7

Computational Epigenomics: From Fundamental Research to Disease Prediction and Risk Assessment

Mohamed-Amin Choukrallah • Florian Martin • Nicolas Sierro • Julia Hoeng • Nikolai V. Ivanov • Manuel C. Peitsch

PMI R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

Author for correspondence: Mohamed-Amin Choukrallah, PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland.
E-mail: MohamedAmin.Choukrallah@pmi.com

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch7

Abstract: Over the past two decades, rapid advances in DNA sequencing technologies have allowed genome-wide interrogation of epigenetic features. The epigenome landscape encompasses a growing number of chemical properties of DNA and DNA-associated proteins; these properties are tissue-specific, distinctive for disease state and sensitive to environmental exposures. The epigenetic field has rapidly evolved from basic research investigations, aiming to understand the nature and function of epigenetic marks, to clinical and preclinical applications, where vast epigenetic information is used for risk assessment and disease prediction. The large diversity of epigenetic marks is mirrored by the complex variability of their genomic patterns and distributions. Mining of large-scale genomic datasets relies strongly on computational approaches and statistical models that should be carefully selected and adapted to fit the nature of the signals analyzed and the hypotheses tested. Here, we review recent advances in computational approaches used to analyze epigenetic data, with an emphasis on histone modifications and DNA methylation. We discuss the standard workflows for data acquisition, processing, and transformation, as well as the computational approaches used to assess statistical significance in comparative analyses. We also discuss the

In: *Computational Biology*. Holger Husi (Editor), Codon Publications, Brisbane, Australia.
ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

prediction methods utilized to associate epigenetic modifications with human disorders and environmental factors.

Keywords: data modeling; disease prediction; DNA methylation; epigenetics; histone modifications.

INTRODUCTION

Gene expression is regulated by the interaction between DNA molecules and DNA-binding proteins such as transcription factors (TFs), coactivator, and corepressor complexes. Some of these complexes modify the chromatin structure and its transcription competency. Chemical modifications to DNA and DNA-associated proteins (histones) and non-coding RNAs are considered the main epigenetic mechanisms controlling genome activity. The term epigenetics was first coined by Conrad Waddington to describe a set of causal heritable mechanisms that translate genotypes to phenotypes (1). More recent definitions describe epigenetics as modifications that regulate gene expression without altering the DNA sequence.

Epigenetic mechanisms are generally assessed by measuring their associated chemical tags or marks on target molecules, such as the methylation of cytosine or the acetylation of histone residues. The epigenome of a cell can be defined as the combination of all epigenetic marks at a given time across the genome that synergistically dictates the usage of the underlying DNA sequence. Given the large number of known epigenetic marks, and probably a much larger number of unknown ones, as well as the limitations of current epigenomics methods, the epigenome cannot be assessed as a whole, and current studies capture snapshots of only a small fraction of it. Epigenetic marks are dynamic and reversible and can undergo rapid changes during development and in response to various exposures, including drug treatment. Epigenetic alterations are also associated with a number of diseases and can be used as diagnostic and prognostic biomarkers of disease onset and progression, respectively.

The majority of epigenetic studies seek to identify changes between experimental conditions and further leverage this information to explain other molecular or physiological alterations. The results of such comparative studies mainly rely on the statistical approach and selection criteria used. Here, we review current knowledge of epigenetic mechanisms, with a focus on DNA methylation and histone modifications. We describe the workflows used to process and interpret epigenetic data generated by next-generation sequencing technologies. We also discuss the main computational approaches applied to identify differentially regulated loci and prediction methods used to associate epigenetic changes with human disorders or environmental exposures.

DNA MODIFICATIONS

Cytosine methylation at the 5-carbon position (5mC) is the most frequent DNA modification in eukaryotes. In mammals, 5mC occurs almost exclusively in the context of CpG dinucleotides (2), with the cytosines in both strands usually being methylated. The majority of CpG sites in mammalian genomes are methylated except at active regulatory elements (REs) (3, 4). 5mC is catalyzed by DNA methyltransferases, DNMT1, DNMT3a, and DNMT3b. DNMT1 is a maintenance enzyme that ensures the inheritance of 5mC patterns during DNA replication, while DNMT3a and DNMT3b catalyze de novo DNA methylation (5). Formation of 5mC is reversible and can be converted by ten–eleven translocation (TET) enzymes to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) through three consecutive oxidation reactions (6), ultimately leading to the unmethylated cytosine. Initially, 5mC oxidation derivatives were considered as intermediates in the process of DNA demethylation. However, recent investigations indicate that they may represent distinct epigenetic states with regulatory functions (7). Although 5mC was historically associated with gene silencing, genome-wide investigations have shown that 5mC readout depends on the genomic context. While a high level of DNA methylation at REs is indicative of transcriptional silencing, gene bodies show high levels of DNA methylation regardless of their expression status (3, 8). In addition to cytosines, eukaryotic DNA can also be methylated at the nitrogen-6 position of adenosine bases (6mA) (9). In contrast to cytosine modifications, adenosine modifications have received less attention and will not be discussed in this chapter.

ASSESSMENT OF CYTOSINE MODIFICATIONS

Cytosine modifications can be assessed by various methods (10) involving two main technologies, high-throughput sequencing and methylation arrays. While methylation arrays are restricted to annotated loci such as promoters and a fraction of known enhancers, sequencing methods can potentially cover every cytosine in the genome. Whole-genome bisulfite sequencing (WGBS) is currently the gold standard technique for assessing cytosine modifications at single-base resolution across the entire genome. WGBS is based on the bisulfite reaction that converts unmodified cytosines (uCs) into uracils while 5mC and 5hmC bases are protected from the conversion (Figure 1). After DNA amplification and high-throughput sequencing, uCs are read as thymines, whereas 5mC and 5hmC are read as cytosines. Given that WGBS cannot distinguish 5mC from 5hmC (11), the measured signal represents the sum of both modifications. However, the contribution of each modification strongly depends on its relative abundance in the investigated tissue or cell type. 5hmC levels are generally very low in mammalian cells and vary across cell types and tissues. 5hmC is abundant in the brain but extremely low in blood and spleen and almost undetectable in cultured cell lines (12). 5hmC levels can be assessed by a subtractive approach through the combination of oxidative bisulfite sequencing (oxBS-seq) (13, 14) and bisulfite sequencing (BS-seq). oxBS-seq implies an oxidation step that converts 5hmC into 5fC, which is further converted by the bisulfite reaction into uracil and read as thymine after sequencing, similar to uCs. Therefore, oxBS-seq identifies real 5mC, while BS-seq identifies 5mC + 5hmC (Figure 1). Consequently, subtracting the oxBS-seq signal from the BS-seq signal allows the computation of 5hmC levels (15), on the condition that the oxBS- and BS-conversion rates are very close to 100%.

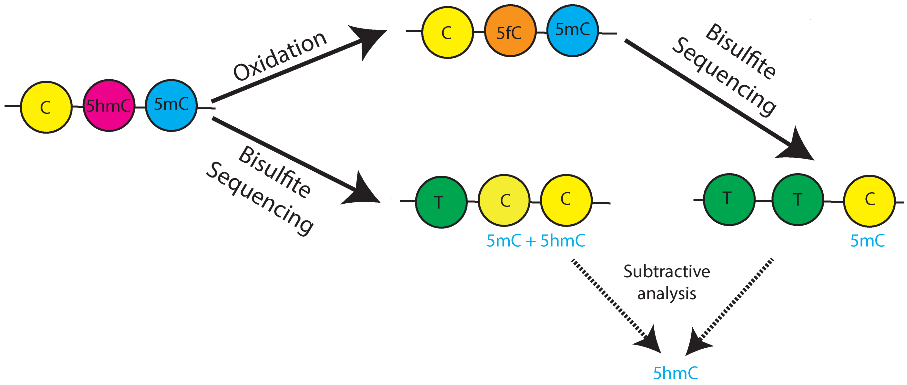

Figure 1 During bisulfite sequencing (BS-seq), unmodified cytosines (C) are read as thymines (T), while methylated (5mC) and hydroxymethylated cytosines (5hmC) are protected from bisulfite conversion and read as C. In this scenario, BS-seq does not discriminate 5mC from 5hmC. Oxidative bisulfite sequencing (oxBS-seq) includes an oxidation step, during which 5hmC is converted to 5fC and read as T after bisulfite sequencing, similar to unmodified C, while only 5mC is read as C. 5hmC proportions are computed by subtracting oxBS-seq signals from BS-seq signals.

The majority of DNA-methylation investigations are based on bisulfite conversion and assume that 5mC is the major cytosine modification and, therefore, neglect the contribution of 5hmC. The current chapter discusses only WGBS analysis, which is applicable to all BS-seq data, and the term DNA methylation will refer to 5mC + 5hmC as measured by BS-seq, unless otherwise stated.

Sequence alignment, read count, and methylation calling

The first step in analyzing DNA methylation data is the alignment (mapping) of sequencing reads to the reference genome. To maximize the rate of read mapping, it is recommended to trim sequencing adapters and low-quality bases at read ends. This process can be performed by using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) or cutadapt tools (16). C-to-T bisulfite conversion results in reduced complexity of the converted DNA and subsequent loss of complementarity with the reference genome. The bisulfite reaction and subsequent DNA amplification produce four individual strands from a single original fragment. Additionally, bisulfite libraries can be directional or non-directional, with the first approach preserving strand specificity in contrast to the second (17). BS-seq analysis tools such as Bismark (18) and QuasR (19) take into consideration these parameters and try to identify the best unique alignment by running four alignment processes simultaneously. Firstly, reads are C-to-T or G-to-A (reverse strand) converted and aligned to an equivalently converted genome. The alignment process is time-consuming and requires considerable computing resources. For large studies, a high-performance computing cluster is required.

As the assessment of cytosine status strongly depends on C-to-T conversion, the bisulfite conversion efficiency must be controlled for every experiment. Spiking samples with unmethylated lambda phage DNA provide an accurate estimation of bisulfite conversion as all cytosines in this genome should be converted. High-quality experiments produce conversion rates greater than 99%.

In classical WGBS experiments, where 5mC and 5hmC cannot be distinguished, both modifications are reported as methylated cytosines. In this context, the methylation level of cytosines is reported as the ratio of the number of reads with “C” (5mc + 5hmC) over the number of reads with either “C” or “T” (5mC + 5hmC + C). These reads originate from a population of cells with variable methylation states. Therefore, the methylation ratio ranges from 0 to 1, where 0 corresponds to a fully unmethylated state and 1 indicates a fully methylated state. Tools such as Bismark and QuasR produce count matrices containing the number of methylated and unmethylated reads for every cytosine that can be used for further analysis. In studies combining the oxBS-seq and BS-seq approaches, 5mC, 5hmC, and uC proportions can be computed using maximum likelihood estimates (20) or binomial modeling (21). It is important to mention that calculating the simple difference between BS and oxBS signals as an estimate of 5hmC can produce negative proportions and sums (5mC + 5hmC + uC) greater than 1. Such inconsistencies may simply represent sequencing artifacts or low-coverage biases and have no biological significance.

DNA methylation patterns

Cytosine methylation, as measured by WGBS in the human (3) and mouse (4) genomes, has shown that 5mC occurs mostly in the context of CpG dinucleotides, while non-CpG methylation is a rare event. The 5mC frequency of individual CpGs has a bimodal distribution, with the majority of CpGs being highly methylated and a small subset of CpGs showing an unmethylated state. In addition to these two categories, a third population of CpGs shows an intermediate range of methylation ranging from 10 to 50% (Figure 2). At the genome scale, the methylome can be segmented into three distinct classes: fully methylated regions (FMRs),

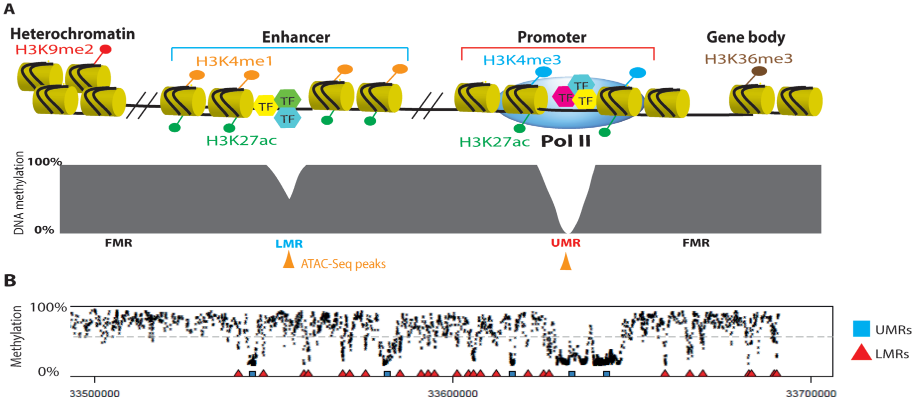

Figure 2 Epigenetic landscape. (A) Genomic distribution of the main epigenetic marks. Histones are depicted as yellow cylinders; black lines represent DNA, and modifications of histone tails are shown as circles. Transcription factors (TF) are indicated by hexagons. Grey blocks represent DNA methylation patterns. UMRs, LMRs and FMRs show the unmethylated regions, low-methylated regions and fully methylated regions, respectively. Assay for transposase-accessible chromatin using sequencing (ATAC-Seq) peaks are depicted in orange. (**B**) Genome browser snapshot illustrating DNA methylation patterns. ac: acetylation; H: histone; K: lysine; me: methylation; Pol II: RNA polymerase II.

unmethylated regions (UMRs), and low-methylated regions (LMRs) (4). FMRs represent 90% of the genome and are enriched at inter- and intragenic regions. UMRs correspond to the majority of CpG islands and active promoters, while LMRs exhibit enhancer features such as specific histone marks and binding of TFs (4) (Figure 2). While the majority of genomic regions fit this classification, some cell types contain contiguous regions showing disordered states of methylation ranging from 0 to 100%, with little similarity between neighboring CpGs (3). These loci were termed "partially methylated domains" (PMDs). It is important to mention that PMDs and LMRs are two distinct methylation profiles, and particular attention should be paid to the behavior of PMDs in comparative investigations. Methylation distribution in PMDs may affect the identification of LMRs (22) and the computing of differential methylation between experimental groups. The presence of PMDs can be evaluated using, for instance, the MethylSeekR package (22), and whether or not to exclude them from the analysis depends on the study purpose.

Computing differential methylation

After methylation calling, the next step is to compare methylation profiles between experimental groups to identify differentially methylated cytosines (DMCs) or differentially methylated regions (DMRs). Selecting the appropriate statistical model is the most important step in computing differential methylation for studies with biological replicates. The different statistical methods used to call DMCs/DMRs were summarized in an excellent review by Wreczycka et al. (23), which addressed additional aspects of DNA methylation analysis. In general, regression models are the best choice for comparing methylation profiles in studies with several replicates per experimental group. The selected method should take into consideration intra-group variability, which is more pronounced in in vivo and clinical studies. Beta-binomial distribution is a natural choice for computing differential methylation when biological replicates are available as it can correct for technical sampling and intra-group variability. A number of tools, such as DSS (24) and RADMeth (25), are based on the beta-binomial distribution. These tools compute differential methylation for single cytosines and require predefined file formatting. For more flexibility, the beta-binomial distribution is implemented in a number of R packages such as TailRank (https://cran.r-project.org/web/packages/TailRank/index.html) and AOD (https://cran.r-project.org/web/packages/aod/index.html).

DMRs are usually called based on the FDR-adjusted P-value from the fitted statistical model. The value of methylation difference can also be used as an additional selection parameter. As the outcome of this approach is based on the cut-offs used, it is important to have a global quantitative view of data distribution by generating volcano or simple scatter plots. Another important parameter in calling DMRs is the read coverage at the investigated position, because accurate evaluation of methylation differences between samples requires decent read coverage. In our own work, we set the minimum number of reads to 15; however, this parameter can be changed depending on the data at hand. Computing differential methylation should also take into consideration a number of covariates such as age, sex, and other potential confounding factors. Finally, genetic variations can also affect methylation status, and particular attention should be paid to C/T single-nucleotide polymorphism (SNPs).

In addition to differential methylation, increased variability in methylation levels can also be observed at some loci in response to exposures (26, 27) or in relation to some diseases (28). These variably methylated regions (VMRs) have been suggested to be regions of stochastic epigenetic variations (29) that may indicate a certain degree of flexibility in the control of local chromatin structure. VMRs have been observed to occur preferentially at enhancers and 3′-untranslated regions (3′UTRs) (30), suggesting a potential role in gene regulation. VMRs can be called simply by calculating the variance; however, this approach is sensitive to intra-individual and technical variations. The multiple hypothesis testing approach has been suggested to call VMRs by distinguishing biological variability from intra-individual variations (31). Although this approach was applied to methylation arrays, it can also be adapted to sequencing data.

DNA methylation in disease research and risk assessment

Genome-wide association studies (GWAS) have been designed to identify risk-associated SNPs that can be used as prediction tools in clinical investigations or for personalized medicine. Similarly, epigenome-wide association studies (EWAS) aim to derive potential associations between epigenetic marks and a particular trait, disease or exposure–response profile. To date, the vast majority, if not all, of EWAS have been based on DNA methylation. Therefore, these investigations should rather be termed "methylome-wide association studies" (MWAS). MWAS have been mainly conducted in the context of tumorigenesis, where the methylome of cancer cells is characterized by global hypomethylation except at some CpG islands that undergo hypermethylation (32). MWAS have been also conducted in relation to various diseases and phenotypes. The EWASdb database records 1319 MWAS associated with 302 diseases and/or phenotypes, including autoimmune, metabolic, and exposure-related disorders, to name a few (33).

To date, most MWAS have been based on methylation arrays that interrogate mainly annotated regions and poorly cover the complex network of distal REs. Given the central role of distal REs in genome regulation, it is crucial to interrogate the association of these loci with the traits of interest. For example, WGBS investigations in the mouse lung showed that cigarette smoke exposure mainly alters DNA methylation at candidate enhancers (identified as LMRs), while promoters are less affected (34). The importance of distal RE is also illustrated by the fact that the majority of GWAS-identified hits are located in non-coding regions with potential regulatory function, arguing for their informative value in both GWAS and EWAS.

Ideally, a comprehensive MWAS would assess cytosine status at a genome-wide level using WGBS and oxBS-seq in parallel to discriminate 5mC from 5hmC. However, this scenario requires high read coverage to accurately evaluate methylation variations. Finally, an adequate sample size is required to assure sufficient power to detect methylation differences (35). These requirements make whole-genome investigations costly for studies involving large cohorts. Alternatively, cytosine methylation can be investigated for a defined set of genomic targets using capture techniques (36). This approach allows the design of custom sets of loci to address specific needs and to increase the read coverage per site, while reducing the cost.

MWAS must also take into consideration inter- and intra-individual variations in DNA methylation levels. Unlike genetic information, where all cell types share the same genome, the epigenome varies between cell types and tissues. Thus, epigenome profiling in peripheral sources such as blood and saliva may not recapitulate the variations occurring in specific target organs. Cell heterogeneity in liquid biopsies may also complicate the use of DNA methylation variations as reliable biomarkers. Additionally, epigenetic marks change over time, are sensitive to environmental factors and health status, and may be affected by genetic variants. The aforementioned confounding factors and many others should be considered during the experimental design and computational framework.

DNA methylation results are generally reported as the difference in mean methylation ratios between experimental groups, with P-values derived from a sound statistical model. In most MWAS, the effect size is modest. It is rare to observe cytosines moving from the unmethylated state to fully methylated state or vice versa except when comparing methylomes from different cellular origins (37). Usually, the association between the response variable (disease, exposure, and the like) and explanatory variable (DNA methylation level) relies on the P-value, while the effect size is neglected, thus reducing the applicability of MWAS in personalized medicine. For example, the cg03636183 CpG site located in the *F2RL3* gene is considered a strong marker of cigarette-smoke exposure in blood samples. The median methylation level of this CpG is 95% in never smokers and 83% in smokers (38). Despite the methylation difference of 12%, this site still belongs to the category of fully methylated CpGs and can hardly be used to distinguish smokers from non-smokers. However, this site and many others are reproducibly found to be differentially methylated in independent cohorts in relation to smoking. Given the binary nature of 5mC at the allele level, these reproducible, but weak, variations can be explained by the cellular heterogeneity of blood samples. Some DNA methylation variation may reflect the cell-type composition of blood samples (39, 40). As mentioned earlier, the measured methylation level represents the average of events occurring in a population of cells. Therefore, cell-type-specific variations may be diluted in the averaged bulk signal. It has been shown that many loci, including cg03636183 in *F2RL3* and cg05575921 in *AHRR*, exhibit distinct patterns of smoking-associated methylation variations across blood-cell types (41). Investigating DNA methylation variations in specific cell types may reduce the bias linked to cell heterogeneity and allow more accurate detection of cell-type-specific DMRs or DMCs.

Leveraging epigenetic associations to causal biologic mechanisms is still challenging. DNA methylation variations can be the cause or consequence of the investigated phenotype. This complex interaction is illustrated by the chronology of promoter hypermethylation in cancer cells. It has been reported that some transcriptionally silenced promoters in healthy cells become aberrantly hypermethylated during tumorigenesis, implying that the hypermethylation of some loci is likely the consequence, rather than the cause, of tumorigenesis (42). Despite the lack of clear causality to cancer etiology, DNA methylation levels of a limited set of loci have been used to develop diagnostic tests for colorectal, prostate, and bladder cancers (43). Cologuard®, a DNA methylation-based diagnostic kit (Exact Sciences Corporation, WI, USA), was the first stool DNA screening test approved by the U.S. Food and Drug Administration for colorectal cancer.

Machine learning in MWAS

Machine-learning (ML) algorithms are promising tools for identifying methylome variations predictive or indicative of certain phenotypes or exposures. These algorithms seek to identify a set of loci (features) whose methylation levels can be used as a signature to categorize samples from different experimental groups (classification methods) or to estimate continuous metrics such as age (regression methods). In the context of MWAS, classification algorithms have been mainly used to classify cancer samples. Random forest (RF)-based supervised learning is one of the most used ML algorithm in MWAS (37, 44, 45). For example, this algorithm has been used to construct a DNA methylation signature based on 20 loci for stratifying different types of brain metastasis. This signature also showed a good performance on samples from a test set that was not used to train the model (37). The good classification power of this signature is probably due to the cell-type-specific DNA methylation patterns of primary tumors. The RF algorithm has been also used to build DNA methylation signatures to classify different tumor types, including breast, kidney, and thyroid carcinomas (44), and to classify central nervous system tumors (45). DNA methylation has been also used to classify subtypes and predict treatment outcome in patients with childhood acute lymphoblastic leukemia using the nearest shrunken centroids (NSC) approach.

Regression algorithms have been also applied to methylome data, mainly in the context of age prediction. DNA methylation of a limited set of CpGs has been used to build age predictors in humans (46, 47) and mice (48, 49). One of the first epigenetic predictors of age, termed the Horvath clock (46), is a multi-tissue predictor based on 353 CpGs and can estimate chronological age in test samples with a median error of 3.6 years. This model has been derived from 8000 methylomes using elastic-net regression. After this pioneering work, a number of other DNA methylation clocks have been developed using other tissues and regression algorithms (50). Regularized linear regressions are the most used algorithms for building age predictors. The regression method selected depends on the data at hand and the questions to be answered. Although the most accurate DNA methylation clocks are derived from elastic-net regression, the beneficial effects of anti-aging interventions are better computed by ridge regression-based clocks (49). ML approaches have mainly been applied to data generated by methylation arrays and are only starting to be used for sequencing datasets. In the context of sequencing datasets, the aforementioned limitations for computing differential methylation are also valid for ML approaches, and particular attention should be paid to poorly covered sites.

CHROMATIN REGULATION

Chromatin is a DNA–protein complex, the primary function of which is to organize the genetic material in a compact form to fit into the nucleus. The fundamental chromatin unit, the nucleosome, consists of 147 DNA base pairs wrapped around histone octamers. Multiple histone residues, mainly at histone tails, can undergo covalent post-translational modifications (PTMs), including methylation, phosphorylation, acetylation, and SUMOylation. PTM regulation involves three families of epigenetic enzymes: writers that catalyze the addition of various chemical groups, readers that recognize and interpret these modifications, and erasers that remove them to ensure dynamic epigenetic regulation (51).

Genome-wide mapping of PTMs has identified their functional association with chromatin properties, transcriptional competency, DNA-damage repair, and DNA replication. The combination of PTMs at a particular locus shapes the local chromatin structure and modulates transcriptional activity. This combinatorial regulatory code has been termed "histone code." For example, trimethylation of lysine 4 of histone 3 (H3K4me3) marks actively transcribed promoters, while monomethylation of the same residue (H3K4me1) marks active enhancers. Acetylation of any histone residue (e.g., H3K27ac, H3K9ac or H3K14ac) is always associated with active REs (Figure 2). Other PTMs are associated with transcriptional repression. For example, H3K9me2 is a key marker of heterochromatin domains (52), and H3K27me3 indicates polycomb group silenced loci (53).

Profiling histone modifications by ChIP sequencing

Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) is a powerful technique for profiling the genomic distribution of PTMs and other DNA-binding proteins such as TFs and epigenetic enzymes. ChIP-seq involves an immunoprecipitation (IP) step using antibodies directed against the target protein. The captured DNA is further subjected to next-generation sequencing, and the resulting reads are mapped to the reference genome to identify the binding sites of target proteins (Figure 3). The general assumption is that target protein binding sites will produce more reads than the rest of the genome, which will be covered by the sequencing noise/background captured by unspecific binding of the IP antibody. The sequencing noise is generally assessed by sequencing a fraction of the input chromatin prior to the IP step. This noise is not uniform and reflects local chromatin accessibility, amplification, and mappability biases. The interpretability of ChIP-seq experiments strongly depends on antibody specificity, the amount of starting material, and epitope integrity after cell lysis and chromatin shearing. The impact of these parameters is reflected by the signal-to-noise ratio in the sequencing data.

Once the sequencing reads are aligned to the reference genome, the next step is to identify genomic regions that are enriched for the target protein. This step is usually termed peak-calling, because the first ChIP-seq experiments were mainly designed to map TFs and resulted in very short enriched regions (0.5 kb to 1 kb) with a peak shape and clear summit (maximum read density) when visualized on genome browsers. However, not all ChIP-seq experiments generate narrow peaks. Some PTMs such as heterochromatin marks are uniformly enriched in very large regions with no clear summit, while other PTMs such as active promoter marks (e.g., H3K4me3 and H3K9ac) are enriched in relatively short regions (1 kb to 2 kb) with a clear local maximum read density. More complex patterns include a mixture of narrow peaks and diffused regions such as the H3K27me3 mark. The majority of peak-calling tools (listed in two reviews (54, 55) were designed to detect narrow peaks and may not perform accurately on ChIP-seq experiments with broad and diffuse enriched regions. However, some tools such as the popular MACS (56) and Epic (57) have included new parameters to model mixed enrichment events in recent updates.

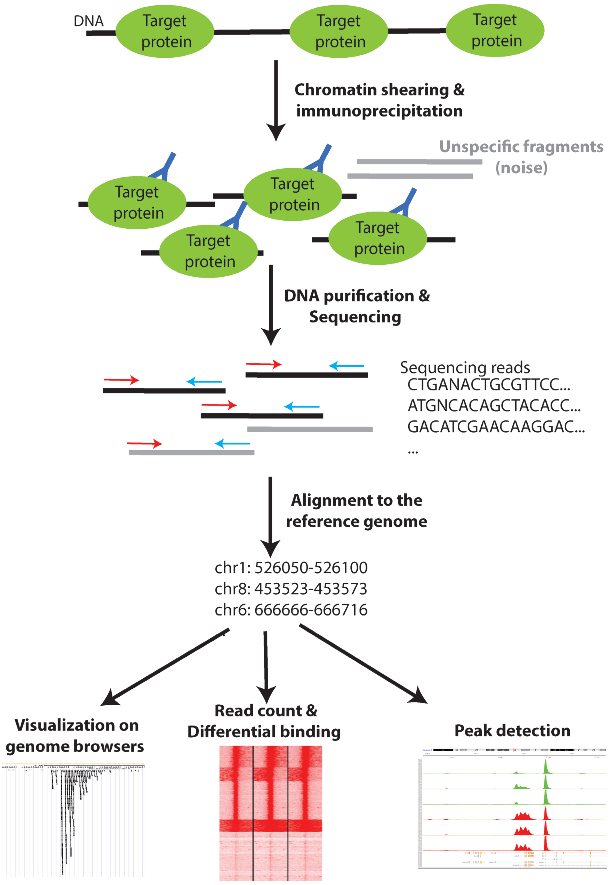

Figure 3 ChIP-seq workflow. Sheared chromatin is incubated with an antibody directed against the target protein. Upon purification, the captured DNA is then subjected to high-throughput sequencing, and the resulting reads are aligned to the reference genome. Aligned reads can be visualized on genome browsers and computed to identify binding sites of target proteins.

Modeling the background distribution of reads is an important step in peak detection and can be performed from the input control, but not all studies include this control. Consequently, the majority of algorithms model the intrinsic background of ChIP samples. While this approach performs well for narrow-peak experiments, it provides poor results for diffuse enriched regions. In our opinion, the input control should be included in all ChIP-seq experiments. A simple scatter plot comparison of read counts over genomic windows from ChIP samples versus the input control (Figure 4) provides a primary evaluation of ChIP-seq quality. Additionally, we believe that the input control is mandatory for investigating heterochromatin marks, given their genomic distribution. Finally, particular attention should be paid to repetitive elements that produce very short peaks with a high number of reads, as these peaks represent sequencing biases rather than binding events.

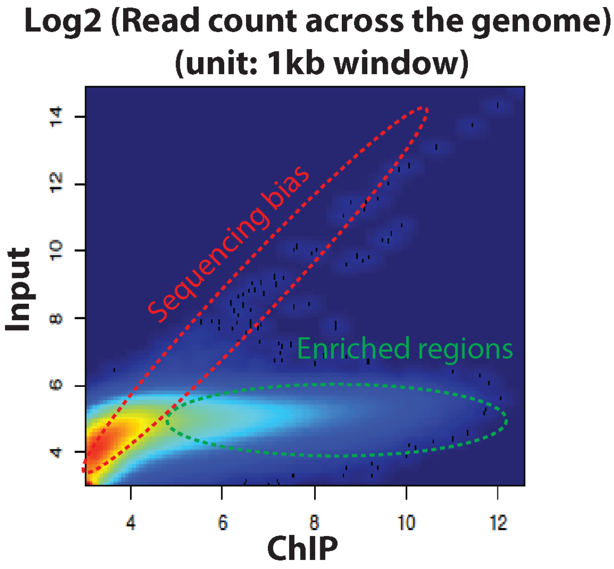

Figure 4 Example of a high-quality ChIP-seq experiment. The genome-wide ChIP signal (number of reads per 1-kb window) is plotted against the corresponding input control. Enriched/bound loci (indicated in green) show higher read numbers in the ChIP sample than in the input control. Unbound loci show low read numbers in the ChIP sample. Reads originated from unspecific binding and/or sequencing biases are present equally in ChIP and input samples (indicated in red).

The initial goal of ChIP-seq experiments was to investigate the genomic distribution of DNA-binding proteins in the context of basic research, and the first ChIP-seq studies rarely included biological replicates. Comparative analyses mainly consisted of detecting differentially enriched regions at defined coordinates, such as annotated promoters based on read count cut-offs. With the continuous decrease in sequencing costs and widespread application of the technique, including in clinical investigations, most recent studies include biological replicates. A number of approaches have been suggested to leverage biological replicates to improve the accuracy of peak detection (58). Most of these methods compare the overlap between peaks detected independently in the different replicates and select confident peaks based on reproducibility. While this approach is convenient for identifying highly confident-enriched regions, it is not suitable for identifying significantly differentially enriched regions based on read counts in comparative analyses (e.g., case vs. control).

To identify differentially enriched regions between experimental groups, we suggest that the analysis should include all potentially bound regions, even those with low confidence. If a peak caller is used, peaks from different replicates and experimental groups can be merged to build a unique set of loci. A more holistic approach consists of assessing differential binding/enrichment at genomic windows along the chromosomes. Once a consensus set of loci is defined, read counts can be generated for all replicates. Differential binding can then be computed based on read counts similar to differential expression in RNA sequencing data.

This step can be performed using the DEseq2 package (59), which uses negative binomial distribution to compute the statistical significance between groups. Other packages such as Diffbind and MMDiff have been developed specifically for differential ChIP-seq analysis. Diffbind uses DEseq2 internally but offers the possibility to integrate input controls, while MMDiff takes in account the distribution of reads within the enriched regions. The choice of which approach to use is dictated by the questions to be answered, number of replicates, and availability of control experiments. Although the majority of available tools perform a normalization step, it is important to ensure the scaling of unequal datasets by library size.

Genome-wide chromatin investigations are rarely conducted in clinical studies because of the complexity of chromatin properties, amount of starting material required, and multiplicity of processing steps. Additionally, histone modification patterns are cell-type-specific and need to be generated from target organs rather than from peripheral sources, which restricts the investigations to postoperative and post-mortem samples. A search for clinical trials involving chromatin among the 308,830 clinical trial records available in the ClinincalTrials.gov database resulted in only 82 and 16 hits for the terms chromatin and ChIP-seq, respectively. The recent adaptation of the ChIP-seq protocol to small cancer biopsies (60) may, however, facilitate the future use of this approach in clinical studies.

Assessing chromatin accessibility by ATAC-seq

The assay for transposase-accessible chromatin using sequencing (61) (ATAC-seq) allows detection of accessible (i.e., open) chromatin regions, which are mainly active REs and TF-binding sites. ATAC-seq is based on a process called tagmentation, which involves simultaneous fragmentation and sequencing-adapter ligation. This reaction is carried out with a hyperactive mutant of Tn5 transposase that inserts sequencing adapters into open chromatin regions. Reads produced from these regions during high-throughput sequencing are used to detect peaks, similar to ChIP-seq data. While ChIP-seq ideally requires a few million cells, a standard ATAC-seq experiment requires only 50,000 cells, making it more suitable for studies with a limited amount of starting material. Although ATAC-seq provides no information about the identity of the binding proteins, ATAC-seq-enriched regions show high overlap with active RE-associated PTMs such as H3K4me3 and H3K27ac. ATAC-seq has been recently used to investigate open chromatin distribution in 23 cancer types (62).

CONCLUSION

Advances in sequencing technologies have enabled scientists to reveal the striking immensity of gene-regulation mechanisms and, particularly, the large repertoire of epigenetic pathways. Although a number of these mechanisms are now well understood, many others remain to be elucidated. For example, the human genome codes for hundreds of TFs, but only a small fraction of them have been studied (63). Similarly, the roles of many histone and DNA modifications remain unclear. Transcriptional alterations play a central role in almost all human disorders, and these alterations are very likely preceded by changes in epigenetic patterns and TF binding and/or activity.

The diversity of measurable epigenetic marks holds the promise of using epigenetic events as early markers of human disorders and for providing mechanistic clues to disease etiology. However, this initial excitement about epigenetic markers has been tempered by the complexity of their biological outcomes and their interactions with other molecular signals (e.g., gene expression). At the molecular level, most, if not all, epigenetic marks are binary, and their variations in some loci can, in theory, be used to monitor a number of biological processes. However, most of the observed epigenetic changes in EWAS are modest and reflect the average of events occurring in a heterogeneous population of cells. Advances in single-cell investigations may help unveil more reliable epigenetic markers as exemplified by the recent characterization of DNA methylation profiles of circulating tumor cells using single-cell methylomes (64) and single-cell ChIP-seq investigation of breast cancer heterogeneity (65). Another limitation of currently available EWAS is the poor investigation of non-coding regions that contain most of the distal REs and represent the vast majority of disease-associated variations at the genetic level.

In our opinion, improvement of EWAS outcomes should be articulated around three main axes: reduction of cell-type heterogeneity, increase in genome coverage, and combination of a larger panel of epigenetic marks. Overcoming these challenges will require massive computational and technical efforts in both academic and industrial research. Generating interpretable genome-wide data from low cell number or single-cell samples will likely be the next breakthrough in clinical investigations. This new type of data will require the development of new computational approaches prioritizing personalized assessment rather than group comparisons. Computational investigations should also leverage the diversity of epigenetic marks together with other omics data to better understand the flow of events leading to disease onset, possibly identifying combinatorial markers for disease progression and drug response.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced.

REFERENCES

1. Waddington CH. The epigenotype. 1942. Int J Epidemiol. 2012 Feb;41(1):10–13. http://dx.doi.org/10.1093/ije/dyr184
2. Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R, et al. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development (Cambridge, England). 1996 Oct;122:3195–205.
3. 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 Nov 19;462:315–22. http://dx.doi.org/10.1038/nature08514

4. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011 Dec 14;480:490–5. http://dx.doi.org/10.1038/nature10716
5. Lyko F. The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. Nat Rev Genet. 2018 Feb;19(2):81–92. http://dx.doi.org/10.1038/nrg.2017.80
6. Wu H, Zhang Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. 2011 Dec 1;25(23):2436–52. http://dx.doi.org/10.1101/gad.179184.111
7. Hill PWS, Amouroux R, Hajkova P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story. Genomics. 2014 Nov 1;104(5):324–33. http://dx.doi.org/10.1016/j.ygeno.2014.08.012
8. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015 Apr 9; 520(7546):243–7. http://dx.doi.org/10.1038/nature14176
9. Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. Nature. 2016 Apr 21;532(7599):329–33. http://dx.doi.org/10.1038/nature17640
10. Epigenomics in tobacco risk assessment: Opportunities for integrated new approaches – ScienceDirect [Internet]. [cited 2019 Jun 7]. Available from: https://www.sciencedirect.com/science/article/pii/S2468202018300573
11. Nestor C, Ruzov A, Meehan R, Dunican D. Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. BioTechniques. 2010 Apr;48(4):317–19. http://dx.doi.org/10.2144/000113403
12. Nestor CE, Ottaviano R, Reddington J, Sproul D, Reinhardt D, Dunican D, et al. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome Res. 2012 Mar;22(3):467–77. http://dx.doi.org/10.1101/gr.126417.111
13. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012 May 18;336(6083):934–7. http://dx.doi.org/10.1126/science.1220671
14. Booth MJ, Ost TWB, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc. 2013 Oct;8(10):1841–51. http://dx.doi.org/10.1038/nprot.2013.115
15. Skvortsova K, Zotenko E, Luu P-L, Gould CM, Nair SS, Clark SJ, et al. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. Epigenetics Chromatin [Internet]. 2017 Apr 20 [cited 2018 Oct 15];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5397694/
16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011 May 2;17(1):10–12. http://dx.doi.org/10.14806/ej.17.1.200
17. Chen P-Y, Cokus SJ, Pellegrini M. BS Seeker: Precise mapping for bisulfite sequencing. BMC Bioinformatics. 2010 Apr 23;11:203. http://dx.doi.org/10.1186/1471-2105-11-203
18. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571–2. http://dx.doi.org/10.1093/bioinformatics/btr167
19. Gaidatzis D, Lerch A, Hahne F, Stadler MB. QuasR: Quantification and annotation of short reads in R. Bioinformatics. 2015 Apr 1;31:1130–2. http://dx.doi.org/10.1093/bioinformatics/btu781
20. Kiihl SF, Martinez-Garrido MJ, Domingo-Relloso A, Bermudez J, Tellez-Plaza M. MLML2R: An R package for maximum likelihood estimation of DNA methylation and hydroxymethylation proportions. Stat Appl Genet Mol Biol. 2019 Jan 17;18(1):pii:/j/sagmb.2019. http://dx.doi.org/10.1515/sagmb-2018-0031
21. Xu Z, Taylor JA, Leung Y-K, Ho S-M, Niu L. oxBS-MLE: An efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. Bioinformatics. 2016 Jan;32(23):3667–9. http://dx.doi.org/10.1093/bioinformatics/btw527
22. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013 Sep;41(16):e155. http://dx.doi.org/10.1093/nar/gkt599

23. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol. 2017 Nov 10;261:105–15. http://dx.doi.org/10.1016/j.jbiotec.2017.08.007
24. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014 Apr;42(8):e69. http://dx.doi.org/10.1093/nar/gku154
25. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics. 2014 Jun 24;15:215. http://dx.doi.org/10.1186/1471-2105-15-215
26. Jenkins TG, James ER, Alonso DF, Hoidal JR, Murphy PJ, Hotaling JM, et al. Cigarette smoking significantly alters sperm DNA methylation patterns. Andrology. 2017 Nov;5(6):1089–99. http://dx.doi.org/10.1111/andr.12416
27. Vaz M, Hwang SY, Kagiampakis I, Phallen J, Patil A, O'Hagan HM, et al. Chronic cigarette smoke-induced epigenomic changes precede sensitation of bronchial epithelial cells to single-step transformation by KRAS mutations. Cancer Cell. 2017 Sep 11;32(3):360–376.e6. http://dx.doi.org/10.1016/j.ccell.2017.08.006
28. Webster AP, Plant D, Ecker S, Zufferey F, Bell JT, Feber A, et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. Genome Med. 2018 Sep 4;10(1):64. http://dx.doi.org/10.1186/s13073-018-0575-9
29. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. PNAS. 2010 Jan 26;107(Suppl 1):1757–64. http://dx.doi.org/10.1073/pnas.0906183107
30. Garg P, Joshi RS, Watson C, Sharp AJ. A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. PLoS Genet. 2018 Oct 1;14(10):e1007707. http://dx.doi.org/10.1371/journal.pgen.1007707
31. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012 Jan;13(1):166–78. http://dx.doi.org/10.1093/biostatistics/kxr013
32. Sproul D, Meehan RR. Genomic insights into cancer-associated aberrant CpG island hypermethylation. Brief Funct Genomics. 2013 May;12(3):174–90. http://dx.doi.org/10.1093/bfgp/els063
33. Liu D, Zhao L, Wang Z, Zhou X, Fan X, Li Y, et al. EWASdb: Epigenome-wide association study database. Nucleic Acids Res. 2019 Jan 8;47(D1):D989–93. http://dx.doi.org/10.1093/nar/gky942
34. Choukrallah M-A, Sierro N, Martin F, Baumer K, Thomas J, Ouadi S, et al. Tobacco Heating System 2.2 has a limited impact on DNA methylation of candidate enhancers in mouse lung compared with cigarette smoke. Food Chem Toxicol. 2019 Jan 1;123:501–10. http://dx.doi.org/10.1016/j.fct.2018.11.020
35. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011 Jul 12;12(8):529–41. http://dx.doi.org/10.1038/nrg3000
36. 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 Jul 13;43(12):e81. http://dx.doi.org/10.1093/nar/gkv244
37. Orozco JIJ, Knijnenburg TA, Manughian-Peter AO, Salomon MP, Barkhoudarian G, Jalas JR, et al. Epigenetic profiling for the molecular classification of metastatic brain tumors. Nat Commun [Internet]. 2018 Nov 6 [cited 2019 Jun 14];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219520/
38. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011 Apr 8;88:450–7. http://dx.doi.org/10.1016/j.ajhg.2011.03.003
39. Bauer M, Linsel G, Fink B, Offenberg K, Hahn AM, Sack U, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics. 2015;7:81. http://dx.doi.org/10.1186/s13148-015-0113-1
40. Bauer M, Fink B, Thürmann L, Eszlinger M, Herberth G, Lehmann I. Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation. Clin Epigenetics. 2015;7:83. http://dx.doi.org/10.1186/s13148-016-0249-7

41. Su D, Wang X, Campbell MR, Porter DK, Pittman GS, Bennett BD, et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. PLoS One. 2016;11(12):e0166486. http://dx.doi.org/10.1371/journal.pone.0166486
42. Sproul D, Nestor C, Culley J, Dickson JH, Dixon JM, Harrison DJ, et al. Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. Proc Natl Acad Sci U S A. 2011 Mar 15;108:4364–9. http://dx.doi.org/10.1073/pnas.1013224108
43. Kronfol MM, Dozmorov MG, Huang R, Slattum PW, McClay JL. The role of epigenomics in personalized medicine. Expert Rev Precis Med Drug Dev. 2017;2(1):33–45. http://dx.doi.org/10.1080/23808993.2017.1284557
44. Celli F, Cumbo F, Weitschek E. Classification of large DNA methylation datasets for identifying cancer drivers. Big Data Research. 2018 Sep;13:21–8. http://dx.doi.org/10.1016/j.bdr.2018.02.005
45. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018 Mar;555(7697):469–74.
46. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115. http://dx.doi.org/10.1186/gb-2013-14-10-r115
47. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013 Jan 24;49(2):359–67. http://dx.doi.org/10.1016/j.molcel.2012.10.016
48. Multi-tissue DNA methylation age predictor in mouse. PubMed – NCBI [Internet]. [cited 2019 Apr 11]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28399939
49. Thompson MJ, Chwiałkowska K, Rubbi L, Lusis AJ, Davis RC, Srivastava A, et al. A multi-tissue full lifespan epigenetic clock for mice. Aging (Albany NY). 2018 Oct 21;10(10):2832–54. http://dx.doi.org/10.18632/aging.101590
50. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 2014 Feb 3;15(2):R24. http://dx.doi.org/10.1186/gb-2014-15-2-r24
51. Kouzarides T. Chromatin modifications and their function. Cell. 2007 Feb 23;128(4):693–705. http://dx.doi.org/10.1016/j.cell.2007.02.005
52. Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature. 2001 Mar 1;410(6824):116–20. http://dx.doi.org/10.1038/35065132
53. Bernstein E, Duncan EM, Masui O, Gil J, Heard E, Allis CD. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Mol Cell Biol. 2006 Apr;26(7):2560–9. http://dx.doi.org/10.1128/MCB.26.7.2560-2569.2006
54. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-Seq peak detection. PLoS One. 2010 Jul 8;5(7):e11471. http://dx.doi.org/10.1371/journal.pone.0011471
55. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat Methods. 2009 Nov;6(11 Suppl):S22–32. http://dx.doi.org/10.1038/nmeth.1371
56. Zhang Y. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137. http://dx.doi.org/10.1186/gb-2008-9-9-r137
57. Stovner EB, Sætrom P. epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics. 2019 Mar 28. http://dx.doi.org/10.1093/bioinformatics/btz232
58. Yang Y, Fear J, Hu J, Haecker I, Zhou L, Renne R, et al. Leveraging biological replicates to improve analysis in ChIP-seq experiments. Comput Struct Biotechnol J [Internet]. 2014 Jan 31 [cited 2019 Jun 16];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3962196/
59. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550. http://dx.doi.org/10.1186/s13059-014-0550-8
60. Singh AA, Schuurman K, Nevedomskaya E, Stelloo S, Linder S, Droog M, et al. Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. Life Sci Alliance [Internet]. 2018 Dec 28 [cited 2019 Jun 19];2(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6311467/
61. Buenrostro J, Wu B, Chang H, Greenleaf W. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015 Jan 5;109:21.29.1–21.29.9. http://dx.doi.org/10.1002/0471142727.mb2129s109

62. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018 Oct 26;362(6413):eaav1898. http://dx.doi.org/10.1126/science.aav1898
63. Li YF, Altman RB. Systematic target function annotation of human transcription factors. BMC Biol. 2018 10;16(1):4. http://dx.doi.org/10.1186/s12915-017-0469-0
64. Gkountela S, Castro-Giner F, Szczerba BM, Vetter M, Landin J, Scherrer R, et al. Circulating tumor cell clustering shapes DNA methylation to enable metastasis seeding. Cell. 2019 Jan 10;176(1–2):98–112.e14. http://dx.doi.org/10.1016/j.cell.2018.11.046
65. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet. 2019 Jun;51(6):1060. http://dx.doi.org/10.1038/s41588-019-0424-9