Title

Integrating ChIP-seq with other functional genomics data.

Journal

Briefings in Functional Genomics, 17(2)

Authors

Jiang, Shan

Mortazavi, Ali

Publication Date

2018-03-01

DOI

10.1093/bfgp/ely002

Peer reviewed

OXFORD

Briefings in Functional Genomics, 17(2), 2018, 104–115

doi: 10.1093/bfgp/ely002
Review paper

Integrating ChIP-seq with other functional genomics data

Shan Jiang and Ali Mortazavi

Corresponding author: Ali Mortazavi, Department of Developmental and Cell Biology, 2300 Biological Sciences 3, University of California, Irvine, CA 92697, USA. Tel: (949)824-6762; E-mail: ali.mortazavi@uci.edu

Abstract

Transcription is regulated by transcription factor (TF) binding at promoters and distal regulatory elements and histone modifications that control the accessibility of these elements. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become the standard assay for identifying genome-wide protein–DNA interactions *in vitro* and *in vivo*. As large-scale ChIP-seq data sets have been collected for different TFs and histone modifications, their potential to predict gene expression can be used to test hypotheses about the mechanisms of gene regulation. In addition, complementary functional genomics assays provide a global view of chromatin accessibility and long-range *cis*-regulatory interactions that are being combined with TF binding and histone remodeling to study the regulation of gene expression. Thus, ChIP-seq analysis is now widely integrated with other functional genomics assays to better understand gene regulatory mechanisms. In this review, we discuss advances and challenges in integrating ChIP-seq data to identify context-specific chromatin states associated with gene activity. We describe the overall computational design of integrating ChIP-seq data with other functional genomics assays. We also discuss the challenges of extending these methods to low-input ChIP-seq assays and related single-cell assays.

Key words: chip-seq; integrative analysis; chromatin states; self-organizing maps; hidden Markov models

Introduction

DNA–protein interactions and epigenetic modifications are crucial for transcriptional regulation. Genome-wide profiling of transcription factor (TF)-binding sites, regions with covalently modified histones and other DNA-binding proteins reveal cell- or tissue-, species- and disease-specific *cis*-regulatory repertoires, which are vital for understanding gene regulation. Chromatin immunoprecipitation (ChIP) methodologies [1–3] use an antibody that recognizes a TF or histone modification to pull down attached DNA for identifying binding locations. With the rapid development of sequencing technology, chromatin immunoprecipitation followed by sequencing (ChIP-seq) [2–5] has become the most common and effective assay to identify bound loci genome-wide *in vitro* and *in vivo*. The basic computational pipeline and software for analyzing ChIP-seq data have been established and optimized alongside advances in sequencing library preparation and ChIP-seq techniques [6–8], including read quality control, alignment, peak calling and evaluation of reproducibility. ChIP peaks can be visualized using genome browsers as a simple quality check of signal over known true positives. Confirmed peaks can be further analyzed with differential density analysis for different treatments, gene-associated annotation, motif discovery and other downstream analyses. Limitations and advances in these steps are reviewed in detail elsewhere [9].

However, the binding of one TF alone is rarely enough to directly infer functional effects on the gene expression levels of neighboring genes, which are typically under the combinatorial control of multiple TFs. Therefore, ChIP-seq data are often actively integrated with other functional genomic techniques to decipher

Shan Jiang is a PhD student in the Department of Developmental and Cell Biology at the University of California, Irvine. Her research focuses on comparative genomics of gene regulation during embryonic stem cell differentiation.
Ali Mortazavi is an associate professor in the Department of Developmental and Cell Biology at the University of California, Irvine. His research focuses on applications of functional and comparative genomics to study of cell differentiation and development.

the basic regulatory control of gene expression by incorporating open chromatin regions, long-range chromatin interactions and SNP (single-nucleotide polymorphism) variants. With the increasing availability of multiple ChIP-seq data sets [10, 11], as well as data sets from other genome-wide assays, the power of integrative computational analysis is of ever-increasing interest. In this review, we discuss the application of probabilistic models and machine learning methods to the analysis of TF and histone modification ChIP data simultaneously to identify chromatin patterns across multiple genomes and cell types. We also focus on the computational integration of ChIP-seq with other functional genomic assays such as RNA sequencing (RNA-seq) for gene expression levels, ATAC-/DNase-seq/FAIRE-seq for chromatin accessibility, and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)/Hi-C for chromatin interactions that affect regulation of gene expression. Finally, we discuss the development of ChIP-seq assays that use low amounts of input materials, and their further application in the emerging field of integrative analysis of single-cell sequencing functional genomics data.

Identifying distinct chromatin states using histone modifications and TF occupancy

Histone modifications are often found in recurring combinations at promoters, enhancers and repressed regions. These combinations are referred to as 'chromatin states' and can be used to annotate regulatory regions in genomes [12, 13]. For example, H3K4me1 alone marks primed enhancers, while H3K4me1 combined with H3K27ac mark active enhancers. Promoters are characterized by a detectable level of H3K4me3 coupled with a high ratio of H3K4me3 to H3K4me1. Furthermore, H3K36me3 histone modifications and RNA polymerase (Pol) II ChIP signal are associated with transcribed regions, while the presence of H3K27me3 or H3K9me3 is associated with repressive chromatin states (Figure 1) [14, 15]. The goal of software packages analyzing chromatin states is to first discover these relationships in the data, and to then check for changes in states assigned to a particular region in different cell types. Large-scale data sets produced by ENCODE [10] and Roadmap Epigenomics [11] have been used to train and to test with statistical or machine learning methods that assign chromatin states to genomic segments (typically 100 bp or longer). These state assignments can then be interpreted through comparisons with known annotations and gene expression.

Hidden Markov models (HMMs) were originally developed for speech recognition, but have since been used extensively in other fields to identify hidden states from observed signal data [16]. In genomics studies, it has been successfully applied to gene annotation [17] and protein domain characterization [18]. HMMseg [19] was the earliest software package to partition and annotate a genome by training HMMs on functional genomics data. However, this tool can only identify two states ('active' or 'inactive'), which limits its application in annotating chromatin states in greater detail, e.g. active/poised promoters and enhancers. ChromHMM [13] and Segway [20] were developed with the goal of capturing more comprehensive combinatorial patterns of multiple histone modifications, RNA Pol II binding and insulator CTCF binding genome-wide (Figure 2). ChromHMM segments the genome into minimum 200 bp intervals (default) and converts raw read counts into binary code using a product of independent Bernoulli random variables for each interval, which are then used to train a HMM. Similarly, Segway was developed based on dynamic Bayesian networks. It transforms raw read counts to coverage signal and can segment the genome down to 1 bp resolution, although 100 bp segments are more practical. Additional tools have been developed to extend and speed up the identification of chromatin states. For example, TreeHMM [21] also uses binary vectors but is position-dependent when inferring chromatin patterns during cell differentiation and across different cell types. hiHMM [22] uses a hierarchically linked infinite HMM model to not only identify chromatin states across multiple ChIP-seq data sets but also address species variance for cross-species inference. diHMM [23] inherits from ChromHMM but uses a hierarchical HMM to identify combinatorial patterns at variable length scale that range from nucleosome-level to higher-order domain-level states. Another joint analysis platform, IDEAS [24, 25], can infer chromatin states using both position-dependency and cell-type-specific cases at multiple range scales, and can run faster than both ChromHMM and Segway using single core mode. Additional tools have been developed for comparing chromatin patterns between different experimental treatments [26] and expanding the comprehensiveness of epigenomic maps [27]. The combinatorial patterns generated by these methods have been correlated with gene expression profiles to find context-specific signatures across cell types using linear regression model [24, 25]. However, the difficulty of interpreting large numbers of states has led to a practical preference for models with lower numbers of states. Typically, the focus is on the discovered states rather than their transition probabilities, unlike more traditional applications of HMM to gene annotation. The assumption is that a limited number of chromatin states and a small number of histone markers combinations covering significant fractions of the genome will capture most of the biologically relevant features.

While useful for predicting chromatin states, HMM-based methods have been relatively less successful when applied to a large number of TFs with restricted, presumably combinatorial binding patterns, which cover small fractions of the genome. Self-organizing maps (SOMs) are an alternative, unsupervised machine learning method for integratively analyzing such high-dimensional, comparatively sparse data. SOMs consist of individual units (which can be thought of as either neurons or mini-clusters) arranged on a scaffold that is trained with data to capture the high-density parts of high-dimensional data sets while preserving similarity relationships, i.e. data that are close in the input will also be close on the SOM. Chromatin SOMs identify TF-TF localization and co-binding pairs of TFs across cell types and tissues [28]. SOMs have been trained on the same data as chromHMM and Segway in ENCODE, namely, histone modification markers, RNA Pol II and CTCF. These are then overlaid post-training with additional data such as EP300 ChIP-seq signals to confirm cell-type-specific and commonly shared enhancer activity of groups of DNA segments [29]. For example, a trained SOM would distinguish open chromatin regions from promoters and enhancers based on their difference in H3K4me3 and H3K4me1 signal density (Figure 3). The individual units in SOM maps can be grouped into map regions called metaclusters [29, 30], which can then be analyzed for their ChIP-seq signal enrichments and used to automatically identify sets of potentially co-regulated regions [29]. Once a unit or metacluster of interest has been identified, proximal genes can be associated with bound DNA elements by using tools like GREAT [31] and Homer [32], and their gene expression profiles can be correlated [24] and visualized together with DNA element activity. Co-associated genes can then be analyzed for gene ontology enrichment using GREAT and Homer, but other tools such as DAVID [33] and Metascape [34] can also be applied to identify potential functional enrichments. While SOM does not impose a state transition model like HMMs, it

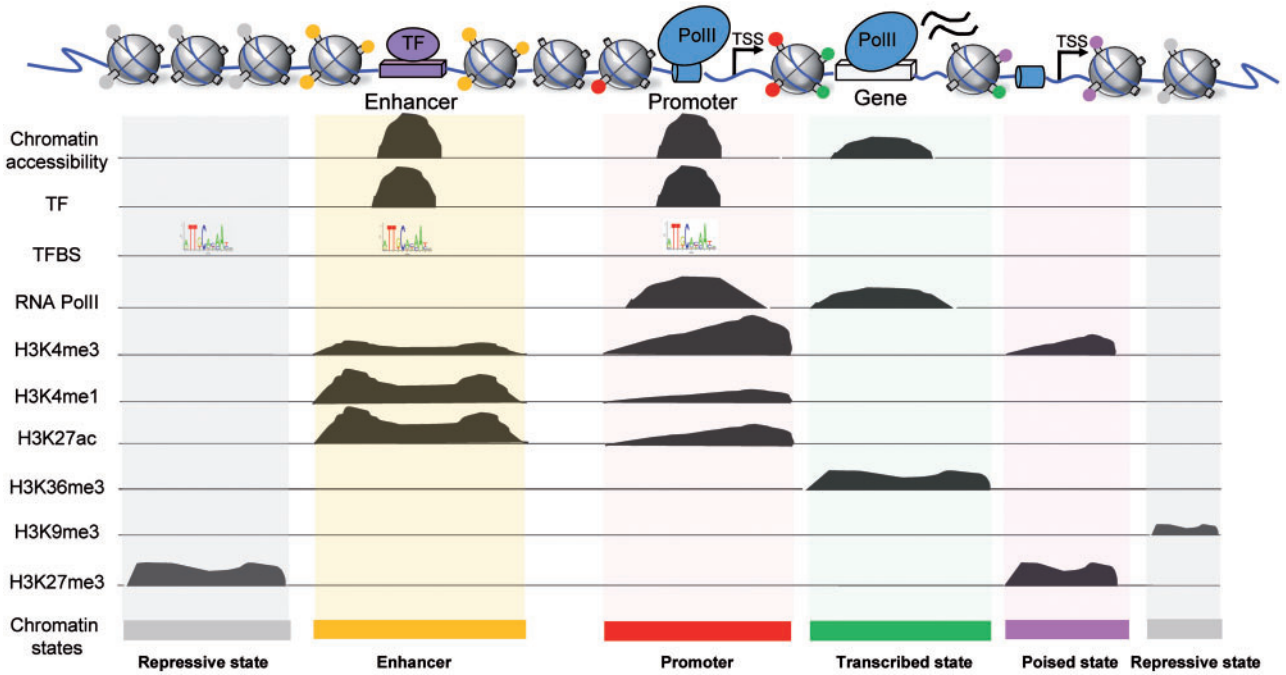


Figure 1. Chromatin states are defined by different combinations of histone modifications, TFs and RNA Pol II binding. In this example, a typical repressive state (gray) is characterized by high H3K27me3 signal or H3K9me3 signal, an enhancer state (yellow) would show a high occupancy ratio of H3K4me1 to H3K4me3 as well as high H3K27ac and the promoter state (red) would show a high occupancy ratio of H3K4me3 to H3K4me1 as well as RNA Pol II binding at the promoter, whereas poised promoter state (magenta) would show the occupancy of H3K4me3 and H3K27me3 bivalent modifications. Actively transcribed region (green) is characterized by a high occupancy of H3K36me3 with some RNA Pol II binding along the gene body.

recovers similar high-level states at the level of metaclusters but allows for further granular mining of 'microstates' corresponding to specific chromatin profiles in individual such as distinct combination of TFs that are present in small sections of the genome [29]. SOM can therefore be used to deeply data-mine for complicated relationships in highly dimensional ChIP-seq data sets.

Incorporating chromatin accessibility with ChIP-seq

Eukaryotic chromatin is tightly packaged into nucleosomes, and the positioning of nucleosomes regulated by TFs and histone modifications shows dynamic patterns during cell differentiation and development [35]. Specific proteins, often called pioneer factors, can control nucleosome repositioning via recruitment of chromatin remodelers, thus exposing *cis*-regulatory elements to lineage- or cell-type-specific TFs that activate or repress gene expression [15, 36]. Additionally, nucleosomes with H3.3/H2A.Z histone variants show hypermobility, which make them less stable and the DNA more easily accessible for TFs binding [37, 38]. Histone-depleted regions are referred to as open chromatin (Figure 1), and several sequencing assays have been developed to capture chromatin accessibility directly at high resolution such as DNase-seq [39–41], FAIRE-seq [42, 43] and ATAC-seq [44]. MNase-seq [35, 45, 46] is a related assay for identifying DNA regions occupied by nucleosomes instead of detecting open chromatin regions directly. DNase- and ATAC-seq depend on enzymatic digestion and Tn5 transposase insertion, respectively, to detect open chromatin regions *in vivo*. Both of them have a higher signal-to-noise ratio than the other methods, and ATAC-seq has become increasingly popular because of its ease of use. All of these methods need deep sequencing (about 50–100 million reads per sample) to get accurate, high-resolution profiles. The basic computational pipeline for open chromatin assays includes reads alignment, visualization for QC, peak calling and footprint analysis for DNase- and ATAC-seq or nucleosome profiling for MNase- and ATAC-seq (each step has been reviewed in detail elsewhere) [35, 47]. Specific software packages have been developed to detect signal-enriched regions for each assay. For example, Hotspot [48] detects DNase I hypersensitive regions for DNase-seq; GeneTrack [49] and DANPOS [50] do nucleosome calling for MNase-seq; NucleoATAC [51] calls nucleosome positions and occupancy for ATAC-seq. In addition, tools developed for ChIP-seq and DNase-seq peak calling also work effectively for ATAC-seq, such as MACS [52], Hotspot [48] and Homer [32]. DNAse-seq open chromatin data have been used alongside histone modification ChIP-seq data to define chromatin states using HMMs and SOMs in the ENCODE project [10, 29].

Deeper sequencing of open chromatin data to 200–500 million reads per sample can also be used to detect TF-binding occupancy 'footprints' at nucleotide resolution [35]. The ability of DNase- and ATAC-seq to perform footprint calling is the consequence of TF occupancy protecting DNA from nuclease cleavage and Tn5 transposition, which results in small stretches of fewer cuts within otherwise open regions. The sequences within these footprints can be compared with known motifs for identification [53–55]. The power of footprinting is that a single experiment can identify the binding sites for hundreds of TFs, a task that would be still gargantuan with hundreds of TF-specific ChIP-seq experiments. However, many TF motifs are similar to each other and can be difficult to distinguish based on sequence alone. For these cases, ChIP-seq of selected TFs can be used to validate the footprints when they are critical to the inferred gene regulatory networks [56]. Additionally, histone modification ChIP-seq data can be mapped

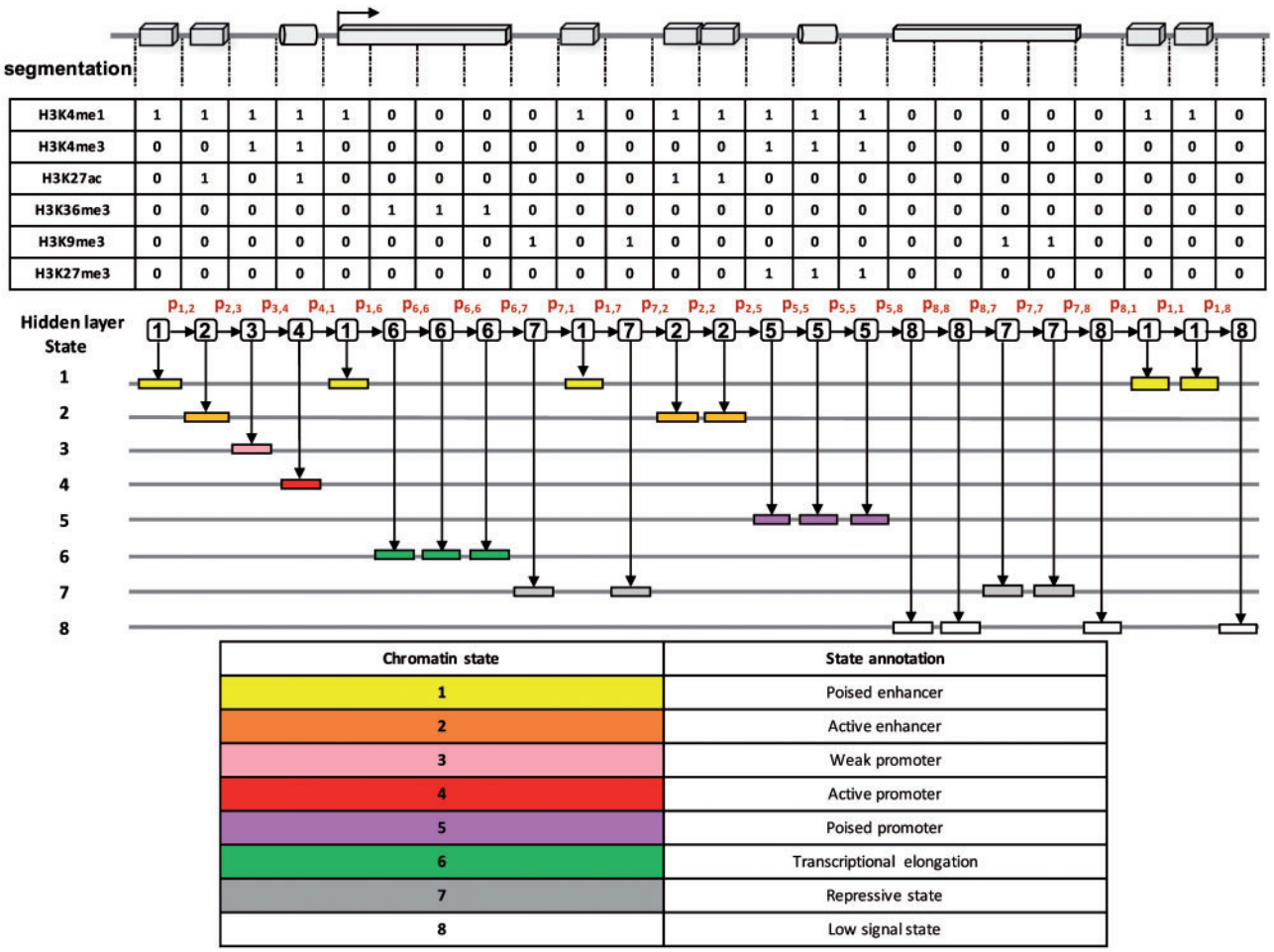
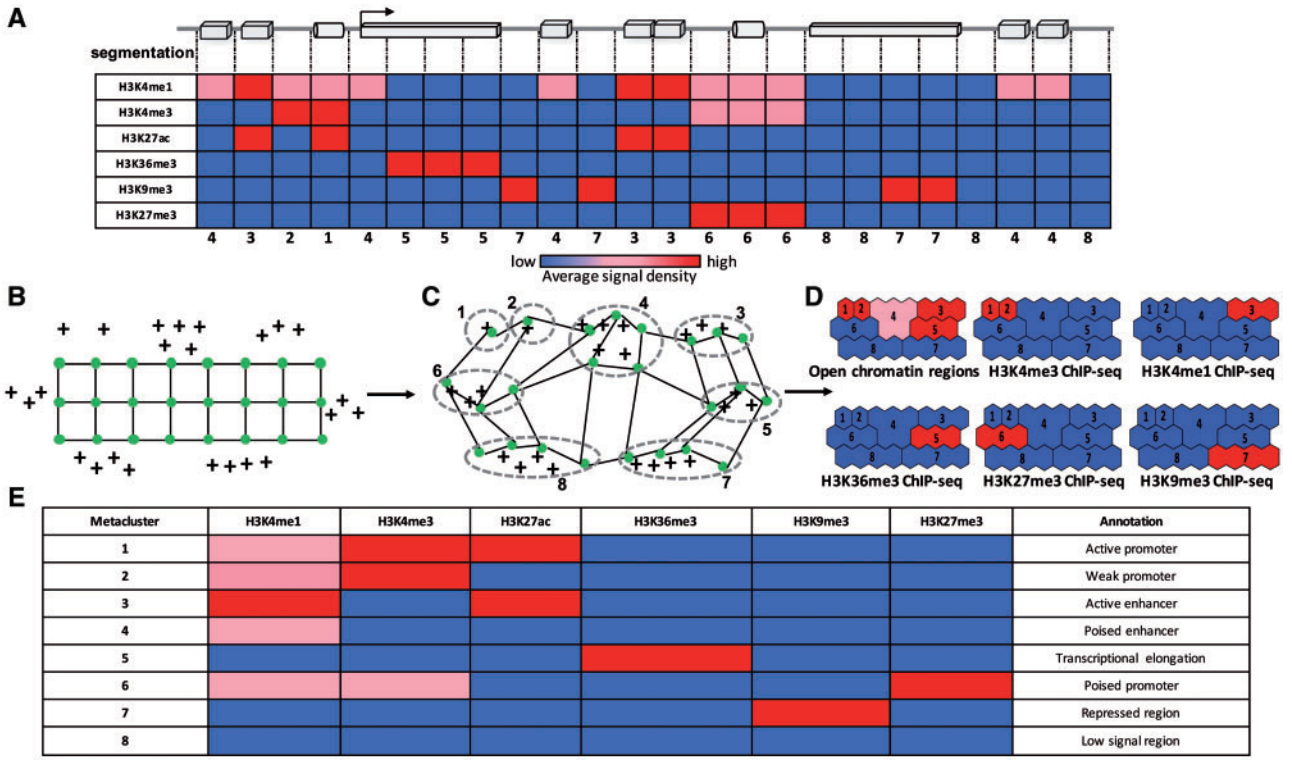


Figure 2. Graphical structure of annotating chromatin states using a HMM method such as ChromHMM. The genome is split into nonoverlapping segments, and ChIP-seq signal for histone modifications is binarized (0 or 1) and collected for each segment, which are further built into input matrix for HMM training. The hidden state of the current segment is dependent on the state of the previous one, and the transition probabilities (in red) of changing from one state to another are learnt from training on the input matrix. ChromHMM outputs trained hidden states for each segmentation, which are then interpreted as chromatin states based on the chromatin profile and gene annotations, such as active promoter/enhancer, transcriptional elongation or repressive states.

to open chromatin peaks to confirm the chromatin state of regulatory elements [44, 57–59]. The profiling of chromatin accessibility and TFs/histone occupancy has revealed that *cis*-regulatory elements show both transitory and stable activity during development and differentiation process for different lineages [60, 61]. Integrative analysis of chromatin accessibility and TFs occupancy from ChIP-seq has revealed that the two processes are not necessarily synchronous. Some TFs commonly referred to as pioneer factors can induce and remodel chromatin accessibility [62–65]. On the other hand, chromatin can be opened and activated before TF binding [48] or closed well after the TF has ceased to be bound. As open chromatin assays such as ATAC-seq are relatively easier to do and require less starting material than ChIP-seq, we expect that an increasing number of studies will start with open chromatin data followed with selected ChIP-seq for TFs and/or histone modifications. These data will be analyzed integratively with additional packages developed to facilitate their joint analysis.

Integrative analysis of gene expression with ChIP-seq

Most users of ChIP-seq data are interested in understanding the impact of TF binding or histone modifications on the expression of nearby genes, and therefore, ChIP-seq and RNA-seq are analyzed jointly to estimate this effect [6, 7, 14, 66, 67]. In the ideal case, a high ChIP-seq signal of a transcriptional activator would be found near highly expressed genes, while a high ChIP-seq signal of a repressor would be found near silenced genes. In another case, differentially expressed genes are first identified and classified into upregulated or downregulated genes between different experimental treatments. Then, differential TF and epigenetic occupancy are correlated with differential gene expression levels. TF-binding peaks and histone modification-enriched regions are associated with genes based on which gene is nearest, or using a particular distance radius. However, TF and epigenetic occupancy alone are seldom effective in predicting nearby target gene expression level accurately because (a) they cannot account for posttranscriptional turnover of the transcript, (b) it is difficult to accurately associate ChIP-seq peaks with their target genes and (c) we may not have the ChIP-seq data for all of the TFs controlling the expression of the target genes. One study has reported that the binding signal of 12 embryonic stem cell (ESC) TFs can explain 65% of the variance in mES gene expression, and the correlation coefficient between predicted and observed gene expression is 0.8 [68]. However, the predictive power of the same set of TFs in differentiated mES



Metacluster	H3K4me1	H3K4me3	H3K27ac	H3K36me3	H3K9me3	H3K27me3	Annotation
1							Active promoter
2							Weak promoter
3							Active enhancer
4							Poised enhancer
5							Transcriptional elongation
6							Poised promoter
7							Repressed region
8							Low signal region

Figure 3. Graphical structure of annotating chromatin states using SOMs. **(A)** The genome is split into nonoverlapping segments, and ChIP-seq signal for histone modifications is collected for each segment to build a signal matrix for SOM training, where each segment represents a vector of signal. **(B)** At the beginning of training, the map consists of a grid of regularly spaced or randomly initialized units (green dots) that we wish to fit to the data, which are signal vectors (black plus signs) spread in high-dimensional space. **(C)** For each training step, a signal vector is selected and the closest unit is found. The best matching unit is pulled as well as other surrounding units toward to the selected signal vector, which causes the map to adapt itself to match the data distribution in the space. **(D)** The trained SOM map is divided into metacluster regions (metaclusters 1–8) that represent combinations of signal enrichments. **(E)** Metaclusters are then assigned chromatin state labels by inspection based on annotations and the combinations of signal enrichments as in the HMM case.

decreased dramatically ($r = 0.2$) [68], and they can only explain 30% of gene expression variance in GM12878 [69]. In addition, while histone modifications alone can explain high gene expression variance in human CD4 T+ cells ($r = 0.7$), combinatorial histone modification combinations show different predictive power [70]. Inferring the effect of TFs on expression is complicated by the fact that TFs may activate a subset of target genes but repress others. Furthermore, TFs and histone marks have different power in predicting gene expression levels [71–73]. Thus, this approach is only practical for predicting gene expression in well-studied systems, where there are plenty of TFs and histone modifications data sets available that can be selected based on biological significance.

Efforts have been made to integrate chromatin accessibility data and ChIP-seq together to predict gene expression, and this combination is more accurate than using ChIP-seq alone [69]. However, the asynchrony between binding and chromatin accessibility also accounts for the less than perfect correlation between changes in these metrics and changes in gene expression. This is because transcription is the sum total of the multitude of effects of chromatin remodelers, TFs co-occupancy, different combination of histone marks and even DNA methylation, which are laborious to capture and profile simultaneously. Using regression models of RNA-seq, ChIP-seq and chromatin accessibility data, gene expression can be predicted from TFs/histone binding [69] and ChIP-seq-identified TF-binding motifs in open chromatin regions [74]. Mixed linear models of gene expression correlated with chromatin accessibility corrected with ChIP-seq TF binding can predict TF triggering or binding before chromatin remodeling [75]. Furthermore, TF-TFs co-occupancy can be predicted using support vector machines (SVMs) trained on open chromatin, histone markers and TFs ChIP-seq data [76]. The predictive power of integrated chromatin feature data can also be extended to the inference of gene regulatory networks. In one recent study [77], chromatin feature data were not only used to predict gene expression but also to predict the activation status of regulatory elements and further infer a context-specific gene regulatory network. The expression of TFs, target genes and chromatin remodelers as well as the accessibility of *cis*-regulatory elements and TF motifs in regulatory elements are integrated together and fed into a statistical Paired Expression and Chromatin Accessibility (PECA) model. This model predicts active *cis*-regulatory elements, TF expression and expression of related target genes within the same context-specific gene regulatory network, which are confirmed by knocking down key TFs in the network [78]. Although combining TF/histone modification ChIP-seq and chromatin accessibility data is an effective strategy for predicting gene expression and inferring gene regulatory networks, more software packages and platforms are still needed to be developed for integrating data from different functional assays. We expect that the next generation of packages will improve the predictive power of ChIP-seq for gene expression prediction using evermore sophisticated and robust statistical methods.

Incorporating long-range chromatin interactions with ChIP-seq

Most gene regulatory analyses only consider the effects of histone modifications and TFs on the nearest gene, thus not taking into account long-range interactions of *cis*-regulatory elements with more distal genes. Promoters and enhancers are physically coupled with target genes by chromatin loops mediated by TFs, cohesin, mediator and some noncoding RNAs to control gene expression [79–83]. A single promoter or enhancer can interact with multiple enhancers or promoters within the same chromatin loops [10, 84]. Recruitment of cofactors such as EP300 by TFs ultimately mediates these complex promoter–enhancer interactions. Chromosome conformation capture (3C)-based sequencing assays such as Hi-C [85, 86] and ChIA-PET [87] can be used to detect these long-range interactions. In particular, ChIA-PET combines ChIP and 3C-based methods to detect chromatin interactions between sites bound by specific proteins such as RNA Pol II or CTCF on a genome-wide scale [79, 88], but requires hundreds of millions of cells as starting materials. Compared with ChIA-PET, Hi-C can capture all sites interactions in the genome but at the expense of deep sequencing, as it needs at least a billion reads to achieve 1 kb resolution in mammalian genomes [85, 86, 89]. ChIA-PET can capture promoter–enhancer, promoter–promoter and enhancer–enhancer interactions that involve RNA Pol II directly, while Hi-C identifies TADs (topologically associated domains) in chromatin structure. Newer methods such as HiChIP [90] and PLAC-seq [91] combine the advantages of ChIA-PET and Hi-C to capture long-range interactions more efficiently and accurately. 3C-based methods and the basic computational analysis pipelines for each of the techniques have been reviewed previously [92, 93].

Although the mechanisms of long-range interactions are not completely understood, it is known that TFs and histone modifications are actively involved in the interactions and may help alter the chromatin structures [94]. By coupling ChIP-seq with long-range interaction data, studies find that TFs such as CTCF and YY1 are highly enriched in interacting loci or the boundaries of TADs in long-range interactions [86, 88, 89, 95–101]. Multiple studies have reported that CTCF can also co-bind with other TFs to form lineage—or cell-type—specific long-range interactions and activate context-specific gene expression [101–104]. It has also been shown that disruptions to TF binding at TADs boundaries or *cis*-regulatory elements, whether caused by mutations, methylation of TF-binding sites or deletion of a TF, can cause remodeling of chromatin interactions and abnormal expression of target genes, which may lead to disease [105, 106]. To integrate ChIP-seq data with ChIP-based long-range interaction data (i.e. ChIA-PET), peak callers are used to find TF co-binding and histone modifications in anchor sites of PETs [79, 107, 108]. For example, RNA Pol II ChIA-PET detects promoter–promoter and enhancer–promoter interactions directly. Enhancers or promoters can be further confirmed by comparing ChIP signal between H3K4me3 and H3K4me1 modifications [79]. In addition, distal enhancers have been thought to interact with promoters via cohesin-associated CTCF-CTCF loops that also insulate enhancers from genes that they are not supposed to target. The insulators are identified by overlapping anchor sites of cohesin ChIA-PET with cohesin and CTCF ChIP signal, while active enhancers are marked with H3K27ac ChIP signal [108]. Specific TFs co-binding patterns involved in the *cis*-interactions can be detected with ChIP-seq peak calling in the anchor regions [107]. Furthermore, differential promoter–promoter, enhancer–promoter and enhancer–enhancer interactions can be identified using ChIP-seq of histone modifications and comparing ChIP signal between conditions [79, 108]. For example, CTCF ChIP-seq signal at the anchor sites of cohesin PETs was used to confirm CTCF-CTCF loops in hESC. Although CTCF-CTCF loops are highly conserved between naïve and primed ES cells, the loop structures are different in terms of enhancer–promoter and enhancer–enhancer interactions, as can be seen by comparing H3K27ac ChIP-seq signal between the two states [108]. A popular strategy is to segment the genome into TADs using HiC when available, or predicting TADs using CTCF and/or cohesin component ChIP-seq to constrain interactions between TFs and *cis*-regulatory elements within these ~100–1000 kb regions [109]. By matching ChIP-seq peaks of CTCF and cohesin complex proteins to non-ChIP-based long-range interaction data, like Hi-C, TAD boundaries can be defined and TADs can be segmented into sub-transcription units more accurately [108]. Although TADs have relatively conserved segmentation structure during cell development and differentiation [105, 110], the intra-TAD interactions and epigenetic states of TADs are less stable in terms of outside stimulus and differentiation conditions [110, 111]. By comparing normalized ChIP signal of histone modifications within TADs before and after treatment, it is possible to define activated or repressed TAD states that are then correlated with differentially expressed genes within the same TADs. As ChIP-seq has been performed routinely in many laboratories and large consortiums such as the ENCODE [10] and modENCODE [112] projects, many ChIP-seq data sets are available for public use. Frequent chromatin interaction loci ('hubs') and TAD boundaries can be predicted accurately from published histone ChIP-seq data integrated with customized Hi-C [113]. Interestingly, a recent study shows that cohesin loss causes loop domains to disappear based on Hi-C data, but CTCF and histone modification ChIP-seq data show that their patterns are unaffected. The disappearance of loop domains only affects the expression levels of a small percentage of genes, which suggests that cohesin-mediated loops only have modest effects on transcription for most genes and that super-enhancers of genes seem to keep their activity intact without cohesin looping [114]. Thus, given the complex relationship between long-range interactions and gene expression, more studies applying Hi-C/ChIA-PET coupled with ChIP-seq are needed to understand the exact role of chromatin loops in gene expression and to further categorize genes based on their response to the disruption of loop formation.

Predicting regulatory sequence variants by integrative analysis with ChIP-seq

Sequence variants or SNPs are known to be associated with genetic traits and diseases [115, 116]. Most SNPs identified by genome-wide association studies as associated with traits or diseases are found outside of protein-coding regions, with the majority of these noncoding SNPs located in open chromatin regions [117, 118]. As open chromatin regions map to enhancers and promoters, noncoding SNPs in the accessible regions may interrupt or strengthen protein–DNA interactions by introducing sequence variants into binding motifs, and thus causing gene expression and traits to vary between individuals. Indeed, multiple studies have reported that many disease-causing nucleotide changes are in TF-binding sites and affect TF-DNA-binding events [10, 119–131]. The interruption in TF binding can not only influence proximal gene expression but also that of distal genes [122, 125, 129, 132]. However, only a minority of

differential TF-DNA-binding causes can be explained by sequence variation in binding motifs [133]. Besides, allelic occupancy profiling of >20 TFs using ChIP-seq data revealed that only a small proportion of these events have sequencing variants in binding motifs for specific TFs [134]. Although local variants in motifs are not necessarily affecting specific TF binding, sequence context is still an important source of differential TF-DNA binding. For example, proximal sequence changes may influence cooperative TF-TF binding [133, 135–138], and distal variants can affect TF-DNA and TF-TF interactions by changing chromatin state and conformation [133, 139–141].

Many efforts have been made to integrate ChIP-seq and other experimental data to predict regulatory sequence variants. One of the most straightforward methods is to match SNPs to known TF-binding motifs from database such as JASPAR [142] and TRANSFAC [143], or to look for putative TF-bindings sites using HMMs. The binding affinity score can be calculated based on a position weight matrix representation of the motif. When comparing the motif affinity score between two alleles, a greater motif score difference indicates that the variant is more likely to be regulatory [144–146]. However, these methods rely on known TF-binding sites and do not leverage the predictive power of chromatin signatures to filter out a large set of false-positive predictions. Recent studies have successfully integrated ChIP-seq and DNase-seq data into predictive analyses without relying on TF-binding motifs databases [147, 148]. In these studies, peak calls from ChIP-seq and DNase-seq are scanned for *k*-mers of a given length, and the putative regulatory sequences are used to train a SVM to predict the regulatory power of any *k*-mer sequence. The weighted sequences can then be used to predict the impacts of single-nucleotide changes on regulatory activity in the variant sequences [147]. Another version of this method is to weigh the predictive power of *k*-mer sequences and compute DNase-seq covariates from ChIP-seq data using regression methods. The trained *k*-mers and DNase-seq signals are then used to predict ChIP-seq binding signals at two alleles. By comparing the predicted ChIP-seq signal between the reference and variant alleles, the variant can be predicted to be regulatory or not [148]. Other studies have applied deep learning methods such as convolutional neural nets to more comprehensively integrate sequence variants, chromatin states, chromatin accessibility and even RNA-binding protein data to predict which regulatory variants will be functional [149, 150]. Some regulatory variants are disease-associated, and we can predict the effect of those variants on the binding affinity of TFs by evaluating the change in score for the motif [149]. We expect additional work on the development algorithms that can predict potential causal disease variants from the integration of functional genomics data, which will require experimental validation. The validation data in turn will be of great value for training the next set of methods to analyze variants from ChIP-seq data.

ChIP-seq integrative analysis in the era of low cell count and single-cell genomics

ChIP-seq has been the standard method for identifying genome-wide protein–DNA interactions when a specific antibody is available [151]. However, the traditional ChIP-seq technique requires a large amount of starting material (preferably >10 million of cells) to get high-resolution profiles, which limits its applicability for small organisms, rare cell types and single cells. Efforts have been made to optimize the ChIP-seq protocol for a low amount of starting materials, which successfully detect TF-binding signals with as few as 5000 cells [152] and H3K4me3-binding signals with only 500 cells [153]. Although these methods generate binding profiles at a good resolution with a small number of cells, the experimental procedures are still time-consuming and costly. Owing to the need for high polymerase chain reaction amplification in the low-input ChIP protocols, the number of identical aligned reads needs to be carefully corrected for during data analysis. The low-input ChIP-seq peaks can also be compared with open chromatin regions from ATAC-seq to show high correlation between enhancer histone modifications and open chromatin regions. By doing motif discovery analysis, people also identify lineage-specific TF binding to lineage-specific open chromatin regions. TF expression levels have been observed to correlate with differential open chromatin regions accessibility across cell types [153]. Another advancement in low-input ChIP-seq technique is to couple ChIP and Tn5 transposase tagmentation to add sequencing adapters to the bead-bound chromatin in a single step [154]. This protocol is both fast as well as cost-effective, and it successfully identifies TF binding with 100 000 cells and histone markers with 10 000 cells. The ChIP signal needs to be normalized to genomic tagmentated DNA to remove tagmentation bias. However, the protocol also benefits from Tn5 insertions in open chromatin regions to detect TF footprints and nucleosome positioning [154].

Single-cell epigenetics is a rapidly emerging area because of the development of new techniques [155, 156]. While we know that TF binding, histone modifications, chromatin accessibility, DNA methylation and long-range interactions work together to generate context-specific patterns, these results are primarily based on experiments with bulk samples. Individual cells may have different epigenetic patterns that influence their random behaviors [157]. Therefore, many single-cell epigenetics assays [156] have been developed to study this, including scATAC-seq [158, 159], scHi-C [160] and scBS-seq [161–163]. In addition, several techniques have been developed to couple multiple functional assays together to get transcriptomic and epigenetic data from the same cell simultaneously [164–166]. Compared with these methods, single-cell ChIP-seq seems more limited because of the technical difficulties of working from so little material. Only one protocol has successfully performed ChIP-seq at single-cell level [167], identifying hundreds of histone modification peaks per cell. The authors successfully distinguished three cell types by doing unsupervised hierarchical clustering and identifying subpopulations with different chromatin signatures. However, the low input and antibody sensitivity cause single-cell ChIP-seq to suffer from high technical variance and low sensitivity across individual cells. Similarly, recent advances in single-cell ATAC-seq [158, 159] successfully identified individual open chromatin regions in single cells, with the downside of low signal-to-noise compared with bulk ATAC-seq. However, scACTAC-seq reads are aggregated to be validated when comparing with bulk ATAC-seq data, which shows less technical variance and higher sensitivity compared with single-cell ChIP-seq. The high background IP noise probably limits scChIP-seq to histone modifications, and extensive computational analysis needs to be carried out to remove the noise in peak calling. The strategy used for now is to segment ChIP-ed DNA for peak calling for individual cells and then cluster cells based on fractions of reads in known ChIP peaks from bulk samples. Thus, the analysis is still performed at low-input level rather than the true single-cell level [9]. Future studies need to develop methods to remove IP noise and improve solid peak calling in individual

cells, as bulk analysis methods cannot be applied directly in single-cell assays. We can further expect that methods will appear combining single-cell ChIP-seq and single-cell RNA-seq from the same cells, which will open up new possibilities when working from mixed cell types and difficult-to-obtain samples.

Future direction and conclusion

ChIP-seq has become the standard method for profiling protein–DNA binding over the past decade, and it has been actively integrated with newer functional genomics assays such as RNA-seq, DNase/ATAC-seq and Hi-C/ChIA-PET to generate models of gene regulation. In the best studies, the integrative analysis is validated with a series of validation experiments to show that the binding of particular TFs is critical for target genes expression. As ATAC-seq and RNA-seq protocols continue to become easier, we expect that ChIP-seq will be routinely integrated with these functional genomics assays. While most current studies compare different 'static' cell types, transcription changes temporally in response to stimuli that involve changes in TF binding, and will become more often the subject of study using ChIP-seq during development and/or stimulation. ChIP-seq following perturbations will also become more routine, and will need to be integrated when building predictive models to identify potentially active *cis*-regulatory elements and key TFs, which would guide experimental validation and will feed back into further model building.

Another challenge in ChIP-seq integrative analysis will be how to incorporate long-range interaction and gene expression data into the chromatin state analyses that are being done with HMMs and SOMs. Currently, all of these analyses include multiple ChIP-seq data sets and can incorporate chromatin accessibility but are not designed to incorporate connectivity between distant regions or gene expression data as part of their training as opposed to post-training analysis and annotation. A challenge is that while at least ChIA-PET and HiC are working in a similar 'feature space' of chromatin as ChIP-seq, regular RNA-seq is measuring the steady state of transcripts, which is affected by several posttranscriptional processes such as mRNA turnover mediated by microRNAs. As chromatin will always be more predictive of transcriptional initiation, it may be more fruitful to compare the predicted models of expression to GRO-seq and other measurements of transcriptional activity than regular RNA-seq.

In recent years, ChIP-seq techniques for low-input materials have been developed to expand its applications to rare tissues or cell types, and even single-cell studies. Other functional genomics assays have also been developed at single-cell level to answer new biological questions. However, the integrative analysis of single-cell ChIP-seq with these functional genomics assays in single cells is a difficult challenge. One reason is that the experimental protocols to capture protein binding, transcriptomes and DNA methylation data from the same cell are still not available. However, it may still be worthwhile to integrate data from scChIP-seq and other functional genomics assays in different individual cells from the same pool based on the assumption that protein-binding profiles would match to the gene expression profiles from the assay because these cells are from the same pool. Once protocols are available to do scRNA-seq and scChIP-seq from the same single cell, algorithms will need to be developed to integrate these single-cell data types together to understand the connection between binding and gene expression heterogeneity in subsets of a cell population. As single-cell data are sparser than bulk data, new statistical methods and tools are required for integration. In a hopefully not-so-distant future where robust single-cell ChIP-seq and RNA-seq are practical, they could become the method of choice for studying samples where the amount of material or the heterogeneity of the population makes the bulk version of these experiments less attractive.

Key Points

- TF and histone modification ChIP-seq data can be used to define chromatin states for annotating regulatory regions in the genome.
- ChIP-seq data can be integrated with chromatin accessibility and long-range interaction data to further decipher mechanisms of gene regulation.
- ChIP-seq data can be integrated with other functional genomics data to predict noncoding regulatory sequence variants.
- Single-cell ChIP-seq promises to reveal the cell-to-cell variability of TF and histone occupancy, but the experimental and computational methods still need to be improved to capture meaningful ChIP signal.

Acknowledgements

The authors thank Dr Weihua Zeng, Rabi Murad, Camden Jansen, Dana Wyman, Katherine Williams and Sorena Rahmanian for kind suggestions on revising the manuscript.

Funding

This work was supported by a NIH New Innovator Award to A.M. (grant number DP2 GM111100).

References

1. Ren B, Robert F, Wyrick JJ, *et al*. Genome-wide location and function of DNA binding proteins. *Science* 2000;**290**(5500): 2306–9.
2. Johnson DS, Mortazavi A, Myers RM, *et al*. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 2007; **316**(5830):1497–502.
3. Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* 2007; **129**(4):823–37.
4. Robertson G, Hirst M, Bainbridge M, *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;**4**(8):651–7.
5. Mikkelsen TS, Ku M, Jaffe DB, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;**448**(7153):553–60.
6. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**(10):669–80.
7. Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* 2012;**13**(12):840–52.
8. Landt SG, Marinov GK, Kundaje A, *et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;**22**(9):1813–31.
9. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform* 2017;**18**:279–90.

10. . The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**:57–74.
11. Kundaje A, Meuleman W, Ernst J, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**(7539): 317–30.
12. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010;**28**(8):817–25.
13. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**(3):215–6.
14. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011;**12**(1):7–18.
15. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* 2013;**49**(5):825–37.
16. Eddy SR. What is a hidden Markov model? *Nat Biotechnol* 2004;**22**(10):1315–6.
17. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 2012;**13**(5):329–42.
18. Finn RD, Coggill P, Eberhardt RY, *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;**44**:D279–85.
19. Day N, Hemmaplardh A, Thurman RE, *et al*. Unsupervised segmentation of continuous genomic data. *Bioinformatics* 2007;**23**(11):1424–6.
20. Hoffman MM, Buske OJ, Wang J, *et al*. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 2012;**9**(5):473–6.
21. Biesinger J, Wang Y, Xie X. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics* 2013;**14**(Suppl 5):S4.
22. Sohn KA, Ho JWK, Djordjevic D, *et al*. HiHMM: bayesian nonparametric joint inference of chromatin state maps. *Bioinformatics* 2015;**31**(13):2066–74.
23. Marco E, Meuleman W, Huang J, *et al*. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. *Nat Commun* 2017;**8**:15011.
24. Zhang Y, An L, Yue F, *et al*. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* 2016;**44**(14):6721–31.
25. Zhang Y, Hardison RC. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res* 2017;**45**(17):9823–36.
26. Yen A, Kellis M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat Commun* 2015;**6**(1):7973.
27. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 2015;**33**(4):364–76.
28. Xie D, Boyle AP, Wu L, *et al*. Dynamic trans-acting factor colocalization in human cells. *Cell* 2013;**155**(3):713–24.
29. Mortazavi A, Pepke S, Jansen C, *et al*. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res* 2013;**23**(12):2136–48.
30. Longabaugh WJR, Zeng W, Zhang JA, *et al*. Bcl11b and combinatorial resolution of cell fate in the T-cell gene regulatory network. *Proc Natl Acad Sci USA* 2017;**114**(23):5800–7.
31. McLean CY, Bristor D, Hiller M, *et al*. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010;**28**(5):495–501.
32. Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**(4):576–89.
33. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;**37**(1):1–13.
34. Tripathi S, Pohl MO, Zhou Y, *et al*. Meta- and orthogonal integration of influenza "omics" data defines a role for UBR4 in virus budding. *Cell Host Microbe* 2015;**18**(6):723–35.
35. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014;**7**(1):33.
36. Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. *Genes Dev* 2014;**28**(24):2679–92.
37. Creyghton MP, Markoulaki S, Levine SS, *et al*. H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment. *Cell* 2008;**135**(4): 649–61.
38. Jin C, Felsenfeld G. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes Dev* 2007;**21**(12):1519–29.
39. Boyle AP, Song L, Lee BK, *et al*. High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res* 2011;**21**(3):456–64.
40. Hesselberth JR, Chen X, Zhang Z, *et al*. Global mapping of protein– DNA interactions *in vivo* by digital genomic footprinting. *Nat Methods* 2009;**6**(4):283–9.
41. Neph S, Vierstra J, Stergachis AB, *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012;**489**(7414):83–90.
42. Simon JM, Giresi PG, Davis IJ, *et al*. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 2012;**7**(2):256–67.
43. Bianco S, Rodrigue S, Murphy BD, *et al*. Global mapping of open chromatin regulatory elements by formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq). *Methods Mol Biol* 2015;**1334**: 261–72.
44. Buenrostro JD, Giresi PG, Zaba LC, *et al*. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;**10**(12):1213–8.
45. Ponts N, Harris EY, Prudhomme J, *et al*. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res* 2010;**20**(2):228–38.
46. Schones DE, Cui KR, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008; **132**(5):887–98.
47. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* 2014;**15**(11):709–21.
48. John S, Sabo PJ, Thurman RE, *et al*. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 2011;**43**(3):264–8.
49. Albert I, Wachi S, Jiang C, *et al*. GeneTrack–a genomic data processing and visualization framework. *Bioinformatics* 2008; **24**(10):1305–6.
50. Chen K, Xi Y, Pan X, *et al*. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* 2013;**23**(2):341–51.
51. Schep AN, Buenrostro JD, Denny SK, *et al*. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* 2015;**25**(11):1757–70.
52. Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-seq (MACS). *Genome Biol* 2008;**9**(9):R137.

53. Piper J, Elze MC, Cauchy P, *et al*. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res* 2013;**41**:e201.
54. Pique-Regi R, Degner JF, Pai AA, *et al*. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 2011;**21**(3):447–55.
55. Jankowski A, Tiuryn J, Prabhakar S. Romulus: robust multistate identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* 2016;**32**(16):2419–26.
56. Ramirez RN, El-Ali NC, Mager MA, *et al*. Dynamic gene regulatory networks of human myeloid differentiation. *Cell Sys* 2017;**4**(4):416–29.
57. Prescott SL, Srinivasan R, Marchetto MC, *et al*. enhancer divergence and cis-regulatory evolution in the human and Chimp neural crest. *Cell* 2015;**163**(1):68–84.
58. Wu J, Huang B, Chen H, *et al*. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* 2016;**534**(7609):652–7.
59. Minoux M, Holwerda S, Vitobello A, *et al*. Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. *Science* 2017;**355**(6332):eaal2913.
60. Stavreva DA, Coulon A, Baek S, *et al*. Dynamics of chromatin accessibility and long-range interactions in response to glucocorticoid pulsing. *Genome Res* 2015;**25**(6):845–57.
61. Su Y, Shin J, Zhong C, *et al*. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nat Neurosci* 2017;**20**(3):476–83.
62. Biddie SC, John S, Sabo PJ, *et al*. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* 2011;**43**(1):145–55.
63. Thurman RE, Rynes E, Humbert R, *et al*. The accessible chromatin landscape of the human genome. *Nature* 2012; **489**(7414):75–82.
64. Sherwood RI, Hashimoto T, O'Donnell CW, *et al*. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 2014;**32**(2):171–8.
65. Takaku M, Grimm SA, Shimbo T, *et al*. GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biol* 2016;**17**(1):36.
66. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 2009;**6**(11 Suppl):S22–32.
67. Angelini C, Costa V. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front Cell Dev Biol* 2014; **2**:51.
68. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci USA* 2009;**106**(51): 21521–6.
69. McLeay RC, Lesluyes T, Cuellar Partida G, *et al*. Genome-wide in silico prediction of gene expression. *Bioinformatics* 2012; **28**(21):2789–96.
70. Karlic R, Chung H-R, Lasserre J, *et al*. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci USA* 2010;**107**(7):2926–31.
71. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 2012;**40**(2):553–68.
72. Cheng C, Yan KK, Yip KY, *et al*. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* 2011;**12**(2):R15.
73. Dong X, Greven MC, Kundaje A, *et al*. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* 2012;**13**(9):R53.
74. Natarajan A, Yardimci GG, Sheffield NC, *et al*. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* 2012;**22**(9):1711–22.
75. Lamparter D, Marbach D, Rueedi R, *et al*. genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility. *PLoS Comput Biol* 2017;**13**(1):e1005311.
76. Liu L, Zhao W, Zhou X. Modeling co-occupancy of transcription factors using chromatin features. *Nucleic Acids Res* 2016; **44**(5):e49.
77. Duren Z, Chen X, Jiang R, *et al*. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci USA* 2017;**114**(25):E4914–23.
78. Wang J, Jiang W, Yan Y, *et al*. Knockdown of EWSR1/FLI1 expression alters the transcriptome of Ewing sarcoma cells *in vitro*. *J Bone Oncol* 2016;**5**(4):153–8.
79. Li G, Ruan X, Auerbach RK, *et al*. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 2012;**148**(1–2):84–98.
80. Harmston N, Lenhard B. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res* 2013;**41**(15): 7185–99.
81. Cavalli G, Misteli T. Functional implications of genome topology. *Nat Struct Mol Biol* 2013;**20**(3):290–9.
82. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 2014;**15**(4):272–86.
83. Böhmdorfer G, Wierzbicki AT. Control of chromatin structure by long noncoding RNA. *Trends Cell Biol* 2015;**25**(10):623–32.
84. Sanyal A, Lajoie BR, Jain G, *et al*. The long-range interaction landscape of gene promoters. *Nature* 2012;**489**(7414):109–13.
85. Lieberman-aiden E, Berkum NL, Van Williams L, *et al*. comprehensive mapping of long-range interactions revelas folding principles of the human genome. *Science* 2009;**326**(5950): 289–94.
86. Dixon JR, Selvaraj S, Yue F, *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**(7398):376–80.
87. Fullwood MJ, Liu MH, Pan YF, *et al*. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 2009; **462**(7269):58–64.
88. Handoko L, Xu H, Li G, *et al*. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 2011;**43**(7): 630–8.
89. Rao SSP, Huntley MH, Durand NC, *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**(7):1665–80.
90. Mumbach MR, Rubin AJ, Flynn RA, *et al*. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* 2016;**13**(11):919–22.
91. Fang R, Yu M, Li G, *et al*. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* 2016;**26**(12):1345–8.
92. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* 2016; **17**(12):743–55.
93. Davies JO, Oudelaar AM, Higgs DR, *et al*. How best to identify chromosomal interactions: a comparison of approaches. *Nat Methods* 2017;**14**(2):125–34.
94. Krivega I, Dean A. Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev* 2012;**22**(2):79–85.

95. Donohoe ME, Zhang LF, Xu N, *et al*. Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol Cell* 2007;**25**(1):43–56.
96. Seitan VC, Faure AJ, Zhan Y, *et al*. Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res* 2013;**23**(12): 2066–77.
97. Giorgetti L, Galupa R, Nora EP, *et al*. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* 2014;**157**(4):950–63.
98. Zuin J, Dixon JR, van der Reijden MI, *et al*. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA* 2014;**111**(3): 996–1001.
99. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014;**15**(4): 234–46.
100. Gómez-Marín C, Tena JJ, Acemel RD, *et al*. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc Natl Acad Sci USA* 2015;**112**(24):7542–7.
101. Beagan JA, Duong MT, Titus KR, *et al*. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* 2017;**27**(7):1139–52.
102. Donohoe ME, Silva SS, Pinter SF, *et al*. The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature* 2009;**460**(7251):128–32.
103. Lee J, Krivega I, Dale RK, *et al*. The LDB1 complex co-opts CTCF for erythroid lineage-specific long-range enhancer interactions. *Cell Rep* 2017;**19**(12):2490–502.
104. Jerković I, Ibrahim DM, Andrey G, *et al*. Genome-wide binding of posterior HOXA/D transcription factors reveals subgrouping and association with CTCF. *PLoS Genet* 2017;**13**: e1006567.
105. Nora EP, Lajoie BR, Schulz EG, *et al*. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012;**485**(7398):381–5.
106. Krijger PH, de Laat W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* 2016; **17**(12):771–82.
107. Zhang Y, Wong CH, Birnbaum RY, *et al*. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* 2013;**504**(7479):306–10.
108. Ji X, Dadon DB, Powell BE, *et al*. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* 2016; **18**(2):262–75.
109. Jost D, Vaillant C, Meister P. Coupling 1D modifications and 3D nuclear organization: data, models and function. *Curr Opin Cell Biol* 2017;**44**:20–7.
110. Le Dily F, Baù D, Pohl A, *et al*. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* 2014; **28**(19):2151–62.
111. Neems DS, Garza-Gongora AG, Smith ED, *et al*. Topologically associated domains enriched for lineage-specific genes reveal expression-dependent nuclear topologies during myogenesis. *Proc Natl Acad Sci USA* 2016; **113**(12):E1691–700.
112. Celniker SE, Dillon LA, Gerstein MB, *et al*. Unlocking the secrets of the genome. *Nature* 2009;**459**(7249):927–30.
113. Huang J, Marco E, Pinello L, *et al*. Predicting chromatin organization using histone marks. *Genome Biol* 2015;**16**(1):162.
114. Rao SS, Huang SC, St Hilaire BG, *et al*. Cohesin Loss Eliminates All Loop Domains. *Cell* 2017;**171**(2):305–20.
115. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**(23):9362–7.
116. Auton A, Brooks LD, Durbin RM, *et al*. A global reference for human genetic variation. *Nature* 2015;**526**(7571):68–74.
117. Maurano MT, Humbert R, Rynes E, *et al*. systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;**337**(6099):1190–5.
118. Schaub MA, Boyle AP, Kundaje A, *et al*. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012;**22**(9):1748–59.
119. Martin DI, Tsai SF, Orkin SH. Increased gamma-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature* 1989;**338**(6214):435–8.
120. Matsuda M, Sakamoto N, Fukumaki Y. Delta-thalassemia caused by disruption of the site for an erythroid- specific transcription factor, GATA-1, in the delta-globin gene promoter. *Blood* 1992;**80**(5):1347–51.
121. De Gobbi M, Viprakasit V, Hughes JR, *et al*. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 2006;**312**(5777):1215–7.
122. Jeong Y, Leskow FC, El-Jaick K, *et al*. Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat Genet* 2008;**40**(11):1348–53.
123. Musunuru K, Strong A, Frank-Kamenetsky M, *et al*. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010;**466**(7307):714–9.
124. Al Zadjali S, Wali Y, Al Lawatiya F, *et al*. The β-globin promoter -71 C>T mutation is a $\beta+$ thalassemic allele. *Eur J Haematol* 2011;**87**(5):457–60.
125. Claussnitzer M, Dankel SN, Klocke B, *et al*. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* 2014;**156**(1–2): 343–58.
126. Kulzer JR, Stitzel ML, Morken MA, *et al*. A common functional regulatory variant at a type 2 diabetes locus upregulates arap1 expression in the pancreatic beta cell. *Am J Hum Genet* 2014;**94**(2):186–97.
127. Weinhold N, Jacobsen A, Schultz N, *et al*. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* 2014;**46**(11):1160–5.
128. Weedon MN, Cebola I, Patch AM, *et al*. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* 2014;**46**(1):61–4.
129. Claussnitzer M, Dankel SN, Kim KH, *et al*. FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* 2015;**373**(10):895–907.
130. Wienert B, Funnell APW, Norton LJ, *et al*. Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nat Commun* 2015;**6**(1): 7085.
131. Wang S, Wu S, Meng Q, *et al*. FAS rs2234767 and rs1800682 polymorphisms jointly contributed to risk of colorectal cancer by affecting SP1/STAT1 complex recruitment to chromatin. *Sci Rep* 2016;**6**(1):19229.
132. Smemo S, Tena JJ, Kim KH, *et al*. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 2014;**507**(7492):371–5.
133. Deplancke B, Alpern D, Gardeux V. The genetics of transcription factor DNA binding variation. *Cell* 2016;**166**(3):538–54.
134. Reddy TE, Gertz J, Pauli F, *et al*. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* 2012;**22**(5):860–9.

135. Kilpinen H, Waszak SM, Gschwind AR, *et al*. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 2013;**342**(6159): 744–7.
136. Siersbæk R, Rabiee A, Nielsen R, *et al*. Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell Rep* 2014;**7**(5):1443–55.
137. Tijssen MR, Cvejic A, Joshi A, *et al*. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* 2011;**20**(5):597–609.
138. Domcke S, Bardet AF, Adrian Ginno P, *et al*. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* 2015;**528**(7583):575–9.
139. Ding Z, Ni Y, Timmer SW, *et al*. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet* 2014; **10**(11):e1004798.
140. Waszak SM, Delaneau O, Gschwind AR, *et al*. Population variation and genetic control of modular chromatin architecture in humans. *Cell* 2015;**162**(5):1039–50.
141. Grubert F, Zaugg JB, Kasowski M, *et al*. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* 2015;**162**(5):1051–65.
142. Mathelier A, Fornes O, Arenillas DJ, *et al*. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016; **44**(D1):D110–5.
143. Wingender E, Dietze P, Karas H, *et al*. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996;**24**(1):238–41.
144. Andersen MC, Engström PG, Lithwick S, *et al*. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol* 2008;**4**(1):e5.
145. Macintyre G, Bailey J, Haviv I, *et al*. Is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics* 2010; **26**(18):i524–30.
146. Riva A. Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics* 2012;**13**(Suppl 4):S7.
147. Lee D, Gorkin DU, Baker M, *et al*. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 2015;**47**(8):955–61.
148. Zeng H, Hashimoto T, Kang DD, *et al*. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* 2016;**32**(4):490–6.
149. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**(8):831–8.
150. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;**12**(10):931–4.
151. Adli M, Bernstein BE. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc* 2011;**6**(10):1656–68.
152. Shankaranarayanan P, Mendoza-Parra MA, Walia M, *et al*. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat Methods* 2011;**8**(7):565–7.
153. Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, *et al*. Chromatin state dynamics during blood formation. *Science* 2014; **345**(6199):943–9.
154. Schmidl C, Rendeiro AF, Sheffield NC, *et al*. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* 2015;**12**(10):963–5.
155. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015;**16**(12): 716–26.
156. Clark SJ, Lee HJ, Smallwood SA, *et al*. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* 2016;**17**(1):10.
157. Sekelja M, Paulsen J, Collas P. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? *Genome Biol* 2016;**17**(1):54.
158. Buenrostro JD, Wu B, Litzenburger UM, *et al*. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;**523**(7561):486–90.
159. Cusanovich DA, Daza R, Adey A, *et al*. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015;**348**(6237):910–4.
160. Nagano T, Lubling Y, Stevens TJ, *et al*. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**(7469):59–64.
161. Guo H, Zhu P, Wu X, *et al*. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* 2013;**23**(12):2126–35.
162. Smallwood SA, Lee HJ, Angermueller C, *et al*. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014;**11**(8):817–20.
163. Farlik M, Sheffield NC, Nuzzo A, *et al*. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* 2015;**10**(8):1386–97.
164. Macaulay IC, Haerty W, Kumar P, *et al*. G & T-seq: parallel sequencing of single- cell genomes and transcriptomes. *Nat Methods* 2015;**12**(6):519–22.
165. Angermueller C, Clark SJ, Lee HJ, *et al*. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 2016;**13**(3):229–32.
166. Hu Y, Huang K, An Q, *et al*. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* 2016;**17**:88.
167. Rotem A, Ram O, Shoresh N, *et al*. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* 2015;**33**(11):1165–72.