Journal of
Data Mining in Genomics & Proteomics

Wu et al., J Data Mining Genomics Proteomics 2014, 5:1
DOI: 10.4172/2153-0602.1000153

Research Article Open Access

Model-Free Inference for ChIP-Seq Data

Mingqi Wu[1]*, Monique Rijnkels[2] and Faming Liang[3]

[1]Shell Projects and Technology, Shell Technology, Center Houston, 3333 Highway 6 South, Houston, TX, USA
[2]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA
[3]Department of Statistics, Texas A&M University, College Station, TX, USA

Abstract

Due to its higher resolution mapping and stronger ChIP enrichment signals, ChIP-seq tends to replace ChIP-chip technology in studying genome-wide protein-DNA interactions, while the massive digital ChIP-seq data present new challenges to statisticians. To date, most methods proposed in the literature for ChIP-seq data analysis are model based, however, finding a single model workable for all datasets is impossible, given the complexity of biological systems and variations generated in the sequencing process. In this paper, we present a model-free approach, the so-called MICS (Model-free Inference for ChIP-Seq), for ChIP-seq data analysis. MICS has a few advantages over the existing methods: Firstly, MICS avoids assumptions for the data distribution, and thus it maintains high power even when model assumptions for the data are violated. Secondly, MICS employs a simulation-based method in estimating the false discovery rate. Since the simulation-based method works independently of ChIP samples, MICS can perform robustly to variety of ChIP samples; it can produce accurate identification of peak regions, even for those where the enrichment is weak. Thirdly, MICS is very efficient in computation, which takes only a few seconds on a personal computer for a reasonably large dataset. In this paper, we also present a simple semi-empirical method for simulating ChIP-seq data, which allows a better assessment of performance of different approaches for ChIP-seq data analysis. MICS is compared with several existing methods, including MACS, CCAT, PICS, BayesPeak and QuEST, based on real and simulated datasets. The numerical results indicate that MICS can outperform others. Availability: An R package called MICS is available at http://www.stat.tamu.edu/~mqwu.

Keywords: Protein-DNA interactions; ChIP-Sequence Data; MICS method; Data analysis

Introduction

With the development of next-generation sequencing techniques, the ChIP-seq technology tends to replace ChIP-chip in studying genome-wide protein-DNA interactions. It is known that ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel short-read sequencing (seq), while ChIP-chip combines ChIP with microarray (chip) techniques. Compared with ChIP-chip, ChIP-seq can provide higher resolution mapping and stronger protein binding sites signals. A ChIP-seq dataset often consists of tens of millions of sequence reads (known as tags) generated from the ends of DNA fragments, with each tag around 25 to 50 base pairs in length.

Although ChIP-seq is advantageous over ChIP-chip in detecting protein-DNA interactions, it presents new challenges to statisticians. In particular, it is very difficult to find appropriate statistical models for the discrete tag count data. In the literature, a lot of work has been published with attempts to address this issue, with attempts to address this issue [1-5]. Among them, MACS models the ChIP-seq data using a dynamic Poisson distribution, and CisGenome models the data using a negative binomial distribution, which can be viewed as the marginal version of the Poisson model. CCAT develops a linear signal-noise model for the ChIP-seq data under the assumption that the tag counts follow a Poisson distribution. BayesPeak models the structure of the ChIP-seq data using a hidden Markov model and assumes that the data marginally follow a negative binomial distribution. PICS is another Bayesian approach, which identifies the binding regions via a Bayesian hierarchical t-mixture model. A common feature of these methods is that a distribution is assumed for the data, and thus their performance is data dependent. When the assumption is violated, they may perform very badly. Given the complexity of biological systems and variations generated in the sequencing process, it is impossible to find a single model which fits all data well. Hence, a model-free method would be attractive for the ChIP-Seq data. To the best of our knowledge, the only well-known work in this direction is QuEST [6], which is based on the kernel density estimation approach. However, this method suffers from many limitations. For example, it cannot work when there are only treatment samples available; and the score of each peak region obtained by QuEST is proportional to the local amount of tags and thus it often fails to detect some weak ChIP-enriched regions.

In this paper, we present a new model-free method, the so-called MICS (Model-free Inference for ChIP-Seq), for ChIP-seq data analysis. Compared to the model-based methods, MICS possesses several attractive features. Firstly, it avoids assumptions for the data distribution. Hence, it maintains high power even when model assumptions for the data are violated. Secondly, when both treatment and control samples are available, MICS employs a simulation-based method for estimation of the false discovery rate. Since the simulation-based method works independently of the treatment samples, MICS can perform robustly to the variation of treatment samples; it can produce accurate identification of peak regions, even for those where the enrichment is weak. Thirdly, MICS is efficient in computation. For a reasonably large dataset, it takes only a few seconds on a personal computer. Compared to QuEST, as a new model-free approach, MICS is more efficient in computation and also more accurate in peak region identification.

The remainder of this paper is organized as follows. In the Method Section, we describe MICS in details. In the Results Section, we test MICS on two real datasets with comparisons with MACS, CCAT,

***Corresponding author:** Mingqi Wu, Shell Projects and Technology, Shell Technology, Center Houston, TX, USA, Tel: 2815446412; E-mail: Mingqi.Wu@shell.com

Received July 19, 2013; **Accepted** February 21, 2014; **Published** February 24, 2014

Citation: Wu M, Rijnkels M, Liang F (2014) Model-Free Inference for ChIP-Seq Data. J Data Mining Genomics Proteomics 5: 153. doi:10.4172/2153-0602.1000153

PICS, BayesPeak and QuEST. In the Simulation Study Section, A semi-empirical method for simulating ChIP-seq data is first introduced; then based on the simulated data, the robustness of empirical Bayes method and MICS in estimating π_0, the probability of non-peak regions, is compared; and finally the performance of MICS and other methods is evaluated using the simulated data. Finally, we conclude the paper with a brief discussion.

Methods

Data pre-processing

Given a ChIP-seq dataset, we carry out the following steps for pre-processing the raw reads data.

• Counting. The segment of the genome involved in the ChIP-seq experiment is first divided into *N* non-overlapping bins with a fixed length *w*, which should be around the DNA fragments size, e.g. $w = 100\,bp$. Then reads are extended by expected DNA fragment length, and the number of reads (both forward and reverse strand) falling into each bin is counted. We denote the counts by n_{1i} and n_{2i} for the control and ChIP samples, respectively; $i = 1,...,\mathrm{N}$. without loss of generality, we here assume that there is only one replicate available for each of the treatment and control samples. If there are more replicates available, we need to take the sum under each condition before going to the next step.

• Taking difference. In order to reduce the adverse effects of local sequence read bias, we choose to work on the count difference between the ChIP and control samples for each bin; that is, we work on

$$z_i = n_{2i} - n_{1i,} \qquad i = 1,...,N. \tag{1}$$

Please note, if the sequencing depths between ChIP and control samples are much different, a normalization step with respect to their sequencing depths is necessary before taking the difference. It is worth noting that we have successfully used this strategy to reduce probe specific effects for ChIP-chip data [7].

FDR estimation

In this section, we propose a simulation-based method for estimating false discovery rate (FDR). Before describing our method, we give a brief review of the empirical Bayes method.

Empirical Bayes method: Under the framework of empirical Bayes method, the summary counts $z_1,...,z_N$ are assumed to follow a two-component mixture distribution with the mass function given by

$$f(z) = \pi_0 f_0(z) + (1-\pi_0) f_1(z) \tag{2}$$

where f_0 and f_1 are the mass functions of summary counts belonging to the non-peak and peak regions, respectively; and π_0 and $1-\pi_0$ are the corresponding probabilities of the two components.

Considering a discovery rule $\Lambda = \{\mathrm{z}_i \geq \mathrm{z}_0\}$, where $z_0 > 0$, a direct application of the Bayes rule yields the so called *false discovery rate* [8]

$$FDR(\Lambda) = P\{\mathrm{null} \mid \mathrm{z} \in \Lambda\} = \frac{\pi_0\{1-F_0(z_0)\}}{1-F(z_0)} \tag{3}$$

where F_0 and *F* denote the cumulative distribution functions of f_0 and *f*, respectively. To calculate $FDR(\Lambda)$, the quantities π_0, $1-F_0(z_0)$ and $1-F(z_0)$ can be estimated empirically as follows:

• Estimating $1-F(\mathrm{z}_0)$ by

$$1-\tilde{F}(z_0) = \#\{\mathrm{z}_i : \mathrm{z}_i \geq \mathrm{z}_0\} / \mathrm{N}, \tag{4}$$

where $\#\{z_i : z_i \in A\}$ denotes the number of z_i satisfies the condition *A*.

• Estimating $1-F_0(z_0)$ by

$$1-\tilde{F}_0(z_0) = \#\{\mathrm{z}_i : \mathrm{z}_i \leq -\mathrm{z}_0\} / \mathrm{N}', \quad if \quad z_0 > 0 \tag{5}$$

where $N' = 2\#\{z_i : z_i < 0\} + \#\{z_i : z_i = 0\}$ denotes an estimator for the total number of non-peak region counts. This estimator makes use of the symmetry of the distribution of $F_0(z)$. Here we assume that is true for the peak regions.

• Estimating π_0 by

$$\tilde{\pi}_0 = \frac{\tilde{F}(\mathrm{A}_0)}{\tilde{F}_0(\mathrm{A}_0)}, \tag{6}$$

where A_0 can be chosen as a set for which $F_1(A_0) = 0$ holds approximately. For example, we can set $A_0 = \{\mathrm{z}_i : \mid \mathrm{z}_i \mid \leq \mathrm{h}\}$, where *h*=1 or 2 as suggested by Efron [8].

Given the above estimators, $FDR(\Lambda)$ can then be estimated using a plug-in method by

$$\widehat{FDR}(\Lambda) = \frac{\tilde{\pi}_0\{1-\tilde{\mathrm{F}}_0(\mathrm{z}_0)\}}{1-\tilde{\mathrm{F}}(\mathrm{z}_0)} \tag{7}$$

Note that this method also works for the case that only ChIP samples are available. In this case, we have $z_i = n_{2i}$, $1-F(z_0)$ can be estimated as in (4) and $1-F_0(z_0)$ can be estimated by

$$1-\tilde{F}_0(\mathrm{z}_0) = \#\{\mathrm{z}_i : \mathrm{z}_i \leq -\mathrm{z}_0 + 2\,\mathrm{m}_z\} / \mathrm{N}'', \quad if \quad z_0 > m_z,$$

where m_z denotes the mode of the empirical distribution of n_{2i}'s and $N'' = 2\#\{\mathrm{z}_i : \mathrm{z}_i < \mathrm{m}_z\} + \#\{\mathrm{z}_i : z_i = m_z\}$; and $\tilde{\pi}_0$ can be estimated by

$$\tilde{\pi}_0 = \frac{\tilde{F}(A_0)}{\tilde{F}_0(A_0)},$$

with A_0 being chosen as $A_0 = \{\mathrm{z}_i : \mathrm{z}_i \leq \mathrm{m}_z\}$.

A simulation-based method (MICS): Consider the estimator of π_0 given in (6). Define

$$y_i = I(z_i \in A_0), \qquad i = 1,...,N,$$

where $I(\bullet)$ is the indicator function, and *N* is the total count of z_i's. Let $\{\mathrm{z}_1^*,...,\mathrm{z}_{N'}^*\}$ be a collection of z_i's belonging to the component F_0; that is, the set of observations from non-peak regions. Define

$$x_i = I(\mathrm{z}_i^* \in \mathrm{A}_0), \qquad i = 1,...,N',$$

assuming that N' is known. Then estimator of π_0 can be written as a ratio estimator

$$\tilde{\pi}_0 = \frac{\overline{y}}{\overline{x}}$$

where $\overline{y} = \sum_{i=1}^{N} y_i / N$ and $\overline{x} = \sum_{i=1}^{N'} x_i / N'$. The usual approximation for the variance of $\tilde{\pi}_0$ is taken as

$$Var(\tilde{\pi}_0) = \frac{1}{[\mathrm{E}(\bar{x})]^2} Var(\bar{y} - \pi_0 \bar{x}) \tag{8}$$

A direct calculation of (8) yields

$$Var(\tilde{\pi}_0) \approx \frac{1}{N}\left\{(1-\rho)\left[\frac{2\pi_0}{F_0(A_0)} - \pi_0(1+\pi_0)\right]\right\} \tag{9}$$

where ρ denotes the correlation coefficient between $\bar{x}$ and $\bar{y}$. It follows from (9) that a choice of A_0 with a large value of $F_0(A_0) = 0$ is preferred, as which will lead to a small variation of $\tilde{\pi}_0$.

However, since the peak and non-peak regions are not directly separable from z_i's, the condition $F_1(A_0) = 0$ is seldom satisfied for a set A_0 with a large probability of $F_0(A_0)$. Note that the resulting estimate of π_0 will be biased if the condition $F_1(A_0) = 0$ is violated. In practice, (6) often produces an unreasonable estimate, greater than 1.0 or much smaller than 0.8. This problem has also been noted by other authors, see e.g. Ma and Wong [16]. To tackle this difficulty, we propose a simulation based method for estimating $F_0(A_0)$ and π_0 thus and FDR for the case that the control sample is available. The simulation-based method simulates a F_0-distribution based on the control samples, and thus $F_0(A_0)$ can be estimated independently of the ChIP samples. Therefore, the simulation based $F_0(A_0)$ estimator will not be affected by the variation of ChIP samples. This is important. As illustrated in Simulation Study Section, when the IP-enrichments are weak, the simulation-based method still works well, while the empirical Bayes method fails. Given the huge amount of ChIP-seq data, the law of large numbers naturally applies, hence, the reliability of the simulation-based $F_0(A_0)$ estimator is not in doubt. The simulation-based method can be described as follows:

1. (*Histogram estimation*) Estimate the histogram of the control samples. Let $k = \max_i n_{1i}$ denote the maximum number of counts in a single bin for the control sample. Normalize the histogram to be a distribution

$$P(\mathrm{S} = \mathrm{j}) = \hat{\mathrm{p}}_j = \#\{\mathrm{n}_{1i} : \mathrm{n}_{1i} = \mathrm{j}\} / \mathrm{N}, \qquad j = 0,1,...,k, \tag{10}$$

where S denotes a random variable defined on the set $\{0,1,...,\mathrm{k}\}$.

2. (*Null samples Simulation*) For ChIP-seq data analysis, the null hypothesis is that, in essence, there is no distributional difference of counts between the control and ChIP samples. Accordingly, under H_0, we may simulate two control samples, $s_1 = (\mathrm{s}_{11},...,\mathrm{s}_{1M})$ and $s_2 = (\mathrm{s}_{21},...,\mathrm{s}_{2M})$, from a multinomial distribution; that is, generating the values of $s_{1i}'s$ and $s_{2i}'s$ with probability

$$P(\mathrm{S}_{li} = \mathrm{j}) = \hat{\mathrm{p}}_j, \qquad j = 0,...,k; \qquad l = 1,2; \qquad i = 1,...,M,$$

where M is chosen in the order of millions to mimic the real order of ChIP-seq samples.

3. (*FDR estimation*) Take the difference between the two samples to construct the null sample $\mathrm{s}_0 = (\mathrm{s}_{01},...,\mathrm{s}_{0M})$, where

$$s_{0i} = s_{2i} - s_{1i}, \qquad i = 1,...,M. \tag{11}$$

Thus, for any set A_0, $F_0(\mathrm{A}_0)$ can be empirically estimated by,

$$\hat{F}_0(A_0) = \#\{\mathrm{s}_{0i} : \mathrm{s}_{0i} \in \mathrm{A}_0\} / \mathrm{M} \tag{12}$$

Therefore, $1 - F_0(z_0)$ can be estimated by

$$1 - \hat{F}_0(z_0) = \#\{\mathrm{s}_{0i} : \mathrm{s}_{0i} > \mathrm{z}_0\} / \mathrm{M} \tag{13}$$

and π_0 can be estimated by

$$\hat{\pi}_0 = \frac{\tilde{F}(A_0)}{\hat{F}_0(A_0)}, \tag{14}$$

where A_0 can be set as $A_0 = \{\mathrm{s}_{0i} : |\, \mathrm{s}_{0i} \,| \leq \mathrm{h}\}$, with a value of h such that $F_0(\mathrm{A}_0)$ has a probability greater than 0.8. In practice, we often set h=1,2 or 3.

Then the FDR can be estimated in equation (7) with $\tilde{F}_0$ and $\tilde{\pi}_0$ being replaced by $\tilde{F}_0$ and $\tilde{\pi}_0$, respectively.

In order to further improve the estimates, we may independently repeat this procedure J times, and use the average value $\hat{\pi}_0 = \sum_{i=1}^{J} \hat{\pi}_{0i} / J$ as the final estimation of π_0, and the average value $1 - \hat{F}_0(\mathrm{z}_0) = 1 - \sum_{i=1}^{J} \hat{F}_{0i}(z_0) / J$ as the final estimate of $1 - F_0(z_0)$.

Hereafter, this simulation-based method will be called MICS. Compared to the empirical Bayes method, a significant advantage of the new method is that it can estimate of $F_0(\mathrm{A}_0)$ independently of the ChIP samples, and resulting FDR estimate can be more robust to the variation of ChIP samples. Our numerical results indicate that MICS can produce more accurate identification of peak regions, even for those where the enrichment is weak. To support our claims, we have done a series of simulation studies to illustrate the robustness of our method in estimating and identifying peak regions.

Identification of peak regions

Consider a window of 10 bins wide, $W = 10w = 1000\,bp$, which is approximately equal to the length of a bound region. If all bins of the window are significant, i.e., identified as IP-enriched bins, the false probability of identifying the window as a peak region can be calculated as

$$\rho = \sum_{i=d}^{10} C_{10}^{i} q^{i} (1-q)^{10-i}, \tag{15}$$

by modeling the number of falsely discovered bins as a binomial random variable, where q denotes the false discovery rate of each bin, and d denotes a pre-specified threshold value. In this paper, we set d=5, which reflects our belief that a short region consisting of 5 or more true IP-enriched bins should be a true binding region. It is easy to see that ρ forms a valid p-value [9], which corresponds to the tail probability $P(i \geq d)$ with i being the number of falsely discovered bins in a window, and thus can be used for inference of peak regions. For a specified value of ρ, the value of q can be obtained by solving equation (15) using a numerical method or by trial and error. Given the value of q, the cutoff value of z can be identified as $z_c = \min\{z : \tilde{q}(z) \leq q\}$ where $\tilde{q}(z)$ is called q-value [10] and defined by

$$\tilde{q}(\mathrm{z}) = \inf_{\{\Lambda : z \in \Lambda\}} FDR(\Lambda) \tag{16}$$

That is, a bin with $z \geq z_c$ will be identified as significant bins.

Given the significant bins identified via (15) and (16), we consider a scanning procedure for peak calling: A sliding window with width of 10 bins moves along the genome with step size equal to the bin size; if the number of significant bins inside the window is greater than or equal to d, then all bins in the window are considered significant and the center of the window is marked as a candidate position for peak regions. The purpose of the scanning is to remove the false peak regions, which usually consist of only a few isolated significant locations. We find

that the choice of d=5 works well in practice. After scanning, those candidate positions separated by half of the moving window size 500 *bp* or less were merged together to form a predicted peak region, and after merging, the predicted peak regions having length less than half of the moving window size 500 *bp*, were considered spurious and removed.

Finally, we point out that by taking advantage of the separation between forward and reverse strand reads, other post processing steps can be taken to fine tune the peak regions boundaries, Zhang et al. [1] and Ji et al. [2] for more discussion. However, this is not the focus of this paper.

Results

H3K4me3 data

The H3K4me3 data was first studied by Spyrou *et al.* (2009) with BayesPeak algorithm. In that ChIP-seq experiment, a ChIP sample for trimethylated lysine 4 of histone H3 (H3K4me3) from livers of mice primarily of the Black 6 strain was obtained as well as a control sample without immunoprecipitation. A subset of the data, with genomic coordinate in the range of 9.0E7 to 9.7E7 *bp*, will be analyzed here mainly for testing purpose.

The MICS method was first applied to the dataset, 114 peak regions were identified with region FDR cut off at 0.01, i.e. $\rho \le 0.01$. Figure 1 shows the significant peak regions along with the count difference between ChIP and control samples in a certain genomic range. It indicates that MICS seems to offer reasonable results by consistently calling significant regions with relative high count difference along the genome.

For comparison, QuEST, MACS, PICS, BayesPeak and CCAT were also applied to this dataset. The same cutoff value 0.01 for region FDR was used for both MACS and PICS. Unfortunately, QuEST cannot provide FDR estimation since there is less number of tags in control condition than in ChIP condition in this dataset. For BayesPeak, inference is based on marginal posterior probability of each region with a natural threshold value equal to 0.5. With their default settings, QuEST, MACS, PICS and BayesPeak report 110, 76, 94 and 121 peak regions respectively. CCAT failed to report peak regions due to the relatively small number of reads in this dataset, which renders the noise rate inestimable. Figure 2 displays the significant peak regions identified by each method on the same genomic coordinates. The comparison shows that, visually, MICS, QuEST, MACS and BayesPeak produce similar results for this dataset, and PICS is relative conservative, which identifies only a subset of peak regions that identified by others. To get a better understanding for the performance of these methods, two Venn diagrams are drawn in Figure 3. Figure 3a shows that MICS and MACS perform very similarly for this example with the maximum overlap; and MICS shares large overlaps with QuEST, but with considerable number of independent calls. Figure 3b shows the Venn diagram for MICS and two Bayesian approaches. Obviously, PICS is the most conservative method among these three, as the peak regions identified by it are only a subset of others.

We note that for this example, the numbers of peak regions identified by different methods are different. The main reason is that some peak regions obtained by one method are split into multiple regions in another. Under such circumstance, the small adjacent regions are merged together to form a longer peak region when Venn

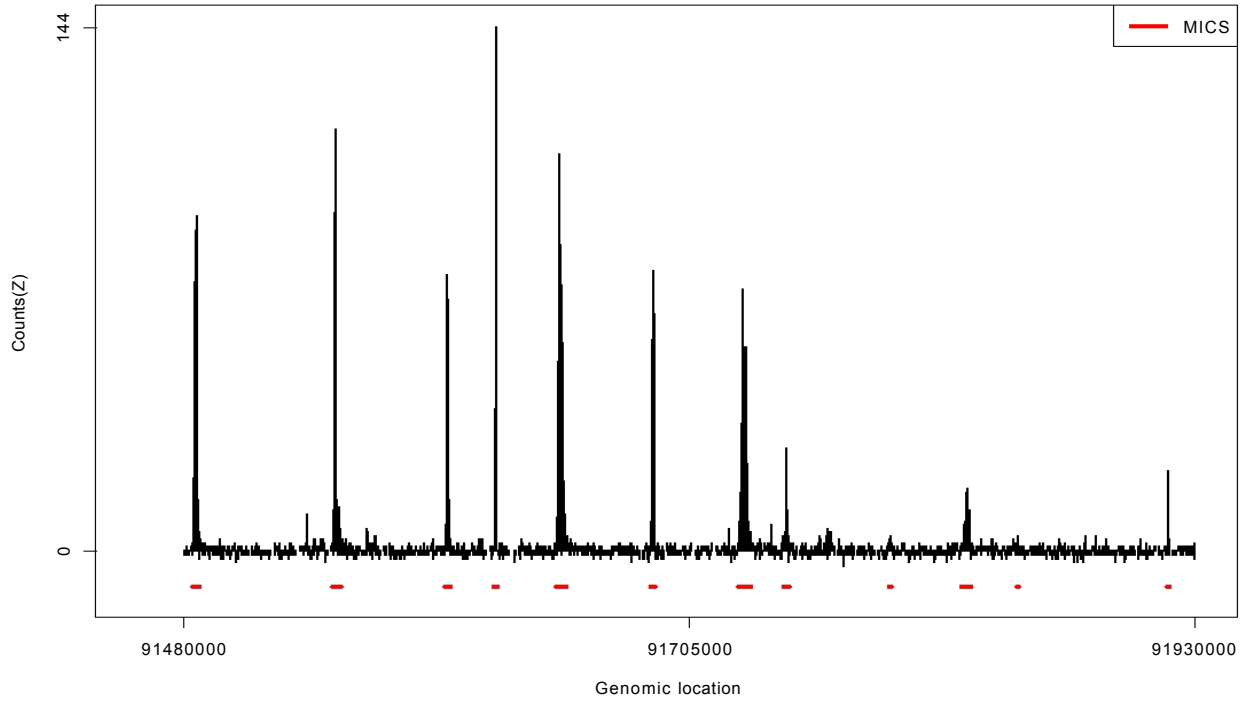

Figure 1: Partial results for the H3K4me3 data by MICS method. The black plots with peaksrepresent the count difference between ChIP and control samples. The red dash lines at the bottomare those significant peak regions identified by MICS with region FDR cutting off at 0.01.

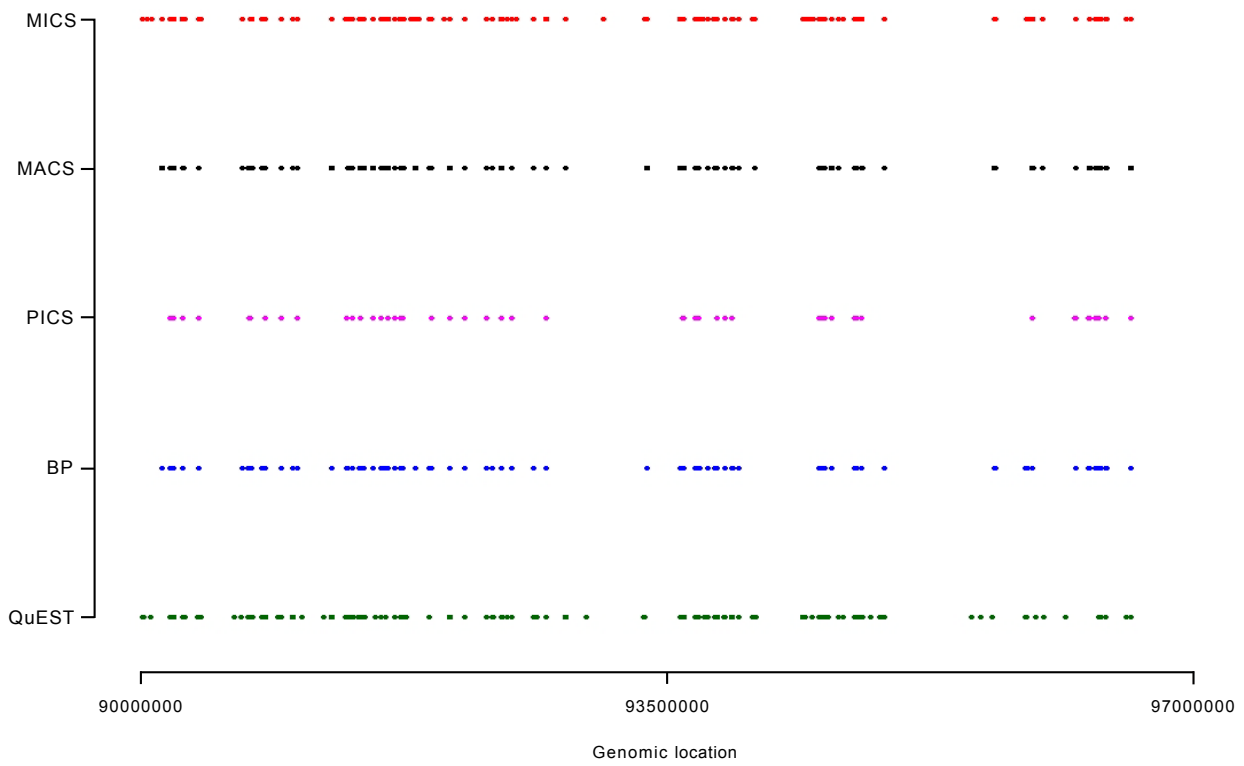

Figure 2: Comparison of MICS, MACS, PICS, BP and QuEST for the H3K4me3 data. The dash lines along the genome represent significant peak regions identified by each method with region FDR cutting off at 0.01.

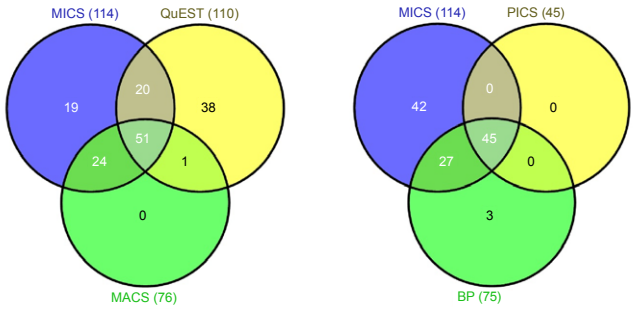

Figure 3: Venn diagram of the peak regions of the H3K4me3 data called by five methods: MICS, QuEST, MACS, PICS and BayesPeak. For some methods, the number of peak regions is different from the one reported in the paper, this is because some peak regions obtained by one method are split into multiple regions in another. Under such circumstance, the small adjacent regions are merged together to form a longer peak region.

diagram is constructed. We also observed that in ChIP-seq literature, the numbers of peak regions reported on the same dataset by the same method are sometimes inconsistent. This is understandable, since comparing peaks found by different methods are non-trivial, and researchers tend to merge the regions by their own criteria. To avoid such subjective post-processing steps, we propose to use the adjusted Rand index as a measure for the similarity of the peak regions resultant from different methods. The adjusted Rand index r is usually used in the literature of clustering and measures the degree of agreement between two partitions of the same set of observations even when the compared partitions have different numbers of clusters. When two partitions are identical, r is 1. When a partition is random, the expectation of r is 0. It is obvious that the problem of peak region identification in the ChIP-seq data analysis can also be viewed as a clustering problem; where the genome was partitioned into a series of segments, non-peak or peak regions, and each of the segments forms a cluster.

The adjusted Rand index is defined as follows. Let Ω denote a set of n observations, let $C = \{c_1, \ldots, c_s\}$ and $C' = \{c'_1, \ldots, c'_t\}$ represent two partitions of Ω, let n_{ij} be the number of observations that are in both cluster c_i and cluster c'_j, let $n_{i\cdot}$ be the number of observations in cluster c_i, and let $n_{\cdot j}$ be the number of observations in cluster c'_j. The adjusted Rand index is

$$r = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}}{\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] \Big/ 2 - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}} \quad (17)$$

A higher value of r means a higher correspondence between the two partitions. When the two partitions are identical, r is 1. When a

partition is random, the expectation of r is 0. Under the generalized hypergeometric model, it can be shown [11] that

$$E\left[\sum_{i,j}\binom{n_{ij}}{2}\right]=\left[\sum_{i}\binom{n_{i\bullet}}{2}\sum_{j}\binom{n_{\bullet j}}{2}\right]\Big/\binom{n}{2}$$

To compare the performance of different methods on the H3K4me3 data, we used the results from MICS method as the standard; that is if the partition is identical to that from MICS, r will be equal to 1. The results are summarized in Table 1. The rank of the methods' similarity with MICS based on adjusted Rand index is, MACS>QuEST>BayesPeak> PICS. They are consistent with our former analysis. MICS and MACS reported very similar results on this dataset with adjusted Rand index r equal to 0.96; while r is only 0.37 for PICS, which indicates a large discrepancy between PICS and MICS. For this particular data, PICS might be too conservative, given the fact that high overlaps exist between MICS and the other three methods.

H3K4me2

We applied MICS and the other five aforementioned methods to an in-house ChIP-seq dataset with both ChIP and control samples available. In this experiment, we studied the presences of Histone H3 lysine 4-di-methylation (H3K4me2) in mammary gland development and functional differentiation. Mammary Epithelial Cell (MEC) enriched organoid preparations were obtained by enzymatic digestion of mammary gland tissue of 12 weeks old virgin animals (staged for estrous cycle at diestrus) and differential sedimentation as described by Fata et al. [12]. We isolated chromatin from these mammary epithelial cells and performed ChIP with an antibody against H3K4me2 (07-030 Upstate-Millipore) according to Wagschal et al. [13]. Sequencing libraries for next generation sequencing were prepared according to Illumina ChIP-seq sample preparation protocol and sequenced on Illumina/Solexa GAII. Raw reads were mapped to mouse reference genome (NCB137/mm9) using Eland (Illumina) with maximally 2 mismatches tolerated.

For comparison, as before, all methods except for QuEST and BayesPeak used the same cutoff value 0.01 for region FDR, and QuEST used its default cutoff setting, while the threshold value 0.5 was used for marginal posterior probability of each region for BayesPeak. MICS, QuEST, MACS, PICS, BayesPeak and CCAT detected 2121, 999, 1963, 2222, 7567 and 5646 peak regions respectively. As mentioned above, the reported number of peak regions is sometimes deceitful due to the different post-processing filtering step. To fairly compare the performance of these methods on the H3K4me2 data, we calculate their adjusted Rand index using MICS partitions as the standard. The results are summarized in Table 2. The comparison shows that, overall, the six methods produced very similar results on this dataset with high adjusted Rand index achieved, and the result of MACS is most similar to that of MICS, with r equal to 0.9999.

Simulation Study

Semi-empirical method for ChIP-seq data simulation

We propose a semi-empirical method for simulating ChIP-seq data. Our method constructs tag positions based on real datasets,

Adjusted Rand Index	Method			
	MACS	QuEST	Bayes Peak	PICS
MICS	0.96	0.82	0.75	0.37

Table 1: Peak regions comparison based on adjusted Rand indices.

Adjusted Rand Index	Method				
	MACS	Bayes Peak	CCAT	QuEST	PICS
MICS	0.9999	0.9998	0.9996	0.9994	0.9981

Table 2: Results comparison based on adjusted Rand indices.

which offers two-fold benefits. It not only reduces the artificial effect of simulated data, but also imitates the location variation along genome of protein-DNA complexes. We illustrate this method through an example. Suppose we want to simulate some ChIP-seq datasets based on the in-house H3K4me2 dataset. Each simulated dataset has one control sample and one ChIP sample with K peak (ChIP-enriched) regions. For our method, the control sample is simply a copy of the control sample from real data. The strand information and genomic position of the first 105 reads of the H3K4me2 control sample are extracted as the control sample, and it is also the background base for constructing ChIP sample, specifically, based on the extracted control sample, we enriched K regions along the genome as our ChIP sample. The detailed simulation steps for ChIP samples are described below.

- Background base construction

The control sample is our building block for the background base of ChIP samples. Let the 10^5 reads denote 10^5 segments, with the length of each segment being equal to the position difference between the current and previous reads. The length of the first segment is the distance between the first read and the starting point, which can be set at a position slightly less than the first read. Next, we permute these segments and concatenate them from the starting point with strand information being carried. These will form the background base of ChIP samples.

- Peaks region enrichment

We randomly pick K reads from the background base as the start point of each peak region, while imposing a separation of at least 30000 *bps* between them. The length of each peak region, L, is uniformly drawn from l_{min} to l_{max} *bps*, with $\frac{1}{2}(l_{\max}+l_{\min})$ close to the average length of binding regions. Since the ChIP-DNA fragments are equally likely to be sequenced from both ends, the reads density around a true binding region should show a bimodal enrichment pattern [1,14]. In order to simulate the bimodal pattern, we enrich each peak region half by half. The first half is dominated by forward reads and the second half is by reverse reads. Let p_f and p_r denote the probability of getting a forward and reverse read respectively. In the first half enrichment, p_f is uniformly drawn from $p_f \sim Unif(\mathrm{p}_{\min},\mathrm{p}_{\max})$, and $p_r = 1-p_f$, where $0.5 < p_{\min} < p_{\max} < 1$. Let l denote the distance from current read to the new read. We model the occurrence of reads along the genome as a Poisson process, therefore l follows an exponential distribution $l \sim \exp(\lambda)$; λ can be uniformly drawn from $\lambda \sim \mathrm{Unif}(\lambda_{\min},\lambda_{\max})$ its inverse reflects the average distance between reads, which can be estimate from real data. After obtaining strand information and l for the new reads, we concatenate them from the start point of peak region until the middle of the region, the enrichment of the first half is completed. For the second half, all simulation steps remain the same except switching the probability of p_f and p_r. Please note, for a given peak region, we assume the probability of getting a dominant read and the mean rate λ to generate l are the same for the first half and the second half, however they could be different from region to region in order to reflect the local fluctuations and bias.

Comparison of π_0 estimation

Following the semi-empirical method, we simulated 10 datasets for comparing the performance of the empirical Bayes method and MICS in estimation of π_0. Each dataset has one control sample and one ChIP sample, and each ChIP sample is enriched with K=300 peak regions. For simplicity, we set the simulation parameters as follows: $(l_{\min}, l_{\max}) = [1000, 1000], (p_{\min}, p_{\max}) = [0.6, 0.9]$ and $(\lambda_{\min}, \lambda_{\max}) = [0.1, 0.1]$, that is, L is fixed to 1000 and, λ is fixed to 0.1. In order to compare the robustness of the two methods to the variation of ChIP samples, we modified the counts of ChIP sample locally in the counting step. Specifically, 3% of the N bins of the ChIP sample are randomly selected, a random count generated from $Pois(2)$ was added to ("+") or subtracted from ("–") the counts of selected bins. The former corresponds to strengthening the IP-enrichments, while the latter corresponds to weakening the IP-enrichments. We tried different cutoff $h(\mathrm{A}_0 = \{\mathrm{z}_i : | \mathrm{z}_i | \leq \mathrm{h}\})$. At each value of h, π_0 and $F_0(A_0)$ were estimated using both methods. The numerical results were summarized in Table 3. The comparison indicates that the empirical Bayes method could offer an acceptable estimate of π_0 under "+" case, while failed under "–" case. In the latter case, it produced unreasonable estimates of π_0, greater than 1.0 at h=0,1,2. However, MICS produced reasonable estimates of π_0 in both cases. The problem with the empirical Bayes method can be qualitatively explained as follows: When IP-enrichments are weak, F_0 and F_1 have much overlap and thus N', which is used as the denominator of $\tilde{F}_0(A_0)$, will be overestimated and $\tilde{F}_0(A_0)$ will be underestimated. This often leads to an estimate of π_0 greater than 1. In MICS, a simulation based method is used for estimating $F_0(A_0)$, which is independent of ChIP samples, and thus the performance of MICS is robust to the variation of ChIP samples.

The true value of π_0 is 0.9918, the estimate $\hat{\pi}_0$ and its standard deviation (the number in the parentheses) were calculated by averaging over the results of 10 datasets. A 3% of N bins of the ChIP sample are randomly selected, and a random count from Pois (2) was subtracted from (case 1) or added to (case 2) the counts of selected bins.

Comparison of MICS with other methods

To have a careful comparison of the performance between MICS and the other methods, we carried out a simulation study to assess the accuracy and efficiency of these algorithms. 10 datasets were simulated, each dataset has one control sample and one ChIP sample with K=30 peak regions. The simulation parameters were fixed at $(l_{\min}, l_{\max}) = (800, 1200), (p_{\min}, p_{\max}) = (0.6, 0.9)$ and $(\lambda_{\min}, \lambda_{\max}) = (0.1, 0.2)$. Figure 4 displays part of H3K4me2 data as well as a simulated dataset based on it. Both real and simulated ChIP sample clearly show a bimodal pattern in the peak regions. For evaluation purpose, we deliberately weaken the enrichment effect by choosing small values in the simulation study, which put greater challenge on the testing methods.

	MICS		Empirical Bayes	
	$\hat{\pi}_0(sd)$	$\hat{F}_0(A_0)$	$\tilde{\pi}_0(sd)$	$\tilde{F}_0(A_0)$
Case 1: subtracting				
0	0.9689 (0.0007)	0.641	**1.0147 (0.0004)**	0.612
1	0.9734 (0.0006)	0.928	**1.0069 (0.0004)**	0.897
2	0.9820 (0.0002)	0.987	**1.0024 (0.0002)**	0.967
3	0.9869 (0.0002)	0.998	0.9972 (0.0002)	0.987
Case 2: adding				
0	0.9691 (0.0010)	0.641	0.9679 (0.0005)	0.642
1	0.9738 (0.0004)	0.928	0.9733 (0.0004)	0.928
2	0.9824 (0.0003)	0.987	0.9823 (0.0002)	0.987
3	0.9872 (0.0001)	0.998	0.9872 (0.0001)	0.998

Table 3: Comparison of π_0 estimates between MICS and empirical Bayes methods at different cutoff values.

Method	ρ	Total	Match	F+	F−	CPU(s)
MICS	0.01	30.0 (0.00)	30.0 (0.00)	0.0 (0.00)	0.0 (0.00)	2.8
	0.05	30.0 (0.00)	30.0 (0.00)	0.0 (0.00)	0.0 (0.00)	2.8
	0.10	30.0 (0.00)	30.0 (0.00)	0.0 (0.00)	0.0 (0.00)	2.8
CCAT	0.01	0.0 (0.00)	0.0 (0.00)	0.0 (0.00)	30.0 (0.00)	0.7
	0.05	28.8 (1.23)	28.8 (1.23)	0.0 (0.00)	1.2 (1.23)	0.7
	0.10	28.8 (1.23)	28.8 (1.23)	0.0 (0.00)	1.2 (1.23)	0.7
MACS	0.01	28.1 (1.37)	28.1 (1.37)	0.0 (0.00)	1.9 (1.37)	12.5
	0.05	30.5 (0.85)	29.9 (0.32)	0.6 (0.70)	0.1 (0.31)	12.5
	0.10	31.1 (1.20)	30.0 (0.00)	1.1 (1.20)	0.0 (0.00)	12.5
PICS	0.01	22.0 (1.94)	22.0 (1.94)	0.0 (0.00)	8.0 (1.94)	8.6
	0.05	22.3 (1.89)	22.3 (1.89)	0.0 (0.00)	7.7 (1.89)	8.6
	0.10	22.3 (1.89)	22.3 (1.89)	0.0 (0.00)	7.7 (1.89)	8.5
QuEST	–	29.7 (0.48)	29.7 (0.48)	0.0 (0.0)	0.3 (0.48)	40.0
Bayes Peak	pp>0.5	34.2 (5.77)	30.0 (0.00)	4.2 (5.77)	0.0 (0.00)	2965.0

Table 4: Computational results for the 10 simulated datasets, where ρ denotes the cutoff value for region FDR; Total denotes the average number of peak regions identified by the method; Match denotes the average number of true peak regions discovered by the method; F+ denotes the average number of false positives; F− denotes the average number of false negatives. pp denotes the marginal posterior probability for each claimed region by Bayes Peak. The number in the parentheses is the standard error. Since the number of reads in control condition is less than that in ChIP condition for our simulated datasets, QuEST will not provide FDR estimation, where we put (−) for ρ.

We applied MICS and all other methods to the 10 simulated datasets. In order to assess the robustness of each method to the cutoff value for region FDR ρ, different choices are used in this study. The computational results are summarized in Table 4. Overall, the results show that MICS outperforms all other methods in terms of accuracy. It identified exactly the same peak regions as the true ones, and this perfect performance is also robust to the choice of the cutoff value for region FDR. In contrast, all the rest methods have some limitations. QuEST and CCAT actually worked very well for these datasets by only missing 0.3 and 1.2 true peak regions on average, without any false positives; MACS also worked decently in this study, it discovered all the 30 true regions when the cutoff value for ρ is released to 0.1, however, as the cutoff value increases, the number of false positives is increased as well; BayesPeak's results are acceptable, though it produced the highest number of false positives on average, all the true peak regions are identified; PICS is a relatively conservative method, even if the cutoff value is increased to 0.1, the number of false negatives is still as high as 7.7 on average.

Below we provide some explanations for the limitation of each of the methods under comparison. Both QuEST and MACS rely on bimodal peak pattern to estimate shift distance for reads. When the enrichment effect is weak and the bimodal pattern is less significant, they are likely to fail. Figure 5 shows a side-by-side comparison for a true peak region that is not detected by both methods and a discovered one, which shows some visual evidence for this claim. To provide some numerical evidence for this finding, we calculated the average value of the simulation parameters for the true positives (T+) and false negatives (F−), and summarized them in Table 5. Please note, MICS and BayesPeak are excluded from this table, since they have detected all the true peak regions. The λ, which reflects the intensity of enrichment effect, and L, the length of regions, are relative small

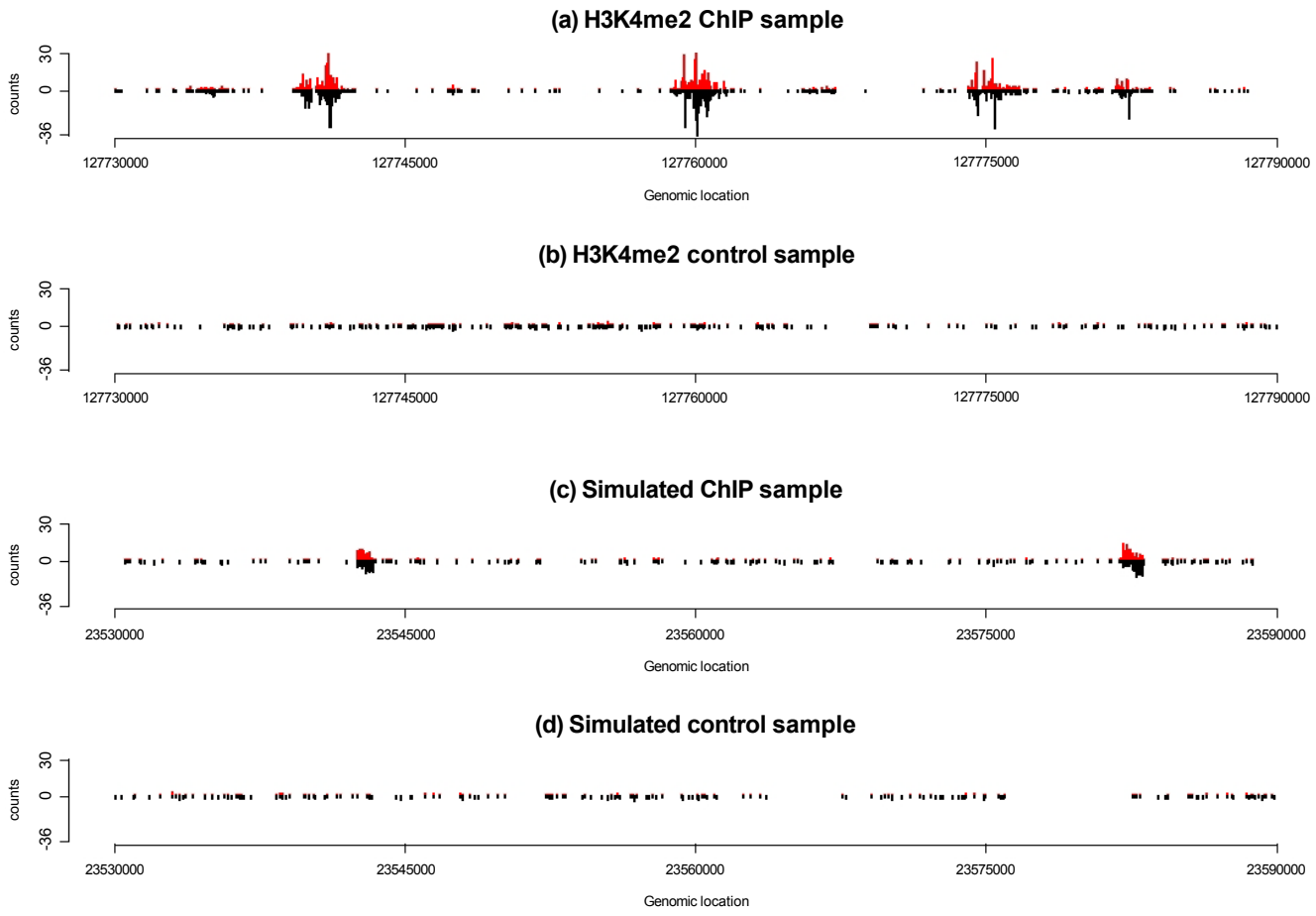

Figure 4: Partial view of the H3K4me3 dataset and a simulated dataset based on it. The red andblack lines represent the forward and reverse reads respectively. A clear bimodal enrichment patterncan be observed in both datasets.

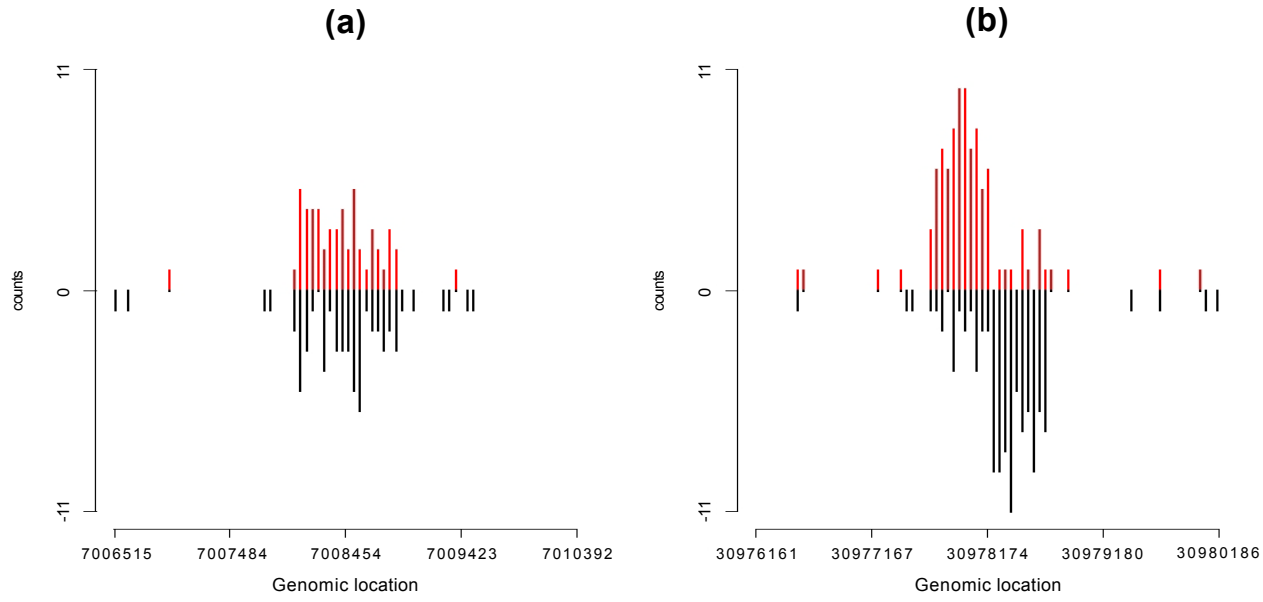

Figure 5: Two simulated peak regions. Both QuEST and MACS failed to detect (a), but succeededin finding (b).

Method	ρ	λ		L		p	
		T+	F−	T+	F−	T+	F−
QuEST	–	0.15 (0.03)	0.10 (0.01)	998 (115)	887 (80)	0.74 (0.09)	0.77 (0.09)
MACS	0.01	0.15 (0.03)	0.12 (0.02)	1002 (114)	926 (118)	0.74 (0.08)	0.70 (0.09)
PICS	0.01	0.15 (0.03)	0.15 (0.03)	977 (112)	1064 (99)	0.72 (0.08)	0.82 (0.05)
CCAT	0.05	0.15 (0.03)	0.14 (0.03)	997 (114)	1039 (156)	0.74 (0.09)	0.74 (0.10)

Table 5: Comparison of the average value of the simulation parameters for true positives(T+) and false negatives(F−), where ρ denotes the cutoff value for region FDR; λ denotes the mean rate for generating the distance from read to read; L denotes the length of peak region; p denotes the probability to generate dominate reads. The number in the parentheses is the standard error. Since the number of T+ and F- are not equal, the interpretation of the standard error should be cautious.

for the false negatives compared to the true positives. This provides a support for our statement, since a smaller value of will render weaker enrichment and shorter length of regions will blur the bimodal pattern. While these trends are not observed in CCAT and PICS. For CCAT, we found that the claimed FDR is incomparable with the FDR resultant from other methods, which deviates much from its nominal value. For example, when the cutoff value $\rho = 0.01$, it could not identify any peak regions; for some of the missed true peak regions, their FDRs can be greater than 0.9. The distribution assumption it relied on could be one reason for this inconsistency. For PICS, we did not observe much difference on parameters for the true positives and false negatives. Given its result that the numbers of regions in each category (Total, Match, F+ and F−) are almost the same as the FDR increases, we suspect that the pre-process step taken by PICS may adversely affect its performance. In the pre-process step of PICS, the genome is segmented into candidate regions based on the count of forward and reverse reads, if some true regions are missed in this step, they will not be reported by this method no matter what cutoff value for FDR is used later. For BayesPeak, a detailed examination of the posterior probability (pp) of those claimed regions reveals that, all the true peak regions have their pp close to 1, which proves the effectiveness of this method. However, BayesPeak is a very sensitive method with high possibility to report many false positives. Some small pseudo-peak regions ($\sim 200bp$), which are "enriched" just due to randomness, are reported with high pp in this simulation study.

In addition to the accuracy of peak region identification, we are also interested in comparing the efficiency of each method in terms of their CPU times cost by a single run. Obviously, MICS and CCAT are much more efficient than MACS, PICS and QuEST, and BayesPeak is the most expensive one. CCAT is a little faster than MICS; however, the subtle difference in efficiency could be due to the programming language they used. CCAT is written in C language and MICS is implemented in *R*.

Discussion

In this paper, we have proposed a new model-free method, MICS, for ChIP-seq data analysis, which carries a few advantages over the existing methods. Firstly, compared with model-based methods, MICS avoids assumptions for the data distribution; therefore, it is more robust to data variations than the model-based methods. Secondly, MICS employs a simulation-based method for estimating $F_0(A_0)$ and thus π_0 and FDR. Since the simulation-based method works independently of ChIP samples, MICS can perform robustly to the variation of ChIP samples. Our numerical results indicate that MICS can produce accurate identification of peak regions, even for those the IP-enrichments are weak. Thirdly, it is computationally efficient. It takes only a few seconds on a personal computer for a reasonably large dataset.

In this paper, we have also presented a semi-empirical method for ChIP-seq data simulation. Our method constructs tag position based on real datasets, which offers two-fold benefits: (i) The artificial effect of simulated data is minimized; and (ii) the information of location variation along the genome of protein-DNA complexes is kept and incorporated into the simulated data through the permutation step. The new simulation method can provide readers a powerful tool for evaluating the performance of different methods for ChIP-seq data analysis.

Finally, we want to point out that, although MICS, as a general method, can be applied to both Histon modification (HM) and transcription factor (TF) binding sites data by properly selecting the bin and window size, given the simulation natural of this method and the relative arbitrariness in bin and window size selection, it may be more suitable and robust for HM binding site identification, since HM considers broader peak regions and therefore less sensitive to the choice of bin and window size, whereas TF binding sites focuses on relatively sharp and narrow peak regions. It will be interesting to conduct a thorough comparison between MICS and algorithms that is specifically suitable for Histon modification data, e.g. MOSAiCS, which we will explore elsewhere [15-17].

References

1. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M (2008) Modeling ChIP sequencing in silico with applications. PLoS Comput Biol 4: e1000158.
2. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26: 1293-1300.
3. Xu H, Handoko L, Wei X, Ye C, Sheng J, et al. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics 26: 1199-1204.
4. Spyrou C, Stark R, Lynch AG, Tavaré S (2009) BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics 10: 299.
5. Zhang X, Robertson G, Krzywinski M, Ning K, Droit A, et al. (2011) PICS: probabilistic inference for ChIP-seq. Biometrics 67: 151-163.
6. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829-834.
7. Wu M, Liang F, Tian Y (2009) Bayesian modeling of ChIP-chip data using latent variables. BMC Bioinformatics 10: 352.
8. Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99: 96-104.
9. Casella G, Berger RL (2002) Statistical Inference (2ndedn). Thomson Learning, USA.
10. Storey JD (2002) A direct approach to false discovery rates. JR Statist Soc B 64: 479-498.
11. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193-218.
12. Fata JE, Mori H, Ewald AJ, Zhang H, Yao E, et al. (2007) The MAPK(ERK-1,2) pathway integrates distinct and antagonistic signals from TGFalpha and FGF7 in morphogenesis of mouse mammary epithelium. Dev Biol 306: 193-207.
13. Wagschal A, Delaval K, Pannetier M, Arnaud P, Feil R (2007) Chromatin Immunoprecipitation (ChIP) on Unfixed Chromatin from Cells and Tissues to Analyze Histone Modifications. CSH Protoc 2007: pdb.
14. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359.
15. Humburg P (2013) Simulation of ChIP-seq experiments.
16. Ma W, Wong WH (2011) The analysis of ChIP-Seq data. Methods Enzymol 497: 51-73.
17. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.