BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 777–783
doi:10.1093/bioinformatics/btq032

Gene expression

Advance Access publication January 28, 2010

A hidden Ising model for ChIP-chip data analysis

Qianxing Mo[1,*] and Faming Liang[2]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10065 and [2]Department of Statistics,Texas A&M University, College Station, TX 88343, USA

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Chromatin immunoprecipitation (ChIP) coupled with tiling microarray (chip) experiments have been used in a wide range of biological studies such as identification of transcription factor binding sites and investigation of DNA methylation and histone modification. Hidden Markov models are widely used to model the spatial dependency of ChIP-chip data. However, parameter estimation for these models is typically either heuristic or suboptimal, leading to inconsistencies in their applications. To overcome this limitation and to develop an efficient software, we propose a hidden ferromagnetic Ising model for ChIP-chip data analysis.

Results: We have developed a simple, but powerful Bayesian hierarchical model for ChIP-chip data via a hidden Ising model. Metropolis within Gibbs sampling algorithm is used to simulate from the posterior distribution of the model parameters. The proposed model naturally incorporates the spatial dependency of the data, and can be used to analyze data with various genomic resolutions and sample sizes. We illustrate the method using three publicly available datasets and various simulated datasets, and compare it with three closely related methods, namely TileMap HMM, tileHMM and BAC. We find that our method performs as well as TileMap HMM and BAC for the high-resolution data from Affymetrix platform, but significantly outperforms the other three methods for the low-resolution data from Agilent platform. Compared with the BAC method which also involves MCMC simulations, our method is computationally much more efficient.

Availability: A software called iChip is freely available at http://www.bioconductor.org/.

Contact: moq@mskcc.org

Received on October 26, 2009; revised on January 4, 2010; accepted on January 22, 2010

1 INTRODUCTION

Identification of the sites of interactions between proteins and genomic DNA is one of the most important topics in biological research. The emergence of high-density tiling microarrays has made it possible to create high-resolution global maps of the *in vivo* interactions between specific proteins and genomes. Tiling arrays use millions of short oligonucleotide probes with various resolutions to interrogate genomic regions of interest. Chromatin immunoprecipitation (ChIP)-chip analysis has become a powerful technique used to determine the binding sites of DNA-binding proteins at a genome-wide basis (Cawley *et al.*, 2004; Boyer *et al.*, 2005).

It is a challenge to analyze ChIP-chip data, due to the huge amount of probes, the small number of replicates and various noises introduced in the experiments. The most important feature of ChIP-chip data is the neighboring dependency of the probe intensities. Quite a few of existing methods in the literature make efforts to incorporate the dependence structure of ChIP-chip data into various statistical models. Sliding window methods may be the most straightforward ones. These methods are to test a hypothesis for each probe or region using the statistic that is built based on all the probes' information within a sliding window of certain genomic distance. The test statistics used are varied. For example, Cawley *et al.* (2004) used Wilcoxon's rank sum statistics, Keles *et al.* (2006) used the moving average of classical *t*-statistics, Ji and Wang (2005) used the moving average of empirical Bayesian *t*-statistics and Buck *et al.*(2005) used the moving average of log2 ratios. Since a huge number of tests are performed and the test statistics are not independent, these approaches leave with a difficult problem for multiple hypothesis testing adjustment. Two significant variants of the sliding window methods are the joint binding deconvolution (JBD; Qi *et al.*, 2006) and the model-based deconvolution (with the software MeDiChI) (Reiss *et al.*, 2008), which fit a local regression for each probe with the nearby probes within a sliding window used as explanatory variables, and the inference is made based on the fitted regression coefficients.

Hidden Markov models (HMMs; Ji and Wang, 2005; Li *et al.*, 2005; Munch *et al.*, 2006; Humburg *et al.*, 2008) are appealing approaches to model the dependency of ChIP-chip data. However, in most of the existing implementations of HMM methods, parameter estimation for the models is typically either *ad hoc* or suboptimal for ChIP-chip data. For example, Li *et al.* (2005) estimated the model parameters using the results from a previous Affymetrix SNP array experiments. Munch *et al.* (2006) estimated the model parameters using the Baum–Welch algorithm (Baum, 1972). Humburg *et al.* (2008) used the Baum–Welch algorithm (Rabiner, 1989) plus Viterbi training (Juang and Rabiner, 1990) to estimate the model parameters. The major limitation of these algorithms, as pointed out by Humburg *et al.* (2008), is that the Baum–Welch algorithm tends to converge to a local maximum of the likelihood function and the Viterbi training algorithm even fails to converge to a local maximum of the likelihood in some cases. Therefore, the parameter estimates and followed inference are often suboptimal to the problem. Markov random fields is a natural generation of Markov processes in which a space index is used to replace the time index. Markov random field models are appealing alternative approaches for ChIP-chip data because the spacial dependency can be naturally modeled.

*To whom correspondence should be addressed.

The Bayesian hierarchical model proposed by Gottardo *et al.* (2008) modeled the probe intensities using a mixture of normal distributions, and the spatial dependency of the probes through a Markov random field prior, which is a Gaussian intrinsic autoregression model (Besag and Kooperberg, 1995). The important feature of this model is that neighboring probes are encouraged to be in the same state (enriched or non-enriched), which is a typical characteristic of ChIP-chip data. The inference is based on the joint probability of neighboring probes, which in fact is an off-model remedy because the model itself is not able to completely simulate the spatial dependency of ChIP-chip data. A limitation of the model used by Gottardo *et al.* (2008) is that it is very computationally intensive: for a dataset with 300 000 probes, it needs about 10 h for 15 000 iterations on a personal computer.

We note that the locally self-clustering behavior of the ChIP-chip data is extremely similar to the phenomenon observed in ferromagnetic materials, which has been extensively studied by physicists. In ferromagnetic materials, many electrons act cooperatively and spin in the same direction. The Ising model, named after the German physicist Ernst Ising, was designed to investigate whether local forces can cause a large number of the electrons to spin in the same direction. For further details of the Ising model, we refer the reader to Kindermann and Snell (1980). Here, we just briefly describe the concept. The 1D Ising model considers a sequence of points on a line. At each point, there is an atom of a magnetic material which at any given moments is in one of two states, 'up' and 'down'. If we think of each probe along the chromosomes is in one of two states, enriched or non-enriched, then the Ising model could be a good fit to ChIP-chip data. In this article, we propose to model ChIP-chip data through a hidden ferromagnetic Ising model. The rationale underlying this modeling will be explained further in later sections. Our method differs from the existing methods that also model the spacial dependency such as HMMs (Ji and Wang, 2005; Li *et al.*, 2005; Munch *et al.*, 2006; Humburg *et al.*, 2008) and the Bayesian hierarchical model (Gottardo *et al.*, 2008) in several respects. First, unlike the existing HMM methods, we model ChIP-chip data in a fully Bayesian framework and simulate the posterior distributions of the model parameters using the Metropolis within Gibbs sampler. Second, our method naturally takes into account the dependence structure of the data, does not need off-model remedy for inference and is computationally much more efficient as compared with the Bayesian hierarchical model (Gottardo *et al.*, 2008). Third, our method makes little assumption about the size of IP-enriched regions, and it is able to detect IP-enriched regions with few to numerous probes. As a result, our method significantly outperforms the existing methods for the data with low-genomic resolutions. We will demonstrate these advantages of our method using three real and various simulated datasets.

The remainder of this article is organized as follows. In Section 2, we describe our new Bayesian hierarchical model and its Markov chain Monte Carlo (MCMC) implementation. In Section 3, we compare our method to three closely related methods, namely TileMap HMM (Ji and Wang, 2005), tileHMM (Humburg *et al.*, 2008) and BAC (Gottardo *et al.*, 2008) using three experimental datasets. In Section 4, we comparatively evaluate the performances of our method and the three alternative methods by a series of simulation studies. In Section 5, we conclude this article with a brief discussion.

2 THE MODEL

Let $\mathbf{y}=(y_1,\dots,y_n)$ be a realization of relative enrichment measurements $\mathbf{Y}=(Y_1,\dots,Y_n)$ on n probes along chromosomes. Here, the definition of Y_i is flexible. It could be any appropriate measurement for comparison of treatment (IP-enriched) and control (non-enriched) samples. For example, y_i could be a $\log_2$ ratio of IP-enriched and control samples for an experiment with single replicate, or a summary statistic constructed at the probe level for an experiment with multiple replicates. The proposed method is designed to model the probe enrichment measurements, instead of the probe intensities, as the latter may have sequence-specific effects due to different binding affinities. A wisely designed statistic of enrichment measurement could remove the sequence-specific effects in preparation of the data $\mathbf{y}$, for example, by subtracting the probe intensities of the control samples from that of the IP-enriched samples. We let each probe associate with a binary latent variable $X_i \in \{-1,1\}$, where $X_i=1$ denotes that the probe belongs to a enriched region, and -1 a non-enriched region. We assume, conditioning on $X_i=x_i$, Y_i follows a normal distribution

$$y_i\,|\,x_i \sim \begin{cases} \mathrm{N}(\mu_a,\sigma^2) & \text{if } x_i=-1, \\ \mathrm{N}(\mu_b,\sigma^2) & \text{if } x_i=1. \end{cases} \tag{1}$$

Furthermore, we assume, conditioning on $\mathbf{X}=(X_1,\dots,X_n)$, $Y_1,\dots,Y_n$ are independent. Thus, we have

$$\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\psi})=\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\mu_a)^2}{2\sigma^2}\right)\right)^{I(x_i=-1)} \left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\mu_b)^2}{2\sigma^2}\right)\right)^{I(x_i=1)}, \tag{2}$$

where $\mathbf{x}=(x_1,\dots,x_n)$ is a realization of $\mathbf{X}$, $\boldsymbol{\psi}=(\mu_a,\mu_b,\sigma)$ denotes a collection of the model parameters and $I(x_i=c)=1$ if $x_i=c$ and 0 otherwise.

2.1 The priors

Let $\lambda=\sigma^{-2}$. To conduct a Bayesian analysis, we assume that β and (μ_a,μ_b,λ) have the following prior densities:

$$\beta\sim\text{Uniform}(0,100), \qquad \pi(\mu_a,\mu_b,\lambda)\propto\frac{1}{\lambda}. \tag{3}$$

The prior imposed on β, which, as a practical matter, is equivalent to uniform$(0,\infty)$, but ensures the posterior to be proper. The hidden state vector $\mathbf{X}$ is modeled by a 1D ferromagnetic Ising model

$$\pi(\mathbf{x}|\beta)=\frac{1}{Z(\beta)}\exp\left(\beta\sum_{i=1}^{n-1}x_i x_{i+1}\right), \tag{4}$$

where β is the interaction parameter, and $Z(\beta)$ is the normalizing constant of the distribution, which has a closed form (Baxter, 1982)

$$Z(\beta)=2^n\left(\cosh(\beta)\right)^{n-1}. \tag{5}$$

When $\beta=0$, it means that there is no interaction between probes; when $\beta<0$, the model (4) tends to generate a spatial pattern with more adjacent probes being in opposite states; when $\beta>0$, the model (4) is called the ferromagnetic Ising model. Only when $\beta>0$, the proposed model reflects the structure of ChIP-chip data. Therefore, we restrict β to be positive. A large positive value of β gives rise to large clusters of probes. Here, we only consider the interactions among the nearest neighbors. When the interaction is beyond the nearest neighbors, the model turns into a high-order Ising model, and $Z(\beta)$ become intractable. For a high-order 1D Ising model, sampling from the posterior distribution of β is extremely difficult considering the huge number of probes that are needed to be processed.

The rationale underlying the model (4) can be justified from the perspective of ChIP-chip experiments as follows. In ChIP-chip experiments, genomic DNA is sheared into fragments with various lengths, typically ranging from 200 to 1000 bp. The DNA fragments bound by binding proteins are enriched through immunoprecipitation (IP). After amplification and labeling, the IP-enriched DNA fragments are further chopped to smaller pieces for hybridizations. As a result of this process, all probes in the vicinity of the binding sites have chances to hybridize to the IP-enriched DNA fragments, resulting in consecutive probes with relatively high intensities. In contrast, if the DNA fragments do not contain binding sites, it is expected that the probe intensities in that genomic regions represented by consecutive probes are primarily composed of background intensities. Therefore, the probes along the chromosomes tend to form clusters, which are made of consecutive probes of the same state (enriched or non-enriched). The cluster sizes depend on the genomic resolution and the lengths of sheared DNA fragments. For example, for an array with 35 bp resolution, each IP-enriched region may consist of approximately 10–58 probes; and for an array with 280 bp resolution, each IP-enriched region may consist of approximately 1–7 probes. This locally self-clustering phenomenon of ChIP-chip data is extremely similar to the cooperative behavior of magnetic materials, which has been modeled by Ising models in physics. If we code the probes belonging to the enriched and non-enriched regions by 1 and -1, respectively, then the hidden state vector introduced in (1) can be viewed as a configuration of an Ising model and can thus be modeled by (4).

We note that an analogy of our model in 2D space, where $\mathbf{Y}$ is observed on a rectangular lattice, has often been used in image analysis and segmentation (see, e.g. Hurn *et al.*, 2003; Ibanez and Simo, 2003).

2.2 The full conditionals

Let $n_a = \sum_{i=1}^{n} I(x_i=-1)$ and $n_b = \sum_{i=1}^{n} I(x_i=1)$ be the total numbers of non-enriched and enriched probes, respectively. Let $\bar{y}_a = \frac{1}{n_a}\sum_{i=1}^{n} I(x_i=-1)y_i$ and $\bar{y}_b = \frac{1}{n_b}\sum_{i=1}^{n} I(x_i=1)y_i$ be the sample means of the measurements for non-enriched and enriched probes, respectively. In addition, let $\mathrm{N}(a,b)$ denote a Gaussian distribution with mean a and variance b, let $\mathrm{Ga}(a,b)$ denote a gamma distribution with mean a/b and variance a/b^2, and let $y\,|\,\cdot$ denote the full conditional distribution of y given everything else in the model. After some algebra, we get the following full conditional distributions:

$$\mu_a\,|\,\cdot \sim \mathrm{N}\left(\bar{y}_a, \frac{1}{n_a\lambda}\right), \tag{6}$$

$$\mu_b\,|\,\cdot \sim \mathrm{N}\left(\bar{y}_b, \frac{1}{n_b\lambda}\right), \tag{7}$$

$$\lambda\,|\,\cdot \sim \mathrm{Ga}\left(\frac{n}{2}, \frac{1}{2}\left(\sum_{i=1}^{n} I(x_i=-1)(y_i-\mu_a)^2 + \sum_{i=1}^{n} I(x_i=1)(y_i-\mu_b)^2\right)\right), \tag{8}$$

$$\pi(x_i=1\,|\,\cdot) = \left(1+\exp\left(\frac{\lambda}{2}\left((y_i-\mu_b)^2-(y_i-\mu_a)^2\right) - 2\beta(x_{i-1}+x_{i+1})\right)\right)^{-1}, \tag{9}$$

$$\pi(\beta\,|\,\cdot) \propto \frac{1}{Z(\beta)}\exp\left(\beta\sum_{i=1}^{n-1} x_i x_{i+1}\right) I(0<\beta<100), \tag{10}$$

where $i=1,\ldots,n$, and $x_0 = x_{n+1} = 0$.

2.3 Metropolis within Gibbs sampling

Given the full conditional distributions (6–9), it is straightforward to simulate from the posterior distributions using a cyclic Gibbs sampler. The posterior distribution of β can be simulated using the Metropolis–Hastings algorithm with a Gaussian random walk proposal. The acceptance probability of the update is

$$a(\beta,\beta') = \min\left(1, \frac{Z(\beta)\exp\left(\beta'\sum_{i=1}^{n-1} x_i x_{i+1}\right)}{Z(\beta')\exp\left(\beta\sum_{i=1}^{n-1} x_i x_{i+1}\right)}\right), \tag{11}$$

where β is the current value and β' is the proposed value. The posterior probabilities of the hidden variable $X_i, i=1,2,\ldots,n$, will be used for inference of binding sites.

3 CASE STUDY: ANALYSIS OF TRANSCRIPTION FACTOR CHIP-CHIP DATA

3.1 The p53, Oct4 and Nanog data

Cawley *et al.* (2004) mapped the binding sites of three transcription factors (TFs), p53, Sp1 and cMyc on human chromosomes 21 and 22 using the Affymetrix tiling arrays with an average of 35 bp resolution. All experiments were done with six replicates. The perfect match (PM) intensities were used as the measurements of the abundance of DNA fragments. The PM intensities were log2 transformed and then quantile normalized within treatment and control replicate groups, respectively (Irizarry *et al.*, 2003). Finally, all arrays were scaled to have the same median value. Here, we focus on the p53-DO1 experiments in which antibody p53-DO1 was used for the IP, and the antibody against bacterial GST was used in the control experiment. The chromosome 22 in the p53 experiments had the most validated regions (nine in total). Therefore, the RT-PCR validated regions on the chromosome 22 were used as the gold standard for comparison.

Boyer *et al.* (2005) used the Agilent promoter arrays with an average of 280 bp resolution to identify the binding sites of three TFs Oct4, Nanog and Sox2. In the experiments, IP-enriched DNA was labeled with Cy5 dye, and whole-cell extract DNA was used as a control and labeled with Cy3 dye. All experiments were done in duplicates. The Cy5 and Cy3 intensities were log2 transformed and then quantile normalized. Oct4 and Nanog data were used for the purpose of illustration and comparison.

3.2 A comparative evaluation of the four methods

Both TileMap HMM (Ji and Wong, 2005) and tileHMM (Humburg *et al.*, 2008) use a two-step approach for the data with multiple replicates. In the first step, probe-level statistics that measure the relative abundance of DNA fragments in the treatment and control experiments are calculated. In the second step, the regularized t-statistics are used to build a HMM, in which they are treated as the outcomes of enriched and non-enriched hidden states. Following these approaches, we also used a two-step approach for the data with multiple replicates. We first calculated a statistic for each probe and then modeled the statistics using the proposed method. Regarding the probe-level statistics, there are several choices available, for instance, the moderated t-statistic (Smyth, 2004), the empirical Bayes t-statistic (Ji and Wong, 2005) and the shrinkage t-statistic (Opgen-Rhein and Strimmer, 2007). All these statistics are constructed by borrowing information from all probes on the microarray to estimate individual variances, resulting in increased performance compared with the ordinary t-statistic. To compare the performances of the proposed method and the alternative methods, we used the same statistic for all the three methods. Here, we chose the moderated t-statistic, which has been implemented in the limma package in R (http://www.bioconductor.org/). For this reason, we will call the moderated t-statistic limma t-statistic. The limma t-statistic has a robust behavior and good performance even for small numbers of replicates (Smyth, 2004; Opgen-Rhein and Strimmer, 2007). Therefore, we used this statistic as the enrichment measurement of DNA fragments.

For comparison of our method and the three alternative methods, we first evaluated these methods using a high-genomic resolution dataset—the p53 chromosome 22 data, and then two low-genomic resolution datasets—the Nanog and Oct4 data. To use our method, we need to initialize two model parameters, the latent vector $\mathbf{X}$ and the interaction parameter β.

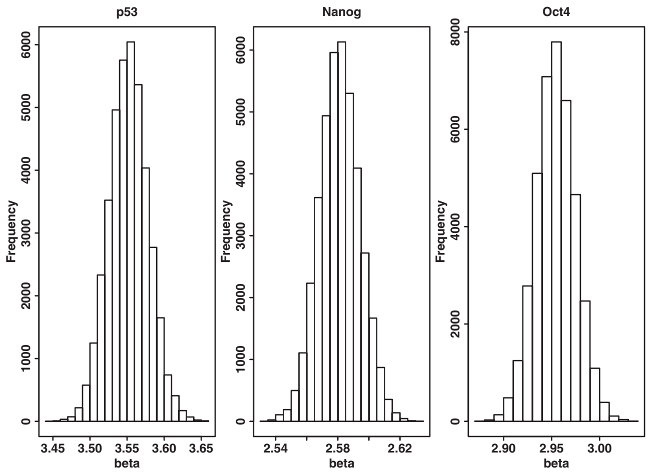

Fig. 1. The histograms of posterior β samples for the p53 chromosome 22, Nanog and Oct4 data.

For both parameters, prior information can be used to set the initial values. It is generally known that the enrichment measurement of an enriched probe is typically several standard deviations (SDs) higher than the mean value. Therefore, we set the initial sate of each probe to 1 if its limma t-statistic is >2 SDs of the mean value and 0 otherwise. The parameter β plays an important role on the model (4). It affects the cluster number and size of the probes. A probe cluster is defined as consecutive probes in the same state (enriched or non-enriched). We can think of a genome that is partitioned into a series of segments, non-enriched and enriched regions, and each of the segments is a cluster. When β is small, the neighboring effect of probes is weak. As a result, the probes tend to form many small clusters. When β increases, the number of probe clusters decreases, and the cluster size increases. We set the initial value of β to 3.0 and updated β using the Metropolis–Hastings algorithm with a Gaussian random walk proposal with SD 0.01. The other parameters were updated using a cyclic Gibbs sampler.

To apply TileMap HMM and BAC to ChIP-chip data, an important parameter to be determined in advance is the size of the sliding window. Both TileMap HMM and BAC have been used to analyze the p53 data, thus we just used the default parameters given by the authors. For the Agilent Oct4 and Nanog data, we followed the suggestions by the authors, adjusting the sliding window size according to the probe resolution and the average length of the sheared DNA fragments. Boyer *et al.* (2005) reported that the maximum sheared DNA fragment size was ∼550 bp, which suggests that the DNA fragments would approximately cover three to five probes, given that the probe resolution is ∼280 bp. In fact, we have tried different window sizes for the two methods and found the results were best when the size was set to this range. This is consistent with our estimation. The tileHMM method uses t emission distribution in the HMM, thus it needs to choose a degree of freedom for the t-distribution. We also tried different values in order to obtain the best results. Therefore, for the three alternative methods, we report the best results we could obtain. For our method, we let the Metropolis within Gibbs sampler run for 50 000 iterations (the first 10 000 used as burn-in) for the p53 and Nanog data, and 70 000 iterations (the first 30 000 used as burn-in) for the Oct4 data. Figure 1 shows the histograms of the β samples generated from their posterior distributions. The means and standard errors (in the parentheses) of the β parameters are $3.55(0.026), 2.58(0.013)$ and $2.96(0.024)$ for the p53, Nanog and Oct4 data, respectively.

Table 1. Number of enriched regions detected by each method for the p53, Oct4 and Nanog data

		p53		Nanog		Oct4	
Method	Cutoff	V	Total	G	Total	G	Total
Ising	0.95 PP	9	145	212	816	211	476
	0.90 PP	9	156	216	881	231	541
	0.01 FDR	9	158	215	872	216	491
	0.05 FDR	9	180	229	1053	249	708
	Top 60	9	60	33	60	49	60
TileMap HMM	0.95 PP	9	102	117	250	98	150
	0.90 PP	9	109	144	352	122	192
	0.01 FDR	9	104	99	195	79	114
	0.05 FDR	9	139	158	418	136	223
	Top 60	9	60	33	60	37	60
tileHMM	0.95 PP	9	14 788	310	1934	291	2026
	0.90 PP	9	15 027	337	2582	321	3630
	0.01 FDR	0	10	296	1501	221	942
	0.05 FDR	9	14 906	341	2970	322	3669
	Top 60	0	60	32	60	41	60
BAC	0.95 PP	9	89	101	175	10	13
	0.90 PP	9	100	133	246	17	21
	0.01 FDR	9	94	80	137	7	8
	0.05 FDR	9	142	144	287	18	24
	Top 60	9	60	45	60	46	60

The Ising, TileMap HMM and BAC detect all the nine validated (V) regions for the p53 data at fixed posterior probability and FDR cutoffs; tileHMM essentially fails for the p53 data. For the Oct4 and Nanog data, the Ising method detects much more gold (G) regions and total enriched regions than TileMap HMM and BAC; tileHMM detects more gold regions than Ising at the expense of big false positives.

We compared our methods with the three alternative methods using the validated regions or high-confidence enriched regions as the gold standard. For the p53 chromosome 22 data, there were nine RT-PCR validated enriched regions. For the Nanog and Oct4 data, although there were not RT-PCR validated regions, the authors (Boyer *et al.*, 2005) reported 367 potentially enriched regions on 353 genes that were co-occupied by the three TFs, Nanog, Oct4 and Sox2. To detect enriched regions, Boyer *et al.* (2005) calculated a score for each probe using the so-called whole-chip error model. A single probe P-value (SPP) was then calculated for each score, and a probe set P-value (PSP) was calculated for the average score of each probe and its two immediate neighbors. A probe set is said to be enriched if its PSP <0.001 and two of the SPPs<0.005, or the center SPP<0.001 and one of the flanking SPPs<0.1. Although the false discovery rate (FDR) was not estimated for the 367 regions, they still should be a good standard for the comparison considering that these regions are detected by all the three experiments. To ease comparison, we selected fixed posterior probability and FDR cutoffs. The latter was calculated using a direct posterior probability approach as described in Newton *et al.* (2004). Following the approach used by Cawley *et al.* (2004), we merged adjacent enriched probes separated by 500 bp or less into the same enriched region. The results are summarized in Table 1, where 'Ising' refers to our method. For the p53 data, at the posterior probability cutoff 0.9 and FDR cutoff 0.05, all the four methods detect the nine validated enriched regions. Although tileHMM detects all the nine validated regions, it essentially fails for the p53 data. For example, it detects 15 027 enriched regions, which are obviously dominated by false enriched regions. We further looked into the ranking performance of these validated regions. The nine validated regions on chromosome 22 are found by Ising, TileMap HMM and BAC in their top 60 regions. However, none is found by tileHMM, which further supports that most of the enriched regions detected by tileHMM are false positives. For the Oct4 and Nanog data, at the fixed posterior probability and FDR cutoffs, our approach detects much more gold regions and total enriched regions than TileMap HMM and BAC. BAC essentially fails for

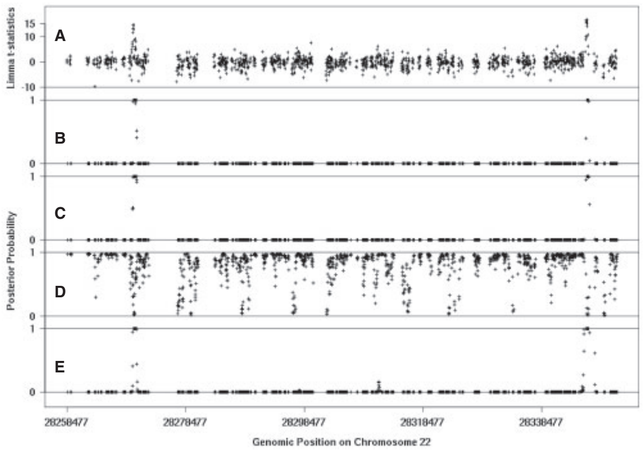

Fig. 2. Limma *t*-statistics and posterior probabilities of probes in enriched state versus genomic positions for the p53 chromosome 22 data. (**A**) The limma *t*-statistics; (**B**–**E**) posterior probabilities for Ising, TileMap HMM, tileHMM and BAC, respectively.

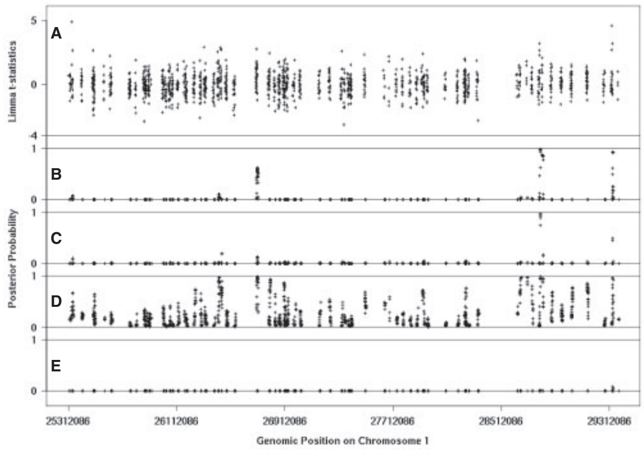

Fig. 3. Limma *t*-statistics and posterior probabilities of probes in enriched state versus genomic positions for the Oct4 data. (**A**) The limma *t*-statistics; (**B**–**E**) posterior probabilities for Ising, TileMap HMM, tileHMM and BAC, respectively.

the Oct4 data because it only detects a small number of enriched regions at various posterior probability and FDR cutoffs. It is pretty clear that most of the enriched regions are not detected by TileMap HMM and BAC for the Oct4 and Nanog data. Although tileHMM detects more gold regions than our method, it achieves this at the expense of a lot of false positives. Therefore, based on these real examples, our approach seems to perform at least as well as the others for the high-resolution data, but significantly outperform the others for the low-resolution data.

To get a better sense of the performance of these methods, it may be necessary to show the results in a typical genomic region. Figures 2 and 3 show the posterior probabilities of probes in enriched state against the genomic positions for a small portion of the p53 and Oct4 data. In general, the posterior probabilities produced by TileMap HMM, BAC and our methods tend to be dichotomized, either close to 0 or 1. As expected, most of the posterior probabilities are close to 0, which corresponds to the non-enriched state. This indicates a clear classification of enriched and non-enriched probes. In contrast, the posterior probabilities generated by tileHMM are not

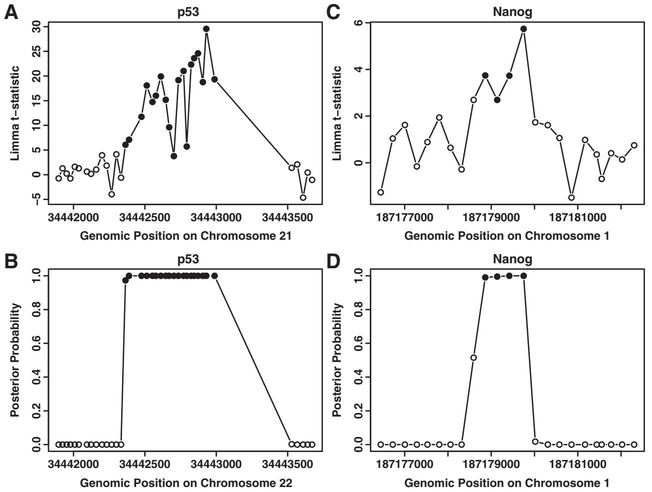

Fig. 4. Distribution of limma *t*-statistics and posterior probabilities of probes in enriched state in typical enriched regions for the p53 data (**A** and **B**) and the Nanog data (**C** and **D**). The enriched region for the p53 data contains about 19 probes due to the high probe resolution (35 bp). In contrast, the enriched region for the Oct4 data only contains 4–5 probes due to the low probe resolution (280 bp).

dichotomized, indicating that it is not able to produce a clear classification of enriched and non-enriched probes. For the high-resolution p53 data, TileMap HMM, BAC and our method work well. The two confirmed enriched regions are successfully detected by all the three methods (Fig. 2). However, most of the posterior probabilities generated by tileHMM are close to 1, and this is obviously unrealistic (Fig. 2). For the low-resolution data, as shown in Figure 3, our method detects more potential enriched regions than TileMap HMM and BAC. TileHMM also fails for this data, because the posterior probabilities generated by that are not dichotomized and have too many high values in the non-enriched regions. Figure 4 shows the typical enriched regions detected by our method for the data with high- and low-genomic resolutions. Due to different genomic resolutions, the numbers of probes used to interrogate the enriched regions are much different. The enriched region for the p53 data in Figure 4 are composed of about 19 probes. In contrast, there are only 4–5 probes in the Nanog enriched region. Our method is able to detect enriched regions with few to numerous probes.

4 SIMULATION STUDIES

In order to have a thorough evaluation of the performance of the Ising model in comparison with the other three methods, we carried out a series of simulations. For a fair comparison, we tried to generate simulated data that reflect the underlying structure of ChIP-chip data, instead of generating particular data to satisfy each model. The probe enrichment measurements in a enriched region typically form a triangular or bell-shaped structure (e.g. see Fig. 4), where the peak location is usually closest to the actual binding site. Following Gottardo *et al.* (2008), we used the function given by $M\exp(-4D_p^2/L^2)$ to model the probe intensities in the enriched regions, where M is the value of the peak, D_p is the genomic distance between probe p and the peak and L is the length of the enriched region.

To generate simulated data with high- and low-genomic resolutions, we used the first 40 000 genomic positions of the Affymetrix chromosome 22 and the Agilent chromosome 1 as the genomic coordinates, respectively. The binding site of a TF is

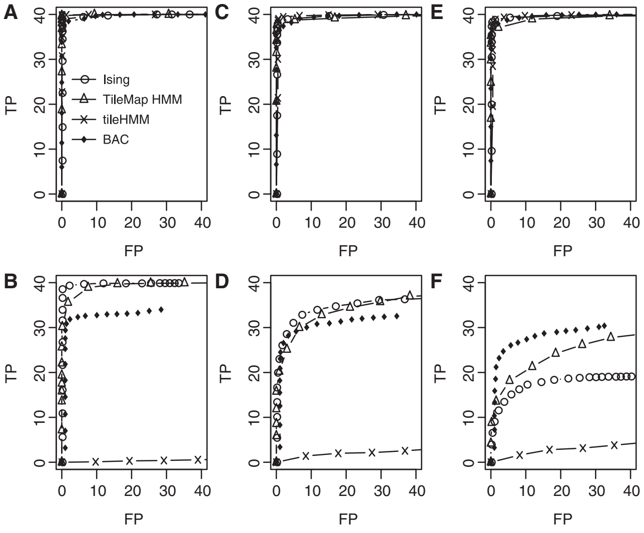

Fig. 5. ROC curves for all the methods applied to the simulated data. (**A, C** and **E**) data were generated using the Affymetrix coordinate (high resolution); (**B**, **D** and **F**) data were generated using the Agilent coordinate (low resolution); (A and B) noises with variance of 1 were added to the data; (C and D) noises with variance of 2.25 were added to the data; (E and F) noises with variance of 4 were added to the data.

typically about several base pairs, thus a probe can completely cover it. Therefore, it is appropriate to choose the genomic position of a probe as a peak position. To generate data for enriched regions, we randomly selected 40 genomic positions as the peak positions. The values for the peaks were randomly generated from a uniform distribution between 3 and 8, and the length of the enriched regions (parameter for L) were randomly generated from a uniform distribution between 400 bp and 1000 bp. The probes in a given enriched region were set to the values according to the function $M\exp(-4D_p^2/L^2)$ if the distances between the probes and the peak are $\leq L/2$ bp, and zero otherwise. In addition, we also added background noises to the probes in the enriched regions. Background noises were generated from the normal distributions with mean 0 and variances of 1, 2.25 and 4, respectively. For the probes in the non-enriched regions, their values were set to pure background noises. We simulated 60 high-resolution and 60 low-resolution datasets. Each dataset contained three replicates for both control and treatment conditions, and each replicate contained 40 000 probes. Among the 60 high- or low-resolution datasets, 20 were created using the background noise with a variance of 1, another 20 with a variance of 2.25 and the rest 20 with a variance of 4. In summary, we generated six different datasets corresponding to the six combinations: two different genomic coordinates × three different background noises.

We evaluated the performance of the four methods by comparing the receiving operating characteristic (ROC) curves for each simulation scenario. The ROC curves show the average number of positive regions against the average number of false positive regions detected by each method when varying the posterior probability cutoff (Fig. 5). For the simulated high-resolution data (Fig. 5A, C and E), regardless the amount of noise added to the data, all the methods perform almost equally well. The area under ROC curve (AUC) is around 1 for each method. This indicates that all the four methods are quite robust to the noise. For the simulated low-resolution data (Fig. 5B, D and F), our method performs the best when the level of noise is low (the variance of noise are equal to 1 or 2.25). When the variance of noise increases to 4, BAC performs the best, followed by TileMap HMM and our method. TileHMM performs the worst for the simulated low-resolution data with an extremely large false positive rate. This can be seen from Figure 5B, D and F, where the corresponding curves are almost flat and close to the horizontal axis. For the low-resolution data, the deteriorated performance of the these methods with increase of the background noise is quite understandable, because the enriched regions typically contain only a few probes, and are thus more easily affected by the noise.

5 DISCUSSION

In this article, we have presented a simple, but powerful Bayesian hierarchical method to model ChIP-chip data through a hidden ferromagnetic Ising model. This approach naturally takes into account the intrinsic dependency of ChIP-chip data, and can be used to analyze data with various genomic resolutions and sample sizes. We comparatively evaluated our method and three closely related methods using three real and various simulated datasets. The results showed that our method performed as well as TileMap HMM and BAC in terms of sensitivity and specificity in detecting IP-enriched regions for the high-resolution data. However, our method significantly outperformed the alternatives for the low-resolution data. The results showed that tileHMM did not work consistently. It essentially failed for the p53, Nanog and Oct4 data because it was not able to distinguish enriched and non-enriched probes (see Figs 2, 3 and Table 1). The regions detected by tileHMM are obviously dominated by false positives. However, it worked as well as the three alternative methods on the simulated high-resolution data (Fig. 5A, C and E). For the simulated low-resolution data, tileHMM had extremely high false positive rates at various posterior probability cutoffs (Fig. 5B, D and F). Since all the four methods essentially make efforts to estimate the state (enriched or non-enriched) of each probe, ideally it is expected that probe-level posterior probabilities should be dichotomized, either close to 1 (for the enriched state) or 0 (for the non-enriched state), which can be used as a criterion for evaluating the performance of these methods. In terms of dichotomization, our method performed the best, followed by TileMap HMM and BAC, and tileHMM performed the worst.

The three alternative methods used for comparison in this article are closely related to the Ising model we use. The Markov model is essentially a 1D Ising model (for details see Kindermann and Snell, 1980). Vice versa, the model we proposed can also be treated as a HMM in which the transition matrix is determined by the interaction parameter β, and the emission probabilities are determined by the Equation (9). In addition, the Gaussian intrinsic auto-regression model used by Gottardo *et al.* (2008) can also be viewed as a continuous extension of the Ising model.

Since all the four methods basically use the same idea to model ChIP-chip data, why does our method outperform the others for the low-resolution data? This could be owed to the way we model the data and the algorithm used for parameter estimation. In our method, Metropolis within Gibbs sampling algorithm is used to estimate the model parameters, which is known to converge to their target distributions when the number of iterations becomes large.

Our method allows for a more precise estimation of the model parameters, which in turn, leads to an improvement in detecting IP-enriched regions. In contrast, tileHMM is a non-Bayesian method, where the parameters are estimated using the Baum–Welch and/or Viterbi training algorithm. It is known that the Baum–Welch algorithm tends to converge to a local maximum of the likelihood function, and the Viterbi training algorithm may not even converge to the local maximum. The inconsistent performance of tileHMM on the real and simulated data indicates that the inferiority of tileHMM is mainly due to its training algorithm. In TileMap HMM, the model parameter estimation is typically *ad hoc*. It assigns a fixed value to the transition probability from an enriched probe i to a non-enriched probe $i+1$ if the two adjacent probes are in a predefined window with certain genomic distance. For example, for the array with 35 bp resolution and average DNA fragments ~1000 bp, TileMap HMM sets the transition probability to 1/28 because it assumes that typical IP-enriched regions contain about 28 probes and 1/28 matches the mean length of a segment in an enriched region. In addition, TileMap HMM tends to be conservative in detecting IP-enriched regions. BAC models the spatial dependency of probes via a Gaussian intrinsic auto-regression model, as aforementioned, which can be viewed as a continuous extension of the Ising model. To further make use of information of neighboring probes, Gottardo *et al.* (2008) included in BAC an extra step to estimate the joint posterior probability of the probes in a sliding window. However, as demonstrated by our numerical examples, this off-model remedy is not very effective, especially for the low-resolution data. In addition, BAC directly models the probe intensity using a mixed-effect model. It assumes that each probe has a probe-specific background intensity, enrichment effect and random error. As a result, there are a huge number of parameters needed to be estimated, which not only significantly increases the computational burden but also makes the model potentially over-fitted, especially when the number of replicates is small.

In summary, we have proposed a hidden ferromagnetic Ising model for ChIP-chip data analysis. The proposed method models the data in a fully Bayesian framework, which enables the model parameters to be estimated in a precise way. This overcomes the difficulty in parameter estimation for the existing HMM methods, leading to a significant improvement in detecting IP-enriched regions, especially for the low-resolution data. In contrast, the parameter estimation in most of the existing HMM methods, is typically either *ad hoc* or suboptimal. BAC also models the ChIP-chip data in a fully Bayesian framework. Compared with BAC, our model is much simpler and the computational burden is significantly reduced. For example, for a dataset with 500 000 probes, to run 15 000 iterations on a 64-bit Linux machine with 2.4 GHz CPU, our method needed ~15 min, but BAC needed ~20 h. In addition, our method makes little assumptions about the size of the IP-enriched regions, and it can be used for the data with various levels of genomic resolution.

Finally, we would like to point out that, with minor modifications, the proposed method may be applicable to other spatially correlated data such as ChIP-seq data (see, e.g. Park, 2009 for a review). To accommodate the discreteness of ChIP-seq data, a Poisson-like model can be used for the summary measurement $\mathbf{Y}$.

ACKNOWLEDGEMENTS

The authors thank Hongkai Ji for helpful discussion about TileMap HMM, and the editor, the associate editor and the referees for their comments which have led to significant improvement of this article.

Funding: National Science Foundation (grant DMS-0607755 to L.F., partially); award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST to L.F.).

Conflict of Interest: None declared.

REFERENCES

Baxter,R.J. (1982) *Exactly Solved Models in Statistical Mechanics*, 1st edn. Academic press, London.

Baum,L.E. (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

Besag,J. and Kooperberg,C. (1995) On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733–746.

Boyer,L.A. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Buck,M.J. *et al.* (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.*, **6**, R97.

Cawley,S. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.

Gottardo,R. *et al.* (2008) A Flexible and powerful Bayesian hierarchical model for ChIP-chip experiments. *Biometrics*, **64**, 468–478.

Humburg,P. *et al.* (2008) Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics*, **9**, 343.

Hurn,M. *et al.* (2003) A tutorial on image analysis. *Lect. Notes Stat.*, **173**, 87–141.

Ibanez,M.V. and Simo,A. (2003) Parameter estimation in Markov random field image modeling with imperfect observations: a comparative study. *Pattern Recogn. Lett.*, **24**, 2377–2389.

Irizarry,R. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Juang,B.H. and Rabiner,L.R. (1990) A segmental k-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech Sign. Process.*, **38**, 1639–1641.

Ji,H. and Wong,W. (2005) Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **18**, 3629–3636.

Keles,S. *et al.* (2006) Multiple testing methods for ChIP-chip high density oligonucleotide array data. *J. Comput. Biol.*, **13**, 579–613.

Kindermann,R. and Snell,J.L. (1980) Markov random fields and their applications. In *Contemporary Mathematics*. Vol. 1. American Mathematical Society, Rhode Island.

Li,W. *et al.* (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21** (Suppl. 1), i274–i282.

Munch,K. *et al.* (2006) A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7**, 239.

Newton,M. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Opgen-Rhein,R. and Strimmer,K. (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, **6**, Article 9.

Park,P. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

Qi,Y. *et al.* (2006) High-resolution computational models of genome binding events. *Nat. Biotechnol.*, **24**, 963–970.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE 1989*, **77**, 257–286.

Reiss,D.J. *et al.* (2008) Model-based deconvolution of genome-wide DNA binding. *Bioinformatics*, **24**, 396–403.

Smyth,G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.