Peak identification for ChIP-seq data with no controls

Yanfeng ZHANG, Bing SU*

State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, the Chinese Academy of Sciences, Kunming 650223, *China*

Abstract: Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is increasingly being used for genome-wide profiling of transcriptional regulation, as this technique enables dissection of the gene regulatory networks. With input as control, a variety of statistical methods have been proposed for identifying the enriched regions in the genome, i.e., the transcriptional factor binding sites and chromatin modifications. However, when there are no controls, whether peak calling is still reliable awaits systematic evaluations. To address this question, we used a Bayesian framework approach to show the effectiveness of peak calling without controls (PCWC). Using several different types of ChIP-seq data, we demonstrated the relatively high accuracy of PCWC with less than a 5% false discovery rate (FDR). Compared with previously published methods, e.g., the model-based analysis of ChIP-seq (MACS), PCWC is reliable with lower FDR. Furthermore, to interpret the biological significance of the called peaks, in combination with microarray gene expression data, gene ontology annotation and subsequent motif discovery, our results indicate PCWC possesses a high efficiency. Additionally, using in silico data, only a small number of peaks were identified, suggesting the significantly low FDR for PCWC.

Keywords: ChIP-seq; Bayesian; Peak calling; Gene regulation

With the advance of high-throughput sequencing technologies, both global survey of the genome structure and transcriptome and transcriptional regulation has become more accurate and sensitive. As one of the more powerful and widely used experimental techniques for DNA-protein interactions *in vivo*, chromatin immunoprecipitation followed by sequencing (ChIP-seq) is one of the early applications of next generation sequencing (NGS). Since the first study in 2007 (Mikkelsen et al), ChIP-seq technology has been increasingly used for mapping DNA-binding sites by transcriptional factors and chromatin modifications. Compared with the earlier method of chromatin immunoprecipitation followed by tiling microarray (ChIP-chip) (Kapranov et al, 2002), ChIP-seq has more advantages, e.g., higher resolution, cost-effectiveness and technical simplification. At the same time, ChIP-seq itself also has several disadvantages, including efficiently computational analysis techniques, sequencing depth requirement, and the like (Park, 2009).

Currently, there are three commonly used types of control samples in ChIP experiments: input DNA (no immunoprecipitation (IP) DNA samples); mock IP DNA (DNA obtained from IP without antibodies); and DNA from non-specific IP (such as IP with immunoglobulin G). Concomitantly, several tools for ChIP-seq analysis have been developed, the majority of which require a matched control sample to determine fold enrichment and significance of peak signals. This idea of fold ratio relative to controls used for ChIP-seq is similar with ChIP-chip data analysis, which in the currently used methods is necessary for peak calling (Rozowsky et al, 2009; Valouev et al, 2008). Meanwhile, several new methods based on Validation Discriminant Analysis (Micsinai et al, 2012), Hidden Markov Models (Choi et al, 2009) or Bayesian Models (Spyrou et al, 2009) combine ChIP-seq and ChIP-chip data for peak identification. Nonetheless, a systematic analysis of ChIP-chip and ChIP-seq datasets revealed that the input data has variable effects on peak finding (Ho et al, 2011), suggesting of the need for a high quality input sample for peak calling. Additionally, sequencing depth significantly

Received: 06 September 2012; Accepted: 31 October 2012
Foundation items: This study was supported by the National 973 project of China (2011CBA01101) and the National Natural Science Foundation of China (30871343 and 31130051)
* Corresponding author, E-mail: sub@mail.kiz.ac.cn

impacts the peak calling, further causing bias of peak identification (Chen et al, 2012). To improve peak calling accuracy, Osmanbeyoglu et al (2012) utilized a strategy of co-regulation binding by integrating multiple sources of biological information. Recently, although several novel tools for peak calling without controls have been made available (Cairns et al, 2011; Fejes et al, 2008; Hower et al, 2011), for example, the model-based analysis of ChIP-seq (MACS) defines a dynamic parameter to capture local biases when a control profile is unavailable (Zhang et al, 2008). No systematic evaluation of peak calling without controls has been carried out, even for MACS or BayesPeak.

Given the bias of control data in the ChIP-seq peak calling, we sought to address whether ChIP-seq peak calling without controls (PCWC) is plausible. Employing the Bayesian theorem and simulation-based empirical estimates of background distribution, we demonstrated that our method ChIP-seq peak identification with no control is reliable, with an FDR lower than 5%. Compared with the Poisson-based MACS method, our Bayesian framework strategies showed lower FDRs, suggesting high selectivity and effectiveness for peak calling. The systematic analysis of PCWC we present in the current study could serve as an alternative strategy for ChIP-seq analysis and subsequent biological interpretation of gene regulation.

MATERIALS AND METHODS

Data sets

In order to present the comprehensive performance of PCWC, we used the FASTQ-formatted ChIP-seq data sets downloaded from Gene Expression Omnibus (GEO) (Edgar et al, 2002), including the ChIP-seq data of transcription factor E2F1 in mESCs (Chen et al, 2008), VDR binding in human lymphoblastoid cell (Ramagopalan et al, 2010) and H3K4me3 in mESCs (Creyghton et al, 2010), as well as the ABI SOLID ChIP-seq data sets of EGR1 ChIP-seq data (Tang et al, 2010) and MNase-seq data in two replicates (Valouev et al, 2011). Furthermore, we integrated the microarray gene expression dataset to evaluate the effectiveness of PCWC. Additionally, we *in silico* generated 1×10^7 36-bp reads using the simreads program of the Rmap package (Smith et al, 2009) to simulate control samples for peak calling. The detailed description for ChIP-seq is shown in Table S1.

Peak calling based on the Bayesian framework

On the genome-wide scale, peaks in the ChIP-seq experiment are those regions significantly enriched by reads. Thus, regarding all uniquely mapped reads of the ChIP-seq as background, the density of reads in peak regions should be dominantly enriched. In the Bayesian framework (Madigan & Ridgeway, 2003), the posterior density for θ is obtained, up to a proportionality constant by multiplying the prior density g(θ) by the likelihood $L(\theta)$ where θ denotes the probability that a region is a true peak:

$$\Pr(\theta \mid data) \propto L(\theta)g(\theta) \quad (1)$$

Based on global deep sequencing that reaches to single nucleotide resolution, a uniform prior distribution is usually supposed, which has been adopted in the RNA-seq analysis (Sultan et al, 2008; Wang et al, 2008). Here, for simplification and feasibility, we assume a uniform distribution for the prior distribution g(θ) ($g(\theta) \sim U$), so $g(\theta) = 1$, for θ , leading to the equation:

$$\Pr(\theta \mid data) = \int L(\theta)d\theta \quad (2)$$

For measuring the significant enrichment of reads in a peak, we considered that the candidate peaks with the number of reads (#reads, a) below a cutoff threshold (c, usually 0.01) would fit the following equation,

$$\begin{aligned}\Pr(\#reads \geq a) &= 1 - \Pr(\#reads < a) \\ &= 1 - \int_0^{1-c} L(\theta)d\theta \ \leq \ \mathrm{c}\end{aligned} \quad (3)$$

where the cumulative density of #reads reaching to a is based on likelihood function $L(\theta)$ and $L(\theta)$ is empirically estimated from the genome-wide scanning of tag-count (10,000 w bp random bins per chromosome).

Following the optimal parameter a below c, we can obtain the read-count (uniquely mapped total reads) with a sliding window of size w/2 bp (w=50 bp in default). Afterward, the #reads in each w/2 bp sliding window size above a is regarded as candidate peaks. Peaks spanning lower than 100 bp are discarded and the neighboring peaks within 1 kb are merged (Guenther et al, 2008) to counteract the shifting effects of aligned tags from forward and reverse reads.

False discovery rate (FDR) estimate for the number of peaks

Similar to methods used by Valouev et al (2008), the overlapped number of called peaks (using the same threshold) between the input dataset and ChIP dataset are the false discovery number, and the FDR is the false discovery number divided by the total number of peaks called in the ChIP experiment.

Motif analysis

We used MEME 4.6.1 to discover motifs (Bailey & Elkan, 1994). Since the running time can be prohibitively long for large sets, the peaks are ranked by the number of uniquely aligned reads and only the top 10% of the peaks were selected for motif discovery. A similar methodological strategy was used by Ramagopalan et al (2010) and Jothi et al (2008). The location of the peak is centered and extended by 100 bp on either side. Motifs

between 5 and 30 bp in length were determined on both strands.

Microarray data preprocessing and analysis

A total of 45 microarray data sets for calcitriol stimulation on human lymphoblastoid cell lines were used for expression analysis. The multi-array average (RMA) expression values were calculated using the Bioconductor "affy" package (Gentleman et al, 2004). The log2 expression signals were used to calculate fold change.

Gene annotation

Genes regulated by trans-acting factors (or by chromatin modification, such as H3K4me3) were defined by no more than 5 kb distance of peaks to Refseq transcription start site (TSS). The enriched analysis of genes was achieved using the DAVID annotation system (Huang da et al, 2009). After Benjamini-Hochberg correction, P<0.01 were considered significantly enriched.

Peak calling based on the Bayesian model and other bioinformatics analyses were implemented in Perl and R (data available upon request).

RESULTS AND DISCUSSION

To demonstrate the effectiveness of ChIP-seq PCWC, we selected five representative ChIP-seq data sets from two platforms (three from Illumina and two from ABI SOLID platforms, respectively) and one *in silico* data (See Materials and Methods for details).

We first applied the Bayesian-based approach on recently published ChIP-seq data for transcription factor E2F1 in mouse embryonic stem cells (mESCs) (Chen et al, 2008), vitamin D receptor (VDR) binding in human lymphoblastoid cell (Ramagopalan et al, 2010) and H3 trimethylated at lysine 4 (H3K4me3) in mESCs (Creyghton et al, 2010), all of which were generated from the Illumina platform. Employing SOAP2 program (Li et al, 2009) with maximal 1 mismatch, the uniquely mapped reads are considered for further analysis. Using a cutoff threshold (c =0.01 or 0.001), the number of reads a based on 10 000 random bins per chromosome with window size (w=50) was estimated (Figure 1). With these two parameters, we can empirically determine the candidate peaks for each data set.

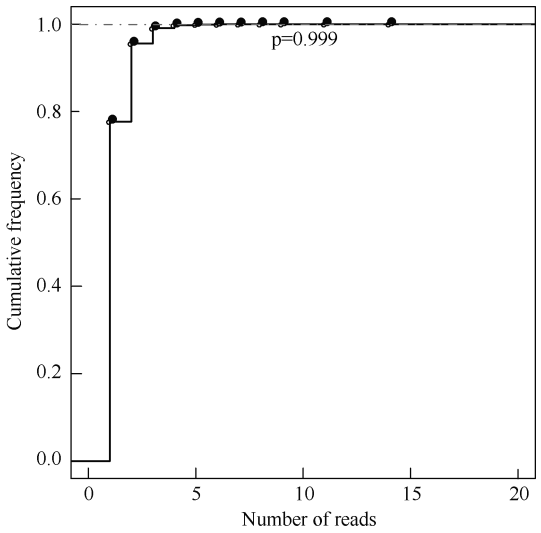

Figure 1 Cumulative density of reads based on genome-wide random bins

Grey dashed line represents the read cutoff for peak calling.

Vitamin D receptor data

For the VDR ChIP-seq samples, it is highly suitable for evaluating the ChIP-seq peak calling without controls due to the inclusion of conditional data (calcitriol treated or not) with deep sequencing (ranging from 7.9×10^6 to 1.46×10^7 uniquely mapped reads, see Table 1) and the available microarray gene expression data set.

Table 1 Summary of the peak calling analysis using the VDR ChIP-seq data

Samples[a]	Description	#Reads	#Uniquely aligned	*Pr* (#reads≥a)≤0.001	# Peaks	Mean length[b]
SRX022390	unstimulated_rep1	18 252 156	13 487 568	7	1 989	633.9
SRX022391	unstimulated_rep2	15 379 663	10 761 914	7	548	708.6
SRX022392	vitaminD_rep1	18 391 888	13 937 615	7	2 920	455.4
SRX022393	vitaminD_rep2	18 965 010	14 613 990	7	2 835	467.3
SRX022394	unstimulated_rep1	12 264 149	10 149 547	7	2 223	519.3
SRX022395	unstimulated_rep2	10 145 026	7 930 475	7	5 298	460.7
SRX022396	vitaminD_rep1	13 302 506	10 657 775	9	6 755	428.9
SRX022397	vitaminD_rep2	14 526 186	11 755 087	9	6 981	434.9
SRX022398	Input1	15 001 356	11 403 930	6	348	–
SRX022399	Input2	13 682 506	11 402 869	6	545	–

[a]: Sample ID is the accession ID for meta-data documented in GEO database; [b]: Mean peak length reflects the resolution of binding sites.

Firstly, we sought to call peaks in both conditions. In the samples not treated with calcitriol, the number of peaks without control is between 548 to 5298, while in the calcitriol-treated samples the number of peaks is between 2 835 to 6 981 (Table 1), consistent with the Ramagopalan's reports (Ramagopalan et al, 2010), where they used MACS with two independent controls for peak calling. Interestingly, our Bayesian-based PCWC get a similar number of peaks compared with the data using MACS with controls.

Next, we conducted analyses on multiple scales. The resolution of peaks (VDR binding sites) is also comparable with the reported study (Table 1). Furthermore, the gene ontology (GO) annotation indicates that immune system development (GO:0002520), lymphocyte activation (GO:0046649) and T cell activation (GO:0042110) are significantly enriched ($P \leq 0.01$, after Benjamini correction), also highly concordant with the report (Ramagopalan et al, 2010). Additionally, previous studies have experimentally validated that *VDR* (auto-regulation) (Zella et al, 2010), *CCNC* (Sinkkonen et al, 2005), *ALOX5* (Seuter et al, 2007), *IRF8* and *PTPN2* (Ramagopalan et al, 2010) modulated by VDR have significant enrichment of peaks (VDR binding) under calcitriol-treated conditions, which were all confirmed using our method (Figure S1), suggesting the effectiveness of our method for ChIP-seq peak calling.

We then integrated the microarray gene expression data (a total of 45 microarrays) to evaluate the accuracy of PCWC. Compared with the untreated condition, a total of 205 genes (fold change≥1.5) were up-regulated with calcitriol treatment, of which 134 (65.4%) genes were significantly enriched with the called peaks. Of the 134 genes, 81 (60.4%) genes enriched with peaks are located in the intronic regions, 33 genes in the TSS regions (TSS±500) bp, consistent with the report that under the condition of the calcitriol stimulation, there is an increased VDR binding in the intronic regions (Ramagopalan et al, 2010), further suggesting the high accuracy of our Bayesian model for PCWC.

For ChIP-seq data analysis, in addition to the integration of gene expression data, the subsequent motif discovery analysis for the enriched regions is also informative to interpret the biological implications (Park, 2009). Thus, we carried out a motif analysis of VDR binding (see Methods). Due to the prohibitively long running time for large data sets, only the top 10% of the peaks were used for motif discovery. Figure 2 shows the motifs discovered by our method, which are nearly identical with the reported data (Ramagopalan et al, 2010). Collectively, the data presented using the VDR dataset indicates that the PCWC is reliable.

Figure 2 Inferred consensus binding motif for VDR

For PCWC, evaluate whether the called peaks are real signals relative to the flanking regions is crucial. We therefore conducted such analysis based on the ratio of peaks to the flanking regions. For each peak, we selected 300 bp region at both the 5' and 3' flanking sequences. As shown in Figure 3, compared with the 5' and 3' flanking regions, the peak regions have much higher (at least 2 times higher) read coverage, indicating the called peaks without control are relatively reliable. Furthermore, the distribution of the uniquely mapped bases (in 1 kb bins) in the genome for the ChIP and input data also indicates that the peak regions in the ChIP data are

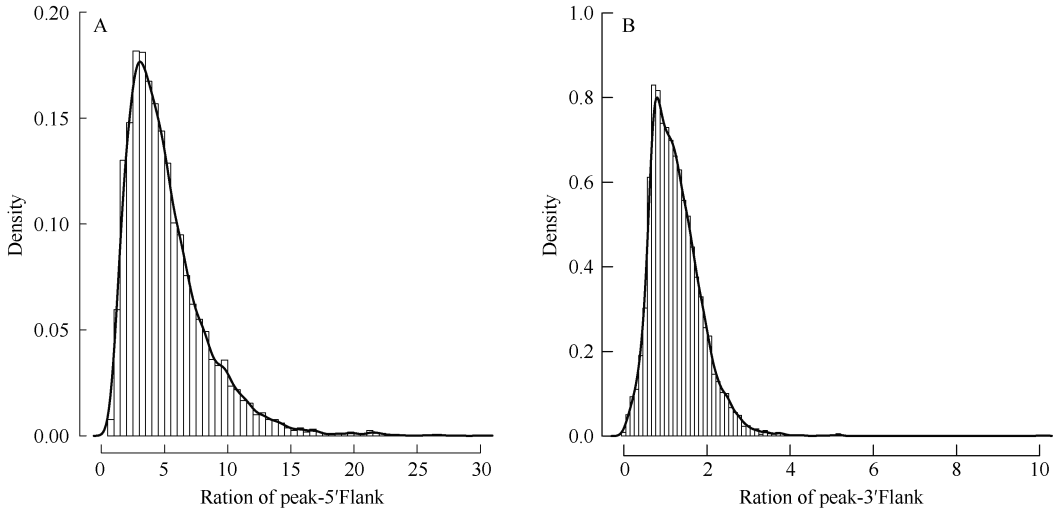

Figure 3 Density of read coverage ratio between peak and flank (5' and 3') regions

enriched in reads, while in the input data, they are relatively uniformly distributed (Figure S2).

We also evaluated the FDR of the Bayesian model for PCWC. Similar with the VDR ChIP-seq data analysis, peak calling ($Pr(\#reads \geq 6) \leq 0.001$) for two independent controls (from the VDR data sets, Table 1) was conducted to evaluate the FDR. We totally identified 348 and 545 peaks in two independent controls (Table 1), respectively. Only less than 3% of peaks in the two controls overlap with the ChIP data, indicating that the FDR for our method is small. It should be noted that the PCWC is also available in the MACS (version 1.4) program. Thus we compared the relative effectiveness of our Bayesian model with the Poisson-based MACS model. Under the default parameter condition, the MACS program failed for peak calling for the two independent controls due to the large *-mfold* (*-mfold*=32) parameter, resulting in unavailable construction of the background model. When we decreased the *-mfold* parameter to 8 (suitable for building the background model), a total of 3 278 and 3 486 candidate peaks are called with the FDRs of 5% and 8%, respectively, which is higher than the FDR of our Bayesian model. The advantages of the Bayesian model relies on the genome-wide scanning to empirically measure the background distribution, while the Poisson-based model is subject to overestimation of the number of candidate peaks, therefore resulting in higher FDRs (Kharchenko et al, 2008).

H3K4me3 data

We also used the H3K4me3 ChIP-seq data (8.77×10^6 uniquely mapped reads, accounting for 69.7%) to compare the peak calling between the Bayesian model and the MACS model. Using the MACS with no controls, a total of 14 346 peaks (Table 2) were identified when the cutoff was set to *P=1e-05* (default). With the use of the Bayesian model ($Pr(\#reads \geq 9) \leq 0.01$), we identified 7 539 peaks, of which 7 518 (99.7%) were overlapped with the MACS model with no controls, indicating a high accuracy for the Bayesian model. To avoid the potentially high false negative rate resulting from the stringent cutoff applied, we also tested a relaxed cutoff for the Bayesian model ($Pr(\#reads \geq 4) \leq 0.05$, and we identified a total of 11,536 peaks. Compared with the MACS model using different cutoffs, the relaxed Bayesian model remained high overlapping rates with the MACS data *P=1e-08*, again indicating the effectiveness of PCWC using our Bayesian model.

Table 2 Comparison of peak callings between Bayesian and MACS models

Options	MACS	Bayesian Overlap	Overlapping（%）
1e-05	14 346	11 312	98.06
1e-06	12 627	11 116	96.36
1e-07	11 663	10 827	93.85
1e-08	10 841	10 432	90.43
1e-09	10 235	10 020	86.86

For the Bayesian model, the cutoff was set to $Pr(\#reads \geq 4) \leq 0.05$.

As the likelihood function used in the Bayesian model is subject to simulation for optimizing the number of reads *a*, we next addressed whether the simulation method based on 10 000 bootstrapping per chromosome with *w*=50 bp random bins would be biased in our method. Firstly, we evaluated the effects of number of bootstrappings (ranging from 1 000 to 50 000), and no bias was observed when analyzing the H3K4me3 ChIP-seq data (Table 3). Similar results were obtained when using the VDR ChIP-seq data (data not shown). We then assessed the effects of random bin size (*w*) and found no bias there either, suggesting that the simulation method itself would not generate bias for the parameters *a* and *c* (Figure 4).

Table 3 Effects of bootstrapping on the cutoff threshold

#Reads	1 000 bootstrap	5 000 bootstrap	10 000 bootstrap	20 000 bootstrap	50 000 bootstrap
1	0.148	0.148	0.144	0.141	0.144
2	0.077	0.083	0.084	0.076	0.081
3	0.052	0.060	0.062	0.052	0.056
4	0.037	0.040	0.048	0.038	0.042
5	0.027	0.028	0.036	0.029	0.031
6	0.023	0.021	0.026	0.021	0.023
7	0.017	0.016	0.018	0.015	0.016
8	0.012	0.011	0.013	0.010	0.011
9[a]	*0.008*	*0.008*	*0.008*	*0.007*	*0.007*
10	0.004	0.005	0.007	0.004	0.005
11	0.002	0.004	0.004	0.003	0.004
13	0.000	0.002	0.003	0.001	0.002

[a]: bolded row indicates that bootstraps are not biased.

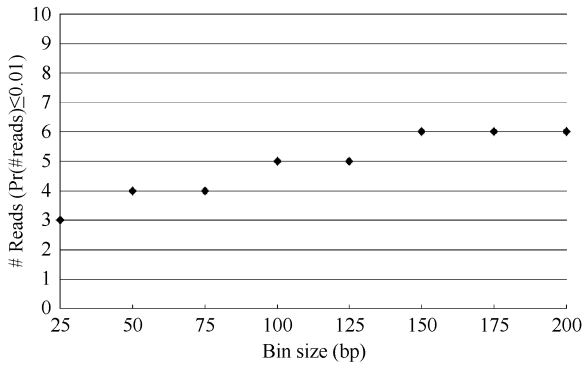

Figure 4 Effects of random bin size on the number of reads *a*

E2F1 data

For the E2F1 ChIP-seq data, a total of 1.188×10^7 reads (~39.0% of total reads) were uniquely mapped onto the genome. Using the PCWC ($Pr(\#reads \geq 9) \leq 0.01$), we identified a total of 11 470 peaks, of which the majority (75.3%, *n*=8 639) span across the annotated Refseq TSSs, consistent with Chen et al's (2008) report that about ~50% of all genes are regulated by E2F1. Our analysis also indicated E2F1 binding regions were very close to TSS (Figure S3), consistent with the previous ChIP-seq and ChIP-chip results (Chen et al, 2008; Xu et al, 2007). The 5 genes (Figure S4) experimentally verified to be positively regulated by E2F1，*Cdc6* (Yan et al, 1998), *Ccne1*, *Ccna2* (DeGregori et al, 1995), *Mcm4* and *Mcm7* (Arata et al, 2000)，were all identified in the E2F1- regulated regions (peaks).

Other data

Using the Illumina platform of ChIP-seq data, we demonstrated the effectiveness and accuracy of PCWC. To further test the effectiveness of the Bayesian model, we used two ABI SOLID platform data sets. The raw colorspace-formatted data were aligned using Bowtie software (Langmead et al, 2009) with maximal 2 mismatches and uniquely mapped reads were used for further analysis.

For the early growth response gene 1 (EGR1) ChIP-seq data (Tang et al, 2010), we identified a total of 7 302 peaks. The KEGG pathway annotation result indicates that genes regulated by EGR1 are significantly enriched in the MAPK, the Wnt and the TGF-beta signaling pathways (hsa04010, hsa04310, hsa04350, respectively), congruent with earlier reported data (Tang et al, 2010).

For the MNase-seq data in two replicates (Valouev et al, 2011), a total of 533 401 and 514 135 peaks were identified, respectively, where 83.12% overlapped between the two replicates, further supporting the effectiveness of peak calling with no control.

***In silico* data**

To demonstrate the effectiveness of ChIP-seq PCWC, we *in silico* generated 1×10^7 36-bp reads using simreads of the Rmap package (Smith et al, 2009) to produce a simulated control sample. As expected, only 428 peaks were generated based on the Bayesian model. Notably, the number of peaks called from the *in silico* data is equivalent with the two independent controls (Table 1), suggesting that the PCWC is sensitive.

CONCLUSIONS

In this study, we demonstrated an informative analysis of ChIP-seq PCWC based on the Bayesian framework approach. By the application to multiple ChIP-seq data sets, we have demonstrated the effectiveness of our method for both transcriptional factor and chromatin modification ChIP-seq experiments. Notably, the demonstration PCWC's effectiveness showed no superiority over the method with controls. The purpose of this study is to demonstrate a potential alternative strategy for designing ChIP-seq experiments and identifying transcriptional factor binding sequences without using controls.

For ChIP-chip analysis, the tiling array may suffer from cross-hybridizations, therefore, the fold ratio for peak calling relative to background is required. By contrast, our analyses indicated the effectiveness of ChIP-seq PCWC, likely due to the finer resolution and greater signal-to-noise ratio of the ChIP-seq data (Rozowsky et al, 2009). Meanwhile, as opposed to the Poisson-based model for ChIP-seq peak calling, our Bayesian model is dependent on genome-wide scanning to empirically measure the background distribution, resulting in the higher selectivity and lower FDRs for the PCWC.

Contrary to RNA-seq, ChIP-seq is more confined by relatively complicated pre-ChIP experiments, e.g., antibody specificity, large amount of cells (usually no less than 1×10^7 cells), formaldehyde cross-link and supersonic shearing, *etc*. If the ChIP experiments ensure high antibody specificity, the following high-throughput sequencing with deep coverage to identify peaks without control is both plausible and robust. While depth-of-sequencing issues (Kharchenko et al, 2008) exist in the ChIP-seq experiments, our analyses suggest that no less than 10 million effective reads (uniquely mapped) are necessary for PCWC.

Moreover, improvements of ChIP-seq experimental strategies without (or with) control data in an appropriate way may be preferable. Based on our survey, two biologically independent replicates is highly replicable (s shown in Table 1) as previously suggested (Park, 2009). Therefore, two replicates of ChIP-seq would be sufficient for peak calling. Meanwhile, if PCWC is determined, the integration with other data types will be essential. For example, the integration of ChIP-seq data with gene expression data (including microarray gene expression

data or RNA-seq data) may maximize the interpretation of gene regulatory network.

Abbreviations: ChIP-seq: chromatin immunoprecipitation followed by sequencing; NGS: next generation sequencing; ChIP-chip: chromatin immunoprecipitation followed by tiling microarray; MACS: model-based analysis of ChIP-seq; PCWC: peak calling without controls; FDR: false discovery rates; GO: gene ontology; VDR: vitamin D receptor; H3K4me3: H3 trimethylated at lysine 4; ESCs: embryonic stem cells

Acknowledgments: We are thankful to Shao-Bin XU (Kunming Institute of Zoology, CAS) for his support on super-computing service, and to Yu-qi ZHAO (Kunming Institute of Zoology, CAS) for his helpful discussion.

References

Arata Y, Fujita M, Ohtani K, Kijima S, Kato J-y. 2000. Cdk2-dependent and -independent Pathways in E2F-mediated S Phase Induction. *J Biol Chem*，**275**(9): 6337-6345.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*，**2**: 28-36.

Cairns J, Spyrou C, Stark R, Smith ML, Lynch AG, Tavare S. 2011. BayesPeak--an R package for analyzing ChIP-seq data. *Bioinformatics*，**27**(5): 713-714.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*，**133**(6): 1106-1117.

Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim TK, He HH, Zieba J, Ruan Y, Bickel PJ, Myers RM, Wold BJ, White KP, Lieb JD, Liu XS. 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods*，**9**(6): 609-614.

Choi H, Nesvizhskii AI, Ghosh D, Qin ZS. 2009. Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. *Bioinformatics*，**25**(14): 1715-1721.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA*，**107**(50): 21931-21936.

DeGregori J, Kowalik T, Nevins JR. 1995. Cellular targets for activation by the E2F1 transcription factor include DNA synthesis- and G1/S-regulatory genes. *Mol Cell Biol*, **15**(8): 4215-4224.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1): 207-210.

Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. 2008. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**(15): 1729-1730.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10): R80.

Guenther MG, Lawton LN, Rozovskaia T, Frampton GM, Levine SS, Volkert TL, Croce CM, Nakamura T, Canaani E, Young RA. 2008. Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. *Genes Dev*, **22**(24): 3403-3408.

Ho JW, Bishop E, Karchenko PV, Negre N, White KP, Park PJ. 2011. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, **12**: 134.

Hower V, Evans SN, Pachter L. 2011. Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics*, **12**: 15.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**(1): 44-57.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res*, **36**(16): 5221-5231.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**(5569): 916-919.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, **26**(12): 1351-1359.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**(3): R25.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**(15): 1966-1967.

Madigan D, Ridgeway G. 2003. Bayesian data analysis. In Ye, N. (eds). The Handbook of Data Mining CRC Press, USA: 103-131.

Micsinai M, Parisi F, Strino F, Asp P, Dynlacht BD, Kluger Y. 2012. Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res*, **40**(9), e70.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**(7153): 553-560.

Osmanbeyoglu H, Hartmaier R, Oesterreich S, Lu X. 2012. Improving ChIP-seq peak-calling for functional co-regulator binding by integrating multiple sources of biological information. *BMC Genomics*, **13**(Suppl 1): S1.

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

Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A, Handunnetthi L, Handel AE, Disanto G, Orton SM, Watson CT, Morahan JM, Giovannoni G, Ponting CP, Ebers GC, Knight JC. 2010. A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res*, **20**(10): 1352-1360.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, **27**(1): 66-75.

Seuter S, Vaisanen S, Radmark O, Carlberg C, Steinhilber D. 2007. Functional characterization of vitamin D responding regions in the human 5-Lipoxygenase gene. *Biochim Biophys Acta*, **1771**(7): 864-872.

Sinkkonen L, Malinen M, Saavalainen K, Vaisanen S, Carlberg C. 2005. Regulation of the human cyclin C gene via multiple vitamin D3-responsive regions in its promoter. *Nucleic Acids Res*, **33**(8): 2440-2451.

Smith AD, Chung W-Y, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ. 2009. Updates to the RMAP short-read mapping software. *Bioinformatics*, **25**(21): 2841-2842.

Spyrou C, Stark R, Lynch AG, Tavare S. 2009. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*, **10**: 299.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**(5891): 956-960.

Tang C, Shi X, Wang W, Zhou D, Tu J, Xie X, Ge Q, Xiao PF, Sun X, Lu Z. 2010. Global analysis of in vivo EGR1-binding sites in erythroleukemia cell using chromatin immunoprecipitation and massively parallel sequencing. *Electrophoresis*, **31**(17): 2936-2943.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*, **5**(9): 829-834.

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature*, **474**(7352): 516-520.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221): 470-476.

Xu X, Bieda M, Jin VX, Rabinovich A, Oberley MJ, Green R, Farnham PJ. 2007. A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res*, **17**(11): 1550-1561.

Yan Z, DeGregori J, Shohet R, Leone G, Stillman B, Nevins JR, Williams RS. 1998. Cdc6 is regulated by E2F and is essential for DNA replication in mammalian cells. *Proc Natl Acad Sci USA*，**95**(7): 3603-3608.

Zella LA, Meyer MB, Nerenz RD, Lee SM, Martowicz ML, Pike JW. 2010. Multifunctional enhancers regulate mouse and human vitamin D receptor gene transcription. *Mol Endocrinol*, **24**(1): 128-147.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**(9): R137.