Published in final edited form as:

Stat Med. 2013 June 15; 32(13): 2292–2307. doi:10.1002/sim.5658.

Epigenetic change detection and pattern recognition via Bayesian hierarchical hidden Markov models

Xinlei Wang^a, Miao Zang^a, and Guanghua Xiao^{b,*†}

^aDepartment of Statistical Science, Southern Methodist University, Dallas, TX 75275, U.S.A.

^bDivision of Biostatistics, Department of Clinical Sciences, The University of Texas Southwestern Medical Center, Dallas, TX 75390, U.S.A.

Abstract

Epigenetics is the study of changes to the genome that can switch genes on or off and determine which proteins are transcribed without altering the DNA sequence. Recently, epigenetic changes have been linked to the development and progression of disease such as psychiatric disorders. High-throughput epigenetic experiments have enabled researchers to measure genome-wide epigenetic profiles and yield data consisting of intensity ratios of immunoprecipitation versus reference samples. The intensity ratios can provide a view of genomic regions where protein binding occur under one experimental condition and further allow us to detect epigenetic alterations through comparison between two different conditions. However, such experiments can be expensive, with only a few replicates available. Moreover, epigenetic data are often spatially correlated with high noise levels. In this paper, we develop a Bayesian hierarchical model, combined with hidden Markov processes with four states for modeling spatial dependence, to detect genomic sites with epigenetic changes from two-sample experiments with paired internal control. One attractive feature of the proposed method is that the four states of the hidden Markov process have well-defined biological meanings and allow us to directly call the change patterns based on the corresponding posterior probabilities. In contrast, none of existing methods can offer this advantage. In addition, the proposed method offers great power in statistical inference by spatial smoothing (via hidden Markov modeling) and information pooling (via hierarchical modeling). Both simulation studies and real data analysis in a cocaine addiction study illustrate the reliability and success of this method.

Keywords

Bayesian hierarchical modeling; epigenetics; epigenetic alteration; Gibbs sampler; hidden Markov model; MCMC; posterior samples; spatial dependence; spatial smoothing

1. Introduction

In biology, epigenetic changes are typically referred to as heritable changes in gene expression or cellular phenotype caused by mechanisms other than changes in the underlying DNA sequence. Recently, epigenetics has been a fast-growing field. This is because epigenetic changes are crucial for the development and differentiation of various cell types in an organism, as well as for normal cellular processes. They can also be responsible for some disease states. Table 1 from [1] provides a brief summary of the diseases caused by epigenetic changes. Detecting epigenetic changes would enable us to understand the roles of epigenetic mechanisms in both normal development and disease. The goal of this paper is to examine the existence of such changes and infer change patterns by analyzing epigenetic data that are often spatially correlated, with high noise levels and only a few replicates available.

The application that motivates this paper is a cocaine addiction study [2], in which binding activities of a transcription factor (TF) of interest, cyclic adenosine monophosphate (cAMP) response element binding protein (CREB; [3]), were measured by experiments using NimbleGen promoter arrays. The experiments were performed on both cocaine-treated and saline-treated mice (i.e., treatment vs. control) to detect cocaine-induced alterations in the TF binding. In these experiments, fresh nucleus accumbens (NAc) punches were processed for ChIP as described in [2]. The samples, both enriched and not enriched by the immunoprecipitation (IP) process, were amplified and labeled with Cy3 (reference) and Cy5 (IP enriched) and then hybridized to the same promoter arrays, with three biological replicates per condition. Each biological replicate was prepared by nucleus accumbens (a major brain reward region) punches pooled from eight mice under one condition to reduce the biological variability and then assayed using one array. See Figure 1 for an illustration of the experiment design (two samples with paired internal control). Also, Figure 2 compares it with two other popular designs in epigenetic studies. Design A compares the IP-enriched sample with normal DNA sample (reference) in each array, which is commonly used to detect TF binding sites. Design B compares IP-enriched samples between two different conditions in each array. It is commonly used to detect epigenetic changes between two conditions. Design C is the design used in our motivating example, involving two-sample experiments with paired internal control. This design allows us to detect not only epigenetic changes but also change patterns.

For illustrative purposes, in Figure 3, we plot several typical patterns of epigenetic alterations that could occur in the application. The horizontal axis represents the probe location in the promoter region of a gene under consideration; the vertical axis represents the corresponding TF binding strength. Black and red lines represent the true epigenetic profiles under the saline (control) condition and the cocaine (treatment) condition, respectively; black and red dots represent epigenetic data observed (truth plus error terms) under saline and cocaine conditions, respectively. Panel A shows a region with cocaine-blocked TF binding: the region is only bound by the TF in saline-treated mice (black line) but not in cocaine-treated mice (red line). Panel B shows a region with cocaine-induced TF binding: it is only bound by the TF in the cocaine condition but not in the saline condition. Panel C shows a region where the cocaine treatment changes the TF binding strength. In addition, we

plot two cases in which no epigenetic alteration occurs: panel D shows a region that is bound by the TF in both conditions, with the same binding strength; and panel E shows a region bound by the TF in neither condition. The goal of the study is to detect genomic locations (i.e., probes and genes) with cocaine-induced epigenetic alterations in the TF binding and to infer which types the changes belong to.

One feature associated with epigenetic data is that levels of binding strength, usually measured by the log ratio of the IP sample intensity and the reference sample intensity

(i.e., $\log \frac{\text{IP}}{\text{REF}}$), often show spatial patterns around the peaks (e.g., approximate triangle patterns as shown in Figure 3). Such data are often coupled with a very low signal-to-noise ratio. To improve efficiency in inference, authors have developed various approaches to incorporate the dependency among neighboring probes (e.g., [4–14]). Among them, hidden Markov models (HMMs) have been a popular choice for spatial modeling of epigenetic data (e.g., [4, 7, 15, 16]). However, most statistical methods previously developed for analysis of epigenetic data focus on identification of TF binding sites by comparing IP-enriched samples versus reference samples (see design A in Figure 2). Few of such methods can be directly applied to detect epigenetic changes and infer their patterns under design C. This might be somewhat strange, especially to readers who are not familiar with such genomic data, because one could naively treat the comparison between two different conditions as if it were the comparison between IP and reference samples. However, several issues arise from doing so. First, design C yields data with a more complicated structure than design A or B, where design A is used as a subdesign under one condition. Also, for design C, each array is only used for one condition, so no array is measured under two conditions simultaneously. Blindly using existing methods or software developed for design A or B with data having different structures could cause serious problems. Secondly, among the existing methods that do not exploit information of the paired data structure in design A or B, many of them make assumptions that are appropriate for enriched/non-enriched probes, which, however, are not suited for probes with/without epigenetic changes between the two conditions (e.g., [7, 11–14, 17]). For example, Mo and Liang [13, 14] proposed to use Ising models to account for spatial dependence. They assumed that the binding strength of all IP-enriched probes (i.e., those in binding regions) follows a common normal distribution and that of all non-IP-enriched probes follow another normal distribution. Similar assumptions are imposed in [11] and [15]. If we apply their methods to detect probes with epigenetic changes, then we have to assume that all the epigenetic changes have a common mean with a certain sign. This may not be realistic because epigenetic changes have various patterns, as shown in Figure 3, and obviously, they have different directions (positive or negative). Thirdly, even for existing methods that seem to fit in the needs of detecting epigenetic changes across two conditions (e.g., [4, 6, 9]), they cannot infer what types of changes occur.

In this paper, we develop a Bayesian hierarchical model to detect spatially dependent epigenetic changes in two-sample experiments with internal control, using high-throughput data generated from promoter arrays. Unlike the whole genome tiling arrays that provide unbiased information for all genomic regions, promoter arrays cover only gene regulatory promoter regions with a higher probe density. The distance between two consecutive probes

within the same promoter region is only 100–200 base pairs, whereas that between probes from different promoter regions (of two distinct genes) is relatively far away, typically at least several hundred kilobase pairs. Hence, it is reasonable to believe that the epigenetic data of different genes are spatially independent.

To account for the inherent dependence within the promoter region of each gene, our model relies on a hidden Markov structure. Under each condition, we use latent binary variables to indicate whether a probe k in the promoter region of a gene j is IP-enriched or not (say, S_{jk}^x and S_{jk}^y for the two conditions x and y , respectively). Here, an ‘enriched’ probe means TF binding activities occur at the site, whereas ‘non-enriched’ means no binding activities.

Instead of modeling the dependence of S_{jk}^x ’s and S_{jk}^y ’s among the different probes of gene j by two Markov processes (one for each condition) that evolve independently, we jointly

model the spatial dependence of $D_{jk} = (S_{jk}^x, S_{jk}^y)$ by one Markov process with four states (0,0), (0,1), (1,0), and (1,1), where 0 means not enriched and 1 enriched. Note that epigenetic changes often occur in some genomic regions rather than isolated locations of probes. Thus, using one Markov process with four states is a natural modeling choice to account for dependence in epigenetic alterations across conditions, as well as dependence in TF binding under each condition. Our model will allow epigenetic change detection and pattern recognition in parallel. This would provide meaningful information to identify the most interesting cases for potential experimental validation, as well as to facilitate biological interpretation.

We will adopt a fully Bayesian approach rather than a classical HMM approach. In the classical approach, estimation of the state variables is conditional on the maximum likelihood estimates of the model parameters. However, in a typical Bayesian approach, the parameters and state variables are jointly distributed random variables so that the posterior estimate of each appropriately reflects the uncertainty of the others. Another advantage is that, combined with hierarchical modeling, the Bayesian approach can naturally reflect the multilevel structure of the data in our motivating application (i.e., the probe, gene, and genome levels), and estimation through sampling from the joint posterior distribution would proceed as before, even though there are many more parameters involved compared with single-level data. By contrast, the classical approach often relies on an expectation–maximization (EM) algorithm for obtaining maximum likelihood estimates of the parameters. However, for such complex data, the EM algorithm often does not have explicit solutions in both E-step and M-step, which may lead to great computational cost.

2. Models and methods

2.1. Hidden Markov modeling

Consider a two-sample experiment involving two groups, treatment versus control (e.g., cocaine-treated vs. saline-treated mice). Let i index arrays, j index genes, and k index probes. For the control group, we have I_x arrays measured, and for the treatment group, we have I_y arrays. For each array, we have J genes in total, and for gene j , there are K_j probes (about 30–50) in its promoter region. For the control (or treatment) group, we let X_{ijk} (or

Y_{ijk}) be the log ratio of red channel (IP enriched) versus green channel (reference) intensities for probe k of gene j in array i , for $i = 1, \dots, I_x$ (or I_y), $j = 1, \dots, J$, and $k = 1, \dots, K_j$. We assume

$$\begin{aligned} \text{Control: } X_{ijk} &= \alpha_{jk} + \epsilon_{ijk}^x, & \epsilon_{ijk}^x &\sim N\left(0, \sigma_{xj}^2\right); \\ \text{Treatment: } Y_{ijk} &= \beta_{jk} + \epsilon_{ijk}^y, & \epsilon_{ijk}^y &\sim N\left(0, \sigma_{yj}^2\right), \end{aligned}$$

where we allow the measurement error terms of the two groups to have different gene-specific variances. Let S_{jk}^x (or S_{jk}^y) indicate whether the k th probe of gene j is IP enriched or not after the saline (or cocaine) treatment: 1 if IP enriched and 0 otherwise. Let D_{jk} denote the joint status of (S_{jk}^x, S_{jk}^y) . Table I gives the detailed information about modeling the probe-level mean intensities (after the log ratio), α_{jk} 's and β_{jk} 's, under each state of D_{jk} .

When $D_{jk} = 0$, the k th probe of gene j is enriched in neither condition (i.e., $S_{jk}^x = S_{jk}^y = 0$). Non-enrichment implies that the probes' mean intensities should be close to 0 and so are assumed to follow normal distributions with a common mean 0 (but different variances under the two conditions). Here, we do not force them to be exactly zero because real data suggest that the mean intensities, even for probes not enriched, vary from zero all the time.

When $D_{jk} = 3$, the k th probe of gene j is IP enriched after both treatments (i.e., $S_{jk}^x = S_{jk}^y = 1$). In this case, we have

$$\begin{aligned} \alpha_{jk} | D_{jk} = 3 &\sim N\left(a_j, \tau_{\alpha 1j}^2\right), \\ \beta_{jk} | D_{jk} = 3 &\sim N\left(b_j, \tau_{\beta 1j}^2\right). \end{aligned}$$

Note that the enriched probes tend to have relatively high mean intensities. So we set the gene-level means a_j and b_j to be nonzero and further from normal distributions with strictly positive means (at the genome level), as will be discussed in Section 2.2. In addition, Table I allows different variances of the mean intensities for enriched and non-enriched probes, as well as for the treatment and control groups.

Obviously, the aforementioned table involves a hidden Markov process (HMP), which assumes that the distribution of each observed data point, X_{ijk} or Y_{ijk} , depends on the unobserved (hidden) variable D_{jk} that takes on integer values from 0 to 3. Further, a Markov chain is used for the evolution of the unobserved state variable among the probes of the same gene, and so the process for D_{jk} is assumed to depend on $D_{j,k-1}$ only. Let Λ be the 4×4 transition matrix, where the (s, t) th element is given by

$$\lambda_{st} = P(D_{jk} = t | D_{j,k-1} = s)$$

for $s \in \{0, 1, 2, 3\}$ and $t \in \{0, 1, 2, 3\}$. Let $\vec{\lambda}_s = (\lambda_{s0}, \lambda_{s1}, \lambda_{s2}, \lambda_{s3})$ (i.e., the s th row of Λ), where $\sum_{t=0}^3 \lambda_{st} = 1$ for $s \in \{0, 1, 2, 3\}$. Let $\vec{\rho} = (\rho_0, \rho_1, \rho_2, \rho_3)$ be the row vector of

stationary probabilities satisfying $\vec{\rho} \Lambda = \vec{\rho}$ and $\sum_{s=0}^3 \rho_s = 1$. For ease in reading, we sometimes use $\rho(s)$ to denote ρ_s and $\lambda(s, t)$ to denote λ_{st} . Let $\mathbf{D}_j \equiv (D_{jk})_{k=1}^n$. Then for gene j ,

$$p(\mathbf{D}_j | \Lambda) = \rho(D_{j1}) \prod_{k=2}^{K_j} \lambda(D_{j,k-1}, D_{jk}), \quad (1)$$

2.2. Prior specification

For the gene-specific means a_j and b_j of IP-enriched probes, we consider normal prior distributions conditional on $\mathbf{D}_j$,

$$p(a_j | \mathbf{D}_j) \propto \left[\phi \left(a_j | \mu_a, \tau_a^2 \right) \right]^{I(n_j^x > 0)}, \quad p(b_j | \mathbf{D}_j) \propto \left[\phi \left(b_j | \mu_b, \tau_b^2 \right) \right]^{I(n_j^y > 0)}, \quad (2)$$

where we use $\phi(x | \mu, \sigma^2)$ to denote the pdf of a normal distribution with mean μ and variance σ^2 , evaluated at x ; the genome-level means μ_a and μ_b are both restricted to be strictly positive; $I(\cdot)$ is the indicator function; $n_j^x = \sum_{k=1}^{K_j} S_{jk}^x$ (i.e., the number of IP-enriched probes of gene j after the saline treatment) and $n_j^y = \sum_{k=1}^{K_j} S_{jk}^y$ (the number of IP-enriched probes of gene j after the cocaine treatment), which are both functions of $\mathbf{D}_j$. When all the probes in the promoter region of gene j are non-enriched in either condition (e.g., see panel E of Figure 3), we have $n_j^x = 0$ or $n_j^y = 0$ so that either of the two conditional priors in (2) becomes flat. This conditional specification is the key to ensure that for any gene with no binding activities in the entire promoter region under one condition, a_j or b_j is irrelevant in the full probability model, and then in the corresponding MCMC iterations where $n_j^x = 0$ or $n_j^y = 0$, they will not be updated.

Let $\Theta_{\text{genome}} = (\mu_a, \mu_b, \tau_a^2, \tau_b^2, \Lambda)$ denote the genome-level parameters. Then

$$\begin{aligned} p(\mathbf{a}, \mathbf{b}, \mathbf{D} | \Theta_{\text{genome}}) &= \prod_{j=1}^J p(a_j, b_j, \mathbf{D}_j | \Theta_{\text{genome}}) \\ &= \prod_{j=1}^J p(a_j | D_j) p(b_j | D_j) p(\mathbf{D}_j | \Lambda), \end{aligned}$$

where $p(\mathbf{D}_j | \Lambda)$ is given by (1).

Let Θ denote all the probe-level, gene-level, and genome-level parameters involved, $\mathbf{X}$ denote the data from the control group, and $\mathbf{Y}$ denote the data from the treatment group. We assume all the variance components are a priori independent. Then the full probability model is given by

$$\begin{aligned}
p(\mathbf{X}, \mathbf{Y}, \Theta) \propto & \prod_{i=1}^{I_x} \prod_{j=1}^J \prod_{k=1}^{K_j} \phi(x_{ijk} | \alpha_{jk}, \sigma_{xj}^2) \cdot \prod_{i=1}^{I_y} \prod_{j=1}^J \prod_{k=1}^{K_j} \phi(y_{ijk} | \beta_{jk}, \sigma_{yj}^2) \\
& \cdot \prod_{j=1}^J \prod_{k=1}^{K_j} \{ [\phi(\alpha_{jk} | 0, \tau_{\alpha 0j}^2)]^{1-S_{jk}^x} [\phi(\alpha_{jk} | a_j, \tau_{\alpha 1j}^2)]^{S_{jk}^x} \} \\
& \cdot \prod_{j=1}^J \prod_{k=1}^{K_j} \{ [\phi(\beta_{jk} | 0, \tau_{\beta 0j}^2)]^{1-S_{jk}^y} [\phi(\beta_{jk} | b_j, \tau_{\beta 1j}^2)]^{S_{jk}^y} \} \\
& \cdot \prod_{j=1}^J \{ [\phi(a_j | \mu_a, \tau_a^2)]^{I(n_j^x > 0)} \cdot [\phi(b_j | \mu_b, \tau_b^2)]^{I(n_j^y > 0)} \cdot \rho(D_{j1}) \prod_{k=2}^{K_j} \lambda(D_{j,k-1}, D_{jk}) \} \\
& \cdot \prod_{j=1}^J [p(\sigma_{xj}^2) p(\sigma_{yj}^2) p(\tau_{\alpha 0j}^2) p(\tau_{\alpha 1j}^2) p(\tau_{\beta 0j}^2) p(\tau_{\beta 1j}^2)] \\
& \cdot p(\mu_a, \mu_b) \cdot p(\tau_a^2) p(\tau_b^2) \cdot \prod_{s=0}^3 p(\vec{\lambda}_s),
\end{aligned}$$

where the first line is for the observed data, the second and third lines are for the probe-level parameters, reflecting the information given in Table I, the fourth and fifth lines are for the gene-level parameters, and the sixth line is for the genome-level parameters.

For the genome-level means (μ_a, μ_b) , we use independent noninformative flat priors, $U(0, L_a)$ and $U(0, L_b)$, respectively. For all the variance components, we specify weak inverse gamma priors $\text{IG}(u, v)$. For the s th row $\vec{\lambda}_s$ of the transition matrix Λ , we consider Dirichlet prior $\text{Dir}(\vec{\lambda}_s | \vec{\delta})$. As to the specification of the hyperparameters involved, we can specify the upper bounds $L_a = \max_{ijk} X_{ijk}$ and $L_b = \max_{ijk} Y_{ijk}$. Another way to specify L_a (or L_b) is to find the mean and standard deviation of all X_{ijk} 's (or Y_{ijk} 's), say $\bar{X}$, sd_x (or $\bar{Y}$, sd_y), then set $L_a = \bar{X} + 10sd_x$ and $L_b = \bar{Y} + 10sd_y$. The hyperparameters of the inverse gamma priors u and v are chosen to make the priors very vague, for example, $u = 0.01$ and $v = 0.01$. For the Dirichlet priors, we choose $\vec{\delta} = (1, 1, 1, 1)$ so that they are noninformative.

2.3. Posterior computation

We use MCMC to draw samples from the joint posterior distribution $p(\Theta | \mathbf{X}, \mathbf{Y})$, which is proportional to $p(\mathbf{X}, \mathbf{Y}, \Theta)$. The full conditional posterior distributions, as detailed in the Supporting Information[†], allow us to adopt a Gibbs sampler easily, of which all the steps can be carried out by direct sampling from known distributions. We also provide R implementation of our algorithm in the Supporting Information, along with details about computing speed. We use standard diagnostic techniques [18, 19], such as trace plots, density plots, autocorrelation plots, Gelman and Rubin statistics, and posterior confidence intervals, to check the convergence.

[†]Supporting information may be found in the online version of this article.

3. Statistical inference

As mentioned in Section 1, our primary goal is to identify probes with epigenetic changes and meanwhile infer which types of changes occur. At the probe level, there are three types of epigenetic changes: (1) IP enriched in treatment but not in control; (2) IP enriched in control but not in treatment; and (3) IP-enriched in both conditions but with different magnitudes. We label them types 1, 2, and 3, respectively. As shown in Figure 3, epigenetic changes often occur over certain genomic regions, showing obvious spatial patterns. The proposed Bayesian method offers several advantages. First, by using HMPs, it explicitly models the dependence between neighborhood probes, which allows us to detect local trends easily. Second, through its multilevel structures, it allows for information pooling among different probes of the same promoter region for estimating gene-specific variances. More importantly, it allows for information sharing across the whole genome for estimating mean intensities of IP-enriched probes and the transition matrix of the HMPs, which helps us capture global features among the data. Third, the four states of the HMPs have well-defined biological meanings, which allow us to directly call the type of change based on the corresponding posterior probabilities.

The first two types of epigenetic changes correspond to $D_{jk} = 1$ and 2, respectively. So for probe k in the promoter region of gene j , we define $\psi_{jk}(d) \equiv \Pr(D_{jk} = d | \mathbf{X}, \mathbf{Y})$ for $d = 1, 2$ (i.e., the posterior probability that this probe has the d th state). It can be estimated by

$$\hat{\psi}_{jk}(d) = \frac{\sum_{t=1}^T I(D_{jk}^{(t)} = d)}{T}, \quad d=1, 2; \quad (3)$$

where $D_{jk}^{(t)}$ is the state of the probe in the t th iteration of MCMC and T is the total number of iterations after the burn-in period. For the third type of changes, we define $\psi_{jk}(3) \equiv \Pr(D_{jk} = 3 \ \& \ \delta_{jk} > \delta_0 | \mathbf{X}, \mathbf{Y})$ (i.e., the posterior probability that this probe is enriched in both conditions but with different binding strength), where $\delta_{jk} \equiv |\alpha_{jk} - \beta_{jk}|$ and δ_0 is a predefined threshold that can be chosen according to biological relevance in applications. Then, we can estimate $\psi_{jk}(3)$ by

$$\hat{\psi}_{jk}(3) = \frac{\sum_{t=1}^T I(D_{jk}^{(t)} = 3) I(\delta_{jk}^{(t)} > \delta_0)}{T}, \quad (4)$$

where $\delta_{jk}^{(t)} = |\alpha_{jk}^{(t)} - \beta_{jk}^{(t)}|$ is calculated using posterior samples in the t th iteration. To further distinguish type 3 changes with positive/negative signs, we can define

$\psi_{jk}^+(3) \equiv \Pr(D_{jk}=3 \ \& \ \alpha_{jk} < \beta_{jk} | \mathbf{X}, \mathbf{Y})$ and $\psi_{jk}^-(3) \equiv \Pr(D_{jk}=3 \ \& \ \alpha_{jk} > \beta_{jk} | \mathbf{X}, \mathbf{Y})$, as one reviewer pointed out; then we can estimate them accordingly (i.e., simply change

$I(\delta_{jk}^{(t)} > \delta_0)$ in (4) to $I(\alpha_{jk} < \beta_{jk})$ or $I(\alpha_{jk} > \beta_{jk})$).

To detect sites with each of the three types of epigenetic changes, we can rank genes based on $\hat{\psi}_{jk}(d)$, $d \in \{1, 2, 3\}$, respectively, and select the probes on the top of the ranked list as the predicted locations with the corresponding type of change. Here, we can use Bayesian

false discovery rate (FDR) estimates [20] to guide the choice of a significance cutoff (say, κ). In our context, we can estimate the corresponding FDR for detecting the d th type of change by

$$\hat{FDR}_d(\kappa) = \frac{\sum_{j=1}^J [1 - \hat{\psi}_{jk}(d)] I [\hat{\psi}_{jk}(d) \geq \kappa]}{\sum_{j=1}^J I [\hat{\psi}_{jk}(d) \geq \kappa]}, \quad (5)$$

where $d \in \{1, 2, 3\}$ and $\kappa \in (0, 1)$. We can choose κ so that $\hat{FDR}_d(\kappa) \leq \zeta$ (say, $\zeta = 0.01, 0.05$, or 0.1).

At the probe level, the change pattern of any specific probe can be uniquely defined and must be one of the three types. However, at the gene level, different probes of the same gene could have different types of change (e.g., some probes have type 1 changes, and some others have type 2 changes). Thus, the gene-level inference requires not only a clear definition of the gene-level pattern that researchers are interested in but also a quantitative measure summarizing the corresponding posterior probabilities of the individual probes to the gene level such as average, maximum. For example, if one is interested in identifying genes with probes having any type of epigenetic change, then the gene-level posterior

probability can be defined by $\psi_j = \max_k \sum_{d=1}^3 \psi_{jk}(d)$ for gene j . Or if one is interested in identifying genes with probes having type 3 changes, then the gene-level posterior probability can be defined by $\psi_j(3) = \max_k \psi_{jk}(3)$ or gene j , as will be used in our motivating application. After the summary measure is defined, we proceed as before to detect genes with the changes.

4. Results

4.1. Simulation

Here, we examine the performance of the proposed method, labeled *BHMP* (Bayesian HMP), in detecting sites with epigenetic changes as well as types of changes. We also compared BHMP with four existing methods, labeled *Anova*, *SAM-t*, *TileMap*, and *MA*, respectively. The first is the regular ANOVA method that completely ignores the spatial dependence among probes from the same promoter region, which fits a one-way ANOVA model for each probe separately. The second is the significance analysis of microarrays method (SAM; [21]), where the SAM t -statistic, with a stabilized variance term, has been widely used to identify statistically significant genes in practice. The third method, *TileMap*, involves a two-step approach proposed in [4]: the first step uses a hierarchical empirical Bayes model to compute a test statistic for each probe; the second step uses spatial models (an HMM or a moving average method) to combine the test statistics of neighboring probes within a genomic region and then infer whether the region is of interest or not. We consider the HMM option for *TileMap* because the proposed method uses HMM, too. The last one, *MA*, is the moving average method proposed in [8], with sliding windows of size 5. Note that all the aforementioned methods can detect the sites with changes but cannot recognize which types of changes occur. For example, if a (modified) t -statistic for some probe turns out to be positive, then one can only know that the change in binding strength from the

control to treatment is positive but cannot know whether it is type 1 or 3. So to make the comparison possible, we use the sum $\sum_{d=1}^3 \hat{\psi}_{jk}(d)$ for BHMP to infer whether any change occurs at the probe.

We conducted two simulation studies. In the first study, we simulated data from the proposed Bayesian hierarchical model with HMPs. Thus, we expect BHMP to work well here. In the second study, we generated data using patterns that are similar to what we observe in real data to check the robustness of BHMP. For each dataset, we run 10,000 MCMC iterations, and the first 50% is used as burn-in when applying BHMP. To detect type 3 changes, we need to set the value of δ_0 in (4). In our previous numerical experiments, we tried different values of δ_0 and found that selection of the cutoff was not crucial for the purpose of comparison because the ROC curves do not change much as δ_0 varies. Also, the ranks of the posterior probabilities are not sensitive to the choice of δ_0 so that our inference based on top ranked sites is nonsensitive. As a result, we use a default setting that simply sets the cutoff δ_0 to be the overall difference in mean intensity between the treatment and control across all the probes.

We mention that besides the tables and figures reported in this section, additional results are reported in the Supporting Information, including trace and density plots for convergence detection, results for the gene-level inference, and the genome-level mean estimation.

4.1.1. Simulation study I—Suppose there are three arrays in each of the treatment and control groups, as in our motivating application. For each array, we simulated 1000 genes, with 50 probes within the promoter region of each gene. The genome-level mean of IP-enriched probes in the control is set at 2, and the corresponding mean in the treatment is set at 3. The variances of gene-level mean intensities are set at 0.5 for both groups. The variances of probe-level mean intensities are all set at 0.75. For measurement errors, we consider two noise levels: unit variance (setting I-1) and variance equal to 4 (setting I-2) for both groups. The transition matrix of the underlying HMPs is set to have all diagonal elements equal to 0.85 and off-diagonal elements equal to 0.05.

We first examine the performance of BHMP in estimating probe-level mean intensities. In Table II, we report mean squared errors for two estimators; the first is the simple sample mean, $\bar{X}_{jk}$ for the control and $\bar{Y}_{jk}$ for the treatment; the second is the posterior mean of MCMC samples from the proposed method, $\hat{\alpha}_{jk}$ for α_{jk} (treatment) and $\hat{\beta}_{jk}$ for β_{jk} (control). Clearly, our estimates perform much better than the sample means because of strength borrowing from neighboring probes. Also, as the noise level increases, the mean squared error becomes larger for both estimators. But our estimator seems to be less sensitive to the noise level. In Figure 4, we show scatter plots of mean estimates versus true values for four different categories (enriched/non-enriched $\times$ control/treatment) in setting I-2 as an example, where the black dots are for the naive estimator using sample mean and the red dots are for the posterior mean from BHMP. In all the four cases, the estimates from BHMP follows the straight line $y = x$ more closely, whereas the sample means tend to deviate more from the line.

Figure 5 shows the ROC curves to compare the performance of the five methods (Anova, SAM-t, TileMap, MA, and BHMP) in detecting sites with any epigenetic changes (without distinguishing the types). Table III reports the corresponding values of area under curve (AUC). As expected, BHMP is superior to the other four methods in change detection in this study. Although the performance of the other methods seems to be similar, TileMap is slightly better than SAM-t, which is better than Anova and MA.

Figure 6 shows the ROC curves for detecting the three types of epigenetic changes, using the proposed method BHMP. Table IV reports the corresponding values of AUC. These results show that BHMP can provide excellent performance in pattern recognition as well. Here, it is interesting to observe that the lowest AUC value among the three types in each setting of Table IV is higher than the corresponding overall AUC value in the last column of Table III. This is perhaps because we used the sum of individual posterior probabilities for change detection and so might lose some resolution (i.e., cancellation among the three probabilities can occur for the same probe).

4.1.2. Simulation study II—In the second study, we simulated data by mimicking triangle patterns we observed in real epigenetic data, rather than using data generated from HMMs. To do so, we consider six different patterns; five of them are shown in Figure 3, and the other one has two peaks instead of one within the promoter region of one gene. Again, we assume three arrays in each of the treatment and control groups; and for each array, we simulated 1000 genes, with 50 probes within the promoter region of each gene.

For the first 10% genes, only the middle 13 probes within each of the promoter regions in the control are IP enriched, representing cocaine-blocked TF binding regions (Figure 3A). The mean intensities of these enriched probes are set to be 10 times $\phi(l; 25, 9)$ evaluated in their locations l 's and 0 otherwise. For the second 10% genes, only the middle 13 probes in the treatment are IP enriched, representing cocaine-induced TF binding regions (Figure 3B). The mean intensities of these IP-enriched probes are set to be 15 times $\phi(l; 25, 9)$ evaluated in their locations l 's and 0 otherwise. For the third 10% genes, the middle 13 probes in both control and treatment groups are IP enriched but with different binding strength (Figure 3C). The mean intensities of these probes are the same as the first 20% genes (10 and 15 times of $\phi(l; 25, 9)$ in the control and treatment, respectively). For the fourth 10% genes, the middle 13 probes are IP enriched in both groups but with the same binding strength 10 times of $\phi(l; 25, 9)$, indicating no actual epigenetic changes in this case (Figure 3D). For the fifth 10% genes, the 13 middle IP-enriched probes are shifted to the left by 10 probes in the control and to the right by 10 probes in the treatment group, representing cocaine-induced TF binding translocation (i.e., the TF is bound to a different site of the same promoter region, showing two spikes in epigenetic changes). For the remaining 50% of the genes, all the probes are non-enriched in both groups (Figure 3E).

In this study, we consider three noise levels for measurement errors: SD equal to 0.6 (setting II-1), 0.8 (setting II-2), and 1 (setting II-3) for both groups. We summarize the values of the MSE and AUC in the right panels of Tables II–IV; the bottom panels of Figures 5 and 6 show the ROC curves. Again, BHMP offers much more efficient estimation of probe-level mean intensities than the simple estimator using sample means, and it is less sensitive to the

noise level. In terms of change detection, we can observe the following approximate order in AUC from Table III: $BHMP > TileMap \approx MA > SAM-t > Anova$, among which the first three methods all adopt some spatial smoothing techniques and so offer better performance than the last two. We can also see from Figure 5 that, when the noise level is low, both TileMap and MA are comparable with the proposed BHMP, and they are even slightly better when the false positive rate is controlled to be low. However, when the noise level increases, BHMP consistently dominates all the other methods because the performance of BHMP stays about the same, whereas the others more quickly deteriorate. The noise-resistant feature of BHMP can be also seen in pattern recognition from Table IV.

4.2. Application

The cAMP is derived from adenosine triphosphate and plays an important role in signal transduction. Many biological studies (e.g., [22–24]) have shown the essential role of CREB in cocaine addiction, whereas the cocaine-induced epigenetic changes in CREB binding profiles are still under active investigation. We applied the proposed method to the cocaine addiction study [2] mentioned in Section 1, in which CREB binding activities were measured by NimbleGen MM8 mouse promoter arrays under two experimental conditions (cocaine treatment vs. saline treatment). The transcriptional factor CREB plays a pivotal role in dopamine receptor-mediated nuclear signaling. It is associated with the neuroplasticity in drug addiction and depression. The goal of the study is to gain insight into cocaine-induced transcriptional actions of CREB. Chronic cocaine treatment induces CREB activity in the nucleus accumbens region, and then it feeds back and enhances the effects of cocaine treatment. Further, studies have shown that cocaine treatment will alter the phosphorylation status of CREB and change its binding activities, which leads to both positive and negative reinforcement of the drug addiction [25]. Genes with altered CREB binding profiles in their promoter regions are very likely to be regulated by CREB and play important roles in the molecular mechanism of cocaine addiction, which could be potential therapeutic targets. Thus, researchers are interested in identifying genes with CREB binding alterations. In particular, they are most interested in genes with probes having type 3 changes (binding activity occurs in both treatment and control conditions but with different strength; e.g., pattern C in Figure 3) because they believe this pattern reflects one of the most realistic biological situations.

As mentioned in Section 1, the experiments were performed on both cocaine-treated and saline-treated mice, with three replicates (arrays) in each condition. The data, available at Gene Expression Omnibus with the series number GSE16184, were preprocessed and normalized using MAC2 software [26]. In the original analysis of Renthal *et al.* [2], the data were spatially smoothed using a simple sliding window method and then compared between cocaine and saline treatment conditions within each sliding window [5, 6]. As a result, 1743 genes were identified to have cocaine-induced alteration of CREB binding. However, there was no direct way to further distinguish the change types and identify genes with pattern C, without examining genes one by one via ad hoc approaches (e.g., visual inspection).

To identify genes with probes having type 3 changes, we applied our proposed Bayesian method BHMP to analyze the data, where MCMC was used to draw samples from the joint

posterior distribution with a Gibbs sampler. We used the largest value of the posterior probabilities among the K_j probes to measure the significance of gene j , that is,

$\hat{\psi}_j(3) \equiv \max_{k=1}^{K_j} \hat{\psi}_{jk}(3)$, and then kept those with $\hat{\psi}_j(3) > 0.995$. The total number of MCMC iterations is 4000, and we discarded the first 2000 iterations for burn-in. We checked the convergence by using Gelman and Rubin's statistics and other standard tools. Figure 7 shows the posterior distributions of the genome-level mean intensities of IP-enriched probes in the control and treatment (i.e., μ_a and μ_b). It indicates that cocaine treatment has altered the status of CREB, which leads to a significantly increased CREB binding level under cocaine treatment (μ_b), compared with that under saline treatment (μ_a). This is consistent with the fact that cocaine treatment will increase the level of CREB and hence affect many genes regulated by CREB [27].

In total, we identified 227 genes with probes having type 3 changes using the proposed method. We note that there is no other formal statistical method that can do so. Among them, 83 genes were also identified by the simple comparison in [2] but without recognizing the change patterns, whereas 144 genes were newly identified by our proposed method. We applied the pathway and network analysis on the 144 newly identified genes using Ingenuity IPA Software (Ingenuity Systems Inc., Redwood City, CA, USA) to gain some insight at the system level. Interestingly, the most significant network is the nervous system development related network (Figure 8), and in the network, many genes interact with the CREB gene, according to the current biological knowledge. The results show that the newly identified 144 genes contain many CREB targets, indicating the advantage of the proposed method.

We also found some biologically interesting genes. For example, gene *Rap2a* is previously known to be associated with addiction to other drugs [28] and is a known target of CREB [29]. We show the CREB binding profiles for gene *Rap2a* under saline and cocaine

treatments in Figure 9, along with the posterior probability $\hat{\psi}_{jk}(3)$ in (4) for every probe in the corresponding promoter region. From the figure, we can see that the difference in the binding strength between the two conditions is relatively small so that the gene was not identified by the original comparison in [2]. But using the proposed method, we can identify the gene with very high confidence by jointly modeling the CREB binding profiles in both cocaine-treated and saline-treated conditions. Further, it is known that CREB binds to DNA sequences TGACGTCA, called cAMP response elements in the gene promoter region, and thereby regulates the gene in the downstream. Figure 9 shows that the cAMP response element binding sequence for gene *Rap2a* (TGACGTCA; [30]) is located right in the middle of the region with high posterior probabilities, which confirms that the discovery from our method is real. Similarly, we have found CREB binding sequences in the peak regions of several other genes identified by our method, such as *Clasp2*.

Figure 10 shows another gene example that our method has identified, *Rps11* (40S ribosomal protein S11). This gene is associated with the alteration of CREB binding in social isolation status [31], and so it is likely to be related to cocaine-induced plasticity. Again, the gene was overlooked by the original analysis as the difference in CREB binding strength at the promoter region is relatively small. Similar examples identified by our

method include *Notch1* [32], *Pdyn* [33], and *Fos* [34–37], all of which are well known for their important roles in molecular mechanisms of cocaine-induced neuron plasticity.

In summary, our proposed method BHMP, via a coherent model-based approach, not only enables us to identify specific patterns of epigenetic changes that are of particular interest but also offers great power in detecting such gene targets.

5. Discussion

Motivated by an epigenetic study in drug addiction, we have developed a Bayesian hierarchical approach relying on HMPs to analyze spatially dependent epigenetic profiling data that have a complex multilevel structure (probes nested within gene promoter regions that are nested within a certain genome), as well as a complex design (a two-sample experiment with internal control). Besides offering epigenetic change detection, our Bayesian approach has the ability of pattern recognition based on the posterior. This is a very attractive and unique feature, which allows biologists to automatically classify detected sites and identify those with the most interesting patterns. All the existing methods available for change detection cannot do so.

In terms of epigenetic change detection, we have shown through simulation studies that the proposed method BHMP can offer better efficiency than the competing methods. This is not surprising. As mentioned before, Anova analyzes data for different probes of the same gene in a completely separate way; hence, no strength borrowing or spatial smoothing is carried out here. MA offers one step forward from Anova, by smoothing the observed t -statistics over adjacent probes of the same gene via sliding windows to take into account the spatial dependence. SAM is a variant of two-sample t -test with a variance stabilizing factor, which borrows information from other relevant probes for error estimation. However, it totally ignores the spatial correlation. Tilemap (with the HMM option) uses HMMs to account for spatial dependence among t -statistics within a genomic region, which are computed for each probe based on a hierarchical empirical Bayes model. Obviously, Tilemap enables both strength borrowing and spatial smoothing and so can offer better performance, which has been confirmed by our simulation. Like Tilemap, our method BHMP does both. However, BHMP does not rely on a summary statistic for comparison between two experimental conditions; instead, it explicitly models every possible state under the treatment and control through four-state HMPs and build a hierarchical model to reflect the actual multilevel data structure. By doing so, we fully utilize the information underlying the data and lead to efficiency in statistical inference, in addition to offering the feature of pattern recognition. Through a data example, we have shown that BHMP can identify epigenetically different genomic regions with biologically interesting patterns that cannot be found by other methods and also provide meaningful interpretation.

Another strength of BHMP is that it relies on essentially no tuning parameters. Unlike many other algorithms, users only need to input the normalized log2 ratios. This is convenient for an end user with little or no statistical training. In all our analyses, we have used the default parameterizations specified in Section 2.2. However, if meaningful prior knowledge is

available, certain features of the proposed BHMP may be changed to produce potentially better results that incorporate such knowledge.

Finally, we should mention that with some modification or extension, our methodology can be generally purposed and provide a useful way to analyze spatially correlated data from multiple-condition ChIP experiments with internal control. As one reviewer pointed out, for experiments with S different conditions, we could generalize four-state HMPs to 2^S -state HMPs with straightforward changes in the algorithm. Further, the methodology itself, in principle, has no limit on platforms. The implementation of our methodology is readily available for whole-genome tiling arrays, as long as we model the spatial dependence within each chromosome region instead of the promoter region of each gene. In this case, the model and algorithm remain the same. Also, if we consider the Poisson or negative binomial distribution instead of the commonly used normal assumption, then our methodology could be generalized to ChIP-seq data. However, the computational cost for Bayesian MCMC can be prohibitive for such data without appropriate adaption. Clearly, there is ample space for future work on applying the methodology to ChIP-seq data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NSF grants DMS-0906545 and DMS-0907562, 4R33DA027592.

References

1. Portela A, Esteller M. Epigenetic modifications and human disease. *Nature Biotechnology*. 2010; 28(10):1057–1068. 10.1038/nbt.1685
2. Renthall W, Kumar A, Xiao G, Wilkinson M, Covington HE III, Maze I, Sikder D, Robison AJ, Dietz DM, LaPlant Q, Russo SJ, Vialou V, Chakravarty S, Kodadek TJ, Stack A, Kabbaj M, Nestler EJ. Genome-wide analysis of chromatin regulation by cocaine reveals a role for sirtuins. *Neuron*. 2009; 62(3):335–348. [PubMed: 19447090]
3. Carlezon WA Jr, Duman RS, Nestler EJ. The many faces of CREB. *Trends in Neuroscience*. 2005; 28(8):436–445. <http://dx.doi.org/10.1016/j.tins.2005.06.005>. 10.1016/j.tins.2005.06.005
4. Ji H, Wong W. Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*. 2005; 18:3629–3636. [PubMed: 16046496]
5. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmand TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. *Nature*. 2005; 436:876–880. [PubMed: 15988478]
6. Buck MJ, Nobel AB, Lieb JD. ChIPOTIe: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biology*. 2005; 6:R97. [PubMed: 16277752]
7. Li W, Meyer CA, Liu X. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*. 2005; 21(suppl 1):274–282.
8. Keles S, Van der Laan MJ, Cawley SE. Multiple testing methods for chip-chip high density oligonucleotide array data. *Journal of Computational Biology*. 2006; 13:579–613. [PubMed: 16706714]
9. Keles S. Mixture modeling for genome-wide localization of transcription factors. *Biometrics*. 2007; 63:10–21. [PubMed: 17447925]

10. Reiss DJ, Facciotti MT, Baliga NS. Model-based deconvolution of genome-wide DNA binding. *Bioinformatics*. 2008; 24:396–403. [PubMed: 18056063]
11. Gottardo R, Li W, Johnson WE, Liu XS. A flexible and powerful Bayesian hierarchical model for chip-chip experiments. *Biometrics*. 2008; 64(2):468–478. [PubMed: 17888037]
12. Wu M, Liang F, Tian Y. Bayesian modeling of ChIP-chip data using latent variables. *BMC Bioinformatics*. 2009; 10:352. [PubMed: 19857265]
13. Mo Q, Liang F. Bayesian modeling of chip-chip data through a high-order Ising model. *Biometrics*. 2010; 66(4):1284–1294. <http://dx.doi.org/10.1111/j.1541-0420.2009.01379.x>. 10.1111/j.1541-0420.2009.01379.x [PubMed: 20128774]
14. Mo Q, Liang F. A hidden Ising model for ChIP-chip data analysis. *Bioinformatics*. 2010; 26:777–783. [PubMed: 20110277]
15. Gelfond JA, Gupta M, Ibrahim JG. A Bayesian hidden Markov model for motif discovery through joint modeling of genomic sequence and ChIP-chip data. *Biometrics*. 2009; 65:1087–1095. [PubMed: 19210737]
16. Berard C, Martin-Magniette ML, Brunaud V, Aubourg S, Robin S. Unsupervised classification for tiling arrays: ChIP-chip and transcriptome. *Statistical Applications in Genetics and Molecular Biology*. 2011; 10(1)10.2202/1544-6115.1692
17. Zheng M, Barrera LO, Ren B, Wu YN. ChIP-chip: data, model, and analysis. *Biometrics*. 2007; 63(3):787–796. [PubMed: 17825010]
18. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457–511.
19. Brooks SP, Roberts GO. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*. 1998; 8:319–335.
20. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004; 5(2):155–176. <http://biostatistics.oxfordjournals.org/cgi/content/abstract/5/2/155>. [PubMed: 15054023]
21. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*. 2001; 98(9):5116–5121. <http://www.pnas.org/cgi/content/abstract/98/9/5116>.
22. Carlezon WA Jr, Thome J, Olson VG, Lane-Ladd SB, Brodtkin ES, Hiroi N, Duman RS, Neve RL, Nestler EJ. Regulation of cocaine reward by CREB. *Science*. 1998; 282(5397):2272–2275. [PubMed: 9856954]
23. McClung CA, Nestler EJ. Regulation of gene expression and cocaine reward by CREB and DeltaFosB. *Nature Neuroscience*. 2003; 6(11):1208–1215.
24. Hollander JA, Im HI, Amelio AL, Kocerha J, Bali P, Lu Q, Willoughby D, Wahlestedt C, Konkright MD, Kenny PJ. Striatal microRNA controls cocaine intake through CREB signalling. *Nature*. 2010; 466(7303):197–202. [PubMed: 20613834]
25. Walters CL, Blendy JA. Different requirements for cAMP response element binding protein in positive and negative reinforcing properties of drugs of abuse. *Journal of Neuroscience*. 2001; 21:9438–9444. [PubMed: 11717377]
26. Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai AK, Liu JS, Chen R, Liu XS. Model-based analysis of two-color arrays (MA2C). *Genome Biology*. 2007; 8:R178. [PubMed: 17727723]
27. Nestler EJ. Historical review: molecular and cellular mechanisms of opiate and cocaine addiction. *Trends in Pharmacological Sciences*. 2004; 25(4):210–218. <http://www.sciencedirect.com/science/article/pii/S0165614704000537>. 10.1016/j.tips.2004.02.005 [PubMed: 15063085]
28. Konu O, Kane JK, Barrett T, Vawter MP, Chang R, Ma JZ, Donovan DM, Sharp B, Becker KG, Li MD. Region-specific transcriptional response to chronic nicotine in rat brain. *Brain Research*. 2001; 909(1–2):194–203. [PubMed: 11478936]
29. Kawabe H, Neeb A, Dimova K, Young SM Jr, Takeda M, Katsurabayashi S, Mitkovski M, Malakhova OA, Zhang DE, Umikawa M, Kariya K, Goebels S, Nave KA, Rosenmund C, Jahn O, Rhee J, Brose N. Regulation of rap2a by the ubiquitin ligase nedd4-1 controls neurite development. *Neuron*. 2010; 65(3):358–372. [PubMed: 20159449]
30. Pfenning AR, Schwartz R, Barth AL. A comparative genomics approach to identifying the plasticity transcriptome. *BMC Neuroscience*. 2007; 8:20. [PubMed: 17355637]

31. Wilkinson MB, Xiao G, Kumar A, LaPlant Q, Renthal W, Sikder D, Kodadek TJ, Nestler EJ. Imipramine treatment and resiliency exhibit similar chromatin regulation in the mouse nucleus accumbens in depression models. *Journal of Neuroscience*. 2009; 29:7820–7832. [PubMed: 19535594]
32. Yao H, Duan M, Hu G, Buch S. Platelet-derived growth factor B chain is a novel target gene of cocaine-mediated Notch1 signaling: implications for HIV-associated neurological disorders. *Journal of Neuroscience*. 2011; 31:12449–12454. [PubMed: 21880906]
33. Andersson M, Konradi C, Cenci MA. cAMP response element-binding protein is required for dopamine-dependent gene expression in the intact but not the dopamine-denervated striatum. *Journal of Neuroscience*. 2001; 21:9930–9943. [PubMed: 11739600]
34. Hope BT, Nye HE, Kelz MB, Self DW, Iadarola MJ, Nakabeppu Y, Duman RS, Nestler EJ. Induction of a long-lasting ap-1 complex composed of altered Fos-like proteins in brain by chronic cocaine and other chronic treatments. *Neuron*. 1994; 13(5):1235–1244. [PubMed: 7946359]
35. Moratalla R, Elibol B, Vallejo M, Graybiel AM. Network-level changes in expression of inducible Fos-Jun proteins in the striatum during chronic cocaine treatment and withdrawal. *Neuron*. 1996; 17(1):147–156. [PubMed: 8755486]
36. Moratalla R, Vickers EA, Robertson HA, Cochran BH, Graybiel AM. Coordinate expression of c-fos and jun b is induced in the rat striatum by cocaine. *Journal of Neuroscience*. 1993; 13(2):423–433. [PubMed: 8426221]
37. Steiner H, Gerfen CR. Cocaine-induced c-fos messenger RNA is inversely related to dynorphin expression in striatum. *Journal of Neuroscience*. 1993; 13(12):5066–5081. [PubMed: 7504719]

Figure 1.

An illustration of the experimental design in our motivating example: CTM represents cocaine-treated mice, and STM represents saline-treated mice; each well (oval shape) represents a probe in an array. The experiment has two samples (treatment vs. control), each with three replicates (i.e., arrays). Each replicate has paired internal control: the IP-enriched sample and normal DNA sample (reference) were both hybridized to the same array; the array is then scanned, and two intensity values, corresponding to Cy5 (IP) and Cy3 (reference), respectively, are extracted for each probe in the array.

Figure 2.
Comparison of different study designs.

Figure 3.

An illustration of typical types of epigenetic alterations using simulated data. Panel A shows a region with cocaine-blocked TF binding: the region is only bound by the TF in saline-treated mice (black line) but not in cocaine-treated mice (red line). Panel B shows a region with cocaine-induced TF binding: it is only bound by the TF in the cocaine condition but not in the saline condition. Panel C shows a region where the cocaine treatment changes the TF binding strength. In addition, we plot two cases in which no epigenetic alteration occurs: panel D shows a region that is bound by the TF in both conditions, with the same binding strength; and panel E shows a region bound by the TF in neither condition.

Figure 4.

Simulation results: scatter plots to show the relationship between estimates versus true values of probe-level mean intensities. Red dots are for the posterior mean from BHMP, and black dots are for the naive estimator using the sample mean. The straight line in each panel represents $y = x$. The R^2 value between the posterior mean and the naive estimator is reported in the upper left corner of each scatter plot. All the four R^2 values are moderately strong, but not too strong, indicating the potential improvement that can be achieved by the proposed method.

Figure 5.

Simulation results: ROC curves for detecting sites with any epigenetic changes (without distinguishing the types), to compare BHMP with Anova, SAM-t, TileMap, and MA. The top panels labeled 'I-1' and 'I-2' are for simulation study I with the noise level $\sigma = 1$ and 2, respectively; and the bottom panels labeled 'II-1', 'II-2', and 'II-3' are for simulation study II with the noise level $\sigma = 0.6, 0.8$, and 1, respectively.

Figure 6.

Simulation results: ROC curves for detecting types of epigenetic changes using the proposed method BHMP. Type 1 means IP enriched in treatment but not in control; type 2 means IP enriched in control but not in treatment; and type 3 means IP enriched in both conditions but with different magnitudes.

Figure 7.

The cocaine addiction study: the white bars show the histogram of μ_a (i.e., the genome-level mean intensity of IP-enriched probes under saline treatment); and the shaded bars show the histogram of μ_b (i.e., the genome-level mean intensity of IP-enriched probes under cocaine treatment) formed by posterior samples.

Figure 8.

The most significant gene network identified by Ingenuity IPA Software. In the network, each node represents a gene/molecule, and each edge represents a relationship. The detailed information about different shapes of the nodes can be found from the Ingenuity website. The nodes highlighted in gray are those identified by our proposed method, and the nodes in white are the genes interacting with those in gray based on the current biological knowledge. The CREB gene is highlighted in blue, and its interactions are highlighted in light blue.

Figure 9.

CREB binding profiles at the promoter region of gene *Rap2a* for cocaine-treated (in black) and saline-treated (in red) mice. The line and dots in blue show the posterior probability $\hat{\psi}_{jk}(3)$ (i.e., binding at both conditions but with different strength), of which the scale is indicated at the right side.

Figure 10.

CREB binding profiles at the promoter region of gene *Rps11* for cocaine-treated (in black) and saline-treated (in red) mice. The line and dots in blue show the posterior probability $\hat{\psi}_{jk}(3)$ (i.e., binding at both conditions but with different strength), of which the scale is indicated at the right side.

Table I

A hidden Markov model for analysis of epigenetic data from promoter arrays.

D_{jk}	S_{jk}^x	S_{jk}^y	α_{jk}	β_{jk}
0	0	0	$N(0, \tau_{\alpha 0j}^2)$	$N(0, \tau_{\beta 0j}^2)$
1	0	1	$N(0, \tau_{\alpha 0j}^2)$	$N(b_j, \tau_{\beta 1j}^2)$
2	1	0	$N(a_j, \tau_{\alpha 1j}^2)$	$N(0, \tau_{\beta 0j}^2)$
3	1	1	$N(a_j, \tau_{\alpha 1j}^2)$	$N(b_j, \tau_{\beta 1j}^2)$

Table II

Simulation results: mean squared error (MSE) for estimating probe-level mean intensities.

	MSE for study I		MSE for study II			
	noise level		noise level			
	I-1	I-2	II-1	II-2	II-3	
Setting	$\sigma = 1$	$\sigma = 2$	$\sigma = 0.6$	$\sigma = 0.8$	$\sigma = 1$	
Control	0.32	1.34	0.12	0.21	0.32	
	0.26	0.68	0.04	0.05	0.05	
Treatment	0.33	1.38	0.12	0.22	0.34	
	0.26	0.70	0.04	0.06	0.06	

Simulation results: AUC for detecting sites with any epigenetic changes (without distinguishing the types), to compare the proposed method BHMP with four existing methods Anova, SAM-t, TileMap, and MA.

Table III

	AUC for study I		AUC for study II			
	noise level		noise level			
	I-1	I-2	II-1	II-2	II-3	
Setting	$\sigma = 1$	$\sigma = 2$	$\sigma = 0.6$	$\sigma = 0.8$	$\sigma = 1$	
Anova	0.77	0.66	0.71	0.66	0.64	
SAM-t	0.80	0.69	0.73	0.67	0.65	
TileMap	0.78	0.74	0.85	0.77	0.65	
MA	0.74	0.69	0.81	0.76	0.74	
BHMP	0.94	0.84	0.86	0.86	0.82	

Table IV

Simulation results: AUC for detecting types of epigenetic changes using the proposed method BHMP.

	AUC for study I		AUC for study II			
	noise level		noise level			
	I-1	I-2	II-1	II-2	II-3	
Setting	$\sigma = 1$	$\sigma = 2$	$\sigma = 0.6$	$\sigma = 0.8$	$\sigma = 1$	
Type 1	0.98	0.93	0.80	0.80	0.80	
Type 2	0.98	0.94	0.76	0.76	0.78	
Type 3	0.95	0.90	0.93	0.94	0.93	