Briefings in Bioinformatics, 20(5), 2019, 1865–1877

doi: 10.1093/bib/bby057
Advance Access Publication Date: 4 June 2019
Review article

A practical guide for DNase-seq data analysis: from data management to common applications

Yongjing Liu, Liangyu Fu, Kerstin Kaufmann, Dijun Chen and Ming Chen

Corresponding authors: Ming Chen, Department of Bioinformatics, State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou 310058, China. Tel.: +86(0)571-88206612; Fax: +86(0)571-88206612; E-mail: mchen@zju.edu.cn; Dijun Chen, Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, Germany. E-mail: chendijun2012@gmail.com

Abstract

Deoxyribonuclease I (DNase I)-hypersensitive site sequencing (DNase-seq) has been widely used to determine chromatin accessibility and its underlying regulatory lexicon. However, exploring DNase-seq data requires sophisticated downstream bioinformatics analyses. In this study, we first review computational methods for all of the major steps in DNase-seq data analysis, including experimental design, quality control, read alignment, peak calling, annotation of *cis*-regulatory elements, genomic footprinting and visualization. The challenges associated with each step are highlighted. Next, we provide a practical guideline and a computational pipeline for DNase-seq data analysis by integrating some of these tools. We also discuss the competing techniques and the potential applications of this pipeline for the analysis of analogous experimental data. Finally, we discuss the integration of DNase-seq with other functional genomics techniques.

Key words: DNase-seq; pipeline; chromatin accessibility

Introduction

Deoxyribonuclease I (DNase I) is an enzyme that cuts DNA in a desultory manner. In eukaryotes, chromosomal DNA is packaged into a regularly repeating chain of nucleosomes. These nucleosomes will block DNase I from easily nicking DNA, leading to preferential sensitivity of the accessible nucleosome-free regions to DNase I cleavage. Active genes are more likely to have altered nucleosome state [1], which makes DNase I digestion a great reference measure for mapping genomic regulatory elements. Since the concept of DNase I-hypersensitive sites (DHSs) [2] was coined to describe open chromatin regions, the attention to DNase I studies rapidly reached its zenith in the 1980s but gradually decayed. The main barrier to further progress appears to be the lack of high-throughput analyses, as the data generated by traditional methods were not sufficient enough to draw significant conclusions at a genome-wide scale. However, the interest in DHS profiling has recently revived since the emergence of next-generation sequencing (NGS).

The combination of DNase I digestion and high-throughput sequencing introduces an emerging technique known as DNase I-hypersensitive site sequencing (DNase-seq) [3], which allows genome-wide mapping of DNase I cleavage events at nucleotide resolution and shows an improved signal-to-noise ratio compared with its predecessors. Based on DNase-seq, extensive researches for different purposes have been conducted, including the investigation of nucleosome

Yongjing Liu is a PhD candidate at the Department of Bioinformatics, College of Life Sciences, Zhejiang University, China. His research interests include cancer genomics and multi-omics analysis.
Liangyu Fu is a master of genomics and a guest scientist at Humboldt-Universität zu Berlin. Her research centers on plant epigenetics, including the processing of Hi-C and ChIP-seq data.
Kerstin Kaufmann is a professor at the Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, Germany. Her current research theme is plant developmental genetics and evolution.
Dijun Chen is the corresponding author. He is a postdoctoral research fellow at the Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, with expertise in plant molecular biology.
Ming Chen is the corresponding author. He is the director of the Bioinformatics Lab of Zhejiang University, and a professor at the Department of Bioinformatics, College of Life Sciences, Zhejiang University, China.
Submitted: 8 March 2018; **Received (in revised form):** 6 June 2018

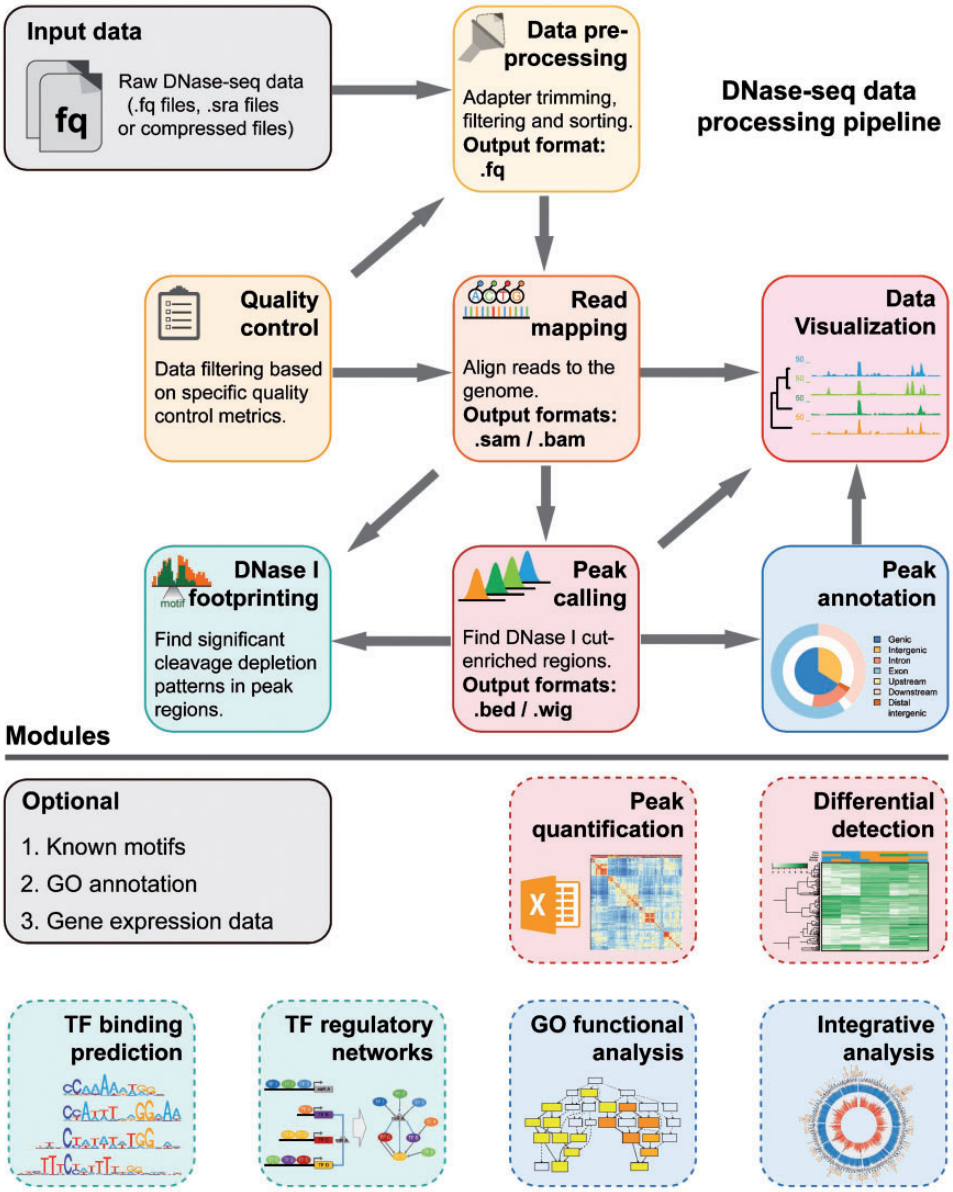

Figure 1. A modularized pipeline for analyzing DNase-seq data. The processing steps (modules) for DNase-seq data are represented as boxes, and the dependence between major modules is indicated by arrows. Several optional modules that need additional data are listed below the line.

positions [4], the identification of genomic regions with nucleosome rotational stability [5] and the recognition of regulatory quantitative trait loci underlying expression variation [6]. Apparently, the application of DNase-seq could be various, and from an analysis point of view, a single computational pipeline cannot satisfy every study. To produce a reliable and reproducible study, one should carefully design the roadmap and try to use the appropriate methods and optimal parameters, as any small change at the beginning may lead to a huge change in result. Therefore, a commented guideline for DNase-seq studies will help people to focus on their scientific inquiries.

The primary purpose of this survey is to offer perspectives on DNase-seq data analysis and review widely used tools in each step. We present a modularized pipeline for the analysis of DNase-seq data in Figure 1, and all major steps in the pipeline are discussed in detail in the following sections. Finally, the common applications of data integration and the potential strategies to reduce the biases are also discussed.

Experimental protocols

Currently, there are two main experimental protocols for DNase-seq, the 'end-capture' protocol [7] developed in Duke University and the 'double-hit' protocol [8] developed in University of Washington. Both protocols are widely used and have successfully been applied for DHS detection in the Encyclopedia of DNA Elements (ENCODE) project [9]. The two protocols share a common architecture (Figure 2), the only essential difference between them is the library preparation step. The 'end-capture' protocol ligates all fragment ends to a specially designed linker and performs an extra MmeI digestion to generate 20 bp tags for amplification and sequencing. The 'double-hit' protocol involves the use of gel electrophoresis to select

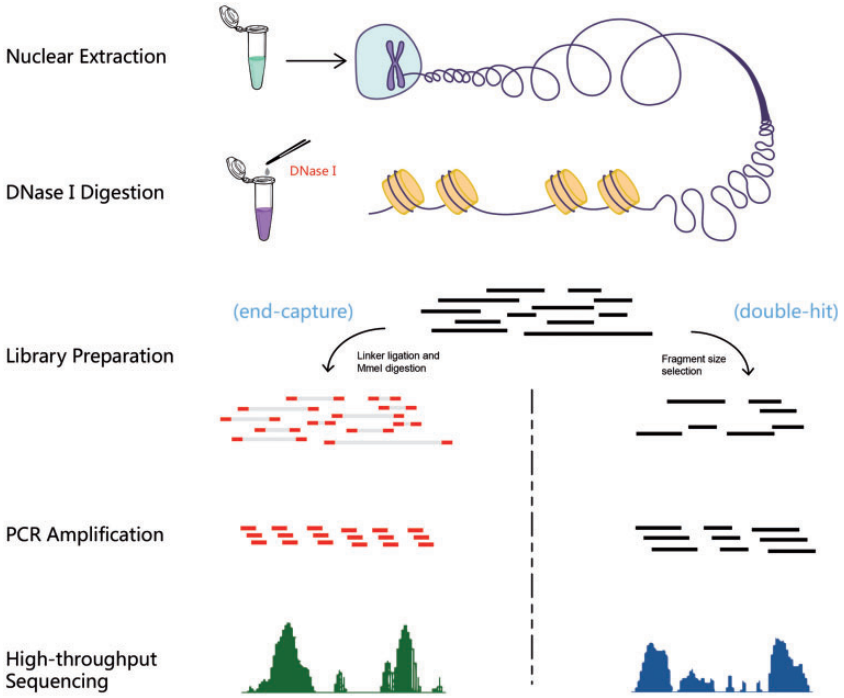

Figure 2. Overview of DNase-seq experimental protocols. Basic steps necessary for a DNase-seq experiment are shown in the figure, including nuclear extraction, DNase I digestion, library preparation, PCR amplification and high-throughput sequencing. Different strategies for library preparation are adopted by end-capture and double-hit methods.

small-sized fragments, as it is argued that close cleavage events are of greater importance. Recommended fragment size is 50–100 bp for 'double-hit' protocol, which is shorter than the 147 bp DNA wrapping one nucleosome and longer than normal linker regions [10]. Both methods can use naked DNA as control. The end-capture protocol keeps the integrity of cleavage events, but the double-hit protocol is now preferred, with longer reads, cleaner data, easier operations and higher specificity for uncovering DHSs.

Based on the two types of protocols, a number of customized protocols are developed for specific aims (Table 1). The choice for protocols should be in line with the experimental design, and minor modifications could be made to meet requirements. Note that DNase I concentration and digestion time will directly influence the digestion level [11]. Important factors including sequencing depth and the use of control samples/replicates should also be taken into consideration when designing the study. Zeng *et al.* [12] suggested the minimal sequencing depth sufficient for accurately profiling for different purposes.

Quality control and data manipulation

Quality control is one of the most important parts in processing DNase-seq data. DNase-seq raw data (generally in FASTQ formatted files) show no fundamental difference with other types of sequencing data; therefore, typical quality metrics for raw reads are applicable. FastQC [13] provides a series of quality control checks on raw sequence data involving GC content, sequence quality and duplicate sequences. Note that, the number of duplicate reads is expected to be high for DNase-seq, as DNase-seq reads concentrate at limited DHSs. Besides, for single-end sequencing, the duplicate reads may arise from different fragments sharing the same end position, which is also common for DNase-seq. According to Meyer *et al.* [14], discarding duplicates is likely to distort read signal quantification. Therefore, users may ignore the 'fail' for 'Duplicate Sequences' or adjust default warn/fail thresholds. Based on the results, low-quality reads and adapter-contaminated reads can then be filtered and trimmed, which can be done by many tools such as Trimmomatic [15]. Del *et al.* [16] have reviewed the read trimming tools for NGS data. Other toolkits like Bbtools [17] and AfterQC [18] provide both quality checking and read preprocessing functions.

After preprocessing, qualified reads need to be mapped to the reference genome for further analysis. The most common methods for read alignment include BWA [19] and Bowtie2 [20]. Post-mapping quality control methods like Picard [21] and SAMStat [22] are often used for assess GC content of mapped reads, read coverage, homopolymer biases and other experimental artifacts. However, as the quality standards are highly specific to the type of used techniques, the outcome of quality control methods might be inadequate to assess the quality of DNase-seq data. While double-hit DNase-seq data and chromatin immunoprecipitation sequencing (ChIP-seq) data are highly similar, specific quantitative measurements of ChIP-seq data quality may be practically not suitable for DNase-seq. For instance, phantompeakqualtools [23] provides three specific metrics for ENCODE data. The PCR bottleneck coefficient measures the distribution of reads, which is inherently skewed to DHSs. The normalized strand cross-correlation coefficient and relative strand cross-correlation coefficient are somewhat meaningless, as DNase-seq data show low-strand asymmetry. Other ENCODE metrics like irreproducible discovery rate and Signal Portion of Tags (SPOT) are suitable for DNase-seq data. According to the DNase-seq data standards provided by ENCODE (https://www.encodeproject.org/data-standards/dnase-seq/), the SPOT score should be no <0.3, which means at

Table 1. DNase-seq experimental protocols

Methods	Considerations	Citations
DNase-seq (end-capture)	Identification of DHSs at whole-genome scale using 20bp fragment ends	[7]
DNase-seq (double-hit)	NGS after size fractionation of DNase I-digested fragments	[12]
DNase-FLASH	Separately sequencing and mapping smaller and larger DNase I fragments from the same DNase I digestion experiment	[13]
DNase I SIM	Efficient and time-saving DNase-seq for plant tissues	[14]
scDNase-seq	DNase-seq analysis for single-cell or limited-cell inputs	[15]

least 30% reads should fall in tag-enriched regions. These ENCODE criteria (also including the proportion of duplicate reads, mitochondrial reads, unique mapped reads, etc.) are recommended when performing a new DNase-seq experiment. Note that the mapping efficiency for end-capture-style reads is always weaker, because of its fixed 20-bp read length.

A specially designed quality control tool, ChiLin [24], provides multiple quality control metrics across different levels for DNase-seq data. ChiLin is an integrative pipeline that automates both the data analysis and quality control of DNase-seq. Specifically, ChiLin collected a comprehensive resource of DNase-seq data, including DHS and ENCODE Blacklist regions [25], and using them for comparisons at the post-peak-calling level.

Peak calling

DNase-seq is developed for large-scale profiling of DHSs, and most DNase-seq studies critically depend on calling DHSs (or 'peaks') from DNase-seq data.

The general aim of peak calling is to find genomic locations in which reads pile up to form pulse shapes, which can be achieved by multiple tools (Table 2). For example, F-Seq [26] is based on a simple approach by dividing the reference genome into nonoverlapping, equal-sized bins and counts the reads starting in each bin. Despite the low resolution, the result of this rough method seriously depends on the segmentation points of bins, so F-Seq applies a Gaussian smoothing kernel with a customized bandwidth considering the bin boundary effects. MACS2 [27] uses a sliding window instead of the separate bins, allows the inclusion of control samples to estimate the background signal and calculates fold changes inside the windows. Benjamini–Hochberg adjustment is applied on *P*-values calculated by dynamic Poisson distribution, and the regions with significant false discovery rate (FDR) values and fold changes are determined as DHSs. Unlike MACS2 that may find a broad peak, the length of DHS peaks is restricted in the Hotspot algorithm [28, 29]. Hotspot calculates z-scores using binomial distribution to find 'hotspot' regions, and a 150-bp peak with maximum density value is identified inside each region. A modified version of hotspot, DNase2Hotspots [30], also provides FDR values for each DHS. ZINBA [31] applies a mixture regression model including a set of covariates such as GC content and background signal to classify each genomic window. ZINBA accounts for factors like copy number and can determine the contributions of components for each window. However, ZINBA was recognized as less sensitive than the other three methods, according to the evaluation by Koohy *et al.* [32]. A more recent peak caller, Dfilter [33], uses a linear filter that maximizes the signal-to-noise z-score as the Hotelling observer for peak detection. DFilter supports integrative signal detection and may have better accuracy under certain circumstances.

The algorithms mentioned above are the most popular peak callers for DNase-seq studies, and most of them have considered the characteristics of DNase-seq or have built-in DNase-specific options. Nonetheless, few algorithms have been designed specifically for DNase-seq peak calling. Owing to the high similarity between DNase-seq reads (especially for double-hit protocol) and ChIP-seq reads, many studies have applied ChIP-seq peak-calling tools for DNase-seq analyses. Reuben *et al.* [34] provided rationale for selecting ChIP-seq peak-calling methods for a given application. In view of the difference in the nature of the techniques, setting the parameters of these tools is vital for improving the fit to DNase-seq data. For ChIP-seq data, the peaks exhibit the enrichment of immunoprecipitationed DNA fragments and may indicate a transcription factor (TF)-binding site. However, researchers are usually interested in the whole protein-binding sequence of ChIP-seq, but the fragment end representing a cleavage site is what matters more in DNase-seq. This is the reason that F-seq uses only the 5′-end base for peak calling. The method is more recommended for end-capture-style data, as all tag lengths are the same (20 bp). The findPeaks function of HOMER [35] also provides a specific option (-style dnase) for end-capture-style data. As for double-hit-style DNase-seq data, the read length is always less than fragment size, and the read distribution cannot reflect real positions of binding sites. To overcome this, two strategies, tag extension (elongate read length to inferred fragment size *n*) or tag shifting (move the reads by *n*/2 toward the center of fragment) are applied widely in ChIP-seq peak-calling algorithms. Although shifting and extension are not needed for DNase-seq analysis, the parameters, however, cannot be simply set as 0. Take MACS2 for instance, the –extsize option can not only represent the extended read length but also be seen as a window easier for reads to pile up. The option '–nomodel –shift -75 –extsize 150' allows user to find peaks using a 150-bp smoothing window, with the center still at the cutting site. For histogram-based methods, the length of each window will affect the outcome even more. The recommended window length is no longer than the nucleosome size to improve resolution and half-nucleosome size for nucleosome position mapping studies.

Results produced by different peak callers may vary greatly [32] because of their different intrinsic models. However, the most pronounced DHSs can always be identified using any algorithms, and the less significant ones account for the main difference. These shallow or narrow peaks are not necessarily false DHSs, as they may be caused by underdigestion, low sequencing depth or even masked by transcription factor binding [10]. To gain more information about potential regulatory regions, sometimes threshold adjustments are needed, including *P*-value/*q*-value, fold-change and absolute signal intensity. A proper combination strategy for different peak callers may also help users have robust data interpretations. Besides, the use of control samples would be useful to achieve better

Table 2. Computational methods for DNase-seq data analysis

Methods	Language/ version[a]	Descriptions	URL
Quality control: Data filtering and bias characteristics of DNase-seq samples before/after mapping			
AfterQC	Python and C/ v0.9.7	Automatically filtering, trimming and quality control for FASTQ data. Generates statistical information with figures	https://github.com/OpenGene/AfterQC
FastQC	Java/v0.11.5	Performs quality control checks on raw sequence data. Provides an HTML report with graphs and tables	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Hotspot	C++ and Shell/ v4.1.0	Calculates the SPOT score. For a sample, a higher SPOT score indicates a higher signal-to-noise ratio. A SPOT of 0.4 or higher is considered high quality for DNase-seq data	http://www.uwencode.org/proj/hotspot/
Picard	Java/2.17.3	Post-mapping quality control. Supports SAM/BAM/CRAM and VCF formats	http://broadinstitute.github.io/picard/
Sickle	C/1.33	Trimming FASTQ files based on quality and length thresholds, includes a paired-end mode	https://github.com/najoshi/sickle
Trimmomatic	Java/0.36	Adapter trimming for illumina reads, containing a variety of functions	http://www.usadellab.org/cms/? page=trimmomatic
Read mapping: Find the locations of DNase-seq reads on a reference genome			
Bowtie2	C++/2.2.9	Aligning NGS reads to a reference genome. Quick and memory-efficient	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA	C/0.7.17	Similar to Bowtie2. Bwa-aln is recommended for shorter reads. Choose bwa-mem for high mapping rate and efficiency	http://bio-bwa.sourceforge.net/
HISAT2	C++/2.1.0	Can handle known SNPs during alignment. Less resource demanding	https://ccb.jhu.edu/software/hisat2/index.shtml
STAR	C++/2.5.3	An aligner that may have better performance for low-quality genomes	https://github.com/alexdobin/STAR
Peak calling: Detect genomic regions with DNase I cleavage enrichment			
DFilter	Matlab/1.6	Detects potential peaks in tag-profile of sequencing data, using an receiver operating characteristic-area under curve (ROC-AUC) maximizing linear filter	http://collaborations.gis.a-star.edu.sg/~cmb6/kumarv1/dfilter/
DNase2Hotspots	C++/NA	Identifies tag-enriched regions (hotspots) of DNase-seq data with corresponding z-scores	https://sourceforge.net/projects/dnase2hotspots/
F-seq	Java/1.85	Calling peaks based on a continuous tag density estimation	http://fureylab.web.unc.edu/software/fseq/
MACS2	Python/2.1.1	A peak caller designed for ChIP-seq, works well with DNase-seq data	https://github.com/taoliu/MACS/
PeaKDEck	Perl/v1.1	A kernel density estimator-based DNase-seq peak caller with an available graphical user interface version	http://www.ccmp.ox.ac.uk/peakdeck
Popera	Python/NA	With similar capabilities as F-Seq	https://github.com/forresthang/Popera
ZINBA	R/2.03.1	A DNase-seq peak caller that allows mixture modeling with a set of genomic factors	https://code.google.com/p/zinba/
Peak analysis: Annotation and quantification for identified peak regions			
CEAS	Python / 1.0.2	Provide genome features at the peak regions. Allow analyses on loci from a custom file	http://liulab.dfci.harvard.edu/CEAS/
ChIPpeakAnno	R / 3.12.4	Annotation, comparison and visualization for peak regions	https://bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html
ChIPseeker	R / 1.15.2	Similar to ChIPpeakAnno. Can be used on any organism if the corresponding TxDb object is available	http://bioconductor.org/packages/release/bioc/html/ChIPseeker.html
GREAT	NA / 3.0.0	A webserver for functional annotation of peak regions	http://great.stanford.edu/public/html/
UROPA	Python & R / 2.0.2-alpha	Annotation for peak regions. Permits linkage of individual queries including prioritization	https://github.molgen.mpg.de/loosolab/UROPA
Footprint detection: *De novo* methods			
DNase2TF	R / 1.0.1	Finds significant footprint patterns within a set of candidate regions	http://sourceforge.net/projects/dnase2tfr/
DNaseR	R / 1.4	Strand-specific digital genomic footprinting for 'double-hit' DNase-seq data	https://bioconductor.riken.jp/packages/3.0/bioc/html/DNaseR.html
fp2012	C++ / NA	Calculate the FOS to find footprints. Lower FOS indicates more significant result	https://github.com/sjneph/footprinting2012
HINT/HINT-BC	Python / 0.11.1	Footprinting for DNase-seq and ATAC-seq data. HINT-BC is the bias-corrected version	http://www.regulatory-genomics.org/hint/

Continued

Table 2. (continued)

Methods	Language/ version[a]	Descriptions	URL
Wellington/ Wellington-bootstrap	Python / 0.2.6	Strand-sensitive footprinter. The length of candidate footprint regions and flanking regions is user-settable Wellington-bootstrap is the version for pairwise analysis	https://pypi.python.org/pypi/pyDNase
Footprint detection: Motif-guide methods			
BinDNase	R / NA	Machine learning method using data of ChIP-seq peaks and hotspot regions	http://research.ics.aalto.fi/csb/software/bindnase/
CENTIPEDE	R / 1.2	Unsupervised footprinter based on Bayesian mixture models	http://centipede.uchicago.edu/
DeFCoM	Python / 1.0.1	SVM-based footprinter for DNase-seq and ATAC-seq data	http://fureylab.web.unc.edu/software/defcom/
FootprintMixture	R / NA	A mixture modeling framework to train bias-corrected or TF-specific footprint models. Calculate FLR	https://ohlerlab.mdc-berlin.de/software/FootprintMixture_109/
MILLIPEDE	R / 1.1.0	Supervised and simplified version of CENTIPEDE	https://users.cs.duke.edu/~amink/software/millipede/
msCentipede	Python / 1.0	Extension of CENTIPEDE that accounts for heterogeneity in cleavage patterns	http://rajanil.github.io/msCentipede/
PIQ	R / v1.3	An easy-to-use set of R scripts to identify TF footprints. Needs an input motif file	http://piq.csail.mit.edu/
Romulus	R / 1.0.2	DNase-seq footprinter that combines the strengths of CENTIPEDE and Wellington	https://github.com/ajank/Romulus
Motif analysis			
EXTREME	Python / 2.0	Online EM algorithm to discover motifs in DNase-seq footprinting data	https://github.com/uci-cbcl/EXTREME
findMotifs.pl	Perl / v4.9	A tool for *de novo* motif discovery and analysis provided by HOMER	http://homer.ucsd.edu/homer/motif/index.html
MEME-ChIP	C / 4.12.0	Performs motif discovery and comprehensive motif analyses on a set of sequences	http://meme-suite.org/tools/meme-chip
SeqGL	R / 1.1.4	A Lasso-based algorithm to detect multiple TF-binding motifs from DNase-seq profiles	https://bitbucket.org/leslielab/seqgl/
Data visualization			
Circos	Perl / 0.69-6	Visualizes genomic data and long-range interactions in a circular layout	http://circos.ca/software/
IGB	Java / 9.0.0	Stand-alone visualization tool for large genomic data sets	http://bioviz.org/igb/
IGV	Java / 2.4.7	Similar to IGB. Uses one-based coordinate system, while IGB uses interbase coordinates	http://software.broadinstitute.org/software/igv/
UCSC Genome Browser	NA	Web-based tool for visualizing a requested portion of a genome at any scale, accompanied by a series of aligned annotation 'tracks'	https://genome.ucsc.edu/
WashU Epigenome Browser	NA / v45.1	Similar to UCSC Genome Browser. Currently hosts ENCODE and Roadmap Epigenomics data for human and model organisms	http://epigenomegateway.wustl.edu/browser/
WebLogo	NA	Creates sequence logos with input data. Use for motif visualization	http://weblogo.threeplusone.com/
Data format conversion			
BEDOPS	C++ / 2.4.30	Convert a variety of formats to BED. It is internally called by the Hotspot program	http://bedops.readthedocs.io/en/latest/
BEDTools	C++ / 2.27.0	Process and manipulate genomic interval files (e.g. BED, GFF and VCF)	http://bedtools.readthedocs.io/en/latest/
SAMtools	C / 1.6	Sorting, indexing, editing and format converting of SAM/BAM/CRAM files	http://samtools.sourceforge.net/
SRA Toolkit	C++ / 2.8.2-1	Download raw sequence data from NCBI SRA and convert SRA files to FASTQ files	https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
UCSC Toolkits	C / v359	Data conversion to generate wig, bigWig, bigBed and bedGraph formats	http://hgdownload.soe.ucsc.edu/admin/exe/
Integrated tools and pipelines			
BBtools	Java / 37.88	A set of handy tools for sequence data, some of which may help to process DNase-seq data −BBduk: fast and memory-efficient tool to remove adapters −BBmap: fast short-read aligner with high sensitivity and accuracy	https://jgi.doe.gov/data-and-tools/bbtools/
ChiLin	NA	A pipeline of DNase-seq/ChIP-seq quality control and analysis	http://cistrome.org/chilin/
CIPHER	NA	An easy-to-use data processing workflow platform designed for NGS data including DNase-seq	https://github.com/c-guzman/cipher-workflow-platform

Continued

Table 2. (continued)

Methods	Language/ version[a]	Descriptions	URL
HOMER	C++ & Perl / v4.9	A widely used suite of tools for NGS analyses, including quality control, peak calling, peak annotation and motif analysis –findPeaks: Peak caller with a specific option for end-capture-style DNase-seq data –findMotifs.pl: Uncovering motifs from FASTA sequence file –annotatePeaks.pl: Annotates BED-format peaks, support custom annotation files and FASTA genome files	http://homer.salk.edu/homer/index.html
pyDNase	Python / 0.2.6	A set of scripts for common uses in DNase-seq analyses and also an implementation of the Wellington footprinting algorithms A detailed tutorial from obtaining data to visualization is provided at the website https://pythonhosted.org/pyDNase/tutorial.html –dnase_ddhs_scorer.py: Calculate DNase I hypersensitivity changes between two DNase-seq data sets –dnase_average_profile.py: Together with dnase_to_javatree-view.py, enable the bias-corrected visualization of footprints	http://pythonhosted.org/pyDNase/index.html

Note: Only several popular tools and DNase-specific tools are listed. Some of the recommended tools are discussed as examples in the manuscript. *NA*: not available.
[a]Programming language and the latest version when we wrote this article.

estimation accuracy. In some cases, the sequencing depth for control samples may be different (often lower), which needs normalization to avoid incorrect estimation of background signal. A common strategy is to uniformly rescale control samples to equal depth, but specific normalization methods [36] is recommended to improve the accuracy of results.

Peak annotation

Peaks identified from DNase-seq are termed as DHSs, which is often correlated with active *cis*-regulatory elements. Therefore, peak annotation of DNase-seq could be vital in providing the basis for a functional interpretation of DHS regions. Peak callers typically generate a BED-format peak file describing the position of each peak and a wiggle signal file describing the size and shape of peaks. The peak file can be directly used in peak annotation tools, while the signal file is mostly used for visualization.

The fundamental of peak annotation methods is to associate peak genomic positions with certain functionally relevant genomic regions or elements, such as TSS (transcription start site), TTS (transcription termination site), promoter, exon, intron, 5′ untranslated region (UTR), 3′ UTR and intergenic regions. For this purpose, almost all annotation tools use at least one of the genome annotation sets, including UCSC Known Genes [37], Ensembl [38], GENCODE [39] and RefSeq [40]. The GENCODE consortium was set up to provide reference genome annotation for the ENCODE project, comprising extensive manual annotation by the HAVANA group and computational annotation by Ensembl. The RefSeq geneset is provided by NCBI and was also integrated into UCSC genome browser. According to Frankish *et al.* [41], GENCODE is recommended over RefSeq for its greater genomic coverage and more variants. In addition, DHSs can be used to annotate genomic regulatory elements such as active enhancers [25].

Generally, peak annotation tools map DHSs to proximity genes to define their target genes, so that additional functional analysis can be performed based on gene annotation. GREAT [42] is such an example, which provide multiple ontologies including Gene Ontology, Disease Ontology, phenotype data, pathway data, gene families, MsigDB [43] Immunologic Signatures and MSigDB Oncogenic Signatures. The latest version of GREAT annotates peaks using Ensembl genes and allows users to adjust the maximum association distance to TSS. In addition to mapping nearest genes by TSSs, ChIPpeakAnno [44] can also annotate peaks to overlapping and flanking genes. ChIPpeakAnno can evaluate the concordance of multiple peak files by finding overlaps among the peaks from biological replicates or different studies. CEAS [45] can take wiggle files as input to calculate average signal intensity near important genomic regions. The annotatePeaks.pl function of HOMER [35] provides a detailed annotation, including CpG islands and repeat elements, with peak score and FDR. AnnotatePeaks.pl includes a variety of options and accepts custom annotation files, even custom genomes, in GTF/GFF format. A most recent tool, UROPA [46], also supports flexible configuration, allowing it to faultlessly simulate the performance of existing tools. UROPA has a unique filtering function for the annotations, and some built-in extend functions that are easy to use.

Occasionally there is a need to integrate more information of specific interest, for example, genetic polymorphisms like single nucleotide polymorphism (SNP) or indels, or epigenetic modifications like histone marks and methylation. Aside from the common annotations, these factors are generally not included as references in annotation tools. One good solution is to upload the peak file to UCSC Genome Browser and check for desired tracks. If corresponding track is unavailable, a GTF file is needed to build a custom track. The custom annotation files can be created by tools like Cufflinks [47] using RNA-seq data. Users can also manually apply BED files obtained from sources like ENCODE [25] to annotate peaks using tools like BEDTools [48].

Footprinting

DHSs indicate nucleosome depletion, which was observed around most transcription factor binding sites [49]. Unwrapping of nucleosomal DNA is an important mechanism for easier access of proteins, leading to the exposure of naked DNA, but the regulatory proteins may bind to DNA thus protect the target sequence from DNase I cleavage. Compared with a DHS, the protected region is often narrow (8–30 bp, mostly depending on the TF-binding motif), leaving a 'footprint' with significantly

reduced signal in comparison with flanking regions. By means of DNase-seq, the signal intensity can be presented at single-nucleotide resolution, allowing a giant leap in the search for the footprints, known as footprinting.

Hesselberth *et al.* [50] performed the first large-scale DNase I footprinting (digital genomic footprinting) to survey the TF-binding sites across yeast genome. Hesselberth et al. used naked DNA as reference to control for cleavage bias, and derived TF-binding motifs *de novo* using the MEME [51] suite. The in-house footprinting software developed by the authors does not work well for large genomes but works well as an inspiration. Several new footprinting methods have emerged in the past few years, which can broadly be classified into two types based on their different methodologies. The motif-guide footprinting methods first seek for candidate TF-binding sites predicted by motif information, and then apply statistical or machine learning models to distinguish the *bona fide* binding sites from false ones. The *de novo* footprinting methods focus on the global detection of wanted signal patterns and have the potential to uncover novel motifs.

Motif-guide methods

CENTIPEDE [52] first scans the human genome for candidate-binding sites based on position weight matrices (PWMs), and then applies a hierarchical mixture model considering the PWM score, the conservation score and the distance to the nearest TSS of the sites. Multinomial distribution is used to model cleavage profiles, and a posterior probability is calculated to determine the binding status for a site. In this study, 10-mer words enriched in the most DNase I-sensitive regions were also used instead of PWMs to locate candidate sites and infer novel motifs. A supervised version of CENTIPEDE with less parameters was developed by Luo *et al.* [53], and another extension of CENTIPEDE accounting for different binding sites and replicates was provided by Raj *et al.* [54]. PIQ [55] and other methods use similar strategy to find candidate sites, and the per-base signals are modeled as a Gaussian process. In PIQ, expectation propagation is applied to obtain the final result for TF-binding estimation. Footprint likelihood ratio (FLR) [56] builds a two-component mixture model consisting of footprint and background parameters, and after training by expectation–maximization (EM) algorithm, the log-transformed footprint-background likelihood ratio for each site is then calculated. BinDNase [57] models DNase-seq signal by standard logistic regression and uses a greedy backward machine learning that extracts features from DNase-seq data for each TF. ChIP-seq peaks and hotspots from ENCODE were obtained to train the model. Similarly, DeFCoM [58] is an SVM-based supervised method that extracts features from windows with decreasing sizes toward the motif center. DeFCoM selects between two different SVM models by bootstrapping in the training process. Another EM-based approach, Romulus [59], takes into account the different binding modes for a single TF and provides a special statistic that may quantify pioneer factor activity.

De novo methods

Right after the first digital genomic footprinting work [50], the same group developed a new approach based on dynamic Bayesian network [60], which is the combination of hidden Markov model (HMM) and Bayesian network. Neph *et al.* [61] presented a footprinting method based on a simple formula to quantify the relative accessibility of interested region. The formula for footprint occupancy score (FOS) is written as $FOS = (C+1)/L + (C+1)/R$, where C represents the signal intensity of interested region, and L/R represents the signal intensity of the left/right flanking region. Protein occupancy may lead to depleted cleavage at the core motif (C in the formula); therefore, the regions with lower FOS are identified as footprints. Wellington [62] applies a similar strategy with only one length parameter representing flanking regions. A binomial cumulative distribution is used to calculate the P-value determining the footprints. A subsequent version, Wellington-bootstrap [63], is designed for pairwise analysis of DNase-seq data sets. DNase2TF [64] is a bit like motif-guide methods, it limits the search space in calculated hotspot regions instead of motif-predicted sites. The seed regions within hotspots are then selected based on local low signal, and the z-score for each seed is calculated to detect footprints. HINT [65] applies an eight-state HMM consisting of the background state, three DNase-states, three histone-states and the footprint state. Two modified versions of HINT are presented in the study of Gusmao *et al.* [66], with less states and even better performance. Based on an evaluation by Gusmao *et al.*, HINT (and its modified versions), DNase2TF, PIQ, Wellington and FOS were top-ranked tools, among which PIQ is the only motif-guided method.

Similar to peak-calling methods, footprints identified by different tools also vary. A justified ensemble method combining different results may strengthen the specificity or coverage, depending on the needs. To improve the confidence of results, TF-binding sites inferred by ChIP-seq are commonly used as reference. However, ChIP-seq is limited to a specific TF and heavily dependent on the quality of available antibodies, making it not fully qualified as a gold-standard benchmark. Besides, DNase-seq has an incontrovertible advantage over ChIP-seq with better resolution at the footprinting level. A protein-binding site nestles in the middle region of a fragment produced by ChIP-seq, and the exact boundary is unknown, leading to a limited resolution depending on the algorithm. This is an innate advantage for DNase-seq, as the cleavage site could be immediately flanking the binding sites [12], making it easier to have a single-nucleotide resolution for footprints. From this perspective, ChIP-exo [67] could work as a better benchmark than ChIP-seq, as the flanking regions of binding sites are digested by exonuclease.

Still, the DNase-seq footprinting approach is also facing several problems. The DNase-seq experiment can be seen as a capture of dynamic activity of cells, and the footprints reflect an average pattern of the cells with different states of nucleosome breathing [68] (the millisecond-scale wrapping and unwrapping of nucleosomal DNA). The influence of cell-state-specific or infrequent binding events can be dampened by a majority of cells, leaving shallow or no footprints. Sung *et al.* [69] discussed a same situation for TFs with short residence and also brought up several other challenges. For example, distinct TFs may have similar binding motifs, making it challenging to assign a TF to the motif shared by two or more TFs.

Common applications and data integration

DHS calling and footprinting have become regular practices for DNase-seq studies; yet, the potential of DNase-seq data is not fully explored. Based on peak calling, He *et al.* [70] performed a differential DNase I hypersensitivity analysis using cancer cells with different conditions, providing a quantitative measure to predict cell-type-specific TF-binding patterns. Based on DNase I footprinting, Neph *et al.* [71] provided an approach that

determines TF occupancy within the promoter proximal regions of TF genes, leading to a hierarchical TF regulatory network. The significance of this approach is evident from the similarity between *de novo*-derived subnetworks and experimentally annotated subnetworks. Emerging new approaches that solely depend on DNase-seq data may provide insights into future directions of DNase-seq data application.

However, biological systems contain complex, dynamic, interactive structures, and a single-layer analysis cannot cover a complete system. To prevent one-sided understanding of mechanisms and biased models, other types of data are often needed for complementary. Integrative analysis has been widely accepted as an effective approach to address questions in complex biological systems. While most data types are connected as components in the evolved central dogma, data for integration should be biologically relevant to avoid producing incorrect mechanisms and misleading results. Also, on-demand integration for specific aims may need matched samples or modified algorithms.

As another important vehicle in cistrome studies, ChIP-seq acts as the most common partner of DNase-seq. For many studies, a DNase-peak or footprint is considered reliable if it is in the vicinity of a ChIP-seq peak, and unrecognized *de novo* motifs may be ascertained by ChIP-seq peaks of a specific factor. Transcriptional activation or regulation requires the repositioning of nucleosomes; therefore, gene expression can also be correlated to DHSs. For example, Natarajan *et al.* [72] combined DNase-seq data with matched gene expression data for prediction of cell-specific gene expression, and He et al. [73] integrated DHS signal, RNA-seq and genomic sequences for predicting enhancer–promoter interactions. Similar to open chromatin status, epigenetic modifications may also indicate transcriptional activity. Activation-associated histone mark H3K4me3 regularly coincide with DHSs, while repression-associated mark H3K27me3 is always negatively correlated with DNase-seq signals [74]. Lazarovici *et al.* [75] showed the potential of DNase-seq in predicting DNA methylation status. In addition to these factors, genetic variants in DHSs are also a popular topic, as the change of sequences may influence the binding of functional proteins. Perera *et al.* [76] examined the promoter DHS regions of 1161 samples across 14 cancer types, found that mutated DHSs are associated with transcription initiation activity and contributed to the alteration of nucleotide excision repair processes. Moyerbrailean *et al.* [77] studied the SNPs in DHS regions and footprints, obtained the finding that most SNPs in regulatory regions are silent and identified functional SNPs in known GWAS hit regions. While not directly related, DNase-seq has the potential to complement chromatin conformation data. DNase-seq may be applied to study the interactions of regulatory elements [78] at topologically associated domains or help estimate A/B compartments [79]. Both Hi-C techniques and DNase-seq are used to measure spatial organization of the genome, but at different scales [80]. Their integration may provide high-resolution chromatin structural information, as well as in-depth characterization of interchromosomal/long-distance DNA interactions.

In the pursuit for a complete understanding of cell biology, molecular biologists have made unremitting efforts to unravel the complex regulatory mechanisms. In the past decades, DNase-seq-related studies have made significant contributions to the area of genomics and epigenomics. Researchers have applied DNase-seq to study the profile of functional regulatory elements, elucidate chromatin remodeling activities and obtain high-resolution chromatin structures. An accessible chromatin landscape involving 2.9 million human DHSs across 125 diverse cell and tissue types was provided by Thurman *et al.* [28, 29], and the relationships of DHSs and several factors, including genetic variation and epigenetic modifications, were explored. In this article, over 110 000 promoters were predicted by integrating H3K4me3 signals and DHSs, and a 32-element tag-density vector (generated by hierarchically clustering of 79 cell types) was assigned to each DHS. Approximately 580 000 distal DHSs were then connected to their target promoters according to their correlation coefficients. Based on DNase-seq, a more recent study [81] established the chromatin accessibility landscape during human early embryogenesis. Embryo cells at several early stages were collected to study the dynamics of chromatin reorganization during preimplantation development. The gradual increase in numbers of DHSs suggested a progressive increase of chromatin accessibility, and the enrichment of binding motifs in DHSs indicated that OCT4 contributes to human zygotic genome activation. The DHS signal was found to be negatively correlated with DNA methylation levels, which is consistent with other studies [82]. These works may serve as great inspirations helping people to unfold new secrets of cells.

Visualization

Visualization is an important practice to convert complex data into an intuitive representation. For DNase-seq studies, providing graphical examples with recognizable inherent patterns of peaks/footprints/motifs will enable a quick understanding of the corresponding issues. The linear representations of interested genomic regions can be done by genome browsers such as UCSC genome browser [83], Integrated Genome Browser (IGB) [84] and Integrative Genomics Viewer (IGV) [85]. These genome browsers can handle all well-known formats for genomic data at different levels and support the loading of custom tracks. Sometimes, linear layout might be less convenient when one wants to keep the genome context and include dispersed data like relationships between genomic positions. Circos [86] is a good tool for visualization of DNase-seq data in circular layout. The genomes can be basically represented by circles, the peaks or footprints can be represented as segments on the circles and the interactions can be shown as ribbons inside the circles. Like the different tracks for genome browsers, Circos supports for multiple different genomes and data types (in the form of concentric circles), but the users have to manually create input files to fit the standards of Circos.

In the absence of the demonstration of genomes, visualization can be in various forms with great flexibility, and extensive tools can be used for plotting and further modification. The bottom line is to clearly illustrate the data with a visible conclusion. A famous counterexample is the 'hairball' network, which can be avoided by edge bundling, node clustering or the use of specific network visualization tools like HivePlots [87]. Moreover, format conversion tools like BEDOPS [88] and UCSC toolkits [89] could be useful as some visualization tools support only the specific formats.

Discussion

DHS profiling is recognized as an effective instrument for the dissection of transcriptional activities. The technique of DHS profiling evolves in tune with the advances in genomic sequencing techniques. DNase I digestion followed by microarray hybridization (DNase-chip/DNase-array) was developed to replace traditional Southern Blot methods but was soon

superseded by DNase-seq. With high coverage and accuracy, DNase-seq will be a major method characterizing chromosome accessibility in years. With the rapid innovation in experimental techniques, variation of the protocols and the understanding of new data, new best practices have been emerging, and the commented pipeline in this review may shorten the preparation time for inexperienced researchers.

Similar to DNase-seq, some other techniques like Assay for Transposase-Accessible Chromatin with high throughput Sequencing (ATAC-seq) [90], MNase-seq [91] and Formaldehyde-Assisted Isolation of Regulatory Elements Sequencing (FAIRE-seq) [92] are also capable of dissecting chromatin accessibility. MNase-seq uses micrococcal nuclease (MNase) to digest chromosome-free regions of DNA, and the isolated fragments for sequencing determine the positions of nucleosomes. The signal profile of MNase-seq is complementary to DNase-seq profiles. MNase-seq typically requires at least 10 million cells and may show considerable variation of signals between technical preparations [93]. FAIRE-seq uses formaldehyde to cross-link chromatin and sonication to shear chromosomal DNA into fragments. Compared with other techniques, FAIRE-seq has lower signal-to-noise ratio, and the experiment heavily depends on fixation efficiency. Compared with the two assays, DNase-seq could achieve better performance in identifying regulatory elements. ATAC-seq is the most promising and appealing alternative to DNase-seq, with easier experimental preparation and much smaller required amount of cells. ATAC-seq uses Tn5 transposase to cut DNA and insert primer DNA sequences. Although there are several disadvantages to this approach [94], like serious mitochondrial DNA contamination, emerging protocols [95, 96] have addressed some of these issues. Nonetheless, DNase-seq has some specific advantages, for example, the preference of DNase I to cut at the minor groove helps better distinguish rotational positioning of nucleosomes [4]. Tsompana *et al.* [97] provided a comparison of these techniques. Meyer *et al.* [14] discussed the intrinsic biases and technical biases for DNase-seq, ChIP-seq, ATAC-seq, MNase-seq and FAIRE-seq. According to the different nature of the enzymes and the different experimental procedures, the pipeline provided in this article is not directly applicable for another type of data. However, the major data processing steps could be similar even for less similar techniques like CLIP-seq [98], ChIP-exo [67], Sono-seq [99] and RIP-seq [100], and researchers can still refer to the guideline in this study while flexibly alter the tools and parameters to get better performance.

Like other techniques, biases for DNase-seq research may arise from a number of sources, and most of them will be reduced by introducing other information. However, one always need to compromise between accuracy and other factors like time and money. For example, the genome used for DNase-seq could have differences from the reference genome, causing biases in the mapping step. The best solution is to perform another whole-genome sequencing for this sample, which is unworthy for most cases. Preferred second-tier solution is to configure the options of BWA or Bowtie2 to allow mismatches. Insignificant biases including the nicking preference of sequences are discounted for most studies. As an example, DNase I may generate 2–4 bp overhang [101] cutting DNA, which is abandoned the blunt-ended procedure, leading to a slight bias for mapping. The improvements of sequencing technologies may also help DNase-seq to overcome some of the limitations. For example, the influence of mixed cells can be avoided by single-cell DNase-seq [15]; long-read sequencing will help further increase the mapping rate, while 36-bp illumina reads are already considered sufficient. However, some of the limitations will not be managed in the near future, like mapping reads to signal artifact regions [102].

With the cost dropping way faster than Moore's law, the NGS techniques allow an enormous increase in the amount of sequence data [103]. Large-scale projects like ENCODE, BLUEPRINT [104] and Roadmap [93] are providing large, publicly available data sets, while more and more DNase-seq data are continuously depositing into the Sequence Read Archive (SRA) [105] or Gene Expression Omnibus (GEO) [106]. Mei *et al.* [107] provided a data browser for public DNase-seq data. For researchers that cannot perform new experiments, these data sources may serve as great assistance. The integration of 'multi-omics' data will eventually form a full chain in dissecting the cellular machinery, and methods like DNase-seq will be an important link for the chromatin part.

Key Points

- DNase-seq is a powerful tool to study chromatin accessibility on a large scale. However, there is wide variation in the applications of DNase-seq, and the optimal workflow could be different for every scenario.
- A modularized DNase-seq data processing pipeline is proposed in this work. Detailed discussion was made on each major step to promote the efficiency and accuracy of DNase-seq studies.
- This article provides a comprehensive review of current computational approaches for DNase-seq analyses.

Funding

This work was supported by the National Key Research and Development Program of China (grant number 2016YFA0501704), National Natural Sciences Foundation of China (grant numbers 31571366 and 31771477), Jiangsu Collaborative Innovation Center for Modern Crop Production and the Fundamental Research Funds for the Central Universities.

References

1. Weintraub H, Groudine M. Chromosomal subunits in active genes have an altered conformation. *Science* 1976;**193**(4256):848–56.
2. Elgin SC. DNAase I-hypersensitive sites of chromatin. *Cell* 1981;**27**(3 Pt 2):413–15.
3. Boyle AP, Davis S, Shulha HP. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008;**132**(2):311–22.
4. Zhong J, Luo K, Winter PS, *et al.* Mapping nucleosome positions using DNase-seq. *Genome Res* 2016;**26**(3):351–64.
5. Winter DR, Song L, Mukherjee S, *et al.* DNase-seq predicts regions of rotational nucleosome stability across diverse human cell types. *Genome Res* 2013;**23**(7):1118–29.
6. Degner JF, Pai AA, Pique-Regi R, *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 2012;**482**(7385):390–4.
7. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010;**2010**(2):pdb.prot5384.
8. Sabo PJ, Kuehn MS, Thurman R, *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* 2006;**3**(7):511–18.

9. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**(7414):57–74.
10. He HH, Meyer CA, Hu SS, *et al*. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* 2014;**11**(1):73–8.
11. McArthur M, Gerum S, Stamatoyannopoulos G. Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse beta-globin LCR. *J Mol Biol* 2001;**313**(1):27–34.
12. Zeng W, Mortazavi A. Technical considerations for functional sequencing assays. *Nat Immunol* 2012;**13**(9):802–7.
13. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
14. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* 2014;**15**(11):709–21.
15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15):2114–20.
16. Del Fabbro C, Scalabrin S, Morgante M, *et al*. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* 2013;**8**(12):e85024.
17. Bushnell B. A suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. 2015. https://sourceforge.net/projects/bbmap/
18. Chen S, Huang T, Zhou Y, *et al*. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 2017;**18**(S3):80.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
21. Fennell T, Wysoker A, Tibbetts K. A set of command line tools for manipulating high-throughput sequencing data. 2013. https://broadinstitute.github.io/picard/
22. Lassmann T, Hayashizaki Y, Daub CO. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* 2011;**27**(1):130–1.
23. Marinov GK, Kundaje A, Park PJ, *et al*. Large-scale quality analysis of published ChIP-seq data. *G3* 2014;**4**(2):209–23.
24. Qin Q, Mei S, Wu Q, *et al*. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics* 2016;**17**(1):404.
25. Hoffman MM, Ernst J, Wilder SP, *et al*. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 2013;**41**(2):827–41.
26. Boyle AP, Guinney J, Crawford GE, *et al*. F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* 2008;**24**(21):2537–8.
27. Zhang Y, Liu T, Meyer CA, *et al*. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**(9):R137.
28. Thurman RE, Rynes E, Humbert R, *et al*. The accessible chromatin landscape of the human genome. *Nature* 2012;**489**(7414):75–82.
29. John S, Sabo PJ, Thurman RE, *et al*. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 2011;**43**(3):264–8.
30. Baek S, Sung MH, Hager GL. Quantitative analysis of genome-wide chromatin remodeling. *Methods Mol Biol* 2012;**833**:433–41.
31. Rashid NU, Giresi PG, Ibrahim JG, *et al*. 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.
32. Koohy H, Down TA, Spivakov M, *et al*. A comparison of peak callers used for DNase-seq data. *PLoS One* 2014;**9**(5):e96303.
33. Kumar V, Muratani M, Rayan NA, *et al*. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol* 2013;**31**(7):615–22.
34. Thomas R, Thomas S, Holloway AK, *et al*. Features that define the best ChIP-seq peak calling algorithms. *Brief Bioinform* 2017;**18**(3):441–50.
35. Heinz S, Benner C, Spann N, *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.
36. Liang K, Keleş S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* 2012;**13**:199.
37. Hsu F, Kent WJ, Clawson H, *et al*. The UCSC known genes. *Bioinformatics* 2006;**22**(9):1036–46.
38. Flicek P, Amode MR, Barrell D, *et al*. Ensembl 2014. *Nucleic Acids Res* 2014;**42(Database issue)**:D749–55.
39. Harrow J, Frankish A, Gonzalez JM, *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;**22**(9):1760–74.
40. Pruitt KD, Brown GR, Hiatt SM, *et al*. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;**42(Database issue)**:D756–63.
41. Frankish A, Uszczynska B, Ritchie GR, *et al*. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* 2015;**16(Suppl 8)**:S2.
42. McLean CY, Bristor D, Hiller M, *et al*. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010;**28**(5):495–501.
43. Liberzon A. A description of the Molecular Signatures Database (MSigDB) web site. *Methods Mol Biol* 2014;**1150**:153–60.
44. Zhu LJ, Gazin C, Lawson ND, *et al*. ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 2010;**11**:237.
45. Shin H, Liu T, Manrai AK, *et al*. CEAS: cis-regulatory element annotation system. *Bioinformatics* 2009;**25**(19):2605–6.
46. Kondili M, Fust A, Preussner J, *et al*. UROPA: a tool for Universal RObust Peak Annotation. *Sci Rep* 2017;**7**(1):2593.
47. Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**(5):511–15.
48. Quinlan AR. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr Protoc Bioinformatics* 2014;**47**(1):11.12.1–34.
49. Kaplan N, Moore IK, Fondufe-Mittendorf Y, *et al*. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 2009;**458**(7236):362–6.
50. Hesselberth JR, Chen X, Zhang Z, *et al*. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 2009;**6**(4):283–9.
51. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37(Web Server issue)**:W202–8.
52. Pique-Regi R, Degner JF, Pai AA, *et al*. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 2011;**21**(3):447–55.
53. Luo K, Hartemink AJ. Using DNase digestion data to accurately identify transcription factor binding sites. *Pac Symp Biocomput* 2013;**80**:80–91.

54. Raj A, Shim H, Gilad Y, *et al*. msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLoS One* 2015;**10**(9):e0138030.
55. Sherwood RI, Hashimoto T, O'Donnell CW, *et al*. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 2014;**32**(2):171–8.
56. Yardimci GG, Frank CL, Crawford GE, *et al*. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res* 2014;**42**(19):11865–78.
57. Kahara J, Lahdesmaki H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics* 2015;**31**(17): 2852–9.
58. Quach B, Furey TS. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics* 2017;**33**:956–63.
59. Jankowski A, Tiuryn J, Prabhakar S. Romulus: robust multistate identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* 2016;**32**(16):2419–26.
60. Chen X, Hoffman MM, Bilmes JA, *et al*. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics* 2010; **26**(12):i334–42.
61. Neph S, Vierstra J, Stergachis AB, *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012;**489**(7414):83–90.
62. Piper J, Elze MC, Cauchy P, *et al*. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res* 2013;**41**(21):e201.
63. Piper J, Assi SA, Cauchy P, *et al*. Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors. *BMC Genomics* 2015;**16**(1):1000.
64. Sung MH, Guertin MJ, Baek S, *et al*. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* 2014;**56**(2):275–85.
65. Gusmao EG, Dieterich C, Zenke M, *et al*. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* 2014;**30**(22):3143–51.
66. Gusmao EG, Allhoff M, Zenke M, *et al*. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Methods* 2016;**13**(4):303–9.
67. Rhee HS, Pugh BF. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* 2012;**Chapter 21**: Unit 21.24.
68. Zlatanova J, Bishop TC, Victor JM, *et al*. The nucleosome family: dynamic and growing. *Structure* 2009;**17**(2):160–71.
69. Sung MH, Baek S, Hager GL. Genome-wide footprinting: ready for prime time? *Nat Methods* 2016;**13**(3):222–8.
70. He HH, Meyer CA, Chen MW, *et al*. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res* 2012;**22**(6):1015–25.
71. Neph S, Stergachis AB, Reynolds A, *et al*. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 2012;**150**(6):1274–86.
72. Natarajan A, Yardimci GG, Sheffield NC, *et al*. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* 2012;**22**(9):1711–22.
73. He B, Chen C, Teng L, *et al*. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci USA* 2014;**111**(21):E2191–9.
74. Shu W, Chen H, Bo X, *et al*. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res* 2011;**39**(17):7428–43.
75. Lazarovici A, Zhou T, Shafer A, *et al*. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci USA* 2013;**110**(16):6376–81.
76. Perera D, Poulos RC, Shah A, *et al*. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* 2016;**532**(7598):259–63.
77. Moyerbrailean GA, Kalita CA, Harvey CT, *et al*. Which genetics variants in DNase-seq footprints are more likely to alter binding? *PLoS Genet* 2016;**12**(2):e1005875.
78. Lan X, Witt H, Katsumura K, *et al*. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Res* 2012;**40**(16):7690–704.
79. Fortin JP, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 2015;**16**:180.
80. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* 2014;**14**(6): 762–75.
81. Gao L, Wu K, Liu Z, *et al*. Chromatin accessibility landscape in human early embryos and its association with evolution. *Cell* 2018;**173**(1):248–59.e15.
82. Wu J, Xu J, Liu B, *et al*. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* 2018;**557**(7704):256–60.
83. Raney BJ, Dreszer TR, Barber GP, *et al*. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC genome browser. *Bioinformatics* 2014;**30**(7):1003–5.
84. Nicol JW, Helt GA, Blanchard SG, Jr, *et al*. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 2009; **25**(20):2730–1.
85. Robinson JT, Thorvaldsdottir H, Winckler W, *et al*. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**(1):24–6.
86. Krzywinski M, Schein J, Birol I, *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;**19**(9): 1639–45.
87. Krzywinski M, Birol I, Jones SJ, *et al*. Hive plots–rational approach to visualizing networks. *Brief Bioinform* 2012;**13**(5): 627–44.
88. Neph S, Kuehn MS, Reynolds AP, *et al*. BEDOPS: high-performance genomic feature operations. *Bioinformatics* 2012;**28**(14):1919–20.
89. Kent WJ, Zweig AS, Barber G, *et al*. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 2010;**26**(17):2204–7.
90. Buenrostro JD, Wu B, Chang HY, *et al*. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;**109**:21.29.1–29.
91. Schones DE, Cui K, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008; **132**(5):887–98.
92. Simon JM, Giresi PG, Davis IJ, *et al*. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 2012;**7**(2):256–67.
93. Rizzo JM, Bard JE, Buck MJ. Standardized collection of MNase-seq experiments enables unbiased dataset comparisons. *BMC Mol Biol* 2012;**13**:15.
94. Sos BC, Fung HL, Gao DR, *et al*. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol* 2016;**17**(1):20.

95. Corces MR, Buenrostro JD, Wu B, *et al*. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* 2016;**48**(10):1193–203.
96. Corces MR, Trevino AE, Hamilton EG, *et al*. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* 2017;**14**(10):959–62.
97. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014;**7**(1):33.
98. Licatalosi DD, Mele A, Fak JJ, *et al*. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008;**456**(7221):464–9.
99. Auerbach RK, Euskirchen G, Rozowsky J, *et al*. Mapping accessible chromatin regions using Sono-seq. *Proc Natl Acad Sci USA* 2009;**106**(35):14926–31.
100. Zhao J, Ohsumi TK, Kung JT, *et al*. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* 2010;**40**(6):939–53.
101. Sollner-Webb B, Melchior W, Jr, Felsenfeld G. DNAase I, DNAase II and staphylococcal nuclease cut at different, yet symmetrically located, sites in the nucleosome core. *Cell* 1978;**14**(3):611–27.
102. Kundaje A. A comprehensive collection of signal artifact blacklist regions in the human genome. 2013. https://sites.google.com/site/anshulkundaje/projects/blacklists
103. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51.
104. Fernandez JM, de la Torre V, Richardson D, *et al*. The BLUEPRINT data analysis portal. *Cell Syst* 2016;**3**(5):491–5.
105. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res* 2011;**39(Database issue)**: D19–21.
106. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013;**41**:D991–5.
107. Mei S, Qin Q, Wu Q, *et al*. Cistrome data browser: a data portal for ChIP-seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* 2017;**45**(D1):D658–62.