The Annals of Applied Statistics
2016, Vol. 10, No. 1, 506–526
DOI: 10.1214/16-AOAS905

CHANGE POINT ANALYSIS OF HISTONE MODIFICATIONS REVEALS EPIGENETIC BLOCKS LINKING TO PHYSICAL DOMAINS

BY MENGJIE CHEN[1,*], HAIFAN LIN[2,†] AND HONGYU ZHAO[3,†]

University of North Carolina at Chapel Hill and Yale University†*

Histone modification is a vital epigenetic mechanism for transcriptional control in eukaryotes. High-throughput techniques have enabled whole-genome analysis of histone modifications in recent years. However, most studies assume one combination of histone modification invariantly translates to one transcriptional output regardless of local chromatin environment. In this study we hypothesize that the genome is organized into local domains that manifest a similar enrichment pattern of histone modification, which leads to orchestrated regulation of expression of genes with relevant biological functions. We propose a multivariate Bayesian Change Point (BCP) model to segment the *Drosophila melanogaster* genome into consecutive blocks on the basis of combinatorial patterns of histone marks. By modeling the sparse distribution of histone marks with a zero-inflated Gaussian mixture, our partitions capture local BLOCKs that manifest a relatively homogeneous enrichment pattern of histone marks. We further characterized BLOCKs by their transcription levels, distribution of genes, degree of co-regulation and GO enrichment. Our results demonstrate that these BLOCKs, although inferred merely from histone modifications, reveal a strong relevance with physical domains, which suggest their important roles in chromatin organization and coordinated gene regulation.

1. Introduction. Epigenetics refers to the study of heritable changes affecting gene expression and other phenotypes that occur without a change in DNA sequence. Epigenetic mechanisms, including chromatin remodeling, histone modification, DNA methylation and binding of nonhistone proteins, provide a fundamental level of transcriptional control. Extensive studies on histone modifications have led to the "histone code" hypothesis that histone modifications do not occur in isolation but rather in a combinatorial manner to provide "ON" or "OFF" signature for transcriptional events [Allis (2007)].

Genome-wide studies using high-throughput technologies such as chromatin immunoprecipitation (ChIP) followed by microarray analysis (ChIP on chip) or

Received May 2014; revised August 2015.

[1]Supported by National Institutes of Health Grants R01 CA082659 and P01 CA142538.

[2]Supported by National Institutes of Health Grant DP1 OD006825.

[3]Supported by National Institutes of Health Grants R01 GM59507, P01 CA154295 and P30 CA016359.

Key words and phrases. Bayesian change point model, Histone modification, chromosomal domain.

deep sequencing (ChIP-seq) have begun to decipher the "histone code" at the genome-wide scale. Currently, a common approach to assess chromatin states using these data is a multivariate Hidden Markov Model (HMM) introduced by Ernst and Kellis (2010), which has been used in several modENCODE and ENCODE project publications [modENCODE Consortium (2010), Kharchenko et al. (2011), Riddle et al. (2011), Eaton et al. (2011)]. This model associates each 200 bp genomic window with a particular state, generating a chromatin-centric annotation. However, a predefined number of states needs to be specified in HMMs and it is difficult to justify and interpret a particular choice. Different studies trying to balance resolution and interpretability based on different criteria often led to different numbers of states, both between different organisms [Ernst and Kellis (2010), modENCODE Consortium (2010)] and within the same organism [Filion et al. (2010), modENCODE Consortium (2010)]. Moreover, HMM summarizes chromatin information by a vector of "emission" probabilities associated with each chromatin state and a vector of "transition" probabilities with which different chromatin states occur in the spatial relationship of each other [Ernst and Kellis (2010)]. These settings assume the homogeneity of hidden states and their transitions across the genome. However, since histone modifications are outcomes of interplay with the local environment, the assumption of spatial homogeneity may not hold at the genome level.

To address the limitations in the HMM-based approaches, we propose an alternative approach to examining combinatorial histone marks at coarse scales. We hypothesize that the genome is organized into local blocks that display regionalized histone signatures. Those blocks may have important roles in the orchestrated regulation of the expression of genes with relevant biological functions. We note that our approach does not require a predefined number of possible states and it identifies local patterns without the assumption on spatial homogeneity.

To computationally infer these blocks, we propose a multivariate Bayesian Change Point (BCP) model which is capable of incorporating both local and global information. The BCP model was first proposed by Barry and Hartigan (1992, 1993) to describe a process where the observations can be considered to arise from a series of contiguous blocks, with distributional parameters different across blocks. One of the inferential goals is to identify the change points separating contiguous blocks. By "assuming probability of any partition is proportional to a product of prior cohesions, one for each block in the partition, and that given the blocks the parameters in different blocks have independent prior distributions" [Barry and Hartigan (1992, 1993)], a fully Bayesian approach can be adopted to detect change points from a sequence of observations. Barry and Hartigan (1992) considered in detail the case where the observations $X_1, \ldots, X_n$ are independent and normally distributed given the sequence of parameters μ_l with $X_i \sim N(\mu_l, \sigma^2)$ where the observations from the same block l have the same μ_l. This method has been used by Erdman and Emerson (2008) to segment microarray data. However, this model cannot be directly applied to infer histone modification blocks because

observed modification data do not follow normal distributions. This is due to the fact that histone modifications are usually observed at a small proportion of the genome locations with a signal at the rest of the genome being (or near) zero [see supplementary figure in Chen, Lin and Zhao (2016)]. To accommodate these unique features, here we present a multivariate BCP model through the introduction of a zero-inflated Gaussian mixture distribution to partition the genome into blocks where each block is relatively homogeneous with respect to histone marks.

1.1. *Outline of the paper.* We organized the paper as follows. In Section 2 we present the methodological details of the BCP model with a mixture prior and an MCMC algorithm to infer the posterior probability. Section 3 presents results from simulation studies. In Section 4 we describe a change point analysis of the *D. melanogaster* genome with multiple histone marks using S2 cell data from the modENCODE project. The identified chromosomal blocks are called BLOCKs in the rest of this article. Then we present two sets of exploratory analysis, Section 4.2 on BLOCKs' relationship with physical domains and Section 4.3 on the functional relevance of BLOCKs. In Section 4.4 we compare our results with HMM. We conclude the paper with a summary and discussion in Section 5.

1.2. *Notation.* We denote the density function of $N(\mu, \sigma^2)$ by $\phi(\cdot|\mu, \sigma)$, and denote the density function of Beta(a, b) by $\psi(\cdot|a, b)$. The Dirac function δ indicates the point mass at 0. For a set S, $\#S$ is the cardinality of S. For a random variable X, $\{X = 1\}$ is the indicator function taking value 1 if $X = 1$ and taking value 0 if $X \neq 1$. The indicator function $\{X = 0\}$ is defined in the same way. The set $\{i + 1, i + 2, \ldots, j\}$ with integers $i < j$ is denoted by $(i : j)$. The function $f(\cdot|\cdot)$ is a generic notation for conditional density when the distribution is clear in the context.

2. Method.

2.1. *A BCP model for block identification.* The observation we have is an $M \times n$ data matrix $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)^T$, where each $\mathbf{X}_m$ for $m = 1, \ldots, M$ is a modification mark with length n. We first describe the likelihood of each $\mathbf{X}_m$ and then combine them together. For notational simplicity, we suppress the subscript and write $\mathbf{X}$ instead of $\mathbf{X}_m$.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector with length n. Create another vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ to indicate whether each X_h is zero or not, that is, $Z_h = 0$ if $X_h = 0$, and $Z_h = 1$ if $X_h \neq 0$. Note $\mathbf{Z}$ is fully determined by $\mathbf{X}$.

For the index set $\{1, \ldots, n\}$, let ρ be a partition of this set, that is, $\rho = \{S_1, \ldots, S_N\}$, with $\{1, \ldots, n\} = \bigcup_{l=1}^{N} S_l$ and $S_{l_1} \cap S_{l_2} = \varnothing$ for all $l_1 \neq l_2$. The number N represents the number of blocks of $\{1, \ldots, n\}$. For the change-point problem, each S_l is a contiguous subset of $\{1, \ldots, n\}$, that is, $S_l = (i : j) = \{i + 1, \ldots, j\}$ for some $i < j$.

2.1.1. *Likelihood.* Given the partition $\rho = \{S_1, \ldots, S_N\}$, X_k follows a mixture distribution $X_k \sim (1-\lambda_l)N(\mu_l, \sigma^2) + \lambda_l \delta$, for $k \in S_l$ and each $l = 1, \ldots, N$. The parameter μ_l is block-specific, while σ is shared among different blocks. The parameter λ_l describes how likely X_k is zero, which varies across different blocks. Thus, given $(\rho, \mu_1, \ldots, \mu_N, \lambda_1, \ldots, \lambda_N, \sigma)$, the likelihood of $(\mathbf{X}, \mathbf{Z})$ can be fully specified. That is,

$$L(\mathbf{X}, \mathbf{Z}|\rho, \mu_1, \ldots, \mu_N, \lambda_1, \ldots, \lambda_N, \sigma) = \prod_{l=1}^{N} f(X_{S_l}, Z_{S_l}|\mu_l, \lambda_l, \sigma), \tag{2.1}$$

where for each l with $S_l = \{i+1, \ldots, j\}$,

$$f(X_{S_l}, Z_{S_l}|\mu_l, \lambda_l, \sigma) \tag{2.2}$$

$$= (1-\lambda_l)^{\#\{k \in S_l : Z_k = 1\}} \lambda_l^{\#\{k \in S_l : Z_k = 0\}} \prod_{\{k \in S_l : Z_k = 1\}} \phi(X_k|\mu_l, \sigma), \tag{2.3}$$

where $X_{S_l} = (X_{i+1}, \ldots, X_j)$ and $Z_{S_l} = (Z_{i+1}, \ldots, Z_j)$.

2.1.2. *Prior.* We proceed to specify the prior distribution on the parameters $(\rho, \mu_1, \ldots, \mu_N, \lambda_1, \ldots, \lambda_N, \sigma)$:

$$\rho \sim \prod_{l=1}^{N} c(S_l), \tag{2.4}$$

$$\mu_l \sim N\big(\mu_0, \sigma_0^2 d_l^{-1}\big) \tag{2.5}$$

$$\text{for each } l \text{ with } S_l = \{i+1, \ldots, j\}, \text{ and } d_l = \#\{k \in S_l : Z_k = 1\},$$

$$\lambda_l \sim \text{Beta}(a, b). \tag{2.6}$$

The prior (2.4) on the partition ρ is called the product partition model, which was originally described in Barry and Hartigan (1993). The quantity $c(S_l)$ is called cohesion. In this paper, $c(S_l)$ is defined to be $c_{(i:j)} = (1-p)^{j-i-1}p$ when $j < n$ and $c_{(i:j)} = (1-p)^{j-i-1}$ when $j = n$, where $0 \le p \le 1$ and $S_l = \{i+1, \ldots, j\}$ as mentioned before. This specification implies that the sequence of change points forms a discrete renewal process with inter-arrival times identically geometrically distributed. The geometric distribution has memoryless property. For histone mark data, it means we assume the possibility of a genomic position (bin) as a boundary for BLOCK is roughly constant. To note, the cohesion prior is a true nonparametric prior on all possible 2^{n-1} partitions for n data points, thus, the number of blocks does not need to be specified and can be inferred from the data. The priors (2.5) and (2.6) are conjugate priors with respect to the likelihood. The prior on the variance σ^2 will be jointly specified with the hyperparameters.

To pursue a fully Bayesian approach, we put priors on the hyperparameters (p, μ_0, σ_0) in (2.4) and (2.5). Define $w = \frac{\sigma^2}{\sigma^2+\sigma_0^2}$. We jointly specify the priors on the hyperparameters together with the prior on σ^2:

$$\mu_0 \sim 1, \qquad -\infty < \mu_0 < \infty, \tag{2.7}$$

$$\sigma^2 \sim \frac{1}{\sigma^2}, \qquad 0 \le \sigma^2 < \infty, \tag{2.8}$$

$$w \sim \frac{1}{w_0}, \qquad 0 \le w \le w_0, \tag{2.9}$$

$$p \sim \frac{1}{p_0}, \qquad 0 \le p \le p_0. \tag{2.10}$$

The priors (2.7), (2.9) and (2.10) are uniform priors. They reflect our ignorance of knowledge. The prior (2.8) can be viewed as a uniform distribution on the logarithmic scale. Notice (2.7) and (2.8) are improper priors. This will not cause a problem in view of our sampling procedure described later.

2.1.3. *Posterior.* Our goal here is to find the posterior distribution of the partition, which is $f(\rho|\mathbb{X}, \mathbb{Z})$. According to Bayes' formula,

$$f(\rho|\mathbb{X}, \mathbb{Z}) = \frac{\prod_{m=1}^{M} f(\mathbf{X}_m, \mathbf{Z}_m|\rho) f(\rho)}{\int \prod_{m=1}^{M} f(\mathbf{X}_m, \mathbf{Z}_m|\rho) f(\rho)\, d\rho}. \tag{2.11}$$

Since the denominator of (2.11) is complicated, we need to use MCMC to sample from the posterior by

$$f(\rho|\mathbb{X}, \mathbb{Z}) \propto \prod_{m=1}^{M} f(\mathbf{X}_m, \mathbf{Z}_m|\rho) f(\rho). \tag{2.12}$$

The conditional density $f(\mathbf{X}, \mathbf{Z}|\rho)$ is found by integrating out the likelihood function (2.1) using the prior of $(\mu_1, \ldots, \mu_N, \lambda_1, \ldots, \lambda_N, \sigma)$ specified in (2.5), (2.6), (2.7), (2.8) and (2.9). The prior $f(\rho)$ is found by integrating out $f(\rho|p)$ specified in (2.4) with respect to (2.10). We first find $f(\rho)$:

$$\begin{aligned} f(\rho) &= \frac{1}{p_0} \int_0^{p_0} f(\rho|p)\, dp \propto \frac{1}{p_0} \int_0^{p_0} \left(\prod_{l=1}^{N} c(S_l) \right) dp \\ &= \frac{1}{p_0} \int_0^{p_0} \left(\prod_{S_l = \{i+1, \ldots, j\}} c_{ij} \right) dp \\ &= \frac{1}{p_0} \int_0^{p_0} p^{N-1} (1-p)^{n-N}\, dp. \end{aligned} \tag{2.13}$$

Then, we continue to find $f(\mathbf{X}, \mathbf{Z}|\rho)$. We first integrate out $(\mu_1, \ldots, \mu_N, \lambda_1, \ldots, \lambda_N)$ in (2.1) using (2.5) and (2.6). Remember $\psi(\lambda_l, b)$ is the density of Beta(a, b). Using (2.2) as the representation of (2.1), we have

$$
\begin{aligned}
&f(\mathbf{X}, \mathbf{Z}|\rho, \mu_0, w, \sigma) \\
&\quad= \prod_{l=1}^{N} \int \prod_{\{k \in S_l : Z_k = 1\}} \phi(X_k|\mu_l, \sigma)\phi\bigl(\mu_l|\mu_0, \sigma_0 d_l^{-1/2}\bigr)\, d\mu_l \\
&\qquad\times \prod_{l=1}^{N} \int_0^1 (1-\lambda_l)^{\#\{k \in S_l : Z_k = 1\}} \lambda_l^{\#\{k \in S_l : Z_k = 0\}} \psi(\lambda_l|a, b)\, d\lambda_l \\
&\quad\propto A \times \bigl(2\pi\sigma^2\bigr)^{-T/2} w^{N/2} \exp\biggl(-\frac{1}{2\sigma^2}\bigl(W + wB + wT(\mu_0 - \bar{X}_T)^2\bigr)\biggr),
\end{aligned}
\tag{2.14}
$$

where

$$
\begin{aligned}
&T = \sum_{k=1}^{n} \{Z_k = 1\}, \\
&\bar{X}_T = T^{-1} \sum_{k=1}^{n} X_k, \\
&\bar{X}_{(i:j), Z_k} = \frac{1}{\#\{Z_k = 1, i < k \le j\}} \sum_{\{k : Z_k = 1, i < k \le j\}} X_k, \\
&W = \sum_{\{(i:j) = S_l \in \rho\}} \sum_{\{k : Z_k = 1, i < k \le j\}} (X_k - \bar{X}_{(i:j), Z_k})^2, \\
&B = \sum_{\{(i:j) = S_l \in \rho\}} \#\{Z_k = 1 : i < k \le j\} (\bar{X}_{(i:j), Z_k} - \bar{X}_T)^2, \\
&A = \prod_{\{(i:j) = S_l \in \rho\}} \frac{\Gamma(a + \#\{Z_k = 0 : i < k \le j\})\Gamma(b + \#\{Z_k = 1 : i < k \le j\})}{\Gamma(a + b + j - i)}.
\end{aligned}
\tag{2.15}
$$

Next, we integrate out (μ_0, w, σ) in (2.14) using priors (2.7), (2.8) and (2.9):

$$
f(\mathbf{X}, \mathbf{Z}|\rho) \tag{2.16}
$$

$$
\propto \frac{1}{w_0} \int_0^{w_0} \int \sigma^{-2} \int f(\mathbf{X}, \mathbf{Z}|\rho, \mu_0, w, \sigma)\, d\mu_0\, d\bigl(\sigma^2\bigr)\, dw \tag{2.17}
$$

$$
\propto A \int_0^{w_0} \frac{w^{(N-1)/2}}{[W + wB]^{(T-1)/2}}\, dw. \tag{2.18}
$$

To model multiple histone marks, $\mathbf{X}_1, \ldots, \mathbf{X}_M$ are independent vectors given the same block structure ρ. As has been calculated in (2.16), for each m,

$$
f(\mathbf{X}_m, \mathbf{Z}_m|\rho) \propto A_m \int_0^{w_0} \frac{w^{(N-1)/2}}{[W_m + wB_m]^{(T_m-1)/2}}\, dw, \tag{2.19}
$$

where a_m, b_m, W_m, B_m T_m and A_m are values for the mth sequence as a, b, W, B, T and A defined above. $\mathbf{Z}_m$ are indicators determined by $\mathbf{X}_m$ and $Z_{k,m}$ is the kth element in $\mathbf{Z}_m$. Combining (2.13) and (2.19), we have

$$f(\rho|\mathbb{X},\mathbb{Z}) \propto \frac{1}{p_0}\int_0^{p_0} p^{N-1}(1-p)^{n-N}\,dp \times \prod_{m=1}^{M} A_m \times \prod_{m=1}^{M}\int_0^{w_0} \frac{w^{(N-1)/2}}{[W_m + wB_m]^{(T_m-1)/2}}\,dw. \tag{2.20}$$

Although an exact implementation of this model is tractable, the calculations are $O(n^3)$. It is prohibitive to evaluate the posterior probability when n is large. We have implemented an MCMC approximation that greatly facilitates the estimation.

2.2. *MCMC algorithm for BCP model inference.* Following Barry and Hartigan (1993), for a partition ρ induced by $\mathbf{U} = (U_1, \ldots, U_n)$, where $U_i = 1$ indicates a change point at position $i+1$, the odds ratio for the conditional probability of a change point at the position $i+1$ is as follows:

$$\frac{P(U_i = 1|\mathbb{X},\mathbb{Z},U_j, j\neq i)}{P(U_i = 0|\mathbb{X},\mathbb{Z},U_j, j\neq i)} = \frac{\int_0^{p_0} p^N(1-p)^{n-N-1}\,dp \times \prod_{m=1}^{M} A_m^1 \int_0^{w_0} \frac{w^{N/2}}{[W_m^1+wB_m^1]^{(T_m-1)/2}}\,dw}{\int_0^{p_0} p^{N-1}(1-p)^{n-N}\,dp \times \prod_{m=1}^{M} A_m^0 \int_0^{w_0} \frac{w^{(N-1)/2}}{[W_m^0+wB_m^0]^{(T_m-1)/2}}\,dw},$$

where W_m^0, B_m^0, W_m^1 and B_m^1 are the within and between block sums of squares obtained for the mth sequence when $U_i = 0$ and $U_i = 1$ respectively, and A_m^0 and A_m^1 are the values of (2.15) obtained for the mth sequence when $U_i = 0$ and $U_i = 1$ respectively. The result is a direct consequence of (2.20).

We then approximate these integrals by an incomplete beta function as follows:

$$\begin{aligned}\frac{P(U_i = 1|\mathbb{X},\mathbb{Z},U_j, j\neq i)}{P(U_i = 0|\mathbb{X},\mathbb{Z},U_j, j\neq i)} &= \prod_{m=1}^{M}\left(\left(\frac{W_m^1}{B_m^1}\right)^{1/2}\left(\frac{W_m^0}{W_m^1}\right)^{(T_m-N-2)/2}\left(\frac{B_m^0}{B_m^1}\right)^{(N+1)/2}\right)\\ &\quad\times \frac{\prod_{m=1}^{M}\int_0^{\frac{B_m^1 w_0/W_m^1}{1+B_m^1 w_0/W_m^1}} x^{(N+2)/2}(1-x)^{(T_m-N-3)/2}\,dx}{\prod_{m=1}^{M}\int_0^{\frac{B_m^0 w_0/W_m^0}{1+B_m^0 w_0/W_m^0}} x^{(N+1)/2}(1-x)^{(T_m-N-2)/2}\,dx}\\ &\quad\times \frac{\int_0^{p_0} p^N(1-p)^{n-N-1}\,dp \times \prod_{m=1}^{M} A_m^1}{\int_0^{p_0} p^{N-1}(1-p)^{n-N}\,dp \times \prod_{m=1}^{M} A_m^0}.\end{aligned}$$

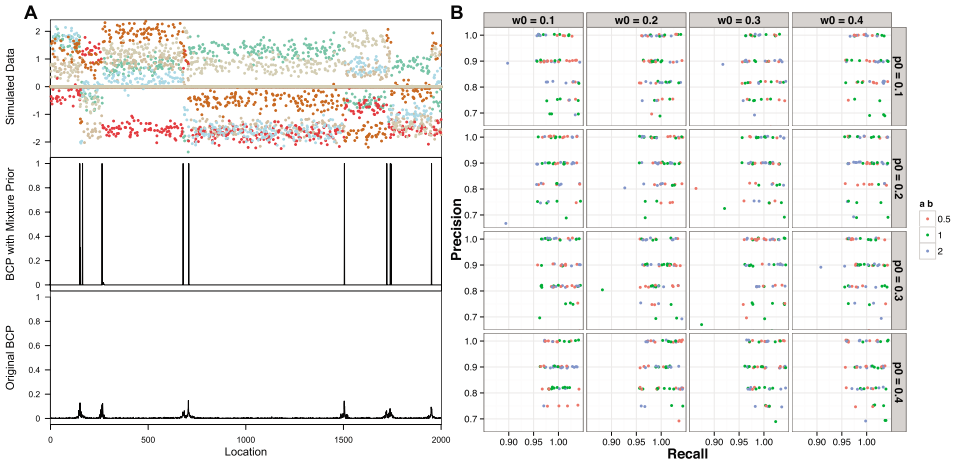

FIG. 1. *Simulation results.* (A) *One example of simulated datasets with posterior probabilities inferred from the proposed BCP model with* $p_0 = 0.1$, $w_0 = 0.1$, $a_m = b_m = 0.5$ *and from the original BCP model using function bcp() in R package bcp.* (B) *Jitter plot for precision and recall rates of BCP model with* 48 *different sets of hyperparameters on* 20 *simulated datasets.*

We initialize U_i to 0 for all $i < n$, with $U_n = 1$. Then we update U_i by passes through data. 500 passes were used in block identification.

3. Simulation studies. First we used simulated data to study the performance of the proposed method. The simulation assumed that there were 10 blocks and six histone modification marks were observed at each one of the 2000 locations in the genome. The lengths of the 10 blocks were ranging from 10 to 1500. (In simulation 1 shown in Figure 1, the lengths are 152, 10, 102, 416, 27, 799, 217, 22, 206 and 49.) We use $X_{(i:j),m}$ to denote the observed signal within a block from $(i+1)$th to jth location for the mth mark. We assumed that each component of the $X_{(i:j),m}$ followed a mixture distribution of $0.2 * N(\mu_{(i:j),m}, 1) + 0.8 * \delta$ where $\mu_{(i:j),m}$ was a random draw from $U(-2, 2)$. These settings are based on the empirical observation that for a specific histone mark, on average, $\sim$20% of the genome display binding peaks with the intensities ranging from -2 to 2 for the normalized data. To apply our method, we need to specify the values of the hyperparameters p, w, a_m and b_m. In the simulation we investigated the sensitivity of the results to the specifications of these parameter values by considering a range of values, with $p = (0.1, 0.2, 0.3, 0.4)$, $w = (0.1, 0.2, 0.3, 0.4)$, and $(a_m, b_m) = \{(1, 1), (2, 2), (0.5, 0.5)\}$. As a result, we considered a total of 48 specifications for (p_0, w_0, a_m, b_m). We simulated 20 datasets. For each simulated dataset, we ran 48 MCMC chains with each chain using one of the 48 different hyperparameters described above. Change points were inferred to be those locations in the genome that had a posterior probability larger than 0.8 (the results were similar under different cutoff values).

TABLE 1
Overview of modENCODE data that were used in this study

modENCODE experiment	Method	Cell line or tissue type	Sample
Genomic distributions of histone modifications	ChIP-chip/ChIP-seq	S2-DRSC, ML-DmBG3-c2	H3K18ac, H3K23ac, H3K27Ac, H3K27Me3, H3K36me1, H3K36me3, H3K4Me3, H3K4me1, H3K4me2, H3K79Me2, H3K79Me1, H3K9ac, H3K9me2, H3K9me3, H4AcTetra, H4K16ac, H4K5ac, H4K8ac
Transcriptional profiling of Drosophila cell lines	RNA-seq	S2-DRSC	
Developmental stage timecourse transcriptional profiling	RNA-seq	Embryo 10–12 h, white pre-pupae 24 h, larvae L1, adult female eclosure 1d	

We then checked the precision and recall rates based on the true and inferred change points from the simulated data. The precision rate is defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and the recall rate is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, where TP is the number of true positives (predicted block boundaries that are true), FP is the number of false positives (predicted boundaries that are not true), and FN is the number of false negatives (undiscovered true block boundaries). In our assessment, if the inferred change point was 3 units or less from one of the true change points, this inference was considered a true positive. As shown in Figure 1(B), the overall posteriors are insensitive to the specified values of the hyperparameters p_0, a_m, b_m. Simulation studies also show that the proposed method is capable of identifying large blocks expanded over 1000 positions as well as small blocks of size around 10 (Figure 1). Moreover, the ability of identifying zero-inflated blocks is significantly boosted by the introduction of the mixture prior (Figure 1).

4. Application to modENCODE epigenome data. All data used in this analysis were generated by the modENCODE project (Table 1). Specifically, we used the preprocessed enrichment score of 18 histone marks in S2 cells from the study "Genomic Distributions of Histone Modifications"; the S2 cell transcriptome data came from the study "Paired End RNA-Seq of *Drosophila* Cell Lines"; the transcriptome data for 9 different developmental stages were drawn from the study "Developmental Stage Timecourse Transcriptional Profiling with RNA-Seq". To identify and characterize blocks from histone marks, we divided the *Drosophila melanogaster* genome into 1000-bp bins and calculated the enrichment level for

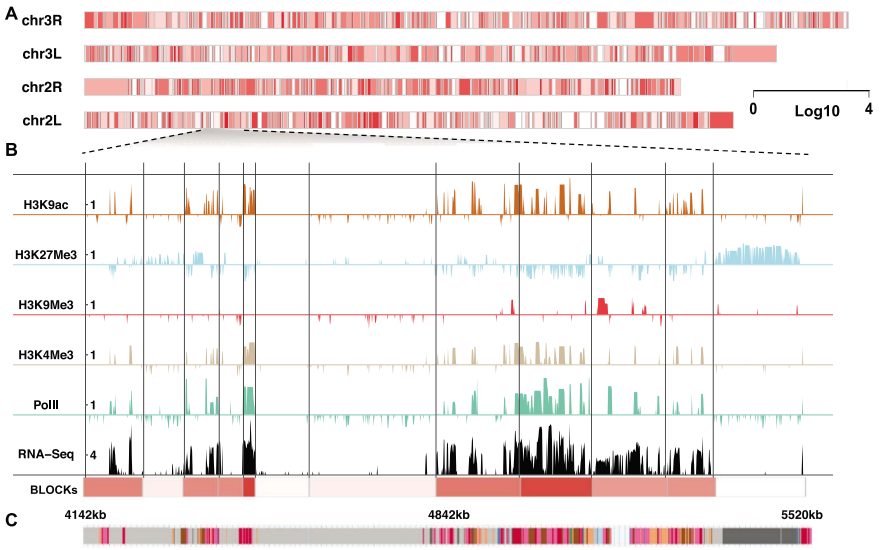

FIG. 2. *BLOCKs inferred from multiple histone marks in the Drosophila melanogaster S2 cell.* (A) *An overview of the BLOCKs in S2 cells with average transcriptional levels shown in the gradient.* (B) *An example of BLOCK characterization at a specific locus on chromosome 2L. BLOCK boundaries are shown as solid black lines. The enrichment levels of several chromatin signatures are shown at 1 kb resolution, including transcription activation marks H3K4m3 and H3K9ac, and transcription repression marks H3K9me3 and H3K27Me3. PolII and RNA-seq counts at* log10 *scale are shown as a reference of the transcription activity.* (C) *"Chromatin states" annotation from Kharchenko et al.* (2011).

each bin by averaging the log2 intensity values of each mark. The transcription level (in the S2 cell and different development stages) was calculated by averaging read counts from replicates.

4.1. *Identification of chromatin blocks based on histone modifications.* We applied the proposed method to 18 histone methylation and acetylation marks in S2 cells. Change points with posterior probability greater than 0.75 were defined as block boundaries. Because chromosome X is distinguished by the high-level enrichment of H4K16ac and H3K36me3 from other chromosomes [Kharchenko et al. (2011)], we applied our model to autosomes only.

A total of 994 blocks were inferred from chromosomes 2L, 2R, 3L and 3R, with 90% of the blocks ranging in size from 21 kb to 247 kb, with a median of 70 kb [called as BLOCKs, see supplementary tables in Chen, Lin and Zhao (2016)]. We observed that BLOCKs captured the combinatorial pattern of histone modifications and reflected local transcriptional activities. We use chr2L:4142–5520 kb as an example to illustrate this (Figure 2). For simplicity, we only show the enrichment levels of several chromatin signatures, including transcription activation marks H3K4m3 and H3K9ac, and transcription repression marks H3K9me3 and

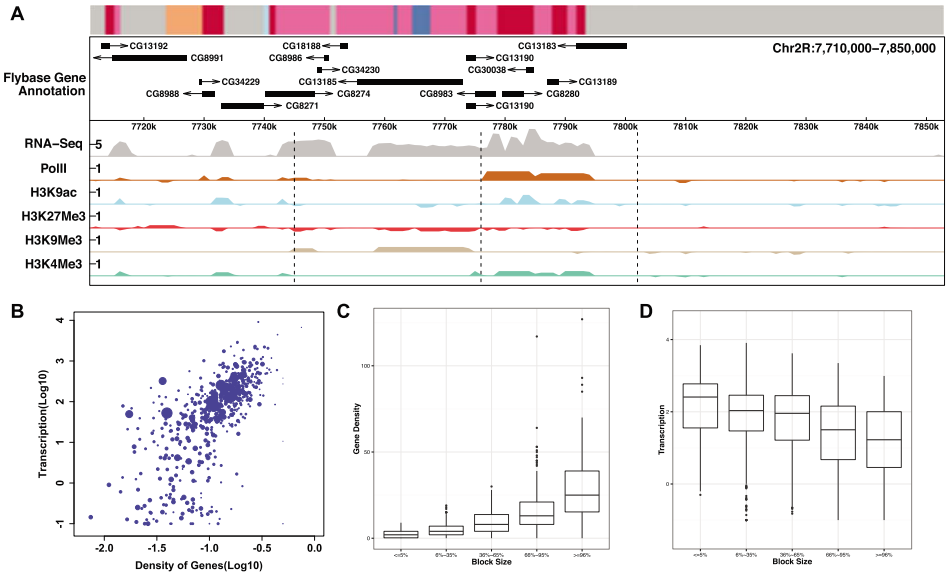

FIG. 3. *BLOCKs characterization*. (A) *A locus on chromosome* 2*R with BLOCKs displaying diverse sizes, gene density and transcription activity* [*corresponded "chromatin states" annotation from Kharchenko et al.* (2011) *shown on the top*]. (B) *Transcription activity vs. gene density with block size shown in the gradient*. (C) *Boxplot for gene density on five block size quantiles*. (D) *Boxplot for transcription activity on five block size quantiles*.

H3K27me3 (see Figure 4 for an example of all marks). PolII enrichment and RNA-seq counts at log10 scale are shown as a reference of transcriptional activity. Compared with "chromatin states" annotation for nonoverlapping 200 bp windows in the genome [Kharchenko et al. (2011)] [Figure 2(C)], BLOCKs depict the genome as local domains at a larger scale. We divided BLOCKs into five quantiles based on their sizes: $\leq 5\%$, $6\% \sim 35\%$, $36\% \sim 65\%$, $66\% \sim 95\%$, $\geq 96\%$ and looked into the transcription activity distributions for each group [Figure 3(E)]. Transcription activities do not show a systematic bias as a function of block size.

4.2. *BLOCK boundaries are potentially physical domain boundaries*. A recent published high-resolution chromosomal contact map on *Drosophila* embryonic nuclei [Sexton et al. (2012)] shows that the entire genome is linearly partitioned into well-demarcated physical domains. We therefore studied the link between physical domains and BLOCKs inferred from histone marks. A total of 994 physical domains were identified from *Drosophila* embryonic nuclei [Sexton et al. (2012)] chromosomes 2L, 2R, 3L and 3R with the sizes ranging from 10 kb to 823 kb and a median of 60 kb. We observed strong association between physical domains and BLOCK boundaries. For example, 36% of BLOCK boundaries are within 10 kb of physical domain boundaries, whereas this proportion never exceeds 26% in 1000 randomized block partitions, and 56% of BLOCK boundaries

are within 20 kb of physical domain boundaries, whereas this proportion never exceeds 42% in 1000 randomized block partitions.

In Sexton et al. (2012), the authors characterized physical domains into four epigenetic classes based on the enrichment of epigenetic marks. Out of the four classes, transcriptional "Active" domains are associated with H3K4me3, H3K36me3 and hyperacetylation, "PcG" domains are associated with the mark H3K27me3, "HP1/Centromere" class is associated with HP1, and "Null" domains are not enriched for any available marks. We explored whether BLOCKs can be aligned to the classification in Sexton et al. (2012). We assigned the four classes to BLOCKs based on enrichment of H3K4me3, H3K27me3 and HP1a. Specifically, BLOCKs with average intensities of HP1a greater than 1 and coverage greater than 10% are classified as "HP1/Centromere" domains, BLOCKs with average intensities of H3K27me3 greater than 0.5 and coverage greater than 25% are classified as "PcG" domains, BLOCKs with average intensities of H3K4me3 greater than 1 and coverage greater than 25% are classified as "Active" domains, and all the remaining ones are characterized as "Null" domains. Figure 4 shows the alignment between BLOCKs and physical domains with epigenetic classes. In 93,835 genomic bins annotated by both BLOCKs and chromHMM, 62,987 have the same assignment, leading to a jaccard index of 0.5. The high concordance between BLOCKs and physical domains suggests that BLOCKs bridge the link between epigenetic domains with topological domains. The difference may be introduced by techniques, data quality and cell types used in these two studies.

Another indirect evidence for BLOCKs as physical domains is the consistency with replication timing. Replication timing refers to the order in which segments of DNA along the length of a chromosome are duplicated. Since the packaging of DNA with proteins into chromatin takes place immediately after the DNA is duplicated, replication timing reflects the order of assembly of chromatin. Recent studies suggest that late-replicating regions generically define not only a repressed but also a physically segregated nuclear compartment. Thus, replication timing is a manifestation of spatial organization of the chromosome. To investigate the association of BLOCKs with replication timing, we compared BLOCKs with the meta peaks of replication origins (10 kb to 285 kb) from cell lines BG3, Kc and S2 analyzed by the modENCODE project. We observed that 69% of meta peaks are within 20 kb of BLOCK boundaries. This statistic agrees with physical domains well since we observed that 60% of meta peaks within 20 kb of physical boundaries were characterized in Sexton et al. (2012).

4.3. *Functional relevance of BLOCKS.* To investigate whether BLOCKs represent domains of functional importance, we performed three different analyses. First, we checked whether genes within each BLOCK tended to be coregulated using transcriptome in L1 larvae and 10–12 h embryo measured by RNA-seq. A total of 11,376 FlyBase genes were used in our analysis. When a gene had multiple isoforms, the longest one was used. We defined the following rules to describe the

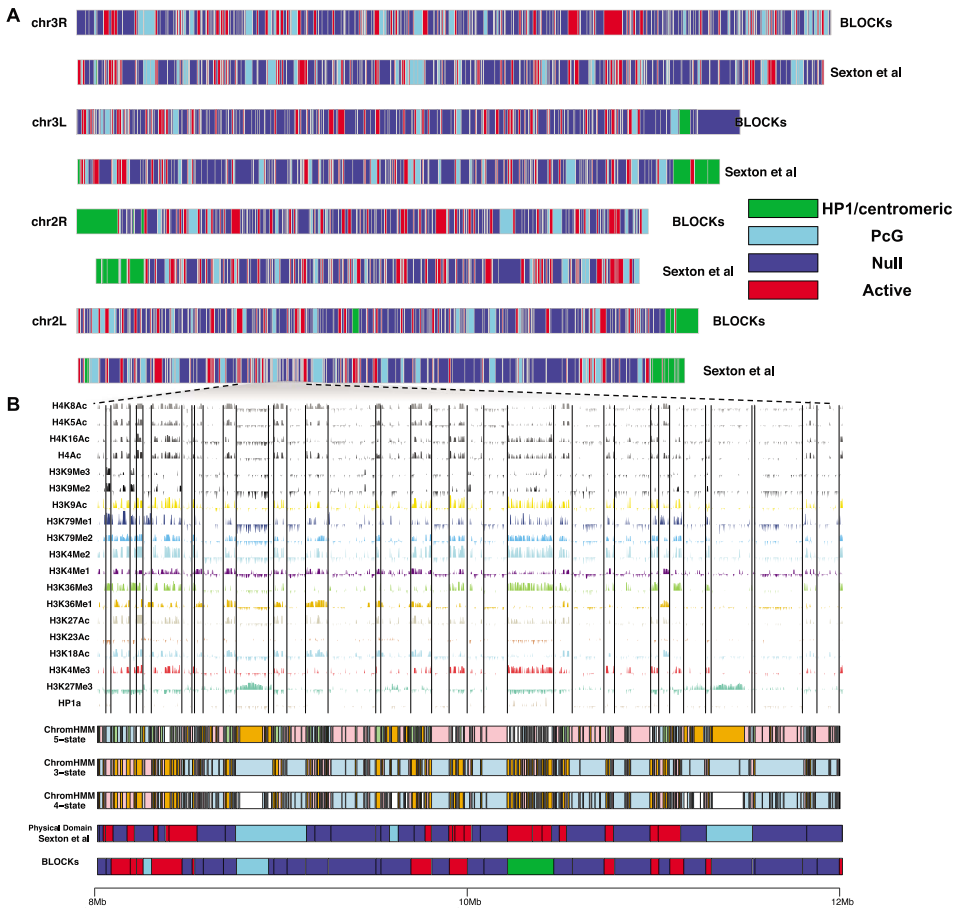

FIG. 4. (A) *The alignment of BLOCKs* (*S2 cells*) *with physical domains in Sexton et al.* (2012) (*embryonic nuclei cells*). (B) *A comparison of ChromHMM, BLOCKs and physical domains at a locus on chromosome* 2*L* (8 *Mb*–12 *Mb*). *BLOCK boundaries are shown as vertical gray lines.*

status change of each gene between L1 larvae stage (and 10–12 h embryo) and S2 cell: genes whose expression increased by more than 2 fold but were not below 10 were categorized as "up-regulated"; those with fold change less than 0.5 but the expression levels were not below 10 as "down-regulated"; and others as "no-change". To examine whether each BLOCK is enriched for genes with a specific status, we used the proportion of blocks that the dominant status accounted for at least 50% of the genes within a block as a test statistic. We observed the percentage of BLOCKs where the dominant status accounted for more than 50% of the genes was 71.8% and 67.6% for L1 larvae and 10–12 h embryo, respectively, with 55.4% of the BLOCKs overlapped between the two comparisons. These observed statistics reach statistical significance when testing against randomly permutated

blocks. For physical domains in Sexton et al. (2012), we observed 68% and 65.8% with dominant coregulation for L1 larvae and 10–12 h embryo, respectively.

Second, we asked whether genes within each BLOCK tended to have similar biological functions. We tested for the enrichment of Gene Ontology (GO) categories within each BLOCK using a hypergeometric test with Bonferroni correction. 51.2% (412 out of 805 BLOCKs with more than 2 genes) were enriched for at least one GO category using a 0.05 cutof, and 1172 GO categories in total are enriched [see supplementary tables in Chen, Lin and Zhao (2016)]. The observed numbers of GO enriched BLOCKs and enriched GO categories were both significantly higher than those from permutated blocks. We further asked which biological processes or functions involve genes that are significantly linearly juxtaposed. We found 86.4% (108/125) of chromatin assembly or disassembly genes (GO:0006333) for *Drosophila* were juxtaposed within a BLOCK located on chr2L: 21344–21579 kb, with a striking p-value of 3.3×10^{-235}. Genes in chitin-based cuticle development (GO:0040003), body morphogenesis (GO:0010171) and proteinaceous extracellular matrix (GO:0005578) were found significantly clustered with over 70% of genes in one BLOCK sharing the same function.

Third, we reasoned if BLOCKs reflected coordinated regulation of genes with relevant biological functions, we would expect that BLOCKs enriched in developmentally specific GO categories would have large variation across different developmental stages, while BLOCKs enriched in "house-keeping" GO categories would display limited fluctuations. We ranked the BLOCKs based on their standard deviation of transcription level across 9 different developmental stages [see supplementary tables in Chen, Lin and Zhao (2016)]. BLOCKs with the top 20% largest deviations and 20% smallest deviations were checked for their GO enrichment respectively, and then were listed in Tables S2 and S3 by their order of statistical significance. Notably, in BLOCKs displaying the most striking changes across different developmental stages, we found GO categories associated with developmental-specific biological processes or functions, such as heart development, structural constituent of chitin-based cuticle, positive regulation of muscle organ development and midgut development, among others. Moreover, metabolism-related functions, such as serine-type endopeptidase activity, peptidyl-dipeptidase activity, etc., display turnover across developmental transcriptomes and are among the top of our list. GO categories associated with "house-keeping" functions, like transferase activity, aminoacylase activity, chromatin assembly and insulin receptor binding, showed limited fluctuations through development. This result provides further evidence on the role of BLOCKs in coordinated regulation.

4.4. *Comparison with ChromHMM.* In this subsection we compare the results from our method with those from a popular HMM-based method, ChromHMM.

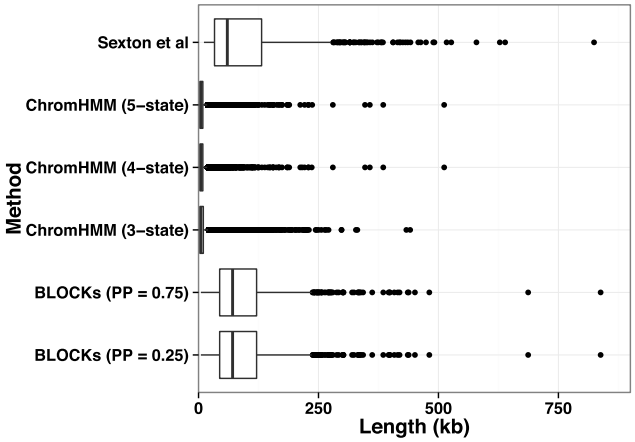

FIG. 5. *Boxplot for the sizes of segments identified using different methods*: *physical domains in embryonic nuclei identified using High-C data* [*Sexton et al.* (2012)], *ChromHMM with* 5, 4 *and* 3 *hidden states and BLOCKs with posterior probability greater than* 0.75 *and* 0.25.

We applied ChromHMM to the same dataset (18 histone modification, 1 kb bins, S2 cell). The data were binarized to fit ChromHMM's requirement of input. More specifically, all intervals with intensities greater than 0 are set to 1 and the remaining are set to 0. To obtain blocks at coarse levels, we explored ChromHMM models by varying the prespecified number of hidden states (from 3 to 18). We observed that a smaller number of hidden states tended to produce blocks with larger sizes. Here we report ChromHMM models with the number of hidden states from 3 to 5. The ChromHMM model with 3 hidden states generates 12,517 segments, the model with 4 hidden states generates 9157 segments, and the model with 5 hidden states generates 12,444 segments. For each ChromHMM model, the sizes of segments range from 2 kb (5% quantile) to 26 kb (95% quantile) and a median of 5 kb. The distributions of sizes of segments from ChromHMM models and BLOCKs are visualized in Figure 5. Therefore, our model has advantages over the HMM models in characterization of histone modification patterns at coarse levels.

4.5. *How robust is the result?* The BCP model used in this paper assumes that different histone marks are independent. However, some histone marks, such as H3K4me3 and H3K4me2, are highly correlated with each other. Moreover, it is known that there exists redundancy and exclusivity between the active and repressive marks. To further explore how the input histone marks will affect the result, we performed the change point analysis with the input of 4 marks, 7 marks and 10 marks, respectively. The marks for each model were selected based on their correlation across the entire genome. As shown in Fig-

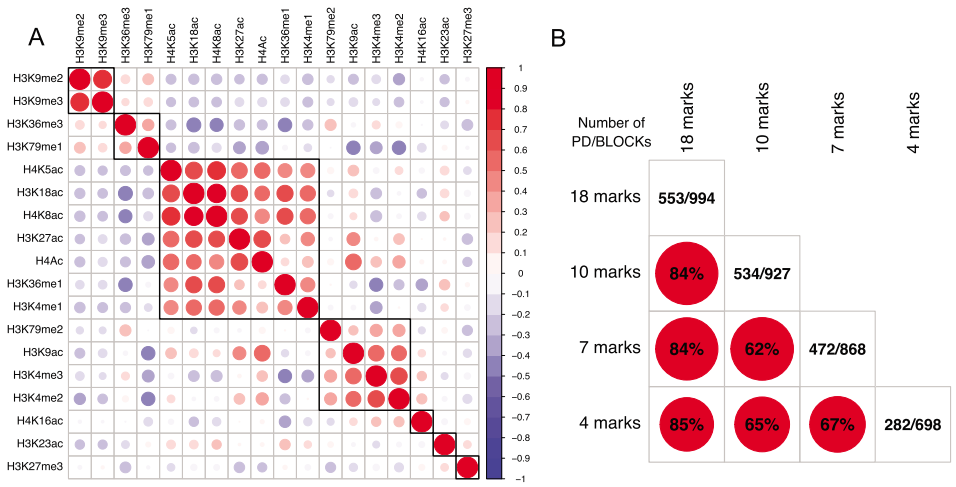

FIG. 6. (A) *Genome-wide correlation plot for* 18 *histone marks in S2 cells. The marks are ordered based on the result of hierarchical clustering.* (B) *Comparison of models with different input histone marks. Each of the off-diagonal elements is the percentage of boundaries* (*within* 20 *kb*) *shared by any pair of the models. The diagonal element is the number of boundaries shared with physical boundaries in Sexton et al.* (2012) (*short as PD*)*/the number of segments detected for each model.*

ure 6(A), there are mainly 7 groups of marks based on their correlation patterns: the first group consists of H3K9me2 and H3K9me3; the second group is featured by H3K36me3 and H3K79me1; the third group consists of H4K5ac, H3K18ac, H4K8ac, H3K27ac, H4Ac, H3K36me1 and H3K4me1; the fourth group is featured by H3K79me2, H3K9ac, H3K4me3 and H3K4me2; whereas three separate groups are formed by H4K16ac, H3K23ac and H3K27me3, respectively. For the 7 marks model, we selected one mark from each of the 7 groups with the input marks as H3K18ac, H3K23ac, H3K27Me3, H3K36me3, H3K4Me3, H3K9me2 and H4K16ac. For the 10 marks model, we further introduced H4, H3K79Me2 and H3K9ac into the 7 marks model. For the 4 marks model, we excluded H3K18ac, H3K36me3 and H4K16ac from the 7 marks model. The 4 marks, 7 marks and 10 marks models identified 698, 868 and 927 blocks, respectively. We observed high consistency between these results and reported BLOCKs obtained with 18 marks, for example, 84% of boundaries from the 10 marks model are within 20 kb of BLOCK boundaries and 84% of boundaries from the 7 marks model are within 20 kb of BLOCK boundaries [see Figure 6(B) for other comparisons].

To investigate how the posterior probability cutoff would affect the characterization of BLOCKs, we varied the threshold and checked the distribution of the sizes. The results were rather stable under different cutoff values. When the cutoff value was set as low as 0.25, only 2 new boundaries were added, leading to a total of 996 blocks.

5. Discussion.

5.1. *Methodological comparisons*. Our BCP model was developed with a different purpose compared to existing methods for analyzing a combinatorial pattern of histone marks. For example, ChromaSig [Hon, Ren and Wang (2008)] was designed to uncover potential regulatory elements through searching for genome-wide frequently occurring chromatin signatures. Spatial clustering [Jaschek and Tanay (2009)] identified novel patterns of local co-occurrence among histone modifications by imposing a spatial K-clustering solution on HMM. Segway [Hoffman et al. (2012)] based on Dynamic Bayesian Networks achieved a breakthrough in precision and resolution in finding known elements and handling of missing data compared to HMM-based approaches. The most recent method of this kind, ChAT [Wang, Lunyak and Jordan (2012)], extends the capabilities of chromatin signatures characterization through an inherent statistical criterion for classification. All these methods tried to detect chromatin signatures associated with a variety of small functional elements. To the best of our knowledge, our model is the first effort to examine histone marks at coarse scales, although no explicit constraint has been put on block size. By separately modeling zero and nonzero signals, our model is able to capture the local enrichment patterns of different sizes implicitly, which is superior than the existing *ad hoc* merging strategy [Wang, Lunyak and Jordan (2012)].

BCP differs substantially from several previously described studies to subdivide the genome at "domain-level". de Wit et al. (2008) reported a study to identify nested chromatin domain structure through a statistical test of each chromatin component. Their chromatin domains are specific for each component or factor, whereas our approach captures domain with a combinatorial pattern of multiple factors. Thurman et al. (2007) used a simple two-state HMM to segment the ENCODE regions into active and repressed domains based on multiple tracks of functional genomic data, including activating and repressive histone modifications, RNA output and DNA replication timing. By using wavelet smoothing, their method focuses on a single scale at a time [Lian et al. (2008)]. In contrast, our analysis focuses on histone modifications only and simultaneously captures enrichment patterns over different scales. BCP is most similar to a four-state CPM model proposed to characterize chromatin accessibility based on tiled microarray DNaseI sensitivity data only [Lian et al. (2008)]. Both methods formulate the segmentation of genome into a change point detection problem. However, these two methods differ in several respects. First, CPM is still a hidden-state model with transition probabilities imposed on segments other than equal-sized bins in HMM, whereas BCP is hidden-state free with emphasis on local patterns. Second, a four-state CPM model was developed to interpret a single track DNaseI array data, while our method was an examination based on multivariate histone modification data. Third, CPM models the DNaseI signal as a continuous mixture of Gaussian at each state, whereas we model a histone binding signal with a zero-inflated Gaussian mixture due to spatial sparsity of binding events.

5.2. *Summary and future directions*. In this paper we have developed a novel multivariate BCP model to partition a genome into contiguous blocks based on histone modifications. It could be extended to analyze chip-sequencing data or applied to other studies with partitioning zero-inflated multiple observation tracks as a task. Our model presents a new approach to examining combinatorial histone marks. Histone marks are not only signatures for functional elements [Kharchenko et al. (2011), Ernst and Kellis (2010)], our results from the *D. melanogaster* S2 cell genome suggest that they are also roadmaps for chromatin organization at coarse scales.

It is worthwhile to further investigate whether BLOCKs and topological domains are substantively different, or if BLOCKs merely redescribe topological domains based on histone marks. Besides the difference introduced by techniques, data quality and cell types, we believe two other possible reasons for imperfect alignment between BLOCKs and physical domains are: (1) the partition is not saturated based on the current profile of histone modifications; (2) the equal weight assigned to different histone modifications in the partition limit the identification of finer domains (a drawback of all current approaches).

It has become increasingly clear that functionally related genes are often located next to one another in the linear genome [Sproul, Gilbert and Bickmore (2005)], resembling DNA operon in bacteria [Chen et al. (2012), Keene (2007)]. This proximity is essential for coordinated gene regulation. Genome-wide expression analysis has identified many clusters of co-expressed genes during *Drosophila* development [Lee and Sonnhammer (2003), Yi, Sze and Thon (2007)], such as the *hox* gene clusters [Duboule (2007)]. One mechanism for this coordinated regulation is that these genes are organized into a chromatin domain that acts as a regulatory unit by the epigenetic mechanism [Kosak and Groudine (2004), Sproul, Gilbert and Bickmore (2005)]. Several such chromatin domains have already been characterized [Kosak and Groudine (2004), Tolhuis et al. (2006), Pickersgill et al. (2006), Orlando and Paro (1993)]. In this study we illustrated the widespread existence of these chromatin domains as BLOCKs that were identified by histone marks.

Last but not least, although we have shown that a substantial portion of BLOCKs can potentially act as regulatory units, this is likely still an underestimate. First, our BLOCKs were identified based on combinatorial patterns of 18 histone marks from the S2 epigenome. We do not know in totality how many histone marks are sufficient to saturate the segmentation. It is likely that more markers, including potentially undiscovered ones, will be needed to get a complete view of the epigenetic landscape. Over 100 histone marks have been discovered with a lot of exclusivity and correlation. Future studies addressing relationships among histone marks will give us more insight into this open question. It is also important to develop block identification methods that can accommodate the dependency structure among marks. Second, when evaluating expression of genes within an individual BLOCK, we used developmental transcriptome from *Drosophila* tissues other than S2 cells, which only presented a weighted average of varying BLOCKs

across different cell types within each developmental stage. In reality, each type of cells is likely to have its distinct pattern of BLOCKs. Third, plasticity in chromosomal modifications has been shown in several reports [Riddle et al. (2011), Eaton et al. (2011), modENCODE Consortium (2010)]. Thus, we would expect BLOCKs are dynamic structures and the percentage of BLOCKs with tendency of coregulation might be even higher if taking into account this plasticity. This conjecture could be tested when more histone marks data across development stages are available. Fourth, with incomplete and inaccurate knowledge on gene functions in the GO database (as well as others) [Khatri, Sirota and Butte (2012)], it is likely many BLOCKs with functional relevance may not stand out just because supporting information does not exist yet. Finally, coordinated regulation is a complex process accomplished by miRNA, transcript factors and other regulatory elements with feedback effect on chromatin organization. Further analysis on binding sites of regulatory elements and their interplay with genes within BLOCKs will shed more lights on understanding the underlying mechanism.

Acknowledgments. We thank the reviewers for their constructive comments and Chao Gao for discussion. The authors thank Yale University Biomedical High Performance Computing Center for computing resources, and NIH grant RR19895 and RR029676-01, which funded the instrumentation.

SUPPLEMENTARY MATERIAL

Supplement A: modENCODEhistone (DOI: 10.1214/16-AOAS905SUPPA; .pdf). Number of enriched regions of 46 histone marks and nonhistone chromosomal proteins from the modENCODE project.

Supplement B: BLOCKs (DOI: 10.1214/16-AOAS905SUPPB; .zip). BLOCKs identified by BCP in S2 cells using posterior probability cutoff 0.75.

Supplement C: EnrichedGenes (DOI: 10.1214/16-AOAS905SUPPC; .zip). Gene lists in GO enriched BLOCKs in S2 cell.

Supplement D: LargestVarianceBLOCKs (DOI: 10.1214/16-AOAS905SUPPD; .zip). BLOCKs with the top 20% largest deviations in the transcription across 9 different developmental stages.

Supplement E: SmallestVarianceBLOCKs (DOI: 10.1214/16-AOAS905SUPPE; .zip). BLOCKs with the top 20% smallest deviations in the transcription across 9 different developmental stages.

REFERENCES

ALLIS, D. (2007). *Epigenetics*. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY.

BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20** 260–279. MR1150343

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

CHEN, M., LIN, H. and ZHAO, H. (2016). Supplement to "Change point analysis of histone modifications reveals epigenetic blocks linking to physical domains." DOI:10.1214/16-AOAS905SUPPA, DOI:10.1214/16-AOAS905SUPPB, DOI:10.1214/16-AOAS905SUPPC, DOI:10.1214/16-AOAS905SUPPD, DOI:10.1214/16-AOAS905SUPPE.

CHEN, D., ZHENG, W., LIN, A., UYHAZI, K., ZHAO, H. and LIN, H. (2012). Pumilio 1 suppresses multiple activators of p53 to safeguard spermatogenesis. *Current Biology* **22** 420–425.

DE WIT, E., BRAUNSCHWEIG, U., GREIL, F., BUSSEMAKER, H. J. and VAN STEENSEL, B. (2008). Global chromatin domain organization of the Drosophila genome. *PLoS Genet.* **4** e1000045.

DUBOULE, D. (2007). The rise and fall of Hox gene clusters. *Development* **134** 2549–2560.

EATON, M. L., PRINZ, J. A., MACALPINE, H. K., TRETYAKOV, G., KHARCHENKO, P. V. et al. (2011). Chromatin signatures of the Drosophila replication program. *Genome Res.* **21** 164–174.

ERDMAN, C. and EMERSON, J. W. (2008). A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics* **24** 2143–2148.

ERNST, J. and KELLIS, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology* **28** 817–826.

FILION, G. J., VAN BEMMEL, J. G., BRAUNSCHWEIG, U., TALHOUT, W., KIND, J. et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143** 212–224.

HOFFMAN, M. M., BUSKE, O. J., WANG, J., WENG, Z., BILMES, J. A. et al. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9** 473–476.

HON, G., REN, B. and WANG, W. (2008). ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.* **4** e1000201, 16. MR2457133

JASCHEK, R. and TANAY, A. (2009). Spatial clustering of multivariate genomic and epigenomic information. *Research in Computational Molecular Biology* **5541** 170–183.

KEENE, J. D. (2007). RNA regulons: Coordination of post-transcriptional events. *Nat. Rev. Genet.* **8** 533–543.

KHARCHENKO, P. V., ALEKSEYENKO, A. A., SCHWARTZ, Y. B., MINODA, A., RIDDLE, N. C. et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471** 480–485.

KHATRI, P., SIROTA, M. and BUTTE, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **8** e1002375.

KOSAK, S. T. and GROUDINE, M. (2004). Gene order and dynamic domains. *Science* **306** 644–647.

LEE, J. M. and SONNHAMMER, E. L. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13** 875–882.

LIAN, H., THOMPSON, W. A., THURMAN, R., STAMATOYANNOPOULOS, J. A., NOBLE, W. S. et al. (2008). Automated mapping of large-scale chromatin structure in ENCODE. *Bioinformatics* **24(17)** 1911–1916.

MODENCODE CONSORTIUM (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330** 1787–1797.

ORLANDO, V. and PARO, R. (1993). Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell* **75** 1187–1198.

PICKERSGILL, H., KALVERDA, B., DE WIT, E., TALHOUT, W., FORNEROD, M. and VAN STEENSEL, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat. Genet.* **38** 1005–1014.

RIDDLE, N. C., MINODA, A., KHARCHENKO, P. V., ALEKSEYENKO, A. A., SCHWARTZ, Y. B. et al. (2011). Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res.* **21** 147–163.

SEXTON, T., YAFFE, E., KENIGSBERG, E., BANTIGNIES, F., LEBLANC, B., HOICHMAN, M. et al. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148** 1–15.

SPROUL, D., GILBERT, N. and BICKMORE, W. A. (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6** 775–781.

THURMAN, R. E., DAY, N., NOBLE, W. S. and STAMATOYANNOPOULOS, J. A. (2007). Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17** 917–927.

TOLHUIS, B., DE WIT, E., MUIJRERS, I., TEUNISSEN, H., TALHOUT, W., VAN STEENSEL, B. and VAN LOHUIZEN, M. (2006). Genome-wide profiling of PRC1 and PRC2 polycomb chromatin binding in Drosophila melanogaster. *Nat. Genet.* **38** 694–699.

WANG, J., LUNYAK, V. V. and JORDAN, I. K. (2012). Chromatin signature discovery via histone modification profile alignments. *Nucleic Acids Res.* **40** 10642–10656.

YI, G., SZE, S.-H. and THON, M. R. (2007). Identifying clusters of functionally related genes in genomes. *Bioinformatics* **23** 1053–1060.

M. CHEN
DEPARTMENT OF BIOSTATISTICS AND GENETICS
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27599
USA
E-MAIL: mengjie@email.unc.edu

H. LIN
YALE STEM CELL CENTER
YALE SCHOOL OF MEDICINE
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
USA
E-MAIL: haifan.lin@yale.edu

H. ZHAO
DEPARTMENT OF BIOSTATISTICS
YALE SCHOOL OF PUBLIC HEALTH
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
USA
E-MAIL: haifan.lin@yale.edu
hongyu.zhao@yale.edu