HHS Public Access

Author manuscript

Stat Appl Genet Mol Biol. Author manuscript; available in PMC 2016 May 13.

Published in final edited form as:

Stat Appl Genet Mol Biol. ; 12(2): 189–205. doi:10.1515/sagmb-2012-0057.

Detection of Epigenetic Changes Using ANOVA with Spatially Varying Coefficients

Guanghua Xiao[1], **Xinlei Wang**[1], **Quincey LaPlant**[2], **Eric Nestler**[2], and **Yang Xie**[1]

[1]UT Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, TX 75390

[2]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, One Gustave Levy Place, New York, NY 10029

Abstract

Identification of genome-wide epigenetic changes, the stable changes in gene function without a change in DNA sequence, under various conditions plays an important role in biomedical research. High-throughput epigenetic experiments are useful tools to measure genome-wide epigenetic changes, but the measured intensity levels from these high-resolution genome-wide epigenetic profiling data are often spatially correlated with high noise levels. In addition, no formal statistical method was developed to compare genome-wide epigenetic changes across multiple conditions. In this study, we consider ANOVA models with spatially varying coefficients, combined with a hierarchical Bayes approach, to explicitly model spatial correlation caused by location-dependent biological effects (i.e., epigenetic changes) and borrow strength among neighboring probes to compare epigenetic changes across multiple conditions. Through simulation studies and applications in drug addiction and depression models, we find that our approach compares favorably with competing methods; it is more efficient in estimation and more effective in detecting epigenetic changes. In addition, it can provide biologically meaningful results.

Keywords

AR1; autoregressive; Bayesian hierarchical model

1 Introduction

Epigenetics can be defined as the study of stable changes in gene function without a change in DNA sequence. Such changes, including histone modifications (methylation or acetylation) and DNA methylation, may have long-lasting effects in cells and cause many diseases such as cancer, depression and drug addiction, to name a few. Identification of genome-wide epigenetic changes under various conditions plays an important role in biology, medicine and evolution research [1–3]. Genome-wide Chromatin Immunoprecipitation (ChIP) experiments including ChIP on microarray chip (ChIP-chip) and ChIP-sequencing (ChIP-seq) are powerful high-throughput methods for analyzing epigenetic modifications and genomic regions bound to regulatory proteins [4–10].

Correspondence: Guanghua Xiao, UT Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, TX 75390, guanghua.xiao@utsouthwestern.edu..

One feature associated with such epigenetic data is that the measured intensity levels are often spatially correlated with high noise levels. Various approaches have been used to incorporate the neighboring dependency of epigenetic data, including sliding window methods combined with certain test statistics (e.g., MAT[11] and ChIPOTle[12]), local regression fitting methods across moving windows (e.g., [6, 13]), hidden Markov models (e.g., [14–16]), and Bayesian (hierarchical) methods (e.g., [17–22]). All of the methods show that accounting for spatial dependence can greatly enhance detection efficiency. Although many statistical methods have been developed to process epigenetic data and identify binding sites of transcription factors (e.g., [19, 20, 22–27]), relatively fewer methods have been developed to identify genome-wide epigenetic changes. For example, ChIPDiff [28] used a hidden Markov model to identify the chromosome regions with epigenetic changes; Tilemap [29] provides options to use either the moving windows or hidden Markov models; and Taslim et al [30] proposed a statistical method that uses mixture models to identify positive and negative differential binding sites. These existing methods were developed to identify epigenetic changes or differential binding sites among different groups. However, some recent epigenetic studies used more complex designs, such as two-way ANOVA settings (see the second motivating example) or time course measurements (for example, modENCODE project measures the epigenetic profiles at different time points or growth stages), and the existing methods could not handle the complex designs efficiently. In this study, we incorporate spatial correlation into a hierarchical Bayesian linear regression model, which can offer great flexibility to study the epigenetic changes under complex experimental designs. More specifically, we consider linear models with spatially varying regression coefficients in spirit of Gelfand *et al.* [31], where location-dependent epigenetic changes (represented by the coefficients) are modeled by a simple autoregressive model.

The paper is arranged as follows. Section 2 describes the research background. In Section 3, we propose a Bayesian approach for detecting epigenetic changes, ANOVA with spatially varying coefficients, in which we set up a hierarchical Bayes model, specify prior distributions and discuss posterior computation and inference. In Section 4, we present numerical results for both one-way and two-way ANOVA settings, where we examine the performance of our approach using simulated data with the first-order autocorrelation (AR1) structures and data that mimic realistic patterns (for the purpose of robustness checking). Section 5 applies the proposed approach to two genome-wide epigenetic data sets and shows its effectiveness and usefulness. Section 6 concludes the paper with a brief discussion.

2 Motivating Examples

Our work is motivated by studies of molecular mechanisms of drug addiction and depression [32–34], which are two of the most common illnesses in the world. Although drug addiction and depression involve many psychological and social factors, they also represent a biological process: repeated exposure of stress or a drug of abuse causes stable changes at molecular and cellular levels in brain, and alters the functioning of individual neurons and larger neural circuits [35]. Increasing evidence suggests that gene expression changes in brain nucleus accumbens regions (NAc, a major brain reward region), which contribute to the pathogenesis and persistence of depression and drug addiction, are mediated in part by epigenetic mechanisms [36, 37]. To better understand how the brain responds to repeated

perturbations (under normal and pathological conditions), epigenetic profiling data were generated from mouse NAc using NimbleGen promoter arrays. The distance between two consecutive probes within same promoter region is only 100 ~ 200 base pairs, while that between probes from different promoter regions (of two distinct genes) are relatively far away, typically at least several hundred kilo-base pairs. Because of this feature, it is reasonable to believe that the epigenetic changes from the same promoter region are spatially dependent, while those of different genes are spatially independent.

2.1 Cocaine addiction study

The first motivating dataset was generated from a cocaine addiction study [32] which contains histone H3 methylation (dimethylK9/K27) data measured by ChIP-chip experiments using NimbleGen MM8 mouse promoter arrays. The experiments were performed on both cocaine and saline treated mice to detect cocaine induced changes in histone modifications. In the experiments, fresh nucleus accumbens (NAc) punches were processed for ChIP as described in [34]. The samples were amplified and labeled , and then hybridized to the promoter arrays with three biological replicates per condition. Each biological replicate was prepared by NAc punches pooled from ten mice to reduce the biological variability. The goal of the study is to identify histone modification changes between cocaine and saline treated samples.

2.2 Depression study

The second motivating dataset was generated from a depression study [33], which also used NimbleGen MM8 mouse promoter arrays to characterize histone H3 methylation (dimethylK9/K27) that occur in the NAc in response to chronic stress with and without antidepressant treatment. In the experiment, the choonic stress is introduced by a “social defeat” mouse model. When housed with an unfamiliar mouse in a wire mesh cate, the undefeated control mice spent most of time interacting socially with an unfamiliar target mouse, while the defeated mice spent less time in close proximity to the target mouse, which is a depression-like symptom. The mice were then divided into treatment groups. For each group (defeated or control), one-half of the animals received imipramine, an antidepressant drug; and the other half in each group received saline as the control.

Under this two-way ANOVA design, we want to study histone methylation in the NAc induced by social defeat and imipramine treatment. Because previous studies have shown that imipramine treatment reverses the social interaction deficit in defeated animals [38, 39], we are particularly interested in identifying genes with significant interactions between the two factors (i.e., defeated/control mice, imipramine/saline treatment).

2.3 Exploratory data analysis

Before applying any formal statistical analysis to the datasets, we first did exploratory analysis on several studies to show data features. Figure 1 presents examples of genes (i.e., CART and CDK5) with signals of cocaine-induced histone methylation and acetylation changes in the NAc from our previous study[34]. The fold changes of these signals (between cocaine and saline treated conditions) in histone H4 acetylation (measured by NimbleGen MM5 promoter arrays) and methylation of histone H3 at K9 and K27 (measured by

NimbleGen MM8 promoter arrays) are plotted by red and blue colored lines, respectively. It clearly shows that the changes occur over specific segments of the chromosome rather than on isolated points, and so they are spatially correlated. More examples can be found in Figure 1B of Renthal *et al.* [34]. This observation is consistent with the previous studies that epigenetic changes have spatial patterns. To further explore these spatial patterns in the cocaine addiction dataset [32], we calculated the first order autocorrelation function (ACF1) in the histone methylation for all the genes in that dataset [32], including those without signals of epigenetic modifications. For the genes with no signals, the autocorrelation might not exist. Figure 2 plots the histograms of ACF1 after the saline and cocaine treatment, respectively. Overall, it indicates positive autocorrelation among adjacent probes, which probably comes from those genes with the biological effects. Currently, most existing statistical methods for ChIP-chip data focus on analysis of transcription factor (TF) binding to identify the binding sites. The TF binding happens at a specific point (several base pairs) on a chromosome. By contrast, the histone modifications take place on a specific segment of the chromosome so that spatial dependence is very likely to occur in regions wherever such real biological effects exist. This is consistent with our observations from the datasets. There is currently a lack of methods to identify the spatially correlated changes in such methylation or acetylation profiles.

Due to the complexity of the NAc region, with respect to its cellular heterogeneity, as for virtually all brain regions, it is extremely difficult to separate the brain regions affected by the psychiatric disorders from the unaffected regions. As a result, dissected brain tissues are usually a mixture of the affected and unaffected tissues, and the measured intensity levels in these *in vivo* studies exhibit very larger variability. Even though NAc punches pooled from multiple mice were used for each biological replicate, the signal to noise ratio is still low. In addition, the experiments were all performed with small sample sizes (3 replicates for each group), and so some important genes might be overlooked because of the lack of statistical power using conventional statistical analysis. Thus, modeling spatial dependence explicitly may offer a great advantage in these studies.

In summary, the goals of cocaine addiction and depression studies are to identify epigenetic changes across multiple conditions. The measurement variability is large for the datasets due to the complexity of the sample preparation, and there is strong spatial correlation in epigenetic changes across the chromosome. Incorporating such correlation into statistical modeling may reduce variability and increase statistical power for the analysis. Motivated by these two examples, next we develop Bayesian linear models with spatially varying coefficients to analyze such epigenetic data.

3 ANOVA with Spatially Varying Coefficients

3.1 Model specification

Let Y_{ijk} be the centralized log ratio of red channel (IP-enriched) vs. green channel (background) intensities for probe k of gene j in array i, for $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K_j$, where I is the number of arrays, J is the number of genes, and K_j is the number of probes for gene j (about 30~50). Here, centralization means that $Y_{ijk}=Y^*_{ijk} - \bar{Y}_j$, where Y^*_{ijk}

is the raw log ratio observed and $\bar{Y}_j = \sum_i \sum_k Y^*_{ijk} / (IK_j)$, so that for each gene *j*, the average of Y_{ijk}s is 0. We consider ANOVA in the form of linear models for gene *j*, which connects the response to the treatments, conditions or blocks,

$$\vec{Y}_{jk} = \mathbf{X}\vec{\beta}_{jk} + \vec{\epsilon}_{jk}, \quad (1)$$

where *k* indexes the order of probes of the gene.

$\vec{Y}_{jk} = (Y_{ijk})_{i=1}^{I}$, $\mathbf{X} = \left(X_{i,0}, \ldots, X_{i,L}\right)_{i=1}^{I}$ is an $I \times (L+1)$ design matrix with $X_{i,0} \equiv 1$, $\vec{\epsilon}_{jk} = (\epsilon_{ijk})_{i=1}^{I}$, and $\vec{\beta}_{jk} = (\beta_{jk,l})_{l=0}^{L}$, where L is the total number of the explanatory variables/covariates under consideration. Assume $\epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2_{jk}\right)$. Let $\boldsymbol{Y}_j = \left(\vec{Y}_{j1}, \ldots, \vec{Y}_{jK_j}\right)_{I \times K_j}$ and $\boldsymbol{\beta}_j = \left(\vec{\beta}_{j1}, \ldots, \vec{\beta}_{jK_j}\right)$ be a $(L+1) \times K_j$ matrix. Then $\vec{\beta}_{jk}$ is a column vector and $\vec{\beta}_{j,l} = \left(\beta_{j1,l}, \ldots \beta_{jK_j,l}\right)$ is a row vector. Let $\boldsymbol{\sigma^2_j} = diag\left(\sigma^2_{j1}, \ldots, \sigma^2_{jK_j}\right)$.

The model in (1) can be used in ChIP-chip experiments under ANOVA setups. In an one-way ANOVA setting, for example, let $X_{i,1}$ equal 1 for the cocaine treatment and 0 for the saline treatment. The regression parameter $\beta_{jk,1}$ is the difference in some epigenetic modification induced by the cocaine treatment. We are interested in whether any $\beta_{jk,1}$ is significantly different from 0 so that we can further infer which genes have a significant difference between the cocaine and saline-treated conditions. In a two-way ANOVA setup, for example, in a depression-related study [33], we have a dataset with two different mice groups: N (normal) and D ("depressed," i.e., chronically stressed), and two types of treatment: C (saline, i.e., control) and T (chronic antidepressant drug administration). So there are 2 × 2 groups: NC, NT, DC, and DT, with replicates within each group. Here, $X_{i,1}$ = 1 for the depressed (stressed) mice and 0 for the normal mice; and $X_{i,2}$ = 1 for the antidepressant treated mice and 0 for the control mice. We want to study the cross effect of depression and drug on the methylation measurements; that is, for a specific gene, whether the two groups of mice react differently to the drug (as opposed to the control) in its epigenetic profiles. So we need an interaction term $X_{i,3} = X_{i,1} \times X_{i,2}$ and testing the corresponding coefficient $\beta_{jk,3} = 0$ is our main interest.

For gene *j*, there often exists significant spatial correlation in real biological effects (such as changes in epigenetic profiles) among its probes. As mentioned in the Introduction, this is because the probes belong to the same promoter region and are arranged in close physical proximity along the chromosome. Thus, to construct the covariance structure, we consider a linear neighboring structure and use a first-order autoregressive model for $\vec{\beta}_{j,l}$. That is, for $j = 1, \ldots, J$, and $l = 0, \ldots, L$,

$$\vec{\beta}_{j,l} | \mathbf{B}_{j,l} \sim N\left(\vec{0}, \mathbf{B}_{j,l}\right). \quad (2)$$

Here, the covariance matrix $\mathbf{B}_{j,l}$ has an AR1 structure: $|\mathbf{B}_{j,l}|=\left(\tau_{j,l}^2\right)^K/\left(1-\rho_{j,l}^2\right)$, namely

$$\mathbf{B}_{j,l}=\frac{\tau_{j,l}^2}{1-\rho_{j,l}^2}\begin{pmatrix}1 & \rho_{j,l} & \cdots & \rho_{j,l}^{K_j-1}\\ \rho_{j,l} & 1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ \rho_{j,l}^{K_j-1} & \cdots & \cdots & 1\end{pmatrix}_{K_j\times K_j}\equiv\frac{\tau_{j,l}^2}{1-\rho_{j,l}^2}AR1\left(\rho_{j,l}\right), \quad (3)$$

where $\tau_{j,l}^2$ is the conditional variance of $\beta_{jk,l}$ given $\beta_{j,k-1,l}$.

We should mention that our model uses the idea of spatially varying coefficients, which is not new [31]. However, accommodating this idea for use with high-volume and high-density epigenetic data under an ANOVA framework is novel. Also, the partial ACF plots from our data suggest that the first order autocorrelation dominates higher-order ones so that AR1 structure appears to be adequate in our applications. In other applications, one may consider second or higher order autoregressive models for the regression coefficients. However, in our other studies, we find that improvement from using higher-order models is small but the computing time may increase substantially.

3.2 Bayesian Framework

Below we describe our Bayesian approach for a specific gene *j*, and hence subscript *j* is omitted for notational brevity (e.g, $\mathbf{Y}$ represents $\mathbf{Y}_j$, $\boldsymbol{\beta}$ represents $\boldsymbol{\beta}_j$, K represent K_j, etc.). Let $\Theta=(\boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\rho}, \boldsymbol{\tau^2})$ be the collection of all (hyper)parameters, where $\boldsymbol{\tau^2}=\left(\tau_0^2,\ldots,\tau_L^2\right)$ and $\boldsymbol{\rho}=(\rho_0,\ldots,\rho_L)$ are introduced by the AR(1) covariance structure of $\overrightarrow{\beta}_l$s, as discussed in (2). Assuming that $\overrightarrow{\beta}_l$s, σ_k^2s, τ_l^2s and ρs are a priori independent for all l or k, the full probability model is given by

$$p\left(\boldsymbol{Y}|\Theta\right)\propto p\left(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\sigma^2}\right)p\left(\boldsymbol{\beta}|\boldsymbol{\tau^2},\boldsymbol{\rho}\right)\pi\left(\boldsymbol{\sigma^2}\right)\pi\left(\boldsymbol{\tau^2},\boldsymbol{\rho}\right)=\prod_{i=1}^{I}\prod_{k=1}^{K}\left\{\frac{1}{\sqrt{2\pi}\sigma_k}exp\left[-\frac{\left(Y_{ik}-\sum_{l=0}^{L}\beta_{k,l}X_{i,l}\right)^2}{2\sigma_k^2}\right]\right\}$$

$$\cdot\prod_{l=0}^{L}\left\{\frac{\sqrt{1-\rho_l^2}}{(2\pi)^{\frac{K}{2}}\left(\tau_l^2\right)^{\frac{K}{2}}}exp\left[-\frac{\left(1-\rho_l^2\right)\beta_{1,l}^2+\sum_{k=2}^{K}\left(\beta_{k,l}-\rho_l\beta_{k-1,l}\right)^2}{2\tau_l^2}\right]\right\}\cdot\prod_{k=1}^{K}\pi\left(\sigma_k^2\right)\cdot\prod_{l=0}^{L}\left[\pi\left(\tau_l^2\right)\pi\left(\rho_l\right)\right],$$

where π's are prior distributions. For all the variance components, we specify conjugate inverse gamma priors, that is, $\sigma_k^2\overset{iid}{\sim}IG\left(\alpha_\sigma,\gamma_\sigma\right)$, $\tau_l^2\overset{iid}{\sim}IG\left(\alpha_\tau,\gamma_\tau\right)$, where the hyperparameters are chosen to make the prior very vague, for example, IG(0.01,0.01). For correlation coefficients, it would be natural to consider noninformative uniform priors : $\rho_l\overset{iid}{\sim}unif\left(-1,1\right)$. Note that all the (hyper)prior distributions are proper, and so the posterior distribution $p(\Theta|\mathbf{Y})$, which is proportional to $p(\mathbf{Y}, \Theta)$, is proper.

We use Markov chain Monte Carlo (MCMC) to draw samples from $p(\Theta|\mathbf{Y})$. Under the conjugate priors for τ_l^2 and σ_k^2's, we can show that the full conditionals of $\overrightarrow{\beta}_l$'s, σ_k^2's and τ_l^2's, except for those of ρ_ls, are available in closed forms, and so can be sampled directly. Thus, we adopt a Gibbs sampler, in which values of ρ_l's can be drawn by a built-in Metropolis-Hastings (M-H) algorithm. Let $\Theta\theta$ denote the collection of all the parameters in Θ except for θ. The full conditional posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\sigma^2}$, $\boldsymbol{\rho}$, $\boldsymbol{\tau^2}$ are given below.

For $l = 0, \ldots, L$,

$$\overrightarrow{\beta}_l|\boldsymbol{Y}, \Theta\backslash\overrightarrow{\beta}_l \sim \mathrm{N}\left(\overrightarrow{b}_l, \mathbf{V}_l\right) \quad (4)$$

where

$$\begin{aligned} \mathbf{V}_l &= \left(\mathbf{B}_l^{-1}+\mathbf{C}_l^{-1}\right)^{-1} \\ \overrightarrow{b}_l &= \mathbf{V}_l\mathbf{C}_l^{-1}\overrightarrow{\psi}_l \\ \mathbf{C}_l &= \frac{\boldsymbol{\sigma^2}}{\sum_{i=1}^I x_{i,l}^2} = \frac{diag\left(\sigma_k^2\right)_{K\times K}}{\sum_{i=1}^I x_{i.l}^2} \\ \overrightarrow{\psi}_l &= \left(\frac{\sum_{i=1}^I\left(y_{ik}-\sum_{l'\neq l}\beta_{k,l'}x_{i,l'}\right)x_{i,l}}{\sum_{i=1}^I x_{i,l}^2}\right)_{k=1}^K \end{aligned} \quad (5)$$

For $k = 1, \ldots, K$,

$$\sigma_k^2|\boldsymbol{Y}, \Theta\backslash\sigma_k^2 \sim IG\left(\alpha_\sigma+\frac{I}{2}, \quad \frac{1}{2}\sum_{i=1}^I\left(y_{ik}-\sum_{l=0}^L\beta_{k,l}X_{i,l}\right)^2+\gamma_\sigma\right).$$

For $l = 0, \ldots, L$,

$$\tau_l^2|\boldsymbol{Y}, \Theta\backslash\tau_l^2 \sim IG\left(\alpha_\tau+\frac{K}{2}, \quad \frac{1}{2}\left[\left(1-\rho_l^2\right)\beta_{1,l}^2+\sum_{k=2}^K(\beta_{k,l}-\rho_l\beta_{k-1,l})^2\right]+\gamma_\tau\right),$$

and

$$\begin{aligned} \rho_l|\boldsymbol{Y}, \Theta\backslash\rho_l &\propto \sqrt{1-\rho_l^2}exp\left[-\frac{\left(1-\rho_l^2\right)\beta_{1,l}^2+\sum_{k=2}^K\left(\beta_{k,l}-\rho_l\beta_{k-1,l}\right)^2}{2\tau_l^2}\right] \\ &\propto \sqrt{1-\rho_l^2}exp\left[-\frac{\left(\sum_{k=2}^K\beta_{k,l}^2-\beta_{1,l}^2\right)\rho_l^2-2\sum_{k=2}^K\beta_{k,l}\beta_{k-1,l}\rho_l}{2\tau_l^2}\right] \end{aligned}$$

Note that due to the mathematical convenience of using the AR1 structure, sampling the regression parameters $\beta_{k,l}$s can be simplified from (4) and (5) without using any matrix inverse operation that may be very computational intensive for large K. That is, for $l = 0, \ldots, L$, and $k = 1, \ldots, K$,

$$\begin{aligned}
\beta_{k,l}|\boldsymbol{Y},\Theta\backslash\beta_{k,l} \sim\ & \mathrm{N}\left(b_{k,l},v_{k,l}\right);\\
b_{k,l}=\ & -\tfrac{\tilde{b}_{k,l}}{2\tilde{a}_{k,l}},\quad v_{k,l}=\tfrac{1}{2\tilde{a}_{k,l}},\\
\tilde{a}_{k,l}=\ & \tfrac{1}{2}\sum_{i=1}^{I}X_{i,l}^{2}/\sigma_{k}^{2}+\tfrac{1}{2}a'/\tau_{l}^{2},\\
\tilde{b}_{k,j}=\ & \sum_{i=1}^{I}X_{i,l}\left(\sum_{l'\neq l}\beta_{k,l'}X_{il'}-Y_{ik}\right)/\sigma_{k}^{2}+\tfrac{1}{2}b'/\tau_{l}^{2}
\end{aligned}$$

where

$$a'=\begin{cases}1 & \text{if } k=1 \text{ or } K\\ 1+\rho_{l}^{2} & \text{otherwies}\end{cases}$$

$$b'=\begin{cases}-2\rho_{l}\beta_{2,l} & \text{if } k=1\\ -2\rho_{l}\beta_{k-1,l} & \text{if } k=K\\ -2\rho_{l}\left(\beta_{k-1,l}+\beta_{k+1,l}\right) & \text{otherwise}\end{cases}$$

3.3 Posterior Inference

Once we obtain posterior draws from $p(\Theta|\mathbf{Y})$, we are primarily interested in statistical inference based on $\beta_{jk,l}$s or their linear functions. For example, in an one-way ANOVA experiment, suppose there are $L+1$ groups in total, of which the reference group is labeled by 0. For probe k in gene j, the mean epigenetic profile of the reference group is given by $\beta_{jk,0}$ and that of Group l is given by $\beta_{jk,0} + \beta_{jk,l}$ for $l = 1, \ldots L$. To compare Group l with the reference group in epigenetic profile, we first compute the posterior probability at the probe level using samples drawn from the posterior distribution, $q_{jk,l} \equiv Pr(|\beta_{jk,l}| > \delta |$ Data) for each probe in gene j, where δ is a cutoff value that can be chosen according to biological relevance in applications. For inference at the gene level (i.e., to infer which genes are epigenetically different between the two groups), we use the largest value of the posterior probabilities among the K_j probes to measure the significance of gene j, that is, $q_{j,l} \equiv max_{k=1}^{K_j} q_{jk,l}$. The genes having the largest $q_{j,l}$s are regarded as epigenetically different genes between Group l and the reference group. Here, our choice of measuring the significance of gene j is made because we are more interested in identifying large local jumps than stable changes across the entire promoter region of the gene in the motivating studies. Obviously, other choices, such as a p-value from some testing procedure or Bayes factor, could be chosen according to biological relevance in other applications. To compare any two nonreference groups (say l_1 and l_2, $l_1 \neq l_2 \neq 0$), we only need to define the posterior probability $q_{jk}(l_1, l_2) \equiv Pr(|\beta_{jk,l_1} - \beta_{jk,l_2}| > \delta |$ Data) at the probe level and probability $q_{jk}(l_1, l_2) \equiv Pr(|\beta_{jk,l_2}| - \beta_{jk,l_2}| < \delta |$ Data) at the level and then as proceed before at the gene level.

For a multi-way ANOVA experiment, we may define epigenetically changed genes accordingly, depending on what we are interested in (e.g., main effects, interaction effects, or epigenetic difference between two specific groups).

4 Simulation

To evaluate the performance in detecting epigenetically changed genes, we compared our proposed method (with spatially varying coefficients, labeled *SVC*) to three other methods, labeled *ANV, ANV-s, SAM* and *Tilemap*, respectively. The first is the regular ANOVA method that completely ignores the spatial dependence among probes, which fits the linear model (1) for each probe separately; the second is the ANOVA method applied to smoothed data using sliding windows of size 5; and the SAM (Significance Analysis of Microar-rays) method [40], where the SAM *t* statistic has been widely used to identify statistically significant genes in practice. The last is the Tilemap method [29] using the option of hidden Markov Models, which is commonly used for detecting epigenetic changes among groups. We conducted two sets of simulation studies, under one-way and two-way ANOVA settings, respectively. Within each set of studies, we first examined the performance using data generated from AR1 correlation structures. Then we generated data using patterns that are similar to what we observe in real data.

4.1 One-Way ANOVA

Settings for Simulation I1-I4—Here, we consider one-way ANOVA models, $y_{ijk} = \beta_{jk,0} + \beta_{jk,1} + X_i + \in_{ijk}$, where $\epsilon_{ijk} \overset{IID}{\sim} \mathrm{N}(0,1)$, $X_I = 0$ for the control group and 1 for the treatment group, $i = 1, \ldots, 6$, $j = 1, \ldots, 2000$, and $k = 1, \ldots 50$ (i.e., 3 replicates for each group; 2000 genes, each with 50 probes). Out of the 2000 genes, 80% are not epigenetically different between the two groups and the corresponding $\beta_{jk,1}$'s are set to 0.

We conducted four simulation studies (I1-I4) for one-way ANOVA models. In the first study I1, we assumed AR1 structures for the spatial correlation in $\overrightarrow{\beta}_{j,l}=\{\beta_{jk,l}\}_{k=1}^{50}$, $l = 0, 1$. We simulated $\overrightarrow{\beta}_{j,0}$ from $\mathrm{N}\left(\overrightarrow{0}, \frac{1}{1-\rho^2} AR1(\rho)\right)$ for all the genes. For the 20% epigenetically changed genes, we simulated the corresponding $\overrightarrow{\beta}_{j,1}$s from $\mathrm{N}\left(\overrightarrow{0}, \frac{1}{1-\rho^2} AR1(\rho)\right)$. We set ρ to be 0, 0.5 and 0.75.

In the next two simulation studies, we considered non-AR1 structures for the differences between the two groups; that is, $\beta_{jk,1}$s for those epigenetically different genes were generated to follow simple geometric patterns. Though the true spatial pattern within a gene remains unknown, we observed from real data that, for many genes, it was either first increasing and then decreasing or nearly flat in some regions. We therefore simulated data using triangle and rectangle patterns to roughly approximate such situations. In Simulation I2, we used the triangle pattern (the left panel of Figure 3), in which the middle probes (probes 16-35) have nonequa $\beta_{jk,1}$s with the peak height $h = 1.5$. In Simulation I3, we used the rectangle pattern (the right panel of Figure 3), in which the middle 30 probes have nonzero flat $\beta_{jk,1}$s with height $h = 1$. All the $\beta_{jk,0}$s were simulated as in Simulation I1 with $\rho = 0.5$.

In Simulation I4, we generated data from more complex spatial patterns to represent certain realistic situations. Environmental changes can induce different types of epigenetic alteration profiles. We considered four patterns, corresponding to distinct types of such alterations. For the purpose of illustration, we describe them in the context of chronic cocaine-induced

alterations in the methylation of lysine 4 of histone 3 (K4) (a transcription factor that plays a key role in the NAc reward pathway) (Figure 4 A-D). Figure 4(A) represents a gene at which cocaine blocked methylation: the gene is only methylated in saline treated mice (black line), but not in cocaine treated mice (red line). Figure 4(B) represents a gene with cocaine induced methylation: it is only methlated in the cocaine condition, but not in the saline condition. Figure 4(C) represents a gene where cocaine treatment decreases the methylation. Figure 4(D) represents a gene in which themethylation occurs at different sites in cocaine and saline conditions..

In the fourth study I4, we simulated 2000 genes again, 20% with real changes (5% for each of the four patterns above) and 80% without any cocaine induced epigenetic changes (Pattern E). In Pattern A and B, the true nonzero epigenetic profiles were both generated from 10 times the normal density function $f_N(\mu = 35, \sigma^2 = 9)$; in Pattern C, they were generated from $20 \times f_N(\mu = 35, \sigma^2 = 9)$ and $10 \times f_N(\mu = 35, \sigma^2 = 9)$, respectively; in Pattern D, the two curves were from $10 \times f_N(\mu = 15, \sigma^2 = 9)$ and $10 f_N(\mu = 35, \sigma^2 = 9)$; and in Pattern E, the true epigenetic profiles under the two conditions were all set to zero.

Results for I1-I4—For detection of epigenetically changed genes between the treatment and control groups, we drew receiver operation characteristics (ROC) curves for the proposed method based on $q_{j,1}$s, where $q_{j,1} \equiv max_{k=1}^{50} q_{jk,1}$ and $q_{jk.1} \equiv Pr\,(|\beta_{jk,1}|>\delta|\text{Data})$, as defined in Section 3.3. We set the cutoff δ to be 1. In our pre-liminary numerical experiments, we tried different values of δ (i.e., 0.2, 0.5, 1, 1.5, 2), and found selection of the cutoff was not crucial for the purpose of comparison since the ROC curves changed little as δ varies. For the Tilemap method, we used the largest posterior probability among all the probes of a gene to measure the probability of the gene with epigenetic changes. For the other three methods *ANV*, *ANV-s* and *SAM*, we used the smallest *p*–value of the corresponding test statistic (for testing $\beta_{jk,1} = 0$) among all the probes of a gene to measure the significance of the gene, and then drew ROC curves.

Figure 5 compares the five methods, *ANV*, *ANV-s, SAM, Tilemap* and *SVC* based on ROC curves under our one-way ANOVA settings. It is easy to see that *ANV* is the worst in all the settings, except for the first case (AR1 with $g=r = 0$), where *ANV* outperforms *ANV-s*. This indicates that simply smoothing data over adjacent probes would hurt the performance in detection when data have no autocorrelation. For data with AR1 correlation structures, as ρ increases, the performance of *ANV-s* and *Tilemap* get better while that of *SAM* becomes worse. The performance of *SVC* seems to be quite stable at different ρ levels and outperforms the others in general. In the next two studies (I2-I3) involving simple patterns of the main effect (i.e, triangle and rectangle), we observe the order *SVC* > *SAM* > *Tilemap,ANV-s*>*ANV* in performance consistently. In the last study I4, which perhaps is the one simulated to resemble the realistic situation most, *SAM* performs worse than *ANV-s, Tilemap* and *SVC*, though it is still much better than *ANV*. The methods *ANV-s* and *SVC,* come across each other when the false positive rate is around 0.3. However, *SVC* is better than *ANV-s* in the left lower corner, which is the region of primary interest for detection since researchers often want to control the false positive rate to be reasonably low. In

addition, the performance of *SVC* appears to be better than or comparable to that of *Tilemap* in Study I4.

Table 1 reports the computed mean squared errors (MSEs, averaged over all the probes) for estimating the regression parameters of the one-way ANOVA models using *ANV*, *ANV-s* and *SVC*. For data with the AR1 correlation structures, *SVC* can improve *ANV* in estimating both the intercept and main effect for all the three ρ values. For the case $\rho = 0$, the gain is from information borrowing among probes through a hierarchical Bayes setup (i.e., Bayesian shrinkage). For $\rho > 0$, it is from both information borrowing and spatial smoothing. By contrast, when estimating the intercept, *ANV-s* improves *ANV* only for $\rho = 0.75$ (moderately high), and it is worse than *ANV* for $\rho = 0$ or 0.5, meaning that smoothing data among adjacent probes may hurt the performance in estimation when the autocorrelation does not exist or is weak. When estimating the main effect, *ANV-s* improves *ANV* for all the three ρ values; but the gain is not as big as that of *SVC*. For $\rho = 0$, the gain of *ANV-s* over *ANV* in estimating the main effect (0.300 vs. 0.662 in MSE) is from those non-epigenetically changed genes (0.141 vs. 0.663) since these genes have zero-valued $\beta_{jk,1}$'s and spatial smoothing is helpful for them. This offsets the loss from smoothing data over the probes of epigenetically changed genes with no autocorrelation (0.933 vs. 0.666). For data with non-AR1 structures, again, the proposed method is consistently the best in both estimating the intercept and main effect, showing that spatial smoothing (even via a simple AR1 model for rough approximation) and strength borrowing (via Bayesian shrinkage) among probes can greatly improve the efficiency of estimation. Note that *ANV-s* improves estimation of the main effect, but not as much as *SVC*.

4.2 Two-way ANOVA

Settings for Simulation II1-II3—Now we proceed to consider two-way ANOVA models, $y_{ijk} = \beta_{jk,0} + \beta_{jk,1} X_{i,1}, + \beta_{jk,2} X_{i,2} + \beta_{jk,3} X_{i,1} X_{i,2} + \in_{ijk}$, where $\epsilon_{ijk} \overset{IID}{\sim} \mathrm{N}(0,1)$, $X_{i,1}$ and $X_{i,2}$ represent two factors, each with two levels (e.g., $X_{i,1} = 0$ for the normal condition and 1 for the deceased condition; and $X_{i,2} = 0$ for the control and 1 for the treatment); $i = 1, \ldots, 12$, $j = 1, \ldots, 2000$, and $k = 1, \ldots 50$ (i.e., 3 replicates for each of the four group; 2000 genes, each with 50 probes). Suppose our primary interest is to study the interaction between $X_{i,1}$ and $X_{i,2}$; that is, we want to first infer whether any $\beta_{jk,3}$ is significantly different from 0 so that we can further infer, for which genes, the two factors interact in the epigenetic profiles. Out of the 2000 genes, 80% are assumed to have no interaction and the corresponding $\beta_{jk,3}$'s are set to $\overrightarrow{0}$.

We conducted three simulation studies (II1-II3) for two-way ANOVA models. In the first study (II1), we assumed AR1 correlation structures for all the parameters. We simulated $\overrightarrow{\beta}_{j,0}$, $\overrightarrow{\beta}_{j,1}$ and $\overrightarrow{\beta}_{j,2}$ from $\mathrm{N}\left(\overrightarrow{0}, \frac{1}{1-\rho^2} AR1(\rho)\right)$ for all the genes. For the 20% genes with interactions, we simulated the corresponding $\overrightarrow{\beta}_{j,3}$s from $\mathrm{N}\left(\overrightarrow{0}, \frac{1}{1-\rho^2} AR1(\rho)\right)$. We set $\rho = 0.5$ again.

In the next two studies, we considered non-AR1 spatial patterns for the nonzero interactions of the 20% genes. In Simulation II2, we used the triangle pattern in which the middle probes

(probes 16-35) have nonequal $\beta_{jk,3}$s with the peak of height $h = 1.25$ or 1.5. In Simulation II3, we used the rectangle pattern, in which the middle 30 probes have flat $\beta_{jk,3}$s with height $h = 0.75, 1, 1.25$. All the other parameters were simulated as in Simulation II1.

Results for II1-II3—Table 2 reports the computed MSEs (averaged over all the probes) for estimating regression parameters of the two-way ANOVA models using *SVC*, *ANV* and *ANV-s*. Note that the two main effects β_1 and β_2 have the same theoretical MSE since they were simulated in the same way under a balanced design. So we combined the results for β_1 and β_2 by taking the average. For estimating both the main effects and interaction, we find that the performance has the order *ANV* < *ANV*-s< *SVC* in all the two-way settings. As to the estimation of the intercept, we can observe *ANV-s*< *ANV* < *SVC*. Clearly, *SVC* is the winner, especially in estimating the interaction, which has substantial improvement over *ANV* and *ANV-s*.

For identifying genes with interactions, Figure 6 reports ROC curves under our two-way settings to compare the methods *ANV*, *ANV-s*, and *SVC* based on the posterior probabilities or p-values related to $\beta_{jk,3}$s, defined similarly to those for $\beta_{jk,1}$s in Section 4.1. Here, *SAM* is excluded for comparison since it is not directly applicable to detect the interaction effects. Again, the order in performance is given by *ANV* < *ANV*-s< *SVC* in all the studies. In II2 and II3 involving the triangle and rectangle patterns, we can observe that as *h* increases (i.e., *h* reflects the effect size of the interaction term), the performance of the proposed method *SVC* gets better, approaching the upper right corner. By contrast, *ANV*-s is not as sensitive to *h* as *SVC*, though it gets a bit better, too. It appears that *ANV* does not change much with *h*, and the performance in detecting the interaction is poor, only slightly better than the method of purely random selection.

5 Applications

5.1 Cocaine addiction study

As mentioned in Section 2.1, the cocaine addiction study used a one-way ANOVA design. The histone H3 methylation data were measured by Nimblegen MM8 Mouse Promoter arrays and normalized by model-based analysis of two-color arrays (MA2C) software[41]. Using our proposed method, we can identify 26 genes (summarized in Table 3) whose methylation levels were decreased after the cocaine treatment with posterior probabilities greater than 0.90. Using the same criterion, Tilemap with the option of Hidden Markov models only identified 3 genes and Tilemap with the option of moving average identified no genes when controlling FDR less than 50%. Among these 26 genes identified by our proposed method, several of them are biologically interesting and cannot be identified using the ordinary ANOVA or SAM t test. For example, chronic ciliary neurotrophic factor (CNTF) is a gene that plays an important role in cocaine addiction [42]. Using the SAM t test, the CNTF gene ranks 843, while using our proposed method, it ranks 14. In this example our proposed method can detect the methylation changes in CNTF which would be likely missed by the SAM method. Furthermore, Figure 7, the promoter plot of the Hoxb3 gene, shows that our proposed method can identify a decrease in methylation over a region of several adjacent probes, while the results from the SAM method are much more variable.

Gene expression data in [34] indicate that Hoxb3 expression is increased after the cocaine treatment, which is consistent with the cocaine-induced epigenetic changes found by our method. In addition, our proposed method has identified the methylation changes in several cocaine addiction associated genes, such as the methylation decrease in CD44 (the cell differentiation antigen) [43] and the methylation increase and decrease in different regions of Gria4 (the gene encoding the glutamate receptor subunit GluR4) [44].

5.2 Depression study

Recall that the depression study, as described in Section 2.2, used a two-way ANOVA design. The histone H3 methylation data were also measured by Nimblegen MM8 Mouse Promoter arrays and normalized by MA2C. We applied the proposed method to the dataset and identified 22 genes, all with the posterior probabilities greater than 0.99 for significant interactions. Although evaluating the results in real data applications is difficult, we find here that the two genes with the highest posterior probabilities of significant interactions, Nxph1 and Accn2, are of great interest. They are the top two ranked by our method, while they rank much lower using the other methods (e.g., Nxph1 and Accn2 rank 157 and 407, respectively, by the regular ANOVA approaches). The Nxph1 gene encodes neurexophilin 1 α-neurexins; and it has been shown in several studies to be associated with neuroticism and severe psychiatric disorders such as suicidal behavior (e.g., [45]). In this study, Figure 8 shows that, for the Nxph1 gene, the mean diffierences in H3 methylation between the imipramine-treated mice and the saline-treated mice are totally different for defeated mice (red line) and control mice (black line), with increased H3 methylation seen under defeat conditions. There exists a strong interaction between the imipramine and social defeat exposures with respect to the epigenetic profiles of Nxph1, which indicates the gene may be one of the relevant targets for treatment. The Accn2 gene (which encodes the amiloride-sensitive cation channel 2, neuronal) is ubiquitously expressed in the nervous system. Our data show that H3 methylation at this gene increases after imipramine treatment under defeat conditions. Links between Accn2 and anxiety-related behavior have been found in mouse models [46]. Further, a recent study has shown that Accn2 knockout mice exhibit reduced depression-related behavior [47] so that inhibition of Accn2 could be a novel strategy for the treatment of depression [48].

6 Discussion

Motivated by important epigenetic studies in drug addiction and depression fields, we have developed ANOVA models with spatially varying coeffients, to analyze epigenetic profiling data with spatial correlation existing in biological effiects of interest instead of measurement errors. Through simulation studies, we have shown that the proposed method *SVC* can offer better efficiency in both parameter estimation and gene detection than the competing methods. This is not surprising. As mentioned before, the regular ANOVA method *(ANV)* analyzes data for different probes of the same gene in a completely separate way, hence no strength borrowing or spatial smoothing is done here. *ANV-s* offers one step forward from *ANV*, by smoothing the observed data 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 ignores the spatial correlation. By contrast, our method *SVC* does both. First of all, through a Bayesian hierarchical approach, *SVC* can borrow information from the other probes of the same gene, which is much better than modeling each probe independently. Secondly, *SVC* can conduct spatial smoothing by modeling the spatial dependence among neighboring probes explicitly through the AR1 correlation structure. Though AR1 can only provide a rough approximation to real spatial patterns, *SVC* is clearly better than regular hierarchical Bayes models using an identity correlation matrix in (2), the latter of which offers the advantage of information borrowing through shrinkage but no spatial modeling of the probes. Through real data examples, we have shown that our method can identify epigenetic changes that might be overlooked by other methods and also provide biologically meaningful results.

In our applications, we used the AR1 model to account for the spatial patterns among neighboring probes. The AR1 model assumes the measurements from adjacent probes are autocorrelated, i.e., the adjacent probes tend to have more similar values than would be expected by random chance. The histograms of the ACF1 in Figure 2 demonstrate that there is positive correlation between adjacent probes, and the means of ACF1 are 0.25 and 0.24 in saline and cocaine-treated conditions, respectively. In addition, we also plotted the histograms of partial autocorrelation of lag 2 (PACF2), to check whether the first-order auto-correlation structure is adequate to capture the spatial patterns. The PACF2 is the correlation in the measurements between a probe and its second neighbor with the linear dependence of its first neighbor removed. The PACF2 histograms (not shown here) indicate that the partial autocorrelations of lag 2 are symmetric around 0, with means equal to −0.02 and −0.03 in saline and cocaine-treated conditions respectively. So the AR1 model is adequate in our applications. However, depending on real data, the higher-order autocorrelation structure, such as AR2, might fit other applications better. In those cases, our model can be easily extended to accommodate the higher-order structures.

A typical characteristic of high throughput data is the existence of significant spatial correlation caused by underlying biological processes. Also, ANOVA designs are frequently used in many scientific experiments to avoid bias in comparison. With (or even without) slight modification or straightforward extension, our method could be generally purposed and provide a useful tool to analyze spatially correlated data from biomedical studies using ANOVA designs. For example, our method, motivated by the applications using promoter arrays, can be applied to whole genome tilling arrays, too. In this case, we would fit the regression model in (1) for all the probes in each chromo-some instead of probes in each promoter region. In principle, the model could also be generalized to ChIP-seq data, if we use Poisson or Negative Binomial distribution instead of normal distribution to model the involved count data. However, due to enormous amounts of data generated from ChIP-Seq experiments, the computation of Bayesian MCMC could be much more intensive for ChIP-Seq data. Therefore, further development is needed to modify the proposed model and algorithm to analyze ChIP-seq data efficiently.

References

1. Qiu J. Epigenetics: unfinished symphony. Nature. 2006; 441(7090):143–5. [PubMed: 16688142]

2. Jones RS. Epigenetics: reversing the 'irreversible'. Nature. 2007; 450(7168):357–9. [PubMed: 18004369]
3. Bird A. Perceptions of epigenetics. Nature. 2007; 447(7143):396–8. [PubMed: 17522671]
4. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. Science. 2000; 290(5500):2306–9. [PubMed: 11125145]
5. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 2001; 409(6819):533–8. [PubMed: 11206552]
6. 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]
7. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
8. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods. 2008; 5:829–834. [PubMed: 19160518]
9. Tuteja G, White P, Schug J, Kaestner KH. Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Res. 2009; 37:e113. [PubMed: 19553195]
10. Mo Q. A fully Bayesian hidden Ising model for ChIP-seq data analysis. Biostatistics. 2011
11. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. PNAS. 2006; 103:12457–12462. [PubMed: 16895995]
12. Buck MJ, Nobel AB, Lieb JD. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol. 2005; 6(11):R97. [PubMed: 16277752]
13. Reiss DJ, Facciotti MT, Baliga NS. Model-based deconvolution of genome-wide DNA binding. Bioinformatics. 2008; 24:396–403. [PubMed: 18056063]
14. Li W, Meyer C, Liu X. a hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics 2005. 2005; 21(suppl I): 274–282.
15. Humburg P, Bulger D, Stone G. Parameter estimation for robust HMM analysis of ChIP-chip data. BMC Bioinformatics. 2008; 9:343. [PubMed: 18706106]
16. Wang X, Zang M, Xiao G. Epigenetic change detection and pattern recognition via Bayesian hierarchical hidden Markov models. Stat Med. 2012 [Epub ahead of print].
17. Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. High-resolution computational models of genome binding events. Nature Biotechnology. 2006; 21:963–970.
18. Keles S. Mixture modeling for genome-wide localization of transcription factors. Biometrics. 2007; 63:10–21. [PubMed: 17447925]
19. Gottardo R, Li W, Johnson WE, Liu XS. A flexible and powerful bayesian hierarchical model for chip-chip experiments. Biometrics. 2008; 64(2):468–78. [PubMed: 17888037]
20. Wu M, Liang F, Tian Y. Bayesian modeling of ChIP-chip data using latent variables. BMC Bioinformatics. 2009; 10:352. [PubMed: 19857265]
21. Mo Q, Liang F. Bayesian modeling of chip-chip data through a high-order ising model. Biometrics. 2010; 66(4):1284–1294. doi:10.1111/j.1541-0420.2009.01379.x. [PubMed: 20128774]
22. Mo Q, Liang F. A hidden Ising model for ChIP-chip data analysis. Bioinformatics. 2010; 26:777–783. [PubMed: 20110277]
23. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature Biotechnology. 2002; 20:835 – 839.
24. Zheng M, Barrera LO, Ren B, Wu YN. ChIP-chip: data, model, and analysis. Biometrics. 2007 Sep. 2007; 63:787–796.

25. Pan W, Wei P, A K. A parametric joint model of DNA-protein binding, gene expression and DNA sequence data to detect target genes of a transcription factor. Pacific Symposium on Biocomputing. 2008:465–476. NIL. [PubMed: 18229708]
26. Wei P, Pan W. Incorporating gene functions into regression analysis of dna-protein binding data and gene expression data to construct transcriptional networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2008; 5:401–415. [PubMed: 18670043]
27. 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]
28. Xu H, Wei CL, Lin F, Sung WK. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics. 2008; 24(20):2344–2349. [PubMed: 18667444]
29. Ji H, Wong W. Tilemap: Create chromosomal map of tiling array hybridizations. Bioinformatics. 2005; 18:3629–3636. [PubMed: 16046496]
30. Taslim C, Wu J, Yan P, Singer G, Parvin J, Huang T, Lin S, Huang K. Comparative study on ChIP-seq data: normalization and binding pattern characterization. Bioinformatics. 2009; 25(18):2334–2340. [PubMed: 19561022]
31. Gelfand AE, KIM HJ, SIRMANS CF, BANERJEE S. Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association. 2003; 98:387–396.
32. LaPlant Q, Vialou V, Covington HE 3rd, Dumitriu D, Feng J, Warren BL, Maze I, Dietz DM, Watts EL, Iniguez SD, Koo JW, Mouzon E, Renthal W, Hollis F, Wang H, Noonan MA, Ren Y, Eisch AJ, Bolanos CA, Kabbaj M, Xiao G, Neve RL, Hurd YL, Oosting RS, Fan G, Morrison JH, Nestler EJ. Dnmt3a regulates emotional behavior and spine plasticity in the nucleus accumbens. Nat Neurosci. 2010; 13(9):1137–43. [PubMed: 20729844]
33. 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. J Neurosci. 2009; 29(24):7820–32. [PubMed: 19535594]
34. Renthal W, Kumar A, Xiao G, Wilkinson M, Covington r H E, Maze I, Sikder D, Robison AJ, LaPlant Q, Dietz DM, 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):335–48. [PubMed: 19447090]
35. Nestler EJ, Aghajanian GK. Molecular and cellular basis of addiction. Science. 1997; 278(5335): 58–63. [PubMed: 9311927]
36. Tsankova N, Renthal W, Kumar A, Nestler EJ. Epigenetic regulation in psychiatric disorders. Nat. Rev. Neurosci. 2007; 8:355–367. [PubMed: 17453016]
37. Renthal W, Nestler EJ. Epigenetic mechanisms in drug addiction. Trends Mol Med. 2008; 14:341–350. [PubMed: 18635399]
38. Tsankova NM, Berton O, Renthal W, Kumar A, Neve RL, Nestler EJ. Sustained hippocampal chromatin regulation in a mouse model of depression and antidepressant action. Nat. Neurosci. 2006; 9:519–525. [PubMed: 16501568]
39. Berton O, McClung CA, Dileone RJ, Krishnan V, Renthal W, Russo SJ, Graham D, Tsankova NM, Bolanos CA, Rios M, Monteggia LM, Self DW, Nestler EJ. Essential role of BDNF in the mesolimbic dopamine pathway in social defeat stress. Science. 2006; 311:864–868. [PubMed: 16469931]
40. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U.S.A. 2001; 98:5116–5121. [PubMed: 11309499]
41. 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 Biol. 2007; 8(8):R178. [PubMed: 17727723] Song; Johnson, Jun S.; Zhu, W Evan; Zhang, Xiaopeng; Li, Xinmin; Manrai, Wei; Liu, Arjun K.; Chen, Jun S.; Liu, Runsheng; Shirley, X. 1R01 HG004069-01/HG/NHGRI NIH HHS/United States 1U01 HG004270-01/HG/NHGRI NIH HHS/United States R01 HG004069-02/HG/NHGRI NIH HHS/United States Research Support, N.I.H., Extramural England Genome biology. Genome Biol. 2007; 8(8):R178. [PubMed: 17727723]

42. Berhow MT, Hiroi N, Kobierski LA, Hyman SE, Nestler EJ. Influence of Cocaine on the JAK-STAT Pathway in the Mesolimbic Dopamine System. J. Neurosci. 1996; 16(24):8019–8026. [PubMed: 8987828]

43. Mash DC, ffrench Mullen J, Adi N, Qin Y, Buck A, Pablo J. Gene expression in human hippocampus from cocaine abusers identifies genes which regulate extracellular matrix remodeling. PLoS One. 2007; 2(11):e1187. [PubMed: 18000554]

44. Hemby SE, Horman B, Tang W. Differential regulation of ionotropic glutamate receptor subunits following cocaine self-administration. Brain Res. 2005; 1064(1-2):75–82. [PubMed: 16277980]

45. van den Oord EJ, Kuo PH, Hartmann AM, Webb BT, Moller HJ, Hettema JM, Giegling I, Bukszar J, Rujescu D. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. Arch. Gen. Psychiatry. 2008; 65:1062–1071. [PubMed: 18762592]

46. Wemmie JA, Coryell MW, Askwith CC, Lamani E, Leonard AS, Sigmund CD, Welsh MJ. Overexpression of acid-sensing ion channel 1a in transgenic mice increases acquired fear-related behavior. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(10):3621–3626. doi:10.1073/pnas.0308753101. [PubMed: 14988500]

47. Coryell MW, Wunsch AM, Haenfler JM, Allen JE, Schnizler M, Ziemann AE, Cook MN, Dunning JP, Price MP, Rainier JD, Liu Z, Light AR, Langbehn DR, Wemmie JA. Acid-sensing ion channel-1a in the amygdala, a novel therapeutic target in depression-related behavior. J. Neurosci. 2009; 29(17):5381–5388. doi:10.1523/JNEUROSCI.0360-09.2009. [PubMed: 19403806]

48. Flight MH. Mood disorders: Channel inhibitor shows antidepressant potential. Nat Rev Drug Discov. 2009; 8(7):540–540.

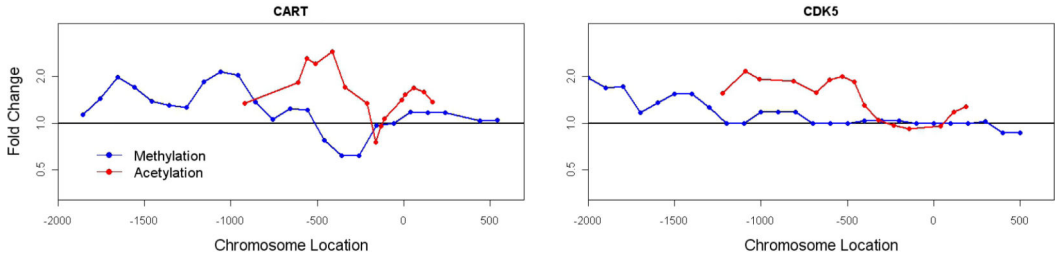

Figure 1.

Cocaine-induced epigenetic changes including H4 acetylation and H3methylation at the promoter regions of (A) Cart and (B) Cdk5 genes (reproduced from [34])

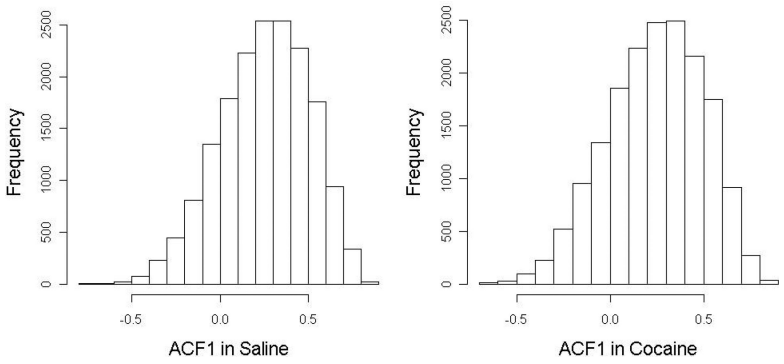

Figure 2.

Cocaine addiction study: the histograms of the ACF1 of methylation levels in the promoter regions for genes in saline (left panel) and cocaine (right panel) treated mice.

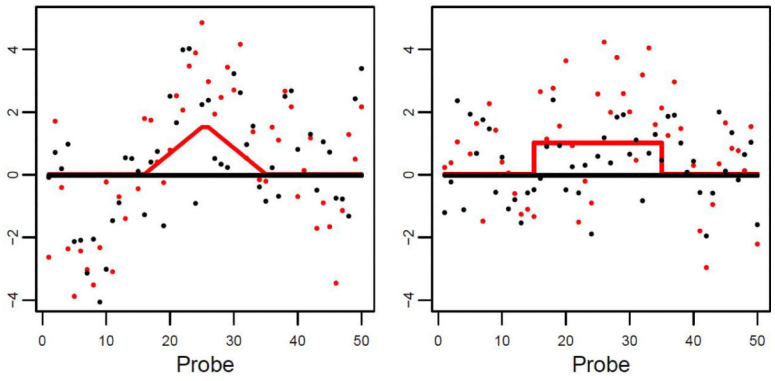

Figure 3.

Illustration of the triangle and rectangle patterns in Simulation I2-I3 under one-way ANOVA. The red line represents the true epigenetic changes (between the treatment and control groups) and red dots represent the simulated data (truth plus error term) for epigenetically changed genes, while the black line and dots are for genes not epigenetically changed.

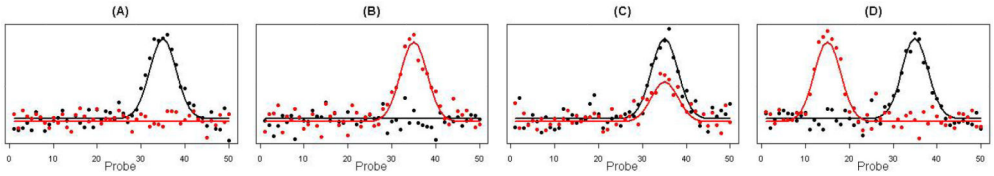

Figure 4.

Illustration of four different types of epigenetic changes in Simulation I4 under one-way ANOVA. Black and red lines represent the true epigenetic profiles under the saline (control) condition and the cocaine (treatment) condition, respectively; while black dots and red dots represent the simulated data (truth plus error terms) under saline and cocaine conditions respectively.

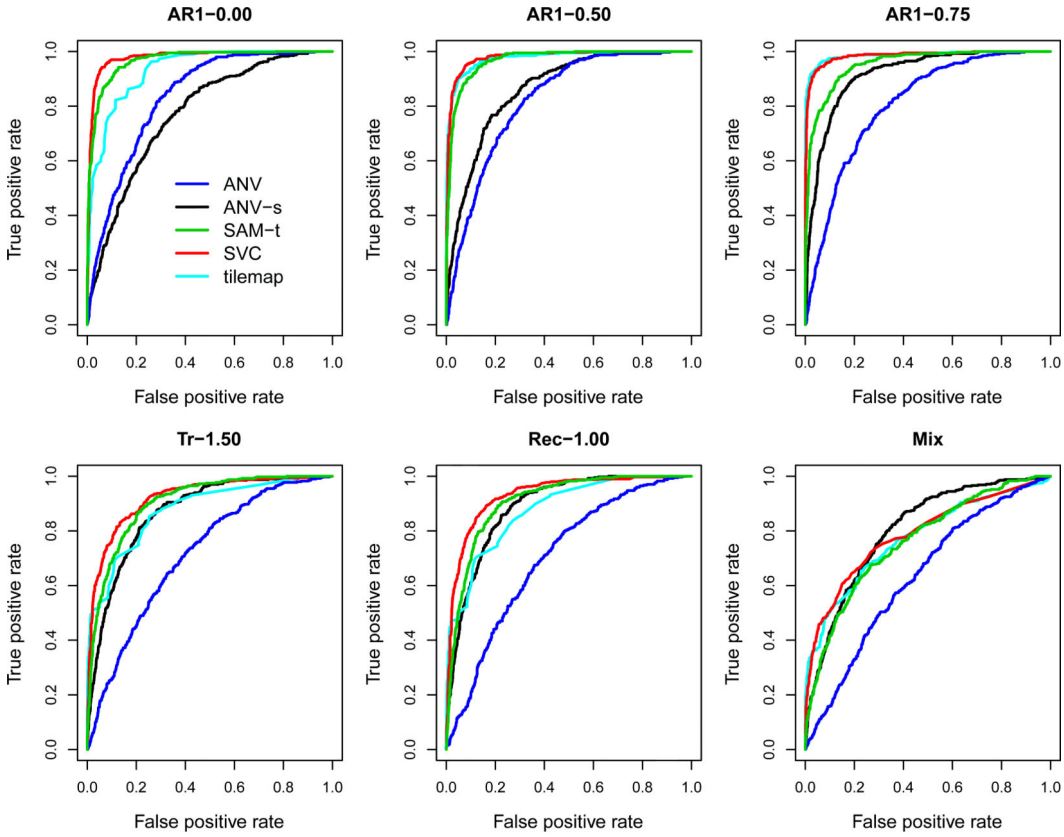

Figure 5.

One-way ANOVA settings: ROC curves for detecting epigenetically changed genes using *ANV* (regular ANOVA), *ANV-s* (ANOVA with smoothed *Y*), *SAM, SVC* (the proposed method) and *Tilemap*. "AR1" represents the first-order autocorrelation and the number after "-" indicates the value of the correlation ρ. "Tr" represents triangle and "Re" represents rectangle, and the number after "-" indicates the height. "Mix" represents the mixed five patterns A-E in I4.

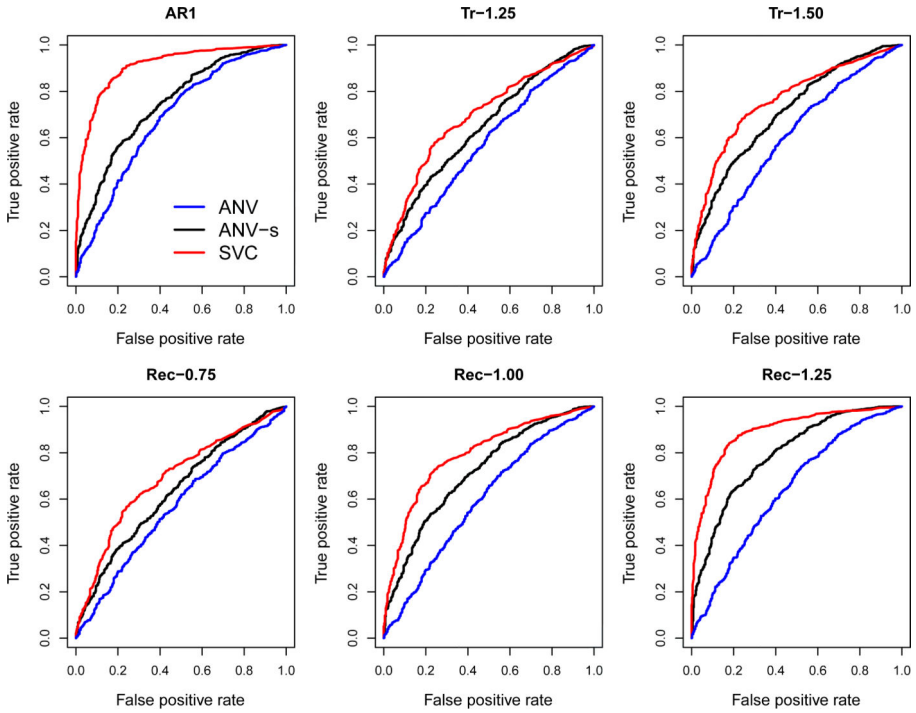

Figure 6.

Two-way ANOVA settings: ROC curves for detecting genes with interactions using *ANV* (regular ANOVA), *ANV-s* (ANOVA with smoothed *Y*), *SVC* (the proposed method). “AR1” represents the first-order autocorrelation with $\rho = 0.5$. “Tr” represents triangle and “Rec” represents rectangle, and the number after “-” indicates the height.

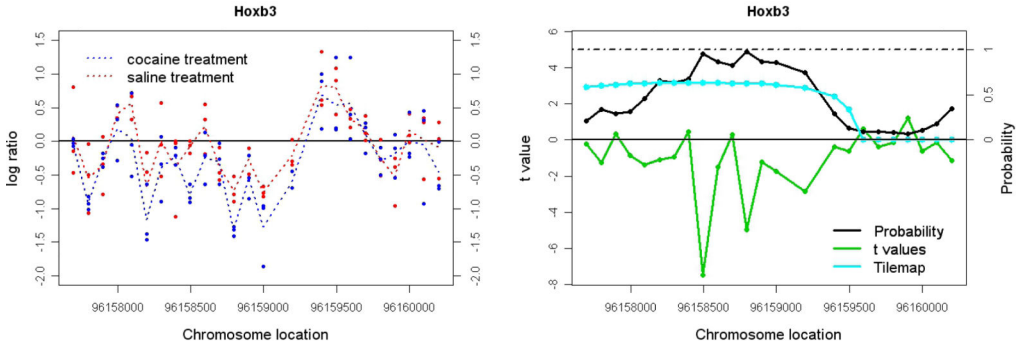

Figure 7.

Cocaine addiction study: promoter plot for Hoxb3. In the left panel, the red and blue dots represent the histone modifications for triplicates in the saline and cocaine treatment, respectively, while the red and blue dotted lines plot the mean values correspondingly. In the right panel, the black line plots the posterior probability of decreasing the methylation level after the cocaine treatment, the green line plots the SAM t values and the light blue line represents the posterior probability from Tilemap.

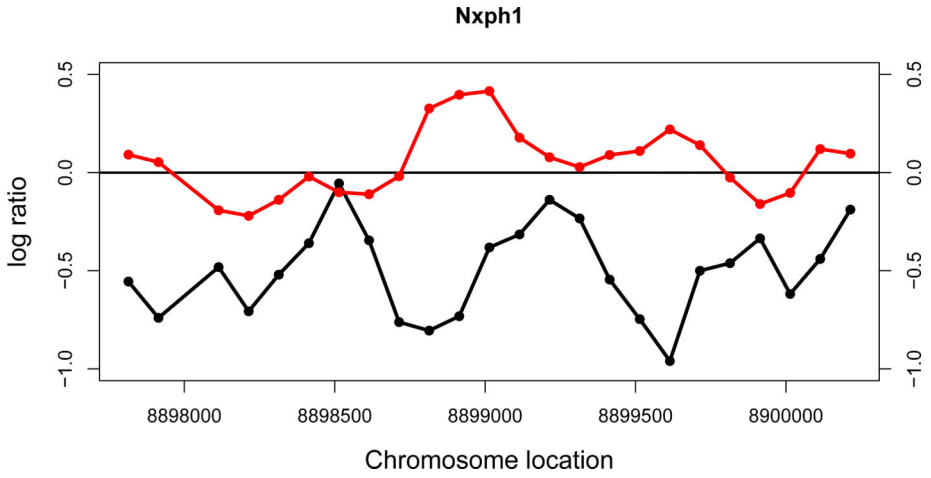

Figure 8.

Depression study: the mean differences in H3 methylation between imipramine- and saline-treated mice. The black line represents the difference for control animals, and the red line represents the difference for defeated animals.

Table 1

One-way ANOVA settings: MSEs for estimating regression parameters using *ANV* (regular ANOVA), *ANV-s* (ANOVA with smoothed *Y*) and *SVC* (the proposed method). “AR1” represents the first-order autocorrelation and the number after “-” indicates the value of the correlation ρ. “Tr” represents triangle and “Rec” represents rectangle, and the number after “-” indicates the height. “Mix” represents the mixed five patterns A-E in I4.

		Intercept			Main Effect		
Sim. #	**Pattern**	**ANV**	***ANV-s***	***SVC***	***ANV***	***ANV-s***	***SVC***
I1	AR1-0	0.332	0.856	0.209	0.662	0.300	0.124
	AR1-0.5	0.333	0.503	0.234	0.663	0.229	0.111
	AR1-0.75	0.333	0.284	0.276	0.666	0.184	0.090
I2	Tr-1.5	0.332	0.502	0.229	0.663	0.140	0.062
I3	Rec-1	0.332	0.504	0.230	0.660	0.142	0.062
I4	Mix	0.330	0.070	0.040	0.660	0.140	0.050

Table 2

Two-way ANOVA settings: MSEs for estimating regression parameters using *SVC* (the proposed method), *ANV* (regular ANOVA) and *ANV-s* (ANOVA with smoothed *Y*). "AR1" represents the first-order autocorrelation and the number after "-" indicates the value of the correlation ρ. "Tr" represents triangle and "Rec" represents rectangle, and the number after "-" indicates the height.

		Intercept			Main Effects			Interaction		
	Pattern	***ANV***	***ANV-s***	***SVC***	***ANV***	***ANV-s***	***SVC***	***ANV***	***ANV-s***	***SVC***
II1	AR1-0.5	0.50	0.54	0.31	1.00	0.64	0.38	1.98	0.76	0.49
II2	Tr-1.25	0.50	0.54	0.31	1.00	0.64	0.35	1.99	0.41	0.19
	Tr-1.5	0.50	0.54	0.31	0.99	0.64	0.36	1.98	0.42	0.22
II3	Rec-0.75	0.50	0.54	0.31	0.99	0.64	0.35	1.98	0.43	0.18
	Rec-1	0.50	0.53	0.31	1.00	0.64	0.36	1.99	0.43	0.23
	Rec-1.25	0.50	0.54	0.32	1.00	0.64	0.37	1.99	0.45	0.27

Stat Appl Genet Mol Biol. Author manuscript; available in PMC 2016 May 13.

Table 3

A list of the 26 genes with posterior probabilites greater than 0.90 in the proposed method

SEQ_ID	accession	Name	Probability
chr7:29890930-29893430	NM_009944	Cox7a1	0.990
chr9:62473866-62476366	BC058716	Itga11	0.980
chr16:88639987-88642487	NM_010671	Krtap13	0.970
chr11:96157698-96160358	NM_010458	Hoxb3	0.970
chr9:7556459-7558959	NM_008611	Mmp8	0.960
chr17:37754642-37757142	NM_146831	Olfr133	0.960
chr11:79319714-79322214	NM_019409	Omg	0.960
chr2:86395165-86397665	NM_207674	Olfr1082	0.950
chr19:53651758-53654258	NM_007790	Cspg6	0.950
chr5:65771465-65773965	NM_133697	1110003E01Rik	0.941
chr17:36921935-36924435	NM_001011721	Olfr102	0.941
chr4:119133663-119136163	NM_008190	Guca2a	0.931
chr5:27378179-27380679	NM_207282	B930011P16Rik	0.931
chr2:87170372-87172872	NM_146348	Olfr1121	0.931
chr19:12845472-12847972	NM_053007	Cntf	0.931
chr16:26904707-26907207	NM_001013761	Gm606	0.931
chr9:66847857-66850357	NM_024427	Tpm1	0.931
chr2:102702004-102704504	NM_001039150	Cd44	0.921
chr4:62009729-62013375	NM_021498	Pole3	0.921
chr4:43651034-43653534	BC042470	Npr2	0.921
chr2:11621388-11624200	NM_008358	Il15ra	0.911
chr11:98855349-98857849	NM_010517	Igfbp4	0.911
chr9:4795642-4798142	NM_019691	Gria4	0.901
chr17:37197904-37200404	NM_146287	Olfr114	0.901
chr19:44384636-44387136	NM_183216	Scd4	0.901
chr9:59548114-59550614	NM_027838	Senp8	0.901