BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 20 2008, pages 2344–2349
doi:10.1093/bioinformatics/btn402

Gene expression

An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data

Han Xu[1,2], Chia-Lin Wei[3], Feng Lin[2,*] and Wing-Kin Sung[1,4,*]

[1]Computational & Mathematical Biology Group, Genome Institute of Singapore, 138672 Singapore, [2]School of Computer Engineering, Nanyang Technological University, 637553 Singapore, [3]Genome Technology & Biology Group, Genome Institute of Singapore, 138672 Singapore and [4]School of Computing, National University of Singapore, 117543 Singapore

Received on April 9, 2008; revised on July 13, 2008; accepted on July 28, 2008

Advance Access publication July 29, 2008

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Epigenetic modifications are one of the critical factors to regulate gene expression and genome function. Among different epigenetic modifications, the differential histone modification sites (DHMSs) are of great interest to study the dynamic nature of epigenetic and gene expression regulations among various cell types, stages or environmental responses. To capture the histone modifications at whole genome scale, ChIP-seq technology is becoming a robust and comprehensive approach. Thus the DHMSs are potentially identifiable by comparing two ChIP-seq libraries. However, little has been addressed on this issue in literature.

Results: Aiming at identifying DHMSs, we propose an approach called ChIPDiff for the genome-wide comparison of histone modification sites identified by ChIP-seq. Based on the observations of ChIP fragment counts, the proposed approach employs a hidden Markov model (HMM) to infer the states of histone modification changes at each genomic location. We evaluated the performance of ChIPDiff by comparing the H3K27me3 modification sites between mouse embryonic stem cell (ESC) and neural progenitor cell (NPC). We demonstrated that the H3K27me3 DHMSs identified by our approach are of high sensitivity, specificity and technical reproducibility. ChIPDiff was further applied to uncover the differential H3K4me3 and H3K36me3 sites between different cell states. Interesting biological discoveries were achieved from such comparison in our study.

Availability: http://cmb.gis.a-star.edu.sg/ChIPSeq/tools.htm

Contact: asflin@ntu.edu.sg; sungk@gis.a-star.edu.sg

Supplementary information: Supplementary methods and data are available at *Bioinformatics* online.

1 INTRODUCTION

Eukaryotic DNA is packaged into a chromatin structure consisting of repeating nucleosomes by wrapping DNA around histones. The histones are subject to a large number of post-translational modifications such as methylation, acetylation, phosphorylation and ubiquitination. The histone modifications are implicated in influencing gene expression and genome function. Considerable evidence suggests that several histone methylation types play crucial roles in biological processes (Martin and Zhang, 2005). A well-known example is the repression of development regulators by tri-methylation of histone H3 lysine 27 (H3K27me3, or K27) in mammalian embryonic stem cell (ESC) to maintain stemness and cell puripotency (Bernstein *et al*., 2006; Boyer *et al*., 2006). Some epigenetic stem cell signature of K27 is also found to be cancer specific (Widschwendter *et al*., 2007). Moreover, the tri- and di-methylation of H3 lysine 9, are implicated in silencing the tumor suppressor genes in cancer cells (McGarvey *et al*., 2006). In the light of this, the specific genomic locations with differential intensity of histone modifications, which are called differential histone modification sites (DHMSs) in this article, are of great interest in the comparative study among various cell types, stages or environmental response.

The histone modification signals can be captured by chromatin immunoprecipitation (ChIP), in which an antibody is used to enrich DNA fragments from modification sites. Several ChIP-based techniques, including ChIP-chip, ChIP-PET and ChIP-SAGE, have been developed in the past decade for the study of histone modification or transcription factor binding in large genomic regions (Impey *et al*., 2004; Kim and Ren, 2006; Wei *et al*., 2006). With the recent advances of ultrahigh-throughput sequencing technologies such as Illumina/Solexa sequencing, ChIP-seq is becoming one of the main approaches for its high coverage, high resolution and low cost, as demonstrated in several published work (Barski *et al*., 2007; Johnson *et al*., 2007; Mardis, 2007). The basic idea of ChIP-seq is to read the sequence of one end of a ChIP-enriched DNA fragment, followed by mapping the short read called *tag* to the genome assembly in order to find the genomic location of the fragment. Millions of tags sequenced from a ChIP library are mapped and form a genome-wide profile in which ChIP fragment counts are overrepresented in histone modification sites or transcription factor binding sites.

Inspired by the success of ChIP-seq in identifying histone modification sites in a single library, we asked if the DHMSs could be identified by computationally comparing two ChIP-seq libraries generated from different cell types or experimental conditions. Mikkelsen *et al.* (2007) mapped the H3K4me3 (K4) and K27 sites in mouse ESC, neural progenitor cell (NPC) and embryonic fibroblast (MEF) and compared the occurrence of modification sites in promoter regions across three cell types. A limitation of

*To whom correspondence should be addressed.

their study is that the modification sites are compared qualitatively but not quantitatively. An example demonstrating this limitation is the regulation of Klf4 by K4, which is known to be positively correlated to gene expression. The Klf4 promoter was flagged as 'with K4' in both ESC and NPC by qualitative analysis, hence it could not explain the up-regulation of Klf4 in ESC. On the other hand, quantitative comparison indicated the intensity of K4 in Klf4 promoter is more than 5-fold higher in ESC than in NPC, consistent with the observation of expression change.

To the best of our knowledge, little has been published in literature on the quantitative comparison of two ChIP-seq libraries in genome wide. Triggered by the idea from microarray analysis (Quackenbush, 2002), a simple solution to this problem is to partition the genome into bins and to compute the fold-change of the number of ChIP fragments in each bin. However, fold-change approach is sensitive to the technical variation caused by random sampling of ChIP fragments. In this article, we propose an approach called ChIPDiff to improve the fold-change approach by taking into account the correlation between consecutive bins. We modeled the correlation in a hidden Markov model (HMM) (Rabiner, 1989), in which the transmission probabilities were automatically trained in an unsupervised manner, followed by the inference of the states of histone modification changes using the trained HMM parameters.

To evaluate the performance of ChIPDiff, we first compared the K27 libraries between ESC and NPC based on Mikkelsen *et al.*'s dataset. We identified 4722 K27 DHMS regions in genome wide. Three lines of evidence showed that the performance is satisfactory: (a) sensitivity: in highly conserved non-coding elements (HCNEs) studied by Bernstein *et al*. (2006), 80% of DHMSs inferred from gene expression were identified by ChIPDiff; (b) specificity: based on comparison between non-cell-specific controls, we approximated a false positive rate of 0.19% for the identified DHMS regions; (c) reproducibility: checking the intersection of the results on two independent subsets, we showed that 57.4% of the DHMSs were technically reproducible, conditional on the sequencing depth of 3–4 million tags. The evaluation also demonstrated that our method outperforms fold-change approach and qualitative analysis in all the three aspects.

We further applied ChIPDiff to H3K4me3 (K4) and H3K36me3 (K36) for the discovery of DHMSs on these two types of histone modifications and studied their potential biological roles in stem cell differentiation. Several interesting biological discoveries were achieved in the study.

2 METHOD

2.1 Determining putative histone modification sites

Given two ChIP-seq libraries, L_1 and L_2, the first step for identifying DHMSs is to determine the putative sites that involve histone modifications either in L_1 or L_2. This section details this step.

Tags in the raw data generated from a ChIP-seq experiment were mapped onto the genome to obtain their positions and orientations. Due to the PCR process in ChIP-seq experiments, multiple tags may be derived from a single ChIP fragment. To remove the redundancy, tags mapped to the same position with the same orientation were treated as a single copy. Note that in ChIP-seq protocol a tag is retrieved by sequencing one end of the ChIP fragment, of which the median length is around 200 bp (Barski *et al*., 2007; Robertson *et al*., 2007). Hence, we approximated the center of corresponding ChIP fragment by shifting the tag position by 100 bp towards its orientation. The whole genome was partitioned into 1 kb bins and the number of centers of ChIP fragments were counted in each bin.

After the above preprocessing procedures, a profile of ChIP fragment counts was generated. Considering a genome with m bins, the profiles of L_1 and L_2 are represented as $X_1=\{x_{1,1},x_{1,2},\ldots,x_{1,m}\}$ and $X_2=\{x_{2,1},x_{2,2},\ldots,x_{2,m}\}$, respectively, where $x_{i,j}$ is the fragment count at the j-th bin in L_i. To depict the combined enrichment of fragments in each bin, we defined a score F normalized against the sequencing depth:

$$F(i)=\frac{x_{1,i}}{n_1}+\frac{x_{2,i}}{n_2}, \qquad i=1,2,\ldots,m$$

where n_1 and n_2 are the total number of sequenced fragments in L_1 and L_2, i.e. $n_1=\sum_i x_{1,i}$, $n_2=\sum_i x_{2,i}$.

Mikkelsen *et al*. (2007) and Robertson *et al*. (2007) pointed out that not all the bins can be interrogated in the tag mapping procedure, mainly due to the existence of repeat region. Let η denote the fraction of 'valid' bins in the genome, the expectation of the score F at a 'valid' bin is $\sum_i F(i)/(m\times\eta)$, which equals to $2/(m\times\eta)$. (Mikkelsen *et al*., 2007) estimated $\eta\approx 0.7$ for mouse genome. If a bin has an F-score $>2/(m\times\eta)$, we flagged it as a putative histone modification site. Consecutive modification sites within 1 kb apart from each other were merged into histone modification regions.

2.2 Quantitative comparison of modification intensity by fold-change

For the convenience of denotation and description, the methods in the rest of the article will be introduced based on the putative modification regions defined in Section 2.1. Considering a region that consists of k bins, we use the notation $x_{1,i}$, $x_{2,i}$ for the ChIP fragment counts in L_1 and L_2, respectively, at the i-th bin in that region ($i=1,2,\ldots,k$).

Histone modifications exhibit a variety of kinetics and stoichiometries (Gan *et al*., 2007). For a ChIP-seq experiment, we define the modification intensity at the i-th bin in library L_j to be the probability of an arbitrary ChIP fragment captured from the i-th bin in the ChIP process, denoted $p_{j,i}$. Since the extraction and sequencing of ChIP fragments is a random sampling process, the posterior probability of observing $x_{j,i}$ fragments at the i-th bin in library L_j, conditional on the intensity $p_{j,i}$, approximately follows a binomial distribution:

$$\Pr(x_{j,i}|p_{j,i})=\binom{n_j}{x_{j,i}}p_{j,i}^{x_{j,i}}(1-p_{j,i})^{n_j-x_{j,i}} \tag{1}$$

We further assume that the prior probability of $p_{j,i}$ follow a beta distribution:

$$\Pr(p_{j,i})=\frac{1}{B(\alpha,\beta)}p_{j,i}^{\alpha-1}(1-p_{j,i})^{\beta-1} \tag{2}$$

where $B(\alpha,\beta)$ is the β-function. Note that β-distribution is the conjugate prior of binomial (Raiffa and Schaifer, 2000), hence the conditional probability $\Pr(p_{j,i}|x_{j,i})$ also follows a β-distribution, with the expectation $E(p_{j,i}|x_{j,i})=(\alpha+x_{j,i})/(\alpha+\beta+n_j)$. In our application, the parameters α and β are set to be 1 and m, respectively, where m is the total number of bins in the genome. (see the Supplementary Methods for details).

We define a DHMS as a bin in which the ratio of intensities between L_1 and L_2 is larger than τ (L_1-enriched DHMS) or smaller than $1/\tau$ (L_2-enriched DHMS), where τ is a predetermined threshold, and $\tau\geq 1.0$. A simple solution for identifying DHMSs is to estimate the fold-change of expected intensity (preferably in term of log-ratio) from the ChIP fragment counts, as follow:

$$\log\left[\frac{E(p_{1,i}|x_{1,i})}{E(p_{2,i}|x_{2,i})}\right]=\log\left[\frac{(\alpha+x_{1,i})(\alpha+\beta+n_2)}{(\alpha+x_{2,i})(\alpha+\beta+n_1)}\right] \tag{3}$$

An example of the log-ratio estimation based on (3) is shown in Figure 1a.

A drawback of the fold-change approach is that it is prone to the technical variation caused by random sampling. Figure 1b shows an RI-plot (Quackenbush, 2002) to depict the variation of the log-ratio dependent on the intensity. When the intensity is relatively small, the variation of log-ratio becomes too high, which may result in considerable false positives.

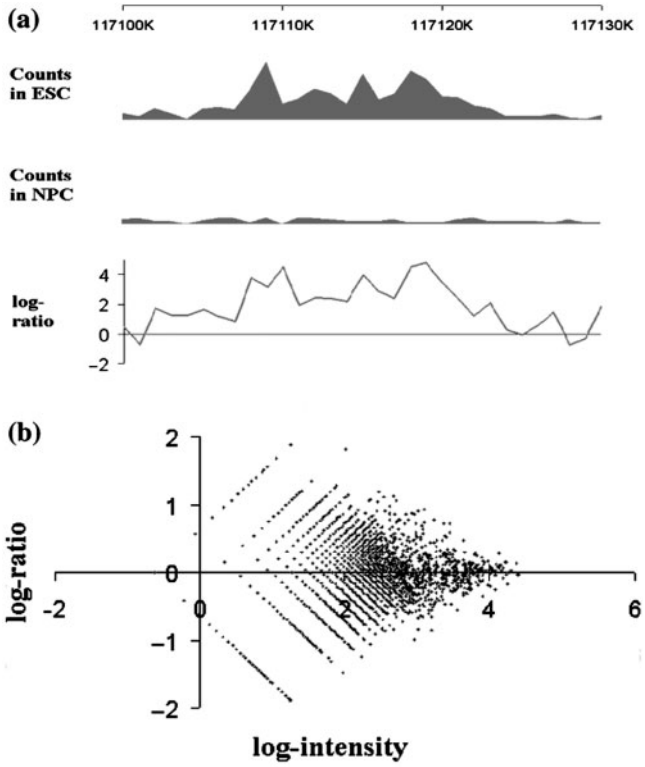


Fig. 1. (**a**) An example of the log-ratio estimation of H3K27me3 intensity between mouse ESC and NPC. Bin size set to be 1k; displayed genomic region range from chr14:117 100 000 to 117 130 000; data retrieved from Mikkelsen *et al.*'s (2007) dataset; (**b**) An RI-plot for chromosome 19 in K27 data.

2.3 An HMM-based approach for identifying DHMSs

Histone modifications usually occur in continuous regions that span a few to hundreds or even thousands of nucleosides. Hence, one may expect strong correlation between consecutive bins in the measurements of intensity changes. This argument is supported by our observations from ChIP-seq profile. As an example, the log-ratio profile in Figure 1a has an autocorrelation of 0.84. In ChIP-chip data analysis, Li *et al*. (2005) have designed an HMM to model the correlation of signals between consecutive probes and successfully applied it for the identification of p53 binding site, suggesting the potential ability of HMM for identifying DHMSs in our study. Here, we propose a HMM-based approach called ChIPDiff to solve the problem.

We denote s_i to be the state of histone modification change at the i-th bin ($i = 1, 2, \ldots, k$). Based on the definition of DHMS in Section 2.2, the state s_i takes one of the following three values:

- α_0: non-differential site, if $1/\tau \leq p_{1,i}/p_{2,i} \leq \tau$;
- $\alpha_1 : L_1$-enriched DHMS, if $p_{1,i}/p_{2,i} > \tau$;
- $\alpha_2 : L_2$-enriched DHMS, if $p_{1,i}/p_{2,i} < 1/\tau$.

We modeled the inter-bin correlation as a first-order Markov chain such that $\Pr(s_i|s_0, s_1, \ldots, s_{i-1}) = \Pr(s_i|s_{i-1})$, where S_0 is the start state before the first bin in the region. A HMM was implemented to infer the posterior probability distribution of the states from the observations of fragment counts. The HMM is characterized by three features: the prior probability of the start state S_0, the emission probability $\Pr(x_{1,i}, x_{2,i}|s_i)$ and the transmission probability $\Pr(s_i|s_{i-1})$.

The initial state S_0 is fixed to take the value α_0 since we assume the region starts from the genomic locations where histone modification is depleted in both libraries.

We derived the emission probability $\Pr(x_{1,i}, x_{2,i}|s_i)$ by integrating $p_{1,i}$ and $p_{2,i}$ over all possible values constrained by S_i:

$$\Pr(x_{1,i}, x_{2,i}|s_i) = \frac{\iint_{s_i} \Pr(x_{1,i}|p_{1,i})\Pr(x_{2,i}|p_{2,i})\Pr(p_{1,i})\Pr(p_{2,i})dp_{1,i}dp_{2,i}}{\iint_{s_i} \Pr(p_{1,i})\Pr(p_{2,i})dp_{1,i}dp_{2,i}} \quad (4)$$

Readers may refer to the Supplementary Methods for the detailed derivation. In Equation (4), $\Pr(x_{j,i}|p_{j,i})$ follows binomial distribution defined in (1) and $\Pr(p_{j,i})$ follows β-distribution defined in (2).

The transmission probability table was trained using the Baum–Welch algorithm (Baum *et al*., 1970), which takes expectation maximization (EM) steps to iteratively estimate the parameters of the HMM from hidden states in an unsupervised manner. In the training process, the transmission parameters were initialized to be uniform and the start state S_0 and the emission probabilities were fixed as described above. Since the transmission probability table is identical in the whole genome, it was trained by cumulating the transmission frequencies in all the putative histone modification regions.

In the last step of ChIPDiff, the probability distributions of the states in each bin were inferred using forward–backward algorithm. Bins with posterior probability larger than a confidence threshold $\rho (0 < \rho < 1)$ for $s_i = \alpha_1$ or $S_i = \alpha_2$ were identified as DHMSs. Consecutive DHMSs with no gap between them were merged into DHMS regions.

The most computationally expensive step in ChIPDiff is the training of the transmission probability table. Two strategies were employed to reduce the computational cost: (a) the integrals for emission probabilities were numerically computed and were compiled into a lookup table, prior to the training of the HMM; (b) we allowed the transmission probability table to be trained based on a subset randomly selected from the putative histone modification regions.

3 RESULTS

We applied ChIPDiff to Mikkelsen *et al.*'s dataset publicly available at http://www.broad.mit.edu/seq_platform/chip/. The performance of ChIPDiff was evaluated by comparing the H3K27me3 (K27) libraries between mouse ESC and NPC. We further applied ChIPDiff to H3K4me3 (K4) and H3K36me3 (K36) data for the discovery on DHMSs and studied their potential biological roles in stem cell differentiation.

3.1 Evaluation on H3K27me3 data

H3K27me3 was selected for the evaluation since its DHMSs in HCNEs have been implicated in literature (Bernstein *et al*., 2006). Moreover, K27 preferentially marks gene region and functions as a repressor, which facilitated our indirect validation using expression data. We compared the K27 ESC library and NPC library with ChIPDiff, in which the fold-change threshold τ was set to be 3.0 and the confidence threshold ρ was set to be 0.95. The HMM was trained with 10 000 randomly selected histone modification regions. A total of 26 230 bins were identified to be DHMSs, corresponding to 4722 continuous regions. Among them, 3833 (81.2%) regions are ESC enriched and 889 (18.8%) are NPC enriched, implying a global trend of K27 depletion upon cell differentiation.

We first assessed the capability of ChIPDiff in identifying the biologically significant DHMSs, i.e. sensitivity. Bernstein *et al*. (2006) have shown that K27 is enriched in HCNEs in ESC, repressing a number of development regulators to maintain the

stemness of the cell. These histone marks are depleted in diverse differentiated cells. From HCNEs, we selected 223 genes of which the expressions were studied by (Mikkelsen *et al*., 2007). Since K27 functions as a gene repressor, we reasoned that some of those HCNE genes marked by K27 will be up-regulated in NPC and DHMSs should be identified at these genes. As expected, a subset containing 30 genes were determined to be up-regulated with the criterion of more than 4-fold. Among them, 24 (80%) are marked by DHMSs identified by ChIPDiff in promoter region ± 1 kb from transcription start site (TSS). In contrast, only 37 (19.2%) out of the 193 genes that are not up-regulated in NPC are marked by DHMSs.

To test the specificity of ChIPDiff result, we need to estimate the fraction of falsely identified DHMS regions that are not cell specific. For this purpose, we partitioned each library into two technical replicates: $L_{ESC,rep1}$ and $L_{ESC,rep2}$ for ESC, $L_{NPC,rep1}$ and $L_{NPC,rep2}$ for NPC. The replicates consist of tags retrieved from different lanes in ChIP-seq experiments, with similar sequencing depth (see Supplementary Table 4 for the assignment of lanes to replicates). We generated two new libraries by merging the tags in $L_{ESC,rep1}$ and $L_{NPC,rep1}$, $L_{ESC,rep2}$ and $L_{NPC,rep2}$, respectively. Since the replicates are of similar sequencing depth, the difference between these two libraries should not be cell specific and only reflect the technical variations in the experiments. Comparing these non-cell-specific controls, nine differential regions were identified by ChIPDiff. Hence, we approximated a false positive rate of 0.19% (9/4722) for the DHMS regions identified in cell-specific comparison.

We also tested the reproducibility by conducting two independent passes of cell-specific comparison: $L_{ESC,rep1}$ versus $L_{NPC,rep1}$ and $L_{ESC,rep2}$ versus $L_{NPC,rep2}$. To measure the reproducibility, we defined a score as the ratio of the number of DHMSs identified in both passes, to the average number of DHMSs in individual pass. As the result, we obtained a reproducibility score of 57.4% for ChIPDiff. Note that the reproducibility is conditional on the sequencing depth of the replicates, which ranges from 3 to 4 million tags in our assessment (Supplementary Table 4).

To compare the performance among different methods, we repeated the sensitivity, specificity and reproducibility tests for fold-change and qualitative method. In qualitative method, K27 modification sites were identified for ESC and NPC individually using the binning approach proposed by (Mikkelsen *et al*., 2007), and bins marked as K27 site in only one cell type were identified to be DHMSs. Consecutive DHMSs were merged into DHMS regions as well. For a fair comparison, the thresholds were adjusted to allow similar number of DHMS regions to be identified for all three methods (The numbers are not identical because the thresholds take discrete values). The evaluation results are summarized in Table 1. ChIPDiff outperformed the other two methods in all three aspects. Fold-change approach and qualitative method had much higher false positive rates, indicating these methods are sensitive to technical variation and experimental bias.

Table 1. Comparison of the performance of ChIPDiff, fold-change approach and qualitative method, based on H3K27me3 data

	ChIPDiff	fold-change	qualitative method
No. of DHMS regions in cell-specific comparison	4722	4,958	4,790
FPR estimated from non-cell-specific control (%)	0.19	10.8	52.3
Detection rate on HCNE DHMSs (%)	80.0	63.3	73.3
Reproducibility score (%)	57.4	23.4	43.8

FPR refers to false positive rate.

Table 2. A summary of DHMSs identified from H3K4me3 and H3K36me3 libraries

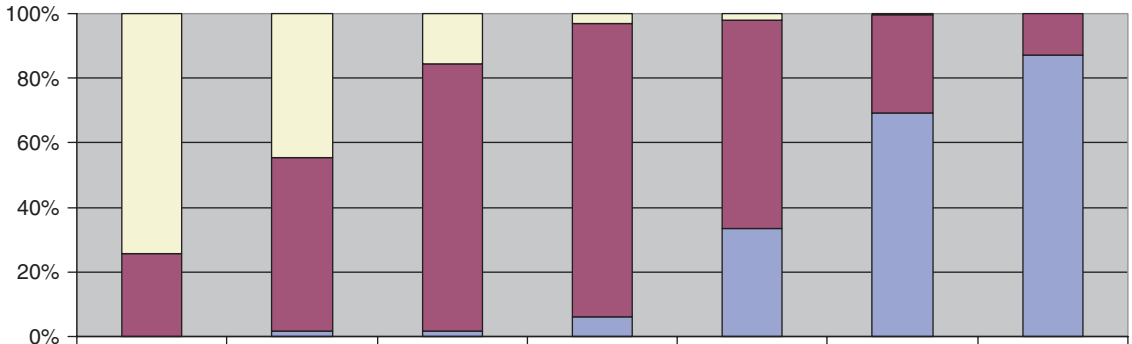
	H3K4me3		H3K36me3	
	ESC enriched	NPC enriched	ESC enriched	NPC enriched
No. of DHMS bins	32 384	3742	15 111	16 719
No. of DHMS regions	12 976	1768	1158	1228
No. of RefSeq genes marked	3877	211	747	417

3.2 Application to H3K4me3 and H3K36me3 data

We extended our study to trimethylations on K4 and K36. Both histone modification types positively regulate gene expression but in different manner. Guenther *et al*. (2007) revealed K4 marks the active promoters where the transcription of the genes is initiated, while K36 occupies the gene region as a hallmark of elongation. Our previous study (Zhao *et al*., 2007) also showed that K4, together with K27, establishes distinct genomic domains of active and inactive chromatin structures in human ESC. Thus, it attracted our interest to study the DHMSs of these histone modifications between ESC and NPC. Moreover, K4 sites usually appear in punctated pattern sharply around TSS in ChIP-seq profile, while K36 sites appear in a much broader pattern, providing a comprehensive test-bed for evaluating the adaptability of our approach to diverse histone modification types.

We processed the libraries with the same ChIPDiff configurations as mentioned in Section 3.1. The results are summarized in Table 2. Consecutive bins identified as DHMSs were merged into regions. Strikingly, the number of ESC-enriched K4 DHMSs is much larger than NPC-enriched ones. Considering such imbalance was also observed for K27, we hypothesized it may be associated with the bivalent chromatin structure marked by K4 and K27 (Bernstein *et al*., 2006). In further analysis, we found 1961 (51.2%), out of 3833 ESC-enriched K27 DHMS regions overlap with ESC-enriched K4 DHMSs. In contrast, K36 and K27 seemed to be mutually exclusive: only 8 (0.21%) of these 3833 regions overlap with ESC-enriched K36 DHMSs.

To study the correlation between DHMSs and gene expression, we annotated the RefSeq genes with DHMS regions and expression data published by Mikkelsen *et al*. (2007). RefSeq genes were retrieved from UCSC database (Pruitt *et al*., 2005). To remove the redundancy, the longest ORF was selected for gene annotation if multiple transcripts are mapped to the same gene, which resulted in 18 795 unique genes in total. As shown in Figure 2, K4 and K36 co-regulate the gene expression with strong significance.



	K4&K36 in ESC	only K36 enriched in ESC	only K4 Enriched in ESC	None	only K4 enriched in NPC	only K36 enriched in NPC	K4&K36 enriched in NPC
up-regulated in ESC	313	76	402	290	2	1	0
not differentially expressed	109	91	2124	8480	66	80	10
up-regulated in NPC	0	3	44	572	34	183	67

Fig. 2. Combinatorial effect of H3K4me3 and H3K36me3 on gene expression between ESC and NPC. Up/down-regulations were determined by the criterion of 4-fold change.

This observation is consistent with the conclusion by Guenther *et al*. (2007). Among 1,085 genes up-regulated in ESC, 791 (72.9%) are associated with ESC-enriched K4 or K36 DHMSs, suggesting that the gene expression is potentially predictable from DHMSs. Notably, two key transcription factors in ESC, Nanog and Oct4, are marked by DHMSs of both K4 and K36, implying the critical roles played by these histone modification marks in ESC by interfering the transcription regulatory network.

4 DISCUSSION AND FUTURE WORK

In this article, we developed an HMM-based approach called ChIPDiff for the genome-wide quantitative comparisons of histone modifications derived by sequencing approach. Using stem cell as a model system, we applied ChIPDiff to identify the DHMSs between ESC and NPC. We showed that ChIPDiff is able to render robust and technically reproducible results though evaluation with K27 data. We further applied ChIPDiff to K4 and K36 data and achieved several interesting biological discoveries. The experimental results indicate that ChIPDiff is a useful tool for the comparative study of histone modifications between cell types or different biological treatments.

Nevertheless, there are several limitations. First, the bin size was set to be 1 kb in ChIPDiff, of which the resolution is relatively low when considering the nucleosome size of 200 bp (including the linker). The resolution, however, is limited due the sequencing depth: if we reduce the bin size, there would not be enough fragment counts to be included in a bin for a reliable prediction. Second, the specificity, sensitivity and repeatability of our approach were evaluated based on technical replicates or a limited list of DHMSs inferred from biological knowledge and gene expression. There might be an argument on whether these data provide a 'golden' standard for the evaluation. In fact, such a 'golden' standard is very difficult to define for most biological data. Third, we used the total number of ChIP fragments for the normalization against sequencing depth. This normalization procedure is subject to the noise level of ChIP experiment. As an alternative, qPCR measurements (Ding and Cantor, 2004) on a few 'control' sites may provide a better way for normalization. And finally, in the preprocessing step, multiple tags retrieved from different fragments and mapped to the same genomic location were counted only once, which may result in error in quantitative measurement. We look forward to deeper investigations on these problems.

Except for histone modification, ChIP-seq has been widely used in the research on transcription factor binding, chromatin accessibility and nucleosome locations. It would be interesting to evaluate the possibility of expanding the applications of ChIPDiff to these genomic features. Unlike histone modifications, transcription factors bind to a specific genomic location with a very sharp ChIP-seq pattern, hence the correlation between consecutive bins is relatively weaker. We are studying new mathematical model for the quantitative comparison of transcription factor binding in future work.

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Baum,L.E. *et al.* (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

Bernstein,B.E. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

Boyer,L.A. *et al*. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.

Ding,C. and Cantor,C.R. (2004) Quantitative analysis of nucleic acids – the last few years of progress. *J. Biochem. Mol. Biol.*, **37**, 1–10.

Gan,Q. *et al*. (2007) Concise review: epigenetic mechanism contribute to pluripotency and cell lineage determination of embryonic stem cells. *Stem Cell*, **25**, 2–9.

Guenther,M.G. *et al.* (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.

Impey,S. *et al*. (2004) Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell*, **119**, 1041–1054.

Johnson,D.S. *et al*. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kim,T.H. and Ren,B. (2006) Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.*, **7**, 81–102.

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 (ISMB2005)*, **21** (Suppl. 1), i274–i282.

Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
Martin,C. and Zhang,Y. (2005) The diverse functions of histone lysine methylation. *Nat. Rev. Mol. Cell Biol.*, **6**, 838–849.
McGarvey,K.M. *et al*. (2006) Silenced tumor suppressor genes reactivated by DNA demthylation do not return to a fully euchromatic chromatin state. *Cancer Res.*, **66**, 3541–3549.
Mikkelsen,T.S. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
Pruitt,K.D. *et al*. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.
Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
Raiffa,H. and Schaifer,R. (2000) *Applied Statistical Decision Theory*. Wiley, New York.
Robertson,G. *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
Wei,C.L. *et al*. (2006) A global mapping of p53 transcription factor binding sites in the human genome. *Cell*, **124**, 207–219.
Widschwendter,M. *et al*. (2007) Epigenetic stem cell signature in cancer. *Nat. Genet.*, **39**, 157–158.
Zhao,X.D. *et al*. (2007) Whole-genome mapping of histone h3 lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell*, **1**, 286–298.