{\mathbf{Z}}\left(E_{\mathbf{W}|\mathbf{Z}}\left(\log\left(f(\mathbf{y},\mathbf{W},\mathbf{Z}|\mathbf{x};\boldsymbol{\Psi})\right) |\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)} \right) |\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)} \right),\\
&= Q_0(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)}) + Q_1(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}) + Q_2(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}) + Q_3(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}), \qquad (5)
\end{aligned}
$$

where $Q_0(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)})$, $Q_1(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)})$, $Q_2(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)})$, and $Q_3(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)})$ are defined in Appendix A. In the E-step of the EM algorithm, we compute the posterior probabilities from (5). The quantities $Pr\left(Z_j = r|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$ and $Pr\left(Z_{j-1} = r, Z_j = s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$ (from Appendix A) can be calculated through the Forward-Backward algorithm (Appendix B) and $Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{..j},\mathbf{x};\boldsymbol{\Psi}^{(t)}) = f_{(2,l)}(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}^{(t)}_{(2,l)})\delta^{(t)}_l / \sum_{k=1}^{L} f_{(2,k)}(\mathbf{y}_{..j}|\mathbf{x}_k;\boldsymbol{\psi}^{(t)}_{(2,k)})\delta^{(t)}_k$ for $l = 1,\ldots,L$.

The Q function is maximized with respect to the parameters $\boldsymbol{\Psi} = (\boldsymbol{\pi}',\boldsymbol{\gamma}',\boldsymbol{\delta}',\beta_1,\beta_3,\lambda_1,\lambda_3)'$ during the M-step of the algorithm. Estimates for the initial and transition probabilities can be directly calculated as $\hat{\pi}_r^{(t+1)} = Pr\left(Z_1 = r|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$ and $\hat{\gamma}_{rs}^{(t+1)} = \sum_{j=2}^{M} Pr(Z_{j-1} = r, Z_j = s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}) / \sum_{j=2}^{M} Pr(Z_{j-1} = r|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)})$, respectively, restricted to $\sum_{r=1}^{3} \hat{\pi}_r^{(t+1)} = 1$ and $\sum_{s=1}^{3} \hat{\gamma}_{rs}^{(t+1)} = 1$, for $r \in \{1,2,3\}$. We perform conditional maximizations to compute estimates of the remaining model parameters $(\boldsymbol{\delta}',\beta_1,\beta_3,\lambda_1,\lambda_3)'$. First, mixture proportions can be estimated as $\hat{\delta}_l^{(t+1)} = \sum_{j=1}^{M} Pr(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}) Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{..j},\mathbf{x};\boldsymbol{\Psi}^{(t)}) / \sum_{j=1}^{M} Pr(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)})$. Estimating $(\beta_1,\beta_3,\lambda_1,\lambda_3)'$ from (5) can be seen as obtaining parameter estimates from a series of weighted NB regression models with shared mean and dispersion

parameters. Here, we jointly estimate these quantities via the gradient-based algorithm BFGS (Fletcher 2013).

The proposed model and its estimation scheme are robust to situations where certain differential combinatorial patterns of enrichment are rare (see Figure 6 in Section 5.3 for an example). This unique characteristic results from the fact that ChIP-seq experiments often provide enough data (usually $M > 10^7$ non-overlapping windows of 250 bp fixed size for the human reference genome) to estimate the model parameters $(\beta_1, \beta_3, \lambda_1, \lambda_3)'$, which are shared across all L mixture components and HMM states. If pruning differential combinatorial patterns of the differential mixture component is of interest, the optimal number of mixture components L^*, $L^* < L$, can be selected via the Bayesian Information Criterion (BIC) for HMMs (Zucchini et al. 2017). We observed that selecting the optimal number of mixture components based on BIC agrees with the pruning of rare differential combinatorial patterns that we would not biologically expect to observe in real data (see the Online Supplementary Material B3 for a discussion and examples).

To obtain the parameter estimates $\hat{\boldsymbol{\Psi}}$, the EM algorithm iterates until the maximum absolute relative change in the parameter estimates three iterations apart is less than 10^{-3} for three consecutive iterations. To reduce the computation time, we make use of a rejection controlled EM algorithm (RCEM, Ma et al. (2006)) with threshold 0.05. Briefly, the RCEM algorithm substantially reduces the dimensionality of the data during the M-step by randomly assigning a zero posterior probability to genomic windows unlikely to belong to each of the HMM states. In addition, the current estimation set up allows genomic windows exhibiting equal distribution of read counts to have their posterior probability aggregated during the M-step of the algorithm. Often, the distribution of read counts along the genome is highly concentrated on a particular set of values, such as 0, 1, and 2 for instance. Genomic windows exhibiting a particular pattern of counts across samples and conditions can have their posterior probability aggregated during the M-step, which further reduces the dimensionality of the objective function during the numerical optimization and leads to a fast gradient-based optimization. Genomic windows with assigned zero posterior probability due to the RCEM approach do not contribute to the final set of aggregated posterior probabilities.

Once the algorithm reaches convergence, the final set of HMM posterior probabilities can be used to segment the genome into consensus background, differential, or consensus enrichment windows. Approaches that control the total false discovery rate (FDR) via posterior probabilities (Efron et al. 2001) or that estimate the most likely sequence of hidden states (Viterbi 1967) can be used for such purposes. Let $\hat{\rho}_{j2} = Pr\left(Z_j = 2|\mathbf{y}, \mathbf{x}; \hat{\mathbf{\Psi}}\right)$ denote the estimated posterior probability that the jth genomic window belongs to the differential HMM state, $j = 1, \ldots, M$. For a cutoff of posterior probability α, the total FDR is $\sum_{j=1}^{M}(1-\hat{\rho}_{j2})I(\hat{\rho}_{j2} \geq 1-\alpha)/\sum_{j=1}^{M} I(\hat{\rho}_{j2} \geq 1-\alpha)$, where $I(\cdot)$ is an indicator function. The posterior probability cutoff is then chosen by controlling the total FDR. Differential regions of enrichment are formed by merging adjacent windows that either meet a given FDR threshold level for the differential HMM state or belong to the same Viterbi's predicted state. A discussion about these two approaches under the proposed model is presented in the Online Supplementary Material B4. Additional details of proposed EM algorithm and the implemented code are available in the Online Supplementary Materials B5 and A3, respectively.

4 Simulation Studies

We evaluate the presented model in two independent simulation studies of broad epigenomic marks. In the first study (Section 4.1), we simulated read count-based data to assess the precision of the parameter estimation scheme, the performance of differential peak detection, and the accuracy of the classification of specific differential combinatorial patterns of enrichment within differential peaks. In the second simulation study (Section 4.2), we utilize the simulation pipeline presented in Lun & Smyth (2015) to generate synthetic ChIP-seq reads from *in silico* experiments with broad differential peaks. The aim of the second simulation study was to directly compare our model with other DPCs in a more realistic scenario with broad peaks, while also avoiding the assumption of a specific parametric model for the data. Code to reproduce this simulation study are detailed in the Online Supplementary Material A4.

4.1 Read Count Simulation

Read counts were simulated under different scenarios that varied regarding the type of histone modification mark mimicked (H3K36me3 and H3K27me3), genome length (M, 10^5, 5×10^5, and 10^6 windows), number of conditions (G, 2, 3, and 4), and number of replicates per condition (n, 1, 2, and 4). For each scenario, we further assessed our model under different Signal-to-Noise Ratio (SNR) levels. Here, we define the SNR as the ratio between the means of consensus enrichment and consensus background emission distributions. Mean and dispersion parameters used in this simulation study were estimated from ENCODE data and are presented in the Online Supplementary Material C1 for all the scenarios. Different SNR levels were defined by decreasing the ratio of the means in decrements of 10% while maintaining the mean-variance relationship. Read counts were assumed to follow a NB distribution and were simulated using a first-order Markov chain with 2^G states, representing every combination of background and enrichment across G conditions. We aimed to assess whether our model was able to assign all $2^G - 2$ simulated differential states to the differential HMM state, while maintaining a precise parameter estimation scheme and accurate classification of differential combinatorial patterns.

4.1.1 Simulation Results

Table 1 shows the true values and the average relative bias of parameter estimates (and the 2.5^{th}, 97.5^{th} percentiles) from a hundred simulated data sets relative to the scenario of H3K27me3 with 10^5 genomic windows (see the Online Supplementary Material C1 for additional results and parameter specifications). Results are shown for different levels of SNR, number of conditions, and number of replicates per condition. Overall, no significant differences regarding the relative bias of parameter estimates were observed across simulations under different genome lengths. Depending on the number of conditions, the observed relative bias and the range of the reported percentiles tended to decrease as more replicates were included in the analyses. This effect was particularly significant in scenarios with four conditions with respect to parameters β_3 and λ_3. In general, scenarios with higher SNR showed lower relative bias and variability of the parameter estimates in comparison to sce-

narios with lower SNR, regardless of the number of conditions or replicates per condition. In scenarios with lower SNR levels or higher number of conditions, these results also highlight the importance of experimental replicates to achieve precise parameter estimates. These results show that the proposed estimation approach via EM algorithm led to precise parameter estimates and was robust to a data generating mechanism that was different than the one assumed by the proposed model.

Next, we assessed the sensitivity of our method to detect simulated differential regions of enrichment. First, differential regions were defined from the HMM posterior probabilities pertaining to the differential state by controlling the total FDR as defined in Section 3.2. For different nominal FDR threshold levels, the model sensitivity was estimated as the proportion of windows correctly assigned as differential out of the total number of simulated differential windows. Additionally, the observed FDR was calculated as the proportion of genomic windows incorrectly called as differential out of the total number of called differential windows. Figure 2A shows the average observed true positive rate (y-axis) and the observed FDR (x-axis) for different nominal FDR levels across a hundred simulated data relative to the scenario of H3K27me3 with $M = 10^5$ genomic windows. Results are shown for different levels of SNR, number of conditions, and number of replicates per condition. Overall, we observed that the number of replicates per condition played a major role on the sensitivity levels of the model, in which scenarios with two and four replicates had the best results regardless of the number of conditions and SNR levels. For scenarios with either high number of conditions or low SNR levels, more replicates were needed to achieve higher sensitivity.

Finally, we used the estimated mixture model posterior probabilities $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \hat{\mathbf{\Psi}})$, $j = 1, \ldots, M$, to classify the differential combinatorial patterns of enrichment of detected differential windows. To this end, we first calculated the maximum estimated mixture model posterior probability across all L components to determine the most likely differential combinatorial pattern from genomic windows assigned to be part of the differential state. Then, we compared the window-based classification with the true window-based simulated states from the Markov Chain (states $2, \ldots, G-1$). Figure 2B shows the confu-

Table 1: Read count simulation. True values and average relative bias of parameter estimates (and 2.5^{th}, 97.5^{th} percentiles) across a hundred simulated data sets are shown for H3K27me3 with 10^5 genomic windows. Scenarios with observed SNR and 70% of observed SNR are shown.

				One Replicate		Two Replicates		Four Replicates	
SNR	Conditions	Parameter	True	R. Bias	($P_{2.5}$,$P_{97.5}$)	R. Bias	($P_{2.5}$,$P_{97.5}$)	R. Bias	($P_{2.5}$,$P_{97.5}$)
70% of	Two	β_1	1.116	0.000	(-0.004, 0.004)	0.000	(-0.002, 0.002)	0.000	(-0.002, 0.002)
Observed		β_3	0.808	-0.001	(-0.012, 0.008)	0.000	(-0.005, 0.006)	0.000	(-0.003, 0.004)
SNR		λ_1	1.281	0.000	(-0.016, 0.013)	0.000	(-0.010, 0.010)	0.000	(-0.008, 0.007)
		λ_3	-0.232	-0.001	(-0.109, 0.110)	0.000	(-0.078, 0.067)	0.000	(-0.058, 0.048)
	Three	β_1	1.116	0.002	(-0.005, 0.013)	-0.003	(-0.007, 0.001)	-0.001	(-0.002, 0.001)
		β_3	0.808	-0.156	(-0.233, -0.098)	-0.014	(-0.025, -0.005)	-0.001	(-0.004, 0.003)
		λ_1	1.281	0.005	(-0.033, 0.032)	0.004	(-0.006, 0.014)	0.001	(-0.006, 0.007)
		λ_3	-0.232	0.898	(0.679, 1.083)	0.130	(0.012, 0.242)	0.015	(-0.033, 0.061)
	Four	β_1	1.116	0.001	(-0.005, 0.020)	-0.012	(-0.017, -0.008)	-0.010	(-0.014, -0.007)
		β_3	0.808	-0.145	(-0.256, -0.120)	-0.087	(-0.098, -0.075)	-0.032	(-0.062, -0.015)
		λ_1	1.281	0.016	(-0.028, 0.033)	0.028	(0.016, 0.041)	0.017	(0.007, 0.028)
		λ_3	-0.232	0.947	(0.819, 1.083)	0.769	(0.672, 0.877)	0.366	(0.211, 0.613)
Observed	Two	β_1	1.116	0.000	(-0.004, 0.004)	0.000	(-0.003, 0.002)	0.000	(-0.002, 0.002)
SNR		β_3	1.165	0.000	(-0.005, 0.005)	0.000	(-0.003, 0.003)	0.000	(-0.002, 0.003)
		λ_1	1.281	0.000	(-0.015, 0.018)	-0.001	(-0.010, 0.008)	0.001	(-0.004, 0.007)
		λ_3	0.124	-0.012	(-0.231, 0.174)	0.008	(-0.108, 0.129)	-0.008	(-0.107, 0.085)
	Three	β_1	1.116	-0.004	(-0.008, 0.001)	-0.001	(-0.003, 0.002)	0.000	(-0.002, 0.002)
		β_3	1.165	-0.010	(-0.019, -0.002)	0.000	(-0.003, 0.003)	0.000	(-0.002, 0.002)
		λ_1	1.281	-0.001	(-0.012, 0.012)	0.000	(-0.008, 0.009)	0.000	(-0.006, 0.007)
		λ_3	0.124	-0.422	(-0.697, -0.068)	-0.026	(-0.171, 0.083)	-0.003	(-0.085, 0.081)
	Four	β_1	1.116	-0.013	(-0.019, -0.008)	-0.008	(-0.014, -0.003)	0.000	(-0.002, 0.001)
		β_3	1.165	-0.111	(-0.121, -0.101)	-0.005	(-0.010, -0.001)	0.000	(-0.002, 0.002)
		λ_1	1.281	-0.011	(-0.025, 0.001)	0.009	(-0.002, 0.018)	0.001	(-0.004, 0.006)
		λ_3	0.124	-3.526	(-3.782, -3.239)	-0.438	(-0.740, -0.217)	-0.009	(-0.084, 0.059)

sion matrices of classified (x-axis) and simulated (y-axis) differential windows for a scenario with three conditions and data simulated from H3K27me3 with 10^5 genomic windows. Differential combinatorial patterns of enrichment are represented by the sequences of letters 'E' (enrichment) and 'B' (background), such that each letter corresponds to the status of a given condition. The number of windows (averaged over a all simulated data sets) is shown as entries of the matrices and represented by the color scale. Darker colors on the diagonal entries indicate better agreement between simulated and classified patterns. By utilizing the posterior probabilities from the mixture model, we observed a good performance when classifying the differential combinatorial pattern of enrichment from differential windows. Results were best under scenarios with higher number of replicates or SNR.

Overall, simulated scenarios with higher number of replicates or a higher SNR led to less biased and more precise parameter estimates, higher accuracy of differential peak detection, and best classification of the differential direction of enrichment. To the best of our knowledge, the classification capability of the proposed model in settings with more than two conditions is a novelty not yet available in any other method for the detection of differential protein-DNA binding sites. Although Lun & Smyth (2015) presented a DPC tailored for multiple conditions, its current implementation does not allow the classification of differential combinatorial patterns of enrichment under three or more conditions.

4.2 Sequencing Read Simulation

We performed a second simulation study aiming to compare the proposed model with the current DPCs such as ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR, as none of these methods were designed to accept matrices of window-based read counts as input. To this end, we used the simulation pipeline presented by Lun & Smyth (2015) where data were generated in a more general scheme without a particular read count model assumption. Here, sequencing reads from broad ChIP-seq experiments were generated for two conditions and two replicates per condition. For the differential peaks callers ChIPComp and DiffBind that require sets of candidate regions, we followed the analyses presented by Lun & Smyth (2015) and called peaks in advance using HOMER. Those were then used as input in the respective software for differential call. A hundred simulated data sets were generated and

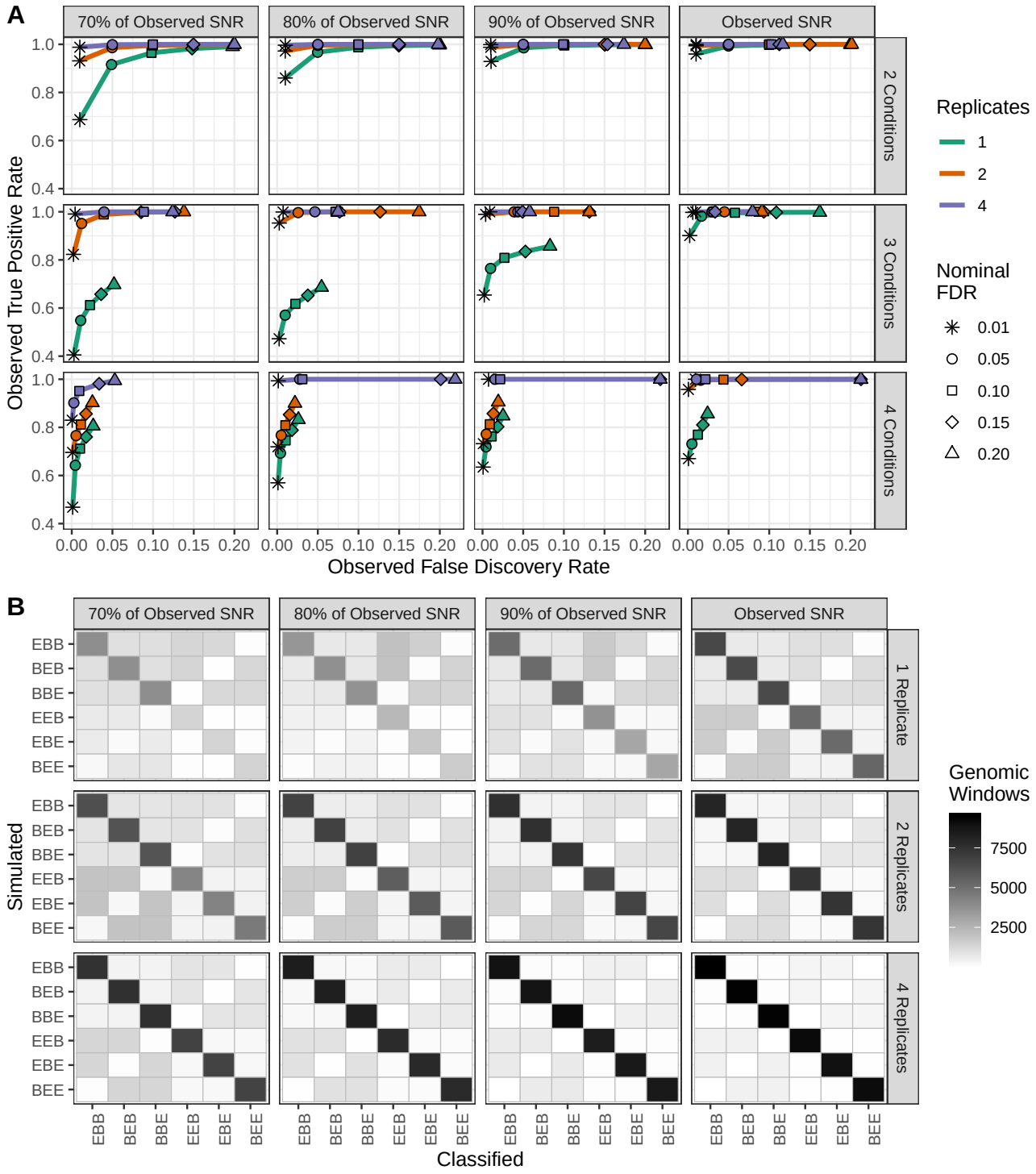

Figure 2: Read count simulation. (A): average observed TPR and FDR for different nominal FDR levels for simulated scenarios of H3K27me3 with 10^5 windows. (B): confusion matrices for the 3 conditions scenario. On x- and y-axes, the labels indicate the classified and simulated patterns, respectively (e.g., BEB denotes enrichment in conditions 2 only). Darker colors on the diagonal indicate better agreement.

peaks were called by all the methods under multiple nominal FDR thresholds. For our method and RSEG, window-based posterior probabilities were used to control the total FDR as described in Section 3.2. For illustration purposes, we refer to the method presented in this article as mixNBHMM.

4.2.1 Simulation Results

Figure 3 shows the main results of our second simulation study. Out of 100 simulated data sets, RSEG either failed to analyze the data due to internal errors or called the entire genome as differential in 26 and in 3 instances, respectively. Similar issues have been previously reported in other studies (Starmer & Magnuson 2016). We observed that our method showed the highest observed sensitivity among all DPCs, regardless of the nominal FDR thesholding level, while maintaining a moderate observed FDR (panel A). Methods such as diffReps, RSEG, and THOR showed higher observed FDR levels than the nominal threshold due to the excessive number of differential peaks called outside true differential regions (shaded area in panel D). While diffReps and THOR called an excessive number of short and discontiguous peaks, RSEG called regions that were usually wider than the observed differential enrichment regions. These results are further illustrated in Figure 3B, where we present the average ratio of the number of called and simulated peaks (y-axis) and the average number of called peaks intersecting true differential regions (x-axis). Regarding the computation time, the HMM-based algorithms RSEG and THOR appeared to be the most computationally intensive and required longer amounts of time to analyze the data. In Figure 3C, we present the box plots of computing time (in minutes) across a hundred simulated data sets for all benchmarked methods. While still being an HMM-based algorithm, the method presented in this article was among the fastest tools for differential peak detection due to the implemented strategies to improve the computation time of the EM algorithm (Section 3.2). Figure 3D shows an example of a genomic region with simulated data and called peaks from various methods using nominal FDR threshold 0.05. As shown, our method was able to consistently cover most of true differential regions with broad peaks while exhibiting a limited number of false discoveries (see Figure C.8 in the Online Supplementary Material C2 for additional examples and results).

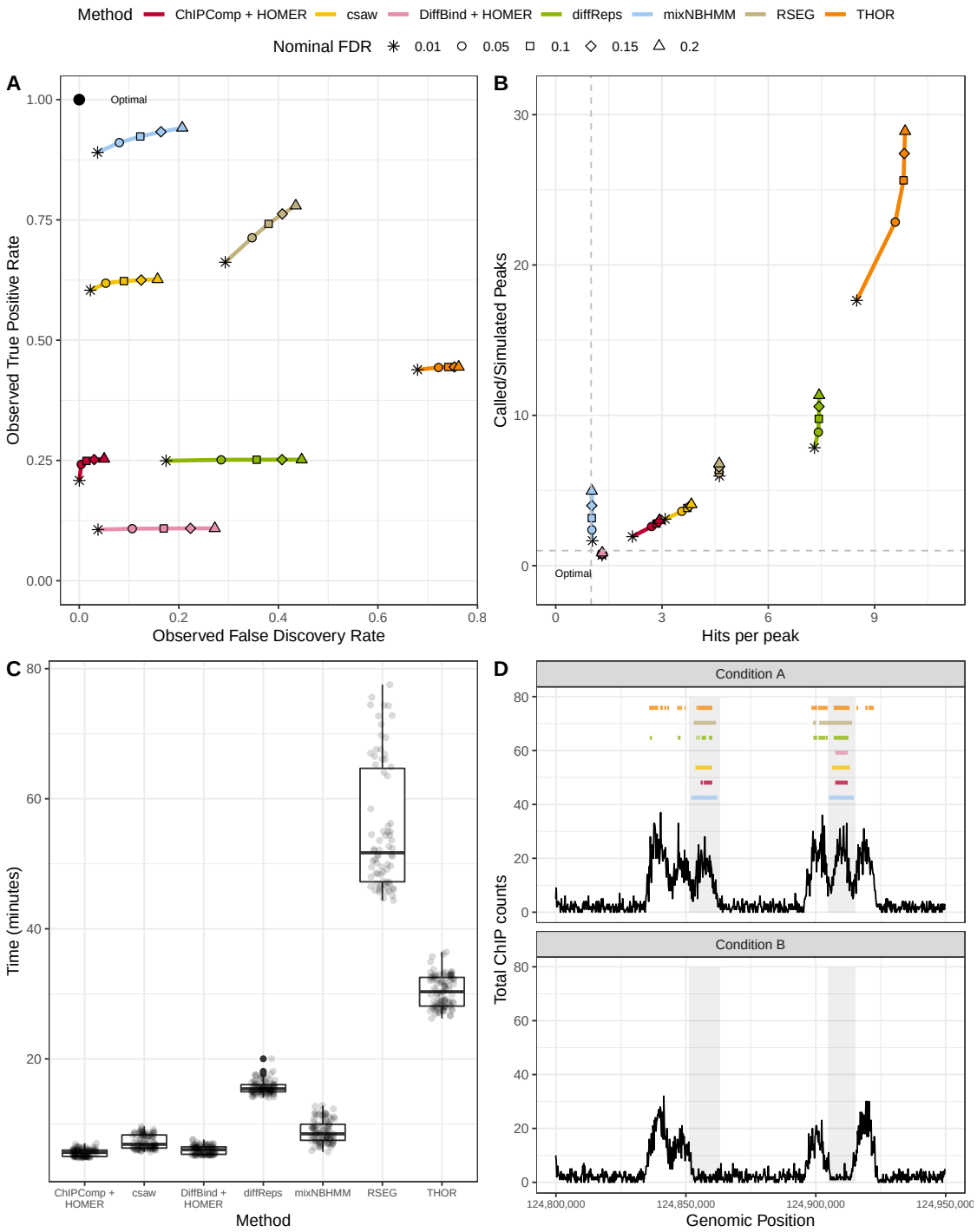

Figure 3: Sequencing read-based simulation from the csaw pipeline. (A): average observed sensitivity and FDR for various methods. (B): scatter plot of average ratio of called and simulated peaks (y-axis) and number of called peaks intersecting true differential regions (x-axis). (C): box plot of computing time (in minutes) for various algorithms. (D): an example of differential peak calls under a nominal FDR control of 0.05. Shaded areas indicate true differential peaks.

5 Application to ENCODE Data

We applied our method to ChIP-seq data from the ENCODE Consortium (Section 2) to detect differential regions of enrichment of several epigenomic marks across distinct cell lines. First, we analyzed broad data from the histone modifications H3K36me3 and H3K27me3 as well as data from the enhancer EZH2 (Section 5.1). Secondly, we assessed the performance of the presented model on ChIP-seq experiments from the transcription factor CTCF and the histone modifications H3K27ac and H3K4me3 characterized by short peaks (Section 5.2). For H3K27ac and H3K4me3, enrichment peaks are usually deposited on the promoter regions of actively transcribed genes and several studies have associated their role with gene transcription (Creyghton et al. 2010, Lauberth et al. 2013). The transcription factor CTCF is a protein that binds to short DNA motifs and is responsible for several cellular processes that include the regulation of the chromatin 3D structure and mRNA splicing (Shukla et al. 2011). Two technical replicates for each epigenomic mark were used in the analysis. Using RNA-seq data, we assessed the practical significance of our results by associating the detection and classification of differential combinatorial patterns from called peaks of H3K36me3, H3K27me3, and EZH2 with the direction of gene expression (Section 5.3; see the Online Supplementary Material D1 for the analysis steps using RNA-seq data).

We sought to compare the genome-wide performance of the presented model with the current DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR. For the methods that require a set of candidate regions to be specified *a priori*, ChIPComp and DiffBind, peaks were called in advance using MACS (Zhang et al. 2008) and used as input to the software for differential call. We benchmarked methods regarding the coverage of differentially transcribed gene bodies, the number and average size of differential peak calls, $\log_2$ fold change (LFC) of read counts, Spearman correlation of $\log_2$-transformed read counts between cell lines, and computation time. Metrics for sensitivity and specificity were defined on the window level and based on the coverage of differentially transcribed gene bodies by called peaks. For broad marks, read counts were computed using non-overlapping windows of 500bp. For the remaining short marks, we computed read counts using non-overlapping windows of 250bp. Results presented in this section pertain to the analysis of two cell lines,

namely Helas3 and Hepg2, as only csaw was designed to call differential peaks from more than two conditions. A discussion about the choice of the window size is presented in the Online Supplementary Material D2. Results from the analysis of more than two cell lines are presented in the Online Supplementary Material D3. Data accessing code, data preprocessing steps, method-specific parameters, and code to replicate the presented results are detailed in the Online Supplementary Material A.

5.1 Analysis of ChIP-seq Data From Broad Marks

First, methods were benchmarked regarding the coverage of differentially transcribed gene bodies by their differential peak calls. The histone modification H3K36me3 is known to be associated with gene transcription and enriched regions of this mark are usually deposited on actively transcribed gene bodies. Hence, the location of differential peaks of H3K36me3 is expected to agree with the location of differentially expressed genes. Following the ideas presented by Steinhauser et al. (2016) and Ji et al. (2013), we define a set of protein coding genes exhibiting $|\text{LFC}| > 2$ of ChIP-seq read counts between the two analyzed cell lines as true differentially transcribed genes. Results using different threshold levels are presented in the Online Supplementary Material D1 and agree with those presented here. Protein-coding genes with total read count across cell lines under the 25th percentile were excluded from the analysis. Normalization by the median log-ratios of each replicate over the geometric mean was performed to avoid spurious differences due to sequencing depth.

In Figure 4A, we show receiver operating characteristic (ROC) curves for various methods and different nominal levels of FDR threshold for differential peaks of H3K36me3. Our method had the best overall performance among all DPCs as its differential peaks were able to cover most of differentially transcribed gene bodies while still maintaining a low number of false positives. Methods that tended to call short peaks, such as ChIPComp and DiffBind, were the ones with the lowest sensitivity among all methods. ChIPComp and DiffBind have been previously shown to be highly dependent on the set of candidate peaks and to perform best in scenarios with short peaks (Steinhauser et al. 2016, Lun & Smyth 2015). Examples of differential peak calls for H3K36me3 under a nominal FDR control of 0.05 are shown in Figure 4B. In Figure 4C, we show the average size (x-axis) and the number (y-

axis) of called differential H3K27me3 peaks for various methods and different nominal FDR levels. The presented model and RSEG, two of the HMM-based methods, tended to call a limited number of broad differential H3K27me3 peaks among all methods (bottom right corner of panel C). However, we observed that differential peaks called by RSEG often did not correspond to the observed differential region of enrichment (Figure 4D). This behavior of RSEG was previously observed by Starmer & Magnuson (2016) and agrees with the results presented in our simulated study (Figure 3). In Figure 4F, we show examples of differential peak calls for the enhancer EZH2. Regarding the computing time (Figure 4E), our method was among the fastest algorithms due to our computational scheme, taking approximately an hour to jointly analyze genome-wide data.

5.2 Analysis of ChIP-seq Data From Short Marks

We further evaluated the performance of the proposed method on data sets characterized by short peaks, namely the histone modifications H3K4me3 and H3K27ac and the transcription factor CTCF. The goal of our analysis was to assess whether our method was robust to different types of data and still able to call short differential regions of enrichment. In these scenarios, differential peaks are usually observed in isolated genomic regions and exhibit a high SNR. It has been shown that certain HMM-based approaches, including RSEG, have low accuracy under short histone modification marks and TF data (Hocking et al. 2016). As we show, the model proposed in this article performs comparably to the evaluated DPCs known to perform best in short data (ChIPComp, DiffBind; Steinhauser et al. (2016)) and appeared to be more efficient regarding the computation time in certain scenarios.

We calculated the LFC and the Spearman correlation between cell lines Helas3 and Hepg2 based on ChIP-seq read counts mapped onto differential peaks called by each method. Read counts were previously normalized by the median log-ratios of each replicate over the geometric mean to avoid spurious differences due to sequencing depth. As these marks are characterized by short peaks, ideal methods would show high absolute LFC and negative correlation between read counts mapped on differential peaks. Figure 5 shows the main results from our analysis using short data sets. In Figure 5 panels A and C we show the median LFC and the Spearman correlation of ChIP-seq counts for differential CTCF and

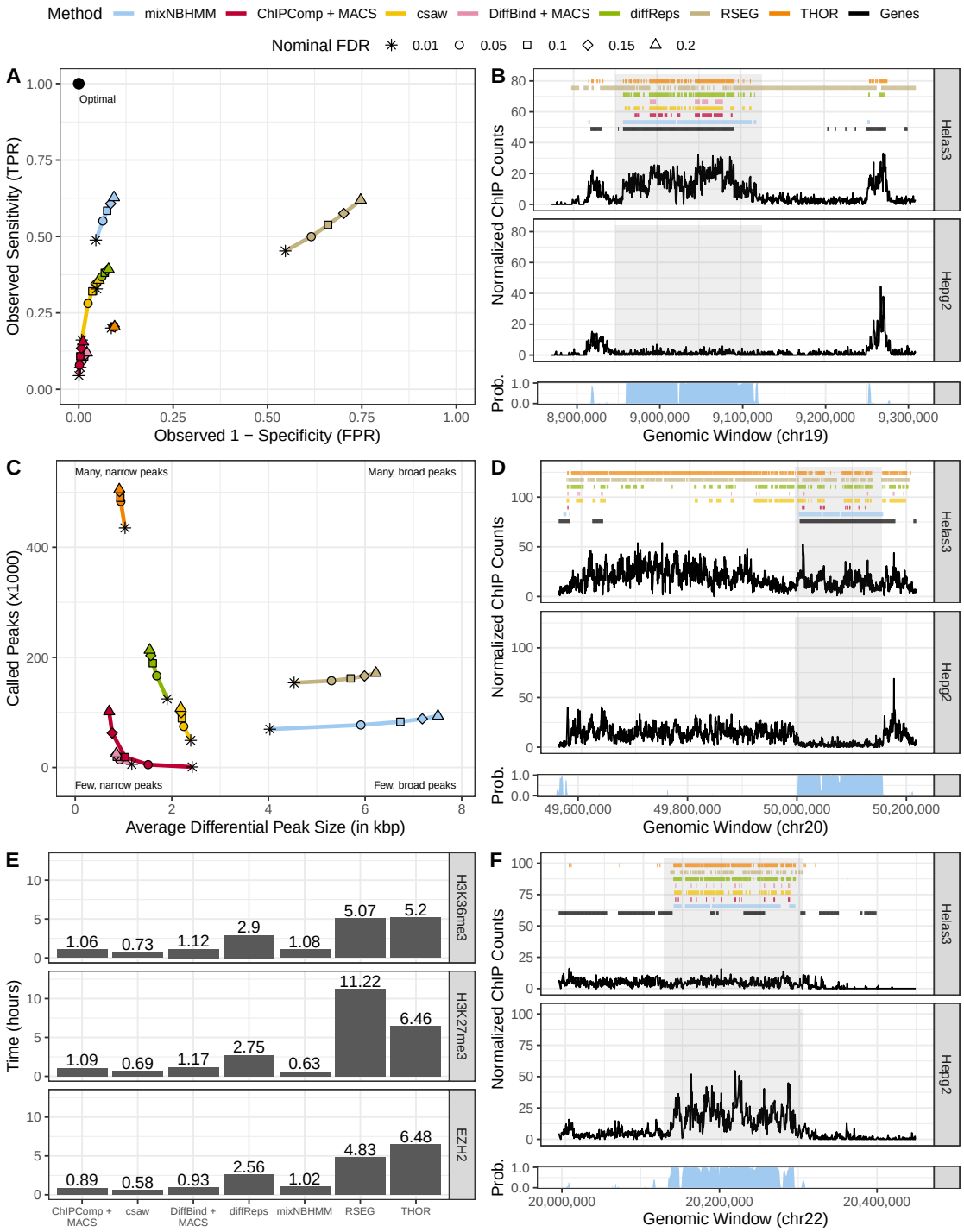

Figure 4: Analysis of broad ENCODE data. (A): ROC curves of H3K36me3 differential peak calls. (C): average number (y-axis) and size (x-axis) of H3K27me3 called peaks for various methods and different nominal FDR thresholds. (B), (D), and (F): example of peak calls from H3K36me3, H3K27me3, and EZH2, respectively, under a nominal FDR control of 0.05. Posterior probabilities of the HMM differential state are shown at the bottom of each panel. (E): computing time of genome-wide analysis from various methods.

H3K4me3 peak calls (sorted by the absolute LFC), respectively, under a nominal FDR control of 0.05. We present separate curves regarding the signal of observed enrichment to better characterize the direction of change. The results show that the HMM-based methods RSEG and THOR were among those with the lowest absolute LFC among all methods, which confirms their suboptimal performance in the scenario of short peaks (Hocking et al. 2016). In addition, we observed that ChIPComp had the best performance overall as it was able to call differential peaks with the highest absolute LFC and the lowest correlation between read counts of the two analyzed cell lines. The presented model was able to properly call truly short differential peaks (Figure 5 panels B, D, and F) and was comparable to the non-HMM based methods regarding the computation time (Figure 5E). In all data sets, our method was able to jointly call genome-wide differential peaks in less than 1.5 hour.

5.3 Genomic Segmentation and Classification of Chromatin States

Lastly, we analyzed data from the cell line Helas3 to segment its genome regarding the activity of marks H3K36me3, H3K27me3, and EZH2. We considered each mark as a separate experimental condition ($G = 3$), where now we seek to classify joint local chromatin states in Helas3 based upon the presence or absence of enrichment from each mark. It is known that EZH2 catalyzes the methylation of H3K27me3, a repressive mark, and H3K36me3 is associated with transcribed genes (Section 2). Hence, we expected regions of enrichment in consensus for these marks to be rare and differential regions to be mostly represented by either transcribed chromatin states (enrichment for H3K36me3 alone) or repressed chromatin states (enrichment co-occurrence for H3K27me3 and EZH2). The analyses presented in this section highlight the applicability of our method in the context of genomic segmentation (Pinello et al. 2014), a distinct problem not tackled by current DPCs.

Figure 6 shows the main results of our analyses. First, we segmented the genome using the Viterbi sequence of most likely HMM states to understand the distribution of genomic regions associated with consensus background, differential, and consensus enrichment states. The Viterbi algorithm allows the simultaneous segmentation of the genome regarding the three HMM states without the necessity of choosing state-specific FDR control thresholds. While the majority of the genomic regions exhibited no enrichment for any of the analyzed

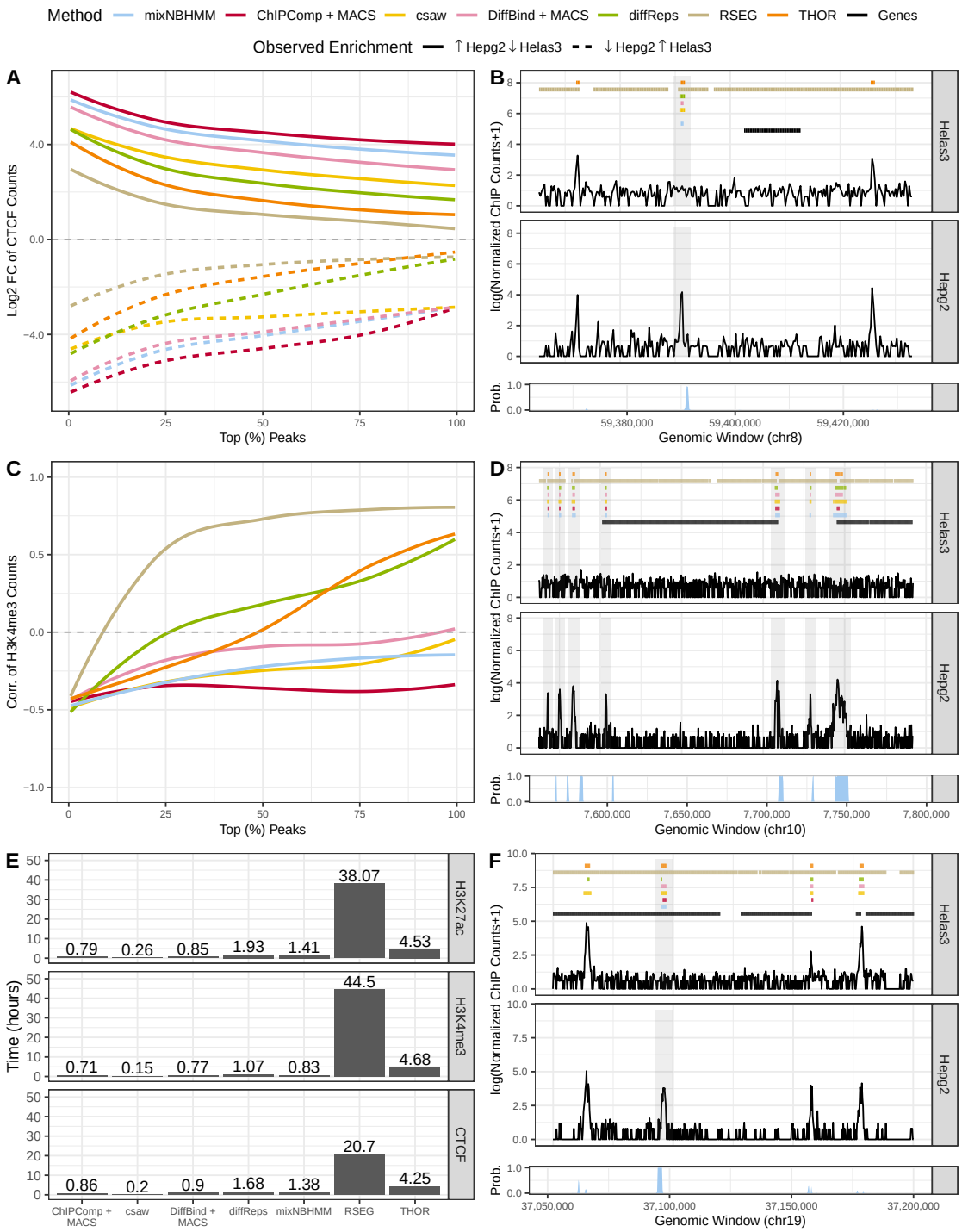

Figure 5: Analysis of short ENCODE data. (A) and (C): median LFC and correlation between cell lines of ChIP-seq counts from differential peaks for CTCF and H3K4me3, respectively. (B), (D), and (F): example of peak calls from CTCF, H3K4me3, and H3K27ac, respectively. Posterior probabilities of the HMM differential state are shown at the bottom of each panel. (E): computing time of genome-wide analysis from various methods. Results are shown under a nominal FDR control of 0.05.

marks (approximately 1500 Mbp of the genome), regions of consensus enrichment were rare and represented only 2% of the analyzed genome (panel A), as expected. Consensus background and differential regions mostly covered intergenic (66%) and protein-coding genic regions (61%), respectively. Given the roles of the analyzed marks (Section 2), these results suggest that protein-coding genes overlapping differential regions should be either silenced (e.g. genes associated with repressed chromatin states) or highly expressed (e.g. genes associated with transcribed chromatin states). Next, we assessed the relative frequency of the differential combinatorial patterns across the three epigenomic data sets. To assign the combinatorial pattern of a differential window, we chose the combination pertaining to the mixture component with the maximum mixture model estimated posterior probability, $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \mathbf{\Psi}^{(t)})$, $j = 1, \ldots, M$. As expected, differential genomic windows were mostly representing either transcribed chromatin states or repressed chromatin states (panel B). All four differential combinatorial patterns expected to be rare had associated mixture proportion estimates less than 0.02 (see the Online Supplementary Material B3 for a discussion on pruning rare states).

Lastly, for genes overlapping differential regions associated with either transcribed chromatin states or repressed chromatin states, we computed the distribution of transcripts per million (TPM) values from matching RNA-seq experiments (panel C). We observed that genes associated with transcribed chromatin states had a significantly higher distribution of TPM counts than genes associated with repressed chromatin states. These results agree with the biological roles of the analyzed epigenomic marks. In panel D, we show an example of a genomic region, its ChIP-seq associated read counts, and the track of differential peaks called by our model. Different colors represent different classified combinatorial patterns of enrichment (peaks) and levels of TPM counts (genes). Our method was not only able to detect broad differential regions of enrichment, but also correctly classified the differential combinatorial pattern under the three types of broad epigenomic marks. In addition, the classification agreed with the gene expression levels of gene bodies covered by either transcribed or repressed chromatin states. Such a novelty is not yet available in any other DPC method presented in the literature.

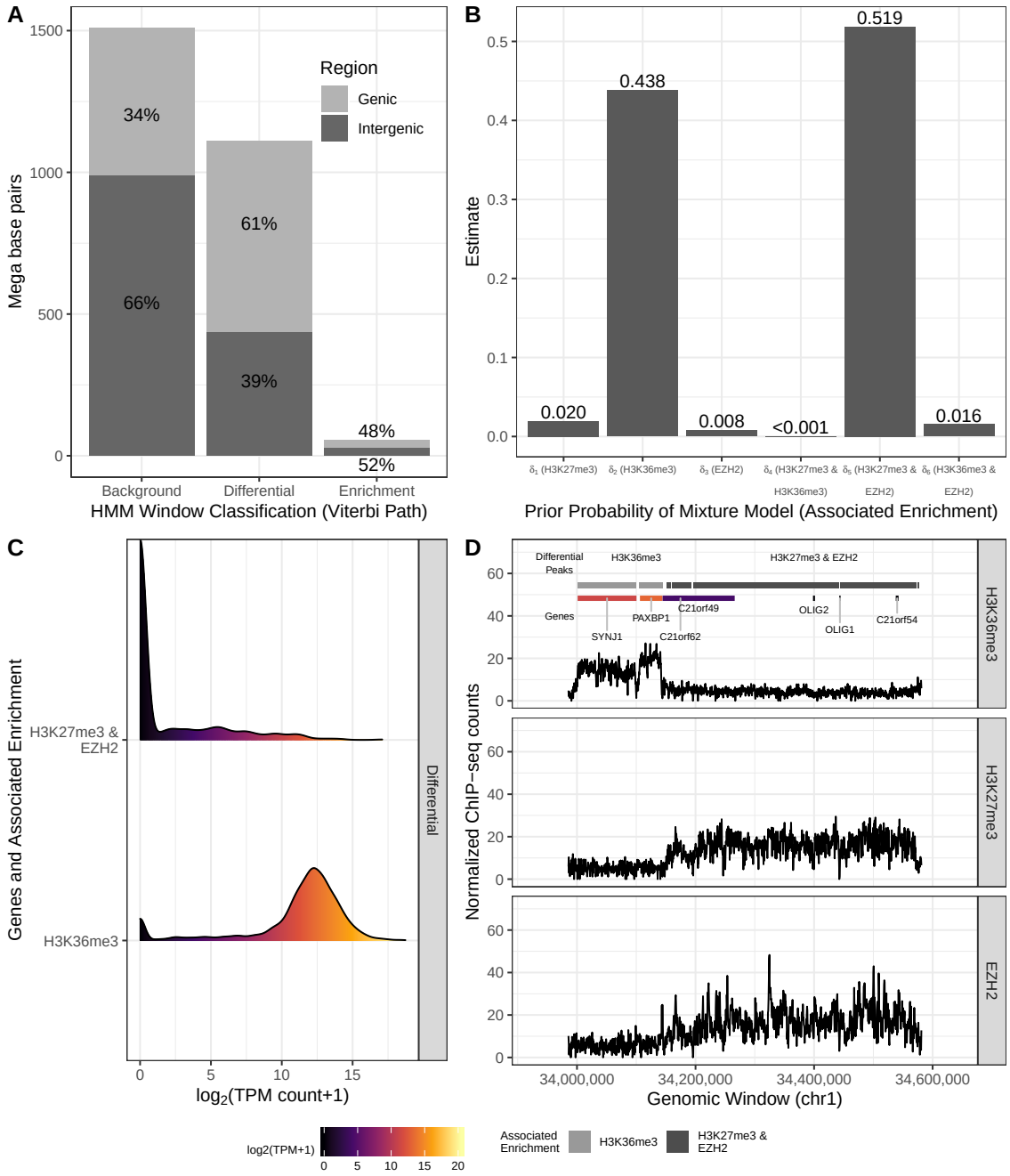

Figure 6: Genomic chromatin state segmentation and classification. (A): distribution of base pairs (y-axis) and the Viterbi sequence of states (x-axis). (B): estimated mixture probabilities and the associated differential combinatorial patterns. (C): density estimate from expression of genes intersecting differential peaks associated with the enrichment of H3K36me3 alone or the enrichment of H3K27me3 and EZH2 in consensus. (D): example of a genomic region with differential peaks and genes, colored according to their expression level.

6 Discussion

Here, we presented a flexible and efficient statistical model designed to call differential regions of enrichment from ChIP-seq experiments with multiple replicates and multiple conditions. Our model has three main advantages over current methods tailored for differential peak detection. First, it uses an HMM-based approach that accounts for the longitudinal aspects of ChIP-seq data with broad signal and does not require any pre-specified set of candidate peaks. Its computation time was comparable to some of the fastest current algorithms and was at least 5 times faster than current HMM-based algorithms (Figure 4). Second, it generalizes the problem of differential peak detection for multiple conditions. Third, our presented HMM allows the classification of differential combinatorial patterns of enrichment from its embedded mixture model and the associated posterior probabilities under any number of conditions. This novelty is not yet available in any other algorithm presented in the literature. As shown in Section 5.3, our predictions agreed with the biological roles of the analyzed marks and with the expression from the associated genes. Such a tool could trigger, for instance, new insights on the field of epigenomics as investigators are often interested in making sense of the roles of epigenomic marks on the human genome. Other epigenomic assays that assess genome-wide chromatin accessibility, such as ATAC-seq, DNase-seq, and FAIRE-seq (Tsompana & Buck 2014), could also benefit from the presented method. Data resulting from these assays are similar in nature but aim to identify regions that alternatively are not bound to proteins. Our software has been implemented into an R package and is available for download (see Section A3 of the Online Supplementary Material for details).

References

Allhoff, M., Seré, K., F. Pires, J., Zenke, M. & G. Costa, I. (2016), 'Differential peak calling of chip-seq signals with replicates with thor', *Nucleic acids research* **44**(20), e153–e153.

Bannister, A. J. & Kouzarides, T. (2011), 'Regulation of chromatin by histone modifications', *Cell research* **21**(3), 381.

Chen, L., Wang, C., Qin, Z. S. & Wu, H. (2015), 'A novel statistical method for quantitative comparison of multiple chip-seq datasets', *Bioinformatics* **31**(12), 1889–1896.

Clouaire, T., Webb, S. & Bird, A. (2014), 'Cfp1 is required for gene expression-dependent h3k4 trimethylation and h3k9 acetylation in embryonic stem cells', *Genome biology* **15**(9), 451.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A. et al. (2010), 'Histone h3k27ac separates active from poised enhancers and predicts developmental state', *Proceedings of the National Academy of Sciences* **107**(50), 21931–21936.

Cui, K., Zang, C., Roh, T.-Y., Schones, D. E., Childs, R. W., Peng, W. & Zhao, K. (2009), 'Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation', *Cell stem cell* **4**(1), 80–93.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001), 'Empirical bayes analysis of a microarray experiment', *Journal of the American statistical association* **96**(456), 1151–1160.

Fazzari, M. J. & Greally, J. M. (2004), 'Epigenomics: beyond cpg islands', *Nature Reviews Genetics* **5**(6), 446.

Fletcher, R. (2013), *Practical methods of optimization*, John Wiley & Sons.

Hocking, T. D., Goerner-Potvin, P., Morin, A., Shao, X., Pastinen, T. & Bourque, G. (2016), 'Optimizing chip-seq peak detectors using visual labels and supervised machine learning', *Bioinformatics* **33**(4), 491–499.

Ji, H., Li, X., Wang, Q.-f. & Ning, Y. (2013), 'Differential principal component analysis of chip-seq', *Proceedings of the National Academy of Sciences* **110**(17), 6789–6794.

Kim, J., Lee, Y., Lu, X., Song, B., Fong, K.-W., Cao, Q., Licht, J. D., Zhao, J. C. & Yu, J. (2018), 'Polycomb-and methylation-independent roles of ezh2 as a transcription activator', *Cell reports* **25**(10), 2808–2820.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J. et al. (2015), 'Integrative analysis of 111 reference human epigenomes', *Nature* **518**(7539), 317.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P. et al. (2012), 'Chip-seq guidelines and practices of the encode and modencode consortia', *Genome research* **22**(9), 1813–1831.

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H. & Roeder, R. G. (2013), 'H3k4me3 interactions with taf3 regulate preinitiation complex assembly and selective gene activation', *Cell* **152**(5), 1021–1036.

Liu, X., Wang, C., Liu, W., Li, J., Li, C., Kou, X., Chen, J., Zhao, Y., Gao, H., Wang, H. et al. (2016), 'Distinct features of h3k4me3 and h3k27me3 chromatin domains in preimplantation embryos', *Nature* **537**(7621), 558.

Lun, A. T. & Smyth, G. K. (2015), 'csaw: a bioconductor package for differential binding analysis of chip-seq data using sliding windows', *Nucleic acids research* **44**(5), e45–e45.

Ma, P., Castillo-Davis, C. I., Zhong, W. & Liu, J. S. (2006), 'A data-driven clustering method for time course gene expression data', *Nucleic acids research* **34**(4), 1261–1269.

Margueron, R. & Reinberg, D. (2011), 'The polycomb complex prc2 and its mark in life', *Nature* **469**(7330), 343.

Pfister, S. X., Markkanen, E., Jiang, Y., Sarkar, S., Woodcock, M., Orlando, G., Mavrommati, I., Pai, C.-C., Zalmas, L.-P., Drobnitzky, N. et al. (2015), 'Inhibiting wee1 selectively kills histone h3k36me3-deficient cancers by dntp starvation', *Cancer cell* **28**(5), 557–568.

Pinello, L., Xu, J., Orkin, S. H. & Yuan, G.-C. (2014), 'Analysis of chromatin-state plasticity identifies cell-type–specific regulators of h3k27me3 patterns', *Proceedings of the National Academy of Sciences* **111**(3), E344–E353.

Portela, A. & Esteller, M. (2010), 'Epigenetic modifications and human disease', *Nature biotechnology* **28**(10), 1057.

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J. & Nestler, E. J. (2013), 'diffreps: detecting differential chromatin modification sites from chip-seq data with biological replicates', *PloS one* **8**(6), e65598.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. & Oberdoerffer, S. (2011), 'Ctcf-promoted rna polymerase ii pausing links dna methylation to splicing', *Nature* **479**(7371), 74.

Song, Q. & Smith, A. D. (2011), 'Identifying dispersed epigenomic domains from chip-seq data', *Bioinformatics* **27**(6), 870–871.

Stark, R. & Brown, G. (2011), 'Diffbind: differential binding analysis of chip-seq peak data', *R package version* **100**, 4–3.

Starmer, J. & Magnuson, T. (2016), 'Detecting broad domains and narrow peaks in chip-seq data with hiddendomains', *BMC bioinformatics* **17**(1), 144.

Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. (2016), 'A comprehensive comparison of tools for differential chip-seq analysis', *Briefings in bioinformatics* **17**(6), 953–966.

Tsompana, M. & Buck, M. J. (2014), 'Chromatin accessibility: a window into the genome', *Epigenetics & chromatin* **7**(1), 33.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G., Otte, A. P. et al. (2002), 'The polycomb group protein ezh2 is involved in progression of prostate cancer', *Nature* **419**(6907), 624.

Viterbi, A. (1967), 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE transactions on Information Theory* **13**(2), 260–269.

Wei, Y., Xia, W., Zhang, Z., Liu, J., Wang, H., Adsay, N. V., Albarracin, C., Yu, D., Abbruzzese, J. L., Mills, G. B. et al. (2008), 'Loss of trimethylation at lysine 27 of histone h3 is a predictor of poor outcome in breast, ovarian, and pancreatic cancers', *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center* **47**(9), 701–706.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. et al. (2008), 'Model-based analysis of ChIP-Seq (MACS)', *Genome biology* **9**(9), R137.

Zucchini, W., MacDonald, I. L. & Langrock, R. (2017), *Hidden Markov models for time series: an introduction using R*, Chapman and Hall/CRC.

Appendices

A Q-function of the EM algorithm

$$Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}\right) = Q_0\left(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)}\right) + Q_1\left(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}\right) + Q_2\left(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}\right) + Q_3\left(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}\right)$$

such that

$$\begin{aligned}
Q_0\left(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{r=1}^{3}\left\{Pr\left(Z_1 = r|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log(\pi_r)\right\} + \\
&\quad + \sum_{j=2}^{M}\sum_{r=1}^{3}\sum_{s=1}^{3}\left\{Pr\left(Z_{j-1} = r, Z_j = s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log(\gamma_{rs})\right\}, \\
Q_1\left(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{j=1}^{M} Pr\left(Z_j = 1|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right), \\
Q_2\left(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{j=1}^{M} Pr\left(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\sum_{l=1}^{L} Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{..j},\mathbf{x};\boldsymbol{\Psi}^{(t)})\log f_{(2,l)}\left(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}\right) + \\
&\quad + \sum_{j=1}^{M} Pr\left(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\sum_{l=1}^{L} Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{..j},\mathbf{x};\boldsymbol{\Psi}^{(t)})\log(\delta_l), \quad \text{and} \\
Q_3\left(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{j=1}^{M} Pr\left(Z_j = 3|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right).
\end{aligned}$$

in which $f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right)$, $f_{(2,l)}\left(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}\right)$, and $f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right)$ are defined in equations (2) and (3), respectively..

B Forward, Backward, and Related Probabilities

Define, for $j = 1, \ldots, M$, the forward probabilities as

$$
\begin{aligned}
&f^p_{11} = \pi_1 f_1(\mathbf{y}_{..1}|\boldsymbol{\psi}_1), f^p_{12} = \pi_2 f_2(\mathbf{y}_{..1}|\mathbf{x};\boldsymbol{\psi}_2), f^p_{13} = \pi_3 f_3(\mathbf{y}_{..1}|\boldsymbol{\psi}_3),\\
&f^p_{j1} = \sum_{l=1}^{3} \gamma_{l1} f^p_{(j-1)l} f_1(\mathbf{y}_{..j}|\boldsymbol{\psi}_1),\\
&f^p_{j2} = \sum_{l=1}^{3} \gamma_{l2} f^p_{(j-1)l} f_2(\mathbf{y}_{..j}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2),\\
&f^p_{j3} = \sum_{l=1}^{3} \gamma_{l3} f^p_{(j-1)l} f_3(\mathbf{y}_{..j}|\boldsymbol{\psi}_3).
\end{aligned}
$$

Conversely, for $j = 1, \ldots, M$, define the backward probabilities as

$$
\begin{aligned}
b^p_{Mk} &= 1, \qquad \forall k = 1, 2, 3,\\
b^p_{j1} &= \sum_{l=1}^{3} \gamma_{1l} b^p_{(j+1)l} f_1(\mathbf{y}_{..(j+1)}|\boldsymbol{\psi}_1),\\
b^p_{j2} &= \sum_{l=1}^{3} \gamma_{2l} b^p_{(j+1)l} f_2(\mathbf{y}_{..(j+1)}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2),\\
b^p_{j3} &= \sum_{l=1}^{3} \gamma_{3l} b^p_{(j+1)l} f_3(\mathbf{y}_{..(j+1)}|\boldsymbol{\psi}_3).
\end{aligned}
$$

Then, we have the following posterior probabilities

$$
\begin{aligned}
P\left(Z_j = k|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f^p_{jk} b^p_{jk}}{\sum_{l=1}^{3} f^p_{Ml}}, \qquad \forall j = 1, \ldots, M \quad \text{and} \quad k = 1, 2, 3,\\
P\left(Z_{j-1} = l, Z_j = 1|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f^p_{(j-1)l} \gamma_{l1} f_1(\mathbf{y}_{..j}|\boldsymbol{\psi}_1) b^p_{j1}}{\sum_{l=1}^{3} f^p_{Ml}},\\
P\left(Z_{j-1} = l, Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f^p_{(j-1)l} \gamma_{l2} f_2(\mathbf{y}_{..j}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2) b^p_{j2}}{\sum_{l=1}^{3} f^p_{Ml}},\\
P\left(Z_{j-1} = l, Z_j = 3|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f^p_{(j-1)l} \gamma_{l3} f_3(\mathbf{y}_{..j}|\boldsymbol{\psi}_3) b^p_{j3}}{\sum_{l=1}^{3} f^p_{Ml}}, \forall j = 2, \ldots, M \text{ and } l = 1, 2, 3.
\end{aligned}
$$