Methods 187 (2021) 44–53

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Methods for ChIP-seq analysis: A practical workflow and advanced applications

Ryuichiro Nakato [a,*], Toyonori Sakata [b]

[a] Laboratory of Computational Genomics, Institute for Quantitative Biosciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan
[b] Laboratory of Genome Structure and Function, Institute for Quantitative Biosciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

ARTICLE INFO

Keywords:
ChIP-seq
Histone modifications
Chromatin state
Single-cell analysis
Quality assessment
Machine learning

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a central method in epigenomic research. Genome-wide analysis of histone modifications, such as enhancer analysis and genome-wide chromatin state annotation, enables systematic analysis of how the epigenomic landscape contributes to cell identity, development, lineage specification, and disease. In this review, we first present a typical ChIP-seq analysis workflow, from quality assessment to chromatin-state annotation. We focus on practical, rather than theoretical, approaches for biological studies. Next, we outline various advanced ChIP-seq applications and introduce several state-of-the-art methods, including prediction of gene expression level and chromatin loops from epigenome data and data imputation. Finally, we discuss recently developed single-cell ChIP-seq analysis methodologies that elucidate the cellular diversity within complex tissues and cancers.

1. Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis is a key technology in epigenomic research. This method uses an antibody for a specific DNA-binding protein or a histone modification to identify enriched loci within a genome [1,2]. Histone modifications are used in the ChIP-seq analysis field to dissect the characteristics and the biological functions of epigenetic signatures. Advances in next-generation sequencing (NGS) technology and computational analysis enable us to systematically understand how the epigenomic landscape contributes to cell identity [3], development [4], lineage specification [5–8], cancer [9], and other diseases [10,11].

Five "core histone marks", proposed by Roadmap Epigenomics Consortium [12], are widely used for ChIP-seq analysis:

- H3 lysine 4 monomethylation (H3K4me1) or H3 lysine 27 acetylation (H3K27ac), which is associated with enhancer regions;
- H3 lysine 4 trimethylation (H3K4me3), which is associated with promoter regions;
- H3 lysine 36 trimethylation (H3K36me3), which is associated with transcribed regions in gene bodies;
- H3 lysine 27 trimethylation (H3K27me3), which is associated with Polycomb repression; and
- H3 lysine 9 trimethylation (H3K9me3), which is associated with heterochromatin.

In addition to genome-wide identification of specific epigenome marks (e.g., enhancers) in a specific cell-line [13], core histone mark enrichment profiles are used to segment and annotate whole-genome regions into distinct "chromatin states," which represent more detailed characteristic epigenetic signatures (e.g., weak transcription and poised promoter) [3]. Maturation of high-quality ChIP-seq databases by large consortia such as ENCODE [14], the Roadmap Epigenomics Consortium [12], and the International Human Epigenome Consortium (IHEC) [15] accelerate chromatin state annotation for various cell lines and tissues. Many studies leverage the accumulated epigenomic information to infer additional genome dynamics using machine-learning approaches.

In this review, we first address the major steps in a typical ChIP-seq computational analysis workflow. Because there are numerous important studies in this field, we focus on outlining the concept for each step by referencing previous important reviews instead of describing each method. Next, we introduce several advanced ChIP-seq applications for histone modifications, including prediction of gene expression level and enhancer-promoter looping, and data imputation. Finally, we discuss recently developed methodologies for single-cell ChIP-seq (scChIP-seq) analysis that elucidate the cellular diversity within complex tissues and

* Corresponding author.
E-mail addresses: rnakato@iam.u-tokyo.ac.jp (R. Nakato), tsakata@iam.u-tokyo.ac.jp (T. Sakata).

https://doi.org/10.1016/j.ymeth.2020.03.005
Received 18 February 2020; Received in revised form 17 March 2020; Accepted 18 March 2020
Available online 30 March 2020

cancers.

2. ChIP-seq analysis workflow

In this section, we describe the step-by-step workflow of a typical ChIP-seq analysis (Fig. 1). Also, see our previous review [16] for details and considerations for each step.

2.1. Environmental setup

Computational tools for NGS analysis are written in various computational languages such as C++, R, Python, Java, and Perl. Each language requires a different setup method. While most are executed on Linux systems, Mac terminal and Windows Subsystem for Linux (WSL) can also be used (note that some specific errors might occur due to different library names and dependencies). One major upcoming issue is that Python2 will not be maintained after 2020 (https://www.python.org/doc/sunset-python-2/). There are several tools that require Python2 but have not been updated to Python3 (e.g., Peakzilla [17] and ChromTIME [18]). In the near future, users will have to consider replacing these packages for newer alternatives. If users want to keep using these older applications (e.g., because current analysis pipelines for big projects are difficult to modify), virtual environments like Docker (https://www.docker.com/) or Singularity (https://sylabs.io/) provide a secure, isolated analysis environment. Several computational tools and analysis pipelines are released as Docker images, which are downloadable pre-compiled computational environments. These images remove difficult extraction and installation.

2.2. Downloading ChIP-seq data from public databases

Multiple public databases are available to download ChIP-seq data of histone modifications (Table 1). We recently published an epigenome database for human endothelial cells (entry 4 in Table 1), which contains 424 histone modification ChIP-seq and 67 RNA sequencing (RNA-seq) datasets obtained from nine blood vessel types from the human body [19]. Various data types are available (e.g., reads, mapfiles, bigwig files, and peak lists), that are suitable as ChIP-seq analysis tutorial data.

Table 1
Public ChIP-seq databases.

Database	URL	Reference
ENCODE portal	https://www.encodeproject.org/	[118]
ROADMAP epigenome database	http://www.roadmapepigenomics.org/	[12]
IHEC Data Portal	https://epigenomesportal.ca/ihec/	[119]
Epigenome database for human endothelial cells	https://rnakato.github.io/HumanEndothelialEpigenome/	[19]

2.3. Technical considerations of ChIP-seq analysis for histone modifications

The reliability of a ChIP analysis is governed by antibody quality, including specificity and signal-to-noise ratio (S/N) [20]. Since the false-positive enriched sites derived from nonspecific antibody-DNA binding may confound the analysis, unexpected ChIP-seq results should be validated using multiple antibodies [21].

While most ChIP-seq tools are designed for sharp peaks that are located at specific genomic positions, such as transcription start sites (TSSs), some histone modifications are associated with large genomic domains, resulting in broadly distributed enrichment regions [1]. H3K27me3 and H3K36me3 enrichments distribute across several hundred kilobases, while H3K9me3 peaks often expand to a few megabases. The enhancer markers, H3K27ac and H3K4me1, produce sharp peaks, but sometimes construct broadly enriched regions called “super enhancers” [22]. H3K4me3 promoter markers can also cover broad domains in mouse oocytes [23]. This peak shape and broadness variation

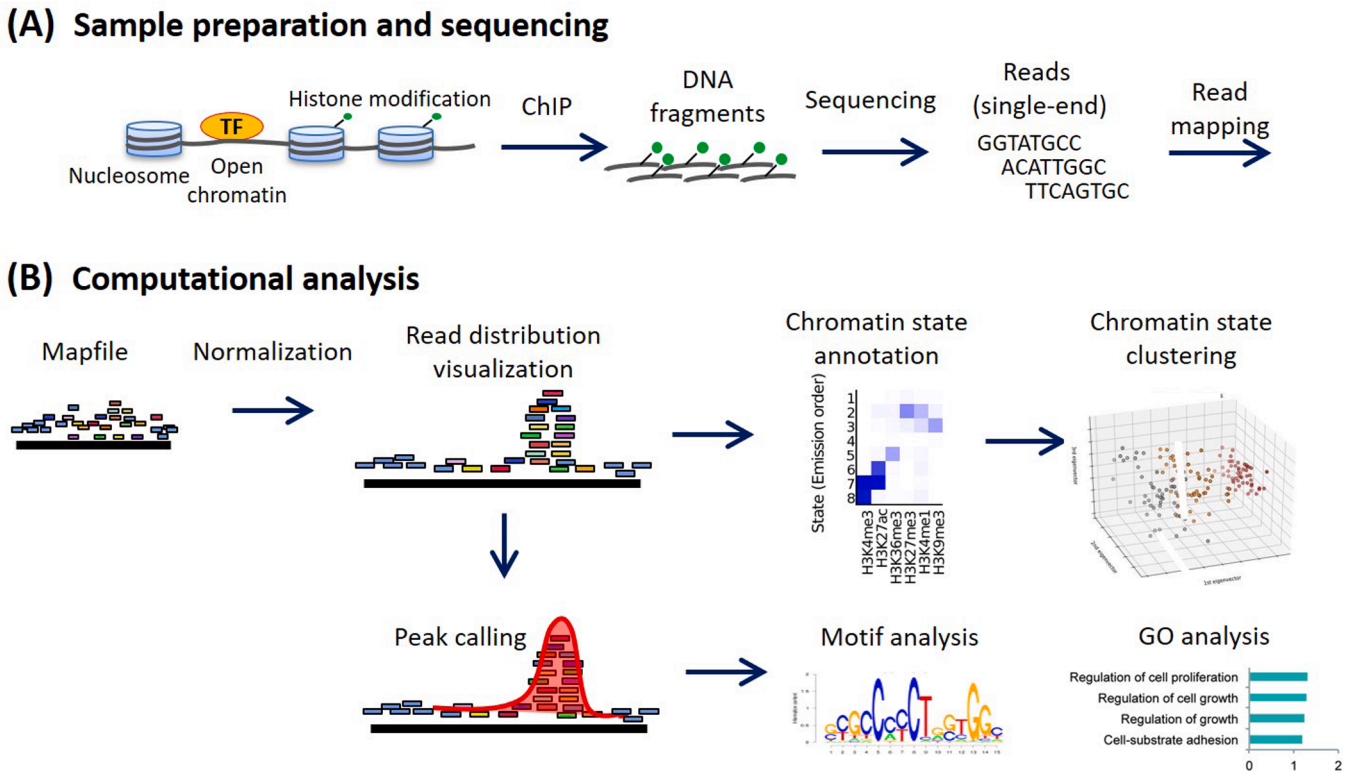

Fig. 1. ChIP-seq analysis workflow. (A) Sample preparation and sequencing. (B) Computational analysis in a canonical ChIP-seq analysis. Various analyses are implemented using normalized read distribution.

affects the choice of optimal computational tools. For example, ROSE [24] is specifically designed to detect super-enhancer sites, which congregate multiple enhancer sites close together. Music [25] can estimate the average sample peak width to be investigated.

2.4. Read mapping

The sequenced reads (FASTQ or CSFSATQ format) are mapped using tools such as Bowtie [26], Bowtie2 [27], or BWA [28]. Bowtie2 and BWA can consider indels (insertions and deletions) by gapped alignments, which is appropriate for long and/or paired-end reads (see [29] for a comparison of mapping tools and parameters). There are several output formats for map files, such as SAM, BAM, CRAM and tagAlign. While the BAM format is the most widely used so far, the more space-efficient CRAM format is maturing and will likely be the next standard (https://www.ga4gh.org/cram/). After alignment, reads mapped to the same genomic positions are filtered as redundant reads, and the remaining nonredundant reads are used for analysis.

2.5. Peak calling

The peak-calling step identifies significantly enriched loci (peaks) in the genome. Peak-calling results are generally returned in BED format. Although ChIP-seq peaks do not have strand information, it can be estimated from the gene information when focusing on the histone marks that are enriched around TSS, for instance. While MACS2 [30] is the most commonly used peak-calling tool, numerous peak-calling tools were recently developed (see [16,31,32] for reviews). However, no tool can achieve 100% accuracy. Therefore, a practical strategy is to obtain a large number of peaks with a relaxed threshold that contain true positives and noise, and then extract subgroups using another way to improve specificity, e.g. selecting consistent signal among biological replicates using the Irreproducible Discovery Rate (IDR) [20].

2.6. ChIP-seq data quality assessment

Quality check (QC) of ChIP-seq samples is critical to judge whether sequencing data are of high quality and suitable for further analyses. Various quantitative QC measures have been developed [16,20]. Among them, the particularly important metrics are:

- Mapping ratio, which reflects read quality and the proportion of sequenced reads that are derived from true genomic DNA. For example, the mapping ratio for samples sequenced by Illumina HiSeq System (e.g., Hiseq2500) should be over 80%. The exception is a sample for non-DNA-binding proteins such as IgG, which often has a lower mapping ratio (~60%).
- Read depth (the number of nonredundant mapped reads). Sufficient read depth depends on the genome size and the antibody S/N ratio [1]. The ENCODE consortium suggested at least 10 million uniquely mapped reads as a minimum to analyze sharp-mode peaks of human samples [20]. Broad histone marks often have weaker S/N and require more reads (e.g., >40 million for human) as a practical minimum for peak calling [33].
- Library complexity (the proportion of nonredundant reads). It ranges from 0 to 1.0, and the ENCODE consortium suggested the complexity > 0.8 for 10 million mapped reads [20]. Lower values (less than 0.6) indicate excessive PCR amplification from a small amount of initial DNA [16].
- The normalized strand coefficient (NSC, obtained by SSP [34]), a S/N indicator for both sharp and broad marks (phantompeakqualtools [20] can only calculate NSC for sharp marks). In-depth validation using>1,000 publicly available ChIP-seq datasets for multiple species suggested that the recommended threshold value is NSC > 5.0 and NSC > 1.5 for sharp and broad marks, respectively [34]. Input samples should have a low S/N and therefore NSC values should be <2.0.
- Background uniformity (Bu) [34]. Bu reflects the read distribution bias in background regions and ranges from 0 to 1.0. Low values (less than 0.8) suggest that the read distribution is more congregated or biased than expected, resulting in numerous false positives in obtained peaks [35]. For the genome that has extensive copy-number variations (e.g., MCF-7 cells), a relaxed threshold value (>0.6) is desirable.
- GC summit bias, reflecting biases during immunoprecipitation and PCR amplification [35]. In general, the GC summit of typical ChIP-seq data becomes similar to the reference genome (e.g., ~50% for human [19]). Unexpected GC-rich summit (e.g., over 60% for human) is often manifested due to PCR amplification biases [35] and/or false-positive peaks derived from 'hyper-ChIPable' regions associated with CpG islands [36].

Fig. 2 shows the QC metrics for six histone modifications in 127 cell types from the ROADMAP project, consisting of sharp (H3K27ac, H3K4me1 and H3K4me3) and broad marks (H3K27me3, H3K36me3, and H3K9me3), along with input samples. Though the number of peaks can be an initial indicator of a successful ChIP experiment, peak number is not suitable for comparison among ChIP samples because it strongly depends on the peak calling threshold. Additionally, a single enriched region can be often divided into multiple subregions, particularly in broad marks [37]. For example, H3K4me3 has the lowest peak number distribution, but the largest SSP-NSC distribution, indicative of a smaller number of stronger peaks (Fig. 2A). Fig. 2A also shows that there are a few samples that are highly GC-rich (>60%) or have a low Bu score (<0.8). This result suggests that the obtained peaks from the samples are less reliable and should be handled with care in the comparative analysis. Finally, inter-sample comparison using correlation heatmaps based on read distribution (Fig. 2B) is a good way to identify other suspicious samples, such as mislabeled samples or other human-introduced errors.

2.7. Visualization

Having developed various statistical methods and quality metrics for ChIP-seq data, visual inspection of read distribution is effective to intuitively assess and analyze the obtained data, e.g., detecting suspicious peaks derived from hyper-ChIPable regions [36]. For that, interactive visualization tools such as Integrated Genome Viewer (IGV) [38] or SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/) are available. Several web servers (e.g. UCSC genome browser [39] and WashU Epigenome Browser [40]) can integrate the obtained ChIP-seq results with other annotation data, such as evolutionary conservation and gene expression in various tissues.

2.8. Normalization for comparative analysis

Read normalization is essential to mitigate technical variance before comparative analysis [35]. Simple total read normalization is commonly used, which scales the sample read number to be the same. The underlying assumption is that the difference in mapped reads among samples is sufficiently smaller than the total read number. This assumption is not always satisfied, and therefore, several methods have been developed to identify differentially enriched regions between two conditions, some of which are specifically designed for histone modification data [41,42]. Since the obtained results vary considerably among tools due to the underlying statistical assumptions, the choice of method will crucially impact the outcome [43].

Quantitative comparison across more than two groups is more complicated. When the expected S/N value is similar among samples, statistical methods for differential gene expression analysis can be used [44]. It is also possible to utilize quantile normalization [19] when the S/N for most common peaks is similar among samples (e.g., a single

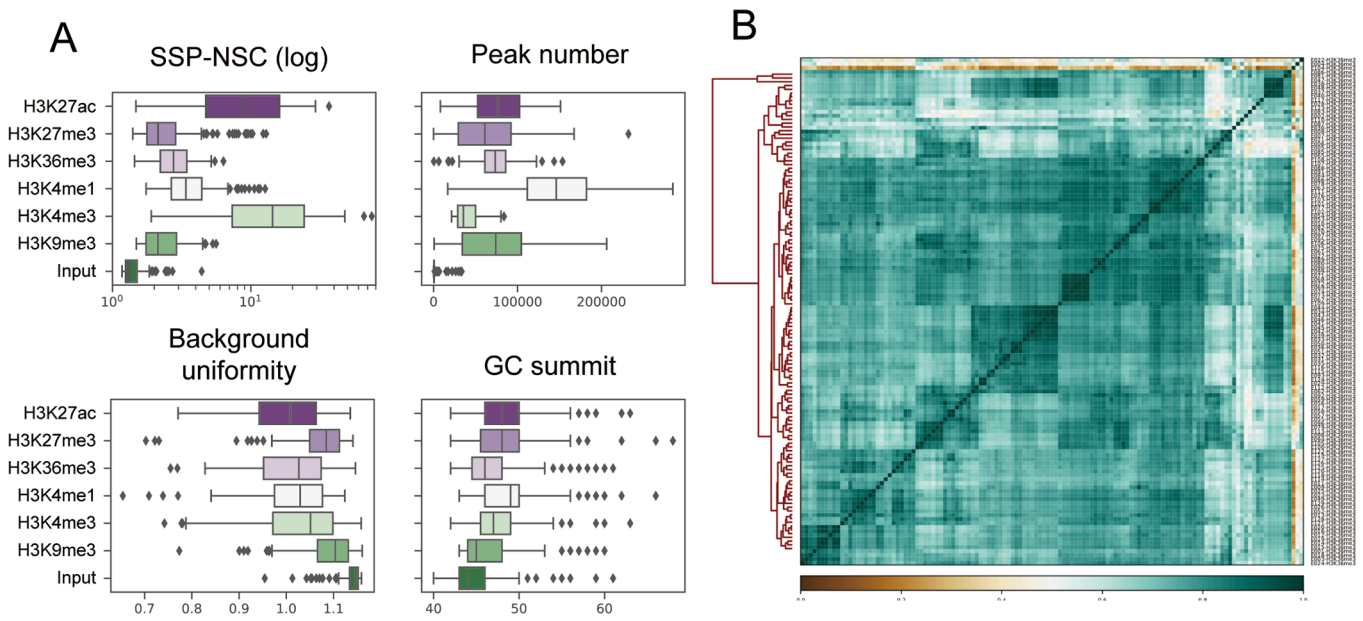

Fig. 2. QC analysis example using ROADMAP histone modification data. (A) Distribution of four QC scores for six histone modifications and input samples: NSC (log-scale) obtained by SSP; Peak number obtained by MACS2; Background uniformity obtained by SSP; and GC summits obtained by DROMPAplus. Data are from reference [34]. (B) Pearson-correlation heatmap of H3K36me3 read distribution for 117 cell types from Roadmap epigenome database, generated by deepTools [116]. Some samples with low correlation are highlighted in beige.

antibody for all samples). If the S/N highly varies among samples (e.g., between with and without stimulation), consider spike-in analysis (also called calibration analysis) [45,46]. This method is a wet-based solution that adds the same amount of DNA from a different species to all samples before or after immunoprecipitation and estimates the weight coefficient based on the number of derived reads. In contrast to computational normalization methods that are limited to relative differences, spike-in ChIP-seq enables investigation of absolute-level differences [16]. However, quantitative ChIP-seq comparisons are still often confounded by intrinsic noisiness and variability caused by multi-step sample preparation, even after normalization [43]. In this case, simple binary comparisons (identifying common or unique peaks) might be desirable, though some false positiveS/Negatives will likely occur in the obtained results.

2.9. ChIP-seq analysis with DROMPAplus

Fig. 3 shows the normalized read distribution of histone modifications generated by DROMPAplus (see "Software availability" section). This is an update of DROMPA [47,48], a ChIP-seq pipeline tool that meets various needs, including quality check, normalization, statistical analysis, and visualization of multiple ChIP-seq samples. DROMPA can be used for any species whose genomic sequence is available, and has been applied to various ChIP-seq studies for human [49–51], mouse [52,53], chicken [54] and yeast [55,56]. It outputs the visualization in conventional PDF format, which is preferable for many users, especially when sharing results (e.g., on a cloud storage) with other collaborators who do not have a strong bioinformatics background, because no additional programs are required.

DROMPAplus has many valuable features. First, it accepts various input map file formats, including CRAM. Second, DROMPAplus can visualize two samples in one line (Fig. 3A), which delineates the co-occurrence (e.g., H3K4me3 and H3K27ac) and exclusivity (e.g., H3K27me3 and H3K36me3) of read enrichment, with chromatin loops obtained from 3C-based assays (see section 3.2). Third, automatic estimation of fragment length from single-end reads using SSP [34]. Forth, it is highly customizable for track heights, axis limits, and display features. For example, the software can depict ChIP/input enrichment with highlighting the enriched regions (Fig. 3B) with linear or log-scale. Fifth, DROMPAplus can support spike-in normalization as well as total read normalization. Finally, it is completely rewritten in C++, which is more flexible than C and computationally faster than Python and R.

2.10. Functional analysis

Motif analysis investigates the sequence specificity inherent in called peaks or specific epigenome regions (e.g., enhancer sites), and estimates the likely transcription factor binding sites within identified regions [57]. Generally, motif analysis methods can be classified into two types: *de novo* motif discovery that identifies potential new binding motifs for unknown factors appearing in a large fraction of peaks [58]; and motif scanning that estimates and ranks the similarity of supplied DNA sequences against all known canonical motifs within a database [59]. ChIP-seq peaks can also be used in functional enrichment analysis. This analysis binarily labels or quantitatively ranks nearby genes as potential targets and groups them by gene ontology or KEGG pathway [60–62].

2.11. Chromatin-state annotation

Chromatin-state annotation, also called semi-automated genomic annotation (SAGA), classifies all genomic regions by characteristic epigenomic patterns, such as promoters, enhancers, transcribed regions, and repressed regions, using an unsupervised machine-learning approach [63]. Obtained clusters are manually annotated as chromatin states. Typical region-specific analysis (e.g., enhancer analyses [19,64]) narrows down the target genomic regions to be investigated. In contrast, chromatin-state annotation segments the genome and assigns chromatin states to whole-genome regions using a hidden Markov model [65–67] or a dynamic Bayesian network [68]. In this analysis, the biologically optimal number of states is unknown and must be experimentally defined. That is, more abundant states cause difficulty when interpreting obtained clusters. In fact, numerous states may not capture sufficiently distinct epigenetic characters [69]. Thus, up to 15 states may be appropriate.

The obtained chromatin states are further extended for various downstream analyses. For example, ChromDiff [70], EpiCompare [71],

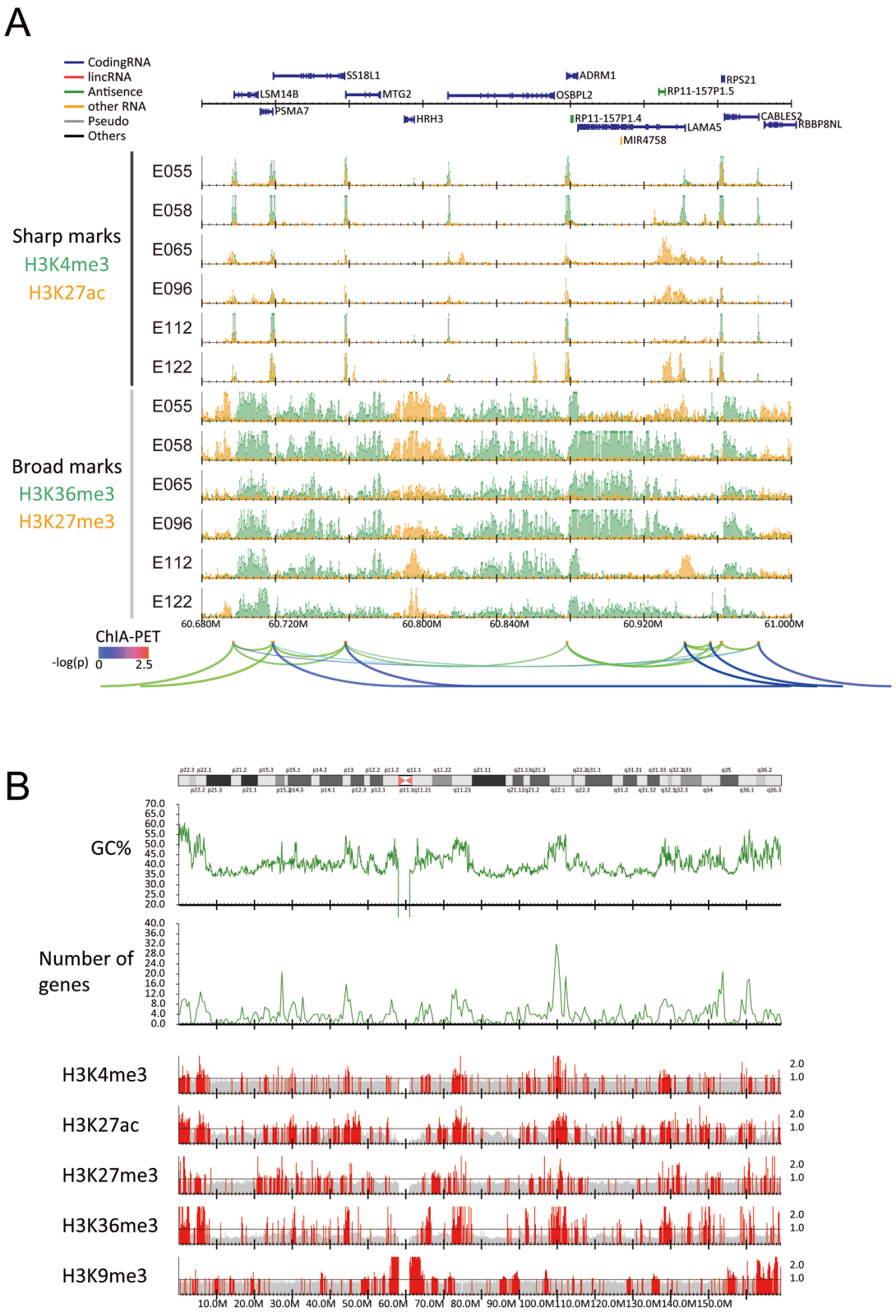

Fig. 3. Visualizing multiple ChIP-seq samples using DROMPAplus. (A) Normalized read distribution of sharp histone marks (top) and broad marks (bottom) for E055 (foreskin fibroblasts), E058 (foreskin keratinocytes), E065 (aorta), E096 (lung), E112 (thymus), and E122 (Human Umbilical Vein Endothelial Cells: HUVEC) cells obtained from the Roadmap epigenome database. Two histone marks are depicted in a single track with different colors (green and orange). RNA Pol II-mediated chromatin loops (based on ChIA-PET data) in HUVECs [117] are represented by arches. (B) Visualization of five core histone modifications (E122) for human chromosome 7. The ChIP/input enrichment distribution (100-kb windows), the GC contents (500 kb windows) and gene numbers (500 kb windows) are plotted. Windows with ChIP/control > 1 are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and ChromDet [72] combine and cluster derived epigenomic landscapes across multiple cell types to explore tissue or cell type-specific epigenomic regions. A probabilistic clustering approach is also adopted to capture chromatin state dynamics across multiple cell lines [73] or time points [18,74]. Graph-based regularization (GBR) integrates chromatin interaction information for chromatin-state annotation [63]. Generated chromatin state information is then used to interpret individual genetic variations [75,76] and understand epigenetic variation in evolution [77].

3. Advanced applications

Because abundant ChIP-seq data are available for several well-studied cell types, it is useful to leverage information from these cell types to infer genome dynamics or to annotate the epigenetic landscape of other cell types with fewer additional experiments. Increasing evidence suggests that epigenetic information is highly correlated with, and can be used to predict, gene expression and chromosomal conformation. In this section, we briefly describe tools for advanced applications of ChIP-seq analysis for histone modifications, which are more experimental and theoretical than the tools introduced in section 2.

3.1. Gene expression prediction from the epigenome

Various machine learning-based approaches have been developed to quantitatively infer gene expression levels based on the epigenetic information obtained by ChIP-seq experiments. For instance, Karlic *et al.* applied a linear regression model to histone modification enrichments at promoter sites to predict gene expression in CD4 + T-cells [78]. They utilized nineteen histone modifications and suggested that as few as three promoter site modifications are sufficient to model gene expression [78]. Dong *et al.* used non-linear models, such as multivariate adaptive regression splines (MARS) and random forests, to map eleven histone modifications and DNase I hypersensitivity in seven human cell lines [79] and successfully predicted gene expression level (Pearson coefficient $r = 0.83$ with observed data). These models simply consider the epigenetic pattern at promoter sites and do not account for enhancer site information. In contrast, DeepExpression [80] utilizes HiChIP data [81], a high-throughput technique for capturing protein-centric chromosome loops, to consider enhancers and enhancer-promoter interactions. There are also several tools that use convolutional neural networks (CNN) to predict gene expression [82] or differential gene regulation patterns [83]. See reference [82] for a detailed discussion regarding the comparison of these gene expression prediction programs. Considering that the preparation of a single RNA-seq sample requires relatively lower cost compared with that of ChIP-seq samples of multiple histone modifications and HiChIP data, the main purpose of these studies is to elucidate the combinatorial roles of histone modifications in gene regulation, rather than the prediction of gene expression level itself.

3.2. Prediction of chromatin interactions from epigenome data

Because recent evidence suggests that single nucleotide polymorphisms (SNPs) in enhancers can cause genetic diseases and cancer [84,85], there is a great demand for genome-wide analysis to characterize the role of enhancers in specific cell lines. However, genome-wide pairing of enhancers and target genes is not a trivial task. Indeed, enhancers do not necessarily regulate the nearest genes, and some enhancers are distant from TSSs [86]. While Chromosome Conformation Capture (3C) assays, such as Hi-C [87], HiChIP [81], and ChIA-PET [88], are available to quantify spatial proximity across an entire genome, computational tools for pairing enhancers and target genes keep evolving. Hariprakash and Ferrari classified gene-enhancer pairing tools into four categories [89]: correlation-based, supervised learning-based, regression-based, and score-based. The key differences are "whether multiple enhancers are considered for each gene" and "whether multiple epigenetic data are considered for each enhancer/promoter site". Correlation-based methods estimate the interaction strength for all-by-all enhancer-promoter pairs, while regression-based methods assume that multiple enhancers contribute to a single gene. Supervised learning-based and score-based methods can combine multiple ChIP-seq datasets and other information types for each site (e.g., evolutionary conservation). While these tools focus on enhancer-promoter interactions, there are many other chromatin interactions, such as enhancer-enhancer loops and weak chromatin aggregation via phase separation [90]. In contrast, CITD [91] and DRAGON [92] comprehensively decipher three-dimensional genome organization from epigenetic data using wavelet transformation and potential energy functions, respectively.

These statistical approaches aim to find consistent patterns in epigenetic data associated with spatial chromatin contacts and predict them without any previous knowledge of genomic architecture. The limitation of these methods is that genomic interactions are considered as qualitative, rather than quantitative, despite their dynamic nature [93]. It was also reported that the current methods involve a training bias due to sharing information of genomic architecture between training and validation datasets [94]. Nevertheless, because the number of tools is rapidly growing, future methods might achieve sufficient accuracy that identifying enhancer-promoter interactions via 3C-based data will be unnecessary.

3.3. Data imputation: Reconstruction and denoising ChIP-seq data

One analytical challenge in large-scale ChIP-seq analysis arises from biases and batch effects in ChIP-seq data. Because machine-learning approaches are sensitive to noise in training data, it is unavoidable that some ChIP-seq samples will be identified as moderate quality or rejected as low-quality data (resulting in missing data), especially in cases where multiple laboratories were responsible for data acquisition (e.g., the large consortium project). If biological samples are precious (e.g. primary cells and clinical samples), it might be practically difficult to collect more samples. In this case, "data imputation" methods may be appropriate. These methods utilize many epigenetic data from other closely related cell types for data de-noising or reconstruction.

"Data de-noising" aims to improve existing ChIP-seq sample quality by identifying and removing noise from the data. For example, Coda [95] encodes a generative noise process and recovers signals in ChIP-seq data using convolutional neural networks. "Data reconstruction" aims to generate missing ChIP-seq data from the large dataset *in silico*. ChromImpute [96] is a pioneering tool that trains a regression tree to infer signal from each missing experiment using the ten most correlated cell types. PREDICTD [97] and Avocado [98] leverage tensor decomposition to impute multiple ChIP-seq data simultaneously. Several prediction tools for transcription factor binding sites are also proposed [99–101].

These data imputation approaches are potential computational alternatives to real ChIP-seq experiments, and might open the way to collect epigenomic data for all possible cell types and environmental conditions that are clearly impossible in biology. At the present stage, there are the limitations for the prediction of sample-specific signals that do not correlate with the other samples and for the incorporation of genetic variation [96]. Because 'a prior expectation of signal' by the imputation across the genome is informative even when high-quality datasets are available [96], the combined use of observed and imputed data is a practically good strategy. Although this approach is computationally challenging, publicly available high-quality data from diverse cell types (Table 1) encourages to accomplish that.

4. Single-cell ChIP-seq analysis

Recent evidence suggests many cells types, including normal immune cells, serve an essential accessory function in complex tissues and tumors [102]. To elucidate this cellular heterogeneity and cell fate

trajectories in developmental processes, various single-cell assays have been developed [103]. Among them, scChIP-seq enables genome-wide profiling of histone modifications and other chromatin-binding proteins at single-cell resolution from low-input samples. Recently, multiple approaches for single-cell labeling and ChIP-seq library preparation have been developed (Table 2) which use microfluidic systems, Tn5 transposase tagmentation, and ChIP-free strategies.

4.1. Microfluidic system-based analysis

The first scChIP-seq method, scDrop-ChIP [104], uses microfluidic systems for cell labeling combined with canonical ChIP methods to generate ~ 800 non-duplicated reads per cell. The more recently developed droplet microfluidic method [105] provides higher resolution, producing ~ 10,000 non-duplicated reads per cell. The limitation of these methods is that the specialized microfluidic devices are not usually available for most laboratories.

4.2. Tagmentation-based analysis

Tagmentation-based library preparation using Tn5 transposase has been widely used for various NGS assays, including ChIP-seq. sc-itChIP-seq [106] employs tagmentation for single-cell labeling and library preparation before the canonical ChIP experiment. This method generates ~ 9000 non-duplicated reads per cell. Because the experimental procedure is similar to the canonical ChIP-seq method, this method is much easier to use than scDrop-ChIP.

4.3. ChIP-free methods

Several ChIP-free strategies have been developed for scChIP-seq. Single-cell chromatin immunocleavage sequencing (scChIC-seq) [107] and single-cell uliCUT&RUN [108] are based on the CUT&RUN method [109] that employs MNase and protein A fusion proteins to detect cleaved target sites with a specific antibody. These methods generate ~ 4,100 non-duplicated reads per cell and require several canonical steps for library preparation. However, these methods are limited by low read-mapping rates (~6%). Three similar methods, called CUT&Tag [110], ACT-seq [111], and CoBATCH [112], have been developed. These methods use a Tn5 transposase and protein A fusion protein. During library preparation, the primary antibody is captured by the fusion protein after binding the target protein on chromosomes. Then, Tn5 transposase is activated for tagmentation at the protein binding sites. The advantage of these methods is that protein binding site detection and library preparation are performed simultaneously, which drastically reduces experimental procedures and time. Further, these methods are less subject to technical biases introduced by an immunoprecipitation step. Moreover, these methods show ~ 97% mapping rates and generate ~ 12,000 non-duplicated reads per cell. Thus, this ChIP-free method has potential for high-throughput and high-quality scChIP-seq analysis.

Finally, chromatin integration labelling followed by sequencing (ChIL-seq) [113] is another ChIP-free method that is based on immunostaining rather than ChIP. The method uses a secondary antibody probe conjugated with dsDNA, which contains a T7 RNA polymerase promoter, an NGS adapter sequence, and a Tn5 binding sequence. After capturing the first antibody, the probe DNA sequence is integrated into the target binding sites by Tn5 transposase. Then, the integrated regions are amplified by *in situ* transcription, followed by RNA purification and library preparation. The method can be used for single-cell analysis, but likely needs several optimizations to achieve high-throughput sequencing.

Additional scChIP-seq methods will be developed in future, such as simultaneous detection of multiple histone modifications and/or other chromatin-binding proteins. These advances will enable to capture co-localization of gene-regulating factors on chromosomes in each cell.

5. Concluding remarks

In this review, we discussed ChIP-seq analysis concepts and methods for histone modifications mainly from the computational aspect. We presented a step-by-step workflow for canonical analysis, from quality assessment to chromatin-state annotation, highlighting key points associated with each step. Then, we discussed several advanced ChIP-seq applications that use machine-learning approaches. Because of the increasing availability of epigenomic data from large consortia and other projects, tools for identifying novel genomic features using these data will continue to gain attention.

Almost all the methods introduced in the “Advanced applications” section use supervised machine-learning approaches, e.g., deep CNN. One limitation is that these methods require many samples from each cell line to develop training data, resulting in heavy demand for ChIP-seq data. These approaches also require large computational power and abundant disk storage in the analysis environment. Cloud computing could be one solution to overcome this limitation. Using cloud computing, researchers can share petabyte data sets and computational environments, which dramatically reduces the computational cost for the large-scale re-analysis of public data [114]. Another limitation is the accuracy of the input data. The training data derived from raw samples are often obtained from various NGS assays and contain some or abundant technical/biological noise, which hampers effective training. While data imputation methodology aims to partly overcome this limitation, it is necessary to develop experimentally validated “gold standard datasets” for training data to assess tool performance.

A next challenge is integration with other NGS assays (e.g., DNA methylation, accessible regions, and 3D genome folding). Some tools introduced in this review aim to integrate multiple datasets. Of note, three-dimensional genome information, such as Hi-C, is particularly important because it enables us to consider enhancer-promoter interactions and topologically associating domains (TADs) information, which are closely related to epigenetic characteristics [115]. Finally, we discussed recently developed methodologies for scChIP-seq analysis. Multiple scChIP-seq protocols developed recently prompt the development of corresponding computational methodologies. Though it is difficult to directly apply tools designed for bulk ChIP-seq to scChIP-seq data [16], it is essential to make full use of the knowledge accumulated from past bulk ChIP-seq analyses.

Table 2
scChIP-seq methods.

Method	Strategy	Cell condition	Mapping rate (average)	Non-duplicated reads (average)	Reference
scDrop-ChIP	ChIP and microfluidic system	Native	70%	796	[104]
sc-itChIP-seq	ChIP and tagmentation	Native and fixed	94%	9,016	[106]
scChIC-seq	ChIP-free (MNase cleavage)	Native	6%	4,079	[107]
CUT&Tag	ChIP-free (tagmentation)	Native	97%	10,104	[110]
ACT-seq	ChIP-free (tagmentation)	Native	83%	2,497	[111]
CoBATCH	ChIP-free (tagmentation)	Native and fixed	94%	12,000	[112]

6. Software availability

DROMPAplus is written in C++ and runs from a single launch command on conventional Linux systems. The source code, manual, and precompiled Docker image are available on GitHub (https://github.com/rnakato/DROMPAplus).

Author contributions

RN conceived this project and organized the manuscript. TS drafted the single-cell analysis section of the manuscript. Both authors read and approved the final manuscript.

Acknowledgments

This work was supported by grants-in-aid for Scientific Research (17H06331 to R.N.).

References

[1] P.J. Park, ChIP-seq: advantages and challenges of a maturing technology, Nat. Rev. Genet. 10 (10) (2009) 669–680.
[2] T.S. Furey, ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions, Nat. Rev. Genet. 13 (12) (2012) 840–852.
[3] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh, L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B. E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types, Nature 473 (7345) (2011) 43–49.
[4] K. Yamaguchi, M. Hada, Y. Fukuda, E. Inoue, Y. Makino, Y. Katou, K. Shirahige, Y. Okada, Re-evaluating the Localization of Sperm-Retained Histones Revealed the Modification-Dependent Accumulation in Specific Genome Regions, Cell Rep 23 (13) (2018) 3920–3932.
[5] T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E.S. Lander, B.E. Bernstein, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (7153) (2007) 553–560.
[6] W. Xie, M.D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, J.W. Whitaker, S. Tian, R.D. Hawkins, D. Leung, H. Yang, T. Wang, A.Y. Lee, S.A. Swanson, J. Zhang, Y. Zhu, A. Kim, J.R. Nery, M.A. Urich, S. Kuan, C.A. Yen, S. Klugman, P. Yu, K. Suknuntha, N.E. Propson, H. Chen, L.E. Edsall, U. Wagner, Y. Li, Z. Ye, A. Kulkarni, Z. Xuan, W.Y. Chung, N.C. Chi, J.E. Antosiewicz-Bourget, I. Slukvin, R. Stewart, M.Q. Zhang, W. Wang, J.A. Thomson, J.R. Ecker, B. Ren, Epigenomic analysis of multilineage differentiation of human embryonic stem cells, Cell 153 (5) (2013) 1134–1148.
[7] J. Zhu, M. Adli, J.Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P.L. De Jager, D.A. Bennett, J.A. Houmard, D.M. Muoio, T. T. Onder, R. Camahort, C.A. Cowan, A. Meissner, C.B. Epstein, N. Shoresh, B. E. Bernstein, Genome-wide chromatin state transitions associated with developmental and environmental cues, Cell 152 (3) (2013) 642–654.
[8] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D.A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, I. Amit, Immunogenetics, Chromatin state dynamics during blood formation, Science 345 (6199) (2014) 943–949.
[9] Z. Zhao, A. Shilatifard, Epigenetic modifications of histones in cancer, Genome Biol. 20 (1) (2019) 245.
[10] K.K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W.J. Housley, S. Beik, N. Shoresh, H. Whitton, R.J. Ryan, A.A. Shishkin, M. Hatan, M.J. Carrasco-Alfonso, D. Mayer, C.J. Luckey, N.A. Patsopoulos, P.L. De Jager, V.K. Kuchroo, C. B. Epstein, M.J. Daly, D.A. Hafler, B.E. Bernstein, Genetic and epigenetic fine mapping of causal autoimmune disease variants, Nature 518 (7539) (2015) 337–343.
[11] W. Sun, J. Poschmann, R. Cruz-Herrera Del Rosario, N.N. Parikshak, H.S. Hajan, V. Kumar, R. Ramasamy, T.G. Belgard, B. Elanggovan, C.C.Y. Wong, J. Mill, D.H. Geschwind, S. Prabhakar, Histone Acetylome-wide Association Study of Autism Spectrum Disorder, Cell 167(5) (2016) 1385-1397 e11.
[12] C. Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, V. Amin, J. W. Whitaker, M.D. Schultz, L.D. Ward, A. Sarkar, G. Quon, R.S. Sandstrom, M. L. Eaton, Y.C. Wu, A.R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C.B. Epstein, E. Gjoneska, D. Leung, W. Xie, R.D. Hawkins, R. Lister, C. Hong, P. Gascard, A.J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R.S. Hansen, R. Kaul, P.J. Sabo, M.S. Bansal, A. Carles, J.R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A.R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S.J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R.C. Sallari, K. T. Siebenthall, N.A. Sinnott-Armstrong, M. Stevens, R.E. Thurman, J. Wu, B. Zhang, X. Zhou, A.E. Beaudet, L.A. Boyer, P.L. De Jager, P.J. Farnham, S. J. Fisher, D. Haussler, S.J. Jones, W. Li, M.A. Marra, M.T. McManus, S. Sunyaev, J.A. Thomson, T.D. Tlsty, L.H. Tsai, W. Wang, R.A. Waterland, M.Q. Zhang, L. H. Chadwick, B.E. Bernstein, J.F. Costello, J.R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J.A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes, Nature 518 (7539) (2015) 317–330.
[13] S. Heinz, C.E. Romanoski, C. Benner, C.K. Glass, The selection and function of cell type-specific enhancers, Nat Rev Mol Cell Biol 16 (3) (2015) 144–154.
[14] E.P. Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57–74.
[15] H.G. Stunnenberg, C. International Human Epigenome, M. Hirst, The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery, Cell 167(7) (2016) 1897.
[16] R. Nakato, K. Shirahige, Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation, Brief Bioinform. 18 (2) (2017) 279–290.
[17] A.F. Bardet, J. Steinmann, S. Bafna, J.A. Knoblich, J. Zeitlinger, A. Stark, Identification of transcription factor binding sites from ChIP-seq data at high resolution, Bioinformatics 29 (21) (2013) 2705–2713.
[18] P. Fiziev, J. Ernst, ChromTime: modeling spatio-temporal dynamics of chromatin marks, Genome Biol. 19 (1) (2018) 109.
[19] R. Nakato, Y. Wada, R. Nakaki, G. Nagae, Y. Katou, S. Tsutsumi, N. Nakajima, H. Fukuhara, A. Iguchi, T. Kohro, Y. Kanki, Y. Saito, M. Kobayashi, A. Izumi-Taguchi, N. Osato, K. Tatsuno, A. Kamio, Y. Hayashi-Takanaka, H. Wada, S. Ohta, M. Aikawa, H. Nakajima, M. Nakamura, R.C. McGee, K.W. Heppner, T. Kawakatsu, M. Genno, H. Yanase, H. Kume, T. Senbonmatsu, Y. Homma, S. Nishimura, T. Mitsuyama, H. Aburatani, H. Kimura, K. Shirahige, Comprehensive epigenome characterization reveals diverse transcriptional regulation across human vascular endothelial cells, Epigenetics Chromatin 12 (1) (2019) 77.
[20] S.G. Landt, G.K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J.B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A.J. Hartemink, M. M. Hoffman, V.R. Iyer, Y.L. Jung, S. Karmakar, M. Kellis, P.V. Kharchenko, Q. Li, T. Liu, X.S. Liu, L. Ma, A. Milosavljevic, R.M. Myers, P.J. Park, M.J. Pazin, M. D. Perry, D. Raha, T.E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M.Y. Tolstorukov, K.P. White, S. Xi, P.J. Farnham, J. D. Lieb, B.J. Wold, M. Snyder, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, Genome Res. 22 (9) (2012) 1813–1831.
[21] J.W. Ho, Y.L. Jung, T. Liu, B.H. Alver, S. Lee, K. Ikegami, K.A. Sohn, A. Minoda, M.Y. Tolstorukov, A. Appert, S.C. Parker, T. Gu, A. Kundaje, N.C. Riddle, E. Bishop, T.A. Egelhofer, S.S. Hu, A.A. Alekseyenko, A. Rechtsteiner, D. Asker, J. A. Belsky, S.K. Bowman, Q.B. Chen, R.A. Chen, D.S. Day, Y. Dong, A.C. Dose, X. Duan, C.B. Epstein, S. Ercan, E.A. Feingold, F. Ferrari, J.M. Garrigues, N. Gehlenborg, P.J. Good, P. Haseley, D. He, M. Herrmann, M.M. Hoffman, T. E. Jeffers, P.V. Kharchenko, P. Kolasinska-Zwierz, C.V. Kotwaliwale, N. Kumar, S. A. Langley, E.N. Larschan, I. Latorre, M.W. Libbrecht, X. Lin, R. Park, M.J. Pazin, H.N. Pham, A. Plachetka, B. Qin, Y.B. Schwartz, N. Shoresh, P. Stempor, A. Vielle, C. Wang, C.M. Whittle, H. Xue, R.E. Kingston, J.H. Kim, B.E. Bernstein, A. F. Dernburg, V. Pirrotta, M.I. Kuroda, W.S. Noble, T.D. Tullius, M. Kellis, D. M. MacAlpine, S. Strome, S.C. Elgin, X.S. Liu, J.D. Lieb, J. Ahringer, G.H. Karpen, P.J. Park, Comparative analysis of metazoan chromatin organization, Nature 512 (7515) (2014) 449–452.
[22] D. Hnisz, B.J. Abraham, T.I. Lee, A. Lau, V. Saint-Andre, A.A. Sigova, H.A. Hoke, R.A. Young, Super-enhancers in the control of cell identity and disease, Cell 155 (4) (2013) 934–947.
[23] J.A. Dahl, I. Jung, H. Aanes, G.D. Greggains, A. Manaf, M. Lerdrup, G. Li, S. Kuan, B. Li, A.Y. Lee, S. Preissl, I. Jermstad, M.H. Haugen, R. Suganthan, M. Bjoras, K. Hansen, K.T. Dalen, P. Fedorcsak, B. Ren, A. Klungland, Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition, Nature 537 (7621) (2016) 548–552.
[24] J. Loven, H.A. Hoke, C.Y. Lin, A. Lau, D.A. Orlando, C.R. Vakoc, J.E. Bradner, T. I. Lee, R.A. Young, Selective inhibition of tumor oncogenes by disruption of super-enhancers, Cell 153 (2) (2013) 320–334.
[25] A. Harmanci, J. Rozowsky, M. Gerstein, MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework, Genome Biol 15 (10) (2014) 474.
[26] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol 10 (3) (2009) R25.
[27] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat Methods 9 (4) (2012) 357–359.
[28] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[29] M. Smolka, P. Rescheneder, M.C. Schatz, A. von Haeseler, F.J. Sedlazeck, Teaser: Individualized benchmarking and optimization of read mapping results for NGS data, Genome Biol 16 (2015) 235.
[30] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, X.S. Liu, Model-based analysis of ChIP-Seq (MACS), Genome Biol. 9 (9) (2008) R137.
[31] R. Thomas, S. Thomas, A.K. Holloway, K.S. Pollard, Features that define the best ChIP-seq peak calling algorithms, Brief Bioinform. 18 (3) (2017) 441–450.
[32] T.D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, L.L. Elo, A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments, BMC Genomics 10 (2009) 618.
[33] Y.L. Jung, L.J. Luquette, J.W. Ho, F. Ferrari, M. Tolstorukov, A. Minoda, R. Issner, C.B. Epstein, G.H. Karpen, M.I. Kuroda, P.J. Park, Impact of sequencing depth in ChIP-seq experiments, Nucleic Acids Res. 42 (9) (2014), e74.

[34] R. Nakato, K. Shirahige, Sensitive and robust assessment of ChIP-seq read distribution using a strand-shift profile, Bioinformatics 34 (14) (2018) 2356–2363.
[35] C.A. Meyer, X.S. Liu, Identifying and mitigating bias in next-generation sequencing methods for chromatin biology, Nat. Rev. Genet. 15 (11) (2014) 709–721.
[36] K. Wreczycka, V. Franke, B. Uyar, R. Wurmus, S. Bulut, B. Tursun, A. Akalin, HOT or not: examining the basis of high-occupancy target regions, Nucleic Acids Res. 47 (11) (2019) 5735–5745.
[37] A.F. Bardet, Q. He, J. Zeitlinger, A. Stark, A computational pipeline for comparative ChIP-seq analyses, Nat. Protoc. 7 (1) (2011) 45–61.
[38] H. Thorvaldsdottir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief Bioinform 14 (2) (2013) 178–192.
[39] C.M. Lee, G.P. Barber, J. Casper, H. Clawson, M. Diekhans, J.N. Gonzalez, A. S. Hinrichs, B.T. Lee, L.R. Nassar, C.C. Powell, B.J. Raney, K.R. Rosenbloom, D. Schmelter, M.L. Speir, A.S. Zweig, D. Haussler, M. Haeussler, R.M. Kuhn, W. J. Kent, UCSC Genome Browser enters 20th year, Nucleic Acids Res 48 (D1) (2020) D756–D761.
[40] D. Li, S. Hsu, D. Purushotham, R.L. Sears, T. Wang, WashU Epigenome Browser update 2019, Nucleic Acids Res 47 (W1) (2019) W158–W165.
[41] M. Heinig, M. Colome-Tatche, A. Taudt, C. Rintisch, S. Schafer, M. Pravenec, N. Hubner, M. Vingron, F. Johannes, histoneHMM: Differential analysis of histone modifications with broad genomic footprints, BMC Bioinf. 16 (2015) 60.
[42] H. Ashoor, C. Louis-Brennetot, I. Janoueix-Lerosey, V.B. Bajic, V. Boeva, HMCan-diff: a method to detect changes in histone modifications in cells with different genetic characteristics, Nucleic Acids Res 45 (8) (2017), e58.
[43] S. Steinhauser, N. Kurzawa, R. Eils, C. Herrmann, A comprehensive comparison of tools for differential ChIP-seq analysis, Brief Bioinform 17 (6) (2016) 953–966.
[44] X. Zhou, H. Lindsay, M.D. Robinson, Robustly detecting differential expression in RNA sequencing data using observation weights, Nucleic Acids Res 42 (11) (2014), e91.
[45] N. Bonhoure, G. Bounova, D. Bernasconi, V. Praz, F. Lammers, D. Canella, I. M. Willis, W. Herr, N. Hernandez, M. Delorenzi, X.C. Cycli, Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization, Genome Res. 24 (7) (2014) 1157–1168.
[46] B. Hu, N. Petela, A. Kurze, K.L. Chan, C. Chapard, K. Nasmyth, Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq, Nucleic Acids Res 43 (20) (2015), e132.
[47] R. Nakato, T. Itoh, K. Shirahige, DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data, Genes Cells 18 (7) (2013) 589–601.
[48] R. Nakato, K. Shirahige, Statistical Analysis and Quality Assessment of ChIP-seq Data with DROMPA, Methods Mol Biol 1672 (2018) 631–643.
[49] M.A. Deardorff, M. Bando, R. Nakato, E. Watrin, T. Itoh, M. Minamino, K. Saitoh, M. Komata, Y. Katou, D. Clark, K.E. Cole, E. De Baere, C. Decroos, N. Di Donato, S. Ernst, L.J. Francey, Y. Gyftodimou, K. Hirashima, M. Hullings, Y. Ishikawa, C. Jaulin, M. Kaur, T. Kiyono, P.M. Lombardi, L. Magnaghi-Jaulin, G.R. Mortier, N. Nozaki, M.B. Petersen, H. Seimiya, V.M. Siu, Y. Suzuki, K. Takagaki, J.J. Wilde, P.J. Willems, C. Prigent, G. Gillessen-Kaesbach, D.W. Christianson, F.J. Kaiser, L. G. Jackson, T. Hirota, I.D. Krantz, K. Shirahige, HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle, Nature 489 (7415) (2012) 313–317.
[50] J.Y. Bleuyard, M. Fournier, R. Nakato, A.M. Couturier, Y. Katou, C. Ralf, S. S. Hester, D. Dominguez, D. Rhodes, T.C. Humphrey, K. Shirahige, F. Esashi, MRG15-mediated tethering of PALB2 to unperturbed chromatin protects active genes from genotoxic stress, Proc Natl Acad Sci U S A 114 (29) (2017) 7671–7676.
[51] M. Minamino, M. Ishibashi, R. Nakato, K. Akiyama, H. Tanaka, Y. Kato, L. Negishi, T. Hirota, T. Sutani, M. Bando, K. Shirahige, Esco1 Acetylates Cohesin via a Mechanism Different from That of Esco2, Curr Biol 25 (13) (2015) 1694–1706.
[52] A. Sakai, R. Nakato, Y. Ling, X. Hou, N. Hara, T. Iijima, Y. Yanagawa, R. Kuwano, S. Okuda, K. Shirahige, S. Sugiyama, Genome-Wide Target Analyses of Otx2 Homeoprotein in Postnatal Cortex, Front Neurosci 11 (2017) 307.
[53] R. Takii, M. Fujimoto, K. Tan, E. Takaki, N. Hayashida, R. Nakato, K. Shirahige, A. Nakai, ATF1 modulates the heat shock response by regulating the stress-inducible heat shock factor 1 transcription complex, Mol Cell Biol 35 (1) (2015) 11–25.
[54] S. Ueda, I.R. Cordeiro, Y. Moriyama, C. Nishimori, K.I. Kai, R. Yu, R. Nakato, K. Shirahige, M. Tanaka, Cux2 refines the forelimb field by controlling expression of Raldh2 and Hox genes, Biol Open 8 (2) (2019).
[55] A. Tazumi, M. Fukuura, R. Nakato, A. Kishimoto, T. Takenaka, S. Ogawa, J. H. Song, T.S. Takahashi, T. Nakagawa, K. Shirahige, H. Masukata, Telomere-binding protein Taz1 controls global replication timing through its localization near late replication origins in fission yeast, Genes Dev 26 (18) (2012) 2050–2062.
[56] K. Jeppsson, K.K. Carlborg, R. Nakato, D.G. Berta, I. Lilienthal, T. Kanno, A. Lindqvist, M.C. Brink, N.P. Dantuma, Y. Katou, K. Shirahige, C. Sjogren, The chromosomal association of the Smc5/6 complex depends on cohesion and predicts the level of sister chromatid entanglement, PLoS Genet 10 (10) (2014), e1004680.
[57] T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, J. Zhang, Practical guidelines for the comprehensive analysis of ChIP-seq data, PLoS Comput Biol 9 (11) (2013), e1003326.
[58] B. Liu, J. Yang, Y. Li, A. McDermaid, Q. Ma, An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data, Brief Bioinform 19 (5) (2018) 1069–1081.
[59] M. Thomas-Chollier, A. Hufton, M. Heinig, S. O'Keeffe, N.E. Masri, H.G. Roider, T. Manke, M. Vingron, Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs, Nat. Protoc. 6 (12) (2011) 1860–1869.
[60] C.Y. McLean, D. Bristor, M. Hiller, S.L. Clarke, B.T. Schaar, C.B. Lowe, A. M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions, Nat. Biotechnol. 28 (5) (2010) 495–501.
[61] R.P. Welch, C. Lee, P.M. Imbriano, S. Patil, T.E. Weymouth, R.A. Smith, L.J. Scott, M.A. Sartor, ChIP-Enrich: gene set enrichment testing for ChIP-seq data, Nucleic Acids Res. 42 (13) (2014), e105.
[62] S. Li, C. Wan, R. Zheng, J. Fan, X. Dong, C.A. Meyer, X.S. Liu, Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks, Nucleic Acids Res 47 (W1) (2019) W206–W211.
[63] M.W. Libbrecht, F. Ay, M.M. Hoffman, D.M. Gilbert, J.A. Bilmes, W.S. Noble, Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression, Genome Res. 25 (4) (2015) 544–557.
[64] D. Pellacani, M. Bilenky, N. Kannan, A. Heravi-Moussavi, D. Knapp, S. Gakkhar, M. Moksa, A. Carles, R. Moore, A.J. Mungall, M.A. Marra, S.J.M. Jones, S. Aparicio, M. Hirst, C.J. Eaves, Analysis of normal human mammary epigenomes reveals cell-specific active enhancer states and associated transcription factor networks, Cell Rep. 17 (8) (2016) 2060–2074.
[65] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization, Nat. Methods 9 (3) (2012) 215–216.
[66] A. Mammana, H.R. Chung, Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome, Genome Biol 16 (2015) 151.
[67] Y. Zhang, L. An, F. Yue, R.C. Hardison, Jointly characterizing epigenetic dynamics across multiple human cell types, Nucleic Acids Res. 44 (14) (2016) 6721–6731.
[68] R.C.W. Chan, M.W. Libbrecht, E.G. Roberts, J.A. Bilmes, W.S. Noble, M. M. Hoffman, Segway 2.0: Gaussian mixture models and minibatch training, Bioinformatics 34 (4) (2018) 669–671.
[69] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, V. Amin, J.W. Whitaker, M.D. Schultz, L.D. Ward, A. Sarkar, G. Quon, R.S. Sandstrom, M.L. Eaton, Y.C. Wu, A.R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R.A. Harris, N. Shoresh, C.B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A.J. Mungall, R. Moore, E. Chuah, A. Tam, T.K. Canfield, R.S. Hansen, R. Kaul, P.J. Sabo, M.S. Bansal, A. Carles, J.R. Dixon, K.H. Farh, S. Feizi, R. Karlic, A.R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T.R. Mercer, S.J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R.C. Sallari, K.T. Siebenthall, N.A. Sinnott-Armstrong, M. Stevens, R.E. Thurman, J. Wu, B. Zhang, X. Zhou, A.E. Beaudet, L.A. Boyer, P.L. De Jager, P.J. Farnham, S.J. Fisher, D. Haussler, S.J. Jones, W. Li, M.A. Marra, M.T. McManus, S. Sunyaev, J. A. Thomson, T.D. Tlsty, L.H. Tsai, W. Wang, R.A. Waterland, M.Q. Zhang, L.H. Chadwick, B.E. Bernstein, J.F. Costello, J.R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J.A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes, Nature 518(7539) (2015) 317-30.
[70] A. Yen, M. Kellis, Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type, Nat. Commun. 6 (2015) 7973.
[71] Y. He, T. Wang, EpiCompare: an online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features, Bioinformatics 33 (20) (2017) 3268–3275.
[72] E. Carrillo-de-Santa-Pau, D. Juan, V. Pancaldi, F. Were, I. Martin-Subero, D. Rico, A. Valencia, B. Consortium, Automatic identification of informative regions with epigenomic changes associated to hematopoiesis, Nucleic Acids Res 45(16) (2017) 9244-9259.
[73] S. Roy, R. Sridharan, Chromatin module inference on cellular trajectories identifies key transition points and poised epigenetic states in diverse developmental processes, Genome Res. 27 (7) (2017) 1250–1262.
[74] P. Yu, S. Xiao, X. Xin, C.X. Song, W. Huang, D. McDee, T. Tanaka, T. Wang, C. He, S. Zhong, Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation, Genome Res. 23 (2) (2013) 352–364.
[75] F. Grubert, J.B. Zaugg, M. Kasowski, O. Ursu, D.V. Spacek, A.R. Martin, P. Greenside, R. Srivas, D.H. Phanstiel, A. Pekowska, N. Heidari, G. Euskirchen, W. Huber, J.K. Pritchard, C.D. Bustamante, L.M. Steinmetz, A. Kundaje, M. Snyder, Genetic control of chromatin states in humans involves local and distal chromosomal interactions, Cell 162 (5) (2015) 1051–1065.
[76] Y. Zhang, Epigenetic combinatorial patterns predict disease variants, Front. Genet. 8 (2017) 71.
[77] B. Gulko, A. Siepel, An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness consequences, Nat. Genet. 51 (2) (2019) 335–342.
[78] R. Karlic, H.R. Chung, J. Lasserre, K. Vlahovicek, M. Vingron, Histone modification levels are predictive for gene expression, Proc. Natl. Acad. Sci. U S A 107 (7) (2010) 2926–2931.
[79] X. Dong, M.C. Greven, A. Kundaje, S. Djebali, J.B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigo, E. Birney, Z. Weng, Modeling gene expression using chromatin features in various cellular contexts, Genome Biol. 13 (9) (2012) R53.

[80] W. Zeng, Y. Wang, R. Jiang, Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network, Bioinformatics 36 (2) (2020) 496–503.
[81] M.R. Mumbach, A.J. Rubin, R.A. Flynn, C. Dai, P.A. Khavari, W.J. Greenleaf, H. Y. Chang, HiChIP: efficient and sensitive analysis of protein-directed genome architecture, Nat. Methods 13 (11) (2016) 919–922.
[82] R. Singh, J. Lanchantin, G. Robins, Y. Qi, DeepChrome: deep-learning for predicting gene expression from histone modifications, Bioinformatics 32 (17) (2016) i639–i648.
[83] A. Sekhon, R. Singh, Y. Qi, DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications, Bioinformatics 34 (17) (2018) i891–i900.
[84] H. Chen, C. Li, X. Peng, Z. Zhou, J.N. Weinstein, N. Cancer Genome Atlas Research, H. Liang, A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples, Cell 173(2) (2018) 386-399 e12.
[85] Y. Murakawa, M. Yoshihara, H. Kawaji, M. Nishikawa, H. Zayed, H. Suzuki, C. Fantom, Y. Hayashizaki, Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases, Trends Genet. 32 (2) (2016) 76–88.
[86] Z. Tang, O.J. Luo, X. Li, M. Zheng, J.J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L.M. Sachs, X. Ruan, C.L. Wei, E.T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, Y. Ruan, CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription, Cell 163 (7) (2015) 1611–1627.
[87] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E.S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326 (5950) (2009) 289–293.
[88] M.J. Fullwood, M.H. Liu, Y.F. Pan, J. Liu, H. Xu, Y.B. Mohamed, Y.L. Orlov, S. Velkov, A. Ho, P.H. Mei, E.G. Chew, P.Y. Huang, W.J. Welboren, Y. Han, H. S. Ooi, P.N. Ariyaratne, V.B. Vega, Y. Luo, P.Y. Tan, P.Y. Choy, K.D. Wansa, B. Zhao, K.S. Lim, S.C. Leow, J.S. Yow, R. Joseph, H. Li, K.V. Desai, J.S. Thomsen, Y.K. Lee, R.K. Karuturi, T. Herve, G. Bourque, H.G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.K. Sung, E.T. Liu, C.L. Wei, E. Cheung, Y. Ruan, An oestrogen-receptor-alpha-bound human chromatin interactome, Nature 462 (7269) (2009) 58–64.
[89] J.M. Hariprakash, F. Ferrari, Computational Biology Solutions to Identify Enhancers-target Gene Pairs, Comput. Struct. Biotechnol. J. 17 (2019) 821–831.
[90] B.R. Sabari, A. Dall'Agnese, A. Boija, I.A. Klein, E.L. Coffey, K. Shrinivas, B. J. Abraham, N.M. Hannett, A.V. Zamudio, J.C. Manteiga, C.H. Li, Y.E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T.I. Lee, R.G. Cisse II, P. A. Roeder, A.K. Sharp, R.A. Young Chakraborty, Coactivator condensation at super-enhancers links phase separation and gene control, Science 361 (6400) (2018).
[91] Y. Chen, Y. Wang, Z. Xuan, M. Chen, M.Q. Zhang, De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles, Nucleic Acids Res. 44 (11) (2016), e106.
[92] Y. Qi, B. Zhang, Predicting three-dimensional genome organization with chromatin states, PLoS Comput. Bio. 15 (6) (2019), e1007024.
[93] Y. Atlasi, H.G. Stunnenberg, The interplay of epigenetic marks during stem cell differentiation and development, Nat. Rev. Genet. 18 (11) (2017) 643–658.
[94] P.S. Belokopytova, M.A. Nuriddinov, E.A. Mozheiko, D. Fishman, V. Fishman, Quantitative prediction of enhancer-promoter interactions, Genome Res. 30 (1) (2020) 72–84.
[95] P.W. Koh, E. Pierson, A. Kundaje, Denoising genome-wide histone ChIP-seq with convolutional neural networks, Bioinformatics 33 (14) (2017) i225–i233.
[96] J. Ernst, M. Kellis, Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues, Nat. Biotechnol. 33 (4) (2015) 364–376.
[97] T.J. Durham, M.W. Libbrecht, J.J. Howbert, J. Bilmes, W.S. Noble, PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition, Nat. Commun. 9 (1) (2018) 1402.
[98] J. Schreiber, T. Durham, J. Bilmes, W.S. Noble, Multi-scale deep tensor factorization learns a latent representation of the human epigenome, bioRxiv (2019).
[99] J. Keilwagen, S. Posch, J. Grau, Accurate prediction of cell type-specific transcription factor binding, Genome Biol. 20 (1) (2019) 9.
[100] D. Quang, X. Xie, FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data, Methods 166 (2019) 40–47.
[101] M. Karimzadeh, M.M. Hoffman, Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome, BioRxiv (2019).
[102] E. Azizi, A.J. Carr, G. Plitas, A.E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R.M. Fromme, P. Dao, P. T. McKenney, R.C. Wasti, K. Kadaveru, L. Mazutis, A.Y. Rudensky, D. Pe'er, Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment, Cell 174 (5) (2018) 1293–1308, e36.
[103] T. Stuart, R. Satija, Integrative single-cell analysis, Nat Rev Genet 20 (5) (2019) 257–272.
[104] A. Rotem, O. Ram, N. Shoresh, R.A. Sperling, A. Goren, D.A. Weitz, B.E. Bernstein, Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state, Nat Biotechnol 33 (11) (2015) 1165–1172.
[105] K. Grosselin, A. Durand, J. Marsolier, A. Poitou, E. Marangoni, F. Nemati, A. Dahmani, S. Lameiras, F. Reyal, O. Frenoy, Y. Pousse, M. Reichen, A. Woolfe, C. Brenan, A.D. Griffiths, C. Vallot, A. Gerard, High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer, Nat Genet 51 (6) (2019) 1060–1066.
[106] S. Ai, H. Xiong, C.C. Li, Y. Luo, Q. Shi, Y. Liu, X. Yu, C. Li, A. He, Profiling chromatin states using single-cell itChIP-seq, Nat Cell Biol 21 (9) (2019) 1164–1172.
[107] W.L. Ku, K. Nakamura, W. Gao, K. Cui, G. Hu, Q. Tang, B. Ni, K. Zhao, Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification, Nat Methods 16 (4) (2019) 323–325.
[108] S.J. Hainer, A. Boskovic, K.N. McCannell, O.J. Rando, T.G. Fazzio, Profiling of Pluripotency Factors in Single Cells and Early Embryos, Cell 177 (5) (2019) 1319–1329, e11.
[109] P.J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites, Elife 6 (2017).
[110] H.S. Kaya-Okur, S.J. Wu, C.A. Codomo, E.S. Pledger, T.D. Bryson, J.G. Henikoff, K. Ahmad, S. Henikoff, CUT&Tag for efficient epigenomic profiling of small samples and single cells, Nat. Commun.. 10 (1) (2019) 1930.
[111] B. Carter, W.L. Ku, J.Y. Kang, G. Hu, J. Perrie, Q. Tang, K. Zhao, Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq), Nat. Commun.. 10 (1) (2019) 3747.
[112] Q. Wang, H. Xiong, S. Ai, X. Yu, Y. Liu, J. Zhang, A. He, CoBATCH for High-Throughput Single-Cell Epigenomic Profiling, Mol. Cell 76 (1) (2019) 206–216, e7.
[113] A. Harada, K. Maehara, T. Handa, Y. Arimura, J. Nogami, Y. Hayashi-Takanaka, K. Shirahige, H. Kurumizaka, H. Kimura, Y. Ohkawa, A chromatin integration labelling method enables epigenomic profiling with lower input, Nat. Cell Biol. 21 (2) (2019) 287–296.
[114] B. Langmead, A. Nellore, Cloud computing for genomic data analysis and collaboration, Nat. Rev. Genet. 19 (4) (2018) 208–219.
[115] J.R. Dixon, D.U. Gorkin, B. Ren, Chromatin domains: the unit of chromosome organization, Mol. Cell 62 (5) (2016) 668–680.
[116] F. Ramirez, D.P. Ryan, B. Gruning, V. Bhardwaj, F. Kilpert, A.S. Richter, S. Heyne, F. Dundar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis, Nucleic Acids Res 44 (W1) (2016) W160–W165.
[117] A. Papantonis, T. Kohro, S. Baboo, J.D. Larkin, B. Deng, P. Short, S. Tsutsumi, S. Taylor, Y. Kanki, M. Kobayashi, G. Li, H.M. Poh, X. Ruan, H. Aburatani, Y. Ruan, T. Kodama, Y. Wada, P.R. Cook, TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed, EMBO J. 31 (23) (2012) 4404–4414.
[118] C.A. Davis, B.C. Hitz, C.A. Sloan, E.T. Chan, J.M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U.K. Baymuradov, A.K. Narayanan, K.C. Onate, K. Graham, S. R. Miyasato, T.R. Dreszer, J.S. Strattan, O. Jolanki, F.Y. Tanaka, J.M. Cherry, The Encyclopedia of DNA elements (ENCODE): data portal update, Nucleic Acids Res. 46 (D1) (2018) D794–D801.
[119] D. Bujold, D.A.L. Morais, C. Gauthier, C. Cote, M. Caron, T. Kwan, K.C. Chen, J. Laperle, A.N. Markovits, T. Pastinen, B. Caron, A. Veilleux, P.E. Jacques, G. Bourque, The international human epigenome consortium data portal, Cell Syst. 3 (5) (2016) 496–499, e2.