BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 9 2010, pages 1199–1204
doi:10.1093/bioinformatics/btq128

Gene expression Advance Access publication April 5, 2010

A signal–noise model for significance analysis of ChIP-seq with negative control

Han Xu[1,2], Lusy Handoko[3], Xueliang Wei[4], Chaopeng Ye[3], Jianpeng Sheng[5], Chia-Lin Wei[3], Feng Lin[2,*] and Wing-Kin Sung[1,4,*]

[1]Computational & Mathematical Biology Group, Genome Institute of Singapore, 138672, [2]School of Computer Engineering, Nanyang Technological University, 637553, [3]Genome Technology & Biology Group, Genome Institute of Singapore, 138672, [4]School of Computing, National University of Singapore, 117543 and [5]School of Biological Science, Nanyang Technological University, 637551, Singapore

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: ChIP-seq is becoming the main approach to the genome-wide study of protein–DNA interactions and histone modifications. Existing informatics tools perform well to extract strong ChIP-enriched sites. However, two questions remain to be answered: (i) to which extent is a ChIP-seq experiment able to reveal the weak ChIP-enriched sites? (ii) are the weak sites biologically meaningful? To answer these questions, it is necessary to identify the weak ChIP signals from background noise.

Results: We propose a linear signal–noise model, in which a noise rate was introduced to represent the fraction of noise in a ChIP library. We developed an iterative algorithm to estimate the noise rate using a control library, and derived a library-swapping strategy for the false discovery rate estimation. These approaches were integrated in a general-purpose framework, named CCAT (Control-based ChIP-seq Analysis Tool), for the significance analysis of ChIP-seq. Applications to H3K4me3 and H3K36me3 datasets showed that CCAT predicted significantly more ChIP-enriched sites that the previous methods did. With the high sensitivity of CCAT prediction, we revealed distinct chromatin features associated to the strong and weak H3K4me3 sites.

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

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

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on October 7, 2009; revised on March 17, 2010; accepted on March 18, 2010

1 INTRODUCTION

With the advances of ultra-high-throughput sequencing technologies in the past 3 years, ChIP-seq is becoming the main approach to the genome-wide study of protein–DNA interactions and histone modifications (Barski *et al.*, 2007; Johnson *et al.*, 2007). In ChIP-seq analysis, the biologically interesting sites are identified by searching for the genomic loci where the reads sequenced from ChIP DNA are over-represented. A number of ChIP-seq analysis tools have been developed for this purpose (Ji *et al.*, 2008; Jothi *et al.*, 2008; Robertson *et al.*, 2007; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Zang *et al.*, 2009; Zhang *et al.*, 2008). All these tools perform well to extract strong ChIP-enriched sites. However, two questions remain to be answered: (i) to which extent is a ChIP-seq experiment able to reveal the weak ChIP-enriched sites? (ii) are the weak sites biologically meaningful? To answer these questions, it is necessary to establish a signal–noise model to separate the weak ChIP signals from background noise.

In a ChIP-seq experiment, majority of the unbound DNA fragments are washed out in the immune-precipitation procedure. The ChIP-processed library is enriched by the fragments pulled down from the genomic loci with high chance of protein–DNA interactions or histone modifications. However, considerable 'non-useful' fragments remain in the library due to the random protein–DNA or antibody–DNA contacts that are not position-specific. Reads sequenced from these fragments are widely spread in the genome, and are considered background noise in addition to the real signal of ChIP enrichment. In early ChIP-seq application without a negative control, distribution of the noise was assumed to be uniform (Robertson *et al.*, 2007). However, recent studies showed that the uniform model is too ideal due to the existence of sequencing and mapping biases, chromatin structure and genome copy number variations (Vega *et al.*, 2009; Zhang *et al.*, 2008). Adjustment of these intrinsic biases requires a negative control, which could be generated using non-specific antibody or input DNA.

To utilize a negative control in ChIP-seq analysis, Ji *et al.* (2008) suggested normalizing the control to a level equivalent to the background noise of the ChIP library. They incorporated the normalization factor into the Bernoulli probability of a binomial distribution for statistical test, and computed false discovery rate (FDR) by *P*-value correction. We extended their idea to a linear signal–noise model, in which a noise rate was introduced to represent the fraction of noise in the ChIP library. We developed an iterative algorithm to estimate the noise rate using a control library, and derived a library-swapping strategy for the FDR estimation. With spike-in simulation datasets, we showed the proposed iterative algorithm can well estimate the noise rate with relative error $<5\%$ under practical sequencing depth. Moreover, spike-in simulation also indicated that the library-swapping approach outperformed the *P*-value correction method in FDR estimation.

We integrated our approaches in a general-purpose framework, named CCAT (Control-based ChIP-seq Analysis Tool), for the significance analysis of ChIP-seq. The CCAT framework was

*To whom correspondence should be addressed.

applied to the H3K4me3 and H3K36me3 datasets (Mikkelsen *et al.*, 2007), and identified significantly more ChIP-enriched sites than previous methods did. Quantitative RT–PCR and comparison to gene annotation validated the reliability of CCAT predictions. With the high sensitivity of CCAT prediction, we revealed distinct chromatin features associated to the strong and weak H3K4me3 sites. In summary, our results imply that the weak sites detectable in a ChIP-seq experiment were under-estimated in previous studies, and the identification of these weak sites may lead to novel discoveries of biological interest.

2 METHOD

2.1 The signal–noise model of ChIP-seq

Considering a control library with M uniquely aligned reads that were mapped to a genome of length L, the control profile can be represented as a vector of independent observations $[c_1^+, c_1^-, \ldots, c_L^+, c_L^-]^T$, where c_x^θ refers to the number of control reads mapped to the genomic location x in direction θ ($x \in [1, L]; \theta \in \{+, -\}$). Due to the random sampling process in ChIP-seq experiment, c_x^θ approximately follow a Poisson distribution (Robertson *et al.*, 2007; Zhang *et al.*, 2008):

$$c_x^\theta \sim \text{Poisson}(u_x^\theta M)$$

where $u_x^\theta$is a normalized factor such that $\sum_{x,\theta} u_x^\theta = 1$.

Next, we consider a ChIP library with N uniquely aligned reads. Similarly, the number of ChIP reads mapped to the genomic location x in direction θ, denoted d_x^θ, follows a Poisson distribution:

$$d_x^\theta \sim \text{Poisson}(v_x^\theta N)$$

We modeled v_x^θ to be the linear combination of the signal s_x^θ and the noise n_x^θ, such that:

$$v_x^\theta = (1-\alpha)s_x^\theta + \alpha n_x^\theta$$

where $\sum_{x,\theta} s_x^\theta = 1$ and $\sum_{x,\theta} n_x^\theta = 1$.

To employ the control library as independent observations of background noise, we made an assumption that the normalized control vector $[u_1^+, u_1^-, \ldots, u_L^+, u_L^-]^T$and the noise vector $[n_1^+, n_1^-, \ldots, n_L^+, n_L^-]^T$ follow the same multivariate hyper distribution π, which models the intrinsic read bias of ChIP-seq. The variation of π reflects slight differences of read bias between ChIP and control libraries, which could be introduced to the experiment prior to random sampling. Note that π is conditional on the chromatin structure, mapping bias and copy number variations. Therefore, the ChIP and control libraries need to be generated on the same cell type, with the same chromatin preparation procedure and identical configurations of alignment software. A line of evidence supporting the above assumption was given by Rozowsky *et al.* (2009). They generated scatter-plot of ChIP and control read counts on a list of genomic regions with no detectable ChIP enrichment. As the result, a linear and nearly symmetric scatter pattern was observed.

For the convenience of description, we denote E to represent the set of genomic locations with real ChIP signal, hence $\overline{E}$ refers to the background with $s_x^+ = 0$ and $s_x^- = 0$. Based on the signal–noise model and the assumption of intrinsic read bias, we have the following proposition:

PROPOSITION 1. Given a genomic region $r = [a, b]$, and $r \subset \overline{E}$, the observation vectors $\mathbf{c}_r = [c_a^+, c_a^-, \ldots, c_b^+, c_b^-]^T$ and $\mathbf{d}_r = [d_a^+, d_a^-, \ldots, d_b^+, d_b^-]^T$ follow the same multivariate distribution under the hyperprior π, if $M = \alpha N$.

Proof of Proposition 1 is given in Supplementary Material of this article. This proposition indicates that the read counts from the control library provide unbiased measurements of the noise level in the ChIP library when the sample sizes satisfy the condition $M = \alpha N$. In this context, the parameter α is called noise rate, which ranges from 0 to 1 and refers to the fraction of noise in the ChIP library. Theoretically, the noise rate is associated with the normalization factor r_0 in Ji *et al.*'s approach (2008), such that $\alpha = r_0 M/N$. Generally speaking, a smaller noise rate implies better ChIP quality in the experiment. In the worst situations where the noise rate approaches 1.0, all the ChIP reads are noise hence no ChIP-enriched site is detectable. Therefore, estimation of the noise rate provides an assessment of the data quality (see Supplementary Material for more details), and further facilitates the subsequent significance analysis.

2.2 Estimating noise rate

If a set of background regions R is known in prior and there is sufficient reads in R, the noise rate can be approximated as the ratio of ChIP read counts to control read counts in these regions, normalized against the sequencing depths:

$$\alpha \approx \frac{\sum_{r \subset R} \sum_{x \in r, \theta} d_x^\theta}{\sum_{r \subset R} \sum_{x \in r, \theta} c_x^\theta} \times \frac{M}{N} \quad (1)$$

Unfortunately, a predefined background region set is unavailable for most ChIP-seq applications. Ji *et al.* (2008) selected the background to be the regions with small read counts. However, due to the intrinsic bias of ChIP-seq, the read counts may also be small for some ChIP-enriched regions and may be relatively large for some background regions. Therefore, a better solution is to determine the background by comparing the ChIP and control libraries, rather than by using the absolute read counts. In our approach, we first partitioned the whole genome into non-overlapping 1 kb bins. The background regions were then selected to be the bins with ChIP read counts less than the expected noise read counts estimated from the control. That is:

$$R^* = \left\{ r : \sum_{x \in r, \theta} d_x^\theta < \alpha N \left(\sum_{x \in r, \theta} c_x^\theta \right) / M \right\} \quad (2)$$

Based on Equations (1 and 2), we propose an iterative algorithm for estimating noise rate. In our algorithm, the ChIP and control data were divided into two subsets: D^+ consisting of sense reads, and D^- consisting of antisense reads. All the reads were shifted by $l/2$ bp towards their orientations, where l is the average DNA fragment length and could be approximated either experimentally or computationally (Robertson *et al.*, 2007; Zhang *et al.*, 2008). In each step of iteration, a set of background regions were selected using D^+, followed by updating noise rate using D^-. The iterative algorithm is described as follows.

Initialization: divide the datasets into D^+ and D^-; partition the whole genome into non-overlapping 1 kb bins; set $\alpha_0 = 1.0$.

The i-th iteration: (i) count the reads from D^+ for the bins, and select a set of background bins R_i^* based on Equation (2), where $\alpha_0 = \alpha_{i-1}$; (ii) count the reads in D^- for the bins in R_i^* and compute α_i based on Equation (1).

Termination: $\alpha_i > \alpha_{i-1}$ or i larger than a threshold.

In our test with real datasets, the above algorithm converged very fast within 10 iterations (Supplementary Fig. S1).

2.3 FDR estimation by library swapping

Several tools have been proposed in the literature for ChIP-seq analysis with negative control (Ji *et al.*, 2008; Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Zhang *et al.*, 2008). In general, these tools perform three steps: (i) select candidate sites; (ii) rank sites based on certain significance measurement; and (iii) determine the cutoff threshold. Most methods predict ChIP-enriched sites based on the read counts in a local region. For these cases, steps (i) and (ii) can be represented as a function f that maps the read count vectors of a local region to a significance measure. Preferentially, the threshold in step (iii) is determined with FDR control. In this article, we take the form of pFDR (also called Bayesian FDR) proposed by Storey (2003). Given a significance function f, and a list of non-overlapping regions determined based on a threshold t, the pFDR is represented as:

$$\text{pFDR}(t) = \Pr(r \subset \overline{E} | f(\mathbf{c}_r, \mathbf{d}_r) > t) \quad (3)$$

In Equation (3), r represents a genomic region, and c_r and d_r are the corresponding read count vectors, as defined in Proposition 1.

Based on Proposition 1 and Equation 3, we have:

PROPOSITION 2. Under the conditions:

(a) $\Pr(f(\mathbf{d}_r,\mathbf{c}_r)>t|r\not\subset\overline{E}) << \Pr(f(\mathbf{c}_r,\mathbf{d}_r)>t|r\not\subset\overline{E})$, and

(b) $M=\alpha N$;

The pFDR for a threshold t can be approximated by:

$$\text{pFDR}(t)=\frac{\#\{f(\mathbf{d}_r,\mathbf{c}_r)>t\}}{\#\{f(\mathbf{c}_r,\mathbf{d}_r)>t\}}$$

where the operator $\#\{\Delta\}$ is defined to be the total number of regions satisfying the condition Δ.

For the proof of Proposition 2, readers may refer to Supplementary Material of this article. Proposition 2 underlies a library-swapping strategy of FDR estimation, in which the null distribution of significance measurement is empirically estimated by exchanging the ChIP and control libraries, i.e. control versus ChIP. The library-swapping strategy was first introduced by Zhang *et al.* (2008) in the software MACS (Model based Analysis for ChIP-Seq). However, as mentioned by Zhang *et al.*, the estimated FDR would be biased if the read sample size from ChIP and control are not balanced. In Proposition 2, we give two necessary conditions for the correctness of library-swapping strategy.

The first condition indicates that the significance function f need to be defined such that the distributions of $f(\mathbf{c}_r,\mathbf{d}_r)$ and $f(\mathbf{d}_r,\mathbf{c}_r)$ are well-separated for the ChIP-enriched regions. For example, if we define $f(\mathbf{c}_r,\mathbf{d}_r)=\|\mathbf{c}_r\|+\|\mathbf{d}_r\|$ ($\|.\|$ is the magnitude operator), the library-swapping strategy is not applicable since the distributions of $f(\mathbf{c}_r,\mathbf{d}_r)$ and $f(\mathbf{d}_r,\mathbf{c}_r)$are identical. On the other hand, if we define $f(\mathbf{c}_r,\mathbf{d}_r)=\|\mathbf{d}_r\|-\|\mathbf{c}_r\|$, library-swapping is applicable for $t>0$ since we expect $f(\mathbf{c}_r,\mathbf{d}_r)>0>f(\mathbf{d}_r,\mathbf{c}_r)$for most ChIP-enriched regions.

The second condition, $M=\alpha N$, applies a constraint of sample size for ChIP and control, which explains the estimation bias with unbalanced sample size. When $M\neq\alpha N$, the constraint condition can be simply satisfied after a re-sampling procedure by which a subset of reads were randomly retrieved for processing. In detail, if $\alpha N>M$, we retrieve M/α ChIP reads and M control reads; otherwise, N ChIP reads and αN control reads are retrieved.

2.4 The CCAT framework

In this section, we propose a general-purpose framework, named CCAT (Control based ChIP-seq Analysis Tool), for the significance analysis of ChIP-seq with negative control. The CCAT framework started with the estimation of noise rate α using the iterative algorithm, followed by a random re-sampling procedure to retrieve a subset of data that satisfies the condition $M=\alpha N$, as described in Section 2.3. Next, we exchanged the ChIP and control libraries of the re-sampled data to generate a swapped dataset. Both the re-sampled dataset and the swapped dataset were processed by a region identification module and a region scoring module (see below). A ranking list of candidate ChIP-enriched sites and a list of control sites were generated from the re-sampled dataset and the swapped dataset, respectively. Following Proposition 2, FDRs were estimated for the candidate sites using the control sites as reference. The flowchart of CCAT is shown in Figure 1.

In our implementation of region identification module, the ChIP-seq reads were first shifted by $l/2$ bp toward their orientations, where l is the average DNA fragment length. A sliding window was then applied to scan the whole genome. The size of the window was set to be $2l$ for applications to motif-specific transcription factors, and ranged from hundreds to thousands base pairs for histone modification study. The shifted ChIP reads and control reads were counted within the window in each step of sliding. We flagged the genomic position at the center of the window if the ChIP read count was more than twice the control read count, corresponding to a minimum expected signal–noise ratio of 1.0. Consecutive flagged positions were merged into non-overlapping candidate regions. Although the ChIP-enriched sites were defined in term of regions in CCAT, we also applied a peak-finding algorithm (Chen *et al.*, 2008) to each candidate region for a high-resolution representation of ChIP enrichment site, which facilitates the motif study in some applications. The region scoring module in CCAT supports multiple options of significance scores, including fold-change, binomial P-value (Ji *et al.*, 2008; Rozowsky *et al.*, 2009), Poisson P-value (Zhang *et al.*, 2008), and normalized difference score (Nix *et al.*, 2008). Details of these scores are given in the Supplementary Material.

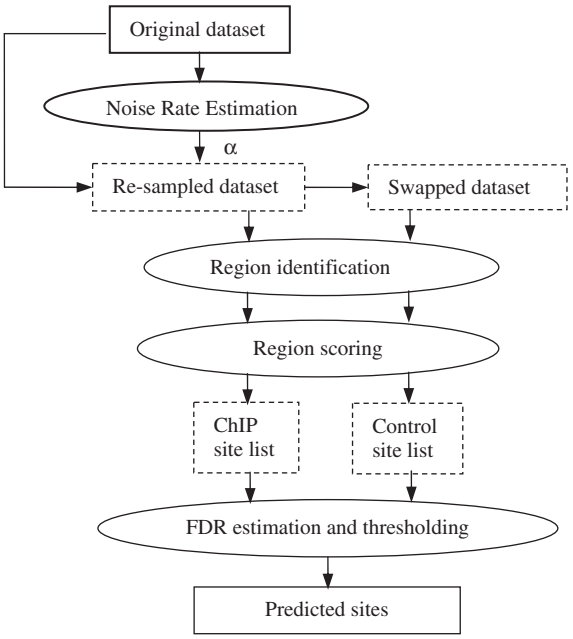

Fig. 1. The flowchart of CCAT. Ovals refer to computational steps; rectangles with solid lines refer to the input and output data; rectangles with dash lines refer to the intermediate data.

3 RESULTS

3.1 Spike-in simulation

To define a standard that is comprehensive enough for the evaluation of computational approaches, we generated spike-in datasets to assess the accuracy of noise rate estimation and FDR estimation. The idea of 'spike-in' here is to computationally add ChIP-enrichment signal to a control background, so that the spike-in loci can be referred as standard in the evaluation (Nix *et al.*, 2008). Two Nanog ChIP-seq datasets published by different groups (Chen *et al.*, 2008; Marson *et al.*, 2008) were included for generating spike-in ChIP library. These datasets can be treated as biological replicates. We defined the spike-in regions as the binding sites predicted from Marson *et al.*'s dataset, and retrieved spike-in reads for those regions from Chen *et al.*'s dataset. By this, 16 688 spike-in regions were defined, corresponding to 464 757 reads. Due to the variation between replicates, we found reads are not over-represented in some of the spike-in regions, which assembles the false negatives in real datasets. The background reads in the spike-in ChIP library were retrieved from a GFP dataset (Chen *et al.*, 2008) prepared with non-specific antibody. The noise rate of the spike-in ChIP library was adjustable by merging different proportion of the GFP reads with the spike-in reads. For the control, we generated a WCE (whole-cell extract) dataset, available at http://cmb.gis.a-star.edu.sg/ChIPSeq/tools.htm. Note that the background reads in

the spike-in ChIP and control libraries were retrieved from different sources in order to better simulate the real ChIP-seq experiment, in which the control was prepared with a different antibody or simply input DNA without antibody. All the libraries used for the spike-in dataset were prepared on mESC (mouse embryonic stem cell), and were mapped by Eland software with identical configurations (26 bp, 2 mismatches). In addition to the Nanog data, we also generated H3K4me3 spike-in data in the same manner. A total of 19 631 spike-in regions were defined based on Marson *et al.*'s H3K4me3 dataset (2008), and 3 171 004 spike-in reads were retrieved from another dataset published by Mikkelsen *et al.* (2007). A summary of the libraries included for the spike-in data generation is provided in Supplementary Table S1.

To assess the accuracy of noise rate estimation, different proportions (10%, 20%, … ,100%) of GFP library were used as the background for spike-in. Therefore, the real noise rate equals to the fraction of GFP reads in the spike-in ChIP library. In Figure 2a, the relative estimation error of the iterative algorithm was plotted against the actual noise rate. We found the errors are positive in most of the cases, implying over-estimation of noise rate. The explanation is that some of the background regions determined by Equation (2) contain weak but undetectable signal due to the issue of sequencing depth. Nevertheless, the relative errors are <5% under the practical sequencing depth, indicating the proposed iterative algorithm reasonably estimated the noise rate. Comparing to the noise rate derived from Ji *et al.*'s normalization factor, our estimation also achieved better accuracy (Supplementary Fig. S2).

Next, we compared the library-swapping based FDR estimation with previous methods. To date, the approaches of FDR estimation for control-based ChIP-seq data mainly fall into two categories: (i) correcting binomial *P*-values using Benjamini–Hochberg (B–H) correction (Benjamini and Hochberg, 1995; Rozowsky *et al.*, 2009) or Storey's method (Nix *et al.*, 2008; Storey, 2002); (ii) estimating expected fraction of false positives by generating empirical background. The library-swapping approach mentioned in Section 2.3 belongs to the latter. Another example of the second category is the eFDR proposed by Nix *et al.* (2008). To calculate eFDR, control library were randomly split in two halves. The eFDR was estimated as the ratio of the number of control-enriched regions (control 1 versus control 2) to the number of experimentally observed enriched regions (ChIP versus control2).

To assess the accuracy of FDR estimated by different approaches, we employed a spike-in datasets configured as follow: 3 million GFP reads were randomly retrieved and were merged with the Nanog or H3K4me3 spike-in reads to generate ChIP library; 3 million WCE reads were used as control (for eFDR, additional 3 million WCE reads were used as an alternative control). By this configuration, the sample sizes of ChIP and control were balanced. The whole genome was partitioned into 1 kb non-overlapping regions and the FDR were calculated on the region basis. For a fair comparison, we used binomial *P*-values as the significance score for all methods. Prior to calculating *P*-value, we followed Nix *et al.*'s (2008) suggestion to exclude the regions with read counts smaller than 10. This filtering step is necessary for *P*-value correction methods, which assume that the *P*-values are uniformly distributed under null hypothesis.

The accuracy of FDR estimation was compared in a dynamic range from 0.001 to 0.2. The results are largely consistent for Nanog (Fig. 2b) and H3K4me3 (Fig. 2c) spike-in datasets. We found that the empirical approaches significantly outperformed the

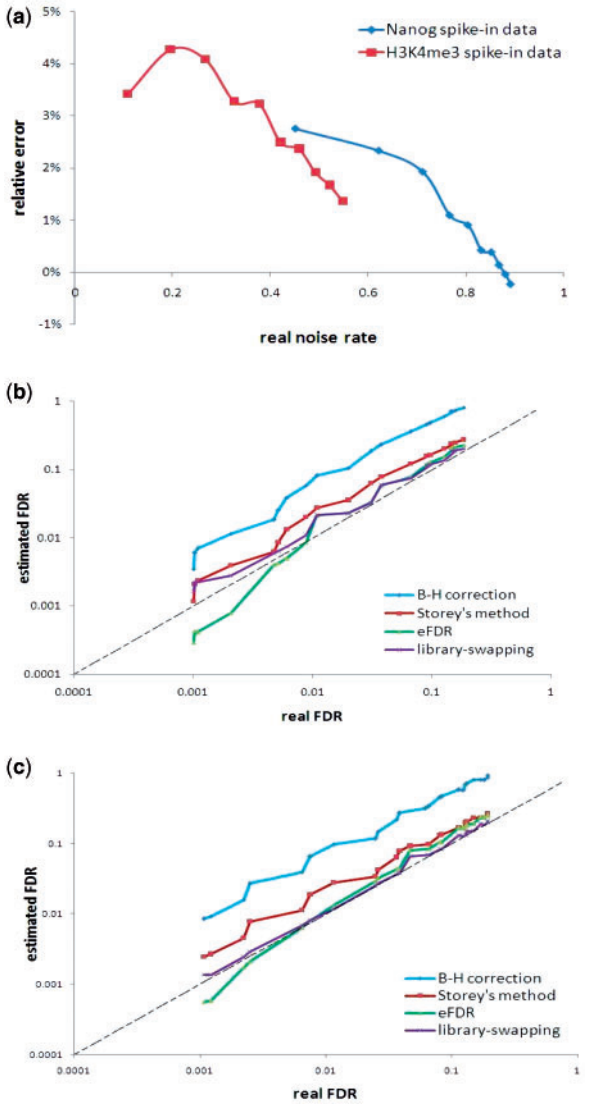

Fig. 2. (**a**) The relative error of noise rate estimation by proposed iterative algorithm; (**b**) and (**c**) comparison of FDR estimation methods using (b) Nanog and (c) H3K4me3 spike-in datasets.

P-value correction methods, for which the FDR was constantly overestimated. Although both empirical approaches performed well when FDR >0.01, eFDR appeared a non-linear characteristic of estimation bias, and under-estimated when FDR <0.01. This may ascribe to slight bias between GFP and WCE libraries prior to random sampling. Such bias could not be modeled using the empirical background that was generated by control splitting. On the contrary, the library-swapping approach seemed to reasonably compensate this type of bias and achieved satisfactory performance both in accuracy and linearity.

3.2 Application to histone modification datasets

We applied CCAT to the mESC H3K4me3 (K4) and H3K36me3 (K36) histone modification datasets (Mikkelsen *et al.*, 2007). The GFP library published by Chen *et al.* (2008) was used as the

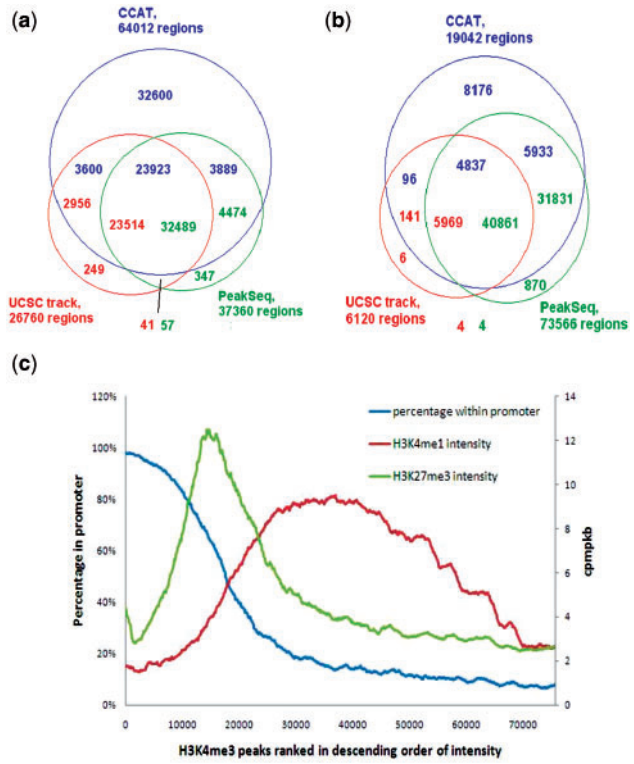

Fig. 3. (**a** and **b**) Venn diagram comparing the predictions by three methods for (a) H3K4me3 and (b) H3K36me3 dataset; (**c**) distinct chromatin features associated with strong and weak H3K4me3 sites. The histone modification intensities are measured with count per million reads per kilobasepair (cpmpkb). The curves were smoothed by a sliding window of length 1000.

negative control. In literature, K4 was known to be a promoter marker and K36 has been shown to occupy the gene region as a hallmark of elongation (Guenther *et al.*, 2007). Our previous study also showed the gene expression is potentially predictable from K4 and K36 marks (Xu *et al.*, 2008). Based on a better measurement of ChIP-seq background level and more reliable FDR estimation, the purpose of our application is to study the weak K4 and K36 sites that may not be predictable by previous methods. In CCAT configuration, we used the normalized difference score (Nix *et al.*, 2008) as the significance measure. The FDR cut-off was set to be 0.05. The size of sliding window was set to be 500 bp for K4 and 2 kb for K36. This setting was determined due to the fact that K4-enriched sites usually appear peak patterns, while K36 signals spread broader regions. As the result, we identified 64 012 K4 regions, corresponding to 75 620 individual peaks (Supplementary Table S2). For K36, 19 042 K36 regions were identified (Supplementary Table S3).

To assess the sensitivity, we overlapped CCAT identified regions with the UCSC histone modification track (http://genome.ucsc.edu/) predicted by Mikkelsen *et al.* (2007), as well as the predictions by a recently published tool PeakSeq (Rozowsky *et al.*, 2009), with FDR = 0.05. As shown in Figure 3a and b, CCAT predicted 100% more K4 sites and 70% more K36 sites than these two methods did. We further extended our comparison to include MACs (Zhang *et al.*, 2008) and SICER (Zang *et al.,* 2009). Again, pair-wise comparison showed that the CCAT prediction is highly sensitive and almost forms a superset of the predicted sites by the other methods, for both K4 and K36 datasets (Supplementary Fig. S3 and Supplementary Material).

Next, we tested the reliability of our predictions, i.e. specificity. For K4, we randomly selected 15 sites from 32 600 regions exclusively identified by CCAT in Figure 3a, and validated them using quantitative real-time PCR (qPCR). Ten negative control sites were selected from the genomic regions outside the predicted K4 regions. The qPCR validation results showed that 14 out of 15 predicted sites are ChIP-enriched with >3-fold-change against the median value of controls (Supplementary Fig. S4). For K36, we employed annotated gene regions as reference since K36 is known to mark transcription elongation. A total of 8176 K36 regions exclusively identified by CCAT in Figure 3b, which are mostly weak in intensity, were compared to three gene annotation databases: RefSeq, Ensembl and MCG. We found 7556 (92.4%) of these regions overlap with the genes annotated by at least one database (Supplementary Fig. S5), implying that the weak K36 signal is biologically meaningful and may reflect low-abundant gene expression.

By traditional arguments, H3K4me3 sites are markers of promoters. Surprisingly, only 26 902 (35.6%) out of 75 620 predicted K4 peaks overlap with known promoters (±1 kb from TSS), and majority of the promoter-associated K4 peaks are in the top 20 k of the ranking list. To understand the chromatin features associated with weaker K4 sites, we ranked the K4 peaks in descending order of intensity (ChIP read counts within 1 kb), and compared the list to other two histone modification types: H3K27me3 and H3K4me1 (Fig. 3c). The ChIP-seq datasets published by Mikkelsen *et al.* (2007) and Meissner *et al.* (2008) were employed for the comparison. We found H3K27me3 signals are enriched for the K4 peaks ranked from 10 to 20 k, corresponding to the K4–K27 bivalent domains that have been proven to be crucial to maintain to pluripotency of ESC (Zhao *et al.*, 2007). Meanwhile, H3K27me3 intensities of the top 10k peaks are relatively low in average. We also observed that H3K4me1, an enhancer marker in mammalian cells (Heintzman *et al.*, 2007), is enriched for the peaks ranked after 20 k, and is almost depleted for the top-ranking peaks. Therefore, in addition to the traditional arguments on H3K4me3, these lines of evidence suggest distinct chromatin features associated with strong and weak K4 sites. As a hypothesis, the weak K4 signals in the enhancer regions may ascribe to the DNA-loop-mediated interactions of enhancers and promoters.

4 DISCUSSION

The goal of the research work presented in this article is to investigate the extent to which the weak ChIP signals are detectable in a ChIP-seq experiment. For this purpose, we proposed a linear signal–noise model for ChIP-seq analysis with negative control. The noise rate defined in the model provides a measurement of the data quality, which determines the capability of ChIP-seq to detect weak ChIP-enriched sites. We developed an iterative algorithm to estimate the noise rate using control library, and derived a library-swapping approach for FDR estimation. The performance of the proposed approaches was demonstrated with spike-in datasets. These approaches, integrated in the CCAT framework, were further applied to H3K4me3 and H3K36me3 datasets, and identified significantly more weak sites than previous predictions. Two reasons count for the improvement of sensitivity: first, the estimation of noise

rate led to a better measurement of background noise level, which was overestimated by commonly used approach of 'normalization against sequencing depth'; secondly, as shown in the spike-in simulation, the library-swapping approach improved the accuracy of FDR estimation, for which the previous *P*-value correction methods are too conservative.

Our results showed that the weak ChIP-seq signals may correspond to different genomic features from that of strong signals. Therefore, studying the weak signals could extend the scope of biological discoveries made from ChIP-seq data. Besides our example on histone modifications, another interesting topic is to study the weak transcription factor binding sites (TFBS), which possibly implicate the indirect binding mediated by TF–TF interactions. We expect the CCAT framework would facilitate such study in the future.

A major drawback of CCAT is that only a portion of the original dataset was utilized due to the re-sampling procedure. One solution to this problem is to use a bootstrapping strategy in the region scoring module. The option of bootstrapping is provided in the software package of CCAT. In detail, the original dataset is re-sampled using different random seeds, and the scores of candidate regions are re-computed for a number of bootstrapping passes. The final score for a region is the median of the scores computed in all the bootstrapping passes. By this, the sensitivity and specificity of the ranking list can be improved. A point of note is that the FDR could be slightly over-estimated when the bootstrapping strategy is used.

ACKNOWLEDGEMENTS

The authors thank Prof. Liu X.L. for critical discussion on the methods, and Ms Veeravalli L. for the assistance on sequence alignment.

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in human genome. *Cell*, **129**, 823–937.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.*, **57**, 289–300.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

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

Heintzman,N.D. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

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

Jothi,R. *et al.* (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.

Marson,A. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell,* **134**, 521–533.

Meissner,A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.

Mikkelsen,T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Nix,D.A. *et al.* (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, **9**, 523.

Robertson,G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Rozowsky,J. *et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

Storey,J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.

Vega,V.B. *et al.* (2009) Inherent signals in sequencing-based chromatin-immunoprecipitation control libraries. *PLoS One*, **4**, e5241.

Xu,H. *et al.* (2008) Genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.

Zang,C. *et al.* (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, **9**, R137.