Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

Review Article

Single molecule and single cell epigenomics

Byung-Ryool Hyun, John L. McElwee, Paul D. Soloway *

Division of Nutritional Sciences, Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 5 April 2014
Revised 19 August 2014
Accepted 27 August 2014
Available online xxxx

Keywords:
Single cell
Single molecule
Epigenomics
Epigenetics
Microfluidics
Nanofluidics

ABSTRACT

Dynamically regulated changes in chromatin states are vital for normal development and can produce disease when they go awry. Accordingly, much effort has been devoted to characterizing these states under normal and pathological conditions. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the most widely used method to characterize where in the genome transcription factors, modified histones, modified nucleotides and chromatin binding proteins are found; bisulfite sequencing (BS-seq) and its variants are commonly used to characterize the locations of DNA modifications. Though very powerful, these methods are not without limitations. Notably, they are best at characterizing one chromatin feature at a time, yet chromatin features arise and function in combination. Investigators commonly superimpose separate ChIP-seq or BS-seq datasets, and then infer where chromatin features are found together. While these inferences might be correct, they can be misleading when the chromatin source has distinct cell types, or when a given cell type exhibits any cell to cell variation in chromatin state. These ambiguities can be eliminated by robust methods that directly characterize the existence and genomic locations of combinations of chromatin features in very small inputs of cells or ideally, single cells. Here we review single molecule epigenomic methods under development to overcome these limitations, the technical challenges associated with single molecule methods and their potential application to single cells.

1. Epigenetics and epigenomics

The term "epigenetics" was originally coined by Conrad Waddington [1], and can be defined as the heritable changes in traits that occur without direct alteration of the underlying primary DNA sequence. At the molecular level, epigenetic phenomena are regulated by so-called epigenetic marks that include covalent changes to DNA and chromatin proteins, and remodeling of nucleosome position on the DNA. These are often controlled by non-coding RNAs and can respond to environmental influences. The epigenome refers to the constellation of epigenetic marks that exist in a cell and their chromosomal locations, which varies by cell type and developmental stage. Normal epigenomic states are important for cell-type specific gene expression and a variety of cellular and developmental processes. In humans, the aberrant placement of epigenetic marks has been linked to many diseases, including cancer (reviewed in [2]). The best studied DNA modification in mammals is 5-methylcytosine (5mC), which plays an important role in genomic imprinting, and the suppression of transposable elements [3]. Recently, it has been shown that 5mC can be converted to 5-hydroxymethylcytosine (5hmC) by the TET (Ten-Eleven-Translocation) family of proteins [4]. In addition to converting 5mC to 5hmC, TET proteins can catalyze further hydroxylation of 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in an enzymatic activity-dependent manner [5,6]. Numerous histone modifications exist primarily on the protruding amino-terminal tails of histones [7]; these include methylation (mono-, di-, trimethylation), acylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, carboxylation, glutathionylation, O-GlcNAcylation, hydroxylation, formylation and citrullination. The covalent modifications are placed by "writer" activities (reviewed in [8–11]), removed by "eraser" activities (reviewed in [12–14]). They exert their cellular effects by altering chromatin structure and by recruiting "reader" factors (reviewed in [15,16]), which in turn recruit additional factors necessary for epigenetically regulated processes.

2. Epigenomic methods

Characterization of epigenomes has been greatly enhanced by use of various high-throughput sequencing technologies, which have been used to construct detailed epigenomic landscapes in many systems. The currently used, highest resolution methods include bisulfite sequencing (BS-seq) to assess 5mC [17–20], vari-

* Corresponding author.
E-mail address: soloway@cornell.edu (P.D. Soloway).

http://dx.doi.org/10.1016/j.ymeth.2014.08.015
1046-2023/

ants of BS-seq to assess 5hmC [21,22], and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for the analysis of histone modifications and chromatin binding proteins [23–33]. Additional methods have been developed to assess epigenetically influenced chromatin states diagnostic for transcriptionally regulatory domains including DNAse-seq [34], FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) [35] and Sono-seq [36]. Among the technical challenges associated with these methods are biased amplification of bisulfite treated DNAs [37,38] and the problem of antibody quality for ChIP studies. Antibody specificity can limit ChIP-seq accuracy and reproducibility. Standards have been established to validate antibody quality [32,39] that include blotting approaches. Semi-synthetic DNA-barcoded nucleosome libraries may also prove useful [40] for antibody validation, and emerging microfluidic-based platforms [41] and the NanoString nCounter platform [42] have the potential for streamlining this process. Synthetic affinity reagents [43] may ultimately replace antibodies that have batch-to-batch variation. Two issues that are more fundamental than these technical challenges are (1) that the cited epigenomic methods generally provide information about a single epigenetic feature at a time, rather than the combinations of epigenomic features that exist in the genome; and (2) that they work best with an abundance of materials, and not as well with inputs from few or even single cells. The remainder of the discussion focuses on why these are limitations, and approaches under development to overcome them.

3. Combinatorial chromatin states

In describing the histone code hypothesis, Strahl and Allis suggested "that multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions" [44]. The strongest support for this at the time was provided by observations that the N-terminal tails of histones H3 and H4 contain multiple sites where covalent modifications may be placed. Since then, many lines of evidence lent support for the histone code hypothesis, and in particular, that combinations of chromatin modifications, and not just single modifications alone, are critical for genomic regulation. First, mass spectrometric analysis revealed that individual histones do in fact carry multiple modifications on their N-terminal tails [45,46]. Second, plant homeodomain (PHD) and bromodomain (BRD) motifs within so-called "reader" proteins that bind to specific modified histone amino acids, bind their targets in a manner that is sensitive to the modification states of nearby amino acids [47]. The histone acetyltransferase p300 in particular was shown to display this property [40]. Third, the reader protein Bromodomain PHD-finger Transcription Factor (BPTF), which has a methyl lysine-binding PHD motif adjacent to an acetyl lysine-binding BRD motif, was shown to bind H3K4me2/3 containing nucleosomes preferentially when the nucleosome also carried H4K16ac [48]. The observation that modifications on two independent histone molecules together affected BPTF binding highlights the fact that biologically relevant combinatorial states are not limited to those in *cis* on a single histone molecule. Fourth, combinations of histone modifications influence the biochemical activities of factors that bind and further modify histones. For example, the demethylase KDM7A that targets methylated forms of H3K9 and H3K27 for demethylation [49] contains a PHD motif that binds H3K4me3, suggesting that KDM7A is directed to its H3K9me and H3K27me targets in chromatin by adjacent H3K4me3 [50].

The histone code hypothesis can be extended to include effects coordinated with DNA modifications, as the combined importance of DNA and histone modifications to gene expression has been documented. The NuRD complex contains methyl binding domain (MBD) proteins, which bind 5mC and 5hmC, histone deacetylases (HDAC) and chromatin remodeling activity [51]. Gene silencing by HDAC activity in these complexes is enabled by MBD recruitment of the complex to modified DNA [52]. Given the cross talk among chromatin modifications, it should come as no surprise that their effects are coordinated by mechanisms that sense the modifications in combination. As the number of known reader proteins [53] and chromatin modifications [6,54] increases, so does the potential complexity of the histone code, or more broadly, the chromatin code. These trends elevate the importance of identifying and mapping the genomic locations of combinations of chromatin features in order to understand how those features regulate genomic information in normal and disease states.

4. Technologies that overcome some limitations of ChIP-seq and BS-seq

4.1. Re-ChIP and ChIP-BS-seq

The most widely used ChIP protocols query chromatin sources for chromatin features one at a time. Several sets of efforts have characterized where in the genome combinations of chromatin features can be found. One of these used sequential- or re-ChIP experiments, whereby chromatin immunoprecipitated with a first antibody was subjected to re-precipitation with a second antibody before analyzing the DNA [55–62]. In one application of re-ChIP, a bivalent state comprising H3K4me3 and H3K27me3 modifications at genes important for lineage specification was found in pluripotent stem cells [63]; in another application, histone variants H3.3 and H2AZ were found together on active promoters, enhancers and insulator regions [64]. Re-ChIP methods require large inputs of chromatin, given the inefficiencies with which each antibody precipitates the chromatin, and in some cases, the low abundance of the chromatin feature. There are few examples of whole genome re-ChIP studies. Studies with more than two sequential ChIP reactions will likely require antibodies or other affinity reagents with dissociation constants well below those of existing reagents in order to have high enough modification capture efficiencies; small reaction volumes that enable use of high concentrations of chromatin and capture reagents; and improvements in library preparation or sequencing methods that make most efficient use of the DNA isolated by ChIP.

In other efforts to define coincidence between 5mC and H3K27me3, DNA isolated by anti H3K27me3 ChIP was subjected to bisulfite sequencing [65]. In principle, this strategy can be applied to any DNAs isolated from a single or re-ChIP experiment, if sufficient amounts of DNA are recovered. Single molecule sequencing technologies have enabled the identification of DNA modifications, including 5mC and 5hmC without bisulfite and other chemistries (see below [66–68]). These provide an alternate means of characterizing the sequences and DNA modification states of chromatin isolated by ChIP.

4.2. Statistical methods

With high quality, high coverage sequencing results in hand, the workflow for ChIP-seq data analysis consists of read mapping and peak calling, including peak modeling and identification (reviewed in [69]). Statistical strategies have been combined with single ChIP experiments to identify sites where combinations of chromatin features are likely to be found [42,70–74]. These include use of hierarchical or *k*-means cluster analyses, heat map analysis, and Venn and Euler diagram analyses, (reviewed in [69]), however, direct observations using methods specifically designed to report combinations of chromatin features will be needed to validate