CHAPTER

4 BIOINFORMATICS OF EPIGENOMIC DATA GENERATED FROM NEXT-GENERATION SEQUENCING

Fei-Man Hsu[1], Moloya Gohain[2], Pearl Chang[2], Jui-Hsien Lu[2], Pao-Yang Chen[2]

The University of Tokyo, Chiba, Japan[1]; Academia Sinica, Taipei, Taiwan[2]

CHAPTER OUTLINE

Epigenetics in Human Disease. https://doi.org/10.1016/B978-0-12-812215-0.00004-2
Copyright © 2018 Elsevier Inc. All rights reserved.

4.1 INTRODUCTION

Epigenetics is the study of heritable changes in transcription without altering DNA sequence. Such changes in the genome constitute the epigenome. Epigenetic modifications may alter DNA accessibility and chromatin structure thereby regulating gene expression. Next-generation sequencing (NGS), also known as massively parallel sequencing and deep sequencing, has revolutionized genomic research. In Fig. 4.1, we summarize the epigenomic components and the associated NGS-based technologies in this chapter.

As an epigenomic regulator, DNA methylation is a chemical modification involved in a repressive state of the chromatin. It maintains genomic stability by repressing transposons and repeat elements. The effect of DNA methylation depends on its genomic location in the genome. Biological processes such as genomic imprinting, X chromosome inactivation, mitotic recombination, and chromosome rearrangement are closely associated with DNA methylation. In cancerous cells, a global hypomethylated state is known to disrupt mitotic recombination and chromosome rearrangement, causing aneuploidy and disrupting cellular homeostasis.

The DNA methylation modifications in the genome constitute the methylome. Genome-wide methylome profiling usually includes analysis steps such as quantification of DNA methylation levels, identification of differentially methylated regions (DMRs), and visualization of the methylome. Methylome data can be generated by high-throughput sequencing or microarray-based techniques.

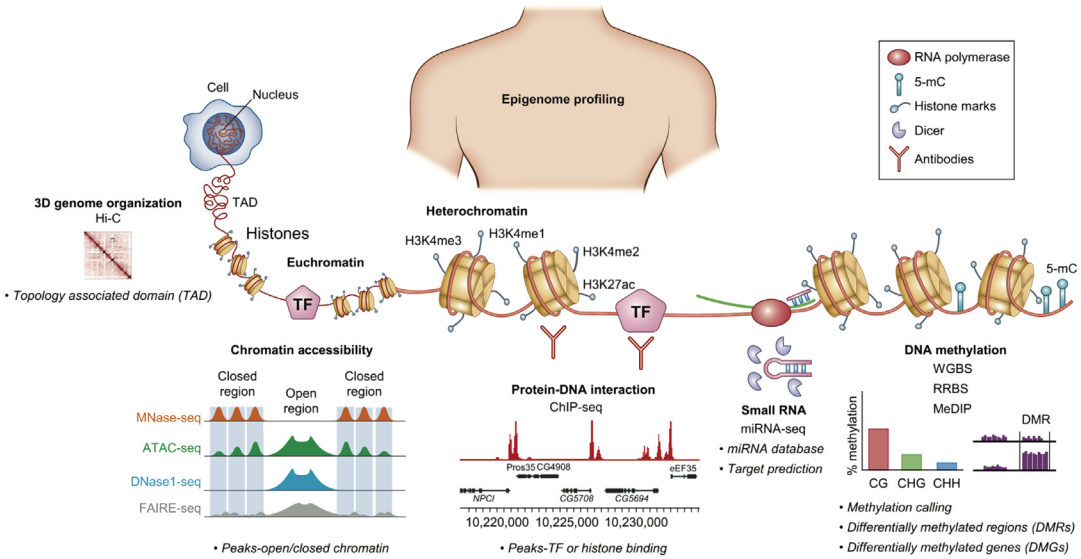

FIGURE 4.1

Schematic representation of next-generation sequencing technologies for profiling different epigenetic components. *5-mC*, 5 methylcytosine; *MeDIP*, methylated DNA immunoprecipitation; *RRBS*, reduced representation bisulfite sequencing; *TF*, transcription factor; *WGBS*, whole genome bisulfite sequencing.

The data from non-bisulfite-conversion methods, such as methylation-sensitive restriction enzymes sequencing (MRE-seq) [1] and methylated DNA immunoprecipitation sequencing (MeDIP-seq) [2], are usually analyzed by comparing the relative abundance of fragments. Whole genome bisulfite sequencing (WGBS) [3] and reduced representation bisulfite sequencing (RRBS) [4] are the state-of-art approaches because they provide measurements of absolute methylation level in single-base resolution. In clinical research, targeted BS-seq and methylation array are particularly useful.

The chromatin comprises DNA packaged with histones forming nucleosomes. The densely packed regions form the "heterochromatin" and represent the less accessible part of the genome, whereas the loosely packed regions form the "euchromatin," which is easily accessible to transcription factors (TFs). These chromatin states are regulated by histone modifications and the changes in accessibility affect binding of TFs. Coupled with NGS, chromatin accessibility can be studied using Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) to predict nucleosome positioning and chromatin state. Furthermore, DNA associated with histones or TFs can be profiled by chromatin immunoprecipitation sequencing (ChIP-seq).

The spatial organization of the chromatin in the nucleus allows physical interaction of different genomic regions with the help of chromatin-binding proteins. The interactions are frequent in regions called topologically associating domains (TADs). Deregulation of these interactions is implicated in genome instability, aging, and cancer [5]. Using chromosome conformation capture techniques (3C techniques), we can characterize these interactions. 3C techniques in combination with high-throughput sequencing (Hi-C) can profile these interactions genome-wide.

Small RNAs (sRNAs) have a genome-wide impact on transcriptional and translational regulation [6]. sRNA sequencing has revealed regulatory roles in growth, development, and differentiation to cell-fate determination by maintaining chromosome stability and regulating gene expression. Over the past decade, sRNAs, particularly micro RNA (miRNA) dysregulation has been shown in cancer and in neurological, cardiovascular, and developmental disorders in various in vitro and in vivo models.

Dysregulation of the epigenome has been linked to a number of diseases and disorders in human, such as cancer, developmental defects, neuropathy, and cardiomyopathy. Epigenomic studies shed light on the diverse aspects of a complex regulatory framework. It is therefore important to understand the experimental techniques, the biology, and the analytical approaches. With rapid advancements in NGS, the enormous amount of epigenomic data challenges existing computational pipelines and data management. Understanding the basic principles involved in the study of different epigenetic marks can be instrumental to meet these challenges. Table 4.1 is a summary of epigenomic regulators and selected tools for NGS-based studies.

In this chapter, we describe different NGS technologies directed to different epigenomic components and mechanisms, including DNA methylation, chromatin accessibility, sRNA regulation, and chromosomal interaction. For each epigenomic mechanism, we described the common computational work flow and the available bioinformatics resources. Additionally, we illustrate applications of each technology with a few case studies.

4.2 PREPROCESSING DATA FROM NEXT-GENERATION SEQUENCING

Here we focused on Illumina-based data, which are wildly used in the research societies. The data output from an Illumina sequencer are often in FASTQ format, including four lines per record (also called "read"):

- line1. Starting with "@" followed by sequence identifier
- line2. The raw sequence letters A, C, G, T, and N
- line3. Starting with "+" followed by sequence identifier
- line4. The quality values called Phred score of the sequence in line2

When multiple libraries were pooled in a sequencing run, a demultiplexing process is required to separate the reads according to their barcodes. One can use customized tools for demultiplexing, such as bcl2fastq [7] and BaseSpace [8]. The demultiplexed reads are then submitted to a quality control step. FastQC [9] is a popular software package for performing quality control of reads, and information such as base composition, quality per base, and overrepresented sequences will be demonstrated. According to the FastQC report, one can decide the parameters for the following adapter-trimming process; detect overrepresented sequences in the FastQC report tend to be $3'$ and $5'$ adapters, and bases with low quality, e.g., Phred score < 30, should be trimmed. Cutadapt [10] is a well-known adapter-trimming tool.

4.3 READ ALIGNMENT

Preprocessed reads are to be aligned to the reference genome. The NGS data discussed in this chapter are to be aligned directly against the genome sequences, except for BS-seq, because the original

Table 4.1 Epigenome Regulators and Selected Tools for NGS-Based Studies

Epigenomic Marks	NGS Technology	Tools/Pipeline/Databases
DNA methylation	Whole genome bisulfite sequencing Reduced representation bisulfite sequencing	Aligner: • BS-Seeker2 (wraps Bowtie, Bowtie2) • Bismarck (wraps Bowtie, Bowtie2) • Bisulfighter (wraps LAST) Generalized profiling: • MethGO • BSPAT • GBSA DMR finding: • BSmooth • methylKit • BiSeq • methylPipe
	Methylated DNA immunoprecipitation sequencing (MeDIP-seq)	Aligner: • Bowtie2 • BWA • SOAP Peak-calling: • MACS • MACS2 DMR finding: • MEDIPS
DNA−protein interaction	Chromatin immunoprecipitation assay-sequencing (ChIP-seq)	Narrow peak-calling tools: • Unique Peaks • Homer • MACS2 bdgdiff Broad peak-calling tools: • MACS bdgbroadcall • ODIN-bin • RSEG • SICER • diffReps-nb Tools for both types of peaks: • ChIPComp • DiffBind • MAnorm
Chromatin accessibility	ATAC-seq FAIRE-seq DNase-seq MNase-seq	Peak-calling tools: • MACS • ZINBA • Hotspot • Homer • F-seq • R module ATAC-seq pipeline • DESeq2
RNA regulation	Micro RNA sequencing (miRNA-seq)	Aligner: • Bowtie2, BWA, Maq, Stampy, NovoAlign, SOAP, GNUMAP Databases: • miRBase • miR2Disease

Continued

Table 4.1 Epigenome Regulators and Selected Tools for NGS-Based Studies—cont'd

Epigenomic Marks	NGS Technology	Tools/Pipeline/Databases
		RNA secondary structure prediction • ViennaRNA Package miRNA target prediction: • PicTar • miRanda miRNA gene prediction/expression profiling: • miRDeep/miRDeep2 • miRanalyzer • miRExpress • miRTRAP • miRTools • miRNAkey
Chromosomal interaction	Hi-C ChIA-PET	Contact matrix generation: • HOMER • HiTC Differential interaction analysis: • diffHic

DMR, *differentially methylated region;* NGS, *next-generation sequencing.*

sequences have been chemically modified therefore converted genome sequences are to be used (see Section 4.4). Bowtie2 [11], BWA [12], SOAP [13], and LAST [14] are available short read aligners. SAM file (or its binary format BAM file) and MAF (multiple alignment format) are output file formats, which store the alignment information such as the genomic position, mismatch, and alignment score. One can extract "uniquely mapped" reads from alignments with SAMtools [15], that is, to exclude reads aligned to multiple positions. For BS-seq, ChIP-seq, and ATAC-seq, duplicated reads with completely same sequences could be removed at this step to avoid potential bias from polymerase chain reaction (PCR) amplification. After read alignment, ChIP-seq, sRNA-seq, and ATAC-seq pipelines include one step "peak calling" with different requirements according to the data specificity.

4.4 PROFILING DNA METHYLATION

4.4.1 DNA METHYLATION

DNA methylation refers to adding one methyl group to the 5th carbon of a cytosine (C), forming 5-methylcytosine (5mC). The methyl group is transferred to cytosine by DNA methyltransferases (DNMTs). DNA methylation can occur in the symmetric CG and CHG contexts and in the asymmetric CHH context ("H" represents A, C, or T). In mammalian genomes, where methylation mainly occurs on CG dinucleotides, DNMT3 de novo methylates cytosine [16,17], and DNMT1 maintains DNA methylation during cell division [17,18]. Methylation at a promoter region may repress gene expression through altering the chromatin structure or blocking transcription initiation [19].

Several biological processes are known to be regulated by DNA methylation. During mammalian development, genomic imprinting is a phenomenon through which genes are expressed in a parent-of-origin manner, that is, one of the two parental alleles is silenced by DNA methylation [20]. Loss of imprinting of *IGF2* (insulin-like growth factor 2) is associated with Beckwith–Wiedemann syndrome

(BWS) and increases the risk of colorectal cancer [21]. X chromosome inactivation is also directed by DNA methylation, in which one of the two X chromosomes in a female genome is packed into heterochromatin to compensate for the extra dosage [19]. The global hypomethylation in cancerous cells results in mitotic recombination and chromosome rearrangement, and aneuploid cells emerge [22]. In plant embryonic development, the endosperm, which contains three copies of each chromosome, is shown to be hypomethylated compared with embryos [23]. Unlike plants and animals, in which promoter and gene body methylation marks are common and provide an additional level of gene regulation, fungi have most DNA methylations occurring in repeat elements and transposons to stabilize the genome [24].

4.4.2 EXPERIMENTAL APPROACHES

Because DNA methylation plays critical roles in biological processes, several experimental approaches have been developed to profile DNA methylation. High-performance liquid chromatography (HPLC) was first used to detect the average 5mC level in a DNA sequence or a genome [25]. Double-stranded DNA is first hydrolyzed into single nucleotides, dissolved in a liquid solvent and pumped into a stationary phase column. While passing through the column, C, 5mC and other nucleotides have different retention times due to different interaction strengths with the static phase. With standard compounds as references, 5mC can be collected at specific time points and then quantified. This method is accurate for quantification but cannot differentiate methylation distribution.

With advances in microarray and NGS technology, genome-wide DNA methylation profiling approaches are therefore doable [26,27]. Restriction enzymes recognize and cut specific sequences. Even with the same cutting site, the "cutting" can be induced or blocked by methylation. MREs might only cleave the target sequence of one methylation state and leave another intact. MRE digestion coupling with NGS (MRE-seq) can reveal the location of CpG sites of the same methylation state within any pair of two recognition sites [1]. This method estimates the relative DNA methylation levels and is limited to the number and distribution of the CpG-containing recognition sites.

Affinity enrichment-based methods use methyl-CpG-binding domain (MBD) proteins or 5mC-specific antibodies (as in MeDIP, methylated DNA immunoprecipitation) to enrich methylated DNA fragments [2,28]. These fragments can then be evaluated using tiling arrays (MeDIP-chip [29]) or NGS (MeDIP-seq [2]); that is, the sequence abundance represents the relative methylation level. These methods can begin with a small amount of starting DNA material, which is important for clinical research and for obtaining the genome-wide methylation profile. Nevertheless, the results may be biased by an uneven distribution of CpG sites; moreover, because the resolution of MeDIP-seq is 100−300 bp, the exact context of methylated site (CG, CHG, and CHH) cannot be distinguished.

Bisulfite treatment of DNA can reveal the methylation status at a single-base resolution of the DNA sequences and can be used to measure absolute methylation level (percent methylated cells within a pooled cell population). In the bisulfite sequencing (BS-seq) protocol, bisulfite conversion is the key step during which the sodium bisulfite chemical converts C into uracil (U) while 5mC is protected by methylation and remains unchanged. In a subsequent PCR, the U eventually converts into thymine (T) (Fig. 4.2). When coupled with Sanger sequencing, the methylation state of each C inside a target DNA sequence is profiled, that is, clonal BS-seq. To profile genome-wide DNA methylation, in WGBS, genomic DNA is fragmented, end-repaired, A-tailed (add an adenine base to 3′end) and ligated with sequencing adapters [3]. These DNA fragments are then size-selected to compromise the sequencer, treated with sodium bisulfite, and PCR-amplified; next, the final library is sequenced.

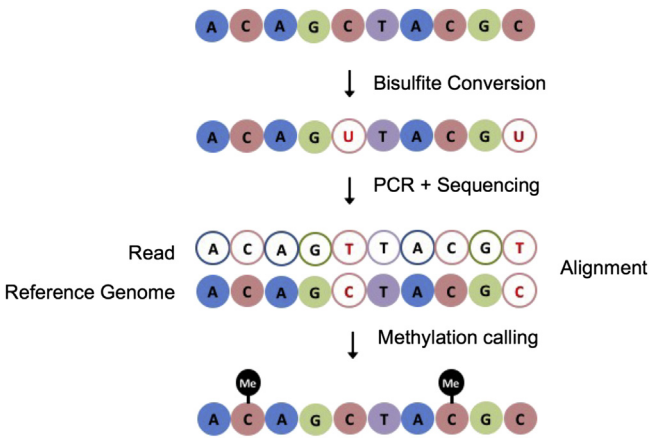

FIGURE 4.2 Principle of BS-seq for Study of the Methylome.

A sequence containing four Cs with an unknown methylation status is bisulfite-converted, PCR-amplified, and sequenced. C will be read as T, whereas 5mC will remain C. After sequence alignment, the methylation status of each C inside this sequence can be characterized by comparison with the reference genome. *PCR*, polymerase chain reaction.

To investigate the mammalian methylome at a lower cost, Meissner et al. developed RRBS, which integrates *Msp*I restriction enzyme digestion, bisulfite conversion, and NGS to analyze the methylation patterns of specific fragments [4]. A size selection of *Msp*I-digested fragments between 40 and 220 bps was found to cover 85% of CpG islands (CGIs), mostly in promoters, which compose only 1%−3% of the mammalian genome, thereby significantly decreasing the amount of sequencing. RRBS has been widely used in profiling large-scale samples. Pellegrini et al. performed RRBS in 90 inbred mouse strains; conducted an integrative analysis that included genome-wide expression levels, proteomics, metabolomics, and 68 clinical traits; and performed epigenome-wide association studies (EWASs) [30].

Targeted epigenomic sequencing is also available. For instance, Li et al. provided a capture-based BS-seq to include predefined genomic regions only [31]. This is especially cost-efficient for clinical research, which has large sample size with definite genomic regions of interest.

4.4.3 METHYLOME

Methylome represents the information of DNA methylation of all cytosines in a genome. The first WGBS study in 2008 reported the bulk methylation level within the CG, CHG, and CHH contexts in the *Arabidopsis* genome; the global methylation pattern in wild-type and methylation-related mutants; and specific sites associated with gene expression [32]. Lister et al. published the human methylome in two human cell lines H1-hESC and the differentiated cells from fibroblast IMR90, and found that in H1-hESC, more 5mCs are found in a non-CG context [33]. Hsieh et al. compared *Arabidopsis* endosperm and embryo methylomes and found that virtually the entire endosperm genome is demethylated, coupled with extensive local non-CG hypermethylation of small interfering

RNA (siRNA)-targeted sequences [23]. In 2013, two maize studies reported that the maize genome is highly methylated, and a specific "CHH island" was found upstream of transcription start sites (TSSs) [34,35].

In addition to global profiling, case studies that compare the methylation pattern between samples provide insight on tissue specificity or developmental control. For instance, *DNMT1* loss-of-function is lethal in humans, suggesting DNA methylation is vital for human embryonic development [36]. Recent studies performed WGBS in human primordial germ cells [37,38]. The time-lapse recording shows two waves of demethylation in mammalian germline development, with the first eliminating the epigenetic memory from the parents and the second removing the memory of early embryonic development. Some persistent methylated regions are found to escape from the reprogramming, suggesting they have indispensable roles in development. This result provides evidence that during early embryonic development, the germline genome lacks methylation protection and is easily affected by the environment.

4.4.4 BS-SEQ DATA ANALYSIS

Genome-wide DNA methylation profiling is computationally intensive. The general workflow for the bioinformatics analysis includes data processing, quantification of DNA methylation levels, general profiling, identification of DMRs, and visualization of the methylome (Fig. 4.3) [27]. The data from

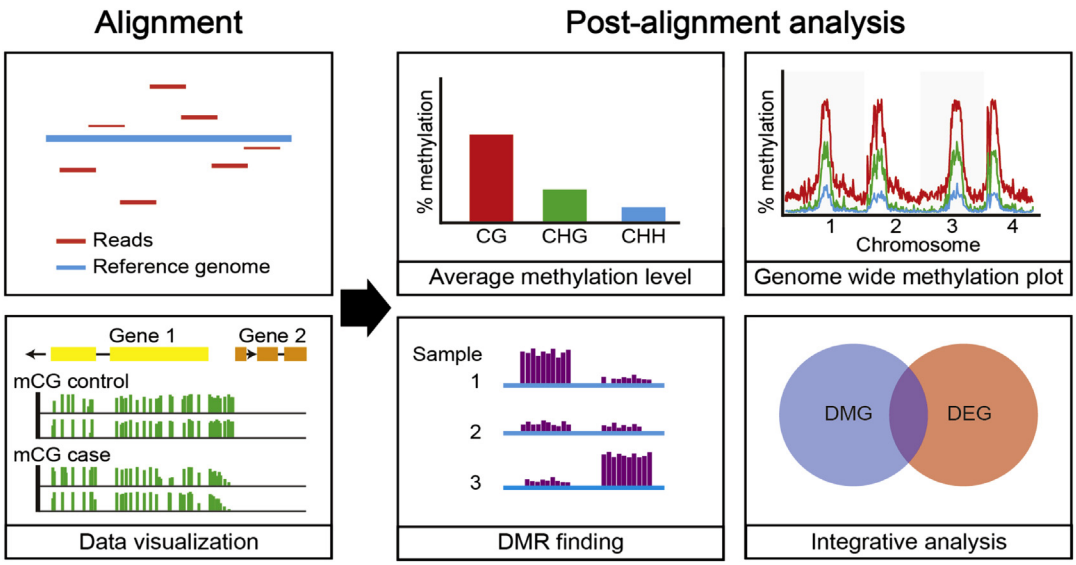

FIGURE 4.3

Analysis pipeline of BS-seq. The reads can be mapped to the reference genome with bisulfite aligners, and the methylation status of each cytosine inside the genome can be viewed with the genome browser. Postalignment analysis such as methylation level calculation, correlation with gene expression RNA-seq data, and differentially methylated region (DMR) finding can then be applied. *DEG*, differentially expressed gene; *DMG, differentially methylated gene*.

other non-bisulfite-conversion methods, such as MRE-seq and MeDIP-seq, are usually analyzed by comparing the relative abundance of fragments. Peak-calling tools such as MACS could be used to identify methylation-enriched regions [39] and DMRs could be identified by comparing peak intensity with a tool such as MEDIPS [40]. Bisulfite-converted data analysis involves methylation calling at individual Cs, and statistical testing is required to assess differential methylation between methylomes. Here, we focus on the bioinformatic analyses of bisulfite-converted data, in particular, WGBS and RRBS.

In BS-seq, the methylation information from a genome is stored in the FASTQ format. These bisulfite-converted read records are processed through several steps, including adapter trimming, a quality assessment of reads, reads alignment, and methylation calling. In particular, mapping bisulfite-converted reads to the reference genome is challenging for the following three reasons:

1. reduced sequence complexity,
2. asymmetric C-to-T alignments,
3. bisulfite-converted Watson and Crick strands that are not complementary to each other because bisulfite conversion occurs only at Cs (not Gs).

Bisulfite sequencing aligners are mostly based on one of two algorithms: wild cards and three-letter algorithms. Wild-card aligners such as Bisulfighter [41] and GSNAP [42] substitute Cs with Cs or Ts in the reference genome, and reads with both Cs and Ts can be aligned [43]. This method results in a higher genomic coverage but might be biased toward higher methylation levels. The three-letter aligners convert all Cs in the reference genome and in the reads into Ts; thus, standard aligners with lower mappability can be adopted because of reduced sequence complexity [44].

The bisulfite aligner generally outputs alignments, along with the methylation calling information of each C with sequence context information, e.g., the CGmap file (Fig. 4.4A) in BS-Seeker2 [44].

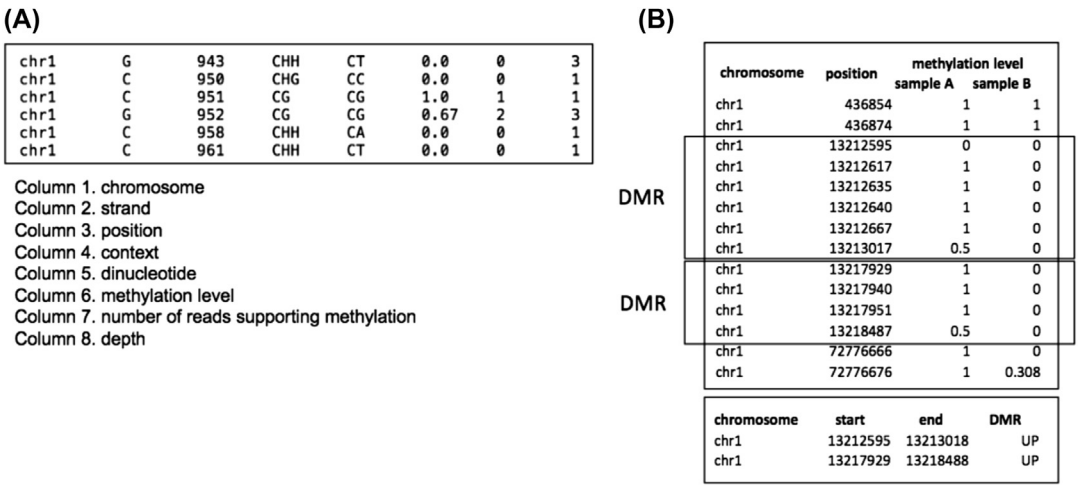

FIGURE 4.4 Example of Data From BS-seq.

(A) CGmap format from BS-seeker2. (B) Differentially methylated cytosines in serial comprise of differentially methylated regions (DMRs).

Users can select sites with sufficient reads coverage (depth, X), calculate the average methylation level, and generate informative plots using postalignment analysis software tools such as MethGo [45], BSPAT [46], and GBSA [47]. WGBS and RRBS generate methylation calls at each C as an estimate of the percentage of cells with methylation. For sample comparison, common sites with sufficient coverage in all samples are needed. Statistical tests are used to identify differentially methylated loci in comparisons (Fig. 4.4B). For studies without replicates, Fisher's exact test is generally adopted. A comparison with no replicates completely ignores within-group variations, resulting in an overstatement of the differences and a high false-positive rate.

DMRs are genomic regions that exhibit a different methylation status between two groups of samples (Fig. 4.4B). Although the prediction of individual loci to be differentially methylated can be noisy and are strongly affected by the sequencing depth, identifying DMR based on several loci within a small genome region can provide relatively robust results. Generally, the DMR detection algorithm adopts a sliding window of several hundreds of base pairs across the genome to survey candidate DMRs, and the most common approach is to perform Fisher's exact test CpG-wise [48]. As the coverage of each sample may be different, only sites covered by all samples are comparable. Comparing statistics such as T-scores from a *t*-test or *P*-value is necessary to test for significant methylation differences. In the BSmooth software, a beta-binomial is assumed to be the suitable model for replicated bisulfite sequencing data [48]. The observation is assumed to be binomially distributed, whereas the methylated proportion at a particular site can vary across samples. The differences at an individual site can be small but may expand and persist across a region, which is a candidate DMR. Therefore, DMRs are determined with greater statistical power and are more informative. Several DMR-finding tools are available, such as methylKit [49], BiSeq [50], and methylPipe [51].

4.4.5 PROFILING 5-HYDROXYMETHYLATION

In addition to 5mC, 5-hydroxymethylation (5hmC) has been shown to be important during mammalian germline development [37]. 5hmC is produced via the oxidation of 5mC catalyzed by the ten-eleven translocation (TET) family of proteins [52]. TET-assisted bisulfite sequencing (TAB-seq) has been used to generate genome-wide 5hmC profiles at a single-base resolution in human and mouse embryonic stem cells [53]. 5hmC is protected from TET protein-mediated oxidation, whereas 5mC is oxidized by the Tet1 enzyme to 5-carboxylcytosine (5caC). 5caC and unmethylated C are susceptible to bisulfite conversion and thus are sequenced as T, whereas 5hmC is sequenced as C. 5hmC data can be analyzed by the same bioinformatic pipelines as those for BS-seq.

4.4.6 QUALITY ASSESSMENT OF BS-SEQ

To assess the quality of BS-seq data, the methylation level of the spiked-in lambda phage DNA could be viewed as the "bisulfite unconversion rate" because lambda phage DNA contains no 5mC. Usually the unconversion rate should be controlled within 1.0%. The correlation of per base methylation level between technical replicates could be used to verify the concordance of a BS-seq sample.

4.4.7 APPLICATION OF BS-SEQ IN CANCER RESEARCH

Promoter hypermethylation has been shown to be important for tumorigenesis through transcriptional silencing of tumor suppressor genes. Owing to the low availability of cancer tissue and high cost of

WGBS, in clinical research it is more often to use methylation array, e.g., Illumina Infinium 450K, MeDIP-seq, or RRBS to allow more replicates comparison for target methylation sites. For example, Ashktorab et al. adapted RRBS in colorectal cancer tissues from African-American patients and identified novel CpG hypermethylation sites in genes involved in Wnt/β-catenin, PI3k/AKT, VEGF, and JAK/STAT3 pathways [54].

4.4.8 CONCLUSION

DNA methylation can modulate gene expression without any change in DNA sequence. With its heritability and specificity, the DNA methylation at specific genomic loci can be served as a candidate biomarker for cancer diagnosis and epigenetic disease prediction. With a reduction in sequencing costs, in the future, it would be possible to construct personal methylomes, even on a single-cell scale. Therefore, improved speed and accuracy for large-scale BS-seq data analyses would be critical.

4.5 ASSESSING DNA–PROTEIN INTERACTION IN THE CHROMATIN-CHIP-SEQ

The chromatin consists of DNA wrapped around core histones H2A, H2B, H3, and H4 forming the beads-on-string structure of nucleosomes. Histones can be chemically modified via lysine acetylation, lysine and arginine methylation, serine and threonine phosphorylation, lysine ubiquitination, and sumoylation. These chemical modifications in histones can alter chromatin structures. Together with TFs, these histone modifications regulate gene expression. ChIP is an effective method to identify DNA sequences associated with TFs and chromatin modifications. DNA fragments coprecipitated with the target modification/protein are enriched using antibodies with affinity purification. Previously, enriched DNA fragments were assayed with a DNA microarray ("chip"), known as ChIP-on-chip. Recent advances in NGS gives rise to the ChIP-seq technique. Compared with ChIP-on-chip, ChIP-seq is a relatively precise way to identify DNA–protein interactions because associated DNA sequences are sequenced to the nucleotide resolution [55]. ChIP-seq results can be integrated with other genome-wide data, including gene expression by RNA-seq, DNA methylation by BS-seq, and chromatin accessibility by ATAC-seq [56–59] for integrative analyses.

The success of a ChIP-seq experiment relies largely on the quality of the sequencing library construction. One major challenge is to acquire antibodies with high specificity against the target protein or histone mark because imprecise immunoprecipitation caused by antibody cross-reactivity to nontargets will lead to background noise and variability. A properly prepared library is composed of nonredundant DNA fragments that are representative of the genome, and the library quality is crucial to accurately identify associated DNA sequences. For data computation, a sufficient sequencing depth (minimal 40–50 million reads) and the use of appropriate peak-calling algorithms are critical to reach a balance between sensitivity and specificity. Moreover, quality assessment across ChIP-seq replicates is essential to eliminate noises and false positives.

4.5.1 PREPARING CHIP-SEQ SEQUENCING SAMPLES

Library preparation for ChIP-seq involves the isolation of protein-bound DNA from chromatin, followed by sonication and purification with immunoprecipitation. After immunoprecipitation, the

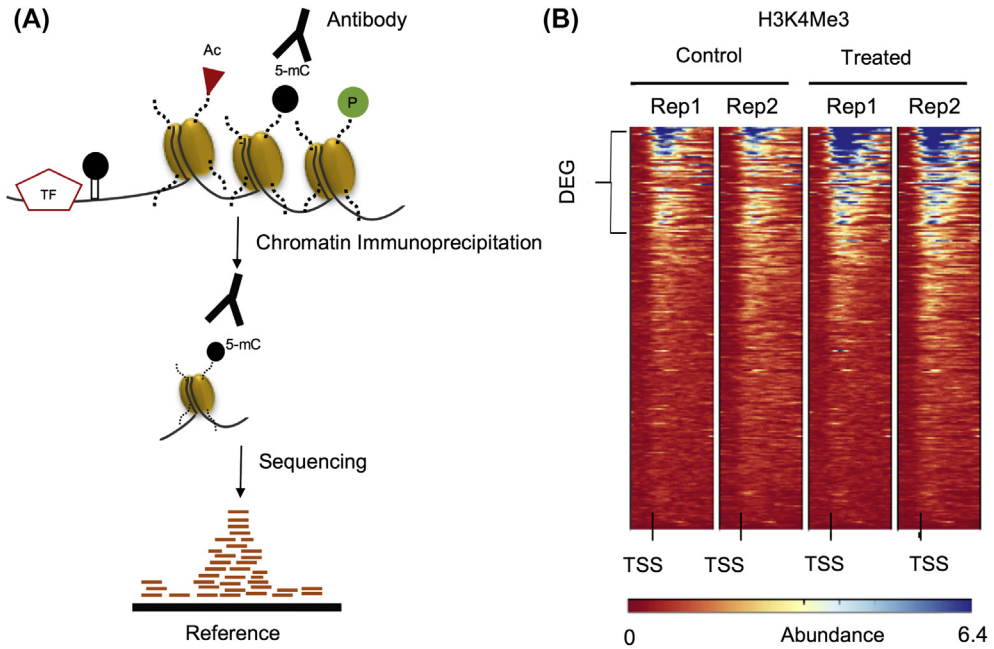

FIGURE 4.5 ChIP-Seq Identifies Histone Marks and Other Protein–DNA Interactions on the Genome.

(A) Illustration of the key steps of ChIP-seq. *Ac*, acetylation; *5-mC*, 5-methylcytosine; *P*, phosphorylation. (B) Analysis of the H3K4me3-enriched regions of differentially expressed genes (DEGs). A heat map showing differential enrichment of H3K4me3, known as a biomarker for active promoter regions. H3K4me3 is enriched near the TSS of a set of DEGs (*brackets*) for both control and treatment samples. H3K4me3 is more abundant in treatment samples, suggesting that the differential gene expression may be a result of H3K4me3 modulation. *Rep*, biological replicates.

captured DNA fragments are sequenced using NGS. Sample preparation relies on the nature of the biological question, the choice of antibodies, the material to be sequenced, the fragment length, and the required depth of sequencing. The ChIP-seq library sample construction involves cross-linking, cell lysis and fragmentation, immunoprecipitation, cross-link reversal, adapter ligation, and quality checks. The key steps in ChIP-seq are illustrated in Fig. 4.5A. Approximately 10 million cells are required for a ChIP experiment, although the on-going development of new techniques such as nano-ChIP-seq and ultralow-input native ChIP-seq can analyze a sample as small as 1000–10,000 cells [60,61]. At the cell lysis stage, removing cytosolic proteins are essential to reduce background binding and increase sensitivity. DNA fragments of ideally 150–500 bp are obtained by sonication or nuclease digestion. Choosing an appropriate antibody is critical to the success of ChIP-seq because the antibody binding specificity determines whether immunoprecipitation can precisely pull down the target DNA. Based on an assessment conducted as part of ENCODE projects, more than one-fifth of 200 antibodies failed specificity tests or immunoprecipitation experiments.

4.5.2 IDENTIFYING DNA SEQUENCES ASSOCIATED WITH PROTEINS OR HISTONE MODIFICATIONS

ChIP-seq data analysis involves one major step—peak calling, a computational method that identifies genomic regions enriched with aligned reads. These peak regions are the targets of protein binding or specific histone modifications. An ideal library sample is composed of every target genomic region, but practically immunoprecipitation or PCR amplification might skew the fragment population. Bioinformatic tools have been developed for peak calling; some are designed to look for identifying peaks from sharp, narrow areas of TF binding, whereas others are for broad, large-sized regions (e.g., ones with histone modification marks) (Fig. 4.6).

Steinhauser and coworkers provide a guideline for choosing tools for ChIP-seq data according to different data types and biological background [62], see Table 4.2 for a summary of the bioinformatic tools. Recommended tools for identifying narrow peaks include Homer [63], DiffBind [64], and ChIPComp [65], whereas tools such as diffReps [66], RSEG [67], and SICER [68] can be used for broad peaks. Some tools are equipped with options to detect both types of peaks, including ChIPComp, DiffBind, and MAnorm [69]. Misuse of tools/algorithms can obscure subsequent analyses because some tools call a large number of small regions, and others aggregate them and report large differential domains larger than one kilobase. Although sharp, small regions are more reflective of the real size of TF-binding motifs, misuse of tools that report large domains can result in false findings. Indeed, it was found that tools such as diffReps or RSEG that have been established to detect differential histone modifications perform poorly with a TF data set. In a test with diffReps, a tool capable of analyzing both replicate and nonreplicate data types, less than half of the differential regions were identified when replicate data were the input, compared with nonreplicate data [62]. Although tools are available for nonreplicate data (such as MAnorm and Homer; Table 4.2), it is highly recommended to use replicate ChIP-seq datasets to achieve a consensus for differentially enriched regions as shown in Fig. 4.5B. The heat map shows differential enrichment of H3K4me3 near the TSS of a set of differentially expressed genes (DEGs) for both control and treatment samples.

The detection of genomic regions with differential read abundance between samples of different cell types or treatments requires differential peak calling (DPC). Currently available DPC methods use either a two-stage or a one-stage approach. For two-stage DPC methods, the candidate peak regions detected in each sample are first determined and later analyzed in the second stage with methods

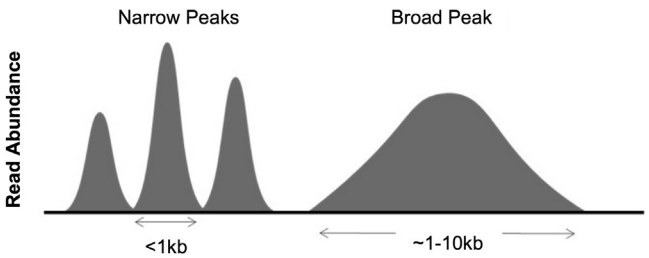

FIGURE 4.6

Illustration of two types of peaks in the ChIP-seq datasets. Narrow peaks are generally associated with TF binding, and broad peaks indicate regions with histone modification marks.

Table 4.2 Tools Suitable for Different Types of ChIP-Seq Datasets

Signal Type	Replicates	Tools	Significance Measure
Sharp	Yes	ChIPComp	Posterior probability
		DiffBind	*P*-value or FDR
	No	MAnorm	*P*-value
		Unique peaks	-
		Homer	FDR or *P*-value
		MACS2 bdgdiff	log10 likelihood ratio
		ODIN-bin	*P*-value
Broad	Yes	ChIPComp	Posterior probability
		DiffBind	*P*-value or FDR
		diffReps-nb	*P*-value
	No	MAnorm	*P*-value
		Unique peaks	–
		MACS bdgbroadcall	log10 likelihood ratio
		RSEG	–
		SICER	FDR

Tools are categorized according to the shape of the signal (Narrow or Broad Peaks) to be detected and presence of replicates.

tailored for the differential expression analysis of RNA-seq data. These methods process data from technical and biological replicates in their second step. Such tools include DBChIP [70] and MAnorm [69] for two-stage differential peak callers. One drawback of the two-stage methods is that the differential peaks for comparison have been predefined in the first stage, and therefore, subtle changes of smaller peaks within the predefined larger peaks cannot be detected. This may be problematic for analyzing ChIP-seq data of histone modification. To resolve this issue, one-stage DPC methods can be applied instead; these methods are based on sliding window approaches or segmentation methods such as hidden Markov model (HMM) [71–73]; examples for two-stage differential peak callers are ChIPDiff [71], ODIN [72], and THOR [74]. Currently, window-based DPC methods support the analysis of technical or biological replicates.

4.5.3 QUALITY ASSESSMENT OF CHIP-SEQ DATA

Accurate peak calling or identification of differential enrichment regions of ChIP-seq is particularly important because the immunoprecipitation step is prone to variations in experimental conditions and antibody-binding specificity. The noise level can be substantial enough to mask real signals and result in fewer peaks. One widely used estimate of the signal-to-noise ratio is the FRiP (fraction of reads in peaks), which are the proportions of sequencing reads, out of the total reads, that are located in enriched regions. Usually, a minority of reads aligns to enriched genomic regions (i.e., peaks), and the remaining reads represent the background. The percentage of reads falling in peak regions, therefore, serves as an indicator for the success of the immunoprecipitation. The ENCODE project exercises a minimal 1% FRiP threshold guideline [75].

An appropriate control data set is essential for ChIP-seq data analysis predominantly because of varied DNA fragmentation during sonication and antibody cross-reactivity. Moreover, some open chromatin regions can be overrepresented in sonicated samples. Common control samples are either an "input" or a "mock" ChIP reaction. An input control is extracted DNA that has been cross-linked and fragmented but not immunoprecipitated. A mock control is prepared using an antibody that binds to nontarget, nonnuclear proteins. Nonspecific immunoglobulin G (IgG) antibodies are often used in the mock reaction; therefore, the mock control is also called the "IgG control." Moreover, an internal control as a spike-in aliquot of chromatin from a different species might be important to precisely determine the normalization factor and enable comparisons across samples. In addition to controls, ideally, at least two biological replicates should be included in ChIP-seq experiments to access the reproducibility by measures such as the Pearson correlation coefficient. Statistical approaches such as the irreproducible discovery rate (IDR) can also be used to identify consistent signals among replicates [76,77]. Because significant peaks are likely genuine signals for the underlying biology, they are expected to have a high consistency between replicates; however, peaks with low significance are more likely noises and are expected to have low consistency. IDR measures the consistency of peaks of both high and low significance between two samples from replicates. The value of IDR describes the expected probability that a signal is false.

4.5.4 CHIP-SEQ IN CANCER RESEARCH

Mutations in enzymes involved in epigenetic modulation are frequently observed in cancer; in fact, approximately one-third of genes driving tumor development are related to chromatin structure and function [78]. For instance, changes in histone H3 methylation were found to be associated with tumorigenesis [79]. KDM3A (lysine demethylase 3A) modulates histone methylation by demethylating mono- or di-methylated H3K9 (lysine 9 of histone H3) [80,81]. To study the dynamic response to epigenetic regulation of KDM3 and its role in prostate cancer development, ChIP-seq was used to profile histone methylation [79]. The peak-calling step used MACS [39] and ChIPseek [82], which uses scripts from BEDTools [83]. The comparison of the control and the *KDM3* knockdown showed specific loss of *KDM3A* ChIP-signals but significant gains of H3K9me1and H3K9me2 signals, indicating a global histone demethylation effect on H3K9. Moreover, tumor formation in an orthotopic prostate tumor was abolished in the knockout. This example demonstrated that ChIP-seq can serve as an effective tool to elucidate how epigenetic dysregulation can drive cancer development, providing potential therapeutic solutions to treat cancer.

4.5.5 CONCLUSION

ChIP-seq is widely used to investigate epigenetic features and to map TF-binding motifs. This technique has advanced our understanding of disease mechanisms and provides insights into potential clinical applications. To benefit fully from what ChIP-seq can offer, integrative analysis with another data type or multiple types is key to discover underlying regulatory networks that may lead to novel therapies or drugs. Future development in quality antibodies and validation in immunoprecipitation specificity would greatly improve the accuracy of analyses. Establishment of protocols that require a small amount of tissue or even down to single-cell levels while maintaining high resolution is another critical development for ChIP-seq applications in medicine.

4.6 ANALYSIS OF THE SMALL RNA COMPONENT OF THE EPIGENOME

Small ribonucleic acids (sRNAs) are short, noncoding RNA molecules of approximately 18−30 nucleotides [84]. They control vital cellular processes from growth, development, and differentiation to a cell-fate determination by maintaining chromosome stability and regulating gene expression. sRNAs are largely posttranscriptional regulators of gene expression that elicit translational repression and/or translational cleavage of target RNA [85]. sRNA deregulation therefore disrupts cell physiology with implications proven in cancer, neurodegenerative disorders, viral infections and a host of other diseases. With advances in sequencing technologies, sRNAs can be applied to disease etiology, developing diagnostics and designing therapeutic targets [86].

4.6.1 BIOGENESIS OF SMALL RNA CLASSES

Mature sRNAs are formed by the action of RNA processing enzymes such as Drosha and DICER. The Argonaute proteins bind to sRNAs and recognize specific target mRNAs through sequence complementarity, which then leads to mRNA cleavage or translation inhibition. A 5′ monophosphate and 3′ hydroxyl group at the termini differentiates sRNAs from mRNAs. Many classes of sRNA are formed from larger RNA precursors. The common classes include miRNA and siRNA, which are functionally similar but differ in their biogenesis [87]. miRNAs are 18−24-nt in length, generated by the action of DICER on endogenous single-stranded RNA (ssRNA) with imperfectly base-paired hairpin structures. miRNA forms from one arm of the stem-loop that contains loops to yield ssRNA usually in excess of their complement. A conserved base-pairing occurs between the 3′ UTR of mRNA and the 5′ region of miRNA, called the seed region.

The siRNAs are generated from perfectly base-paired double-stranded RNA (dsRNA) precursors of both exogenous and endogenous origin. Each strand of the RNA duplex forms complementary siRNA in equal abundance. Both miRNA and siRNA interact with the 3′ UTR of target RNAs and silence cytoplasmic mRNA by one of three ways: mRNA degradation, translational repression, or accelerated mRNA decapping. In the light of high-throughput sequencing, novel classes of sRNAs still remain to be discovered.

4.6.2 NEXT-GENERATION SEQUENCING OF SMALL RNA

Northern blotting, qPCR, microarray, etc., have shown diverse roles of sRNA in cell differentiation, growth/proliferation, migration, apoptosis/death, metabolism, and defense [88]. However, NGS achieves single-base resolution [89,90], allowing us to differentiate between related species of sRNA. NGS-based techniques are cost-effective because of reduced manpower and necessary reagents. The cost of sequencing the human genome continues decreasing in recent years. Because of its massively parallel approach, it allows over 300 Gb of DNA to be read on a single run in a relatively short time. The accuracy of NGS may be attributed to the intrinsic use of overlapping reads, as each read is amplified multiple times during library preparation before sequencing. The greater the number of reads overlapping a region, the higher is the coverage, making it more reliable. The data obtained from miRNA sequencing can be used for expression profiling, identification of sequence isoforms and novel miRNAs, prediction of potential miRNA genes and miRNA targets, and functional prediction.

4.6.3 PROFILING MICRO RNAS

The miRNA-seq is the use of NGS to sequence miRNAs. Libraries for miRNA-seq are prepared using protocols and kits available from Illumina, Applied Biosystems (ABI) SOLiD, New England Biolabs (NEB), and TriLink Biotechnologies [91]. Large RNAs are removed after a size selection step. The library prep relies on specific ligation of adapters to miRNA molecules coupled with the size selection of the miRNA enriched from a pool of sRNA on agarose gel. Adapter ligation to the 3′ and 5′ termini is followed by adapter-specific PCR amplification. For multiplexing sequencing, index sequences integrated with PCR primers are incorporated.

In general, sRNA expression profiling can be performed using 1−2 M reads, which can be increased to 10−20 M for identifying novel sRNA [92]. After post-sequencing quality filtering (see Section 4.2 for detail), reads of at least 15−40 nucleotides after trimming are retained for further analysis. Repeated reads of identical sequence are collapsed into a single unique read with a note of the read count. These unique sequences across samples are then merged. The resulting reads are potential sRNAs. Because sRNA reads are short (16-25 bp for miRNAs), a stringent alignment pipeline is used allowing only perfect matches using short read aligners.

4.6.4 QUALITY ASSESSMENT OF SRNA-SEQ DATA

Normalization is an essential preprocessing step in the analysis. Its primary purpose is to ensure that observed differences are because of the biology rather than artifacts resulted from sample handling or processing. An effective normalization technique minimizes technical and experimental bias without introducing noise. The absolute distribution plot of the miRNA count data after alignment and normalization can be visualized using density distribution curves [93]. High consistencies between the distribution profiles of replicate samples are expected for good-quality sRNA-seq data. In addition, unique spike-in sequences are often used as an internal control. When sequences were spiked into a common background reference, the spike-in sequences should be identified as differentially expressed.

4.6.5 PREDICTION OF MICRO RNA IN THE GENOME

miRNA-seq allows identifying and predicting miRNAs in the genome, profiling their expression, elucidating disease associations, and discovering other novel miRNAs. Homology modeling or ab initio methods are commonly applied. Homology-based methods rely on available and experimentally validated miRNAs, which limits the prediction of novel miRNAs. The ab initio approach overcomes this limitation. Both the homology and the ab initio approaches use algorithms for predicting RNA secondary structures. Homology-based tools such as MapMi [94] map known miRNA to genomes taking sequence similarity and RNA secondary structure into account. The ViennaRNA package [95] predicts RNA secondary structures by generating scores, ranks, and graphical outputs of possible hairpins. Additionally, phylogenetic conservation and filtering of other known sRNA classes for detection improve prediction. Ab initio miRNA prediction requires only the primary sequence and operates in a single-sequence or multiple-sequence mode. The single-sequence mode assumes a mature miRNA is formed from the stem of a hairpin with many possible Watson−Crick pairs and few loops. Alternatively, the degree of conservation of the sequence in related species, the presence of

potential cleavage sites of Drosha and DICER, and the thermodynamic stability of the hairpins strengthen the prediction. RNAmicro [96] is a tool that detects secondary structures in a multiple sequence mode. The predicted miRNA can be further validated by evaluating the target mRNAs.

4.6.6 PREDICTING MICRO RNA TARGETS

The expression of miRNA is highly dependent on time, tissue-type, biotic/abiotic stimulus, or developmental stage. miRNAs are therefore potential markers for disease diagnosis. The miRNAs bind to target mRNAs with complementary sequences. These targets may be predicted computationally considering the seed region of the miRNA. The methods may be sequence, structure, or homology based. Because of the short length of miRNAs, sequence complementarity and homology-based predictions yield high false-positives. Structural predictions consider the thermodynamic stability of the miRNA−mRNA duplex along with sequence complementarity. PicTar [97] relies on miRNA conservation across species to identify targets using multiple sequence alignments of the 3′UTR of eight vertebrates. However, programs such as MicroTar [97] do not rely on conservation but use thermodynamic energies of miRNA−mRNA duplexes for predictions. Among other tools, RNA22 [98] searches for patterns in the 3′UTR to predict targets.

To identify miRNAs, sequences can be aligned with annotated miRNAs searched against the latest version of databases such as miRBase [99] or other sRNA databases. miRBase is the gold standard database that catalogs over 28,645 miRNA entries representing hairpin precursor miRNA expressing about 35,828 miRNA products in 223 species. miRBase provides read data-associated annotated miRNAs, allows the filtering of reads by experiment and count and searches for miRNAs by tissue- and stage-specific expression. A manually curated database of miRNA in various human diseases is provided by miR2Disease [99]. Each entry in miR2Disease contains detailed information on a miRNA−disease relationship, including miRNA ID, disease name, a brief description of the miRNA−disease relationship, the miRNA expression pattern in the disease state, the detection method for miRNA expression, experimentally verified miRNA target gene(s), and related literature. The sRNA-RNA-seq alignment program miRge [100] uses a three-step approach to handle unaligned reads from miRNA-seq. First unaligned reads are aligned to full hairpin miRNA library >25 bp. The resulting unaligned sequences are aligned to other noncoding RNA libraries such as tRNA, snoRNA, and rRNA, allowing for a single mismatch. The rest of the unmatched sequences are aligned to coding RNA allowing only identical matches. Finally, to identify isomiRs, unaligned sequences are again aligned to known miRNAs with less stringent criteria to identify isomiRs. In this approach, alternative alignments to other sRNAs are excluded before classifying the sRNA as isomiR, making it not only accurate but also rapid.

A list of common tools available for miRNA data analysis is shown in Table 4.3. miRDeep [101] is limited to organisms such as human, with known reference genomes. miRanalyzer [102] has been widely applied in different organisms via a Web server tool to detect all known miRNAs annotated in miRBase, finding perfect matches against other libraries and predicting novel miRNAs. miRExpress [103] can be used when the reference genome sequence is not available. miRscan helps identify miRNA genes conserved in more than one genome [104]. miRseeker relies on conservation of sequence and structural features across species to predict miRNAs [105]. DSAP [106] is an automated multitask Web service that facilitates comparative miRNA analyses, such as differential expression, cross-species distribution, and phylogenetic distribution. mirTools [107] provides detailed annotation

Table 4.3 Tools Available for Micro RNA (miRNA) Data Analysis

miRNA Tool	Source (URL)	Function
MiRscan	http://genes.mit.edu/mirscan/	miRNA gene profiling
MiRFinder	http://www.bioinformatics.org/mirfinder/	Expression profiling
miRDeep	http://www.australianprostatecentre.org/research/software/mirdeep-star	Profiling limited to available genomes
miRge	https://github.com/BarasLab/miRge	Profiling and discovery
miRanalyzer	https://github.com/shenlab-sinai/miRNA_pipeline_for_miRanalyzer	Profiling and discovery
miRExpress	http://mirexpress.mbc.nctu.edu.tw/	Expression profiling
mirTools	http://www.wzgenomics.cn/mr2_dev/index.php; http://centre.bioinformatics.zj.cn/mirtools/	Profiling and discovery
miRNAkey	http://ibis.tau.ac.il/miRNAkey/	Expression profiling
ViennaRNA Package	https://www.tbi.univie.ac.at/RNA/documentation.html	RNA secondary structure prediction

for known miRNA and allows determination of the relative expression level of all miRNAs, which can be illustrated using a scatter plot. miRNAkey [108] has a user-friendly graphical user interface that can be used for visualizing differentially expressed miRNAs in paired samples. Among other tools available to study sRNAs, iSmaRT (integrative Small RNA Tool-kit) focuses on predicting novel piRNAs and their RNA targets [109].

4.6.7 APPLICATION OF MIRNA-SEQ IN CANCER RESEARCH

Genetic and epigenetic defects in sRNA, particularly miRNAs, or their processing have been linked to many human diseases [110]. Over the past decade, there have been increasing reports of miRNA dysregulation in cancer, neurological, cardiovascular, and developmental disorders. In vitro and in vivo models have shown global miRNA repression in cellular transformation and tumorigenesis [86]. Calin and coworkers [111] associated chronic lymphocytic leukemia (CLL) with the deletion of a section of chromosome 13—containing genes for *miR-15* and *miR-16*. In a majority of CLL cases, these two genes are deleted or downregulated. In cells derived from breast, prostate, lymphoid, and colorectal tumors, *miR-143* and *miR-145* are downregulated [112]. Approximately 60% of human protein-coding genes are targeted by miRNAs, as predicted by computational methods.

Circulating miRNAs have surfaced as useful biomarkers in early diagnosis and monitoring cancer progression in a noninvasive way. Serum *miRNA-141* has been shown to distinguish prostate cancer from healthy controls with 60% sensitivity and 100% specificity, confirming sRNAs as accurate blood-based markers [113]. Likewise, a low ratio of plasma *miR-92a/miR-638* levels can be indicative of leukemia. Taken together, miRNAs have potential use as biomarkers for tumor evaluation. However, the prediction of miRNA—disease associations cannot solely rely on experimental methods because of limitations of time, money, and samples, as well as a lack of specific endogenous

normalizers. Therefore, computational approaches that integrate multiple biological information to predict miRNA–disease association are critically necessary.

4.6.8 CONCLUSION

NGS is a useful tool for identifying sRNA species using computational analysis. The steps described for the analysis of sRNA data have been applied in organisms such as *Arabidopsis*, humans, mice, *Drosophila*, *Caenorhabditis elegans*, etc., to understand their roles and differences in the regulatory mechanisms among different species of sRNA. The deregulation of sRNAs such as miRNAs is related to abnormal epigenetic patterns, including DNA methylation and histone modification. For instance, promoter demethylation induces reactivation of the oncogenes in lung carcinoma, aberrant hypermethylation, and inactivation of *miR-9-1* in human breast cancer. Elucidating the sRNAs, targets and interactions can be a useful tool for designing disease diagnostic, prognostic, and therapeutic tools. RNAi-based mechanisms can be manipulated via RNA interference mechanisms to control the cell cycle. The universe of novel sRNAs still remains to be discovered and their functions unraveled in the light of NGS technologies.

4.7 PROFILING CHROMATIN ACCESSIBILITY USING ATAC-SEQ

Chromatin is composed of arrays of nucleosomes, each of which consists of a histone octamer core that is wrapped by 147 bp of DNA [114,115]. The electrostatic interactions between histone proteins lead to higher-order compact DNA structures. In general, genomic regions with dense nucleosomes are more tightly packed (i.e., "closed") and less accessible to regulatory components that activate gene expression. On the other hand, genes and their promoters located in nucleosome-depleted (i.e., "open") regions are more likely expressed because the DNAs are available to interact with regulators such as TFs and enhancers that turn on gene expression.

Recent advancement in NGS has led to new techniques that enable genome-wide investigations of chromatin accessibility [116–118]. Recently, a technique called ATAC-seq was developed [119] to profile the chromatin accessibility to complement other NGS-based techniques. ATAC-seq identifies open chromatin regions and putative transcription factor binding sites (TFBSs) at single-nucleotide resolution [119–123]. Soon after its development, ATAC-seq was used as a primary method to investigate the human epigenome and regulome in the ENCODE project [124,125]. ATAC-seq requires less amounts of tissue/cells and sample-processing time. For both animals and plants, 500–50,000 cells are adequate, as opposed to the sequencing of micrococcal nuclease–sensitive sites (MNase-seq) or DNaseI hypersensitive sites sequencing (DNaseI-seq) that require at least 1000-fold more material [119,126]. In fact, single-cell ATAC-seq has been demonstrated to be possible with human and mouse cells [123,127–129], yet not without challenges. In addition to the technical difficulties of isolating intact cells or nuclei from tissues, computational challenges arise from variable capture efficiencies and PCR-induced biases. Regardless, this possibility provides a new approach to meet the challenge of current methods for medical investigation. Current methods require tens of millions of cells and cell manipulation, such as immortalization or extensive ex vivo expansion is often a necessity. As a result, the fidelity of investigation can be skewed, and individual variation cannot be addressed. The nature of high resolution and the low sample quantity required for

ATAC-seq may greatly spur advancements in personalized health care because individual genome landscape mapping is now a possibility. Because of its unique ability to analyze single-cell accessibility profiles and its short turnaround time, ATAC-seq is a promising approach for the in-clinic assessment of personal genome accessibility maps with minute clinical samples.

4.7.1 INVESTIGATING CHROMATIN ACCESSIBILITY

Before ATAC-seq, several sequencing techniques were developed in the past decade for accessibility profiling. Micrococcal nuclease (MNase) is an endo-exonuclease that digests DNA without bound proteins or secondary structures. The end product of MNase digestion is a collection of nucleosomes and hence, MNase-seq results in DNA sequences that are wrapped around the core histones [130–132]. DNA regions with a high density of MNase-seq reads represent closed chromatin. In DNaseI-seq, chromatin is treated with the DNaseI endonuclease that preferentially attacks chromatin regions sensitive to the nuclease [133–135]. Through size selection, reads from open chromatin regions can be enriched. FAIRE (formaldehyde-assisted isolation of regulatory elements)-seq [136,137] is a method to identify open regions in the genome. Formaldehyde is used to cross-link chromatin with proteins, followed by phenol–chloroform extraction used to isolate nucleosome open regions in the aqueous phase. Although MNase-seq identifies nucleosome-dense regions, DNaseI-seq and FAIRE-seq are methods used to reveal regions of open chromatin. Common technical hurdles of these techniques include high sample-quantity requirements and processing time. Specifically, MNase-seq requires a minimum of 10 million cells, and the MNase has sequence-specific biases such as AT-rich regions. DNaseI-seq also requires millions of cells, and the endonuclease used in DNaseI-seq also has sequence biases, although its cutting bias is better understood. Moreover, the optimal endonuclease digestion condition must be adjusted for a given cell type and number. With a similar requirement of a sample amount of millions of cells, FAIRE-seq generally yields a lower solution.

4.7.2 PREPARING ATAC-SEQ SAMPLES

ATAC-seq uses Tn5 transposase to enrich DNA fragments from open chromatin regions, followed by NGS. Transposases are enzymes that catalyze transposon movement and preferentially target genomic regions free from nucleosomes. Tn5 is a mutated hyperactive transposase that can simultaneously insert adapter sequences into integrated sites, eliminating additional ligation steps before sequencing (Fig. 4.7). A detailed protocol of library preparation for ATAC-seq has been reported [120]. The major steps are briefly described below: (1) Cell preparation: collection of intact cells of the target cell type; (2) Transposition reaction: cell lysis and incubation with Tn5 transposase provided by the Nextera DNA Sample Preparation Kit from Illumina; (3) PCR amplification: amplification of transposed DNA fragments, typically five cycles; (4) Quantitative PCR: determination of adequate PCR cycles before saturation. Excessive amplification results in size bias and skews toward GC-rich sequences; and (5) Sequencing sample preparation: amplification of the remaining sample from Step 3 for the number of cycles determined in Step 4. Fragment size is determined by gel electrophoresis. The size distribution should be 100–800 bp to maintain a high library complexity. Amplicons are purified using a PCR purification kit before sequencing.

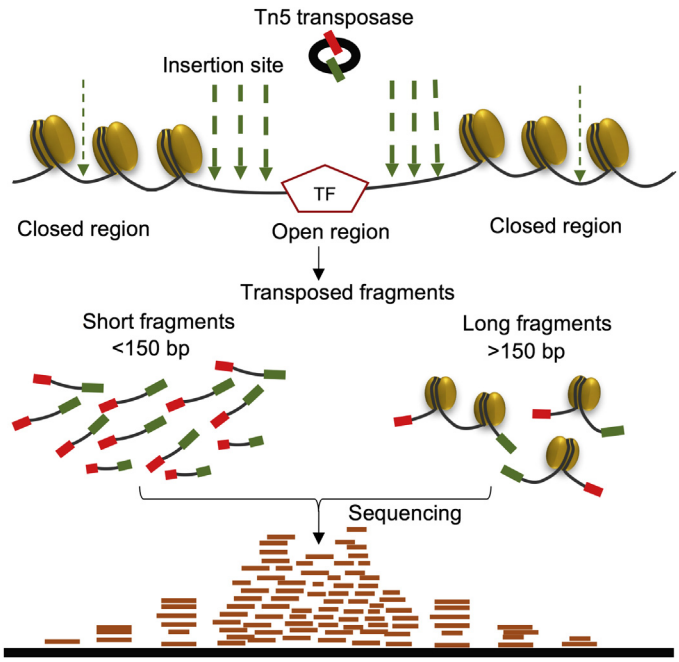

FIGURE 4.7

ATAC-seq reveals different regions of the chromatin structure. Tn5 transposition fragments chromatin. Open regions are more susceptible (*thick arrow*) than closed regions (*thin arrow*) are to Tn5 integration. Short sequences (<150 bp) are transposed fragments from the open region, and nucleosome-associated sequences are longer (>150 bp). The resultant DNA fragments are sequenced, and reads are aligned onto the reference genome. *TF*, transcription factor.

4.7.3 DETERMINING OPEN CHROMATIN REGIONS USING ATAC-SEQ

An overall analysis plan of ATAC-seq data generated by Illumina sequencers is described here, with a focus on the peak-calling step. The read length of fragments originated from open regions is primarily subnucleosomal, approximately $< \sim 150$ bp. The others are longer than 150 bp with characteristic nucleosome-associated periodicity; that is, the fragment size distribution of longer reads shows the enrichment of fragments spanning multiple complete nucleosome units (Fig. 4.7). Removing mitochondrial reads is an important step in ATAC-seq data analysis because high abundance of mitochondrial sequences (usually 20%−80%) is a common issue [120,138]. This issue can be alleviated by software such as the "view" function of SAMtools [15] after sequencing; reads mapped onto the mitochondria genome are removed. Alternatively, a mitochondrial read-removing protocol using the targeted cleavage of DNA fragments with CRISPR/Cas9 has been demonstrated recently [139]. The ATAC-seq sequencing libraries were treated with Cas9 enzyme and guide RNAs that target the human mitochondrial genome before sequencing. The results showed that the targeted cleavage step not only decreased mitochondrial reads by 1.7-fold but also yielded more peaks. Removing mitochondrial fragments during library preparation increases the sequencing efficiency, which also potentially reduces the sequencing cost. In most cases, paired-end sequencing is performed for ATAC-seq. Pair-end 50-cycle reads generally provide accurate alignments, and approximately 50 million mapped reads are sufficient for human samples [119]. The read start sites require adjustment because Tn5 transposase binds as a dimer and inserts adapters separated by 9 bp [140]. Generally, reads aligning to the +strand is offset by +4 bp, and reads aligning to the −strand are offset by −5 bp [119,122].

After mapping, the regions enriched with most reads can be determined by the "peak-calling" step, which is perhaps the most critical step for chromatin accessibility profiling. In ATAC-seq, the open chromatin regions are represented as "peaks" where the maximum number of reads are mapped. ATAC-seq data can reveal both small TFBSs (marked by small peaks) and larger regions of open chromatin (marked by broad peaks). Broad peaks cover broad regions of enrichment, and localized/narrow peaks span a small region of approximately 50−500 bp. In most cases, of chromatin accessibility profiling, the target open chromatin regions would be a few kilo base-pairs or longer and are presented as broad peaks. Open chromatin regions can be inferred from peaks using peak-calling tools. Common tools for region recognition or peak calling include MACS [39], ZINBA [141], Hotspot [142], Homer [63], and F-seq [143]. The MACS2 peak caller, originally designed for ChIP-seq, is a popular tool for ATAC-seq peak calling because of its versatility. It can detect both narrow and broad peaks and also takes into consideration false discovery rate (FDR) and noise. Similar to MACS2, ZINBA calls both broad and narrow regions of enrichment across a range of signal-to-noise ratios. Additionally, it accounts for factors that covary with the background or experimental signal. Hotspot is a tool for identifying the local enrichment of reads mapped to a genome using a binomial distribution model. It can detect regions of enrichment of variable sizes and automatically normalizes for large regions of elevated read levels because of features such as high-copy numbers. Homer was developed to find short (8−12 bp) motifs in large-scale genomic data and is mostly used for ChIP-seq analysis. F-seq is a Java package that detects continuous read density estimation and identifies regions of higher density. F-seq was used to identify broad regions in the ENCODE project, whereas Homer was used to call localized peaks. Recently, an R module was made available that implements ENCODE's ATAC-seq pipeline including F-seq, HOMER, and MACS2 with data visualization using R (available on GitHub).

DPC measures the relative abundance of the ATAC-seq reads of the same genomic region between two samples. This can be achieved by merging replicates (bam files) within each group with less stringent criteria (P-value < 0.1) to obtain union peaks across two biological replicates. The peaks or the center of peaks on summit with $\pm$200 bp can then be used to count reads of each peak from all replicates independently using featureCounts of the DESeq tool. The resulted matrix can be then fed to DESeq2 to generate differential peaks [144−146].

4.7.4 QUALITY ASSESSMENT OF ATAC-SEQ DATA

In the preliminary assessment of the sequencing results, composite plots are used to visualize read abundance as a function of the distance to a particular genetic feature. An increase in read abundance at positions corresponding to accessible regions indicates a good library. For example, TSSs have been demonstrated to be accessible chromatin locations. Hence, DNaseI-, FAIRE- and ATAC-seq data are expected to show an overall increase in abundance at these locations, whereas a decrease at TSSs is expected for MNase-seq data. For ATAC-seq specifically, an additional size distribution plot of inserts (i.e., fragments resulting from Tn5 transposition), can be generated using Picard tools. The size distribution of inserts in a successfully prepared library depicts an array spanning five to six nucleosomal units.

Reads mapped to mitochondria ideally should be less than 50% to ensure a better outcome from the peak-calling process. Furthermore, comparing ATAC-seq peaks with existing DNaseI-seq or FAIRE-seq data to obtain consensus peaks or region will benefit downstream analyses. A comparison of

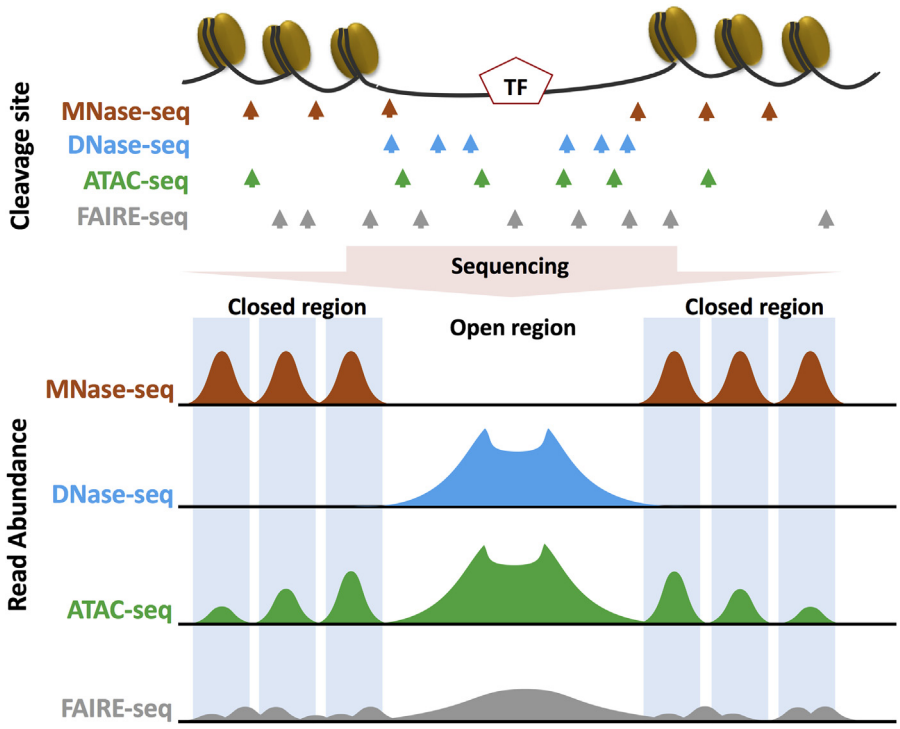

FIGURE 4.8 Assessing Chromatin Accessibility With Different Next-Generation Sequencing (NGS) Techniques.

Comparison of read abundance plots generated by different NGS techniques for chromatin accessibility. MNase-seq identifies nucleosome-associated DNA sequences. Both ATAC-seq and DNaseI-seq identify nucleosome-depleted open chromatin regions as indicated by the broad peak in the center. Differently from DNaseI-seq, ATAC-seq results in additional small peaks near the open region; these peaks cover neighboring nucleosome-dense (closed) regions because size selection is not required for ATAC-seq and hence nucleosome-associated fragments are included in the library. Short peaks of low abundance in FAIRE-seq represent the background noise. *Arrows* stand for cutting/fragmentation sites. *PCR*, polymerase chain reaction; *TF*, transcription factor.

sequencing results generated by these techniques for chromatin accessibility is illustrated in Fig. 4.8. MNase-seq identifies the nucleosome regions. Both DNaseI-seq and ATAC-seq identify nucleosome-depleted open chromatin regions. Notably, small peaks near the open region in ATAC-seq are nucleosome-associated sequences included in the library when no size selection is performed. FAIRE-seq results in a higher noise-to-signal ratio.

When interpreting peak-calling results, it should also be considered that, similar to the exonuclease used in DNaseI-seq, the Tn5 transposase used in ATAC-seq library construction also cleaves DNA in a sequence-dependent manner [116,118]. The peak-calling results in bed files generated together with bedgraphs can be visualized in the Integrative Genome Viewer (IGV) or the UCSC Genome Browser. BEDTools [83] can then be used to look for additional features on the genome.

4.7.5 APPLICATION OF ATAC-SEQ IN CANCER RESEARCH

ATAC-seq has been used to discover the association between cancer subtypes and gene chromatin accessibility. CLL is a common form of leukemia. The progression of CLL shows great heterogeneity that is correlated with commonly used clinical biomarker IGHV (Ig heavy chain V-III region VH26) genes [147,148]. Mutated IGHV genes lead to a less aggressive subtype of CLL (mCLL), and patients without the mutation show a more aggressive subtype (uCLL). An association of CLL subtypes with chromatin accessibility of IGHV genes was discovered by Rendeiro and Schmidl [149] using ATAC-seq. Peak calling was performed with MACS2. Peaks overlapping blacklisted features, i.e., the artifact signal in certain regions of the genome were discarded. Principal component analysis of chromatin accessibility, i.e., accessible regions, clearly identified the IGHV mutation status as the major source of heterogeneity in chromatin accessibility among CLL samples. In addition, it was also found that the variance and distribution of chromatin accessibility across samples of different subtypes were highly gene specific, particularly in cases of CLL-linked genes. Furthermore, TFBSs that coincide with the "dip" of ATAC-seq peaks were identified [119]. This study demonstrated that ATAC-seq serves as valuable technique to discover the association between disease subtypes and chromatin accessibility of specific genes, as well as subtype-specific regulatory elements.

ATAC-seq has also been used to discover changes in accessible regulatory regions in cancerous tissue induced in *Drosophila* [150]. To identify differentially active regulatory regions, the chromatin accessibility map was derived from the peak-calling results performed by the MACS2. With the gene set enrichment analysis, more than 3000 (over-) activated regulatory regions were identified during tumor development, including promoters, enhancers, and insulators. Together with motif discovery for candidate TFs, AP-1 and Stat92 E were found to be key regulators. The complementation of the tumor phenotype by introducing a loss-of-function Stat92 E mutant validated the importance of Stat92 E in tumor development. In addition, nearby target genes of these newly accessible regions are up- or downregulated, suggesting that these are functionally significant regulatory changes. In this case, ATAC-seq was used to identify TFs and regulatory regions driving in vivo tumor development.

4.7.6 CONCLUSION

ATAC-seq is a sensitive method with nucleotide resolution to identify open chromatin regions and putative TFBSs. When combined with other sequencing results, ATAC-seq serves as a great tool to reveal chromatin changes in cells and identify new regulatory elements of diseases. With the recently developed single-cell ATAC-seq technique for human cells, ATAC-seq is a promising new approach for disease diagnosis because of its potential for in-clinic assessments of personal genome accessibility maps with minute clinical samples. The development of robust computational tools specifically tailored for ATAC-seq data analysis is crucial for further advancement and applications of ATAC-seq in medicine.

4.8 CHROMOSOME CONFORMATION CAPTURE

The spatial organization of chromatin can be characterized by 3C techniques. In the nucleus, different genomic loci can be brought nearby in 3-D space by DNA-binding proteins. Hi-C can profile these chromosomal interactions genome-wide and identify TADs within which physical interactions are relatively frequent [151].

The chromatin state of a cell can be preserved by formalin fixation, during which the protein and DNA are cross-linked. The chromatin is then digested with a restriction enzyme, and the DNA ends are filled-in with biotin-labeled nucleotides. After proximal ligation, the DNA ends in the same TADs are ligated with each other. The following reverse cross-link removes proteins from the DNA, and the DNA is sheared into 300−500 bp suitable for NGS platform. Biotin-labeled fragments are pulled down with streptavidin beads, ligated with sequencing adapters, and PCR-amplified. In addition to Hi-C, chromatin interaction analysis with paired-end-tag sequencing (ChIA-PET) can determine the genome-wide chromosomal interactions involved in a particular protein used in the first step of ChIP [152].

The ideal sequencing data from Hi-C and ChIA-PET are "chimeric DNA," that is, the two ends of a read come from different genomic loci. Therefore, paired-end sequencing data is required to decipher information from both ends of a read. After quality control of reads (FastQC and adapter trimming), read1 and read2 are aligned to the reference genome separately. From the paired alignments, i.e., SAM files from the Bowtie2 aligner, the genomic loci of each end of a read can be identified. The intrachromosomal interaction (read1 and read2 come from the same chromosome but are at a distance) is usually more than interchromosomal interactions (read1 and read2 come from different chromosomes) in amount. The orientation of read1 and read2 should be noted. The correct Hi-C read is with read1 and read2 converging to the center of the read, a restriction enzyme cutting site, and the sum of the distance from read1 and read2 to their closest restriction enzyme cutting site should be smaller than the read length. To construct the chromosomal contact matrix, tools such as HOMER [153] and HiTC [154] are developed, and an R package diffHic could detect differential interacting regions [155].

CTCF as a methylation-sensitive insulator was found to be located on the boundary of TADs and separated these gene activity coordinated chromosomal units [151]. In 2016, Flavahan et al. provided evidence that CTCF insulation prevents the activation of oncogenes by distal enhancer elements from different TADs [156]. The authors also found that mutations in the gene *IDH1* (isocitrate dehydrogenase 1) increase the number of methyl groups on CTCF-binding sites and lose the TAD boundary. These results suggest that DNA methylation and chromosome conformation are highly associated and play important roles in gene regulation beyond genome sequence.

4.9 INTEGRATION OF EPIGENOME DATA

Epigenome components together orchestrate gene expression dynamics. For example, transcriptionally active or silent chromatin regions are often marked by DNA methylation and particular histone modifications, associating with specific TADs. Feng et al. coordinated ChIP-seq of histone modifications and Hi-C to reveal that in *Arabidopsis* local chromosome interactive domains are correlated with H3K27me3 and H3K9me2 [157]. Furthermore, in mutants of DNA methylation−related genes, chromatin interaction patterns are altered. These suggest that DNA methylation and histone modification might affect chromosome conformation.

sRNA, DNA methylation, and chromatin states might have confounding effect on gene expression. Hsu et al. showed that in maize, gene promoter regions are often with a feature of increasing CHH methylation, open chromatin, and enriched of sRNA target sequences [158]. Integrative analysis therefore becomes a powerful method to reveal the cross talk among different epigenetic mechanisms.

4.10 PREDICTING TRANSCRIPTIONAL FACTOR BINDING SITES WITH EPIGENOMICS DATA

4.10.1 EPIGENOMIC REGULATION

Gene transcription governs the gene expression network and is orchestrated by the interplay of proteins located in DNA regulatory regions. Binding of TFs is a key step in regulating gene transcription, which directly activates/represses gene expression. The 50−1500 bp *cis*-regulatory elements to which proteins such as TFs tend to bind and initiate gene transcription are called enhancers. Therefore, identifying the location of TFBSs and enhancers can help elucidate the gene regulation network. With the advances in NGS technology, ChIP-seq allows the profiling of protein−DNA binding sites, but this technique is limited by the requirement for high-quality antibodies and laborious experimental procedures [159]. Normally, TFBS can be discovered by peak calling from ChIP-seq data, that is, one experiment for one TF. To save the laborious lab works, more and more research has been published to use epigenomic data for TFBS prediction, e.g., BS-seq, DNaseI-seq, and ATAC-seq, and several computational approaches have been therefore developed to predict TFBSs and enhancers in silico [160].

4.10.2 HIT-BASED TRANSCRIPTION FACTOR BINDING SITE PREDICTION

Typically, TFs bind to short DNA sequences (4−10 bp) across the genome. These short DNA sequences are TFBSs. ChIP-seq and HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment) [161] can precipitate several protein-binding DNA fragments. Position weight matrix (PWM) is a scoring matrix composed of the log likelihood of each nucleotide in a motif (TFBS). With multiple sequence alignment, the PWM is derived from the most frequent TFBSs of a specific TF. Traditionally, PWMs are used in hit-based methods to predict TF-binding. PWM scores are calculated across the genome, and regions beyond the defined threshold are candidate TFBSs. Databases such as JASPAR [162] and TRANSFAC [163] contain PWMs of several TFs, and motif-searching tools such as FIMO [164] implement a statistic model to filter out potential TFBSs in the query DNA sequences. The construction of PWM is based on sequence specificity; therefore, the hit-based method only works well for TFs with a highly specific binding motif. Furthermore, the hit-based method is insufficient to determine whether or not a predicted TFBS is bound by TF in vivo; in other words, tissue-specific TF-binding events are often ignored.

4.10.3 SITE-CENTRIC TRANSCRIPTION FACTOR BINDING SITE PREDICTION

To further differentiate the tissue-specific TF-binding events, recent studies suggest that TF binding is associated with epigenetic signatures such as nucleosome positioning [165], histone modifications [166,167], hypersensitivity to DNaseI (DHS) [134,167] and DNA methylation [168]. The site-centric method requires the result from hit-based motif search to be integrated with other epigenetic information to improve TFBS prediction accuracy and reveal "true-binding" TFBSs. For example, histone marks H3K4me3 and H3K4me1 highlight active promoter and enhancer elements [169], and active TFBSs occur between two regions showing high active histone marks (peak-dip-peak pattern) [170]. With this information, Pique-Regi et al. first detected potential TFBSs with motif searching and then

applied a machine-learning approach to categorizing potential TFBSs into active (with a peak-dip-peak histone pattern) or inactive (without a peak-dip-peak histone pattern) groups [171]. The same principle is applied to DNaseI-seq [160,172] data to integrate open chromatin information, as well as BS-seq data [173] to include DNA methylation as a prior.

4.10.4 SEGMENTATION-BASED TRANSCRIPTION FACTOR BINDING SITE PREDICTION

DNaseI-seq reveals the open chromatin regions across the genome, which are the regions from peak calling. On top of the peaks, a peak-dip-peak region called a "footprint" is thought to be caused by TF binding [174]. In contrast with site-centric methods, which search the entire genome with PWM for potential TFBSs, segmentation-based approaches screen for the footprints with the HMM or sliding window [133,137,175]. This step segments the genome into open and closed chromatin and therefore restricts the searching space for active TF binding. These tools are called "footprinters," and several statistical methods are implemented to model or improve footprint calling.

TFBS prediction has been a popular topic in bioinformatics for a long time, and epigenetic modifications such as chromatin accessibility and histone modification are important factors to enhance the prediction models. Coupling with new techniques, such as ATAC-seq, and the increasing amount of data from ChIP-seq and DNaseI-seq in the ENCODE project may enable a better modeling scenario. Considering data integration, this is a good fit for integrative epigenomic data analysis.

4.11 CASE STUDIES OF EPIGENETICS IN ASSISTED REPRODUCTIVE TECHNOLOGY

Fertility disorders challenge reproduction. Assisted reproductive technology (ART) uses clinical/laboratory techniques on gametes and embryos for reproduction. The common techniques include in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI) (Fig. 4.9).

Although the techniques solve fertility issues, several studies also implicate epigenetic defects in children conceived by ART [176]. Disorders such as BWS, Russell−Silver syndrome (RSS), and Asperger syndrome (AS) occur from epigenetic errors in imprinting or development [177,178]. The general procedure involves fusion of the egg and sperm outside the body, which raises the possibility of epigenetic modifications such as DNA methylation. Epigenetic studies using high-throughput technologies can reveal ART risk factors with greater accuracy and reliability. This section reviews cases of ART-associated epigenetic modifications with a special focus on DNA methylation.

4.11.1 IN VITRO FERTILIZATION-ASSOCIATED TRANSCRIPTOMIC CHANGES

Canovas et al. demonstrated differences in genome-wide expression among in vivo−produced and in vitro−produced pig blastocysts using RNA-seq. Single blastocysts were used from one in vivo group and two in vitro groups. One of the in vitro groups was treated with natural reproductive fluid (Nature-IVF) and the other without reproductive fluid (C-IVF) [179]. The RNA-seq analysis identified 787 DEGs between the in vitro without reproductive fluid (C-IVF) and in vivo, and 621 DEGs between in vitro with reproductive fluid (Nature-IVF) and in vivo. All of these genes were significantly different

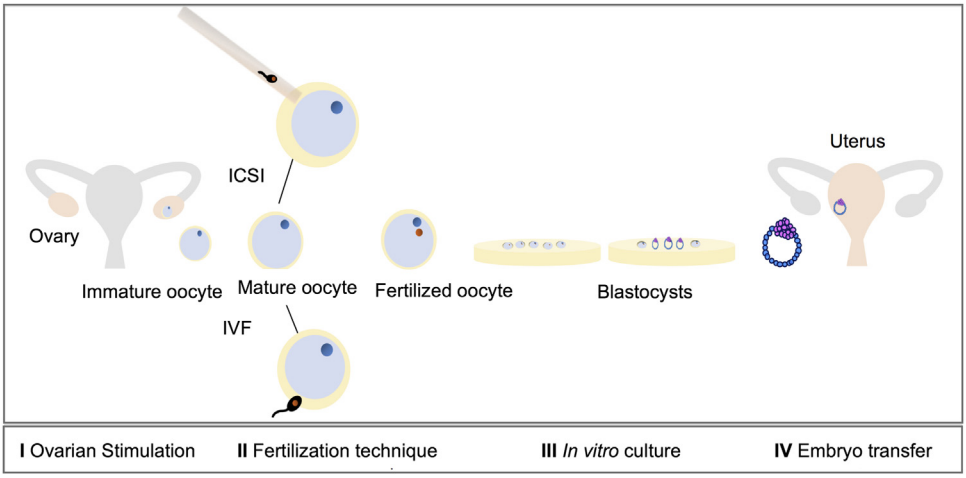

FIGURE 4.9

Schematic of the steps in ART for in vitro fertilization (IVF). *ICSI*, intracytoplasmic sperm injection; *ART*, assisted reproductive technology.

in pair-wise comparisons (adjusted P-value < 0.05, fold change of expression > 1.5). There was a higher number of upregulated genes (534/787—68% in C-IVF embryos and 431/621—69% in Nature-IVF) than downregulated ones. Among these DEGs, there are 334 genes found in both Nature-IVF and C-IVF groups versus in vivo, and several of them are associated with epigenetic reprogramming (down: *DNMT3B*, *DNMT1*; up: *HDAC5*, *KDM5A*), embryo development (down: *CTGF*, *ING2*, *KIT*, *EZH2*; up: *BMP4*, *TLN1*, *ADAR*), cell growth (down: *CDCA5*, *SMC1A*; up: *RB1*, *SMARCA2*), or imprinting (up: *IGF2BP2*, *GNAS*; down: *DIRAS3*). These data indicate that in vitro culture alters embryonic gene expression.

4.11.2 IN VITRO FERTILIZATION-ASSOCIATED DNA METHYLATION CHANGES AT IMPRINTED LOCI

WGBS of single blastocysts from a pig showed differences between in vivo–produced and in vitro–produced embryos [179]. The BS-seq libraries were sequenced at three samples per lane using the Illumina HiSeq 1000. The number of unique alignments in the samples ranged from 13 million to 42 million. In addition, the global methylation level of CpGs were $15.02 \pm 3.3\%$, $11.09 \pm 2.6\%$, and $12.33 \pm 3.6\%$ for the in vitro groups in vitro groups (C-IVF), in vitro groups treated with natural reproductive fluid (Nature-IVF) and in vivo groups, respectively.

The methylation level of the C-IVF group is the highest, suggesting that ART can lead to changes in DNA methylation. This result is consistent with a previous study in which ART-derived

blastocysts had higher methylation levels than did in vivo–derived blastocysts [180]. The global methylation level of <15% suggested few DMRs. For this reason, and to obtain an unbiased measure of differences in genome methylation, they analyzed a tile size of approximately 3 kb defined in SeqMonk. SeqMonk enabled the visualization and analysis of mapped data with 150 CpGs in each tile. After removing tiles without data, 258,885 tiles were extracted from all the samples.

In every pair-wise comparison regions of greater than 5% difference in absolute methylation between all replicates were identified as DMRs, followed by a *t*-test (Benjamini and Hochberg adjusted $P < 0.05$). The result showed 1660 hyper-DMRs and 1901 hypo DMRs between in vivo versus Nature-IVF. Furthermore, 2244 hyper-DMRs and 1511 hypo DMRs were found between in vivo and C-IVF. Among these DMRs, *IGF2R*, a gene differentially methylated in the C-IVF group, is related to the large offspring syndrome, as indicated by the analysis using Ingenuity Pathway Analysis (IPA). After this finding, the researchers focused on targeted imprinted genes. The DMRs of imprinted genes (igDMRs) are expected to maintain constant methylation before implantation embryos. This is to ensure reliable imprinted expression of the associated genes throughout development. In addition, they compared 10 candidates for imprinted region methylation of the three groups using the chi-square test. The result showed that three imprinted genes (*ZAC1*, *PEG10*, and *NNAT*) in C-IVF were more methylated ($P < 0.05$) than in the in vivo groups, and two (*PEG10* and *NNAT*) in C-IVF were more methylated than Nature-IVF. This observation indicated that in vitro culture can affect imprinted gene expression and DNA methylation. In addition, *ZAC1* and *IGF2R* have been reported to be associated with imprinting disorders—transient neonatal diabetes mellitus (TNDM) or RSS, respectively—in patients conceived by ART [181]. Thus, ART can affect the expression of imprinted genes, potentially leading to disorders such as TNDM and RSS.

4.11.3 IN VITRO FERTILIZATION-ASSOCIATED DNA METHYLATION AT INFERTILITY GENES

IVF has been shown to have a close association with infertility genes. Castillo-Fernandez et al. [181a] analyzed whole cord blood cells (WBCs) and cord blood mononuclear cells (CBMCs) from IVF and non-IVF newborn twins and used genome-wide MeDIP-seq. The libraries were subjected to highly parallel 50 bp single-end sequencing on the Illumina GAII platform. All sequencing data were checked for quality using FastQC and then mapped onto the hg19 human genome with BWA after removing duplicates, using quality score Q10 to filter data and producing the mean relative methylation score in terms of reads per million (RPM) in 500 bp bins across the genome. Approximately 11,524,145 windows were used for analysis, and more than 50% of the samples with a RPM value of zero were excluded, resulting in 9,592,803 (WBC) and 9,285,089 (CBMC) bins used in downstream analyses.

After comparing DNA methylation profiles in WBCs and CBMCs, the result showed that at a FDR of 5%, there is one significant DMR in WBCs, which was located approximately 3 kb upstream of *TNP1*, a gene reportedly linked to male infertility [182]. To explore the biological characteristics of the top-ranked results in the IVF epigenome-wide analyses, they selected a more liberal threshold of FDR

of 25%. Forty-six IVF-DMRs were included, the most strongly associated gene is *C9orf3*, a gene related to polycystic ovary syndrome and the development of erectile dysfunction after radiotherapy for prostate cancer in men [183]. The MeDIP-seq result showed that the IVF procedure may change the DNA methylation of infertility genes.

4.11.4 OTHER ART-ASSOCIATED EPIGENOMIC CHANGES

A study comparing in vivo and in vitro conceived F1 mice using ChIP assay showed an increase of lysine 4 methylation (dimethyl Lys4-H3) on the paternal chromatin and a gain in lysine 9 methylation (trimethyl Lys9-H3) on the maternal chromatin at a CTCF site in the imprinting control region [184]. This specific CTCF site also displays de novo DNA methylation on IVF, indicating the link between histone modification and DNA methylation. In addition to IVF, ICSI has been found to cause aberrant chromatin remodeling/decondensation of the male pronucleus in different animal species, such as human [185], monkey [186], and cattle [187].

Recently, the role of miRNAs in IVF has been investigated using a mouse model [188]. Comparative miRNA profiling between embryos resulted from in vivo and in vitro fertilization revealed that dysregulated miRNAs in IVF were mainly associated with carcinogenesis, genetic information processing, glucose metabolism, cytoskeleton organization, and neurogenesis. A specific miRNA, miR-199a-5p, was found consistently downregulated in IVF embryos, and the IVF-induced downregulation in this miRNA was shown to directly result in an elevated glycolytic rate, cell lineage misallocation, and lower fetal survival postimplantation.

NGS has been efficiently used to reveal changes in the transcriptome and epigenome in ART-conceived individuals. The discovered epigenomic changes in DNA methylation, histone modification, chromatin structure, or miRNA point to potential risks of the ART procedure. By understanding mechanism by which the epigenome is affected by the current methods, changes in the procedure can be made to prevent undesired consequences and to improve the success rate of ART.

4.12 SUMMARY

More and more NGS approaches have been developed to study epigenome modifications. Most approaches introduced in this chapter (Table 4.1) compute sequencing read abundance to determine relative quantifications of the epigenomic feature; these methods include MeDIP-seq, ChIP-seq, ATAC-seq, FAIRE-seq, DNase-seq, MNase-seq, miRNA-seq, Hi-C, and ChIA-PET. On the other hand, the bisulfite sequencing–based methods in profiling DNA methylation (WGBS, RRBS) analyzed the fractions of reads that are methylated as an absolute quantification of the percentage of methylated cells in the sample population.

All contents in this chapter together demonstrate that emerging epigenomic NGS data make the gene regulatory network more complete, and bioinformatics not only helps us address biological questions from these data but also makes predictions. With the rapidly increasing amount of sequencing data, development of more powerful and integrative bioinformatics tool is a necessity for effective analysis.

LIST OF ABBREVIATIONS

3C	Chromosome conformation capture
5mC	5-Methylcytosine
ART	Assisted reproductive technology
AS	Asperger syndrome
ATAC-seq	Assay for transposase-accessible chromatin with high-throughput sequencing
BWS	Beckwith–Wiedemann syndrome
BS-seq	Bisulfite sequencing
CGI	CpG island
ChIA-PET	Chromatin interaction analysis with paired-end-tag sequencing
ChIP	Chromatin immunoprecipitation
CLL	Chronic lymphocytic leukemia
DEG	Differentially expressed gene
DMR	Differentially methylated region
DMG	Differentially methylated gene
DNMT	DNA methyltransferase
DPC	Differential peak calling
dsRNA	Double-stranded RNA
ENCODE	Encyclopedia of DNA elements
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements sequencing
FDR	False discovery rate
FRiP	Fraction of reads in peaks
Hi-C	High-throughput chromosome conformation capture
HMM	Hidden Markov model
HT-SELEX	High-throughput systematic evolution of ligands by exponential enrichment
ICSI	Intracytoplasmic sperm injection
IDR	Irreproducible discovery rate
IPA	Ingenuity pathway analysis
IVF	In vitro fertilization
MeDIP-seq	Methylated DNA immunoprecipitation sequencing
miRNA	Micro RNA
MRE-seq	Methylation-sensitive enzyme sequencing
MNase-seq	Micrococcal nuclease sequencing
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
PWM	Position weight matrix
RRBS	Reduced representation bisulfite sequencing
RSS	Russell–Silver syndrome
sRNA	Small ribonucleic acid
siRNA	Small interfering RNA
ssRNA	Single-stranded RNA
TAD	Topologically associating domain
TF	Transcription factor
TFBS	Transcription factor binding site
WBC	Whole cord blood cell
WGBS	Whole genome bisulfite sequencing

ACKNOWLEDGMENTS

We are grateful to Academia Sinica and the Taiwan Ministry of Science and Technology (MOST-103-2313-B-001-003-MY3, MOST-104-2923-B-001-003-MY2 and MOST-103-2633-B-001-002) and the National Health Research Institutes (NHRI-EX103-10324SC) for their financial support.

REFERENCES

[1] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010;466(7303):253−7.
[2] Zhao MT, Whyte JJ, Hopkins GM, Kirk MD, Prather RS. Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA. Cell Reprogram 2014; 16(3):175−84.
[3] Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc 2015;10(3):475−83.
[4] Smith ZD, Gu H, Bock C, Gnirke A, Meissner A. High-throughput bisulfite sequencing in mammalian genomes. Methods 2009;48(3):226−32.
[5] Misteli T. Higher-order genome organization in human disease. Cold Spring Harb Perspect Biol 2010;2(8): a000794.
[6] Mattick JS, Makunin IV. Small regulatory RNAs in mammals. Hum Mol Genet 2005;14(suppl_1): R121−32.
[7] bcl2fastq Conversion Software 2017. Available from: https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.
[8] BaseSpace 2017. Available from: https://basespace.illumina.com/home/index.
[9] FastQC: a quality control tool for high throughput sequence data. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[10] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011;17(1):10−2.
[11] Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min 2015;8(1):1.
[12] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754−60.
[13] Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008; 24(5):713−4.
[14] Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res 2011;21(3):487−93.
[15] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078−9.
[16] Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. Cell 1999;99(3):247−57.
[17] Robertson KD, Keyomarsi K, Gonzales FA, Velicescu M, Jones PA. Differential mRNA expression of the human DNA methyltransferases (DNMTs) 1, 3a and 3b during the G(0)/G(1) to S phase transition in normal and tumor cells. Nucleic Acids Res 2000;28(10):2108−13.
[18] Jones PA, Liang G. Rethinking how DNA methylation patterns are maintained. Nat Rev Genet 2009;10(11): 805−11.
[19] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012; 13(7):484−92.

[20] Surani MA, Barton SC, Norris ML. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature 1984;308(5959):548−50.
[21] Barlow DP, Stoger R, Herrmann BG, Saito K, Schweifer N. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the *Tme* locus. Nature 1991;349(6304):84−7.
[22] Herrera LA, Prada D, Andonegui MA, Duenas-Gonzalez A. The epigenetic origin of aneuploidy. Curr Genom 2008;9(1):43−50.
[23] Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, et al. Genome-wide demethylation of Arabidopsis endosperm. Science 2009;324(5933):1451−4.
[24] Montanini B, Chen PY, Morselli M, Jaroszewicz A, Lopez D, Martin F, et al. Non-exhaustive DNA methylation-mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. Genome Biol 2014;15(7):411.
[25] Lisanti S, Omar WA, Tomaszewski B, De Prins S, Jacobs G, Koppen G, et al. Comparison of methods for quantification of global DNA methylation in human cells and tissues. PLoS One 2013;8(11):e79044.
[26] Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol 2010;28(10):1106−14.
[27] Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. Epigenetics Chromatin 2016;9:26.
[28] Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG. Whole-genome DNA methylation profiling using MethylCap-seq. Methods 2010;52(3):232−6.
[29] Weng YI, Huang TH, Yan PS. Methylated DNA immunoprecipitation and microarray-based analysis: detection of DNA methylation in breast cancer cell lines. Methods Mol Biol 2009;590:165−76.
[30] Orozco LD, Morselli M, Rubbi L, Guo W, Go J, Shi H, et al. Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice. Cell Metab 2015;21(6):905−17.
[31] Li Q, Suzuki M, Wendt J, Patterson N, Eichten SR, Hermanson PJ, et al. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. Nucleic Acids Res 2015;43(12):e81.
[32] Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008;452(7184):215−9.
[33] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462(7271):315−22.
[34] Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, et al. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. Genome Res 2013;23(10):1651−62.
[35] Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: *de novo* DNA methylation in near-gene chromatin regulation in maize. Genome Res 2013;23(4):628−37.
[36] Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat Genet 2015;47(5):469−78.
[37] Gkountela S, Zhang KX, Shafiq TA, Liao WW, Hargan-Calvopina J, Chen PY, et al. DNA demethylation dynamics in the human prenatal germline. Cell 2015;161(6):1425−36.
[38] Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A unique gene regulatory network resets the human germline epigenome for development. Cell 2015;161(6):1453−67.
[39] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.
[40] Lienhard MGC, Morkel M, Herwig R, Chavez L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. Bioinformatics 2014;30(2):284−6.
[41] Saito Y, Tsuji J, Mituyama T. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. Nucleic Acids Res 2014;42(6):e45.
[42] Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol Biol 2016;1418:283−334.

[43] Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet 2012;13(10):705−19.
[44] Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genom 2013;14:774.
[45] Liao WW, Yen MR, Ju E, Hsu FM, Lam L, Chen PY. MethGo: a comprehensive tool for analyzing whole-genome bisulfite sequencing data. BMC Genom 2015;16(Suppl. 12):S11.
[46] Hu K, Ting A, Li J. BSPAT: a fast online tool for DNA methylation co-occurrence pattern analysis based on high-throughput bisulfite sequencing data. BMC Bioinform 2015;16:220.
[47] Benoukraf T, Wongphayak S, Hadi L, Wu M, Soong R. GBSA: a comprehensive software for analysing whole genome bisulfite sequencing data. Nucleic Acids Res 2013;41(4):e55.
[48] Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 2012;13(10):R83.
[49] Akalin A, Kormaksson M, Li S, Garrett-Bakelman F, Figueroa M, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 2012; 13(10):R87.
[50] Hebestreit K, Dugas M, Klein H. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics 2013;29(13):1647−53.
[51] Kishore K, de Pretis S, Lister R, Morelli M, Bianchi V, Amati B, et al. methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. BMC Bioinform 2015;16(313).
[52] Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet 2014;15(10):647−61.
[53] Yu M, Hon GC, Szulwach KE, Song CX, Jin P, Ren B, et al. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nat Protoc 2012;7(12):2159−70.
[54] Ashktorab H, Shakoori A, Zarnogi S, Sun X, Varma S, Lee E, et al. Reduced representation bisulfite sequencing determination of distinctive DNA hypermethylated genes in the progression to colon cancer in African Americans. Gastroenterol Res Pract 2016;2016:2102674.
[55] 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.
[56] Chen S, Jiang SM, Hu F, Xu YJ, Wang T, Mei Q. Foxk2 inhibits non-small cell lung cancer epithelial-mesenchymal transition and proliferation through the repression of different key target genes. Oncol Rep 2017;37(4):2335−47.
[57] McDaniel JM, Varley KE, Gertz J, Savic DS, Roberts BS, Bailey SK, et al. Genomic regulation of invasion by STAT3 in triple negative breast cancer. Oncotarget 2017;8(5):8226−38.
[58] Vareslija D, McBryan J, Fagan A, Redmond AM, Hao Y, Sims AH, et al. Adaptation to AI therapy in breast cancer can induce dynamic alterations in ER activity resulting in estrogen-independent metastatic tumors. Clin Canc Res 2016;22(11):2765−77.
[59] Zhang Z, Shi LH, Dawany N, Kelsen J, Petri MA, Sullivan KE. H3K4 tri-methylation breadth at transcription start sites impacts the transcriptome of systemic lupus erythematosus. Clin Epigenetics 2016;8:13.
[60] Adli M, Bernstein BE. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. Nat Protoc 2011;6(10):1656−68.
[61] Brind'Amour J, Liu S, Hudson M, Chen C, Karimi MM, Lorincz MC. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat Commun 2015;6:6033.
[62] Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform 2016;17(6):953−66.
[63] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, 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.

[64] Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. 2011. Available from: http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.
[65] Chen L, Wang C, Qin ZS, Wu H. A novel statistical method for quantitative comparison of multiple ChIP-Seq datasets. Bioinformatics 2015;31(12):1889−96.
[66] Shen L, Shao N-Y, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-Seq data with biological replicates. PLoS One 2013;8(6):e65598.
[67] Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics 2011; 27(6):870−1.
[68] Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 2009;25(15):1952−8.
[69] Shao Z, Zhang Y, Yuan G-C, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol 2012;13(3):R16.
[70] Liang K, Keles S. Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics 2012;28(1):121−2.
[71] Xu H, Wei C-L, Lin F, Sung W-K. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics 2008;24(20):2344−9.
[72] Allhoff M, Sere K, Chauvistre H, Lin Q, Zenke M, Costa IG. Detecting differential peaks in ChIP-seq signals with ODIN. Bioinformatics 2014;30(24):3467−75.
[73] Heinig M, Colome-Tatche M, Taudt A, Rintisch C, Schafer S, Pravenec M, et al. histoneHMM: differential analysis of histone modifications with broad genomic footprints. BMC Bioinform 2015;16:60.
[74] Allhoff M, Sere K, Pires JF, Zenke M, Costa IG. Differential peak calling of ChIP-seq signals with replicates with THOR. Nucleic Acids Res 2016;44(20):e153.
[75] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22(9):1813−31.
[76] Hu Y, Zhang L, Zhao L, Li J, He SB, Zhou K, et al. Trichostatin a selectively suppresses the cold-induced transcription of the ZmDREB1 gene in maize. PLoS One 2011;6(7):e22132.
[77] Li QH, Brown JB, Huang HY, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann Appl Stat 2011;5(3):1752−79.
[78] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. Science 2013;339(6127):1546−58.
[79] Wilson S, Fan LL, Sahgal N, Qi JF, Filipp FV. The histone demethylase KDM3A regulates the transcriptional program of the androgen receptor in prostate cancer cells. Oncotarget 2017;8(18):30328−43.
[80] Kuroki S, Matoba S, Akiyoshi M, Matsumura Y, Miyachi H, Mise N, et al. Epigenetic regulation of mouse sex determination by the histone demethylase Jmjd1a. Science 2013;341(6150):1106−9.
[81] Yamane K, Toumazou C, Tsukada Y, Erdjument-Bromage H, Tempst P, Wong JM, et al. JHDM2A, a JmjC-containing H3K9 demethylase, facilitates transcription activation by androgen receptor. Cell 2006; 125(3):483−95.
[82] Chen T-W, Li H-P, Lee C-C, Gan R-C, Huang P-J, Wu TH, et al. ChIPseek, a web-based analysis tool for ChIP data. BMC Genom 2014;15:539.
[83] Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinform 2014; 47(11.12):1−34.
[84] Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, et al. The small RNA profile during *Drosophila melanogaster* development. Dev Cell 2003;5(2):337−50.
[85] Garber K. Small RNAs reveal an activating side. Science 2006;314(5800):741−2.
[86] Gong H, Liu CM, Liu DP, Liang CC. The role of small RNAs in human diseases: potential troublemaker and therapeutic tools. Med Res Rev 2005;25(3):361−81.

[87] Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr Biol 2003;13(10):807−18.
[88] Wang H, Ach RA, Curry B. Direct and sensitive miRNA profiling from low-input total RNA. RNA 2007; 13(1):151−9.
[89] Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. Nat Rev Genet 2012;13(5):358−69.
[90] Baker M. RNA interference: from tools to therapies. Nature 2010;464(7292):1225.
[91] Baran-Gale J, Kurtz C, Erdos M, Sison C, Young A, Fannin E, et al. Addressing bias in small RNA library preparation for sequencing: a new protocol recovers microRNAs that evade capture by current methods. Front Genet 2015;6:352.
[92] Illumina. Small RNA sample prep kit support 2017. Available from: https://support.illumina.com/sequencing/sequencing_kits/small_rna_sample_prep_kit/questions.html.
[93] Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. Brief Bioinform 2015; 16(6):950−63.
[94] Guerra-Assuncao JA, Enright AJ. MapMi: automated mapping of microRNA loci. BMC Bioinform 2010; 11:133.
[95] Hofacker IL. RNA secondary structure analysis using the Vienna RNA package. Curr Protoc Bioinform 2009 [chapter 12:unit 12.2].
[96] Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics 2006;22(14):e197−202.
[97] Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet 2005;37(5):495−500.
[98] Loher P, Rigoutsos I. Interactive exploration of RNA22 microRNA target predictions. Bioinformatics 2012; 28(24):3322−3.
[99] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 2009;37(Database issue):D98−104.
[100] Baras AS, Mitchell CJ, Myers JR, Gupta S, Weng LC, Ashton JM, et al. miRge − a multiplexed method of processing small RNA-seq data to determine MicroRNA entropy. PLoS One 2015;10(11):e0143066.
[101] Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 2008;26(4):407−15.
[102] Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 2009; 37(Web Server issue):W68−76.
[103] Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinform 2009;10:328.
[104] Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S. The microRNAs of *Caenorhabditis elegans*. Genes Dev 2003;16:991−1008.
[105] Deleted in review.
[106] Huang PJ, Liu YC, Lee CC, Lin WC, Gan RR, Lyu PC, et al. DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res 2010;38(Web Server issue):W385−91.
[107] Zhu E, Zhao F, Xu G, Hou H, Zhou L, Li X, et al. mirTools: microRNA profiling and discovery based on high-throughput sequencing. Nucleic Acids Res 2010;38(Web Server issue):W392−7.
[108] Ronen R, Gan I, Modai S, Sukacheov A, Dror G, Halperin E, et al. miRNAkey: a software for microRNA deep sequencing analysis. Bioinformatics 2010;26(20):2615−6.
[109] Panero R, Rinaldi A, Memoli D, Nassa G, Ravo M, Rizzo F, et al. iSmaRT: a toolkit for a comprehensive analysis of small RNA-Seq data. Bioinformatics 2017;33(6):938−40.
[110] Esteller M. Non-coding RNAs in human disease. Nat Rev Genet 2011;12(12):861−74.

[111] Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci USA 2002;99(24):15524−9.
[112] Michael MZ, O'Connor SM, van Holst Pellekaan NG, Young GP, James RJ. Reduced accumulation of specific microRNAs in colorectal neoplasia. Mol Cancer Res 2003;1(12):882−91.
[113] Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan E. Circulating micro-RNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci USA 2008;105(30): 10513−8.
[114] Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science 1974;184:868−71.
[115] Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. Nature 2003;423:145−50.
[116] Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin 2014; 7:33.
[117] Sheffield NC, Furey TS. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. Genes 2012;3(4):651−70.
[118] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet 2014;15(11):709−21.
[119] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 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.
[120] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 2015;109. 21.29.1−21.29.9.
[121] Ackermann AM, Wang ZP, Schug J, Naji A, Kaestner KH. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab 2016;5(3):233−44.
[122] Bao X, Rubin AJ, Qu K, Zhang J, Giresi PG, Chang HY, et al. A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. Genome Biol 2015;16:284.
[123] Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res 2015;25(11):1757−70.
[124] Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447(7146):799−816.
[125] Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA 2014;111(17):6131−8.
[126] Lu Z, Hofmeister BT, Vollmers C, DuBois RM, Schmitz RJ. Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Res 2016;45(6):e41.
[127] Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 2015;348(6237):910−4.
[128] Gury-BenAri M, Thaiss CA, Serafini N, Winter DR, Giladi A, Lara-Astiaso D, et al. The spectrum and regulatory landscape of intestinal innate lymphoid cells are shaped by the microbiome. Cell 2016;166(5):1231.
[129] Pott S, Lieb JD. Single-cell ATAC-seq: strength in numbers. Genome Biol 2015;16:172.
[130] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129(4):823−37.
[131] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132(5):887−98.
[132] He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods 2014;11:73−8.

[133] Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 2012;489(7414):83−90.
[134] Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 2009;6(4):283−9.
[135] Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 2011;21(3):456−64.
[136] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17(6):877−85.
[137] Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 2011;21(10): 1757−67.
[138] Sos BC, Fung HL, Gao DR, Osothprarop TF, Kia A, He MM, et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol 2016; 17:20.
[139] Montefiori L, Hernandez L, Zhang Z, Gilad Y, Ober C, Crawford G, et al. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. Sci Rep 2017;7(1):2451.
[140] Adey A, Morrison HG, Asan XX, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 2010;11(12):R119.
[141] Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol 2011;12(7):R67.
[142] John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet 2011;43:264−8.
[143] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132(2):311−22.
[144] Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Gruner BM, et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. Cell 2016;166(2):328−42.
[145] Varet H, Brillet-Gueguen L, Coppee JY, Dillies MA. SARTools: a DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-seq data. PLoS One 2016;11(6):e0157022.
[146] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
[147] Damle RN, Wasil T, Fais F, Ghiotto F, Valetto A, Allen SL, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood 1999;94(6):1840−7.
[148] Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V-H genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood 1999;94(6):1848−54.
[149] Rendeiro AF, Schmidl C, Strefford JC, Walewska R, Davis Z, Farlik M, et al. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. Nat Commun 2016;7:11938.
[150] Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet 2015;11(2):e1004994.
[151] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009; 326(5950):289−93.
[152] Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 2009;462(7269):58−64.
[153] HOMER. Available from: http://homer.ucsd.edu/homer/interactions/index.html.

[154] Servant N, Lajoie B, Nora E, Giorgetti L, Chen C, Heard E. HiTC: exploration of high-throughput 'C' experiments. Bioinformatics 2012;28(21):2843−4.
[155] Lun AT, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinform 2015;16:258.
[156] Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 2016;529(7584):110−4.
[157] Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Mol Cell 2014;55(5): 694−707.
[158] Hsu FM, Yen MR, Wang CT, Lin CY, Wang CR, Chen PY. Optimized reduced representation bisulfite sequencing reveals tissue-specific mCHH islands in maize. Epigenetics Chromatin 2017;10(1):42.
[159] Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. Nat Protoc 2013;8(3):539−54.
[160] Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics 2012;28(1):56−62.
[161] Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell 2013;152(1−2):327−39.
[162] Khan A, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 2018;46:D260−6. https://doi.org/10.1093/nar/gkx1126.
[163] Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 1996;24(1):238−41.
[164] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011; 27(7):1017−8.
[165] He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, et al. Nucleosome dynamics define transcriptional enhancers. Nat Genet 2010;42(4):343−7.
[166] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007;4(8):651−7.
[167] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 2007;39(3):311−8.
[168] Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, et al. Role of DNA methylation in modulating transcription factor occupancy. Cell Rep 2015;12(7):1184−95.
[169] Bell O, Tiwari VK, Thoma NH, Schubeler D. Determinants and dynamics of genome accessibility. Nat Rev Genet 2011;12(8):554−64.
[170] Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol 2009;5(11):e1000566.
[171] Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 2011;21(3):447−55.
[172] Whitington T, Perkins AC, Bailey TL. High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Res 2009;37(1):14−25.
[173] Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, et al. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. Nucleic Acids Res 2015;43(5):2757−66.
[174] Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 1978;5(9):3157−70.
[175] Won KJ, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol 2010;11(1):R7.

[176] Lu YH, Wang N, Jin F. Long-term follow-up of children conceived through assisted reproductive technology. J Zhejiang Univ Sci B 2013;14(5):359−71.
[177] Shufaro Y, Laufer N. Epigenetic concerns in assisted reproduction: update and critical review of the current literature. Fertil Steril 2013;99(3):605−6.
[178] Pinborg A, Henningsen AK, Malchau SS, Loft A. Congenital anomalies after assisted reproductive technology. Fertil Steril 2013;99(2):327−32.
[179] Canovas S, Ivanova E, Romar R, Garcia-Martinez S, Soriano-Ubeda C, Garcia-Vazquez FA, et al. DNA methylation and gene expression changes derived from assisted reproductive technologies can be decreased by reproductive fluids. Elife 2017;6.
[180] Deshmukh RS, Ostrup O, Ostrup E, Vejlsted M, Niemann H, Lucas-Hahn A, et al. DNA methylation in porcine preimplantation embryos developed in vivo and produced by in vitro fertilization, parthenogenetic activation and somatic cell nuclear transfer. Epigenetics 2011;6(2):177−87.
[181] Le Bouc Y, Rossignol S, Azzi S, Steunou V, Netchine I, Gicquel C. Epigenetics, genomic imprinting and assisted reproductive technology. Ann Endocrinol 2010;71(3):237−8.
[181a] Castillo-Fernandez Juan E, et al. DNA methylation changes at infertility genes in newborn twins conceived by *in vitro* fertilisation. Genome Medicine 2017;9:28.
[182] Miyagawa Y, Nishimura H, Tsujimura A, Matsuoka Y, Matsumiya K, Okuyama A, et al. Single-nucleotide polymorphisms and mutation analyses of the TNP1 and TNP2 genes of fertile and infertile human male populations. J Androl 2005;26(6):779−86.
[183] Kerns SL, Ostrer H, Stock R, Li W, Moore J, Pearlman A, et al. Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys 2010;78(5): 1292−300.
[184] Li T, Vu TH, Ulaner GA, Littman E, Ling JQ, Chen HL, et al. IVF results in de novo DNA methylation and histone methylation at an Igf2-H19 imprinting epigenetic switch. Mol Hum Reprod 2005;11(9):631−40.
[185] Hansen M, Kurinczuk JJ, Bower C, Webb S. The risk of major birth defects after intracytoplasmic sperm injection and in vitro fertilization. N Engl J Med 2002;346(10):725−30.
[186] Hewitson L, Simerly C, Dominko T, Schatten G. Cellular and molecular events after *in vitro* fertilization and intracytoplasmic sperm injection. Theriogenology 2000;53(1):95−104.
[187] Dozortsev D, Wakaiama T, Ermilov A, Yanagimachi R. Intracytoplasmic sperm injection in the rat. Zygote 1998;6(2):143−7.
[188] Tan K, Wang XD, Zhang ZN, Miao K, Yu Y, An L, et al. Downregulation of miR-199a-5p disrupts the developmental potential of in vitro-fertilized mouse blastocysts. Biol Reprod 2016;95(3):9.