Quantitative Biology 2017, 5(3): 226–235
DOI 10.1007/s40484-017-0111-8

REVIEW

An introduction to computational tools for differential binding analysis with ChIP-seq data

Shiqi Tu[1,2] and Zhen Shao[1,*]

[1] CAS Key Laboratory of Computational Biology, Collaborative Innovation Center for Genetics and Developmental Biology, CAS-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China.
* Correspondence: shaozhen@picb.ac.cn

Received March 31, 2017; Revised June 5, 2017; Accepted June 8, 2017

***Background*: Gene transcription in eukaryotic cells is collectively controlled by a large panel of chromatin associated proteins and ChIP-seq is now widely used to locate their binding sites along the whole genome. Inferring the differential binding sites of these proteins between biological conditions by comparing the corresponding ChIP-seq samples is of general interest, yet it is still a computationally challenging task.**

***Results*: Here, we briefly review the computational tools developed in recent years for differential binding analysis with ChIP-seq data. The methods are extensively classified by their strategy of statistical modeling and scope of application. Finally, a decision tree is presented for choosing proper tools based on the specific dataset.**

***Conclusions*: Computational tools for differential binding analysis with ChIP-seq data vary significantly with respect to their applicability and performance. This review can serve as a practical guide for readers to select appropriate tools for their own datasets.**

Keywords: ChIP-seq; peak calling; differential binding analysis; computational tools

INTRODUCTION

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has been extensively used to determine the binding sites of chromatin associated proteins and the enrichments for specific histone modifications on a genome-wide scale [1,2]. One of the most important downstream analyses of ChIP-seq data is to identify genomic regions with a significant change in ChIP-seq signal across biological conditions [3]. This analysis is a critical step towards understanding the mechanism regulating dynamic changes of gene expression during tissue development [4–6] and the onset of diseases [7]. Specifically, identifying the genomic loci differentially marked by histone modifications across cell types has been widely used to search for cell type specific *cis*-elements as well as the regulators associated with these elements [4,8,9]. In addition, it has been revealed that lineage master regulators can cooperate with cohesin proteins and transcription factors in signaling pathways to modulate their chromatin occupancy and establish lineage specific gene expression programs [9–11].

In many previous studies, people simply used the overlap between the peaks identified from different ChIP-seq samples to define common and specific peaks [12–14]. However, it has been suggested that the cell type specific peaks defined by this approach may contain a considerable fraction of false positives, and a rigorous comparison based on statistical models specifically designed for this purpose is more recommended [8,15]. In particular, it usually gives more reliable results to perform ChIP-seq data analysis in a quantitative manner, especially for cross-condition comparisons [8,9,16]. For example, several recent studies suggested to quantitatively combine the ChIP-seq signal intensity of a peak, sometimes called peak height, as well as the distance from this peak to a candidate target gene to represent its regulatory potential to the gene [16–18]. It has been shown that such a quantitative measure can integrate with

other observations to better infer the functional impact of the protein's chromatin binding being studied. Another example is that Shao *et al.* performed a systematic comparison of ChIP-seq data between different cell types, and found that, for histone modifications like H3K4me3 and H3K27ac, quantitative changes of ChIP-seq signals strongly correlate with the expression changes of target genes as well as the binding of cell type-specific regulators [8]. Following this direction, Xu *et al.* suggested that, although in mammalian genomes distal enhancer elements often have clearly higher cell type specificity than proximal promoters, it is still important to use the quantitative changes of associated histone modifications to define a high-confidence set of cell type specific enhancers, especially when the difference between the cell types under comparison is mild [9].

Despite the importance of differential binding analysis with ChIP-seq data and the increasing need of methods for this analysis, it is still computationally challenging to reliably assess the statistical significance of changes in signal intensity on a genome-wide scale, due to the high level of noise and variability intrinsic to ChIP-seq data. More specifically, ChIP-seq is a multi-step experiment where biases may be introduced at each step [19,20], leading to a generally limited data reproducibility [21,22]. Among others, the amount of input material, efficiency of antibody, sequencing quality and depth may vary considerably from an experiment to another [19,20]. As a result, ChIP-seq samples typically have quite different signal-to-noise ratios, especially for those generated from different batches and/or labs, which makes it extremely difficult to quantitatively compare the signal intensities between samples [8].

In recent years, quite a number of computational tools have been developed to address the problem [3,15]. These tools take advantage of different statistical techniques and vary in the range of applicability. This review aims to summarize existing computational tools for differential binding analysis with ChIP-seq data, according to their scope of application as well as strategy of statistical modeling (Figures 1 and 2). For simplicity, we primarily focus on the comparison of ChIP-seq samples between two biological conditions, with or without replicates. Besides, we assume the sequencing reads have been appropriately mapped to a reference genome [23].

STRATEGIES FOR DIFFERENTIAL BINDING ANALYSIS WITH ChIP-SEQ DATA

Peak calling for ChIP-seq data

Typically, a considerable proportion of the mapped reads of a ChIP-seq sample are dispersed throughout the genome, while the others cluster together constituting reads-enriched regions, termed peaks (Figure 1, top) [24,25]. To be noted, ChIP-seq reads falling outside of peak regions are predominately contributed by background noise or non-specific binding, while the peaks with significantly elevated ChIP-seq signal intensities generally represent stable binding sites of the protein being ChIPed or genomic regions heavily marked by a specific histone modification. A number of algorithms have been developed to identify significant peaks on a genome-wide scale [25–27]. Here, we provide a brief overview of the available tools for ChIP-seq peak calling. One of the reasons is that a large number of computational tools for differential binding analysis require users to provide a set of pre-defined peaks for the ChIP-seq samples under comparison and then focus on modeling the ChIP-Seq signals at peak regions [8,28,29]. Another reason is that quite several peak calling programs claimed that they could also be applied to identify differential peaks between two ChIP-seq samples by taking one of them as input [25,26]. Thus, the basic concept and strategies of ChIP-seq peak calling can serve as a valuable reference for differential binding analysis.

To be noted, the characteristics of peaks, especially their size, can differ substantially depending on the protein or histone modification targeted in the experiment. For example, the majority of transcription factors and many histone modifications like H3K4me3 and H3K27ac tend to have narrow peaks, with a size ranging from several hundred to a few thousand base pairs [25]. MACS has been widely used to perform peak calling on such ChIP-seq samples, especially those for transcription factors which are usually associated with sharp and isolated peaks [24,25]. For some other histone modifications such as H3K9me3 and H3K36me3, they tend to form broad genomic domains with diffusive ChIP-seq signals, which can span up to hundreds or even thousands of kilo base pairs [30,31]. There are computational tools that are specifically devised to handle such situations. For example, SICER is developed to identify large spatial clusters of ChIP-seq reads [26], while RSEG utilizes a hidden Markov model (HMM) to detect broad epigenomic domains with consecutively elevated ChIP-seq signals [32]. Notably, both MACS and SICER accept a treatment ChIP-seq sample and an optional input sample as the negative control for peak calling. The latter is highly recommended for a practical ChIP-seq experiment design and can be used to account for local biases resulting from read mappability, DNA repeats, local GC content and so on [25].

Many studies choose to generate multiple ChIP-seq samples for the same biological condition, with the aim to assess the variability and reproducibility of ChIP-seq signals. To make a full use of the replication, two

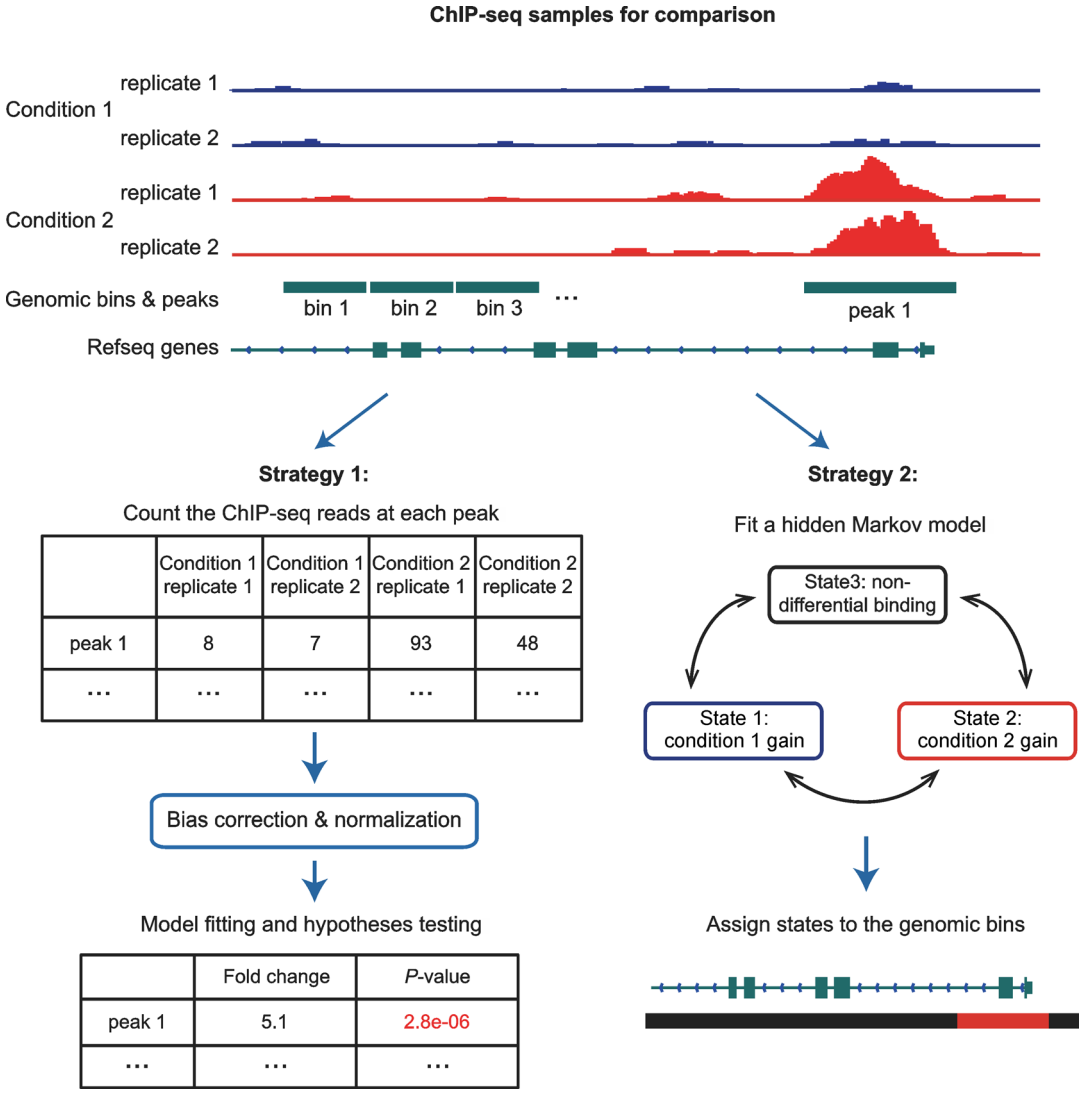

Figure 1. General work flow of two popular strategies for differential binding analysis with ChIP-seq data.

strategies are usually employed to integrate the ChIP-seq replicates and derive a single list of peak regions. One of them directly performs a joint analysis over the replicates, which has been shown to detect peak boundaries with high precision [33]. The other strategy first calls peaks on each individual sample, and then uses measurements such as IDR (irreproducible discovery rate) to select the peak regions with high reproducibility across replicates [21,34].

Differential binding analysis based on pre-defined peaks

Since peak regions with significantly elevated ChIP-seq signals are often of the highest interest across the whole genome, especially for the factors with sharp binding peaks, a lot of computational tools choose to perform differential binding analysis only on the peaks identified from the ChIP-seq samples under comparison [8,28,29,35]. To exploit such tools, users typically need to start with peak calling on each ChIP-seq sample involved. But, it should be noted that peak calling on each individual sample is usually a fundamental step of ChIP-seq data analysis and the obtained peaks could also be used for other analyses [9]. Thus, it will be easy to integrate the results of differential binding analysis with pre-defined peaks with the other analyses. In the start of a differential binding analysis with pre-defined peaks, usually the peak regions of all the samples under comparison are first merged into a consensus set of peaks, which defines the search space where differential ChIP-seq signals are expected to find (Figure 1, lower left). In principle, these peaks serve as the reference genes in a typical differential expression analysis with RNA-seq

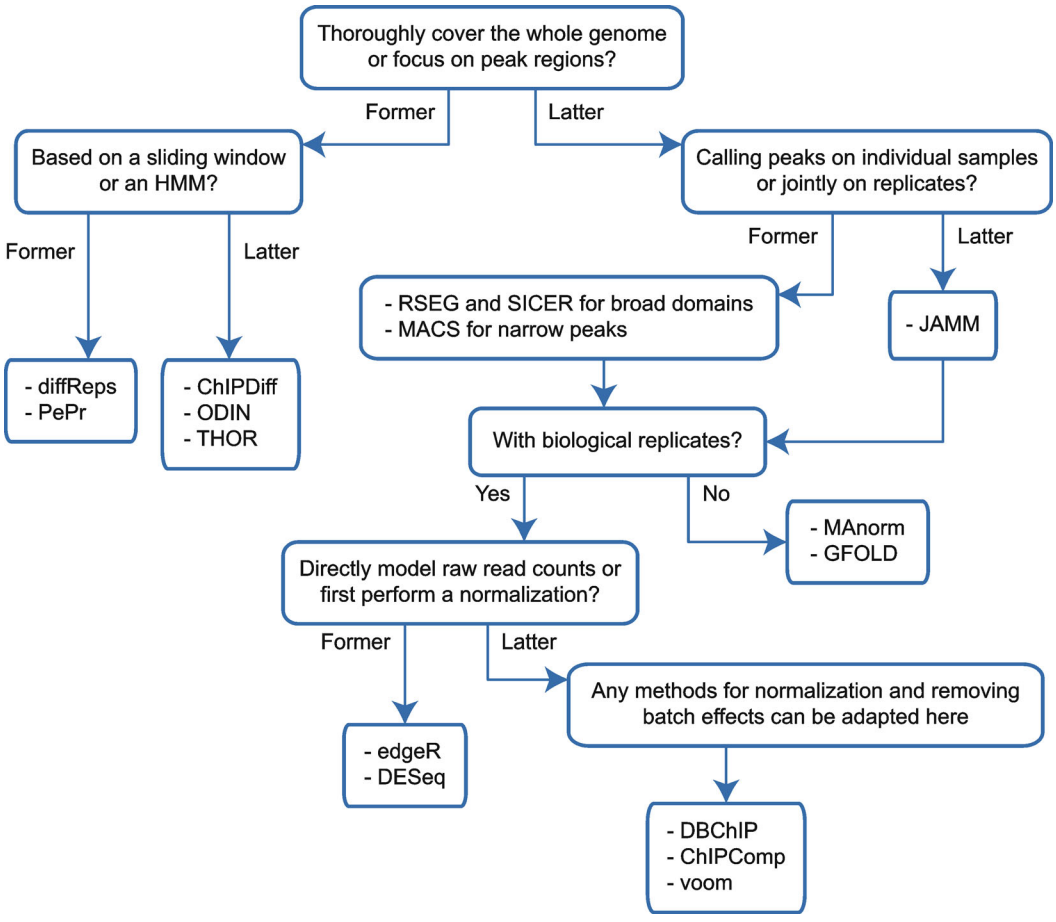

Figure 2. A diagram to classify most of the computational tools for differential binding analysis discussed in the main text, according to their strategy of statistical modeling and range of applicability.

data [36,37]. Therefore, many solid statistical models initially developed for calling differentially expressed genes with RNA-seq data can then be adapted to ChIP-seq data [38].

Biological replicates may not be available in practical studies. In the extreme case, only one ChIP-seq sample is available for each of two conditions. In this case, many peak calling tools such as MACS and SICER can also be used to identify differential peaks, by taking one of the two samples as treatment and the other as negative control [3,25]. Although these methods come up with a *P* value to assess the statistical significance of ChIP-Seq signal change at each detected differential peak, it should be strongly emphasized that, without replication, there is no way in principle to estimate biological variation in the measured signal intensities and, hence, no meaningful inference regarding the population can be made [39]. Therefore, any *P* value deduced in this context has only exploratory value. On this account, it may make more sense to measure the practical significance of signal changes. MACS and SICER calculate a fold change of ChIP-seq signal intensity for each candidate differential peak by normalizing each sample on basis of its library size, which is often inappropriate for ChIP-seq data as different samples may have highly distinct signal-to-noise ratios. Previously, MAnorm proposed to normalize two ChIP-seq samples based on their common peak regions [8]. The method introduces a hypothesis that, when two ChIP-seq samples share a significantly larger number of common peaks than expected by chance, the binding of the targeted protein in the experiment at these common peaks is very likely to be mediated by largely the same mechanism. Hence, no global binding changes should be expected at these peak regions. Based on this hypothesis, MAnorm utilizes the traditional M-A plot, in which the log2 fold ratios of signal intensities are plotted against the average log2-transformed intensities between two samples, and fits a linear model between the M and A values in their common peaks. Then, the linear model is used as a reference to correct the M and A values of all peak regions. Through that, the variation in signal-to-noise ratio across samples is largely moderated, leading to a

more robust estimation of the fold change of ChIP-seq signal intensity [3,8]. Besides MAnorm, in the field of RNA-seq data analysis there are methods that propose to improve the estimation of fold change by taking into account the large variance of M values calculated from small read counts [37,40]. For example, DESeq applies a variance-stabilizing transformation on RNA-seq count data [37], making the M value comparable between transcripts with different expression levels and, thus, providing a more reasonable ranking of differentially expressed genes. Another example is GFOLD, which shrinks the M values calculated from small read counts to 0 under a Bayesian framework [40]. In principle, these methods can be easily adapted to the differential binding analysis with ChIP-seq data.

When ChIP-seq replicates are available, it becomes feasible to model the biological variation at each peak region besides the technical variation introduced during sample preparation and sequencing. This strategy is widely used in the computational tools developed for differential expression analysis of RNA-seq data such as edgeR and DESeq [36,37], which are inherited by DiffBind and DBChIP by adapting their statistical models to ChIP-seq data [28,38]. These tools introduce potential sources of bias like sequencing depth into the model and perform a normalization as well as a differential binding analysis simultaneously. For the majority of other methods, however, normalization between ChIP-seq samples is required prior to conducting a differential binding analysis (Figure 1, lower left). In principle, any normalization approach can be adapted to these methods. To normalize away the most concerning factor across samples, which is sequencing depth, some methods rely on the total or effective read counts [35]. Such approaches tend to perform poorly when peak regions are highly heterogeneous across samples [8]. Besides, peaks associated with very large read counts may considerably bias the normalization result. Normalization methods that avoid using total read counts include those implemented in THOR, MAnorm and DBChIP [8,15,28]. Basically, these methods perform a normalization on basis of peak regions that are supposed to be invariant across samples, such as the promoter regions of house-keeping genes [15] and the observed common peaks [8]. In general, ChIP-seq signals in these regions are more reliable and stable than in the others, and the methods are therefore resistant to individual highly-represented peaks. Moreover, some computational tools can handle additional sources of bias by considering local GC-content, input subtraction and sequence mappability along the genome [15,28,29,32]. In spite of these bias correction and signal normalization procedures at the level of individual samples, ChIP-seq data may still be associated with serious batch effects, especially in the large-scale studies where plenty of samples are involved. COMBAT and ARSyN, which are initially developed to remove batch effects in microarray data [41,42], have been shown to work well with normalized sequencing data [39].

The first computational tools for differential analysis with sequencing count data have used discrete distributions such as the Poisson and negative binomial [28,35,37,43]. The negative binomial distribution can be viewed as a gamma-Poisson hierarchy. From this perspective, it explicitly models the underlying biological variance, which is believed to be the primary cause of the observed over-dispersion in sequencing data, in addition to the technical variance expected from randomly sampling from a pool of molecules [37]. On the other hand, though it follows the nature of count data, the use of discrete distributions is not a requisite for an accurate differential binding analysis. In principle, applying continuous distributions such as the normal distribution to sequencing data analysis is valid as long as the mean-variance relationship for counts is carefully modeled [44]. voom applies a log-transformation to the normalized read counts while learning the global variance structure for transformed values [44]. The method unlocks a repository of statistical methodologies originally devised for microarray data, and has been shown to work well compared with many approaches based on discrete distributions [44,45]. In either case, nearly all the methods have made an effort to reduce the uncertainty associated with peak-specific variance estimates, considering that a highly limited number of replicates is often the case. Among others, a widely adopted strategy is to borrow information between peaks [36,37,44], given the parallel structure in a typical differential binding analysis in which the same model is fitted to each peak. For example, DESeq improves the variance estimates by fitting a mean-variance curve and, hence, sharing information between peaks with close signal levels [37]. voom adapts count data to the empirical Bayes framework implemented in limma and integrates information from all peak regions into a common prior distribution of the variance associated with each peak [44,46].

One-step differential binding analysis without the requirement for pre-defined peaks

Sometimes users are mainly interested in the genomic regions with significant ChIP-seq signal changes across conditions. Thus, they may prefer to perform a one-step differential binding analysis without doing peak calling for each ChIP-seq sample in advance. Following this direction, an alternative strategy of differential binding analysis directly seeks for the changes in ChIP-seq signal intensity throughout the whole genome, without the need of calling peaks beforehand [15,32,47–50]. Approaches

following this strategy address an obvious issue arising from using peak calling based methods, in which the search for differential binding sites is restricted to pre-defined peak regions and artefacts may be introduced when applying a certain cutoff to define peaks. These approaches can be further classified into two categories. One of them scans the whole genome with a sliding window and consecutively performs the same statistical test on the ChIP-seq signals at each window [49,50], where the window size is usually selected to match the typical size of a ChIP-seq signal enriched region. The other class takes advantage of more sophisticated segmentation techniques such as HMM [15,32,47,48], where the genome is fragmented into sequences of bins and a putative hidden state is then inferred for each bin to indicate whether it is associated with differential ChIP-seq signal (Figure 1, lower right). One of the reasons accounting for the superiority of using an HMM is that, for a target bin, it incorporates ChIP-seq signals lying in the vicinity to improve the inference made for the bin. These methods are therefore robust to the selection of bin size and can achieve a wide range of resolution in identifying differential binding sites.

A famous application of HMM in ChIP-seq data analysis is ChromHMM, which takes a set of ChIP-seq samples of multiple chromatin marks generated under the same cellular condition as input, and systematically detects their combinatorial binding patterns as representation of local chromatin states [51,52]. It has been applied to a large panel of human cell types to derive a systematic annotation of chromatin states along the whole genome for each of them [53]. The method leverages the correlation between different marks and, thus, significantly improves the interpretation of observed ChIP-seq signals. However, employing ChromHMM requires the availability of multiple ChIP-seq samples for various marks in a single condition, which seriously limits its applicability. Other HMM based methods, such as THOR [15], are specifically devised to call genomic regions of differential ChIP-seq signals between a pair of biological conditions. THOR accepts ChIP-seq samples for a single mark from two biological conditions, and encodes the significance as well as the direction of ChIP-seq signal changes into the underlying states of an HMM (Figure 1, lower right).

Despite the clear advantages of HMM based methods, they usually make a stronger assumption on the observed data than the methods focusing on peak regions. More specifically, most of these methods train an HMM with a very limited number of hidden states (typically 3 for a comparison between two conditions; Figure 1, lower right) to model the observed ChIP-seq signals [15,32,48], which could be impractical for a real ChIP-seq dataset and may result in a loss of flexibility. In particular, compared with peak calling based methods, HMM based methods may be less sensitive to quantitative changes in ChIP-seq signal intensity between a pair of closely related conditions (e.g., in studies of personal epigenomes [54]).

Practices for selecting the appropriate tools for a custom differential binding analysis

In Figure 2, we use a diagram to depict the major characteristics and applicability of the computational tools introduced in this review. The diagram also serves as a practical decision tree for researchers to choose proper methods depending on their own dataset. Besides, there are several points that we think are necessary to highlight.

Firstly, the computational tools based on pre-defined peaks typically focus on modeling the ChIP-seq signals at peak regions, and, thus, tend to be less sensitive to the variation in signal-to-noise ratio across samples compared to the methods involving background signals in modeling, e.g., the HMM based methods. Hence, for most transcription factors and the histone modifications associated with narrow peaks (e.g., H3K4me3 and H3K9/27ac), methods based on pre-defined peaks should take the priority. On the other hand, HMM based methods may better fit for analysis with the histone modifications typically constituting broad domains (e.g., H3K9me3 and H3K36me3), as the variations of local ChIP-seq signals for these marks may not be quite informative and the HMM based methods can borrow information from flanking genomic regions to help identifying large chromatin domains with a continuous change [47]. Besides, some HMM based methods can also be utilized to detect subtle ChIP-seq signal changes within a broad domain, such as the partial losses/gains of histone modifications [15,48].

Secondly, for the peak calling procedure, JAMM is recommended in the cases where biological replicates are available [33]. It can integrate information from replicates and determine peak boundaries with high precision. Particularly, it could resolve neighboring narrow peaks and, thus, increase the resolution of the downstream differential binding analysis.

Finally, tools such as DESeq and edgeR are originally designed for RNA-seq data, which are expected to have less variability between samples than ChIP-seq data. These methods believe sequencing depth is the only concerning factor that needs to be normalized between samples, and integrate the normalization procedure into the differential analysis [36,37]. In the case where this assumption does not hold (e.g., when the ChIP-seq samples under comparison are generated from different batches or labs), it is wiser to first extensively correct for confounding factors and normalize ChIP-seq signal intensities, prior to performing the differential binding

analysis. The normalization approaches implemented in MAnorm and THOR are recommended, as they are robust to the variation of signal-to-noise ratios across samples [8,15]. For reference, Table 1 summarizes the main features and practical utility of each method shown in Figure 2.

Table 1. Summary of the main characteristics and applicability of the computational tools shown in Figure 2.

Name	Method description	Characteristics and applicability
diffReps [49]	Using a sliding window to scan the whole genome	Multiple statistical tests are designed to handle both of the cases with and without biological replicates
PePr [50]	Using a sliding window to scan the whole genome	The negative binomial test is used to assess differential binding; biological replicates are required
ChIPDiff [47]	Modeling the whole genome with a 3-state HMM using the beta-binomial hierarchy as emission	The Bayesian hierarchy implicitly augments the number of hidden states, making the method sensitive to differential binding; it does not support replicates
ODIN [48]	Modeling the whole genome with a 3-state HMM using the binomial or a mixture of Poisson as emission	Refinement is performed for the specific type of ChIP-seq data, based on whether they are associated with sharp peaks or broad domains; it does not support replicates
THOR [15]	An extension of ODIN using the negative binomial as emission	THOR extends ODIN by supporting biological replicates and providing a series of procedures for bias correction and normalization
RSEG [32]	Using an HMM to identify broad domains with consecutively elevated ChIP-seq signals	The method accepts a treatment sample and an optional input sample as control, suited for histone modifications constituting broad domains such as H3K9me3 and H3K36me3
SICER [26]	Leveraging enrichment information from neighboring regions to identify chromatin domains of enriched ChIP-seq signals	Similar to RSEG, except that the resolution of chromatin domains identified by SICER is explicitly specified by users
MACS [25]	Using the Poisson distribution with a dynamic background level to call ChIP-seq peaks	The dynamic background is used to account for biases of local chromatin regions, suited for most transcription factors and histone modifications associated with sharp peaks
JAMM [33]	Incorporating information from replicate samples to perform peak calling	The method calls peaks jointly on replicates and, thus, improves the precision for determining peak boundaries
MAnorm [8]	Normalizing two ChIP-seq samples based on their common peak regions	The method does not assume that the genome-wide distribution of ChIP-seq signal intensities is invariant across samples and, thus, shows a robust behavior; it is suitable for ChIP-seq samples sharing a significant number of peaks
GFOLD [40]	Given two ChIP-seq samples, modeling the distribution of fold changes in signal levels under a Bayesian framework	GFOLD shrinks considerably a fold change calculated from small read counts to 1, leading to a more reliable ranking of differential peaks
edgeR [36]	Modeling raw counts using the negative binomial distribution and identifying differential peaks; originally developed for RNA-seq data	The method incorporates information from all peaks to estimate the common dispersion parameter, leading to a robust behavior even with the minimal level of replication
DESeq [37]	Modeling raw counts using the negative binomial distribution and identifying differential peaks; originally developed for RNA-seq data	DESeq generalizes edgeR by allowing an arbitrary mean-variance relationship and, thus, is more adaptive to different datasets
DBChIP [28]	Using a generalized linear model with the negative binomial distribution to detect differential peaks	DBChIP is specifically designed for ChIP-seq samples of transcription factors; it can handle experiment designs of arbitrary complexity (not limited to two-condition comparisons)
ChIPComp [29]	Using a generalized linear model with the Poisson distribution to detect differential peaks	ChIPComp is suited for both sharp peaks and broad domains; it can handle experiment designs of arbitrary complexity
voom [44]	Converting count data into normalized continuous values and entering them into the limma package [46] to perform a differential analysis; originally developed for RNA-seq data	voom aims to remove the heteroscedasticity intrinsic to count data by learning the mean-variance relationship and introducing a precision weight for each observation; it unlocks a large repository of tools originally designed for continuous measurements, including the limma [46]

OUTLOOK

ChIP-seq has become the standard technology for determining transcription factor binding sites and histone modification enrichments on a genome wide scale, and the tools for its differential binding analysis are continuing to evolve. Despite the importance of this analysis, the agreement between the results obtained by applying different tools could be surprisingly low [3]. This observation strongly stresses the importance of choosing proper methods based on the specific experimental setting and application scenario. In addition, analytic challenges can emerge under certain contexts. Here we take two scenarios as examples to illustrate the point.

Methods for ChIP-seq data normalization are usually based on the genomic loci with signal levels that are expected to be invariant across samples [8,15], or the assumption that the majority of the loci being analyzed are not differentially bound [28,37]. Such methods, however, may not be appropriate for the case where a global change of chromatin binding takes place (e.g., when an enzyme catalyzing the histone modification under comparison is functionally depleted from the cells) [55]. ChIP-seq data normalization in such cases is difficult to accomplish, due to the lack of "invariant" loci. To our knowledge, no computational tools are currently available to resolve this problem. Recently, spiking experiment was proposed to specifically deal with this problem [55]. In the experimental procedure, a constant, low amount of chromatin sample from a foreign species is added to the chromatin samples of interest prior to the immunoprecipitation step. Then, this ''spike'' genome is used as an internal reference for adjusting ChIP-seq signal levels across samples.

A universal technical problem associated with analyzing count data is that the mathematical theory of discrete distributions (e.g., the Poisson and negative binomial distributions) is far less tractable than that of the normal distribution, which seriously limits the type of analysis that can be performed on sequencing data. For example, a large number of statistical methods based on the normal distribution have been developed to analyze intensity data from microarrays [44], including those for detecting differential expressions [46], modeling random effects [56], testing gene sets [57,58] and so on. However, most of the tools developed for RNA-seq and ChIP-seq data aim at the differential analysis, and only a few of them can handle complicate experiment designs (see Table 1). To re-use the statistical models originally devised for microarray data analysis, voom is proposed to transform sequencing tag counts into the values showing a continuous manner that can be modeled by the normal distribution [44]. These transformed values can then be input to the models for analyzing microarray signal intensities. The normalization and variance structure learning procedures implemented in voom, however, are specifically tailored for RNA-seq data, which are supposed to have less variability and lower noise level than ChIP-seq data. For future methodology studies, a similar method suited for ChIP-seq data is under expectation.

AUTHORS' CONTRIBUTION

T.S.Q. and Z.S. conceived the study and wrote the manuscript.

ACKNOWLEDGEMENTS

This work was supported by the "100-Talent Program" (Y516C11851) and the interdisciplinary innovation team award of Chinese Academy of Science.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Shiqi Tu and Zhen Shao declare that they have no conflict of interests.

This article is a review article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Mardis, E. R. (2007) ChIP-seq: welcome to the new frontier. Nat. Methods, 4, 613–614
2. Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680
3. Steinhauser, S., Kurzawa, N., Eils, R. and Herrmann, C. (2016) A comprehensive comparison of tools for differential ChIP-seq analysis. Brief. Bioinform., 17, 953–966
4. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330
5. Martens, J. H. and Stunnenberg, H. G. (2013) BLUEPRINT: mapping human blood cell epigenomes. Haematologica, 98, 1487–1489
6. Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., *et al.* (2014) Chromatin state dynamics during blood formation. Science, 345, 943–949
7. Koues, O. I., Kowalewski, R. A., Chang, L. W., Pyfrom, S. C., Schmidt, J. A., Luo, H., Sandoval, L. E., Hughes, T. B., Bednarski, J. J., Cashen, A. F., *et al.* (2015) Enhancer sequence variants and transcription-factor deregulation synergize to construct pathogenic regulatory circuits in B-cell lymphoma. Immunity, 42, 186–198
8. Shao, Z., Zhang, Y., Yuan, G. C., Orkin, S. H. and Waxman, D. J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol., 13, R16
9. Xu, J., Shao, Z., Glass, K., Bauer, D. E., Pinello, L., Van Handel, B., Hou, S., Stamatoyannopoulos, J. A., Mikkola, H. K., Yuan, G.

C., *et al.* (2012) Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. Dev. Cell, 23, 796–811

10. Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., Ramsay, R. G., Odom, D. T. and Flicek, P. (2012) Cohesin regulates tissue-specific expression by stabilizing highly occupied *cis*-regulatory modules. Genome Res., 22, 2163–2175
11. Trompouki, E., Bowman, T. V., Lawton, L. N., Fan, Z. P., Wu, D. C., DiBiase, A., Martin, C. S., Cech, J. N., Sessa, A. K., Leblanc, J. L., *et al.* (2011) Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell, 147, 577–589
12. Fujiwara, T., O'Geen, H., Keles, S., Blahnik, K., Linnemann, A. K., Kang, Y. A., Choi, K., Farnham, P. J. and Bresnick, E. H. (2009) Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol. Cell, 36, 667–681
13. Liu, W., Tanasa, B., Tyurina, O. V., Zhou, T. Y., Gassmann, R., Liu, W. T., Ohgi, K. A., Benner, C., Garcia-Bassets, I., Aggarwal, A. K., *et al.* (2010) PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. Nature, 466, 508–512
14. Yu, M., Riva, L., Xie, H., Schindler, Y., Moran, T. B., Cheng, Y., Yu, D., Hardison, R., Weiss, M. J., Orkin, S. H., *et al.* (2009) Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. Mol. Cell, 36, 682–695
15. Allhoff, M., Seré, K., F Pires, J., Zenke, M. and G Costa, I. (2016) Differential peak calling of ChIP-seq signals with replicates with THOR. Nucleic Acids Res., 44, e153
16. Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C. A., Zhang, Y. and Liu, X. S. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat. Protoc., 8, 2502–2515
17. Ouyang, Z., Zhou, Q. and Wong, W. H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc. Natl. Acad. Sci. USA, 106, 21521–21526
18. Wang, S., Zang, C., Xiao, T., Fan, J., Mei, S., Qin, Q., Wu, Q., Li, X., Xu, K., He, H. H., *et al.* (2016) Modeling *cis*-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome Res., 26, 1417–1429
19. Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T. K., He, H. H., Zieba, J., *et al.* (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nat. Methods, 9, 609–614
20. Meyer, C. A. and Liu, X. S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet., 15, 709–721
21. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res., 22, 1813–1831
22. Furey, T. S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet., 13, 840–852
23. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25
24. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9, R137
25. Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X. S. (2012) Identifying ChIP-seq enrichment using MACS. Nat. Protoc., 7, 1728–1740
26. Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K. and Peng, W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics, 25, 1952–1958
27. Wilbanks, E. G. and Facciotti, M. T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One, 5, e11471
28. Liang, K. and Keles, S. (2012) Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics, 28, 121–122
29. Chen, L., Wang, C., Qin, Z. S. and Wu, H. (2015) A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. Bioinformatics, 31, 1889–1896
30. Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. Cell, 129, 823–837
31. Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Peng, W., Zhang, M. Q., *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet., 40, 897–903
32. Song, Q. and Smith, A. D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics, 27, 870–871
33. Ibrahim, M. M., Lacadie, S. A. and Ohler, U. (2015) JAMM: a peak finder for joint analysis of NGS replicates. Bioinformatics, 31, 48–55
34. Li, Q. H., Brown, J. B., Huang, H. and Bickel, P. J. (2011) Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat., 5, 1752–1779
35. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol. Cell, 38, 576–589
36. Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140
37. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106
38. Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., *et al.* (2012) Differential oestrogen receptor binding

is associated with clinical outcome in breast cancer. Nature, 481, 389–393

39. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., *et al.* (2016) A survey of best practices for RNA-seq data analysis. Genome Biol., 17, 13
40. Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X. and Zhang, Y. (2012) GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics, 28, 2782–2788
41. Johnson, W. E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8, 118–127
42. Nueda, M. J., Ferrer, A. and Conesa, A. (2012) ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. Biostatistics, 13, 553–566
43. Robinson, M. D. and Smyth, G. K. (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, 23, 2881–2887
44. Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol., 15, R29
45. Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 91
46. Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol., 3, Article3
47. Xu, H., Wei, C. L., Lin, F. and Sung, W. K. (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics, 24, 2344–2349
48. Allhoff, M., Seré, K., Chauvistré, H., Lin, Q., Zenke, M. and Costa, I. G. (2014) Detecting differential peaks in ChIP-seq signals with ODIN. Bioinformatics, 30, 3467–3475
49. Shen, L., Shao, N. Y., Liu, X., Maze, I., Feng, J. and Nestler, E. J. (2013) diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. PLoS One, 8, e65598
50. Zhang, Y., Lin, Y. H., Johnson, T. D., Rozek, L. S. and Sartor, M. A. (2014) PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. Bioinformatics, 30, 2568–2575
51. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol., 28, 817–825
52. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods, 9, 215–216
53. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473, 43–49
54. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., *et al.* (2013) Extensive variation in chromatin states across humans. Science, 342, 750–752
55. Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I. M., Herr, W., Hernandez, N., Delorenzi, M., *et al.* (2014) Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res., 24, 1157–1168
56. Smyth, G. K., Michaud, J. and Scott, H. S. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics, 21, 2067–2075
57. Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M. L., Visvader, J. E. and Smyth, G. K. (2010) ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics, 26, 2176–2182
58. Wu, D. and Smyth, G. K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res., 40, e133

