Chapter 25

Epigenetic Analysis: ChIP-chip and ChIP-seq

Matteo Pellegrini and Roberto Ferrari

Abstract

The access of transcription factors and the replication machinery to DNA is regulated by the epigenetic state of chromatin. In eukaryotes, this complex layer of regulatory processes includes the direct methylation of DNA, as well as covalent modifications to histones. Using next-generation sequencers, it is now possible to obtain profiles of epigenetic modifications across a genome using chromatin immunoprecipitation followed by sequencing (ChIP-seq). This technique permits the detection of the binding of proteins to specific regions of the genome with high resolution. It can be used to determine the target sequences of transcription factors, as well as the positions of histones with specific modification of their N-terminal tails. Antibodies that selectively bind methylated DNA may also be used to determine the position of methylated cytosines. Here, we present a data analysis pipeline for processing ChIP-seq data, and discuss the limitations and idiosyncrasies of these approaches.

Key words: ChIP-seq, Chromatin immunoprecipitation, Transcription factor binding sites, Peak calling, Histone modification, DNA methylation, Next-generation sequencing, Poisson statistics

1. Introduction

The DNA sequence is the primary blueprint that controls cellular function. However, a complex layer of molecular modifications that are referred to as the epigenetic code affects the transcription and replication of DNA. Epigenetic modifications include the direct methylation of cytosines, as well as modifications to the structure of chromatin. In particular, the N-terminal tails of histones can be modified by a large number of enzymes that add or remove methyl, acetyl, phosphorous, or ubiquitin groups, among others (1). The characterization of the epigenetic state of chromatin is complicated by the fact that each cell type in an organism has a different epigenetic state. In fact, the epigenetic differences

Junbai Wang et al. (eds.), *Next Generation Microarray Bioinformatics: Methods and Protocols*,
Methods in Molecular Biology, vol. 802, DOI 10.1007/978–1-61779–400–1_25,

between cells are fundamental to the generation of diversity between cell types that all arise from a clonal population with identical DNA sequences.

The readout of epigenetic modification on a genome-wide scale can be carried out using chromatin immunoprecipitation techniques (2). In brief, these methods involve the crosslinking of DNA to protein using crosslinking agents as a first step, in order to freeze protein–DNA and protein–protein interactions. Subsequently, the chromatin is sonicated to yield fragments of protein-bound DNA that are typically a few hundred bases long. These fragments are then purified using antibodies that are specific to the particular modification that is being profiled (e.g., a specific modification of the histone tail, or cytosine methyl groups). The immunoprecipitated fraction is isolated, and the crosslinks are reversed to yield the DNA fragments bound to the protein of interest. These fragments are then either hybridized to a microarray (ChIP-chip) or sequenced using a high-throughput sequencing platform (ChIP-seq). The immunoprecipitated fragments are then compared to the fragments that were not selectively immunoprecipitated, often referred to as the input material, to identify sequences that enriched in the former with respect to the latter. These enriched regions correspond to the DNA sequences that are bound by the protein of interest.

Before the advent of next-generation sequencing, ChIP-chip was the standard technique for these types of assays (3). However, for many organisms it is not practical to generate genome-wide tiling arrays, and hence ChIP-chip data sets were often not genome-wide. Furthermore, the ability to detect a binding site in a ChIP-chip experiment is limited by the resolution of the probes on the array. Finally, the signal obtained by hybridization intensities on an array is analog, and it is often difficult to determine levels of enrichment that are statistically significant and hence indicative of true binding sites. Many of these limitations are overcome by using ChIP-seq (4). Since sequencing is not limited in any way by probes, and it is therefore a truly genome-wide approach. The only limitation is that it is impossible to definitively determine the position of a peak if it lies within a sequence that is repeated in the genome. For this reason, often ChIP-seq peaks are only called when they are associated with unique sequences that appear only once in the genome, and this can be a significant limitation since repetitive sequences are very abundant in large genomes such as that of humans. Nonetheless, Chip-seq technology is rapidly eclipsing the older ChiP-chip approach and we therefore present detailed protocols for the analysis of this latter data rather than the former.

2. Materials

In this chapter, we describe the computational protocols for analyzing ChIP-seq data. We will not discuss the experimental protocols for generating ChIP-seq libraries, as these have been published elsewhere.

2.1. Base Calls

From our standpoint, therefore, the material to carry out the analyses, we describe consist of the base calls that are output by the DNA sequencer. For the most common case of data generated by Illumina sequencers, this data consists of tens of millions of short reads that typically range from 36 to 76 bases in length (5). Several data standards have been developed for the encoding of these reads into flat files. The most common is the FASTQ standard which contains both the base calls at each position of the read as well as the quality scores that denote the confidence in the base calls (6) (see Note 1).

2.2. Alignment Software

The second essential material is an alignment tool to align the reads to a reference sequence. Over the past couple of years there has been a proliferation of new alignment tools that are specialized for the rapid alignment of millions of short reads to large reference genomes. These tools include Bowtie (7), Maq (8), and Soap (9) among others (see Note 2). Since the reads contain fragments of DNA from the genome, the alignments do not need to consider gaps (although some of these tools do permit the inclusion of small gaps). Similarly one only expects a few mismatches between the read sequence and reference genome due to base calling errors or polymorphisms in the genome sequence, and all these aligners allow for the inclusion of several mismatches in the alignment. Finally, most of the alignment tools do not explicitly consider base call quality scores when attempting to identify the optimal alignment for a read. However, some tools, such as Bowtie, do consider the quality scores after the alignment has been performed using only the base calls.

2.3. Genome Browser

The other critical tool to enable the analysis and interpretation of ChIP-seq data is a genome browser. This application allows one to zoom and pan to any position in the genome, and view the mapped reads. This is critical for both verifying the data analysis protocols and to generate detailed information for specific loci. Several tools are available for this purpose including the Integrated Genome Browser (10), and the UCSC genome browser (11) among many others (see Note 3). Typically, the data is uploaded in formats that depict either individual reads (e.g., bed format) or the accumulated counts associated

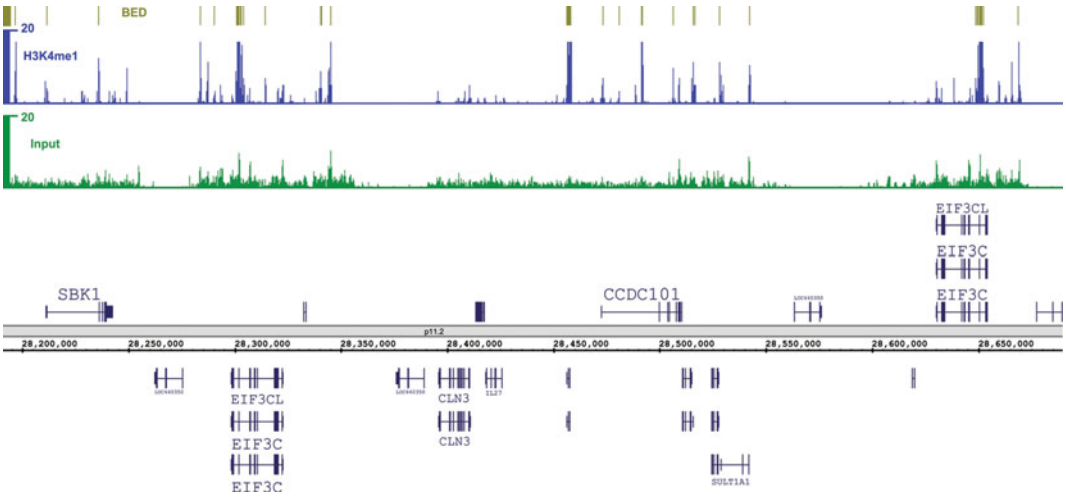

Fig. 1. A sample locus viewed using the UCSC genome browser. The first track from the *top* contains the windows that are found to be significantly enriched in the IP vs. input for H3K4me1, a histone mark. The second track, labeled H3K4me1, shows the counts for each 100 base window. The third track contains the input control. The tracks on the *bottom* contain the gene annotation which indicates the transcriptional start and end sites and the positions of introns for the two genes in this locus.

with reads that overlap a specific base (e.g., wiggle tracks). Examples of the output of these browsers may be seen in Fig. 1.

3. Methods

The methods that we describe will utilize the base calls described above, in conjunction with an alignment tool, to identify all the regions of the genome that containing significant peaks for the particular DNA binding protein that is being tested. Along with a description of the methods for data analysis, we also discuss software that has been developed to visualize the resulting data on the genome.

3.1. Read Alignment

The first step in the data analysis pipeline is to align the reads to a reference genome or other reference sequence of interest. Usually, alignments do not allow for gaps to be inserted between the reads and the reference sequence. For a 36-base reads it is customary to accept all alignments that generate no more than two mismatches between the reads and the reference sequence. The number of allowed mismatches can be adjusted to a higher level for longer reads, but it is difficult to come up with systematic approaches to determine what the optimal number of allowed mismatches should be, and thus this value is nearly always assigned based on ad hoc criteria. Finally, as we discussed above, reads that align with equal

scores to multiple locations on the genome are most often thrown out, since they cannot be unambiguously assigned to a single peak. A variety of approaches have been developed to deal with multiple mapping problems. These include the probabilistic reassignment of reads based on the surrounding region (12) (which assumes that if a read maps to two locations, it is more likely to originate from the one that has more reads mapping in the immediate neighborhood), to the use of representations of the genome that explicitly account for the repeat structure of the sequence (13), to the simple addition of a weight to each read based on the multiplicity of its binding sites. While accounting for repeats is more critical in other applications (such as RNA-seq), in general people have found that it is less important in ChIP-seq applications, and generally none of these more sophisticated approaches are used.

Once the alignments have been completed the next step involves the evaluation of the alignment quality. This is measured using several criteria, the first and most significant of which is the fraction of reads that map to a unique location in the genome. In general, not all reads can map to unique locations because the reference sequence contains repetitive regions and because the sequencing process usually introduces random errors in the base calls. However, a well-prepared ChIP-seq library should yield unique alignments for somewhere around half of the reads. If the actual number is significantly lower (i.e., less than 30%) then this might indicate that there was a problem in the library preparation or the sequencing run. To attempt to optimize the number of reads that map to unique location on the reference sequence, it is common to attempt to trim the end of the reads as these often have lower base calling accuracy. As we see in Fig. 2 for a typical case, the number of mismatches tends to be high at the very start of the reads, low in the middle, and increases toward the end of the read. By trimming these locations it is possible to increase the number of reads that can be uniquely mapped to the genome.

One final consideration that is important for ChIP-seq libraries is that they are often plagued by low complexity. That is, the number of unique reads that are generated by the sequencer is often significantly smaller than the total number of reads, due to the resequencing of the same read multiple times. This phenomenon tends to be more common in ChIP-seq experiments because it is often difficult to produce large quantities of DNA using chromatin immunoprecipitation, due to the limits of the antibody affinity for its target, and potentially due to the limited number of sites where the target protein is bound (see Note 4). However, if we observe the same read multiple times, this does not necessarily imply that the target protein has higher affinity for the corresponding sequence, but could also be due to the fact that the particular read sequence is more efficiently amplified during the library preparation protocol. As a result, to minimize these

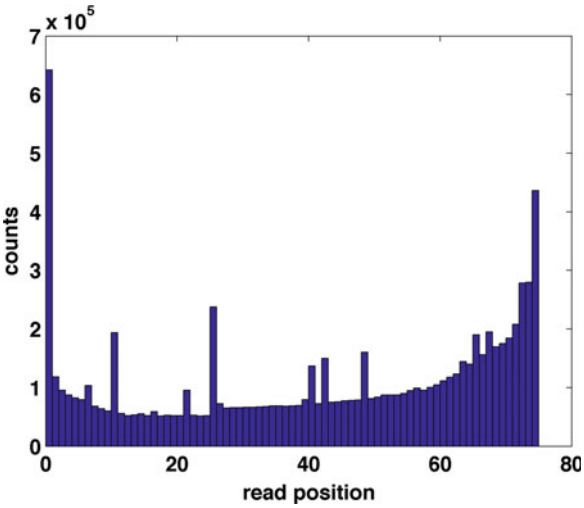

Fig. 2. Mismatch counts as a function of position in read. Reads were aligned to the genome using Bowtie. Up to two mismatches were allowed per alignment. The position of the mismatch along the read is indicated on the *x*-axis, and the total number of mismatches at this position is shown on the *y*-axis. The first base has a significant number of mismatches compared to the first 50 bases. The last ten bases show an increasing number of mismatches. A few positions in the middle of the read also show anomalously high mismatch counts, possibly due to some perturbation to the sequencing cycle during this run.

biases, we usually only align the unique reads in the library, and not the total reads. This may be accomplished by either sorting the reads in the library and selecting unique reads, or by combining reads that map to the same location into a single read that contributes only one count.

3.2. Peak Detection

Once the reads have been aligned to the genome, the binding sites of the target protein can be indentified. To accomplish this it is customary to first tile the genome using windows, within which we attempt to detect peaks. The size of the window is typically between 100 and a couple of hundred bases. This roughly corresponds to the size of the sonicated DNA fragments that are used to generate the ChIP-seq library. Due to the limited sequencing depth (currently 30–40 million reads are produced for each library), and the size of sonication fragments, it is usually not possible to detect peaks with more than 100 base resolution. The tiling can either be sequential, or interleaved.

The counts within each window are determined by computing both the number of reads whose alignment starts directly within the window, as well as reads that align outside, but near the edges of the window. If we assume that each read corresponds to a one to two hundred base DNA fragment, then even reads that align

to a position 100 bases upstream of the window, overlap and contribute to the counts in the window. Each read can either contribute a fractional count to the window, measured by the fraction of the read that overlaps the window, or more simply any level of overlap can lead to a discrete increment of one count. It is also important to realize that reads that map to the negative DNA strand contribute to windows that are upstream of the start site, while reads that map to the positive strand contribute to windows that are downstream of the start site.

To determine whether the counts within a window are significant, it is necessary to compare these to a background level. The most simplistic model is that the background level of each window is simply the average counts for all the windows across the genome. However, it is more customary to sequence a control library, usually referred to as the input library, to estimate the background counts. The input library consists of all the DNA fragments that were not immunoprecipitated during the course of the chromatin immunoprecipitation protocol. It should certainly have a more uniform distribution across the genome than the immunoprecipitated (IP) library, however, recent studies have shown that sonication and DNA purification methods result in biases that often lead to additional peaks around transcription start sites (14). Therefore, comparing the IP libraries with the input can remove some false-positive peaks that are just due to sonication biases. However, in order or this comparison to be meaningful, the input library must first be normalized so that it contains the same total numbers of counts as the IP library (see Note 5).

Once the counts of the IP and input libraries in each window in the genome have been computed, the final step involves that determination of the statistical significance of the increase in IP over input, if any. It is assumed that the counts in each window are approximately distributed according to the Poisson distribution, as the generation of a sequence library fragments from a genome is essentially a Poisson process (15). Therefore, to estimate the probability of observing the IP counts we use the cumulative Poisson distribution with an expected value provided by the input counts. That is, we compute the probability of observing the IP counts, or a higher value, given the expected number provided by the input counts. This approach will be noisy when the input counts are low, or zero. If the input counts are zero we can set the expected distribution to the genome average. This method will generate a *P*-value for each window in the genome. The last step requires one to estimate false-discovery rates (FDRs) based on this *P*-value distribution. There are many statistical approaches for estimating FDRs from *P*-value distribution, and we will not discuss these in detail here other than to provide several references (16, 17).

3.3. Data Visualization

An important component of ChIP-seq data analysis is the visualization of the data on a genome browser. As discussed above there are various tools that can be used for this purpose. Here, we illustrate the use of the UCSC Genome Browser (18). We illustrate a sample locus in Fig. 1. We show tracks for the IP counts the input counts, as well as the regions that are deemed to be significantly enriched in IP vs. input. The data is generated using a variety or formats. The counts files are generated using the wiggle format that describes the chromosome, position, and counts in each window. The significant peaks are displayed using the bed format, which denotes that boundaries of the region with significant enrichment. It is critical to generate these types of files when analyzing ChIP-seq data, to determine whether the peak finding algorithm, and the particular parameters chosen by the user, are in fact yielding reasonable peaks. The tool also allows one to visualize the data in any region of interest in the genome, in order to answer specific question about loci of interest.

3.4. Downstream Analysis

There are a multitude of possible downstream analyses that can be conducted on ChIP-seq data and here we limit ourselves to describe only a small set. It is, for instance, customary to overlay the peaks identified in the ChIP-seq data with positions of transcriptional start sites (TSS), as these can be directly associated regulatory regions. In this regard, it is customary to generate "meta plots" that display the total number of peaks a certain distance from the TSS. For example, in Fig. 3 we show the total number of peaks around the TSS for a specific histone modification. We note right away the modification is enriched around the TSS but depleted right at the TSS. Similar analyses can be performed for any other genomic feature, such as transcription termination sites, intron–exon boundaries, or repeat boundaries.

A slightly different representation of the enrichment around features identifies the average trends along the entire length of the feature (e.g. (19)) (Fig. 3, bottom panel). That is each gene is rescaled so that it is covered by a fixed number of bins (typically 100 or so). The density of peaks in each bin is then computed (i.e., the number of peaks divided by the bin length). The values of the bins are averaged or summed over all the genes in the genome to generate the average trend of peaks across the genome. The same analysis is usually performed on the upstream and downstream regions of the genes, which can comprise 50% or so of the total gene length. The combination of the upstream, gene, and downstream region then generates a comprehensive view of the trends in the data around genes. Thus, unlike the previous plots, these provide a more global view of the peak trends across genes. As before, these types of analyses may be performed across any genomic feature, and not just genes. It may be of interest to generate the average trends across repetitive elements in the genome, or internal exons.

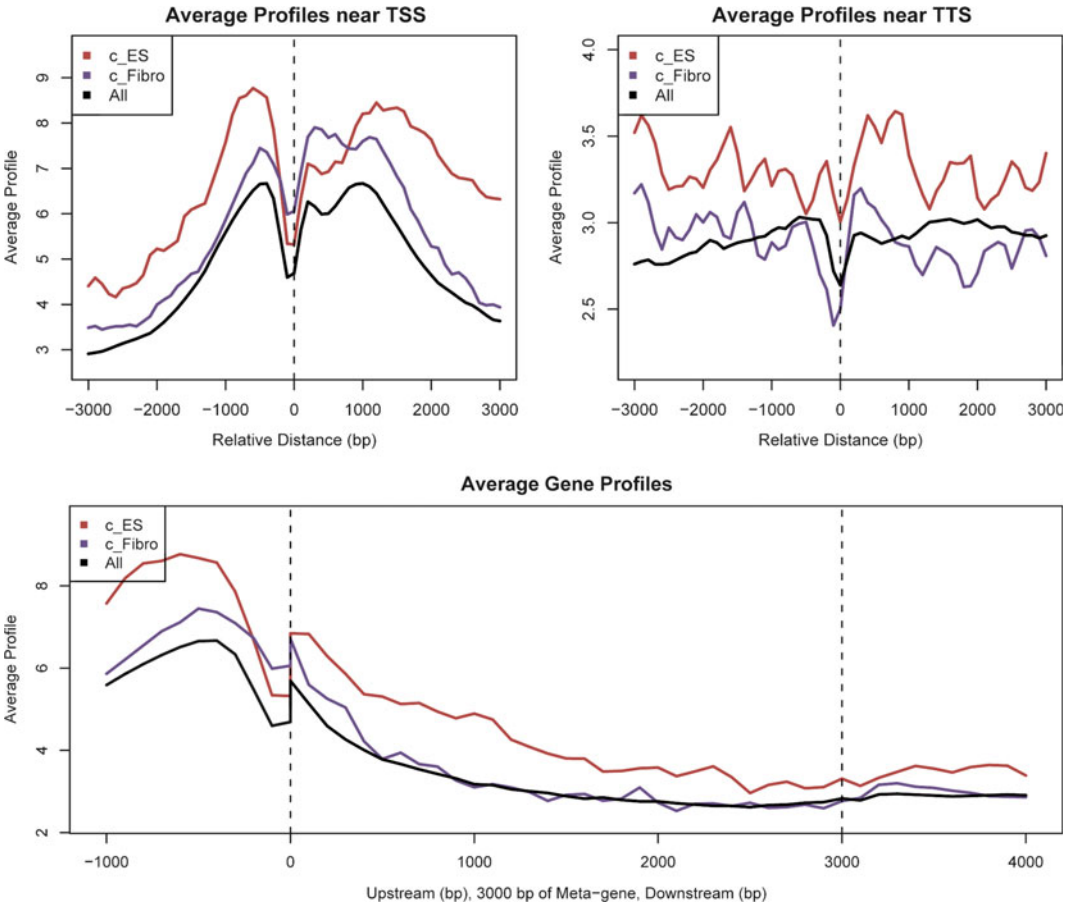

Fig. 3. Average levels of H3K4me1 acetylation at the start and end of genes. This meta-analysis computes the average levels of H3K4me1 in a 6-kb region surrounding the transcriptional start site (*top right*) and end site (*top left*). We see that H3K4me1 positive regions are preferentially located around, but not right over the start sites. In the bottom panel we show a scaled metagene analysis, where all genes have been aligned so that they start at 0 and end at 3,000. The average H3K4me1 levels 1 kb upstream and downstream of all genes are also shown. In all cases, genes are grouped into three groups. c_ES are genes that are differentially induced in embryonic stem cells and c_Fibro are those induced in fibroblasts (24), while All are all the genes.

Another common analysis attempts to summarize the locations of peaks throughout the genome. While the previous two procedures summarize the distribution of peaks around genes, a large fraction of the peaks may lie far from genes, and thus would not be considered in these analyses. To account for these, it is customary to generate a table that describe the fractions of peaks that are within genes, or a certain distance from genes. Such a table might include categories that correspond to regions that are, for example, tens of kilobases away from genes.

Of course the analyses described above are only a small sampling of all the possible downstream analyses that can be attempted on this data. It is also possible to analyze the sequence

composition of peak regions, or search for specific sequence motifs. One might also consider the distribution of peaks across chromosomes to identify large-scale trends. However, a comprehensive description of all of these methodologies lies outside the scope of this chapter (see Note 6).

4. Notes

1. Many aligners do not use base call information and it is therefore often sufficient to simply provide the base calls. These files are sometimes referred to as raw formats and are significantly smaller in size than the FASTQ format.
2. Among the many alignment tools that have become available over the past few years, Bowtie is probably the most popular, as it tends to be one of the fastest, with an efficient indexing scheme that requires relatively small amounts of memory. For a typical mammalian genome the indices built from the reference sequence are around 4 gigabytes, and a single lane of data can be aligned in about an hour.
3. The UCSC genome browser is probably the most widely used browser. It allows users to upload data onto the UCSC site, where it can be compared to data that permanently resides on the server (such as annotation files). However, if the genome of interest is not preloaded in the browser, it is very difficult to upload it onto the browser. Nonetheless, various instances of the browser are maintained by other groups that contain additional genomes (e.g. (20)).
4. To increase the complexity of ChIP-seq libraries it is necessary to immunoprecipitate as much material as possible, which in typical circumstances may require performing multiple immunoprecipitations on batches of millions of cells.
5. Other popular peak calling approaches can be significantly more sophisticated, by taking into consideration the shape of the peak, the length of reads, and the posterior probabilities (21, 22).
6. An example of a suite of tools that may be applied for these types of analyses may be found at ref. 23.

Acknowledgments

The authors would like to thank Professor Bernard L. Mirkin for development of the drug-resistant models of human neuroblastoma cells and for his advice and encouragement, and Jesse Moya for technical assistance. This work was supported by Broad Stem Cell Research Center and Institute of Genomics and Proteomics at UCLA.

References

1. Jenuwein T, Allis CD (2001) Translating the histone code. Science 293:1074–1080.
2. Nelson JD, Denisenko O, Bomsztyk K (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat Protoc 1:179–185.
3. Buck MJ, Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83:349–360.
4. Valouev A, Johnson DS, Sundquist A et al (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5:829–834.
5. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141.
6. Cock PJ, Fields CJ, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771.
7. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
8. http://maq.sourceforge.net/.
9. Li R, Li Y, Kristiansen K et al (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714.
10. Nicol JW, Helt GA, Blanchard SG Jr et al (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 25:2730–2731.
11. Rhead B, Karolchik D, Kuhn RM et al (2010) The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38:D613–619.
12. Clement NL, Snell Q, Clement MJ et al (2010) The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics 26:38–45.
13. Pevzner PA, Tang H (2001) Fragment assembly with double-barreled data. Bioinformatics 17:S225–233.
14. Auerbach RK, Euskirchen G, Rozowsky J et al (2009) Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci U S A 106:14926–14931.
15. Mikkelsen TS, Ku M, Jaffe DB et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560.
16. Benjamini Y, Drai D, Elmer G et al (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125:279–284.
17. Muir WM, Rosa GJ, Pittendrigh BR et al (2009) A mixture model approach for the analysis of small exploratory microarray experiments. Comput Stat Data Anal 53:1566–1576.
18. http://genome.ucsc.edu/.
19. Cokus SJ, Feng S, Zhang X et al (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452:215–219.
20. http://genomes.mcdb.ucla.edu.
21. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137.
22. Spyrou C, Stark R, Lynch AG et al (2009) BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics 10:299.
23. http://liulab.dfci.harvard.edu/CEAS/.
24. Chin MH, Mason MJ, Xie W et al (2009) Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5:111–123.