Computational Molecular Biology
2012, Vol.2, No.2, 1-13 http://cmb.sophiapublisher.com

Research Report Open Access

High-Throughput Computational Approaches to Analyzing Histone Modification Next-Generation Sequencing Data

Jie Lv[1], Hongbo[1], Qiong Wu[1], Y. Zhang[2]

1. School of Life Science and Technology, State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin, 150001, China

2. College of bioinformatics science and technology, Harbin Medical University, China

Corresponding Author email: kigo@hit.edu.cn, yanyou1225@gmail.com; Author

Computational Molecular Biology, 2012, Vol.2, No.2 doi: 10.5376/cmb.2012.02.0002

Abstract Chromatin immunoprecipitation followed by sequencing (ChIP–seq) facilitates systematic analysis of chemical modifications of histone tails. As the cost of next-generation sequencing continues to drop, genome-wide histone modification sequencing becomes a common approach in a variety of research in the epigenetic area. However, challenges of efficient ChIP–seq data analysis are now the main hurdle to interpreting the histone modification ChIP-seq data, calling for continued enhancements of computational approaches. Here we provide a pragmatic overview of available computational approaches for the study of histone modification ChIP-seq data. We present the latest advances of computational methods for systematically detecting and functionally characterizing various types of histone modification ChIP-seq data, discuss the software packages currently available for performing tasks from short read mapping, peak calling to downstream genomic characterization and genome-wide visualization. We also present that the regulatory roles of histone modifications upon gene expression can be inferred by developing algorithms and methods specifically for histone modification ChIP-seq data. Such approaches will facilitate the epigenetic regulatory network construction and provide explicit biological hypothesis for further experiment testing. We also describe some challenges and important directions for histone modification analysis based on ChIP-seq data in the future. We envision that the advances of computational approaches will bring about a brighter future for large-scale histone modification studies.

Keywords Next-generation sequencing; Histone modification; Computational approaches; Peak calling; ChIP sequencing

Background

The nucleosome is the basic unit of chromatin, which includes two copies of each of core histones (H3, H4, H2A and H2B) and 147 bp of DNA wrapped around (McGhee and Felsenfeld, 1980). Histones are evolutionarily conserved proteins with accessible and highly dynamic amino-terminal tails and also bear a histone fold domain that mediates histone-histone interactions. The N-termini of histone tails are extensively modified by up to hundreds of different post-translational modifications including methylation, acetylation and phosphorylation (Kouzarides, 2007). Until now, the biological meanings of most of these covalent modifications was not understood though significant progresses in recent years indicate that methylation and acetylation play important roles in transcriptional regulation. To systematically study the genome-wide patterns of various histone modifications, chromatin immunoprecipitation (ChIP) is usually used to collect DNA fragments isolated from chromatin using antibodies for histone modifications of interest (Collas, 2010). The isolated DNA fragments are followed by hybridization to DNA microarrays or sequencing (Gilchrist et al., 2009).

Understanding the mechanisms of histone modifications in development and disease is of great interest (Kurdistani, 2011; Ikegami et al., 2009; Aoki and Akiyama, 2007). The availability of reference genome sequences and next

Preferred citation for this article:

Lv et al., 2012, High-Throughput Computational Approaches to Analyzing Histone Modification Next-Generation Sequencing Data, Computational Molecular Biology, Vol.2, No.2 1-13 (doi: 10.5376/cmb.2012.02.0002)

Received: 19 Mar., 2012 | Accepted: 19 Jun., 2012 | Published: 02 Jul., 2011

generation sequencing platforms called for approaches to effectively explain high-throughput genome-wide histone modification data (Pepke et al., 2009; Mardis, 2007). In this review, we will describe computational approaches that can analyze the histone modification data produced by next generation sequencing platforms in terms of principle and advantages. We will also illustrate several studies for computationally inferring regulatory roles of histone modifications upon gene expression based on algorithms and methods specifically for histone modification ChIP-seq data. Before we discuss the computational approaches, we will firstly illustrate the history of ChIP-based technologies.

1 Next-Generation Sequencing Techniques in Analyzing Histone Modifications

ChIP is a fundamental approach that has been around for a while (Collas, 2010; Collas, 2009). In brief, DNA is covalently cross-linked to bound proteins. Then, the cross-linked DNA is broken into short fragments. Antibody for specific histone modification of interest is then used to isolate bound DNA. The most prominent problem with the approach is that only individual sites of interest can be studied each time.

The problem is partially tackled in ChIP-chip, one of the earlier methods to study DNA binding proteins on the whole genome. ChIP-chip (Buck and Lieb, 2004; Horak and Snyder, 2002) is a technique which involves immunoprecipitation of DNA using histone modification specific antibody followed by a DNA hybridization array (chip). Though similar to Chromatin immunoprecipitation followed by high-throughput ChIP sequencing (ChIP-seq) in name, its mapping precision is lower than ChIP-seq, and the dynamic range of quantified expression is significantly less than ChIP-seq (Liu et al., 2010). Moreover, all hybridization approaches including ChIP-chip mask repetitive sequences. Though not efficient for genome-wide histone modification studies, ChIP-chip with custom arrays for specific genes or loci is still useful for studies with many experimental conditions.

Later, a high-throughput approach based on ChIP is known as ChIP-SAGE (Schones et al., 2011; Schones and Zhao, 2008). In short, ChIP is carried out and followed by SAGE (Serial Analysis of Gene Expression). More close to ChIP-seq, short sequence tags of 21 bp are extracted from the sequencing library and mapped to a reference genome. The number of tags that are mapped on a genomic region reflects the histone modification level of that region. Since there is no probe hybridization issue involved in the technology, the results of ChIP-SAGE tend to be more quantitative than ChIP-chip, though no direct comparisons of the two techniques are made. However, few studies use ChIP-SAGE for the apparent limitations of the technology and also the advent of a quite more cost-effective yet more sensitive alternative technology, that is, ChIP-seq (Park, 2009).

Today, ChIP-seq is generally the preferred method for studying genome-wide histone modification patterns, which allows tens of millions of DNA targeted by histone modifications to be sequenced in an acceptable time period. ChIP-seq is proven to have low error rates, high specificity and high sensitivity, while keeping cost per library acceptable for researchers (Johnson et al., 2007). Different to ChIP-chip, ChIP-seq completely eliminates potential errors of cross-hybridization. The dominating service provider of ChIP-Seq is illumina, using a high-throughput massively parallel signature sequencing-like technique developed by Solexa (Cuddapah et al., 2009; Whiteford et al., 2009). Briefly, the ChIP DNA is ligated to adaptors followed by limited amplification to generate ~200 ng of DNA that is then bound by hybridization on a solid surface. A short sequence (25~50 bp) for 30~60 million DNA templates is then sequenced from sequence end by 'sequencing-by-synthesis', a modified Sanger sequencing procedure. ChIP-Seq was initially done in CD_4^+ T cells to investigate genome-wide histone modifications (Barski et al., 2007). Conceptually, the number of sequenced reads mapped to a genomic locus is proportional to its histone modification level. Two important merits of ChIP-Seq included less need for PCR amplifications and independence of probe hybridization, making it probably more quantitative and comparable for different genomic regions (Johnson et al., 2007). An additional concern for next generation sequencing based histone modification

profiling is how deeply to sample each library (Liu et al., 2010). Though sequencing with a large scale over saturation which means that further sequencing would fail to discover additional regions above background provides a full coverage and improves confidence of histone modification of interest, sequencing below or up to saturation may be sufficient to keep the sequencing cost acceptable while not significantly decreasing coverage. Refer to Table 1 for a comparison of ChIP-chip, ChIP-SAGE and ChIP-Seq techniques.

Table 1 Comparison of ChIP–chip, ChIP–SAGE and ChIP–Seq

	ChIP-chip	ChIP-SAGE	Chip-seq
Quantification	Limited quantitative and depends on the hybridization efficiency	Quantitative	Quantitative
Resolution	Depends on size of the chromatin fragments for ChIP	Depends on restriction enzyme sites	Depends on the size of the chromatin fragments and sequencing depth
Cost	High for whole-genome tiling arrays	More expensive than ChIP–Seq	Low
Limitation	Only pre-selected genomic regions on a microarray	Recognition sites for the restriction enzyme	Only non-repetitive regions

2 Large Data Resources of ChIP-seq Histone Modifications

Hundreds of ChIP–seq experiments were carried out by the Encyclopedia of DNA Elements (ENCODE) Consortium, which is a valuable data source and provide effective sequencing protocols (Birney et al., 2007). Considering the diversity of cell types being assayed and will be assayed in ENCODE, it is useful to mine knowledge of tissue-specific and/or cell-type-specific histone modification patterns for various genomic elements from the ENCODE data. However, it should be noted that the success of a ChIP experiment highly depends on highly specific antibody to the bound histone modifications (Liu et al., 2010). Antibody quality varies, even between independently prepared lots of the same antibody, as shown in a recent assessment of antibodies in the ENCODE and the model organism ENCODE (mod-ENCODE) projects (Egelhofer et al., 2011). In this study, ~25% failed in specificity tests and 20% failed in immunoprecipitation experiments. Thus, caution is needed to interpret the histone modification ChIP-seq data, especially in comparisons of different histone modification patterns.

modENCODE project was launched to provide a comprehensive encyclopedia of genomic functional elements in the model organisms such as C. elegans and D. melanogaster (Washington et al., 2011; Muers, 2011). The data contents range from gene structure, mRNA and ncRNA expression profiling to transcription factor binding sites, histone modifications and others. All the data is publicly available for download and for publication uses.

The Epigenomics resource (www.ncbi.nlm.nih.gov/epigenomics) at the National Center for Biotechnology Information (NCBI) is a comprehensive public resource for whole-genome histone modification and other epigenetic modification data sets (Fingerman et al., 2011). The data are based on epigenetic modification data from the Gene Expression Omnibus (GEO) database (Barrett and Edgar, 2006). The resource is user-friendly and continues to be updated. The Epigenomics resource is highly integrated with other NCBI databases (Baxevanis, 2008), including the Gene database (Maglott et al., 2011) and PubMed (McEntyre and Lipman, 2001) to facilitate uses. There are over 1100 data tracks encompassing five well-studied species in 2011.

Aiming to catalyze basic biology and disease-oriented research, the NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/) is another public resource of human epigenomic data (Bernstein et al., 2010). The consortium maps histone modifications and other chromatin modifications in various cell types that may represent the normal counterparts of tissues and organ systems involved in human disease. Histone modifications are also assayed by ChIP-seq, which are followed by rigorous specificity tests to ensure antibody specificity. In addition, common cell sources are collectively profiled and compared, ensuring consistency between the different data-collection centers.

3 Tools of Analysis for Next-Generation Histone Modification Data

Analysis of histone modification ChIP-seq data generated on a next generation sequencing platform remains a challenge due partly to the rapid development of many next generation sequencing platforms. Analysis of histone modification data generated by next generation sequencing can be broken into two sections.

3.1 Alignment tools for next-generation histone modification data

Data generated from a next generation sequence platform are base sequences (Illumina Genome Analyzer, 454 FLX) or color space base transitions (SOLiD) together with associated quality scores. The beginning step of analysis of ChIP-seq histone modification data is to align reads in ChIP-seq data downloaded from a public resource or obtain from a service provider to a reference genome assembly. The result of the analysis would be a data set consisting of the genomic coordinates of the read alignments and strand on the reference genome. Many next generation sequencing aligners have been developed to map sequenced reads against a reference genome (Pepke et al., 2009; Kim et al., 2011; Schones et al., 2011; Schones and Zhao, 2008; Hirst and Marra, 2010). The majority of the aligners use a ‘seed and extend’ based algorithm where a sub-string within a read is aligned to either a hash table or more recently a suffix array generated from Burrows–Wheeler transform of the reference genome. Read is ‘extended’ up to the maximum read length on the genome until a match is found. The SAM/BAM file format is a standardized file format that such aligners can output. Though these aligners have minor differences in speed and accuracy, these differences do not have significant impacts on overall mapping rates and accuracies (Wilbanks and Facciotti, 2010). End users can choose one of these aligners according to the advice from other researchers or refer to related paper. The aligned file may be viewed directly on a genome browser or further processed through peak calling. We listed many common short read aligner tools for ChIP-seq data, as shown in Table 2.

Table 2 A subset of short read aligners available for histone modification ChIP-seq alignment

Software tool	Web address
Seed and extend strategy	
MAQ	http://maq.sourceforge.net/
SOAP	http://soap.genomics.org.cn/index.html
SHRiMP	http://compbio.cs.toronto.edu/shrimp/
ZOOM	http://www.bioinfor.com/zoom
BFAST	http://sourceforge.net/projects/bfast/
Using a hash table or more recently a suffix array generated from Burrows–Wheeler transform	
BOWTIE	http://bowtie-bio.sourceforge.net
BWA	http://bio-bwa.sourceforge.net
SOAP2	http://soap.genomics.org.cn/index.html

3.2 Peak calling tools for next-generation histone modification data

Peak calling algorithms transform raw read alignments into peaks - regions of significant tag enrichment. Peaks are considered to be associated with histone modification occupancy, which can be modeled by many peak calling tools (for a recent review see (Pepke et al., 2009)). Some algorithms simply merge mapped tags, while others use strand specific information to find peaks more precisely. Some peak calling tools need a control sequencing ChIP-seq library while others can still work without control. Given there are several known sources of sequencing bias of ChIP-seq, peak calling results without a control library are not reliable. Confidence for mapped peaks is quantified using *p*-value or false discovery rate (FDR), based on the difference of ChIP library and control library, though different peak-calling algorithms differ a lot in details. Generally, such tools can be divided into two parts where two main strategies are used. The first strategy primarily searches for histone modification marks that tend to reach a summit in their genomic distribution such as H3K4me3 or H3K27ac and attempts to model the tag distribution of 'peaks'. Though there are a large number of peak calling software packages, not all of them can meet the need of the first strategy for calling enriched histone modification domains. The second strategy is therefore suitable for histone modifications with more broad distribution patterns such as H3K36me3, which can be detected by peak calling software designed for histone modifications. Publicly available peak-calling algorithms that are suitable for histone modification ChIP-seq data are listed in Table 2 and several related reviews are available elsewhere (Pepke et al., 2009; Wilbanks and Facciotti, 2010; Szalkowski and Schmid, 2011). Other packages not listed in the table may be involved in commercial software packages that also contain peak-calling functionality.

Zang et al. analyzed the score distribution in a genomic background model of random reads, and employed their theory to identify spatial clusters that are unlikely to appear by chance, which was implemented as a software SICER (Zang et al., 2009). Rashid et al. developed ZINBA (Zero-Inflated Negative Binomial Algorithm) to identify enriched genomic regions of ChIP-seq, which models and accounts for factors that co-vary with background or experimental signal, such as G/C content (Rashid et al., 2011). Xu et al. proposed a linear signal–noise model, where a noise rate was introduced (Xu et al., 2010). They developed an iterative algorithm to estimate the noise rate using a control library, and derived a library-swapping strategy to estimate the FDR. The algorithm was implemented as software, named CCAT (Control-based ChIP-seq Analysis Tool). Applications to H3K4me3 and H3K36me3 datasets showed that CCAT predicted significantly more ChIP-enriched sites that the previous methods did. Zhang et al. present a Perl based software Model-based Analysis of ChIP-Seq data, MACS, to analyze ChIP-seq data (Zhang et al., 2008). MACS has a nomodel parameter to support for broad distribution pattern of histone modification such as H3K36me3. Boyle et al. developed F-seq to detect open chromatin regions, which can also be used for histone modification ChIP-seq data (Boyle et al., 2008). The important parameters of these algorithms are shown in Table 3.

The diversity of such peak calling tools is a result of the rapid progress and also diversity of sequencing technology. Researchers undertaking histone modification studies based on ChIP-seq should judge which tool would be most suitable for their own data. However in near future, it is expected that such tools will be standardized in epigenomic research.

4 Differential Histone Modification Region Identification Tools for Next-Generation Histone Modification Data

Differential histone modification sites (DHMSs) are important for studying the dynamic nature of histone modification regulations among various cell types, stages or environmental responses. Though ChIP-seq are less prone to error due to relatively long read length, several procedures such as sample preparation, tags amplification and sequence alignment pose some challenges in comparing different ChIP-seq data to extract true biologically related signals (Taslim et al.,

Table 3 Important parameters for each peak calling algorithm

Algorithm	Important parameters
CCAT	Minimum score: minimum score of normalized difference
	Minimum count: minimum number of read counts at the peak
	Moving Step: step of window sliding
	SlidingWinSize: size of sliding window
	Bootstrap pass: number of passes in the bootstrapping process
MACS	NoLambda: if True, MACS will use fixed background lambda as local lambda for every peak region
	NoModel: whether or not to build the shifting model
	MFold: regions within MFOLD range of high-confidence enrichment ratio against background to build model
	PValue: *p*-value cutoff for peak detection
SICER	WindowSize: size of the windows to scan the genome width
	GapSize: allowed gap in base pairs between islands
	FDR: false discovery rate controlling significance
ZINBA	Selectmodel: Specifying select model = FALSE skips the model selection process altogether and may save a significant amount of time
	extension: average fragment library length (size selected)
	winSize: Selecting a larger window size increases speed of analysis but decreases resolution and sensitivity to detect enrichment
	offset: Smaller non-zero offset distances increase sensitivity but also increase computational builden
	FDR: FDR = TRUE specifies the model to use the FDR threshold rather than posterior probabilities. This typically results in more liberal peak calls. If false, then uses posterior probability to threshold peaks using 1-threshold.
F-seq	FeatureLength: feature length
	Threshold: standard deviations

2009). Though we would expect all differences of samples to reflect the biological conditions, there are more factors that cannot be modeled and thus may bias the results. Effective computational and statistical approaches are necessary to reliably detect differential regions from different ChIP-seq data.

Previously, Xu et al. proposed an approach ChIPDiff for the genome-wide comparison of histone modification enriched regions identified from ChIP-seq (Xu et al., 2008). They employ a hidden Markov model (HMM) to infer the states of histone modification changes at each genomic location. Huang et al. developed an effective framework to identify genome-wide differential histone modification regions (Huang et al., 2011). They implemented a software tool EpiCenter that can efficiently perform relevant data processing (Huang et al., 2011). In addition, Taslim et al. apply a two-step non-linear normalization method based on locally weighted regression (LOESS) approach to compare ChIP-seq data across multiple samples and model the difference using an Exponential-Normal mixture model (Taslim et al., 2009). Though not explicitly designed for histone modification ChIP-seq data, further study should evaluate the power of its application in histone modification ChIP-seq data.

5 Visualization Tools for Next-Generation Histone Modification Data

Many web-based and standalone tools are available for visualization of aligned epigenomic data including histone modification ChIP-seq data sets. The most widely used tool is the Genome Browser maintained

by the University of California Santa Cruz (UCSC) (Dreszer et al., 2012). A local installation of UCSC Genome Browser is favored by many researchers to visualize unpublished ChIP-seq data. Presented as linear tracks in the context of genome annotations, UCSC is one of the early visualization tools for genome-wide data and has influences over later related tools. While extremely powerful for manual genome viewing, it is hard to visualize many large ChIP-seq data simultaneously. A few genome browsers developed later tend to support more for large ChIP-seq data in BAM format, such as GBrowse, GenomeView (Abeel et al., 2012), JBrowse (Skinner et al., 2009) and ABrowse. In addition, standalone tools such as IGV and IGB are also favorable tools to view extremely large aligned ChIP-seq data. Anyone who cannot establish browser-based tools can also use such standalone tools. The useful visualization tools are shown in Table 4.

Table 4 The list of more visualization tools for histone modification ChIP-seq data

Web server / Software	Website / Download Link
UCSC Genome Browser	http://genome.ucsc.edu/
GBrowse	http://www.gbrowse.org/index.html
Ensembl	http://asia.ensembl.org/index.html
GenomeView	http://genomeview.org/
JBrowse	http://jbrowse.org/
ABrowse	http://www.abrowse.org/
Artemis	http://www.sanger.ac.uk/resources/software/artemis/
Avadis Genome Browser	http://www.avadis-ngs.com/features/genome_browser
IGV	http://www.broadinstitute.org/igv/
IGB	http://bioviz.org/igb/

6 Downstream Analysis Tools for Next-Generation Histone Modification Data

Peak calling is generally followed by downstream analyses to annotate and characterize the enriched regions by specific histone modifications. Usually, genomic annotations are also needed to discover potentially interesting associations with enriched histone modification regions. Annotations are available from many public repositories such as UCSC, Ensembl (Flicek et al., 2012) and many scattered websites. To annotate the enriched histone modification regions, it is helpful to show the genomic landscape of such regions in the context of various common genomic annotations such as chromosomes and genes to search for interesting biological associations. For example, finding a possible relationship with enhancers by comparing position of H3K27ac peaks with known enhancers and even by more advanced bioinformatics analysis. Many tools can achieve this, such as Galaxy (Goecks et al., 2010) and Bedtools (Quinlan and Hall, 2010). Some visualization tools such as CEAS (Shin et al., 2009) and ChIPseeqer (Giannopoulou and Elemento, 2011) can support displaying average histone modification enrichment signal within/near genomic elements such as enhancers and gene initiations while do not require peak calling beforehand, which helps biologists to better comprehend histone modification patterns.

In addition, it is also useful to explore enriched histone modification peaks in known genomic elements to obtain a global view of potential regulatory function and localization preference of specific histone modifications without any specific prior knowledge. Particularly, it is of interest to study the preferential target of specific histone modifications in gene structure such as exons, promoters and distal upstream regions. The two complementary approaches are both common in chromatin biology study.

Researchers can choose one or both based on their own biological hypothesis.

7 Regulatory Histone Modifications for Gene Expression

The accumulating histone modification ChIP-seq data enables researchers to carry out global chromatin knowledge mining, which is of great interest in epigenetic field. From a computational perspective, one can leverage such data to analyze interactions among histone modifications by different proposed algorithms. Yu et al. carried out a pioneering study to infer combinatorial relationships among histone modifications and other transcriptional regulators based on the associations with gene expression by a proposed Bayesian network (Wood et al., 2011). They constructed chromatin regulatory networks and inferred many chromatin interaction relationships based on a set of 23 ChIP-seq data in human CD4+ T cells, the most comprehensive histone modification data at that time. A number of further studies show even more complex correlations between histone modifications, genomic elements and gene expression. For example, Karlić et al. used linear regression model to further explore the similar question, finding that genes with promoters of high GC contents and low GC contents are regulated by different sets of histone modifications (Karlic et al., 2010). Costa et al. further applied a mixture of linear regression models on H3K4me3 and H3K27me3 and found that they were more predictive for gene expression compared to transcription factor binding (Beck et al., 2012). Rego et al. used sparse linear regression mixture models to model gene expression and performed an efficient feature selection of transcription factors (do Rego et al., 2012). Using the model, the authors therefore identified blood development related histone modifications and transcription factors (do Rego et al., 2012). Interestingly, these studies modeled gene expression by different computational models, with transformations of tag counts as input. However, other approaches used other derived features such as peak shape and location, as well as signal frequencies to model gene expression. Beck et al. proposed a new strategy that quantifies the ChIP-seq profile, making use of the pattern and location of the signal (Beck et al., 2012). Ucar et al. introduced a subspace clustering algorithm to exhaustively identify combinatorial modification patterns and also identified combinatorial histone modification signatures for different classes of functional DNA elements (Ucar et al., 2011). Altogether, algorithms and models using histone modification ChIP-seq maps of different cells and developmental stages aid in understanding how chromatin modification network regulates gene expression.

8 Conclusions and Perspectives

The importance of histone modifications has motivated the continued accumulation of ChIP-seq data to identify and characterize histone modifications and the combinatorial and regulatory roles in gene expression. ChIP-seq is in fact the standard for identifying genome-wide histone modification landscape. However, technical and computational limitations are still an obstacle to reliably deriving biological knowledge from next-generation ChIP-seq data. Here, we focus on the computational aspects of histone modification ChIP-seq data processing from raw reads to downstream analysis.

In the last five years, next generation sequencing has brought epigenetic studies to a rapid developing era. The study of histone modifications has been evolutionarily changed from methodology to biological explanation. A large number of histone modification ChIP-seq data can be downloaded from public databases, such as NCBI GEO. This development is expected to continue as third generation sequencing platforms will soon be commercialized. However, the second-generation sequencing platforms will still be prevalent for a long time. Therefore, it is necessary to grasp the basic concepts and approaches of processing the second-generation sequencing data.

It is obvious that there are many different tools to carry out same tasks due to continued development of immature next-generation sequencing technology. In near future, it is expected that the processing methods or metrics will be greatly standardized. The key to achieving this end is to study available computational metrics or develop a new metrics, similar to base

quality score used in genomic studies, for enrichment based histone modification data. If developed, a common metric would enable meaningful comparisons among different ChIP-seq experiments, which would be critical to allow for meta-analyses of these rich data sets in future.

Although histone modification data are being accumulated at an unprecedented speed, the development of more efficient computational tools that are necessary to process and integrate a large number of data has lagged a little behind. Differential histone modification identification approaches are not only useful for comparing different biological samples, but also useful for deciphering disease related histone modification patterns. In fact, recent research has proven the power of histone modification markers in diagnosis and therapy (Zhao and Zhang, 2011). To be able to discern the altered histone modification patterns in various diseases, researchers need to continue developing and comparing more powerful tools related to histone modification differential identification based on next-generation sequencing data. More web-based and stand-alone tools with better display effect and more support for a large number of data tracks are highly favorable, which will be useful to compare histone modification data of various types in developmental stages, disease types and so on. We envision that the advances of computational approaches will bring more about a bright future for large-scale histone modification studies.

Acknowledgments

The authors thank National Natural Science Foundation of China for funding. This work is supported by the National Natural Science Foundation of China [31171383, 31371334，31371478], The Fundamental Research Funds for the Central Universities [HIT.NSRIF.2010027] and Natural Science Foundation of Heilongjiang Province [C201217].

References

Abeel T., Van Parys T., Saeys Y., Galagan J., and Van De Peer Y., 2012, GenomeView: a next-generation genome browser, Nucleic Acids Res, 40: e12
http://dx.doi.org/10.1093/nar/gkr995
PMid:22102585 PMCid:PMC3258165

Aoki F., and Akiyama T., 2007, [Involvement of histone modification and histone variants replacement in genome reprogramming during oogenesis and preimplantation development], Tanpakushitsu Kakusan Koso, 52: 2170-2176
PMid:21089289

Barrett T., and Edgar R., 2006, Gene expression omnibus: microarray data storage, submission, retrieval, and analysis, Methods Enzymol, 411: 352-369
http://dx.doi.org/10.1016/S0076-6879(06)11019-8

Barski A., Cuddapah S., Cui K., Roh T.Y., Schones D.E., Wang Z., Wei G., Chepelev I., and Zhao K., 2007, High-resolution profiling of histone methylations in the human genome, Cell, 129: 823-837
http://dx.doi.org/10.1016/j.cell.2007.05.009
PMid:17512414

Baxevanis A.D., 2008, Searching NCBI databases using Entrez, Curr Protoc Bioinformatics, Chapter 1: Unit 1 3

Beck D., Brandl M.B., Boelen L., Unnikrishnan A., Pimanda J.E., and Wong J.W., 2012, Signal analysis for genome-wide maps of histone modifications measured by ChIP-seq, Bioinformatics, 28: 1062-1069
http://dx.doi.org/10.1093/bioinformatics/bts085
PMid:22345622

Bernstein B.E., Stamatoyannopoulos J.A., Costello J.F., Ren B., Milosavljevic A., Meissner A., Kellis M., Marra M.A., Beaudet A.L., Ecker J.R., Farnham P.J., Hirst M., Lander E.S., Mikkelsen T.S., and Thomson J.A., 2010, The NIH Roadmap Epigenomics Mapping Consortium, Nat Biotechnol, 28: 1045-1048
http://dx.doi.org/10.1038/nbt1010-1045
PMid:20944595 PMCid:PMC3607281

Birney E., Stamatoyannopoulos J.A., Dutta A., Guigo R., Gingeras T.R., Margulies E.H., Weng Z., Snyder M., Dermitzakis E.T., Thurman R.E., Kuehn M.S., Taylor C.M., Neph S., Koch C.M., Asthana S., Malhotra A., Adzhubei I., Greenbaum J.A., Andrews R.M., Flicek P., Boyle P.J., Cao H., Carter N.P., Clelland G.K., Davis S., Day N., Dhami P., Dillon S.C., Dorschner M.O., Fiegler H., Giresi P.G., Goldy J., Hawrylycz M., Haydock A., Humbert R., James K.D., Johnson B.E., Johnson E.M., Frum T.T., Rosenzweig E.R., Karnani N., Lee K., Lefebvre G.C., Navas P.A., Neri F., Parker S.C., Sabo P.J., Sandstrom R., Shafer A., Vetrie D., Weaver M., Wilcox S., Yu M., Collins F.S., Dekker J., Lieb J.D., Tullius T.D., Crawford G.E., Sunyaev S., Noble W.S., Dunham I., Denoeud F., Reymond A., Kapranov P., Rozowsky J., Zheng D., Castelo R., Frankish A., Harrow J., Ghosh S., Sandelin A., Hofacker I.L., Baertsch R., Keefe D., Dike S., Cheng J., Hirsch H.A., Sekinger E.A., Lagarde J., Abril J.F., Shahab A., Flamm C., Fried C., Hackermuller J.,

Hertel J., Lindemeyer M., Missal K., Tanzer A., Washietl S., Korbel J., Emanuelsson O., Pedersen J.S., Holroyd N., Taylor R., Swarbreck D., Matthews N., Dickson M.C., Thomas D.J., Weirauch M.T., Gilbert J., et al., 2007, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature, 447: 799-816
http://dx.doi.org/10.1038/nature05874
PMid:17571346

Boyle A.P., Guinney J., Crawford G.E., and Furey T.S., 2008, F-Seq: a feature density estimator for high-throughput sequence tags, Bioinformatics, 24: 2537-2538
http://dx.doi.org/10.1093/bioinformatics/btn480
PMid:18784119 PMCid:PMC2732284

Buck M.J., and Lieb J.D., 2004, ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, Genomics, 83: 349-360
http://dx.doi.org/10.1016/j.ygeno.2003.11.004
PMid:14986705

Collas P., 2009, The state-of-the-art of chromatin immunoprecipitation, Methods Mol Biol, 567: 1-25
http://dx.doi.org/10.1007/978-1-60327-414-2
http://dx.doi.org/10.1007/978-1-60327-414-2_1
PMid:19588082

Collas P., 2010, The current state of chromatin immunoprecipitation, Mol Biotechnol, 45: 87-100
http://dx.doi.org/10.1007/s12033-009-9239-8
PMid:20077036

Cuddapah S., Barski A., Cui K., Schones D.E., Wang Z., Wei G., and Zhao K., 2009, Native chromatin preparation and Illumina/Solexa library construction, Cold Spring Harb Protoc, 2009: pdb prot5237

Do Rego T.G., Roider H.G., De Carvalho F.A., and Costa I.G., 2012, Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models, Bioinformatics, 28: 2297-2303
http://dx.doi.org/10.1093/bioinformatics/bts362
PMid:22730432

Dreszer T.R., Karolchik D., Zweig A.S., Hinrichs A.S., Raney B.J., Kuhn R.M., Meyer L.R., Wong M., Sloan C.A., Rosenbloom K.R., Roe G., Rhead B., Pohl A., Malladi V.S., Li C.H., Learned K., Kirkup V., Hsu F., Harte R.A., Guruvadoo L., Goldman M., Giardine B.M., Fujita P.A., Diekhans M., Cline M.S., Clawson H., Barber G.P., Haussler D., and James Kent W., 2012, The UCSC Genome Browser database: extensions and updates 2011, Nucleic Acids Res, 40: D918-923
http://dx.doi.org/10.1093/nar/gkr1055
PMid:22086951 PMCid:PMC3245018

Egelhofer T.A., Minoda A., Klugman S., Lee K., Kolasinska-Zwierz P., Alekseyenko A.A., Cheung M.S., Day D.S., Gadel S., Gorchakov A.A., Gu T., Kharchenko P.V., Kuan S., Latorre I., Linder-Basso D., Luu Y., Ngo Q., Perry M., Rechtsteiner A., Riddle N.C., Schwartz Y.B., Shanower G.A., Vielle A., Ahringer J., Elgin S.C., Kuroda M.I., Pirrotta V., Ren B., Strome S., Park P.J., Karpen G.H., Hawkins R.D., and Lieb J.D., 2011, An assessment of histone-modification antibody quality, Nat Struct Mol Biol, 18: 91-93
http://dx.doi.org/10.1038/nsmb.1972
PMid:21131980 PMCid:PMC3017233

Fingerman I.M., Mcdaniel L., Zhang X., Ratzat W., Hassan T., Jiang Z., Cohen R.F., and Schuler G.D., 2011, NCBI Epigenomics: a new public resource for exploring epigenomic data sets, Nucleic Acids Res, 39: D908-912
http://dx.doi.org/10.1093/nar/gkq1146
PMid:21075792 PMCid:PMC3013719

Flicek P., Amode M.R., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fairley S., Fitzgerald S., Gil L., Gordon L., Hendrix M., Hourlier T., Johnson N., Kahari A.K., Keefe D., Keenan S., Kinsella R., Komorowska M., Koscielny G., Kulesha E., Larsson P., Longden I., Mclaren W., Muffato M., Overduin B., Pignatelli M., Pritchard B., Riat H.S., Ritchie G.R., Ruffier M., Schuster M., Sobral D., Tang Y.A., Taylor K., Trevanion S., Vandrovcova J., White S., Wilson M., Wilder S.P., Aken B.L., Birney E., Cunningham F., Dunham I., Durbin R., Fernandez-Suarez X.M., Harrow J., Herrero J., Hubbard T.J., Parker A., Proctor G., Spudich G., Vogel J., Yates A., Zadissa A., and Searle S.M., 2012, Ensembl 2012, Nucleic Acids Res, 40: D84-90
http://dx.doi.org/10.1093/nar/gkr991
PMid:22086963 PMCid:PMC3245178

Giannopoulou E.G., and Elemento O., 2011, An integrated ChIP-seq analysis platform with customizable workflows, BMC Bioinformatics, 12: 277
http://dx.doi.org/10.1186/1471-2105-12-277
PMid:21736739 PMCid:PMC3145611

Gilchrist D.A., Fargo D.C., and Adelman K., 2009, Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation, Methods, 48: 398-408
http://dx.doi.org/10.1016/j.ymeth.2009.02.024
PMid:19275938 PMCid:PMC3431615

Goecks J., Nekrutenko A., and Taylor J., 2010, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol, 11: R86
http://dx.doi.org/10.1186/gb-2010-11-8-r86
PMid:20738864 PMCid:PMC2945788

Hirst M., and Marra M.A., 2010, Next generation sequencing based approaches to epigenomics, Brief Funct Genomics, 9: 455-465

http://dx.doi.org/10.1093/bfgp/elq035
PMid:21266347 PMCid:PMC3080743

Horak C.E., and Snyder M., 2002, ChIP-chip: a genomic approach for identifying transcription factor binding sites, Methods Enzymol, 350: 469-483
http://dx.doi.org/10.1016/S0076-6879(02)50979-4

Huang W., Umbach D.M., Vincent Jordan N., Abell A.N., Johnson G.L., and Li L., 2011, Efficiently identifying genome-wide changes with next-generation sequencing data, Nucleic Acids Res, 39: e130
http://dx.doi.org/10.1093/nar/gkr592
PMid:21803788 PMCid:PMC3201882

Ikegami K., Ohgane J., Tanaka S., Yagi S., and Shiota K., 2009, Interplay between DNA methylation, histone modification and chromatin remodeling in stem cells and during development, Int J Dev Biol, 53: 203-214
http://dx.doi.org/10.1387/ijdb.082741ki
PMid:19412882

Johnson D.S., Mortazavi A., Myers R.M., and Wold B., 2007, Genome-wide mapping of in vivo protein-DNA interactions, Science, 316: 1497-1502
http://dx.doi.org/10.1126/science.1141319
PMid:17540862

Karlic R., Chung H.R., Lasserre J., Vlahovicek K., and Vingron M., 2010, Histone modification levels are predictive for gene expression, Proc Natl Acad Sci U S A, 107: 2926-2931
http://dx.doi.org/10.1073/pnas.0909344107
PMid:20133639 PMCid:PMC2814872

Kim H., Kim J., Selby H., Gao D., Tong T., Phang T.L., and Tan A.C., 2011, A short survey of computational analysis methods in analysing ChIP-seq data, Hum Genomics, 5: 117-123
http://dx.doi.org/10.1186/1479-7364-5-2-117
PMid:21296745 PMCid:PMC3525234

Kouzarides T., 2007, Chromatin modifications and their function, Cell, 128: 693-705
http://dx.doi.org/10.1016/j.cell.2007.02.005
PMid:17320507

Kurdistani S.K., 2011, Histone modifications in cancer biology and prognosis, Prog Drug Res, 67: 91-106
PMid:21141726

Liu E.T., Pott S., and Huss M., 2010, Q&A: ChIP-seq technologies and the study of gene regulation, BMC Biol, 8: 56
http://dx.doi.org/10.1186/1741-7007-8-56
PMid:20529237 PMCid:PMC2871264

Maglott D., Ostell J., Pruitt K.D., and Tatusova T., 2011, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res, 39: D52-57
http://dx.doi.org/10.1093/nar/gkq1237
PMid:21115458 PMCid:PMC3013746

Mardis E.R., 2007, ChIP-seq: welcome to the new frontier, Nat Methods, 4: 613-614
http://dx.doi.org/10.1038/nmeth0807-613
PMid:17664943

Mcentyre J., and Lipman D., 2001, PubMed: bridging the information gap, CMAJ, 164: 1317-1319
PMid:11341144 PMCid:PMC81025

Mcghee J.D., and Felsenfeld G., 1980, Nucleosome structure, Annu Rev Biochem, 49: 1115-1156
http://dx.doi.org/10.1146/annurev.bi.49.070180.005343
PMid:6996562

Muers M., 2011, Functional genomics: the modENCODE guide to the genome, Nat Rev Genet, 12: 80
http://dx.doi.org/10.1038/nrg2942
PMid:21245826

Park P.J., 2009, ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet, 10: 669-680
http://dx.doi.org/10.1038/nrg2641
PMid:19736561 PMCid:PMC3191340

Pepke S., Wold B., and Mortazavi A., 2009, Computation for ChIP-seq and RNA-seq studies, Nat Methods, 6: S22-32
http://dx.doi.org/10.1038/nmeth.1371
PMid:19844228

Quinlan A.R., and Hall I.M., 2010, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, 26: 841-842
http://dx.doi.org/10.1093/bioinformatics/btq033
PMid:20110278 PMCid:PMC2832824

Rashid N.U., Giresi P.G., Ibrahim J.G., Sun W., and Lieb J.D., 2011, ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions, Genome Biol, 12: R67
http://dx.doi.org/10.1186/gb-2011-12-7-r67
PMid:21787385 PMCid:PMC3218829

Schones D.E., Cui K., and Cuddapah S., 2011, Genome-wide approaches to studying yeast chromatin modifications, Methods Mol Biol, 759: 61-71
http://dx.doi.org/10.1007/978-1-61779-173-4_4
PMid:21863481

Schones D.E., and Zhao K., 2008, Genome-wide approaches to studying chromatin modifications, Nat Rev Genet, 9: 179-191
http://dx.doi.org/10.1038/nrg2270
PMid:18250624

Shin H., Liu T., Manrai A.K., and Liu X.S., 2009, CEAS: cis-regulatory element annotation system, Bioinformatics, 25: 2605-2606
http://dx.doi.org/10.1093/bioinformatics/btp479
PMid:19689956

Skinner M.E., Uzilov A.V., Stein L.D., Mungall C.J., and Holmes I.H., 2009, JBrowse: a next-generation genome browser, Genome Res, 19: 1630-1638

http://dx.doi.org/10.1101/gr.094607.109
PMid:19570905 PMCid:PMC2752129

Szalkowski A.M., and Schmid C.D., 2011, Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts, Brief Bioinform, 12: 626-633
http://dx.doi.org/10.1093/bib/bbq068
PMid:21059603

Taslim C., Wu J., Yan P., Singer G., Parvin J., Huang T., Lin S., and Huang K., 2009, Comparative study on ChIP-seq data: normalization and binding pattern characterization, Bioinformatics, 25: 2334-2340
http://dx.doi.org/10.1093/bioinformatics/btp384
PMid:19561022 PMCid:PMC2800347

Ucar D., Hu Q., and Tan K., 2011, Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering, Nucleic Acids Res, 39: 4063-4075
http://dx.doi.org/10.1093/nar/gkr016
PMid:21266477 PMCid:PMC3105409

Washington N.L., Stinson E.O., Perry M.D., Ruzanov P., Contrino S., Smith R., Zha Z., Lyne R., Carr A., Lloyd P., Kephart E., Mckay S.J., Micklem G., Stein L.D., and Lewis S.E., 2011, The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details, Database (Oxford), 2011: bar023
http://dx.doi.org/10.1093/database/bar023
PMid:21856757 PMCid:PMC3170170

Whiteford N., Skelly T., Curtis C., Ritchie M.E., Lohr A., Zaranek A.W., Abnizova I., and Brown C., 2009, Swift: primary data analysis for the Illumina Solexa sequencing platform, Bioinformatics, 25: 2194-2199
http://dx.doi.org/10.1093/bioinformatics/btp383
PMid:19549630 PMCid:PMC2734321

Wilbanks E.G., and Facciotti M.T., 2010, Evaluation of algorithm performance in ChIP-seq peak detection, PLoS One, 5: e11471
http://dx.doi.org/10.1371/journal.pone.0011471
PMid:20628599 PMCid:PMC2900203

Wood A.C., Rijsdijk F., Asherson P., and Kuntsi J., 2011, Inferring Causation from Cross-Sectional Data: Examination of the Causal Relationship between Hyperactivity-Impulsivity and Novelty Seeking, Front Genet, 2: 6
http://dx.doi.org/10.3389/fgene.2011.00006
PMid:22303305 PMCid:PMC3268378

Xu H., Handoko L., Wei X., Ye C., Sheng J., Wei C.L., Lin F., and Sung W.K., 2010, A signal-noise model for significance analysis of ChIP-seq with negative control, Bioinformatics, 26: 1199-1204
http://dx.doi.org/10.1093/bioinformatics/btq128
PMid:20371496

Xu H., Wei C.L., Lin F., and Sung W.K., 2008, An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data, Bioinformatics, 24: 2344-2349
http://dx.doi.org/10.1093/bioinformatics/btn402
PMid:18667444

Zang C., Schones D.E., Zeng C., Cui K., Zhao K., and Peng W., 2009, A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, Bioinformatics, 25: 1952-1958
http://dx.doi.org/10.1093/bioinformatics/btp340
PMid:19505939 PMCid:PMC2732366

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., and Liu X.S., 2008, Model-based analysis of ChIP-Seq (MACS), Genome Biol, 9: R137
http://dx.doi.org/10.1186/gb-2008-9-9-r137
PMid:18798982 PMCid:PMC2592715

Zhao Q., and Zhang Y., 2011, Epigenome sequencing comes of age in development, differentiation and disease mechanism research, Epigenomics, 3: 207-220
http://dx.doi.org/10.2217/epi.10.78
PMid:22122282