Biometrics 1, 1–21 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.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Epigenomics, the study of the genome and its interactions with proteins and cellular elements, has become of significant interest in recent years. Such interactions may regulate essential cellular functions, such as gene expression, resulting in downstream phenotypic impact (Kim et al., 2018). Hence, 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 genomic regions of changing (differential) epigenomic activity between conditions as drivers of cell differentiation (Creyghton et al., 2010) and a number of human diseases (Portela and Esteller, 2010). Within differential regions, delineating 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).

To quantify local epigenomic activity, a common high-throughput assay is chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq). ChIP-seq begins with cross-linking DNA and proteins within chromatin structures, which are then fragmented by sonication. DNA fragments bound to the protein of interest are isolated by immunoprecipitation and sequenced via high-throughput sequencing to generate short reads pertaining to the original fragments. Sequences are then mapped onto a reference genome to determine their likely locations of origin. Genomic coordinates containing a high density of mapped reads (enrichment regions, peaks) indicate likely locations of protein-DNA interaction sites, while other regions are referred to as background regions. The read density is often summarized by counting reads mapped onto non-overlapping windows of fixed length (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. However, several

challenges affect their ability to accurately detect regions of differential activity from the wide range of experiments (Section 2). First, differential regions may be both short or broad, causing difficulty for methods optimized for a particular type of peak profile (Stark and Brown, 2011; Chen et al., 2015). Second, pooling experimental replicates often leads to lower specificity in comparison to joint modeling samples from each condition (Song and Smith, 2011). Third, ChIP-seq data is often subject to complex biases that vary across the genome, as differences in local read enrichment may depend on quantities such as local read abundance or GC content (Teng and Irizarry, 2017). DPCs that ignore such effects, or rely on global scaling factors or control subtraction methods (Shen et al., 2013; Allhoff et al., 2016), may be subject to spurious differences due to the lack of non-linear normalization methods (Lun and Smyth, 2015). In ChIP-seq data with broad peaks, a recent comparison showed that methods often detect a large number of either short or false positive peaks (Steinhauser et al., 2016). Moreover, only few methods can detect or classify differential combinatorial patterns across any number of conditions (Stark and Brown, 2011; Chen et al., 2015; Lun and Smyth, 2015).

Here, we present epigraHMM (Figure 1), an efficient and flexible statistical method to identify differential regions of enrichment from epigenomic experiments with diverse signal profiles. We overcome 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 from broad and short ChIP-seq data collected under multi-replicate, multi-condition settings. Second, it models specific combinatorial patterns of enrichment via a finite mixture model emission distribution within the HMM differential state. Each mixture component pertains to a particular differential combinatorial pattern originated by the presence or absence of local enrichment across conditions, where a generalized linear model (GLM) encodes specific differential combinatorial patterns while accounting for sample- and window-specific normalizing offsets. Third, it simultaneously detects and classifies epigenomic changes under

three or more conditions, a novelty not yet available in any other DPC algorithm. epigraHMM offers benefits over current HMM algorithms (Song and Smith, 2011; Allhoff et al., 2016) with an embedded GLM mixture model that allows the modeling of covariates of interest, the inclusion of normalizing offsets for non-linear biases (such as GC-content bias), and a fast estimation scheme via rejection-controlled EM algorithm (RCEM; Ma et al. (2006)).

2. Data

Histones are proteins that condense the DNA in eukaryotic cells into units called nucleosomes and, when enzymatically modified, mediate changes in local DNA packaging and chromatin accessibility. Hence, cellular processes such as gene transcription, gene silencing, and DNA repair are also affected. Proteins that interact with DNA and alter its functional properties are often referred to as epigenomic marks. The histone modifications H3K36me3 and H3K27me3 are example of marks that associate with genomic loci containing transcribed and repressed genes (Liu et al., 2016), respectively, while exhibiting broad enrichment profiles. The enhancer of zeste homolog 2 (EZH2), a component of the complex PRC2 that catalyzes the methylation of H3K27me3 (Margueron and Reinberg, 2011), is another example of a protein characterized by broad enrichment domains and co-occurs with the activity of H3K27me3. Conversely, short peaks from histone modifications H3K27ac and H3K4me3 are usually deposited on the promoter regions of transcribed genes and several studies have associated their role with gene transcription (Creyghton et al., 2010). Similarly, the transcription factor CTCF is a protein that binds to short DNA motifs and is responsible for cellular processes that include the regulation of the chromatin 3D structure (Shukla et al., 2011).

Using ChIP-seq data pertaining to histone modifications H3K27me3 and H3K36me3, and to EZH2 from the ENCODE Consortium (Landt et al., 2012), 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 either call overly fragmented differential peaks or call regions exhibiting no difference in experimental signal between conditions as differential (Sections 4.2 and 5.1). 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). Moreover, most DPCs restrict their application to the analysis of two experimental conditions. For DPCs 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 quantify treatment responses on the epigenomic level or identify sets of processes working together to regulate local chromatin state (Clouaire et al., 2014).

[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 CTCF). In simulated and in real data from the ENCODE Consortium (Sections 4 and 5), our model addresses the issues of the current peak callers in broad data, while being flexible for short peaks 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). Web Appendices A1-A4 present data accession codes, data pre-processing steps, code implementing the method, and scripts to replicate the presented results, respectively.

3. Methods

3.1 *Statistical Model*

Let Y_{hij} denote the random variable pertaining to the 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 $\mathbf{y}_{\cdot\cdot j} = (y_{11j}, \ldots, y_{Gn_Gj})'$ denote the $N \times 1$ vector of window read counts across all samples and conditions, and let $\mathbf{y} = (\mathbf{y}_{\cdot\cdot 1}, \ldots, \mathbf{y}_{\cdot\cdot M})$ 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, differential, and consensus enrichment. 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 $\sum_{l=1}^{L} \delta_l = 1$.

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 $r \in \{1, 2, 3\}$. Let $f_r(\mathbf{y}_{\cdot\cdot j}|\boldsymbol{\psi}_r)$ denote the emission distribution corresponding to the r^{th} hidden state, $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \boldsymbol{\delta}', \boldsymbol{\psi}')'$ denote the vector of all model parameters, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \boldsymbol{\psi}_2', \boldsymbol{\psi}_3')'$ denote each state's set of emission distribution-specific parameters, and $\mathcal{Z}$ denote the set of 3^M possible state paths of length M. The likelihood function of 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 $\mu_{(r,hij)}(1+\mu_{(r,hij)}/\phi_r)$. Conditional on the HMM state, we assume independence of read counts across samples and write the emission distribution of the consensus background and enrichment states as

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

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} may adjust for technical or biological artifacts such as the GC-content bias and allows the non-linear normalization of read counts across genomic windows, replicates, and conditions (Web Appendix B1; Teng and Irizarry (2017)).

In the differential state ($r = 2$), read counts are modeled by a L-component finite mixture model with mixture components that follow a NB distribution, where each component corresponds to a particular differential combinatorial pattern. To define these patterns, consider sets $S_1, \ldots, S_L$ delineating 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 (*e.g.*, 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 $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_{(2,l)}$ denote the parameter vectors pertaining to the differential state and to the l^{th} mixture component, respectively. Conditional on the differential state, we assume independence of read counts across samples and write the finite mixture model emission distribution as

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

where $\mu_{(2,lhij)}$ and $\phi_{(2,lh)}$ 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 $\log(\mu_{(2,lhij)}) = \beta_1 + \beta_3 x_{hl} + u_{hij}$ and $\log(\phi_{(2,lh)}) = \lambda_1 + \lambda_3 x_{hl}$. That is, in the mixture component l, we utilize the same consensus background (consensus enriched) parametrization from (2) in all conditions that $\mathbf{x}_l$ specifies to be background (enriched) in the l^{th} differential combinatorial pattern. Such a parametrization ensures that windows exhibiting differential enrichment across conditions share common means and dispersions between the consensus background enrichment states, an assumption that 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. We evaluate the strength of these assumptions through multiple simulations and a real data benchmarking analysis in Sections 4 and 5.

3.2 *Estimation*

Here, consider a 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 $r = \{1, 3\}$. The data generating mechanism can be then interpreted as read counts pertaining to window j

being sampled from $f_{(2,l)}$, given $\boldsymbol{\psi}_{(2,l)}$ and $\mathbf{x}_l$, when $Z_j = 2$ (differential state) and $W_{jl} = 1$ (l^{th} differential combinatorial pattern). Denoting $\mathcal{W}$ as the set of L^M possible combinations of latent vectors $\mathbf{W}$, the likelihood function of the observed data (1) can be rewritten as

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

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 $Pr\left(Z_{j-1} = r, Z_j = s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$ can be calculated through the Forward-Backward algorithm (Web Appendix B4) 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 $l = 1,\ldots,L$.

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 of 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 estimate the remaining parameters $(\boldsymbol{\delta}',\beta_1,\beta_3,\lambda_1,\lambda_3)'$. First, mixture proportions can be estimated as $\hat{\delta}_l^{(t+1)} = \sum_{j=1}^{M}Pr(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)})Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{\cdot\cdot j},\mathbf{x};\boldsymbol{\Psi}^{(t)})/\sum_{j=1}^{M}Pr(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)})$. Then, estimates of $(\beta_1,\beta_3,\lambda_1,\lambda_3)'$ are obtained in a similar fashion to the parameter estimation of a weighted NB regression model (Web Appendix B4).

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. Here, we use a RCEM algorithm with threshold 0.05, which 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. Moreover, our estimation scheme allows genomic windows with equal distribution of counts across replicates and conditions to have their posterior probability aggregated during the M-step of the algorithm, which leads to a fast gradient-based optimization. Upon convergence, 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 can be used. 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, $j = 1, \ldots, M$. A posterior probability cutoff α is then chosen by controlling the total FDR $\sum_{j=1}^{M}(1 - \hat{\rho}_{j2})I(\hat{\rho}_{j2} \geqslant 1-\alpha)/\sum_{j=1}^{M} I(\hat{\rho}_{j2} \geqslant 1-\alpha)$, where $I(\cdot)$ is an indicator function. Differential peaks are formed by merging adjacent windows that either meet a nominal FDR level for the differential HMM state or belong to the same predicted state (Web Appendix B3). Under this set up, 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 domains with mixture model posterior probabilities.

The estimation scheme is robust to situations where certain differential combinatorial patterns of enrichment are rare (Figure 5), which results from the fact that ChIP-seq experiments often provide enough data to estimate the parameters $(\beta_1, \beta_3, \lambda_1, \lambda_3)'$ shared across all L mixture components and HMM states. If pruning differential combinatorial patterns of the mixture model 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 L^* and combinatorial patterns via BIC agrees with pruning of rare differential combinatorial patterns that one would not expect to observe biologically (Web Appendix B2). In addition, the current implementation of epigraHMM allows for the adjustment for external covariates, such as GC-content and input controls, via model offsets (Web Appendix A3). In real data analyses, however, we did not observe a significant influence of input controls on the sensitivity and specificity of differential peak calls, a fact that has also been noted by others (Lun and Smyth, 2015).

4. Simulation Studies

We evaluated epigraHMM in two simulation studies. First, we simulated read count-based data to assess the precision of the estimation scheme, the performance of differential peak detection, and the accuracy of the classification of differential combinatorial patterns (Section 4.1). Next, we utilized the simulation pipeline from Lun and Smyth (2015) to simulate ChIP-seq reads from experiments with broad differential peaks (Section 4.2). The aim of the second simulation study was to compare epigraHMM with other DPCs in a more realistic scenario with broad peaks, while avoiding the choice of a parametric model for the data.

4.1 *Read Count Simulation*

Read counts were simulated under scenarios that varied regarding the type of histone modification mark (H3K36me3 and H3K27me3), number of 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 epigraHMM under different Signal-to-Noise Ratio (SNR) levels, here defined as the ratio between the means of consensus enrichment and background emission distributions, by decreasing the mean ratio while maintaining the mean-variance relationship. Model parameters were estimated from ENCODE data (Web Appendix C1).

Simulated counts followed a NB distribution and were generated using a first-order Markov chain with 2^G states, representing every combination of background and enrichment across G conditions. We assessed whether epigraHMM was able to assign all $2^G - 2$ differential states to the differential HMM state, while precisely estimating model parameters and accurately classifying differential combinatorial patterns.

Table 1 presents results relative to the H3K27me3 simulation scenario with 10^6 genomic windows and ENCODE-estimated SNR (mean ratio of 3.21; see Web Appendix C1 for additional details). Depending on the number of conditions, the relative bias and the range of the reported percentiles tended to decrease as more replicates were included in the analyses. This effect was more pronounced when estimating β_3 and λ_3 in scenarios with four conditions, which highlights the importance of experimental replicates to achieve precise parameter estimates. Overall, the proposed estimation scheme led to precise parameter estimates and was robust to a data generating mechanism that was different than the one assumed by the proposed model. No significant differences regarding the relative bias of parameter estimates were observed across simulations under different number of genomic windows.

[Table 1 about here.]

4.2 *Sequencing Read Simulation*

Here, we compared epigraHMM with the widely used DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR, a list that covers a variety of algorithms for both broad and short ChIP-seq enrichment profiles, on data simulated with the pipeline from Lun and Smyth (2015) without a particular model assumption. Sequencing reads from ChIP-seq experiments were generated for two conditions and two replicates per condition. For the two-step DPCs ChIPComp and DiffBind, we followed Lun and Smyth (2015) and called peaks in advance using HOMER (Heinz et al., 2010), which were then used as input in the respective software for differential call. A hundred simulated data sets were generated

and peaks were called under multiple nominal FDR thresholds. For epigraHMM and RSEG, window-based posterior probabilities were used to control the total FDR (Section 3.2).

Overall, epigraHMM showed the highest observed sensitivity among all DPCs while maintaining an observed FDR close to the nominal value across a wide range of nominal FDR levels (Figure 2A). Here, the observed FDR is the proportion of windows incorrectly called as differential out of the total number of windows called as differential. Due to the excessive number of false positive differential peak calls, diffReps, RSEG, and THOR showed higher observed FDR levels than the nominal FDR values (Figure 2D). Overall, diffReps and THOR called an excessive number of short and discontiguous peaks while RSEG called regions that were usually wider than the simulated differential peaks and did not correspond to simulated differential enrichment (Figure 2B). Regarding the computation time, epigraHMM was among the fastest DPCs due to the efficient RCEM implementation (Figure 2C). Other HMM-based algorithms, RSEG and THOR, appeared to be the most computationally intensive and required longer amounts of time to analyze the data. In general, epigraHMM was able to consistently cover most of true differential regions with broad peaks while exhibiting a limited number of false discoveries (Figure 2D; Web Appendix C2). Results shown for RSEG do not include 27 cases in which the algorithm failed to analyze the data due to internal errors or called the entire genome as differential. Similar issues of RSEG have been reported by others (Starmer and Magnuson, 2016).

[Figure 2 about here.]

5. Application to ENCODE Data

We applied epigraHMM on ChIP-seq data from the ENCODE Consortium to detect differential peaks of several epigenomic marks across multiple cell lines. We analyzed data characterized by broad peaks (H3K36me3, H3K27me3, and EZH2; Section 5.1) and short

peaks (CTCF, H3K27ac, and H3K4me3; Section 5.2). Two isogenic replicates of each cell line 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 with gene expression (Section 5.3; Web Appendix D1).

We compared the genome-wide performance of epigraHMM with ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR. For two-step DPCs, ChIPComp and DiffBind, we used peak calls from MACS2 as candidate peaks (Zhang et al., 2008). Methods were compared in terms of the coverage of differentially enriched genes, the number and average size of differential peak calls, the $\log_2$ fold change (LFC) and Spearman correlation of $\log_2$ reads counts mapped onto differential peak calls, the computation time, and maximum memory utilized. Metrics for sensitivity and specificity were defined on the window level and based on the coverage of differentially enriched genes by called peaks (Web Appendix D1). Read counts were computed using non-overlapping windows of 500bp and 250bp for broad and short marks, respectively (see Web Appendix D2 for a discussion on window size). Results presented here pertain to the analysis of cell lines Helas3 and Hepg2 (see Web Appendix D3 for additional results). 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 (Web Appendix B1).

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

We compared methods regarding the coverage of differentially enriched genes by differential peak calls for H3K36me3, an epigenomic mark associated with transcribed genes. Following Ji et al. (2013), in which the authors define true differential binding sites from LFCs of ChIP-seq counts to assess model sensitivity (Figure 2E in Ji et al. (2013)), we defined a set of protein coding genes exhibiting $|\text{LFC}| > 2$ of ChIP-seq counts between Helas3 and Hepg2

cell lines as true differential binding sites (see Web Appendix D1 for results using different thresholding values). Median ratio normalization of counts was performed on protein-coding genes after excluding those with total counts under the 25^{th} percentile of the distribution. Figure 3A shows receiver operating characteristic (ROC) curves for various methods and different nominal FDR levels for H3K36me3 differential peak calls. Observed true (false) positive rates were computed on the window-level as the proportion of windows called as differential out of all windows associated (not associated) with differentially enriched genes (Web Appendix D1). Overall, epigraHMM outperformed other DPCs by covering most of the differentially enriched genes while calling a limited number of false positive peaks.

DiffBind, which has been shown to be dependent on the set of candidate peaks (Lun and Smyth, 2015), called the shortest peaks overall (Figures 3B and 3C) and ChIPComp, which has been shown to perform best in scenarios with short marks (Steinhauser et al., 2016), appeared to have limited performance in broad marks such as H3K27me3 and EZH2 (Figures 3D and 3F). epigraHMM and RSEG, two HMM-based methods, called broader differential peaks and exhibited better sensitivity than other methods. Yet, differential peaks from RSEG often did not correspond to observed differential enrichment, a fact that explains its high observed FDR and agrees with our simulations (Figures 3C and 2) and others (Starmer and Magnuson, 2016). Our HMM-based approach with a non-linear normalization via model offsets allowed us to maintain a low observed FDR and a higher sensitivity than other DPCs. Overall, epigraHMM 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 epigraHMM on ChIP-seq samples characterized by short peaks (CTCF, H3K4me3, and H3K27ac). The goal of our analysis was to assess whether

epigraHMM was flexible to different types of data and still able to call short differential regions of enrichment. Here, differential peaks are usually observed in isolated genomic regions and exhibit a high SNR. It has been shown that certain HMM-based algorithms, including RSEG, have low accuracy with such short marks (Hocking et al., 2016).

We calculated the LFC and the Spearman correlation between cell lines Helas3 and Hepg2 of ChIP-seq counts mapped onto differential peaks called by each method. Median ratio normalization of counts was performed prior to computing such metrics. Here, ideal methods would show high absolute LFC and negative correlation between read counts mapped onto their differential peaks. We observed that other HMM-based algorithms, 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 (Figures 4A and 4C). ChIPComp, DiffBind, and csaw, were among those with the best performance overall, as their differential peak calls had the highest absolute LFCs and the lowest Spearman correlation of counts between cell lines. Overall, epigraHMM performed comparably to these methods that are known to perform best in short epigenomic marks while calling differential peaks in less than 1.5 hour (Figures 4B and 4D-4F; Steinhauser et al. (2016)).

[Figure 4 about here.]

5.3 *Genomic Segmentation and Classification of Chromatin States*

We analyzed data from the cell line Helas3 to segment its genome regarding the joint activity of H3K36me3, H3K27me3, and EZH2. We considered each mark as a separate experimental condition ($G = 3$) and jointly classified local chromatin states based upon the presence or absence of enrichment from each mark. Assuming that EZH2 catalyzes the methylation of H3K27me3, a repressive mark, and H3K36me3 associates with transcribed genes (Section 2), we expected consensus enrichment regions to be rare and differential regions to be mostly represented by either transcribed chromatin states (enrichment of H3K36me3 alone) or

repressed chromatin states (co-enrichment for H3K27me3 and EZH2). The analyses presented here 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 observed that while the majority of genomic regions exhibited no enrichment for any of the analyzed marks, regions of consensus enrichment were rare and covered only 2% of the genome (Figure 5A), as expected. Consensus background and differential regions mostly covered intergenic and protein-coding genic regions, respectively. In fact, 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 estimated mixture model proportions less than 0.05, which were then pruned using a model selection scheme via BIC (Figure 5C; Web Appendix B2).

[Figure 5 about here.]

To assess the biological significance of our results, we associated the distribution of transcripts per million (TPM, from matching RNA-seq experiments) of protein-coding genes with combinatorial patterns of overlapping differential peaks. To assign combinatorial patterns to differential peaks, we chose the combination pertaining to the most prevalent mixture component across windows by using the maximum estimated mixture model posterior probability, $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \mathbf{\Psi}^{(t)})$, $j = 1, \ldots, M$. Genes associated with transcribed chromatin states had a significantly higher distribution of TPM counts than genes associated with repressed chromatin states (Figure 5D). In fact, nearly all known Helas3 cell cycle-regulated genes (therefore genes expressed at some stage of the cell cycle; Whitfield et al. (2002)) were associated with high average mixture posterior probabilities of differential enrichment for H3K36me3 alone (Figure 5E).

The expression of such genes are known to correlate with proliferative states of tumors and their study may help characterize their role in cancer (Whitfield et al., 2002). Differential

peaks called by epigraHMM, and their assigned differential combinatorial pattern, often agreed with the biological roles as well as the expression levels of associated genes (Figure 5F; results from ChromHMM with 4 states for comparison). epigraHMM offers the benefit of simultaneously detecting differential peaks and classifying differential combinatorial patterns of enrichment even in the context of genomic segmentation of highly diverse epigenomic marks (Web Appendix D4). By using the BIC for model selection, 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 (Web Appendix B2).

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 accounts for non-linear biases and other potential factors such as GC-content bias or input control samples. Our efficient 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 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. When certain differential combinatorial patterns are rare, our model selection scheme via BIC provides means of pruning specific patterns of interest (Web Appendix B2).

Acknowledgements

The authors wish to thank the editor, associate editor and three referees for helpful comments and suggestions, which have led to an improvement of this article. This research was partially supported by NIH grants GM70335, P01CA142538, P30CA016086, and P50CA058223 and by the Brazilian funding agency CAPES (13195/131).

Data Avalability Statement

The data that support the findings of this paper are openly available in the Encyclopedia of DNA Elements Project portal at https://www.encodeproject.org (Sloan et al., 2016).

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.

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.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell* **38,** 576–589.

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.

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

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.

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.

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.

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

Margueron, R. and Reinberg, D. (2011). The polycomb complex prc2 and its mark in life. *Nature* **469,** 343–349.

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.

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., Lee, B. T., et al. (2016). Encode data at the encode portal. *Nucleic Acids Research* **44,** D726–D732.

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.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell* **13,** 1977–2000.

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 paper at the Biometrics website on Wiley Online Library. MACS2 peak caller is available at https://github.com/macs3-project/MACS.

epigraHMM has been implemented as an R package and it is available for download at https://github.com/plbaldoni/epigraHMM. All of the scripts used in the analyses presented in this paper can be downloaded at https://github.com/plbaldoni/epigraHMMPaper.

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

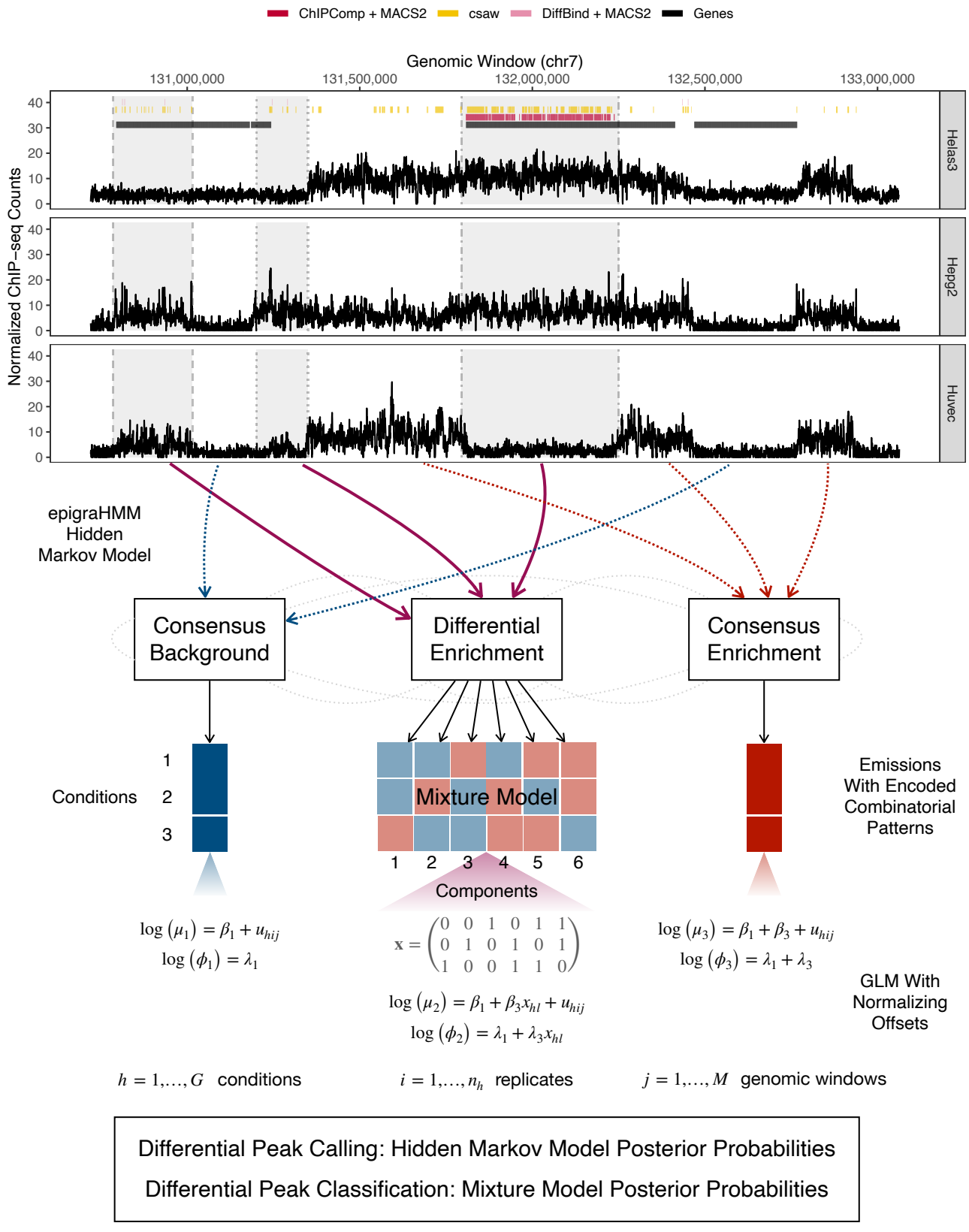

Figure 1: At the top, differential peak calls from current methods (under false discovery rate control of 0.05) for the histone modification H3K27me3 among cell lines Helas3, Hepg2, and Huvec. Shaded regions indicate observed differential enrichment, and each vertical line type bordering each region 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, and any mention of color refers to that version.

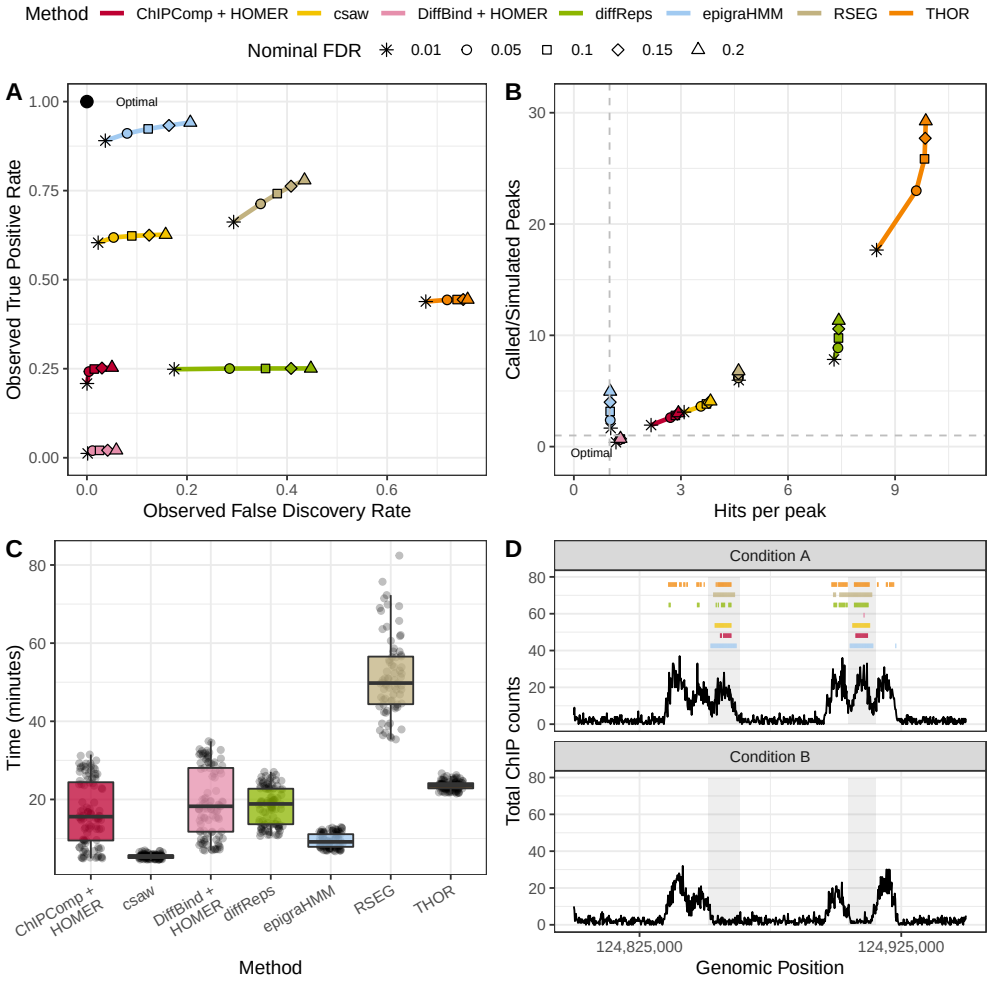

Figure 2: Sequencing read-based simulation from the csaw pipeline. (A): average observed sensitivity and FDR for various methods. (B): scatter plot of average ratio of called and simulated peaks (y-axis) and number of called peaks intersecting true differential regions (x-axis). (C): box plot of computing 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, and any mention of color refers to that version.

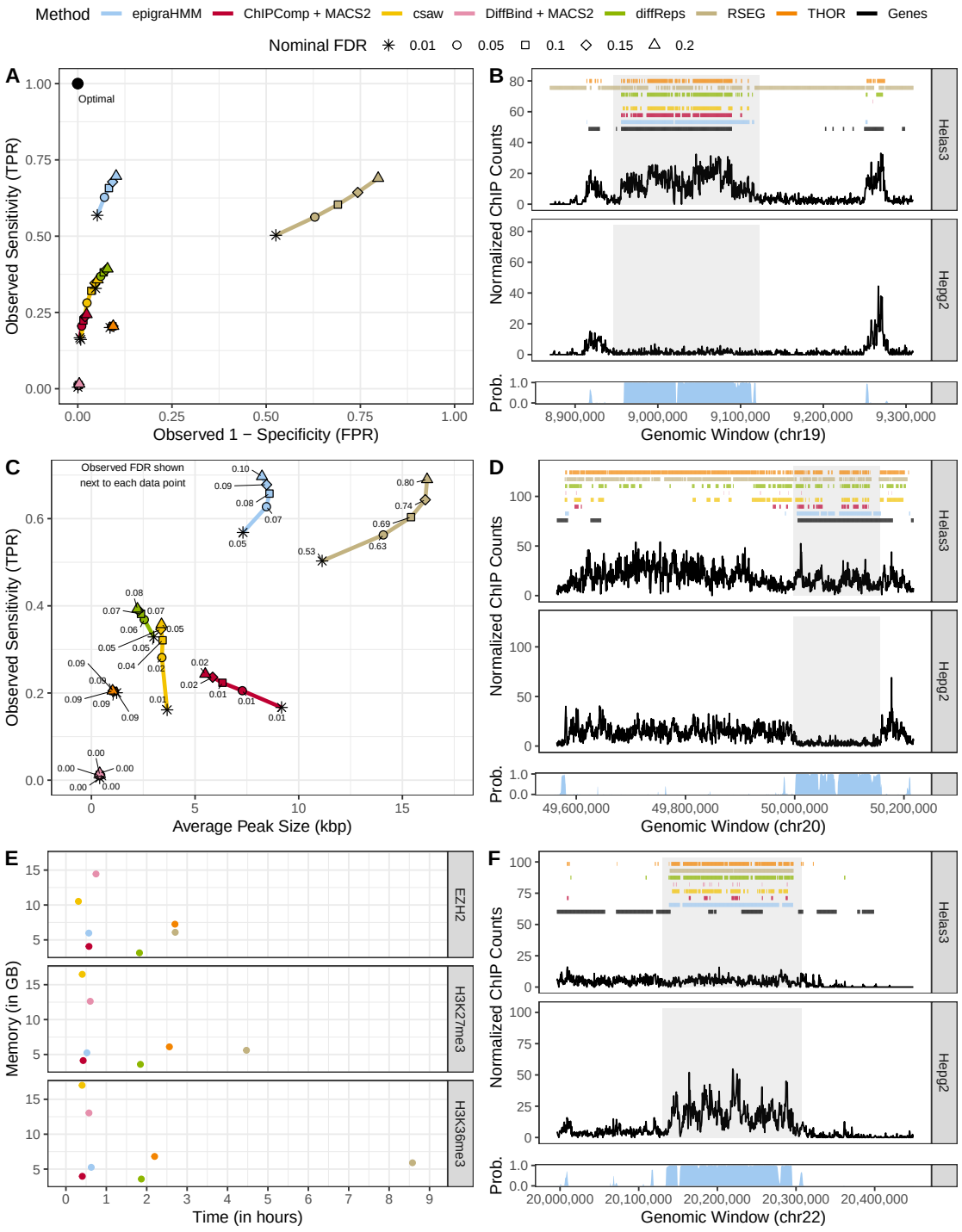

Figure 3: Analysis of broad ENCODE data. (A): ROC curves of DPCs covering differentially enriched genes ($|\text{LFC}| > 2$ of ChIP-seq counts) for H3K36me3. (C): sensitivity (y-axis) and average H3K36me3 differential peak size (kbp; x-axis) of various methods under different nominal FDR thresholds (observed 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. Shaded areas highlight differential peak regions. (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, and any mention of color refers to that version.

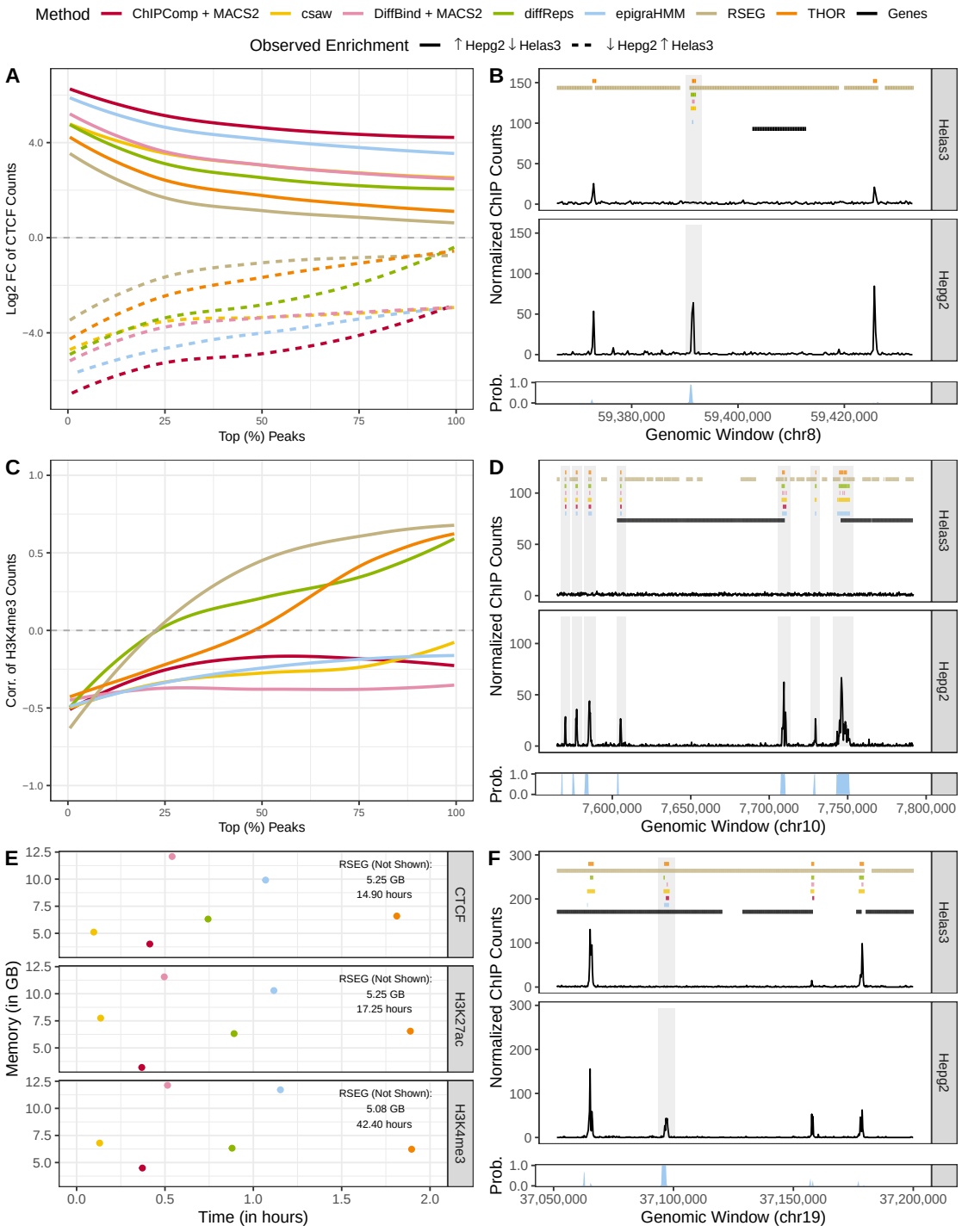

Figure 4: Analysis of short ENCODE data. (A) and (C): median LFC and correlation between cell lines of ChIP-seq counts from differential peaks for CTCF and H3K4me3, respectively. (B), (D), and (F): example of peak calls from CTCF, H3K4me3, and H3K27ac, respectively. Posterior probabilities of the HMM differential state are shown at the bottom of each panel. Shaded areas highlight differential peak regions. (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, and any mention of color refers to that version.

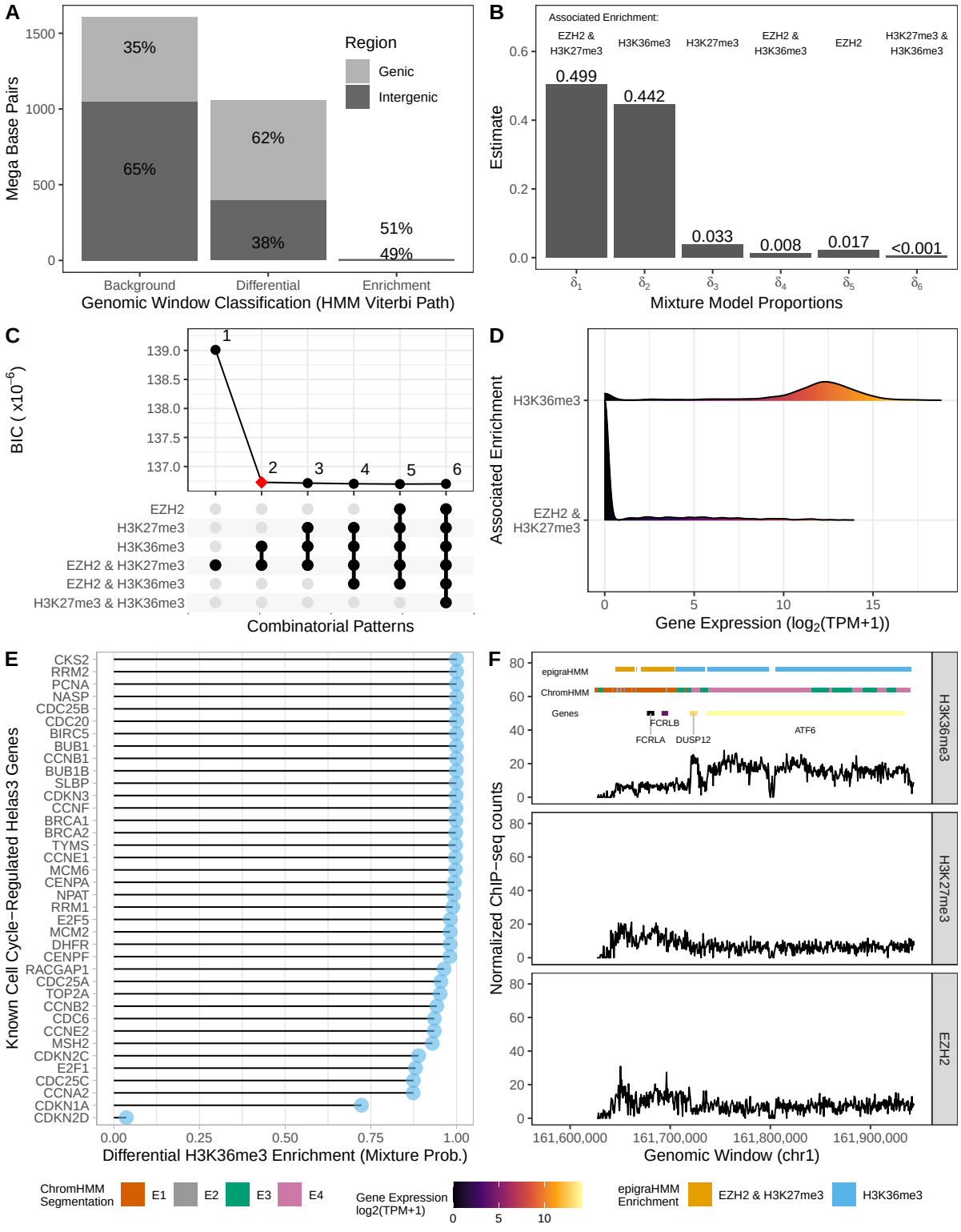

Figure 5: Genomic chromatin state segmentation and classification. (A): distribution of base pairs (y-axis) and the Viterbi sequence of states (x-axis). (B): estimated mixture probabilities and the associated differential combinatorial patterns. (C): choice of best model with two mixture components via 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. (E): cell cycle-regulated genes and their associated average mixture posterior probabilities associated with differential H3K36me3 enrichment pattern of overlapping differential peaks. (F): 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, and any mention of color refers to that version.

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)