Biometrics 1, 1–26 DOI: 00.0000/0.000.0.0.0
May 2020

Efficient Detection and Classification of Epigenomic Changes Under Multiple Conditions

Pedro L. Baldoni*, Naim U. Rashid, and Joseph G. Ibrahim*****

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

**email:* baldoni@email.unc.edu

***email:* naim@unc.edu

****email:* ibrahim@bios.unc.edu

Summary: 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, epigraHMM, exhibits superior performance to current tools and it is among the fastest algorithms available, while allowing the classification of combinatorial patterns of differential epigenomic activity and the characterization of chromatin regulatory states.

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

1. Introduction

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 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. Several landmark articles have identified specific genomic regions of changing (differential) epigenomic activity between conditions as drivers of cell differentiation (Creyghton et al., 2010), cancer progression (Varambally et al., 2002), and a number of human diseases (Portela and Esteller, 2010). Within 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). 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).

To quantify local epigenomic activity on a genome-wide scale, a common high-throughput assay is chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq). 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. Sequences are then mapped onto a reference genome through sequence alignment to determine their likely locations of origin. Genomic coordinates 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 reads mapped onto non-overlapping windows of fixed length tiling the genome (window read counts), forming the basis for downstream analyses. Across multiple conditions, regions exhibiting enrichment in at least one condition, but not across all conditions, indicate the presence of differential activity pertaining to the protein-DNA interaction of interest.

To date, many differential peak callers (DPCs) have been proposed (Song and Smith, 2011; Stark and Brown, 2011; Shen et al., 2013; Chen et al., 2015; Lun and Smyth, 2015; Allhoff et al., 2016). However, several challenges affect their ability to accurately detect regions of differential activity from the wide range of ChIP-seq experiments (Section 2). First, differential regions may be both short or broad in length, causing difficulty for methods optimized for a particular type of signal profile (Stark and Brown, 2011; Chen et al., 2015). Second, methods that pool experimental replicates together (Song and Smith, 2011) often exhibit more false positive calls compared to methods that jointly model replicates from each condition (Steinhauser et al., 2016). Third, the analysis of ChIP-seq data is often subject to complex biases that may vary across the genome, as differences in local read enrichment may depend on quantities such as the total read abundance in a given region or GC content (Teng and Irizarry, 2017). DPCs that disregard such complex effects or that solely rely on sample-specific global scaling factors or control subtraction methods (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 and Smyth, 2015). Reflecting these limitations, a recent 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). Moreover, only few methods are able to simultaneously test for differential activity across three or more conditions (Stark and Brown, 2011; Chen et al., 2015; Lun and

Smyth, 2015), or can classify specific differential combinatorial patterns. Altogether, these limitations can impact the drawing of accurate insights from modern epigenomic studies.

Here, we propose an efficient and flexible statistical method to identify differential regions of enrichment from epigenomic experiments with diverse signal profiles and collected under common multi-replicate, multi-condition settings. Our method overcomes the 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, it captures specific differential combinatorial patterns through a novel finite mixture model emission distribution within the HMM's differential state. 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 (GLM) is used to model the specific differential combinatorial pattern while accounting for sample- and window-specific normalization factors via offsets. Third, it enables the simultaneous detection and classification of epigenomic changes under three or more conditions, a novelty not yet available in any other DPC algorithm. The presented method, epigraHMM (Figure 1), offers additional benefits over current HMM-based DPC algorithms (Song and Smith, 2011; Allhoff et al., 2016) that includes a flexible GLM-based framework with an embedded mixture model, which allows the modeling of covariates of interest, the inclusion of model offsets for non-linear normalization and GC-bias adjustment, and a fast and accurate parameter estimation scheme via rejection-controlled EM algorithm (RCEM).

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 and Kouzarides, 2011). In turn, cellular processes such as gene transcription, gene silencing, DNA repair, replication, and recombination are also affected. 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 cancer studies, where their absence is often observed in multiple cancer types (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 enhancer of zeste homolog 2 (EZH2), a major component of the polycomb complex PRC2 that catalyzes the methylation of H3K27me3 (Margueron and Reinberg, 2011), is another example of a protein 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 and H3K36me3, and to 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 (Figure 1). In line with previous findings (Steinhauser et al., 2016), we observe that even current DPCs designed for broad data (Song and Smith, 2011; Allhoff et al., 2016) tend to detect either overly fragmented differential peaks or call regions exhibiting no difference in experimental signal between conditions as differential (Sections 4.2 and 5.1). The low specificity and sensitivity of such methods may impair the biological interpretation of the resulting peak calls in downstream analyses. Methods that rely on candidate peaks may also exhibit a compromised performance due to the limitations of single-sample peak callers in broad data (Stark and Brown, 2011; Chen et al., 2015). In addition, most 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. We find that the performance of such methods exhibit low sensitivity and specificity in calling differential regions in broad marks (Figure 1).

[Figure 1 about here.]

We assessed the performance of our model on ChIP-seq experiments characterized by broad peaks (H3K36me3, H3K27me3, and EZH2) and short peaks (H3K27ac, H3K4me3, and the transcription factor CTCF). In simulations (Section 4) and in data sets from the ENCODE Consortium (Landt et al., 2012), we show that our model addresses the issues of the current peak callers in broad data (Section 5.1), while being flexible for short peaks (Section 5.2) and comparable to the fastest DPCs regarding the computation time. We show that our method can also be utilized for genomic regulatory state segmentation when studying multiple types of epigenomic processes from a single condition or cell line (Section 5.3). The Web Appendices A1 and A2 present the data accession codes and the data pre-processing steps, respectively. Code implementing the method and to replicate the presented results are available in the Web Appendices 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 j^{th} 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 $N \times M$ matrix of window read counts spanning all windows, samples, and conditions. We assume that each window belongs to one of three possible hidden states: consensus background (state 1), differential (state 2), and consensus enrichment (state 3). Windows exhibiting low (high) enrichment across all conditions will be modeled by an emission distribution pertaining to the consensus background (enrichment) state. Windows exhibiting enrichment under at least one condition, but not all conditions, will be modeled by an emission distribution pertaining to the differential state. If G conditions are of interest, there are $L = 2^G - 2$ possible differential combinatorial patterns of enrichment and background across conditions at a given window. The emission distribution pertaining to the differential state models all L possible differential combinatorial patterns via a mixture model with mixture proportions $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)'$, such that $\sum_{l=1}^{L} \delta_l = 1$.

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 r^{th} 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

$$f_r(\mathbf{y}_{..j}|\boldsymbol{\psi}_r) = \prod_{h=1}^{G}\prod_{i=1}^{n_h} \frac{\Gamma(y_{hij}+\phi_r)}{y_{hij}!\Gamma(\phi_r)} \left(\frac{\phi_r}{\mu_{(r,hij)}+\phi_r}\right)^{\phi_r} \left(\frac{\mu_{(r,hij)}}{\mu_{(r,hij)}+\phi_r}\right)^{y_{hij}}, \quad r \in \{1,3\}, \tag{2}$$

with $y_{hij} \in \{0,1,2,\ldots\}$, 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 offset u_{hij} adjusts for technical or biological artifacts (such as the GC-content bias as presented in Teng and Irizarry (2017), Web Appendix B1) and allows the non-linear normalization of the signal profile across genomic windows, conditions, and samples (Web Appendix B1). 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.

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 l^{th} 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

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

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 mixture component l. 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 l^{th} differential combinatorial pattern. There are several advantages to such a parametrization for the differential emission distribution. For example, it 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. Utilizing a mixture model as the differential state emission distribution avoids the 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. With state-specific parameters, the current implementation of the method allows the direct modeling of continuous covariates (e.g. input controls; Web Appendix A3), for which a state-level testing of their effects on the read count distribution could be performed. In a simulation study and in real data analyses, however, we did not observe a significant improvement in performance in differential peak detection after accounting for the effect of input controls (Web Appendix B), a fact that has also been observed by others (Lun and Smyth, 2015).

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

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

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

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

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 Web Appendix B4. 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 $Pr\left(Z_{j-1}=r,Z_j=s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$, defined in Web Appendix B4, can be calculated through the Forward-Backward algorithm and $Pr(W_{jl}=1|Z_j=2,\mathbf{y}_{\cdot\cdot j},\mathbf{x};\boldsymbol{\Psi}^{(t)}) = f_{(2,l)}(\mathbf{y}_{\cdot\cdot j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}^{(t)})\delta_l^{(t)}/\sum_{k=1}^{L} f_{(2,k)}(\mathbf{y}_{\cdot\cdot j}|\mathbf{x}_k;\boldsymbol{\psi}_{(2,k)}^{(t)})\delta_k^{(t)}$ for $l=1,\dots,L$.

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 $\hat{\delta}_l^{(t+1)}=\sum_{j=1}^{M}Pr(Z_j=2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)})Pr(W_{jl}=1|Z_j=2,\mathbf{y}_{\cdot\cdot j},\mathbf{x};\boldsymbol{\Psi}^{(t)})/\sum_{j=1}^{M}Pr(Z_j=2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)})$. 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 (Web Appendix B4). We jointly estimate these quantities via the algorithm BFGS (Fletcher, 2013).

The estimation scheme is robust to situations where certain differential combinatorial patterns of enrichment are rare (Figure 5). 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 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. 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 (Web Appendix B2).

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 RCEM algorithm 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. 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.

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 j^{th} genomic window belongs to the differential HMM state, $j = 1, \ldots, M$. For a cutoff of posterior probability α, the nominal FDR is $\sum_{j=1}^{M}(1-\hat{\rho}_{j2})I(\hat{\rho}_{j2} \geqslant 1-\alpha)/\sum_{j=1}^{M} I(\hat{\rho}_{j2} \geqslant 1-\alpha)$, 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 nominal 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 Web Appendix B3. Additional details of proposed EM algorithm and the implemented code are available in the Web Appendices B4 and A4, 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 and 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 compare our model with other DPCs in a more realistic scenario with broad peaks, while also avoiding the assumption of a parametric model for the data.

4.1 *Read Count Simulation*

Read counts were simulated under different scenarios that varied regarding the type of histone modification mark (H3K36me3 and H3K27me3), number of genomic windows (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). We further assessed our model under different Signal-to-Noise Ratio (SNR) levels. 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 Web Appendix 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 10^6 genomic windows (see Web Appendix C1 for additional details). Results are shown for the SNR level estimated from the ENCODE project (mean count ratio between peak and enrichment regions of 3.21), different number of conditions, and different number of replicates per condition. Overall, no significant differences regarding the relative bias of parameter estimates were observed across simulations under different number of genomic windows. 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 scenarios with higher number of conditions, these results highlight the importance of experimental replicates to achieve precise parameter estimates. 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.

[Table 1 about here.]

4.2 *Sequencing Read Simulation*

We performed a second simulation study aiming to compare the proposed model with the current DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR. We used the simulation pipeline presented by Lun and 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 and Smyth (2015) and called peaks in advance using HOMER. Peaks were then used as input in the respective software for differential call. A hundred simulated data sets were generated and 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.

4.2.1 *Simulation Results.* Figure 2 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. Problematic cases from RSEG were not considered in this simulation study. Similar issues have been previously reported in other studies (Starmer and Magnuson, 2016). We observed that our method showed the highest observed sensitivity among all DPCs, regardless of the nominal FDR thresholding level, while maintaining a moderate observed FDR (Figure 2A). Here, the observed FDR is the proportion of genomic windows incorrectly called as differential out of the total number of windows called as differential. 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 Figure 2D). 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 2B, 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 2C, 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, our method 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 2D 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 Web Appendix C2 for additional results).

[Figure 2 about here.]

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 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 isogenic 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 Web Appendix D1 for the RNA-seq data processing).

We compared the genome-wide performance of the presented model with the widely used DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR, a list that covers a variety of algorithmic approaches for both broad and short data sets. For the methods that require a set of candidate regions to be specified *a priori*, ChIPComp and DiffBind, peaks were called in advance using MACS2 (Zhang et al. (2008); available at `https://github.com/macs3-project/MACS`) 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 (Web Appendix D1). 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. Analyses were adjusted by input control samples in methods that are designed to do so (ChIPComp, DiffBind, diffReps, and THOR). We did not observe a significant improvement in performance by accounting for input control effect or GC-content bias with epigraHMM and did not adjust our analyses for these effects. A discussion about the choice of the window size is presented in the Web Appendix D2. Results from the analysis of more than two cell lines are presented in the Web Appendix D3. Data accessing code, data pre-processing steps, method-specific parameters, and code to replicate the presented results are detailed in the Web Appendix A.

5.1 *Analysis of ChIP-seq Data From Broad Marks*

Methods were benchmarked regarding the coverage of differentially transcribed gene bodies. The histone modification H3K36me3 is known to be associated with gene transcription and enriched regions of this mark are usually deposited on transcribed gene bodies. Hence, the location of differential peaks of H3K36me3 is expected to agree with the location of differentially expressed genes. Following the analysis presented by Steinhauser et al. (2016), we defined 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 Web Appendix D1 and agree with those presented here. Protein-coding genes with total read count across cell lines under the 25^{th} 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 3A, we show receiver operating characteristic (ROC) curves for various methods and different nominal levels of FDR threshold for differential peaks of H3K36me3. Similar to the analysis of broad histone modification marks presented by Xing et al. (2012), we computed the observed true positive rates (false positive rates) on the window-level as the proportion of windows called as differential out of the total number of windows associated (not associated) with differentially transcribed genes (Web Appendix D1). 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 dependent on the set of candidate peaks and to perform best in scenarios with short peaks (Figure 3B; Steinhauser et al. (2016); Lun and Smyth (2015)). In Figure 3C, we show the observed sensitivity (y-axis) and the average differential peak size (kbp; x-axis) for various methods under different nominal FDR levels (the observed FDR is annotated next to each data point). Our model and RSEG, two HMM-based methods, tended to call broader differential peaks and exhibited better sensitivity than other methods. Yet, differential peaks called by RSEG often did not correspond to differential regions of enrichment (Figure 3D), a behavior that has been noted by others (Starmer and Magnuson, 2016) and also seen in simulated data (Figure 2). Our HMM-based method with a non-linear normalization scheme via model offsets allowed us to maintain a low observed FDR and a higher sensitivity than other DPCs. In Figure 3F, we show examples of differential peak calls for EZH2. Our method was among the most efficient algorithms due to our computational scheme, taking approximately 30 minutes to analyze genome-wide data (Figure 3E).

[Figure 3 about here.]

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)).

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 4 shows the main results from our analysis using short data sets. In Figures 4 A and C we show the median LFC and the Spearman correlation of ChIP-seq counts for differential CTCF and 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 sub optimal 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. Our model was able to properly call truly short differential peaks (Figures 4 B, D, and F) and was comparable to the non-HMM based methods regarding the computation time (Figure 4E), jointly calling differential peaks in less than 1.5 hour.

[Figure 4 about here.]

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 ($G = 3$) and sought to jointly classify 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 (Ernst and Kellis, 2012), a distinct problem not tackled by current DPCs.

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 (Figure 5). While the majority of the genomic regions exhibited no enrichment for any of the analyzed marks, regions of consensus enrichment were rare and represented only 2% of the analyzed genome (Figure 5A), as expected. Consensus background and differential regions mostly covered intergenic (66%) and protein-coding genic regions (61%), respectively. Differential genomic windows were mostly representing either transcribed chromatin states or repressed chromatin states (Figure 5B). All differential combinatorial patterns expected to be rare had associated mixture proportion estimates less than 0.05 and were pruned using our model selection scheme via BIC presented in Web Appendix B2 (Figure 5C).

[Figure 5 about here.]

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). To assign combinatorial patterns to differential peaks, we chose the combination pertaining to the most frequent mixture component across windows by using the maximum estimated mixture model posterior probability, $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \mathbf{\Psi}^{(t)})$, $j = 1, \ldots, M$. For genes overlapping differential regions associated with either transcribed or repressed chromatin states, we computed the distribution of transcripts per million (TPM) from matching RNA-seq experiments. Genes associated with transcribed chromatin states had a significantly higher distribution of TPM counts than genes associated with repressed chromatin states (Figure 5C). In fact, all known marker genes of the Helas3 cell line (Morin et al., 2008) were associated with differential enrichment for H3K36me3, measured by the associated average mixture posterior proportions of the overlapping differential peaks (Figure 5D). We detected broad differential regions of enrichment and the classification of differential combinatorial patterns agreed with their biological roles as well as the expression levels of associated gene bodies. Figure 5E shows an example of a genomic region with differential enrichment for the analyzed marks. We compare our results to ChromHMM with 3 states, a method developed for chromatin segmentation. Our method offers the benefit of simultaneously detecting differential peaks and classifying the combinatorial pattern of enrichment through mixture model posterior probabilities even in the context of genomic segmentation with highly diverse epigenomic marks (see Web Appendix D4). By using the BIC for model selection (Web Appendix B2), one can choose the number of biologically relevant mixture components to be included in the model, a task that may not be as straightforward in methods such as ChromHMM (see Supplementary Figure 4 in Ernst and Kellis (2012)).

6. Discussion

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 local dependency of ChIP-seq counts and is able to precisely detect broad and short differential regions of enrichment. Second, it utilizes a GLM-based framework with model offsets that account for potential non-linear biases in the data as well as a constrained parametrization across HMM states and mixture components. Our implementation of the RCEM algorithm led to genome-wide analyses of ChIP-seq data under a computational time comparable to some of the fastest current methods and was at least 5 times faster than current HMM-based algorithms. Our results highlight the importance of the inclusion of technical/biological replicates in the analysis of epigenomic data. Lastly, our method allows the simultaneous detection and classification of differential combinatorial patterns of enrichment from its embedded mixture model and the associated posterior probabilities under any number of conditions. In situations where certain combinatorial patterns are rare, pruning of specific combinatorial patterns is possible via model selection via BIC (Web Appendix B2). Our software has been implemented into an R package and is available for download (Web Appendix A).

References

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

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research* **21,** 381.

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

Clouaire, T., Webb, S., and Bird, A. (2014). Cfp1 is required for gene expression-dependent h3k4 trimethylation and h3k9 acetylation in embryonic stem cells. *Genome biology* **15,** 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,** 21931–21936.

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

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods* **9,** 215–216.

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

Hocking, T. D., Goerner-Potvin, P., Morin, A., Shao, X., Pastinen, T., and Bourque, G. (2016). Optimizing chip-seq peak detectors using visual labels and supervised machine learning. *Bioinformatics* **33,** 491–499.

Kim, J., Lee, Y., Lu, X., Song, B., Fong, K.-W., Cao, Q., Licht, J. D., Zhao, J. C., and Yu, J. (2018). Polycomb-and methylation-independent roles of ezh2 as a transcription activator. *Cell reports* **25,** 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,** 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,** 1813–1831.

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H., and Roeder, R. G. (2013). H3k4me3 interactions with taf3 regulate preinitiation complex assembly and selective gene activation. *Cell* **152,** 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 pre-implantation embryos. *Nature* **537,** 558.

Lun, A. T. and Smyth, G. K. (2015). csaw: a bioconductor package for differential binding analysis of chip-seq data using sliding windows. *Nucleic acids research* **44,** e45–e45.

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

Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J., and Marra, M. A. (2008). Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques* **45,** 81–94.

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

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

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

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

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

Starmer, J. and Magnuson, T. (2016). Detecting broad domains and narrow peaks in chip-seq data with hiddendomains. *BMC bioinformatics* **17,** 144.

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

Teng, M. and Irizarry, R. A. (2017). Accounting for gc-content bias reduces systematic errors and batch effects in chip-seq data. *Genome research* **27,** 1930–1938.

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,** 624.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory* **13,** 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,** 701–706.

Xing, H., Mo, Y., Liao, W., and Zhang, M. Q. (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS computational biology* **8,** e1002613.

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,** R137.

Supporting Information

Web Appendices A-D, referenced in Sections 1-6, and a link to the implemented software are available with this article at the Biometrics website on Wiley Online Library.

Received May 2020. *Revised May* 2020. *Accepted May* 2020.

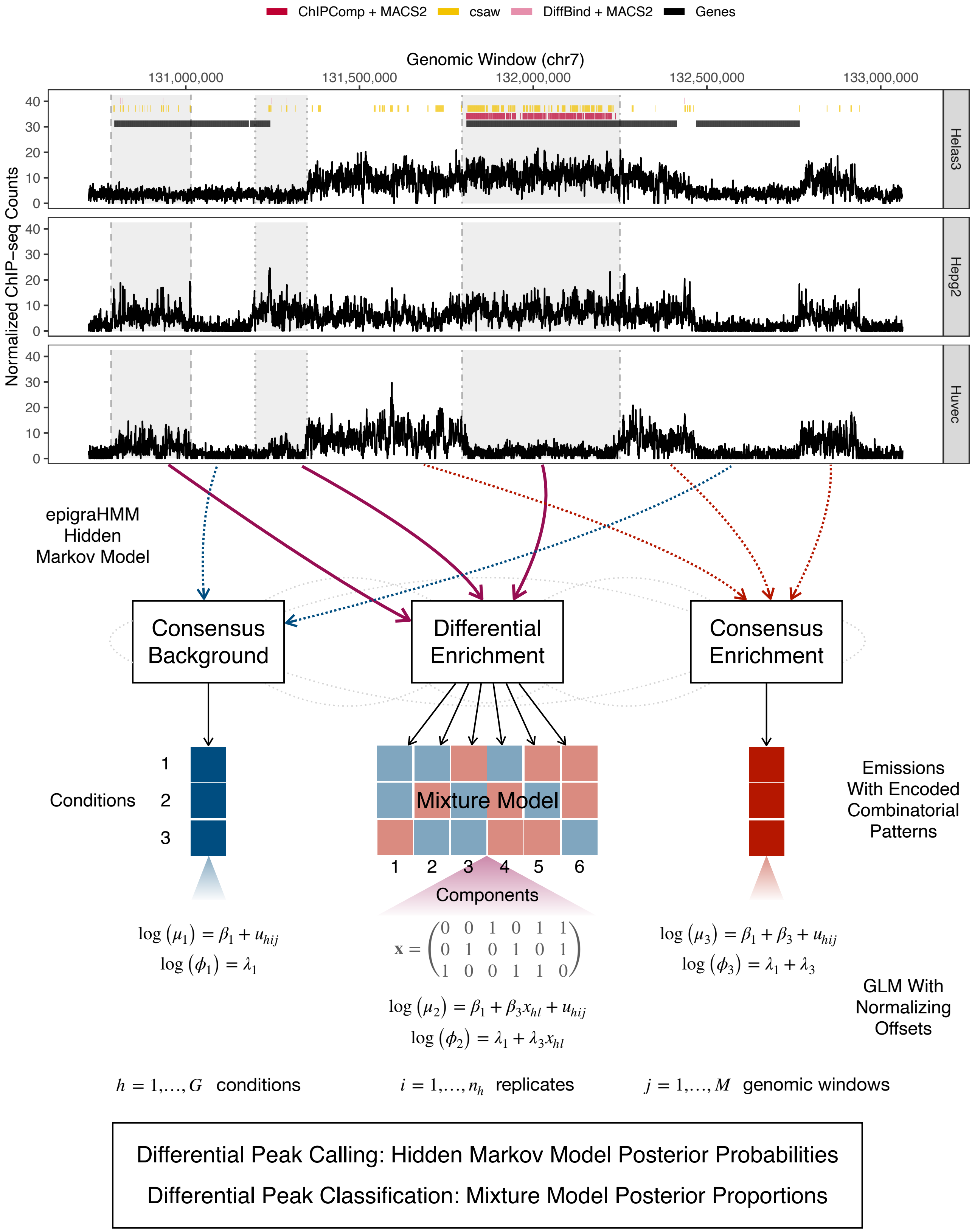
ChIPComp + MACS2
csaw
DiffBind + MACS2
Genes
Genomic Window (chr7)
131,000,000
131,500,000
132,000,000
132,500,000
133,000,000
Normalized ChIP−seq Counts
0
10
20
30
40
Helas3
Hepg2
Huvec
epigraHMM
Hidden
Markov Model
Consensus Background
Differential Enrichment
Consensus Enrichment
Conditions
1
2
3
Mixture Model
1 2 3 4 5 6
Components
Emissions With Encoded Combinatorial Patterns
$\log(\mu_1) = \beta_1 + u_{hij}$
$\log(\phi_1) = \lambda_1$
$\mathbf{x} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$
$\log(\mu_2) = \beta_1 + \beta_3 x_{hl} + u_{hij}$
$\log(\phi_2) = \lambda_1 + \lambda_3 x_{hl}$
$\log(\mu_3) = \beta_1 + \beta_3 + u_{hij}$
$\log(\phi_3) = \lambda_1 + \lambda_3$
GLM With Normalizing Offsets
$h = 1, \ldots, G$ conditions
$i = 1, \ldots, n_h$ replicates
$j = 1, \ldots, M$ genomic windows
Differential Peak Calling: Hidden Markov Model Posterior Probabilities
Differential Peak Classification: Mixture Model Posterior Proportions

Figure 1: At the top, differential peak calls from current methods (under false discovery rate control of 0.05) for the histone modification H3K27me3 among cell lines Helas3, Hepg2, and Huvec. 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 methods would call broad peaks inside shaded regions and no peaks outside them. At the bottom, general overview of the presented method for simultaneous detection and classification of broad and short differential regions of enrichment. This figure appears in color in the electronic version of this article.

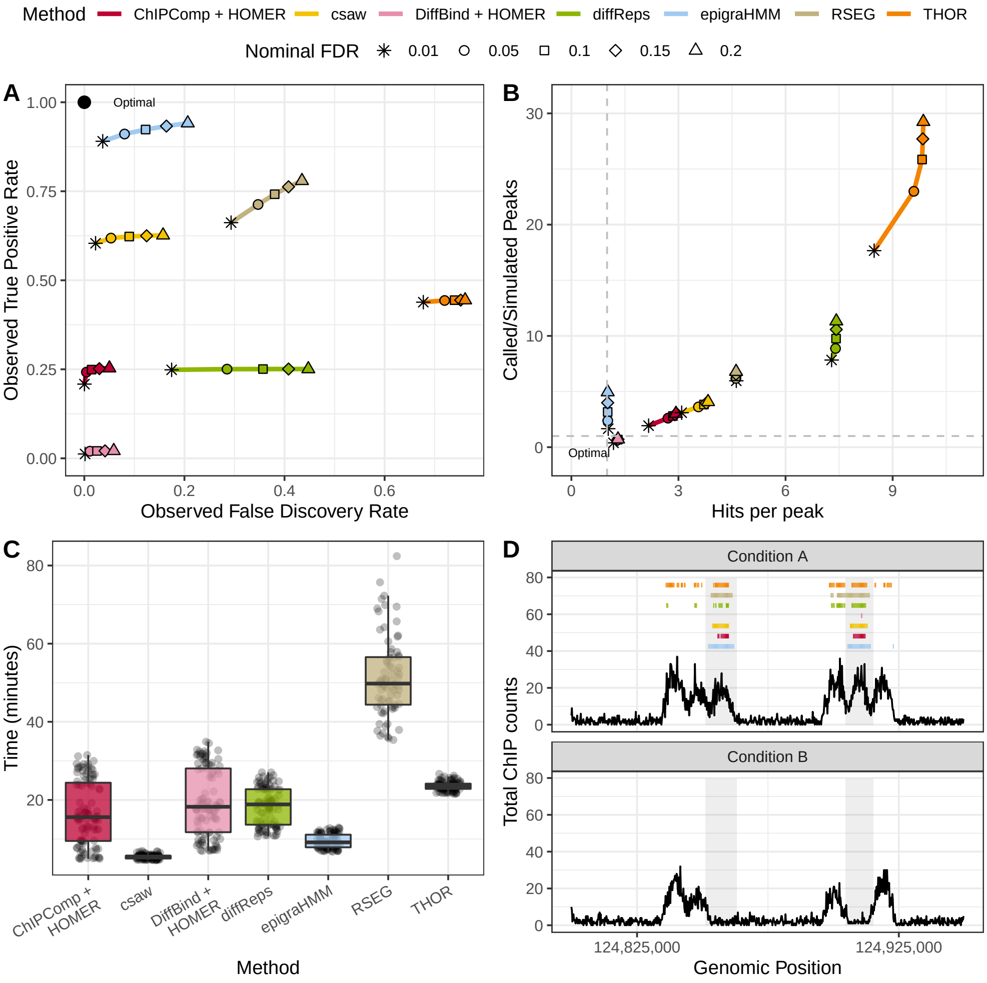

Figure 2: Sequencing read-based simulation from the csaw pipeline. (A): average observed sensitivity and FDR for various methods. (B): scatter plot of average ratio of called and simulated peaks (y-axis) and number of called peaks intersecting true differential regions (x-axis). (C): box plot of computing 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. This figure appears in color in the electronic version of this article.

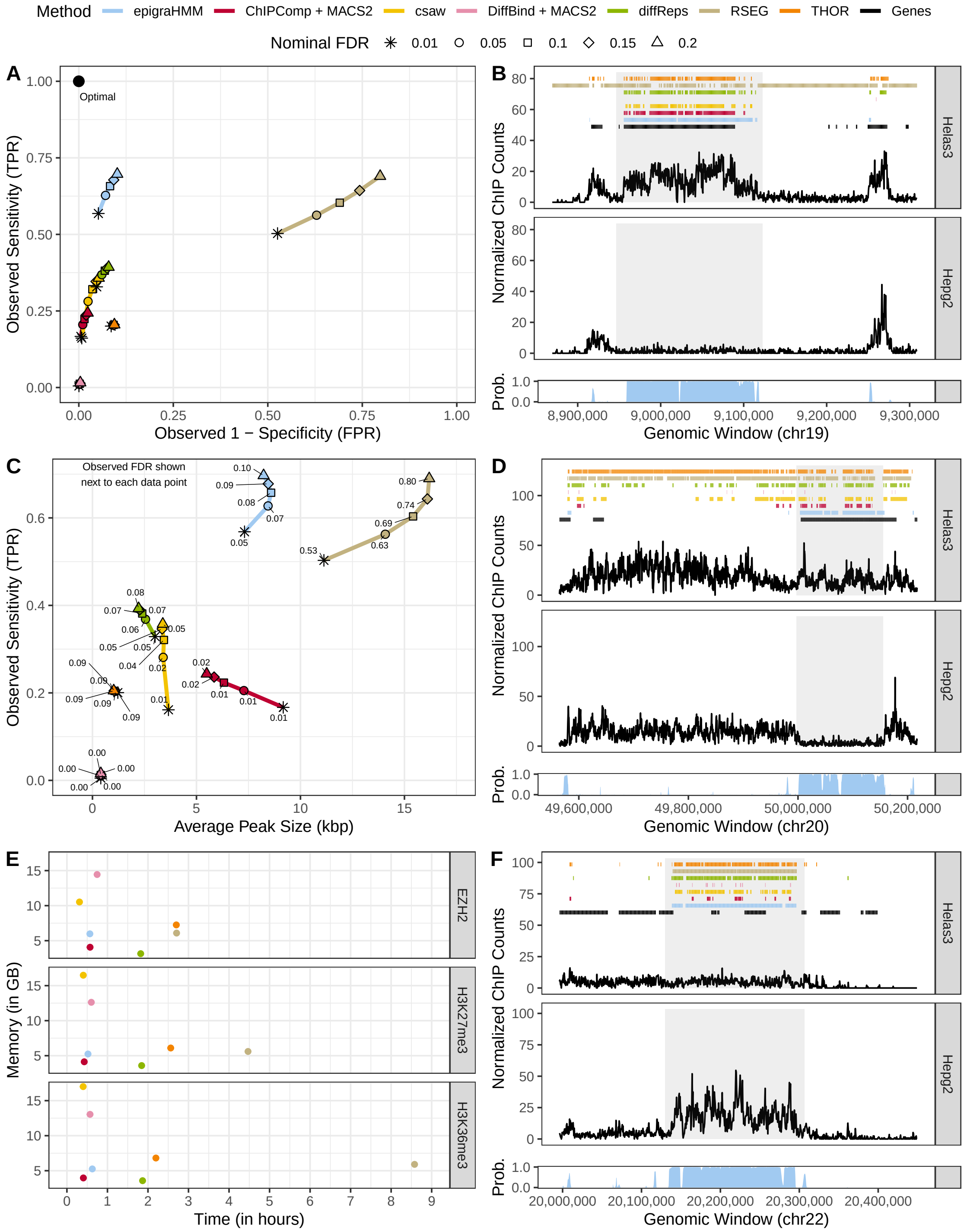

Figure 3: Analysis of broad ENCODE data. (A): ROC curves of H3K36me3 differential peak calls. (C): sensitivity (y-axis) and average H3K36me3 differential peak size (kbp; x-axis) of various methods under different nominal FDR thresholds (observed FDR annotated next to data points). (B), (D), and (F): 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 and peak memory usage of genome-wide analysis from various methods. This figure appears in color in the electronic version of this article.

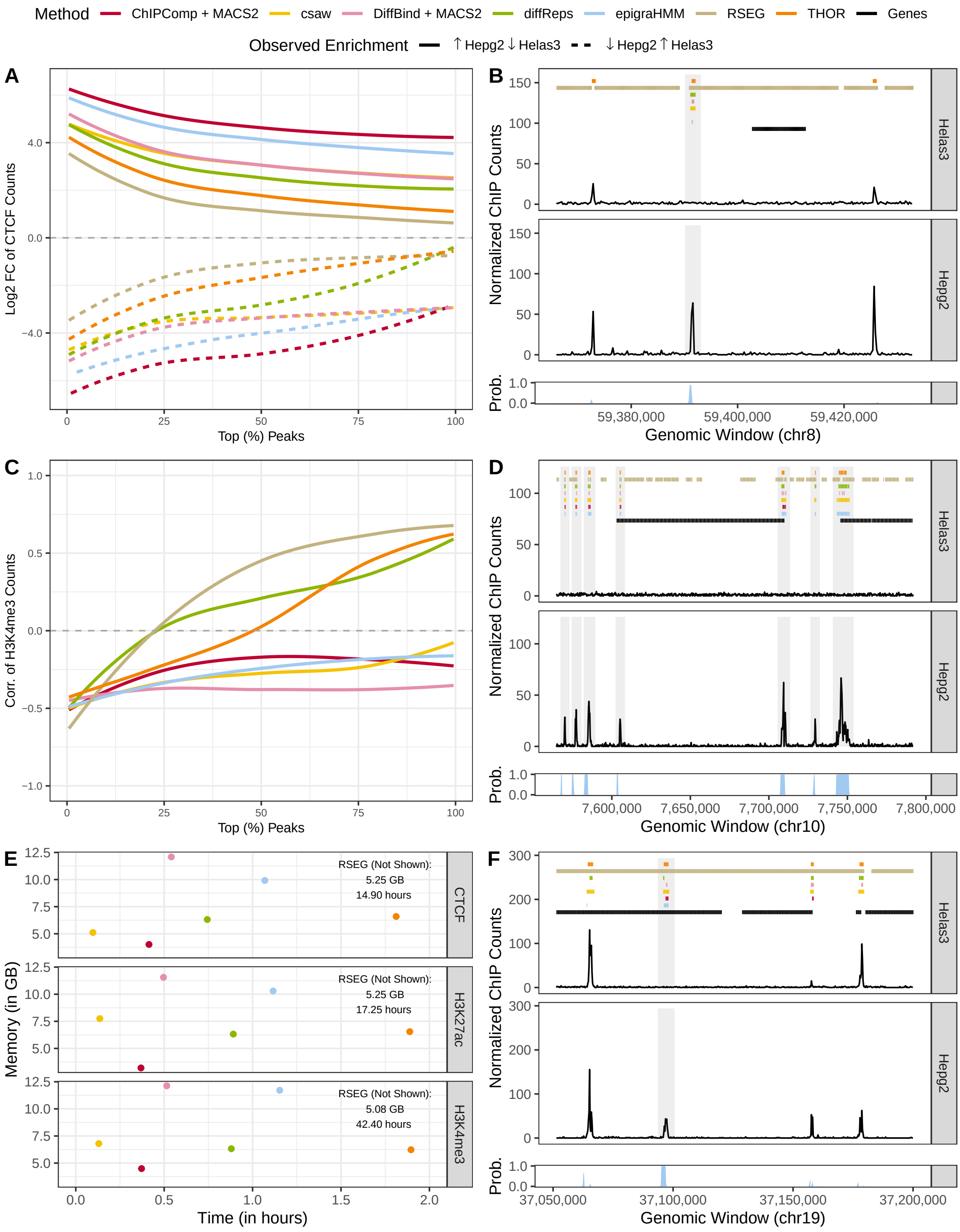

Figure 4: Analysis of short ENCODE data. (A) and (C): median LFC and correlation between cell lines of ChIP-seq counts from differential peaks for CTCF and H3K4me3, respectively. (B), (D), and (F): example of peak calls from CTCF, H3K4me3, and H3K27ac, respectively. Posterior probabilities of the HMM differential state are shown at the bottom of each panel. (E): computing time and peak memory usage of genome-wide analysis from various methods. Results are shown under a nominal FDR control of 0.05. This figure appears in color in the electronic version of this article.

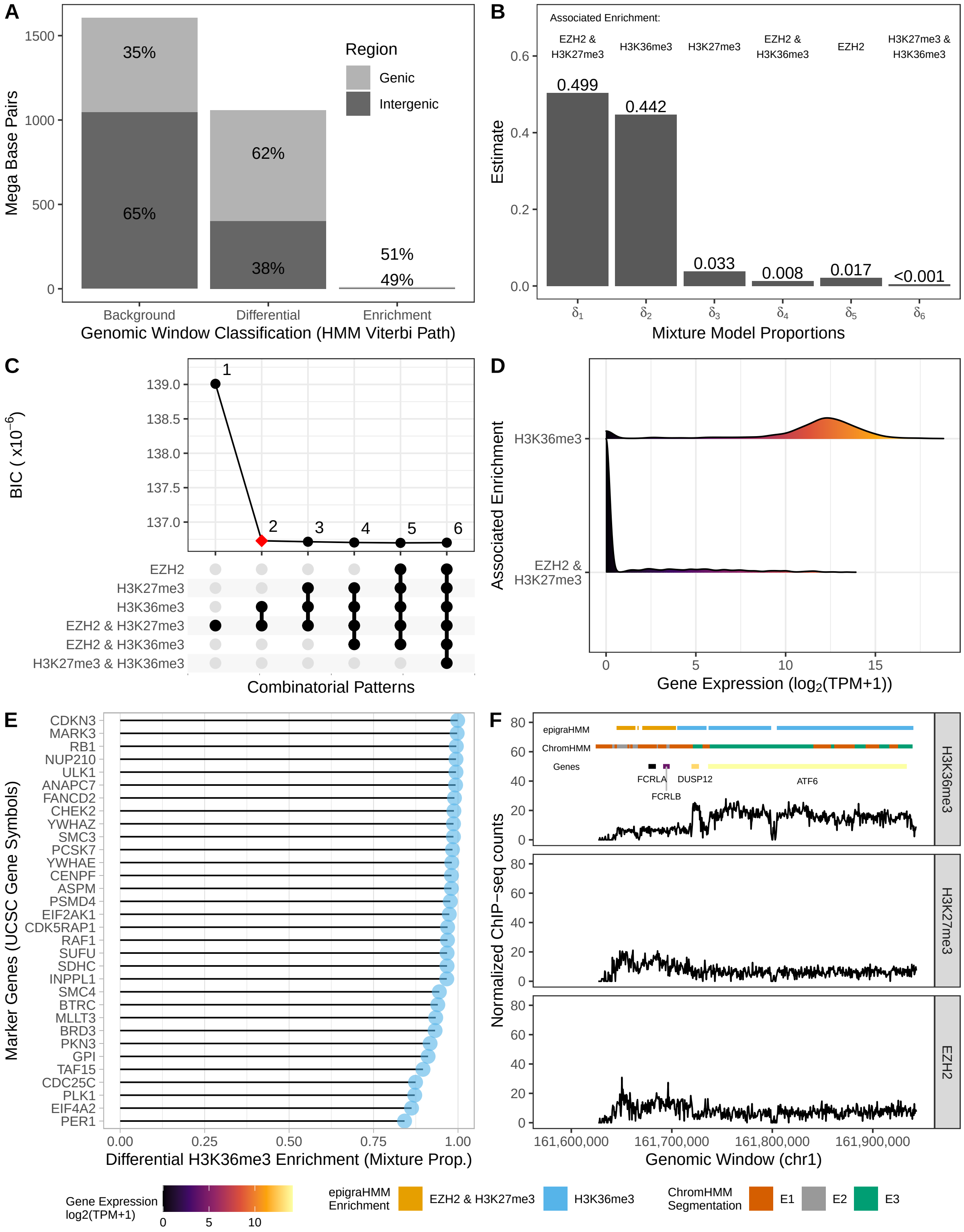

Figure 5: Genomic chromatin state segmentation and classification. (A): distribution of base pairs (y-axis) and the Viterbi sequence of states (x-axis). (B): estimated mixture probabilities and the associated differential combinatorial patterns. (C): choice of best model with two mixture components via BIC. (D): 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): association with differential H3K36me3 enrichment based based on mixture model posterior proportions of known marker genes. (E): example of a genomic region with differential peaks and genes, colored according to their classification and expression levels, respectively. This figure appears in color in the electronic version of this article.

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 data with 10^6 genomic windows and ENCODE-estimated SNR.

			One Replicate		Two Replicates		Four Replicates	
Conditions	Parameter	True Value	Relative Bias	$(P_{2.5}, P_{97.5})$	Relative Bias	$(P_{2.5}, P_{97.5})$	Relative Bias	$(P_{2.5}, P_{97.5})$
Two	β_1	1.116	0.000	(-0.001, 0.001)	0.000	(-0.001, 0.001)	0.000	(-0.001, 0.001)
	β_3	1.165	0.000	(-0.002, 0.001)	0.000	(-0.001, 0.001)	0.000	(-0.001, 0.001)
	λ_1	1.281	0.000	(-0.004, 0.004)	0.000	(-0.003, 0.003)	0.000	(-0.002, 0.002)
	λ_3	0.124	0.004	(-0.057, 0.066)	-0.004	(-0.047, 0.038)	-0.003	(-0.035, 0.034)
Three	β_1	1.116	-0.003	(-0.005, -0.002)	0.000	(-0.001, 0.000)	0.000	(-0.001, 0.001)
	β_3	1.165	-0.007	(-0.010, -0.005)	0.000	(-0.001, 0.001)	0.000	(-0.001, 0.001)
	λ_1	1.281	-0.001	(-0.006, 0.003)	0.001	(-0.002, 0.004)	0.000	(-0.002, 0.002)
	λ_3	0.124	-0.317	(-0.404, -0.223)	-0.023	(-0.057, 0.012)	-0.003	(-0.025, 0.024)
Four	β_1	1.116	-0.014	(-0.015, -0.012)	-0.007	(-0.008, -0.006)	0.000	(-0.001, 0.000)
	β_3	1.165	-0.111	(-0.114, -0.109)	-0.003	(-0.004, -0.002)	0.000	(-0.001, 0.001)
	λ_1	1.281	-0.012	(-0.016, -0.007)	0.007	(0.004, 0.010)	0.000	(-0.001, 0.002)
	λ_3	0.124	-3.520	(-3.589, -3.439)	-0.360	(-0.446, -0.282)	-0.005	(-0.025, 0.017)