QB
Quantitative Biology 2013, 1(1): 54–70
DOI 10.1007/s40484-013-0006-2

REVIEW

Computational methodology for ChIP-seq analysis

Hyunjin Shin[1], Tao Liu[1], Xikun Duan[2], Yong Zhang[2] and X. Shirley Liu[1,*]

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute/Harvard School of Public Health, Boston, MA 02115, USA
[2] Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai 200092, China
* Correspondence: xsliu@jimmy.harvard.edu

Received October 31, 2012; Revised November 26, 2012; Accepted November 26, 2012

Chromatin immunoprecipitation coupled with massive parallel sequencing (ChIP-seq) is a powerful technology to identify the genome-wide locations of DNA binding proteins such as transcription factors or modified histones. As more and more experimental laboratories are adopting ChIP-seq to unravel the transcriptional and epigenetic regulatory mechanisms, computational analyses of ChIP-seq also become increasingly comprehensive and sophisticated. In this article, we review current computational methodology for ChIP-seq analysis, recommend useful algorithms and workflows, and introduce quality control measures at different analytical steps. We also discuss how ChIP-seq could be integrated with other types of genomic assays, such as gene expression profiling and genome-wide association studies, to provide a more comprehensive view of gene regulatory mechanisms in important physiological and pathological processes.

INTRODUCTION

Recent advances in next generation sequencing (NGS) technologies have enabled scientists to investigate a variety of molecular events occurring on the genome with high resolution and accuracy [1–4]. These NGS technologies have been applied to many scientific and clinical areas including detection of genetic variations (e.g., SNP calling) and quantification of RNA transcripts (e.g., RNA-seq). One of the most successful NGS applications is chromatin immunoprecipitation (ChIP) accompanied by NGS, or ChIP-seq, which can map the *in vivo* genome-wide binding sites of DNA-binding proteins such as transcription factors (TFs) or modified histones.

In chromatin immunoprecipitation, cells are lysed [5] and protein-DNA interactions are crosslinked to form covalent bonds by formaldehyde or other chemical reagents. Then the crosslinked DNA is sheared by sonication or DNA-cutting enzymes (e.g., micrococcal nuclease, often called MNase) into 150–500 bp-long fragments. Those DNA fragments crosslinked with the DNA-binding factor of interest are immunoprecipitated using an antibody specific to the factor. ChIP can be applied to a wide range of DNA binding factors, including TFs, transcription co-activators, co-repressors, chromatin regulators, and modified histones. After reverse cross-linking the protein-DNA complexes, the pulled-down DNA fragments are PCR amplified and then subjected to massively parallel sequencing (see Metzker's review for details of various NGS technologies) [1]. Finally, when the resulting ChIP-seq reads are mapped back to the genome, the locations of the factor-DNA interactions can be identified.

ChIP-seq can provide important insights towards gene regulatory process particularly in combination with transcriptomic profiles from expression microarrays or RNA-seq, since ChIP-seq can help identify genes directly regulated by the factor (see the section *Integrate with gene expression profiles* for details). For instance, ChIP-seq and its predecessor ChIP-chip (i.e., ChIP coupled with tiling microarray technologies) have been used to study how nuclear hormone receptors (e.g., androgen receptor or estrogen receptor) and their cofactors (e.g., FoxA1) cooperate to regulate gene expression in prostate and breast cancers [6,7]. ChIP-seq has also been employed in many stem cell studies to associate the core regulatory

circuitries of stem cell TFs such as Oct4, Nanog or cMyc with cell fate [8–12]. These transcriptional regulation studies have led to the recent development of techniques to capture the higher-order chromatin interactions [9,13–22], which provide clues as to functional interactions between TF binding sites and the promoters of their target genes.

Likewise, since the pioneering studies of histone mark ChIP-seq in the human $CD4^+$ T-cells [23,24], researchers have actively adopted ChIP-seq to investigate the biological functions of many histone marks. The ENCODE and modENCODE consortia, for example, conducted ChIP-seq of important histone marks in many cell states in human, worm, and fly [25,26], and used statistical modelling to annotate different chromatin states by the combinatorial patterns of different histone marks [27–29]. These studies help identify previously unannotated functional elements in the genomes.

A number of algorithms for ChIP-seq analysis have been developed. In this review, we will provide an overview of the analytical workflow for ChIP-seq and summarize the key concepts and challenges for each step. In particular, we will introduce useful quality control strategies at different analytical steps, which are useful for data interpretation. We will also discuss methods to integrate ChIP-seq with other types of high-throughput data for more comprehensive understanding of important physiological or pathological processes.

CONSIDERATIONS OF ChIP-seq EXPERIMENTAL DESIGN

The quality of ChIP-seq critically depends on the sensitivity and specificity of the antibody for a DNA-binding factor. Specific antibodies give strong and clean binding enrichment information, while weak and non-specific antibodies have increased background noise. Another important issue related to experimental design is the use of control experiments to adjust the bias caused by chromatin accessibility [30]. In most cases, DNA from chromatin input (i.e., chromatin sample before IP), mock immunoprecipitation (IP) (i.e., IP without antibody) or non-specific IP (e.g., IP against immunoglobulin G) has often been chosen as a control sample. For more detailed topics on ChIP-seq experimental design, the advantage and disadvantage of different types of control, we refer the readers to a recent review on ChIP-seq [31].

Advances in NGS technologies enable ChIP-seq to be conducted at greater genome coverage at lower price [3], and recover weaker binding events. Saturation of ChIP-seq depends on the nature of DNA-binding factor, antibody sensitivity, and specific research focus. It can be evaluated by sub-sampling total sequencing reads, and computing the recovery rate of ChIP-seq peaks [31]. In general, a deeper sequencing is recommended for factors with diffuse binding patterns or repressive functions than those with sharp binding patterns or active functions. It is also important to sequence the IP and control at comparable depth to allow unbiased peak calling. Along with the improvement of NGS technologies, the development of methods for multiplexing samples by barcoding (i.e., multiple samples are processed at a single lane) can increase efficiency without excessive increase in time or cost [32–35].

ChIP-seq DATA ANALYSIS WORKFLOW

A number of computational and statistical tools have been proposed and developed for addressing specific aspects of ChIP-seq analysis. We will describe them in the following subsections. In addition, we would like to highlight integrative analysis platforms such as Cistrome [36] and CisGenome [37,38], which provide comprehensive work environments for users to conduct most of the necessary analyses in one place. Independent algorithms and tools for specific purposes will be introduced in the context of the analytical workflow for ChIP-seq within each of the subsections. It is also important to develop quality control (QC) standards at each analytical step, so we have included suggestions on the potentially useful QC strategies.

Reads mapping

Raw data from NGS platform often appear in fastq format, containing short DNA sequence and quality scores. In general, the first step of ChIP-seq analysis starts with mapping these raw reads to the reference genome. Several algorithms have been developed to quickly map millions to hundreds of millions of short sequencing reads. The popular ones include ELAND (Illumina®), Bowtie [39,40], BWA [41,42], MAQ [43], Stampy [44], Novoalign [45], and SOAP2 [46]. Among them, Bowtie, BWA, and SOAP2 adopt the Burrows-Wheeler transformation, which was originally developed as a data compression technique in the 1990s [39,41]. Choosing appropriate mapping software depends on sequencing platform, speed requirement, and hardware resources. For example, if data come from the SOLiD platform, tools supporting colorspace are the ideal choices, such as BWA. If speed is the major consideration, Bowtie is preferred. For further reading, See Bao et al.'s review paper for more detailed comparisons of the above aligners [47].

Quality control (QC) can be conducted for reads mapping. In order to simplify analysis, usually only reads mapped to one unique location in the genome (called uniquely mapped reads) with minimum allowed mismatches (e.g., up to two mismatches) are kept for

downstream analysis. In case a ChIP-seq read is mapped to multiple locations on the genome, a general solution is to randomly assign one of the locations to it. Often, the ratio of the number of uniquely mapped reads over the total number of ChIP-seq reads can be an assessment of library quality. From statistics of publicly available ChIP-seq datasets, ratios of over 50% suggest good library quality. Another useful QC measure is the number of redundant reads that are mapped to the same genomic coordinates, because high redundancy rate suggests PCR amplification bias from limited ChIP material. The ratio of redundant reads over all mapped reads should ideally be below 50%.

Peak modeling and identification

The sequence reads mapped to the genome are subject to peak calling to detect regions with significant enrichment of ChIP signals with respect to the background (e.g., control if available). For most of ChIP-seq experiments with single-end sequencing, DNA fragments are sequenced from the 5' ends; as a result, bimodal distributions, surrounding the true binding site, are formed from reads mapped on the + and − strands respectively (red and blue curves in Figure 1A). Therefore, to precisely detect the correct binding site, some peak callers empirically model the distance between the + and − strand modes [48,49], and extend the tags towards their 3' direction by the estimated distance (Figure 1B). Then, the pile of the extended tags forms a peak (Figure 1B) and its summit represents the most probable binding location. One QC measure at this step is the ability of peak callers to properly model the +/− mode distance, and failing to model the distance suggests potential biases in the sonication and library construction steps, or that the factor of interest binds to diffuse regions rather than point sources.

Next, peak callers calculate the statistical significance (e.g., *p* value) of the enrichment level of ChIP signals in selected regions comparing to a background model. Due to the discrete nature of NGS data, most of the peak callers adopt the binomial [50], Poisson [48,51,52], negative binomial (almost equivalent to the Poisson based on local average) [37,53] distributions or simulation-based modelling [49,54] to compute statistical significance of the ChIP enrichment over background. A few widely-used peak callers include MACS [48], Sissr [55], SPP [49], and USeq [50]. In our opinion, different distributions are same in essence. For example, negative binomial is a generalized Poisson, and dynamic Poisson based on empirical local lambda is a more generalized version of negative binomial, etc. So they provide similar sensitivity and specificity. But different peak callers have their own rationality, and are different in positional accuracy of predicted binding sites. More detailed description and comparison of different peak callers are available in separate studies [31,56–58].

Detected peaks also need to be checked in terms of quality, where false discovery rate (FDR) or fold change against the background (e.g., control tag counts in the same region) is often used. FDR is defined as the expected proportion of false positive peaks in a list of detected peaks [59–61]. Several peak callers provide empirical [48,50,54,62–64] or model-estimated [37,38,51,52] FDR or *q*-value (minimum FDR at a given *p* value cut-off) for each peak, and 5% FDR is the most commonly accepted

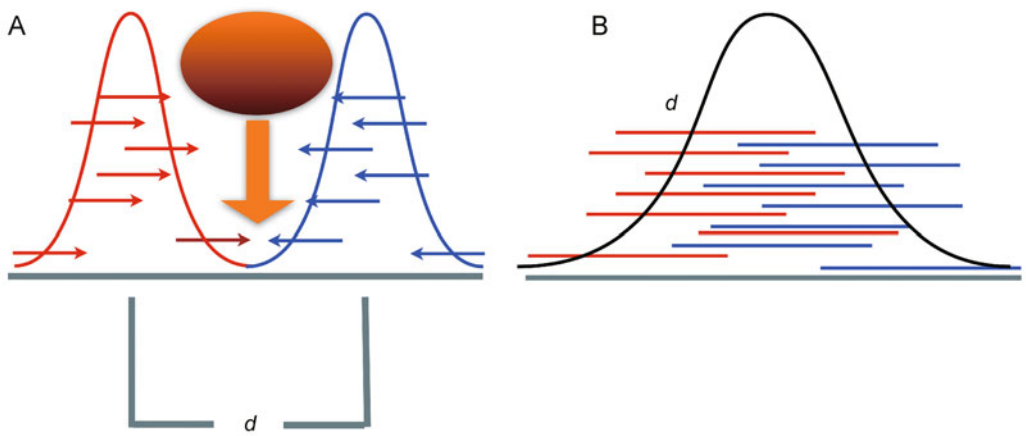

Figure 1. A schematic of peak modelling for ChIP-seq. (A) The bimodal distributions of + and − sequence reads (red and blue arrows, respectively) surrounding a transcription factor binding center (marked by the yellow vertical arrow). The distance (*d*) between the summits of the + and − distributions is considered to be an estimate of the length of DNA fragments pulled down by the antibody. (B) The ChIP enrichment signal can be obtained as the count of the + and − sequence reads extended by the estimated *d* at every base.

value for peaks of good quality. The empirical FDR can be calculated as the number of control peaks passing certain cut-off divided by the number of ChIP-seq peaks passing the same cut-off. Model-estimated FDR can be computed through permutation or random sampling. Fold change, the ratio of tag counts between IP and control in the peak region, is also an intuitive measure of peak quality. A fold change of 5 is generally recommended as a reasonable cut-off, and an enough number (e.g., >50%) of peaks with over 20-fold is an indicator of good ChIP-enrichment.

In addition to the above mentioned peak callers, there are also peak callers with more specialized functions, such as calling positioned nucleosomes from nucleosome-resolution histone mark ChIP-seq (i.e., MNase-digested fragments) [65], identifying diffuse regions enriched by a broad mark [51], combining multiple ChIP-chip or ChIP-seq sets for the same factor for consensus peak calling [66]. As more and more ChIP-seq data (or existing ChIP-chip data) become available, these peak callers with special functions will become more useful for data integration.

Visualization tools

Most people view their ChIP-seq data either as signal profiles or as called peaks on a genome browser. The most widely used visualization environment is the University of California Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu) [67] (Figure 2). In addition to the standard browser functions, this web-based application also provides other important genomic information, including tracks for gene annotation (e.g., refseq or UCSC known genes), evolutionary conservation, annotated SNPs [68,69], and data from NIH funded genomics consortia such as ENCODE [70].

As a web server, UCSC genome browser has limitations in response speed, which are mostly related to the process capability of the server and Internet connection. In contrast, stand alone genome browsers such as IGV (http://www.broadinstitute.org/software/igv/home) [71] and IGB (http://www.bioviz.org/igb/) [72] can efficiently display large data sets and enable the user to visualize genomic datasets from either a local computer or a remote date warehouse and navigate quickly at multiple scales. It also supports the simultaneous display of other types of genomics data tracks such as aligned NGS reads, mutations, copy numbers, gene expression, DNA methylation, and gene annotations as well [71]. Other NGS genome browsers are also available and we refer readers to their individual publications [37,38,73–79].

Use of replicates

Biological replicates can help identify more confident peaks for ChIP-seq experiments. Intuitively, successful ChIP-enrichments should be consistent across biological replicates. Replicates should have similar signal profiles

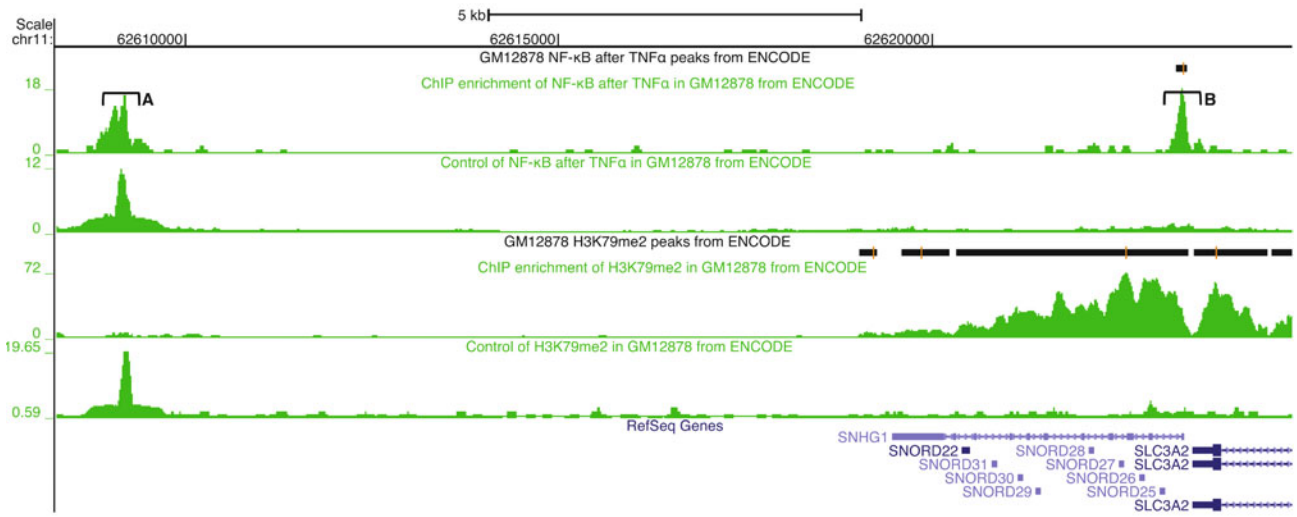

Figure 2. An UCSC genome browser snapshot of ChIP enrichments of NF-κB and H3K79me2 in human lymphoblastoid cell line GM12878 from ENCODE data. The bottom gene track shows RefSeq genes near the binding sites. While NF-κB shows sharp binding patterns (the 1st green track from the top), H3K79me2 diffuses over a broad region (the 3rd green track). The black horizontal bars represent peaks called by MACS. The control tracks of DNA inputs (the 2nd and 4th green tracks) are used to model the background including the chromatin bias in ChIP-seq. The significance of binding site detection is estimated considering the background from the control; therefore, region A (near chr11 62610000) of NF-κB is not identified as a peak because its background level is relatively high compared to region B (near chr11 62625000) although the ChIP enrichments of these regions are similar.

and peak regions. One measure of consistency is the percentage (e.g., > 50%) of overlapping peaks between two replicates, which can be easily visualized using a Venn diagram (Figure 3A). Another measure is the correlation coefficient (e.g., over 0.6 between replicates) of ChIP-seq signals over selected genomic intervals (e.g., every 1 kb or 2 kb) or over the union peak regions of the replicates (Figure 3B).

The ENCODE/modENCODE consortia proposed a third statistical measure of replicate consistency called Irreproducible Discovery Rate (IDR). IDR measures the proportion of inconsistent peaks over the consistent peaks across replicates at certain threshold [80], and also considers whether the peak ranks (based on *p* value or FDR) in the replicates are correlated or not. Thus, IDR informs not only the significance of individual peaks but also the consistency between two replicates. As the decreasing cost of NGS allows increasing number of ChIP-seq experiments to have replicates, IDR will be useful to provide more robust peak calls from the replicates. A typical threshold of IDR is 1% [80].

Use of other prior genomic information for quality assurance

Since the genomic era began in the late 1990s, increasing amount of knowledge, including evolutionary conservation, TF binding motifs, and gene annotations, has been accumulated in the public domain. In this section, we will discuss how to use such knowledge to assess the quality of ChIP-seq data.

Use of evolutionary conservation

C*is*-regulatory elements that harbour TF binding sites are in general under more evolutionary constraint. Therefore, most identified peaks in a successful ChIP-seq experiment show more evolutionary conservation than background sequences in the genome. PhastCons scores [81] (e.g., that downloaded from UCSC genome browser) provide precomputed evolutionary conservation score at every base in a reference genome. When aligning good transcription factor ChIP-seq peak at the peak summit or center, the average PhastCons conservation scores near the summit or center are typically higher than the surrounding regions (Figure 4A).

Use of TF DNA binding motifs

TFs bind to DNA in a sequence-specific manner, so peaks from successful TF ChIP-seq should exhibit significant enrichment of the TF binding motif. Binding motifs can be represented as a position weight matrix or virtualized as a sequence logo, both of which indicate the nucleotide preference of the factor at each motif position (Figure 4B). Currently, several databases of known TF binding motifs are publicly available [82–84]. In addition, many experimental and computational groups continue to

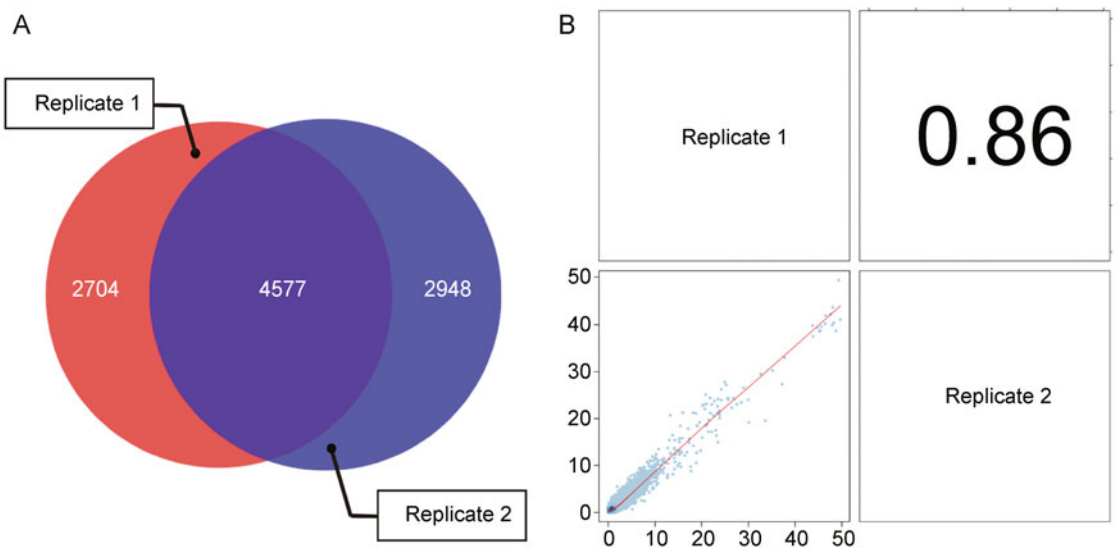

Figure 3. A sample analysis for two replicates of NF-κB ChIP-seq from ENCODE data. (A) The Venn diagram of the peak sets identified from the two individual replicates at the same cut-off for peak calling (7281 and 7525 peaks, respectively). A large intersection indicates high consistency between the replicates (e.g., > 50%). (B) Another consistency measure is to see the pair-wise correlation of ChIP enrichments of the replicates. Each dot represents the average ChIP enrichments of the replicates every 2 kb. Consistent replicates show a tight distribution around the local regression line (red), resulting in a large correlation coefficient (e.g., 0.86 in this example).

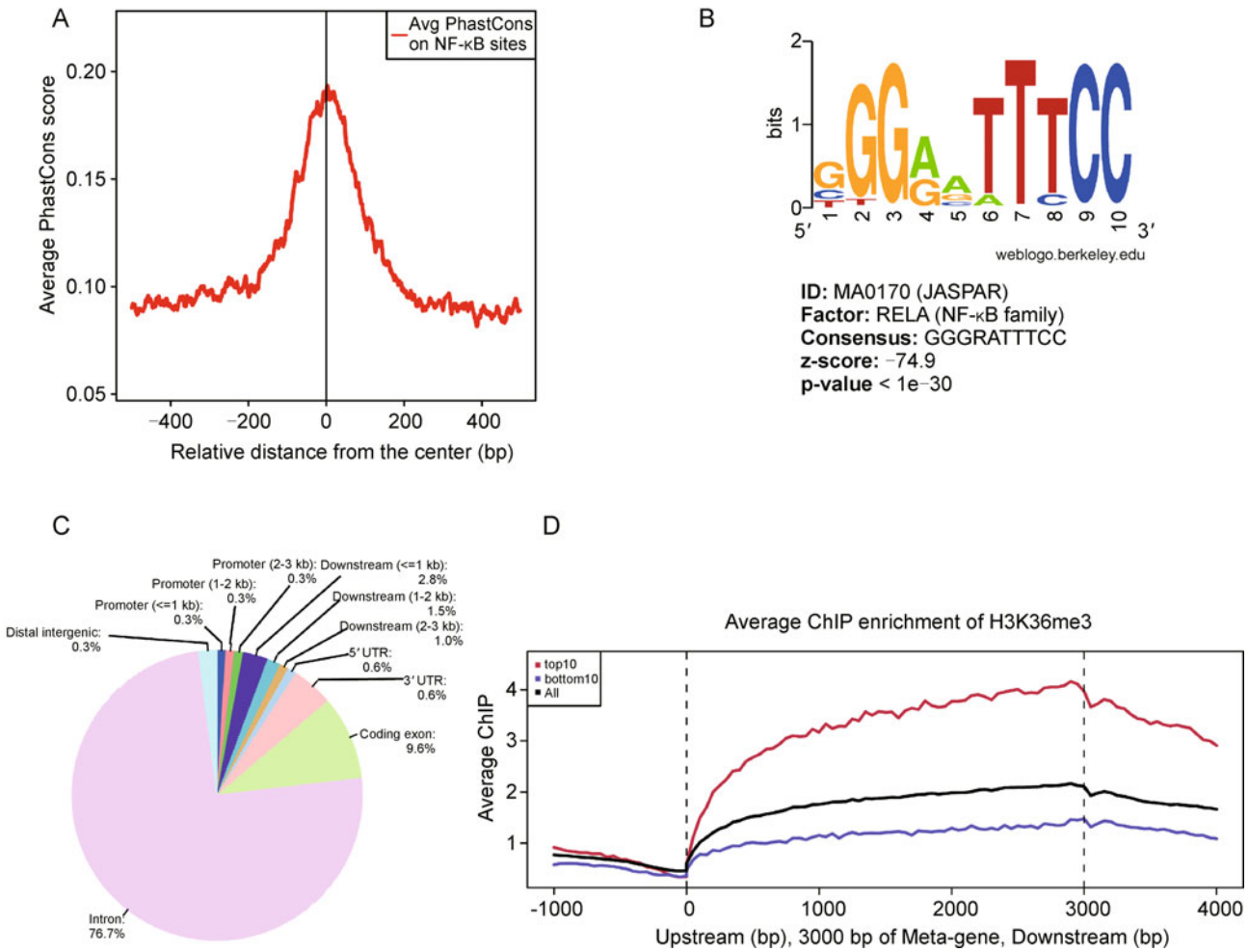

Figure 4. Useful QC measures for ChIP-seq peak calling. (A) The average PhastCons score at TF binding sites can be used to assess the quality of ChIP-seq peak calls. The plot indicates that the center of NF-κB binding sites is more evolutionarily conserved than the background. (B) If ChIP-seq is successful and the factor of interest has specific DNA binding motifs, the motifs should be significantly enriched near the summits of detected binding sites. This example shows that a NF-κB binding motif registered in JASPAR database was found at NF-κB binding sites. (C) The pie chart visualizes the distribution of H3K36me3 peaks over different categories of elements such as promoter, UTRs, coding exon, and intron. Since H3K36me3 is associated with transcriptional elongation, its peaks are primarily present in exons (i.e., 9.6% in the pie chart) and introns (i.e., 76.7%). (D) This trend was also observed using the meta-gene plot of H3K36me3 ChIP enrichment. Every gene was normalized to have the same length of 3 kb and then the average ChIP signal was profiled on the meta-gene including 1 kb upstream and downstream of TSS and TTS. The red and purple lines represent the average ChIP enrichments of H3K36me3 on highly expressed (top 10%, red) and lowly expressed (bottom 10%, purple) genes, respectively, which shows that H3K36me3 is positively correlated with gene expression levels.

derive better or previously uncharacterized TF motifs from new genomic technologies or data [85,86]. Enriched sequence motifs can be identified from either *de novo* methods or known motif scanning [87–90] at TF ChIP-seq peaks. In addition, a motif could be identified with better confidence if its occurrences are more frequently at the peak summits or centers than at the surrounding regions [91]. A common procedure for motif finding is to focus on top ranked ChIP-seq peaks (typically top 1000 ranked by *p* value) in order to avoid noises from weak binding sites.

Use of DNase I hypersensitive sites and HOT regions

Deoxyribonuclease (DNase) I hypersensitive sites are chromatin regions that are more susceptible to cleavage by this DNA cutting enzyme than other regions. DNase I hypersensitivity often indicates DNA accessibility associated with a local reduction in nucleosome occupancy [92–94]. DNase I hypersensitive sites are broadly enriched in bodies of highly expressed genes, and sharply enriched in functional regulatory sequences such as promoters and enhancers, which are targeted by many

TFs. Recently, the ENCODE and Roadmap Epigenomics consortia released the DNase-seq data of many human and mouse cell lines and tissues [25,95]. These data provide a comprehensive repertoire of the locations of TFs, chromatin factors, and histone marks [25,95]. Peaks of a successful ChIP-seq experiment overlap over 80% with the DNase-seq peaks, and this could be another evidence of good data quality.

In published ChIP-seq studies, some regions called High-Occupancy Targets (HOT) are constantly enriched in almost all ChIP experiments. They often lack the binding motif of the factor of interest, and might be caused by protein-protein interactions of unknown factors with the factor of interest or experimental artefacts during the ChIP process [96–98]. It is advisable to filter the HOT regions before downstream analysis, and high percentage of ChIP peaks in the HOT regions raises a red flag on data quality.

Use of the statistics of peak distribution

TF binding sites and modified histone regions are generally enriched for *cis*-regulatory elements. For example, many TFs bind more frequently in promoters than the rest of the genome (Figures 4C and 4D) and histone marks such as H3K36me3 are enriched over the exons of actively transcribed genes. Although these properties vary widely based on the factor or modification being tested, to test the consistency of a priori knowledge and observed enrichment patterns can be used to evaluate ChIP-seq data quality. Several software packages are available to provide summary statistics on the distribution of ChIP-seq peaks or visualize average ChIP-seq enrichment signals on annotated *cis*-regulatory elements [38,99].

Differential peak identification

Transcriptional and epigenetic regulation studies often need to identify differential binding of a factor between two or more biological conditions. For example, a recent work by Wang et al. showed androgen receptor (AR) binding changes when its co-factor FoxA1 was knocked-down in the prostate cell line LNCaP [100]. One simple method is to run peak calling in separate conditions followed by intersection analysis to identify unique peaks for each condition, but it might miscall a region when it is barely above and below the peak calling cut-off in the respective conditions. Another method is to conduct peak calling between the two ChIP-seq conditions treating one as the control [101], but it might erroneously detect a region that is weak in both conditions but significantly enriched in one condition over the other. A more specialized algorithm uses a Hidden Markov Model to detect differential binding through probabilistic modelling of the ChIP-seq profiles in two conditions [102], but the method has limited resolution. Next significant version of MACS: MACS2 (https://github.com/taoliu/MACS/) that attempts to address all these issues based on four paired treatment and control samples is currently available under continuous testing and improvement. There are some existing algorithms initially designed for differential expression studies on RNA-seq data, such as edgeR [103], DESeq [104], and bayseq [105], which can be modified and applied to ChIP-seq as well. As more ChIP-seq data over multiple conditions become available, the increasing importance of differential peak calling will accelerate the effort to develop better algorithms.

INTEGRATIVE ANALYSIS WITH OTHER GENOMIC DATA

Integrate with gene expression profiles

ChIP-seq data of TFs, chromatin factors, and histone marks often need to be interpreted in the context of gene regulation, which requires predicting the target genes that are regulated by the factor and its binding sites. The methods for determining the potential target genes of a factor generally depend on the factor type and binding patterns. For example, most TFs and some histone marks are characterized as sharp binding patterns on either proximal or distal regions to transcription start sites (TSSs). In such case, the distance between each binding site to its nearest gene tends to show a strong association with the expression level or dynamics in expression of the gene. Please keep in mind that, although this strategy works in many cases, it neglects the fact that there would be long-range interactions between *cis*-regulatory elements and their target genes. With more understanding on high-level chromatin structure, we would revise target gene prediction method extensively. On the other hand, for some other histone marks with pervasive enrichment over a region, the coverage as well as the enrichment level of the mark on the promoter, exons, or body of a gene are important variables for determining whether the gene is a target.

One important subject of integrative analysis of ChIP-seq with expression profiles is to infer whether a given factor mainly functions as a transcriptional activator or a repressor. In order to perform such analysis, expression profiling should be conducted where the activity of the factor is perturbed (e.g., through hormone activation or siRNA knockdown). If a histone mark is studied, gene expression can be profiled in the same way by perturbing the activity of the modifying enzyme that is associated with the histone mark. Once the target genes of binding are predicted, another useful analysis is to examine their

potential biological functions. Before experiments are conducted to validate the function of binding, computational analysis of gene annotation/ontology could provide important clues as to whether the target genes are enriched for specific biological processes or pathways.

Target gene prediction

1. Target gene prediction for factors with sharp binding patterns

The effect of TF binding on gene expression often attenuates with the distance to the gene, but could reach hundreds of kb from the target genes [7]. Therefore, distance-based target gene prediction is useful for most TFs and some histone marks with sharp peaks in promoters or enhancers (i.e., distal intergenic or intronic regions). This group of histone marks includes H3K4me1/2/3 (H3K4me3 more enriched at promoters, H3K4me1 more enriched at enhancers, and H3K4me2 enriched at both), H3R17me3, most acetylation marks, and the histone variant H2A.Z. Simple target prediction could be based on the nearest distance (e.g., within 1 kb for promoter binding and within 10 kb for enhancer binding) between factor binding sites and TSSs of genes. If expression profiling with factor activity perturbation is available, limiting genes with significant differential expression between the perturbation conditions could refine the targets. An alternative to using the nearest distance is to count the number of binding sites within a given distance range (e.g., 100 kb) since more binding sites in proximity are thought of as evidence of increased regulatory potential of the factor to the target gene [106]. This approach could be further refined by weighting the different binding sites by their distances to the target gene [107].

2. Target gene prediction for factors with pervasive binding patterns

Chromatin factor and histone marks related to transcriptional elongation or repression often show pervasive enrichment patterns over broad regions. For instance, H3K36me3 is broadly present in the exons of expressed genes while H3K27me3 or PRC2 complex tend to enrich at silenced genes [108]. It is informative to stratify genes according to their expression levels and see the relative enrichment of the mark in each of the strata. Figure 4D displays average ChIP signals of H3K36me3 on the meta-gene (by dividing every gene into the same number of bins) at different gene expression levels in human lymphoblastoid cell line GM12878 [23]. Since H3K36me3 is a transcriptional elongation mark, top 10% expressed genes have significantly higher H3K36me3 enrichment than bottom 10% expressed genes. By computing the ChIP-seq coverage over gene body or exons (normalized by the gene or exon length) and selecting genes with the top (e.g., 30%) coverage, one can predict the H3K36me3 associated genes.

Factor function as transcriptional activator or repressor

The role of a factor as a transcriptional activator or repressor can be inferred by the expression changes of target genes when the activity of the factor is perturbed by activation, overexpression, knockdown, or knockout. For example, if genes with reduced expression in the knockdown or knockout of the factor are more likely to harbor the binding sites of the factor in proximity than genes with increased expression, this will be considered as important evidence that the factor acts as a transcriptional activator.

This concept is illustrated in Figure 5 [109,110]. Using distance of binding to all the genes as background, we could see that in the prostate cancer cell line LNCaP, androgen receptor binds much closer to the up-regulated genes than down-regulated and background genes, which implies that it primarily serves as a transcriptional activator. In contrast, in the breast cancer cell line MCF-7, estrogen receptor appears to be equally close to both down- and up-regulated genes. The statistical significance of these observations can be evaluated using a non-parametric test such as the one-side Kolmogorov-Smirnov test.

Target gene annotation based on gene ontology and biological pathways

Researchers are often interested in seeing whether the target genes of a factor binding are enriched for specific pathways or biological processes. Many gene ontology or annotation analysis tools are publicly available, which are summarized at the Gene Ontology website (http://www.geneontology.org/GO.tools.shtml). A few particularly user-friendly tools that are not in the list include DAVID [111,112] and Panther [113], which take gene list as input, and GREAT [114], which takes binding site coordinates as input. GREAT assigns target genes based on gene-binding distance which might have some disadvantages, but it can detect function or pathway enrichment with better sensitivity because it could have a richer meta-annotation gene list [114]. Gene set enrichment analysis (GSEA) also conducts ontological analysis on the target genes, and its unique gene sets provides more extensive annotation searches [115].

Integrate with other TF ChIP-seq data

TF ChIP-seq experiments often reveal novel associations

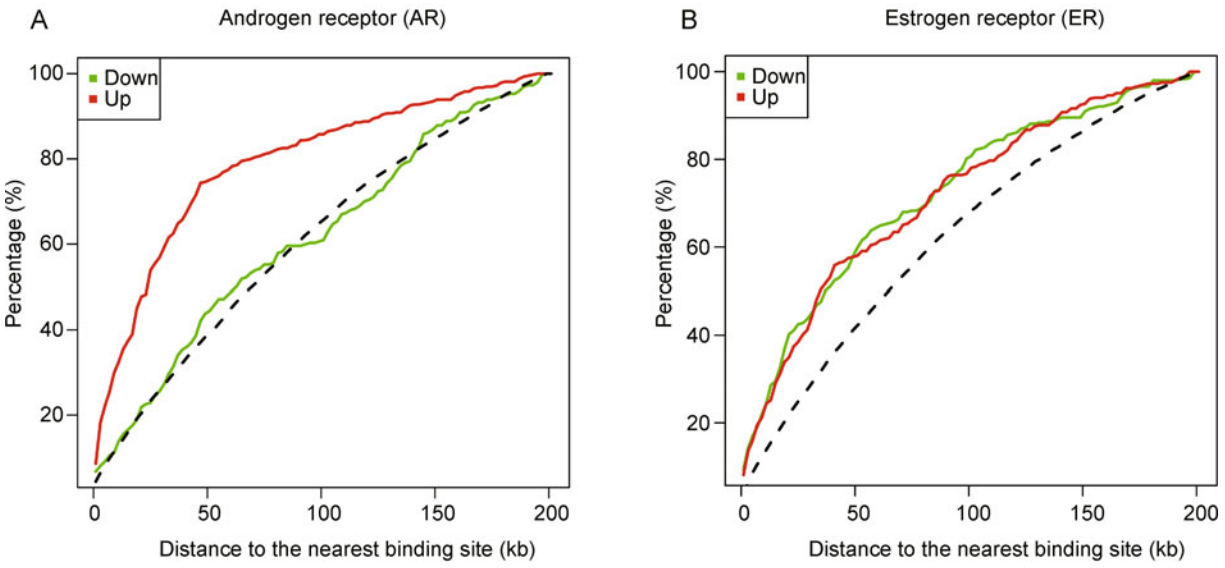

Figure 5. A distance-based method for inferring the role of a sharply binding TF as a transcriptional activator or repressor. (A) Each curve is the cumulative distribution of the distances from genes to the nearest AR binding sites in prostate cancer cell line LNCaP. The red and green colors correspond to the sets of genes that are up- and down-regulated by DHT treatment, respectively. The black dotted line indicates the distance distribution of all genes, which can be used as a background distribution. From this analysis, it can be seen that AR more directly regulates the up-regulated genes than the down-regulated ones. (B) A similar analysis was done with breast cancer cell line MCF-7. When it compared to AR (A), ER regulates the up- and down-regulated genes by E2 treatment almost equally.

between multiple TFs. Motif discovery has been previously suggested as an important ChIP-seq QC measure, and it can also be used to predict collaborating factors [116]. Crosslinking stabilizes not only the covalent link between TFs and DNA, but also that between interacting TFs, which allows ChIP of one factor to enrich the targets of the collaborating factors. If motif discovery on the peaks of one factor finds not only the known motif of its own, but also a significantly enriched motif of another factor, it suggests that the two factors may interact. Sometimes many factors with the same DNA-binding domain share similar motifs, so to pinpoint the exact factor binding to the collaborating motif requires analysis of expression data. If a factor is highly expressed in the cell and correlated in expression with another factor being ChIPed cross a panel of related tissue samples [117], they may be putative collaborating partners. Protein-protein interaction experiments and prior literature might provide additional insights.

The candidate co-factors selected by the aforementioned methods are often further analyzed by ChIP-seq. The Venn diagram (or Euler diagram in case of more than two factors) is also used to see the potential association or colocalization between different DNA binding factors when the ChIP-seq data of these factors are all available. In many cases, the regions co-occupied by collaborative factors (i.e., the intersection in the Venn diagram) are considered to be more important for understanding the detailed gene regulation event governed by these factors together. There are publicly available software to call such colocalized regions from the peak lists of multiple factors in UCSC BED format [36,118,119]. In addition, heatmaps are very useful to display ChIP signals of different factors across their union of binding sites (e.g., from –1 kb to 1 kb from the binding summit or center). Coupled with clustering methods (e.g., hierarchical or *k*-means clustering), heatmap analysis can provide better binding site classification than the Venn diagram by grouping binding sites with similar ChIP enrichment patterns across multiple factors (see Figure 6 for more details [120]).

Integrate ChIP-seq data from multiple organisms

Many biomedical experiments are conducted on model organisms, so it is interesting to investigate whether mechanisms discovered for a factor in model organisms such as round worm or fruit fly still hold true in more complex organisms such as human. As for ChIP-seq, this question can be answered by comparing ChIP-seq profiles of orthologous factor in multiple species. Several studies have found that while the binding sites of the factors diverge extensively between species, the target genes of highly conserved TFs are mostly preserved across species [121,122]. However, some recent studies examining ChIP-seq reads mapping to repetitive regions of the genome also found significant rewiring between factor

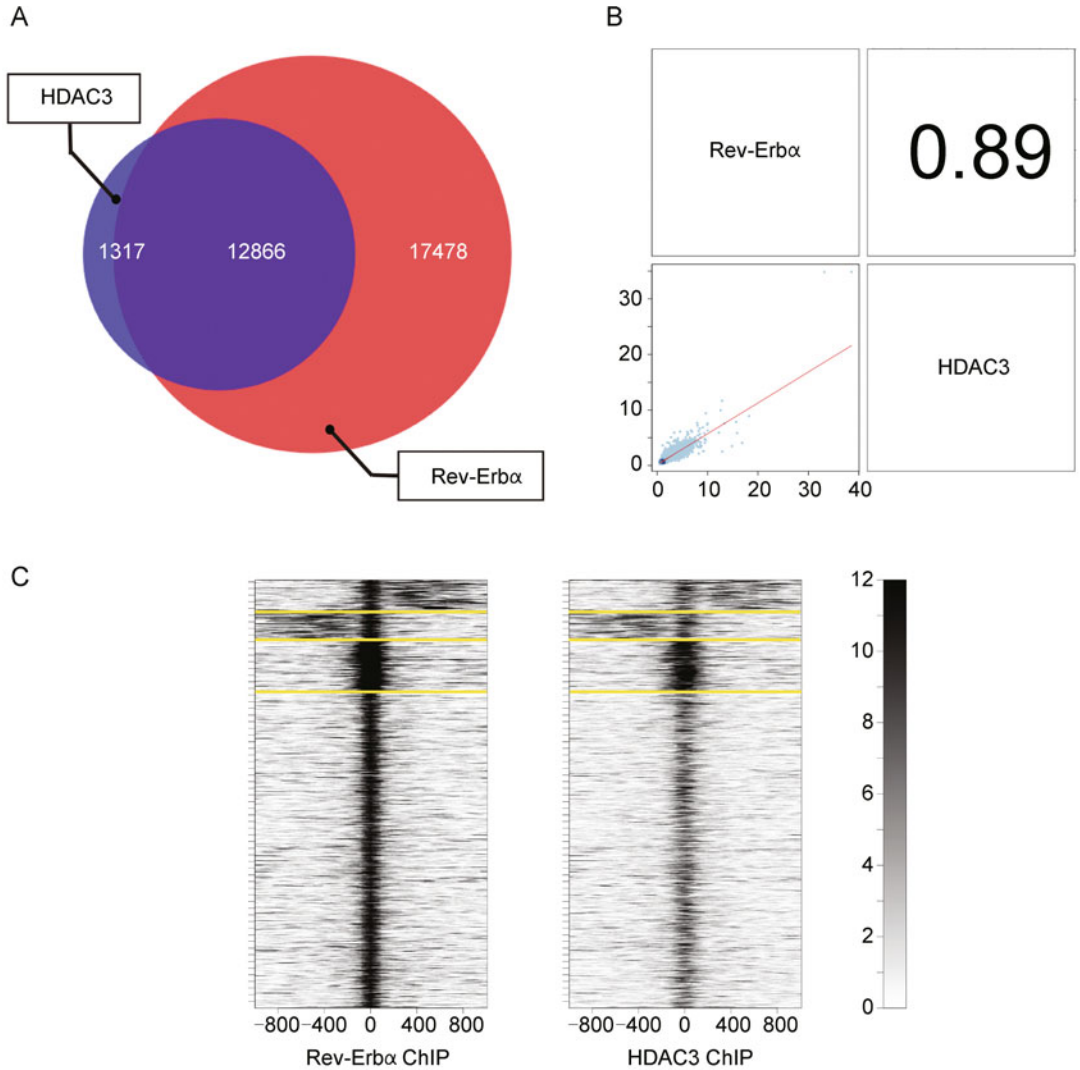

Figure 6. An example analysis for the ChIP-seq data of two collaborating factors. (A) The Venn diagram of Rev-erbα and HDAC3 (histone deacetylase 3) binding in mouse liver. This shows that Rev-erbα and HDAC3 are highly colocalized. However, this example is an extreme case for highly collaborating factors. In general, collaborating factor ChIP-Seq peaks tend to show a smaller intersection than that between two replicates of the same factor. (B) The scatter plot of the ChIP-seq enrichments of the factors at their union binding sites (blue dots) and a regression line (red line). The correlation coefficient of these two factors was also calculated (0.89). (C) Heatmap analysis for Rev-erbα and HDAC3 binding sites. The heatmaps also confirm the high association of these two factors (left and right). The binding sites with similar binding patterns are grouped using *k*-means clustering ($k = 4$) and distinguished by yellow horizontal lines. The 1st and 2nd groups reflect strand bias in binding. The individual groups can be further analyzed by associating with differentially expressed genes.

and target gene over evolution through transposable element turnover [123,124]. To accurately detect enrichment in repeat regions, more efforts are needed to develop algorithms that can better utilize multiply mapped sequence reads.

Integrate with epigenetic ChIP-seq data

Since ChIP-seq of histone methylations [23] and acetylations [24] were first profiled in human $CD4^+$ T-cells, increasing efforts to study histone marks in more cell states and organisms have yielded better understanding of epigenetic regulation. These include the mod/ENCODE and Roadmap Epigenomics consortia and individual studies [23,24,27,125]. For example, it has been revealed that many histone modifications delineate important elements such as active or repressed promoters (e.g., H3K4me3 or H3K27me3, respectively [126]), actively transcribed exons (e.g., H3K36me3 [127]), and active enhancers (e.g., H3K27ac [128] and H3K4me1 [129]).

Also, histone marks such as H4K20me3 and H3K9me3 indicate transcriptionally silenced state over very broad heterochromatin domains. In addition, although different histone marks have different antibody specificities, many, especially histone acetylations, have similar enrichment characteristics and probably redundant regulatory mechanisms [24]. Several groups have attempted to use machine-learning algorithms such as hidden Markov models to infer combinatorial enrichment patterns of histone marks and other chromatin factors [27–29,130,131]. These combinatorial patterns also reduce the redundancy in histone mark profiles. They can be used to segment the whole genome into regions of distinct chromatin signatures, and predict previously unidentified functional elements in the genome.

An interesting follow up is to identify enriched transcription factor motifs on the putative enhancers that are inferred from histone mark profiles. If conducted on different cell types and integrated with gene expression analysis, this could yield important insights to cell-type specific transcription factor activities [29]. Putative transcription factor binding sites can carry enhancer histone marks before active binding of transcription factor. This histone mark pattern often changes upon binding, which is characterized by a displacement of the nucleosome at the binding site and stronger marking and better positioning of the nucleosomes flanking the site. Therefore, motif analysis conducted on the dynamics of histone mark profiles at single-nucleosome resolution can potentially improve the prediction accuracy of transcription factor binding [91,132]. This approach has drawn attention recently as an effective pre-screening measure to find key transcription factors responding to developmental or environmental stimulations [106]. Likewise, in another study, the chromatin state dynamics based on histone mark profiles were used to link enhancers with their target promoters based on correlation estimation. Then, predictions were made for cell-type specific active or repressive TFs through the integration of motif analysis results and gene expression profiles [29].

Integrate with genome variation and disease data

DNA sequence variations such as single nucleotide polymorphisms (SNP) on transcription factor binding sites may influence the binding affinity of the factors. Several studies have attempted to assess the potential impact of allele-specific SNPs or indels on chromatin structure or TF binding sites [133–136]. Interestingly, the majority of disease-associated loci discovered from genome-wide association studies (GWAS) are located on introns or distal intergenetic regions. ChIP-seq data of transcription factors or histone marks can provide hints on the mechanisms of these disease SNPs. For example, H3K4me2 ChIP-seq helped identify a tissue-specific enhancer in the cancer risk loci on the human 8q24 region that might regulate Myc expression [135].

Copy number variation (CNV) studies can also be linked with ChIP-seq analysis [137]. For example, while studying the enrichment of TF binding in amplified genomic regions can reveal novel pathological pathways originating from the CNVs, they might also yield false positive peaks calls [137]. Therefore, caution should be taken to reduce the false positive ChIP-seq peak calls by using CNV information from other sources or proper chromatin input controls.

CONCLUSION

The recent improvements in NGS throughput have dramatically enhanced the dynamic range of ChIP-seq. Illumina® HiSeq™ 2000, for example, can generate up to 200 Gb of sequence data per run. In addition, technological developments in automation, batch processing, and multiplexing also improve data production efficiency and lower cost. We expect increasing number of laboratories to adopt ChIP-seq for gene regulation studies, and each study to generate ChIP-seq data in multiple factors in more physiological or pathological conditions.

Most published ChIP-seq studies in vertebrate species are conducted on cell lines, since tissue or tumour samples often have heterogeneity and insufficient cell count for ChIP-seq (10^6 cells recommended). While some groups isolate relatively homogenous cell population from tissue for ChIP-seq [23,24,138,139], others try to improve the ChIP-seq protocol to work on smaller starting material [140,141]. A recent study proposed a method for single-tube linear DNA amplification that can avoid possible artefacts and bias during the amplification process, so ChIP-seq can be conducted on a few thousand cells [142]. Third generation single-molecule sequencing technologies might be an alternative solution to small sample experiments that we look forward to.

In addition to ChIP-seq, other NGS genomic assays that profile genome-wide chromatin states have also been useful for understanding gene regulation. A comprehensive and unambiguous set of genomic binding sites for a transcription factor can also be revealed by ChIP-exo [143]. DNA methylation profiles from bisulfite sequencing (BS-Seq) can identify both 5-methylcytosines and 5-hydroxymethylcytosinesat base-pair resolution [144–147]. DNase-Seq is a cost effective method for profiling *cis*-regulatory elements in open chromatin regions. In addition, Hi-C [19] or ChIA-PET [16–18] can help infer complex long-range three-dimensional chromatin interactions, for example between promoters and enhancers. Each of the above techniques helps provide additional salient view of the chromatin, and they could be

integrated with ChIP-seq to more clearly understand gene regulation.

For integrative studies, it is essential to establish computational pipelines that perform meta-analysis of multiple datasets using a variety of algorithms and tools. It is also crucial to integrate researchers' unpublished data with relevant publicly available genomic data to effectively refine biological hypotheses. There are already public resources such as the UCSC genome browser that provide processed and curated data for such integrative analysis [70].

As more high-throughput assays on gene expression and DNA binding activities become available, there will be increasing need for systems biology approaches to infer the associative or causal relationships between genes or proteins. Most early systems biology efforts focused on inferring network structures of gene regulation based on co-expression patterns from microarray data [148–152]. From a systems biology point of view, TF binding and epigenomic data can provide additional information on the linkage and directionality between nodes of transcription factors, chromatin factors, and other genes [153–155].

NGS technologies have improved our ability to detect genetic variations in non-coding regions. Also, increasingly improved statistical methods have been proposed to distinguish true variations from sequencing artifacts or errors [156]. As we previously discussed, DNA sequence variations could affect TF binding affinities and nearby gene expression, so ChIP-seq can help to explore the function of GWAS-identified disease loci in non-coding sequences. In addition, the coming years might also see epigenome-wide association studies conducted on tissues of populations to better understand diseases susceptibility or mechanisms.

In conclusion, since its invention, ChIP-seq has become a powerful tool for revealing transcriptional and epigenetic regulation in many cell systems. It still continues to evolve through the development of more advanced sequencing and sample production technologies. For ChIP-seq analysis, many computational and statistical applications have been developed and are now being organized into more comprehensive analytical pipelines. Finally, increasing efforts to integrate ChIP-seq with other types of high-throughput genomic assays will offer a more comprehensive perspective on complex regulatory mechanisms controlling a variety of physiological and pathological processes.

Acknowledgments

This work was supported by the National Basic Research (973) Program of China (No. 2010CB944904), National Natural Science Foundation of China (No. 31028011), and National Institutes of Health (No. HG4069).

REFERENCES

1. Metzker, M. L. (2010) Sequencing technologies — the next generation. Nat. Rev. Genet., 11, 31–46.
2. Ansorge, W. J. (2009) Next-generation DNA sequencing techniques. New Biotechnol., 25, 195–203.
3. Kircher, M., Heyn, P. and Kelso, J. (2011) Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics, 12, 382.
4. Schuster, S. C. (2008) Next-generation sequencing transforms today's biology. Nat. Methods, 5, 16–18.
5. Solomon, M. J., Larsen, P. L. and Varshavsky, A. (1988) Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell, 53, 937–947.
6. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. and Carroll, J. S. (2011) FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat. Genet., 43, 27–33.
7. Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang,Y., Li, W., Carroll, J. S., Liu, X. S. and Brown, M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell, 132, 958–970.
8. Young, R. A. (2011) Control of the embryonic stem cell state. Cell, 144, 940–954.
9. Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., et al. (2010) Mediator and cohesin connect gene expression and chromatin architecture. Nature, 467, 430–435.
10. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell, 133, 1106–1117.
11. Kim, J., Chu, J., Shen, X., Wang, J. and Orkin, S. H. (2008) An extended transcriptional network for pluripotency of embryonic stem cells. Cell, 132, 1049–1061.
12. Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A. and Young, R. A. (2010) c-Myc regulates transcriptional pause release. Cell, 141, 432–445.
13. Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W., Ye, C., Ping, J. L., Mulawadi, F., et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet., 43, 630–638.
14. Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res., 16, 1299–1309.
15. Espinoza, C. A. and Ren, B. (2011) Mapping higher order structure of chromatin domains. Nat. Genet., 43, 615–616.
16. Fullwood, M. J., Han, Y., Wei, C. L., Ruan, X. and Ruan, Y. (2010) Chromatin interaction analysis using paired-end tag sequencing. Curr. Protoc. Mol. Biol., Chapter 21, Unit 21.15.1–25.
17. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462, 58–64.
18. Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., Ariyaratne, P. N., Mohamed, Y. B., Ooi, H. S., Tennakoon, C., et al. (2010) ChIA-PET tool for comprehensive chromatin interaction

analysis with paired-end tag sequencing. Genome Biol., 11, R22.

19. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289–293.
20. Rusk, N. (2009) When ChIA PETs meet Hi-C. Nat. Methods, 6, 863.
21. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J. A., Umlauf, D., Dimitrova, D. S., et al. (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat. Genet., 42, 53–61.
22. Theodorou, V. and Carroll, J. S. (2010) Estrogen receptor action in three dimensions — looping the loop. Breast Cancer Res., 12, 303.
23. 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.
24. 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.
25. The ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol., 9, e1001046.
26. The modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., et al. (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science, 330, 1787–1797.
27. Liu, T., Rechtsteiner, A., Egelhofer, T. A., Vielle, A., Latorre, I., Cheung, M. S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P., et al. (2011) Broad chromosomal domains of histone modification patterns in *C. elegans*. Genome Res., 21, 227–236.
28. Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., et al. (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. Nature, 471, 480–485.
29. 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.
30. Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrançois, P., Struhl, K., Gerstein, M. and Snyder, M. (2009) Mapping accessible chromatin regions using Sono-Seq. Proc. Natl. Acad. Sci. USA, 106, 14926–14931.
31. Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680.
32. de Magalhães, J. P., Finch, C. E. and Janssens, G. (2010) Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. Ageing Res. Rev., 9, 315–323.
33. Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. and Knight, R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods, 5, 235–237.
34. Kim, J. B., Porreca, G. J., Song, L., Greenway, S. C., Gorham, J. M., Church, G. M., Seidman, C. E. and Seidman, J. G. (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science, 316, 1481–1484.
35. Meyer, M. and Kircher M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc., 2010, pdb.prot5448.
36. Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., et al. (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol., 12, R83.
37. Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M. and Wong, W. H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat. Biotechnol., 26, 1293–1300.
38. Ji, H., Jiang, H., Ma, W. and Wong, W. H. (2011) Using CisGenome to analyze ChIP-chip and ChIP-seq data. Curr. Protoc. Bioinformatics, Chapter 2, Unit2.13.
39. 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.
40. Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359.
41. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760.
42. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589–595.
43. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18, 1851–1858.
44. Lunter, G. and Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res., 21, 936–939.
45. Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S. and Robinson, P. N. (2010) Microindel detection in short-read sequence data. Bioinformatics, 26, 722–729.
46. Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25, 1966–1967.
47. Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X. and Song, Y. Q. (2011) Evaluation of next-generation sequencing software in mapping and assembly. J. Hum. Genet., 56, 406–414.
48. 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.
49. Kharchenko, P. V., Tolstorukov, M. Y. and Park, P. J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol., 26, 1351–1359.
50. Nix, D. A., Courdy, S. J. and Boucher, K. M. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics, 9, 523.
51. 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.
52. Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M. B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat. Biotechnol., 27, 66–75.
53. Ji, H. (2010) Computational analysis of ChIP-seq data. Methods Mol. Biol., 674, 143–159.
54. Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M. and Jones, S. J. (2008) FindPeaks 3.1: a tool for identifying areas of

enrichment from massively parallel short-read sequencing technology. Bioinformatics, 24, 1729–1730.

55. Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. (2008) Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res., 36, 5221–5231.
56. Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods, 8, 469–477.
57. Pepke, S., Wold, B. and Mortazavi, A. (2009) Computation for ChIP-seq and RNA-seq studies. Nat. Methods, 6, S22–S32.
58. Wilbanks, E. G. and Facciotti, M. T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS ONE, 5, e11471.
59. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, 57, 289–300.
60. Storey, J. D. (2002) A direct approach to false discovery rates. J. R. Stat. Soc. B, 64, 479–498.
61. Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA, 100, 9440–9445.
62. Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods, 5, 829–834.
63. Tuteja, G., White, P., Schug, J. and Kaestner, K. H. (2009) Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Res., 37, e113.
64. Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. Science, 316, 1497–1502.
65. Zhang, Y., Shin, H., Song, J. S., Lei, Y. and Liu, X. S. (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. BMC Genomics, 9, 537.
66. Chen, Y., Meyer, C. A., Liu, T., Li, W., Liu, J. S. and Liu, X. S. (2011) MM-ChIP enables integrative analysis of cross-platform and between-laboratory ChIP-chip or ChIP-seq data. Genome Biol., 12, R11.
67. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. Genome Res., 12, 996–1006.
68. Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res., 39, D876–D882.
69. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. and Kent, W. J. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res., 32, D493–D496.
70. Raney, B. J., Cline, M. S., Rosenbloom, K. R., Dreszer, T. R., Learned, K., Barber, G. P., Meyer, L. R., Sloan, C. A., Malladi, V. S., Roskin, K. M., et al. (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res., 39, D871–D875.
71. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. (2011) Integrative genomics viewer. Nat. Biotechnol., 29, 24–26.
72. Nicol, J. W., Helt, G. A., Blanchard, S. G. Jr, Raja, A. and Loraine, A. E. (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics, 25, 2730–2731.
73. Donlin, M. J. (2009) Using the Generic Genome Browser (GBrowse). Curr. Protoc. Bioinformatics, Chapter 9, Unit 9.9.
74. Podicheti, R. and Dong, Q. (2011) Administering GBrowse sites with WebGBrowse. Curr. Protoc. Bioinformatics, Chapter 9, Unit 9.14.
75. Huang, W. and Marth, G. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res., 18, 1538–1543.
76. Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. (2010) Tablet — next generation sequence assembly visualization. Bioinformatics, 26, 401–402.
77. Nicol, J. W., Helt, G. A., Blanchard, S. G. Jr, Raja, A. and Loraine, A. E. (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics, 25, 2730–2731.
78. Bao, H., Guo, H., Wang, J., Zhou, R., Lu, X. and Shi, S. (2009) MapView: visualization of short reads alignment on a desktop computer. Bioinformatics, 25, 1554–1555.
79. Lewis, S. E., Searle, S. M., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M. A., et al.. (2002) Apollo: a sequence annotation editor. Genome Biol., 3, RESEARCH0082.
80. 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.
81. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15, 1034–1050.
82. Robasky, K. and Bulyk, M. L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res., 39, D124–D128.
83. Xie, Z., Hu, S., Blackshaw, S., Zhu, H. and Qian, J. (2010) hPDI: a database of experimental human protein-DNA interactions. Bioinformatics, 26, 287–289.
84. Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B. and Sandelin, A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res., 36, D102–D106.
85. AlQuraishi, M. and McAdams, H. H. (2011) Direct inference of protein-DNA interactions using compressed sensing methods. Proc. Natl. Acad. Sci. USA, 108, 14819–14824.
86. Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G. P. and Burge, C. B. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol., 29, 659–664.
87. Bailey, T. L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics, 27, 1653–1659.
88. Machanick, P. and Bailey T. L. (2011) MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics, 27, 1696–1697.
89. Liu, X. S., Brutlag, D. L. and Liu, J. S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol., 20, 835–839.
90. Ma, X., Kulkarni, A., Zhang, Z., Xuan, Z., Serfling, R. and Zhang, M. Q. (2012) A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. Nucleic Acids Res., 40, e50.
91. Meyer, C. A., He, H. H., Brown, M. and Liu, X. S. (2011) BINOCh:

binding inference from nucleosome occupancy changes. Bioinformatics, 27, 1867–1868.

92. Bell, O., Tiwari, V. K., Thomä, N. H. and Schübeler, D. (2011) Determinants and dynamics of genome accessibility. Nat. Rev. Genet., 12, 554–564.
93. Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E. D., Wolfsberg, T. G., et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. USA, 101, 992–997.
94. Sabo, P. J., Humbert, R., Hawrylycz, M., Wallace, J. C., Dorschner, M. O., McArthur, M. and Stamatoyannopoulos, J. A. (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc. Natl. Acad. Sci. USA, 101, 4537–4542.
95. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 28, 1045–1048.
96. Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science, 330, 1775–1787.
97. Moorman, C., Sun, L. V., Wang, J., de Wit, E., Talhout, W., Ward, L. D., Greil, F., Lu, X. J., White, K. P., Bussemaker, H. J., et al. (2006) Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA, 103, 12027–12032.
98. Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., et al. (2011) A *cis*-regulatory map of the *Drosophila* genome. Nature, 471, 527–531.
99. Shin, H., Liu, T., Manrai, A. K. and Liu, X. S. (2009) CEAS: *cis*-regulatory element annotation system. Bioinformatics, 25, 2605–2606.
100. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., et al. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature, 474, 390–394.
101. Cheung, I., Shulha, H. P., Jiang, Y., Matevossian, A., Wang, J., Weng, Z. and Akbarian, S. (2010) Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. Proc. Natl. Acad. Sci. USA, 107, 8824–8829.
102. 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.
103. 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.
104. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106.
105. Hardcastle, T. J. and Kelly, K. A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 11, 422.
106. Verzi, M. P., Shin, H., He, H. H., Sulahian, R., Meyer, C. A., Montgomery, R. K., Fleet, J. C., Brown, M., Liu, X. S. and Shivdasani, R. A. (2010) Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. Dev. Cell, 19, 713–726.
107. Tang, Q., Chen, Y., Meyer, C., Geistlinger, T., Lupien, M., Wang, Q., Liu, T., Zhang, Y., Brown, M. and Liu, X. S. (2011) A comprehensive view of nuclear receptor cancer cistromes. Cancer Res., 71, 6940–6947.
108. The ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447, 799–816.
109. Wang, Q., Li, W., Zhang, Y., Yuan, X., Xu, K., Yu, J., Chen, Z., Beroukhim, R., Wang, H., Lupien, M., et al. (2009) Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell, 138, 245–256.
110. Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., Brodsky, A. S., Keeton, E. K., Fertuck, K. C., Hall, G. F., et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nat. Genet., 38, 1289–1297.
111. Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res., 37, 1–13.
112. Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc., 4, 44–57.
113. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res., 13, 2129–2141.
114. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. and Bejerano, G. (2010) GREAT improves functional interpretation of *cis*-regulatory regions. Nat. Biotechnol., 28, 495–501.
115. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA, 102, 15545–15550.
116. Zhang, Z., Chang, C. W., Goh, W. L., Sung, W. K. and Cheung, E. (2011) CENTDIST: discovery of co-associated factors by motif distribution. Nucleic Acids Res., 39, W391–W399.
117. Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V., Geistlinger, T. R., et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell, 122, 33–43.
118. Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res., 15, 1451–1455.
119. Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842.
120. Feng, D., Liu, T., Sun, Z., Bugge, A., Mullican, S. E., Alenghat, T., Liu, X. S. and Lazar, M. A. (2011) A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism. Science, 331, 1315–1319.
121. Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K. and Fraenkel, E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat. Genet., 39, 730–732.

122. Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science, 328, 1036–1040.
123. Chung, D., Kuan, P. F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E. H., Dewey, C. and Keleş, S. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLOS Comput. Biol., 7, e1002111.
124. Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K. and Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc. Natl. Acad. Sci. USA, 104, 18613–18618.
125. Eaton, M. L., Prinz, J. A., MacAlpine, H. K., Tretyakov, G., Kharchenko, P. V. and MacAlpine, D. M. (2011) Chromatin signatures of the *Drosophila* replication program. Genome Res., 21, 164–174.
126. Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell, 125, 315–326.
127. Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S. and Ahringer, J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet., 41, 376–381.
128. Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA, 107, 21931–21936.
129. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet., 39, 311–318.
130. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol., 28, 817–825.
131. Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A. and Noble, W. S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods, 9, 473–476.
132. He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., et al. (2010) Nucleosome dynamics define transcriptional enhancers. Nat. Genet., 42, 343–347.
133. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., et al. (2010) Variation in transcription factor binding among humans. Science, 328, 232–235.
134. McDaniell, R., Lee, B. K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., Scott, L. J., Morken, M. A., Kucera, K. S., Battenhouse, A., et al. (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. Science, 328, 235–239.
135. Ahmadiyeh, N., Pomerantz, M. M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., He, H. H., Brown, M., Liu, X. S., Davis, M., et al. (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc. Natl. Acad. Sci. USA, 107, 9742–9746.
136. Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E. and Iyer, V. R. (2010) Allele-specific and heritable chromatin signatures in humans. Hum. Mol. Genet., 19, R204–R209.
137. Pickrell, J. K., Gaffney, D. J., Gilad, Y. and Pritchard, J. K. (2011) False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics, 27, 2144–2146.
138. Verzi, M. P., Shin, H., Ho, L. L., Liu, X. S. and Shivdasani, R. A. (2011) Essential and redundant functions of caudal family proteins in activating adult intestinal genes. Mol. Cell. Biol., 31, 2026–2039.
139. Iyengar, S., Ivanov, A. V., Jin, V. X., Rauscher, F. J. 3rd and Farnham, P. J. (2011) Functional analysis of KAP1 genomic recruitment. Mol. Cell. Biol., 31, 1833–1847.
140. O'Geen, H., Echipare, L. and Farnham, P. J. (2011) Using ChIP-seq technology to generate high-resolution profiles of histone modifications. Methods Mol. Biol., 791, 265–286.
141. Adli, M., Zhu, J. and Bernstein, B. E. (2010) Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat. Methods, 7, 615–618.
142. Shankaranarayanan, P., Mendoza-Parra, M. A., Walia, M., Wang, L., Li, N., Trindade, L. M. and Gronemeyer, H. (2011) Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. Nat. Methods, 8, 565–567.
143. Rhee, H. S. and Pugh, B. F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell, 147, 1408–1419.
144. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. and Jacobsen, S. E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature, 452, 215–219.
145. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell, 133, 523–536.
146. Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462, 315–322.
147. Xiang, H., Zhu, J., Chen, Q., Dai, F., Li, X., Li, M., Zhang, H., Zhang, G., Li, D., Dong, Y., et al. (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat. Biotechnol., 28, 516–520.
148. Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., et al. (2003) Computational discovery of gene modules and regulatory networks. Nat. Biotechnol., 21, 1337–1342.
149. Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005) Reverse engineering of regulatory networks in human B cells. Nat. Genet., 37, 382–390.
150. Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. Science, 303, 799–805.
151. Lee, I., Date, S. V., Adai, A. T. and Marcotte, E. M. (2004) A probabilistic functional network of yeast genes. Science, 306, 1555–1558.
152. Liao, J. C., Boscolo, R., Yang, Y. L., Tran, L. M., Sabatti, C. and Roychowdhury, V. P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. USA, 100, 15522–15527.

153. Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B. and Marchal, K. (2006) Inferring transcriptional modules from ChIP-chip, motif and microarray data. Genome Biol., 7, R37.

154. Liu, X., Jessen, W. J., Sivaganesan, S., Aronow, B. J. and Medvedovic, M. (2007) Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. BMC Bioinformatics, 8, 283.

155. Youn, A., Reiss, D. J. and Stuetzle, W. (2010) Learning transcriptional networks from the integration of ChIP-chip and expression data in a non-parametric model. Bioinformatics, 26, 1879–1886.

156. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. and Vogelstein, B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. USA, 108, 9530–9535.