Published in final edited form as:

Methods Mol Biol. 2012 ; 833: 433–441. doi:10.1007/978-1-61779-477-3_26.

Quantitative Analysis of Genome-Wide Chromatin Remodeling

Songjoon Baek¹, Myong-Hee Sung¹, and Gordon L. Hager¹

¹Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

Abstract

Recent high-throughput sequencing technologies have opened the door for genome-wide characterization of chromatin features at an unprecedented resolution. Chromatin accessibility is an important property that regulates protein binding and other nuclear processes. Here, we describe computational methods to analyze chromatin accessibility using DNaseI hypersensitivity by sequencing (DNaseI-seq). Although there are numerous bioinformatic tools to analyze ChIP-seq data, our statistical algorithm was developed specifically to identify significantly accessible genomic regions by handling features of DNaseI hypersensitivity. Without prior knowledge of relevant protein factors, one can discover genome-wide chromatin remodeling events associated with specific conditions or differentiation stages from quantitative analysis of DNaseI hypersensitivity. By performing appropriate subsequent computational analyses on a select subset of remodeled sites, it is also possible to extract information about putative factors that may bind to specific DNA elements within DNaseI hypersensitive sites. These approaches enabled by DNaseI-seq represent a powerful new methodology that reveals mechanisms of transcriptional regulation.

Keywords

Chromatin; Chromatin remodeling; DNaseI hypersensitivity; Global DHS-seq analysis; Global ChIP-seq analysis; Genome-wide; High-throughput deep sequencing; Computational methods; Footprinting

1. Introduction

Chromatin remodeling is the mechanism by which tightly packaged chromatin in the nucleus of a eukaryotic cell is modified such that transcription factors or other nuclear proteins can access their cognate sites on the DNA (1). The chromatin exerts a significant regulatory effect upon the expression of genes, for example. Transcription factor binding sites can be remarkably clustered and access to these elements is tightly regulated in a cell- and condition-specific manner (2–4).

The accessibility of the chromatin template can be measured by the susceptibility of DNA to cutting by DNaseI (3). Chromosomal regions that have an open chromatin structure are hypersensitive to DNaseI cutting and are consequently termed DNaseI hypersensitive sites (DHSs). DHSs often coincide with regulatory elements, such as promoters, enhancers, insulators, silencers, and locus control regions (1, 5–7). DNaseI hypersensitivity coupled with chip (DNaseI-chip) and more recently with sequencing (DNaseI-seq) has provided genome-wide identification of functional regulatory elements (Fig. 1) (3, 8, 9). This

approach makes it possible to achieve a continuous high-resolution profile of DNaseI sensitivity along the genome that represents a map of quantitative chromatin accessibility (2, 10).

If the DNA fragments cut and released by DNaseI are sequenced deeply to yield very large numbers of reads, narrow regions of protection against DNaseI cleavage, caused by transcription factor binding, can be identified as “footprints” (7, 10). Although the cost of sequencing becomes an issue in practice, sufficient tag coverage allows pinpointing of specific binding sites at nucleotide resolution.

This chapter provides a technical description of the computational procedures that can be employed to analyze DNaseI-seq data. These algorithms are categorized into different types of computational analyses: (1) generation of tag density maps for visual exploration; (2) hotspot detection based on a background probability model and calculation of statistical measure of enrichment; (3) artifact filtering; (4) footprint detection for extra-deep sequencing data; (5) annotation of the identified regions with respect to other genomic features or related data; and (6) downstream analyses that the individual studies call for.

2. Data and Sources of Bias

Current high-throughput sequencing of a DNaseI-seq sample routinely produces tens of millions of short sequence reads of 25–80 bp in length. In single-end sequencing, these short reads or tags belong to one end of the DNA fragments that were doubly cut by DNaseI and size selected for 100 bp–1 kb range. Sequence reads are then aligned to the reference genome and regions with densely accumulated tags are putative DHSs (2, 3).

Although DNaseI-seq data are highly reproducible and visually convincing (Fig. 1), there are systematic biases that are readily observed and should be corrected. For example, all the reads are aligned to the reference genome with the assumption that the genomic sequence of the cells used for the experiment is identical to the reference sequence, whereas many commonly used cell lines harbor genomic aberrations, such as polyploidy, translocations, or other mutations. Even karyotypically normal cells may have variations from the reference genome. Amplified regions would contribute more input to the DNA sample and deleted regions would not support any sequence reads.

Another source of bias arises from the highly variable tag “mappability” together with the fact that the initial processing of sequencing data relies on alignment of short reads for assigning their genomic coordinates. A given n-mer may occur at a unique location or at multiple genomic positions. Most alignment algorithms produce genomic coordinates only for uniquely mappable reads, which creates “dark spots” across the genome that directly affect the background probability of observing tags at any given position.

Identification of the genomic regions, where tags are truly enriched over the background, must take into account these and other sources of bias and artifacts in the sequencing data. The objective of an algorithm for DHS detection is to find all of the truly enriched sites while minimizing the false-positive rate. Although various software packages have been developed for calling peaks or enriched sites from ChIP-seq data, most publicly available

software are not optimally designed for handling DNaseI-seq data. An ad hoc solution has been to use a ChIP-seq algorithm that does not require input DNA control (11). In the following section, we describe an algorithm that identifies regions of significant tag enrichment [termed hotspots (2, 10)] from DNaseI-seq data.

3. Methods

3.1. Building a Tag Density Profile for Data Visualization in a Browser

A tag density profile is a direct summary of the sequencing data showing the distribution of tag counts over the genome. This basic information serves as a first-pass assessment of the data, since some data quality problems can be found at the early stage. Because the size-selected DNA fragments (100–1,000 bp) are longer than the sequenced reads (25–80 bp) from the 5' ends, positional adjustments are necessary to estimate the real distribution of fragments in the DNaseI-treated sample. If the average fragment length is known, each sequence tag can be extended in the 3' direction up to that length. When paired-end sequencing is employed, the fragments are obtained directly and can be used to generate the tag (or fragment) density profile. After the adjustment, the tag density profile is calculated simply by counting the number of tags overlapping a given nucleotide (or a bin if a lower resolution is acceptable).

Simultaneous visualization of the tag density profile and identified enriched sites (Fig. 1, see Subheading 3.2) is essential to confirm and understand the data and analysis output. There are several publicly available browser software tools that accept genomic data files in predefined formats and display them in the context of annotation tracks, such as known genes, ESTs, and comparative genomic features.

The University of California Santa Cruz (UCSC) Genome Browser is a popular Web-based browser (<http://genome.ucsc.edu>). It provides a rich database of useful information, including known genes and ncRNA, repeat elements, and publicly available data tracks from the ENCODE project, which can be used to explore the data with additional correlative features. Currently, the UCSC Genome Browser supports various data formats, such as BED, BedGraph, GFF, WIG, and BAM. The UCSC Web site also provides the Table Browser from which one can download data tracks seen in the browser (<http://genome.ucsc.edu/cgi-bin/hgTables>).

Integrated Genome Browser (IGB) is also a powerful browser for high-throughput genomic data (<http://www.bioviz.org/igb/>). IGB supports many reference genomes through the DAS/2 protocol and a custom genome can be created and loaded by the user. An attractive feature of IGB over the UCSC Genome Browser is the speed of browsing: zooming or updating the data tracks is very fast because it runs as a local application and the accompanying annotation tracks are minimal. Moreover, IGB can call upon the same genomic region in the UCSC Genome Browser with a single click, making it easy for the user to identify regions of interest quickly by browsing with IGB and then accessing the vast annotation database of UCSC Genome Browser for more detailed information on the relevant regions.

3.2. Hotspot Detection Algorithm

Our algorithm identifies regions of local enrichment, termed hotspots, of short-read tags mapped to the reference genome as follows (see also Note 1). Enrichment of tags in a 250-bp target window relative to a 200-kb surrounding window (local background) is gauged by the model based on the binomial distribution. Each tag is extended to be 150 bp in length, or the average size from selection, into its strand. Then, the 250-bp target window centered at the extended tag is assigned a z score as explained below using the background window centered at the tag. An unthresholded hotspot is defined as a contiguous cluster of 250-bp windows whose z scores are nominally significant (greater than 2). Once a hotspot is called, the z score of the hotspot is taken to be the highest z score from the constituent tags.

3.2.1. Mappability-Adjusted Z Scores—If there are n observed tags and N tags that overlap the target window and the local background window, respectively, then the probability p of a tag in the background window overlapping the target window is given by the ratio [# of uniquely mappable tags for the 250-bp window]/[# of uniquely mappable tags for the 200-kb window]. Because not all 36-mers (a common read length from sequencers) in a window can be aligned uniquely to the reference genome, p differs greatly from genomic region to region. The expected number of tags overlapping the target window is then $\mu = Np$ after this important step of adjusting for the differential mappability of short reads. The standard deviation of the expectation is: $\sigma = \sqrt{Np(1-p)}$. The z score for the observed number of tags in the target window is defined by: $z = \frac{n - \mu}{\sigma}$.

3.2.2. Peak Detection—Hotspots vary in size because wide regions of DNaseI hypersensitivity can span several kb. We employ a peak finding procedure to detect narrow 150-bp regions of tag enrichment within hotspots. For each hotspot, we compute a new z score “peak z ” similarly as for hotspot detection, but instead using a 150-bp window within the hotspot and the hotspot as the target window and local background window, respectively. The peak z score is calculated by scanning through the hotspot with bp resolution for all possible 150-bp windows. A putative peak is defined as a 150-bp window whose peak z score is above 30. If nearby putative peaks overlap within a hotspot, the 150-bp window with the highest peak z score is selected to be the peak among the cluster of overlapping putative peaks. Therefore, a hotspot may not have any peak or may have multiple peaks (Fig. 1b).

3.2.3. Repeat Sequence Filtering—In sequencing reads, artifacts are inevitably found and typically concentrated on relatively small regions. Before calling hotspots and peaks, these artifacts are eliminated by filtering out sequence reads which overlap satellites, long interspersed repetitive elements (LINE), and short tandem repeats (STR) after extending tags into 150 bp. Repeat masking files can be downloaded from the UCSC Web site <http://hgdownload.cse.ucsc.edu/downloads.html>.

3.2.4. False Discovery Rate Calculations—The final z -score threshold is imposed on candidate hotspots for each data set based on false discovery rate (FDR). To calculate FDR, random sampling of uniquely mappable tags is performed to generate a random data set with the same number of tags as for the observed data set. Hotspots are called for the random data and their z scores are calculated. Then, the FDR for the observed hotspots at a

given z score threshold z_0 is estimated as $[\# \text{ of hotspots with } z > Z_0 \text{ in the random data}]/[\# \text{ of hotspots with } z > z_0 \text{ in the observed data}]$. Hotspots are selected by setting a z -score threshold that corresponds to the desired FDR.

3.2.5. Replicate Concordance—When two biological replicates of the data are available, a replicate concordant set of hotspots is defined as follows (see Note 2). First, for each replicate, minimally thresholded (z score > 2) hotspots are called and nearby hotspots are merged according to a distance threshold (within 150 bp, for example). The initial replicate concordant regions are defined as the intersection of these hotspots between the replicates. Then, the tags from both replicates are combined and the detection algorithm calls FDR-thresholded hotspots from the pooled tags. Finally, FDR-thresholded replicate concordant set of hotspots is defined as the intersection of the FDR-thresholded set of hotspots from the combined data and the set of replicate concordant regions.

3.3 . Digital Footprinting Analysis

Unlike ChIP which identifies binding sites for a given protein, digital footprinting can detect sites bound by any factors from the same sample with a nucleotide precision (7, 10). To find footprints, one looks for narrow regions (8–30 bp) on which DNaseI cleavage is significantly reduced in comparison with that for the immediately surrounding regions. The analysis requires extremely deep sequencing to achieve reasonable coverage of cleavage events. John Stamatoyannopoulos and coworkers used a detection algorithm to identify footprints in DNaseI-seq data from *Saccharomyces cerevisiae*. The software (available at <http://www.nature.com/nmeth/journal/v6/n4/extref/nmeth.1313-S2.zip>) does not scale very well with larger mammalian genomes. We are currently developing computational methods that can be applied to data sets for mammalian genomes.

3.4. Annotation and Integrative/Downstream Analyses

Once the hotspots and the peaks are obtained according to the above procedures, it is often desirable to associate the sites with additional correlative information, such as the closest genes or regions found from other relevant experiments. For instance, one can examine the proportions of accessible sites located at promoters, introns, or intergenic regions. When there are several experimental conditions or treatments, the maximum tag density of hotspots can be used to perform cluster analysis or supervised classification to reveal chromatin accessibility changes across the samples (4).

Motif discovery is a downstream analysis, where one looks for putative protein-binding events within specified DHSs. There are two types of such sequence analyses. One method is simply scanning the DNA sequences for the presence of motifs for known factors. It requires prior knowledge of binding motifs, but the computation is straightforward and focused (FIMO available at <http://meme.sdsc.edu/meme/fimo-intro.html>; MAST at http://meme.nbcr.net/meme4_5_0/mast-intro.html) (12). The other analysis discovers novel motifs enriched in the target DNA sequences, which is computationally very intensive due to the large number of DHSs that are often taken as search input (see Note 3). For preparing input, the UCSC Table Browser can be used to extract the DNA sequences of specified genomic

regions in the FASTA format, although there is a limit on the number of sites. The widely used de novo motif discovery tool, MEME (13, 14), uses an expectation maximization (EM) algorithm and reports ranked potential motifs (see Note 4). Among the motifs from the MEME output, those matching known binding motifs can be identified by motif database query tools, like TomTom (http://meme.nbcr.net/meme4_5_0/cgi-bin/tomtom.cgi) (15).

4. Software

The software package implementing the hotspot detection algorithm is available upon request (see Note 5). The software was written in C++ and compiled by GNU C++ compiler. As output, the program reports the genomic location, maximum tag density, average tag density, and statistics for each hotspot. Accompanying R programs perform some downstream analyses, such as generation of an annotation table for hotspots and rendering Venn diagrams from comparison of multiple data sets.

5. Notes

1. Our algorithms for detecting hotspots and replicate concordance are similar to the two-pass hotspot detection algorithm developed by Robert Thurman for the ENCODE project (2). The two-pass algorithm addresses the issue, where significantly enriched regions are sometimes shadowed by a neighboring extreme hotspot. After obtaining hotspots, it detects a second set of hotspots using the data from which all the tags in the first set of hotspots are deleted. Therefore, the algorithm generally detects more hotspots than our approach while comparison of the significance *z* scores of different-pass hotspots becomes less intuitive.
2. Analogous definitions of replicate concordance can be formulated for three or more replicates.
3. It is often worthwhile to group sites from DNaseI-seq data into more biologically defined subsets. This can be guided by additional data or available information, such as gene expression, proximity to promoters, or colocalization with factor binding or motifs. The input for MEME can be further reduced by taking the top proportion of sites and/or trimming the width of individual sites.
4. Interpretation of an MEME report requires care and proper understanding of the discovery algorithm to sort out the motifs that are likely to be meaningful. Often, the objective is to discover candidate protein factors associated with chromatin remodeling at the DHSs. However, strong signals from sequence-intrinsic motifs that disfavor nucleosome occupancy (thus, abundant in the DNaseI-treated sample) or other repetitive sequence patterns can dominate the top-ranking motifs and mask more subtle but specific motifs that arise from protein binding.
5. The software requires the mappability file of the matching genome. The mappability file for *k*-mers is a binary file with the value (0–255) at the *n*th byte representing the number of occurrences of the *k*-mer sequence starting from the genomic coordinate *n*. Our software uses the same mappability file format as in PeakSeq, and the user can generate a new file using the mappability generation

program of PeakSeq (http://archive.gerstein.02nlab.org/proj/PeakSeq/Mappability_Map/Code/Mappability_Map.tar.gz).

References

1. Wiench M, Miranda TB, and Hager GL (2011) Control of nuclear receptor function by local chromatin structure. *FEBS J*, 278(13): 2211–30. [PubMed: 21501385]
2. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, and Stamatoyannopoulos JA (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet*, 43, 264–268. [PubMed: 21258342]
3. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, and Stamatoyannopoulos JA (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods*, 3, 511–518. [PubMed: 16791208]
4. Siersbaek R, Nielsen R, John S, Sung MH, Baek S, Loft A, Hager GL, and Mandrup S (2011) Adipogenic development is associated with extensive early remodeling of the chromatin landscape and establishment of transcription factor ‘hotspots’. *EMBO J*, 30, 1–14. [PubMed: 21206509]
5. Hager GL, Nagaich AK, Johnson TA, Walker DA, and John S (2004) Dynamics of nuclear receptor movement and transcription. *Biochim. Biophys. Acta*, 1677, 46–51. [PubMed: 15020044]
6. Hager GL, Elbi C, Johnson TA, Voss TC, Nagaich AK, Schiltz RL, Qiu Y, and John S (2006) Chromatin dynamics and the evolution of alternate promoter states. *Chromosome. Res*, 14, 107–116. [PubMed: 16506100]
7. Hager GL (2009) Footprints by deep sequencing. *Nat. Methods*, 6, 254–255. [PubMed: 19333240]
8. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, and Collins FS (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods*, 3, 503–509. [PubMed: 16791207]
9. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, and Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132, 311–322. [PubMed: 18243105]
10. Hesselberth JR, Zhang Z, Sabo PJ, Chen X, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, and Stamatoyannopoulos JA (2009) Global mapping of protein-DNA interactions in vivo by digital genomic foot-printing. *Nat. Methods*, 6, 283–289. [PubMed: 19305407]
11. Ling G, Sugathan A, Mazor T, Fraenkel E, and Waxman DJ (2010) Unbiased, genome-wide in vivo mapping of transcriptional regulatory elements reveals sex differences in chromatin structure associated with sex-specific liver gene expression. *Mol Cell Biol*, 30, 5531–5544. [PubMed: 20876297]
12. Bailey TL and Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14, 48–54. [PubMed: 9520501]
13. Bailey TL, Williams N, Misleh C, and Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34, W369–W373. [PubMed: 16845028]
14. Bailey TL and Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf Intell. Syst. Mol Biol*, 2, 28–36. [PubMed: 7584402]
15. Gupta S, Stamatoyannopoulos JA, Bailey TL, and Noble WS (2007) Quantifying similarity between motifs. *Genome Biol*, 8, R24. [PubMed: 17324271]

Fig. 1. Example sequence tag profiles of DHS elements. **(a)** DNaseI hypersensitive sites (DHSs) are measured by deep sequencing of small, size-selected DNA fragments released by DNaseI treatment from regions of locally reorganized chromatin (2). **(b)** The DNaseI-seq data are presented as tag density profiles, summarizing the number of sequence tags that are aligned to the region. Density values can be normalized by the total number of aligned sequence tags to aid comparison of multiple samples. A DNaseI-seq profile obtained from a mammary carcinoma cell line is shown for a region in chromosome 1 (genome build: mm8). Regions

of increased DNaseI sensitivity are detected by the computational algorithms as “hotspots” (DHS hotspots track). Within these hotspots are found localized peaks of hypersensitivity (DHS peaks track).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript