Chapter 9

Computational Analysis of ChIP-seq Data

Hongkai Ji

Abstract

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) is a new technology to map protein–DNA interactions in a genome. The genome-wide transcription factor binding site and chromatin modification data produced by ChIP-seq provide invaluable information for studying gene regulation. This chapter reviews basic characteristics of ChIP-seq data and introduces a computational procedure to identify protein–DNA interactions from ChIP-seq experiments.

Key words: Transcription factor binding site, high-throughput sequencing, peak detection, false discovery rate.

1. Introduction

Chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing (ChIP-seq) is a new technology to map protein–DNA interactions in genomes (1–4). In this technology, a protein of interest (POI) is cross-linked to chromatin. Chromatin is sheared into small fragments. The POI and its bound chromatin fragments are immunoprecipitated using an antibody specific to the protein. After reversing the cross-links, a DNA sample called "ChIP sample" is obtained. In many studies, a negative control sample is prepared in parallel using a similar protocol that bypasses the immunoprecipitation step. Compared to the control sample, the ChIP sample is enriched in DNA fragments bound by the protein of interest. After size selection and further processing, DNA fragments in the samples are sequenced from both ends using one of the recently developed high-throughput sequencing

I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology 674,
DOI 10.1007/978-1-60761-854-6_9,

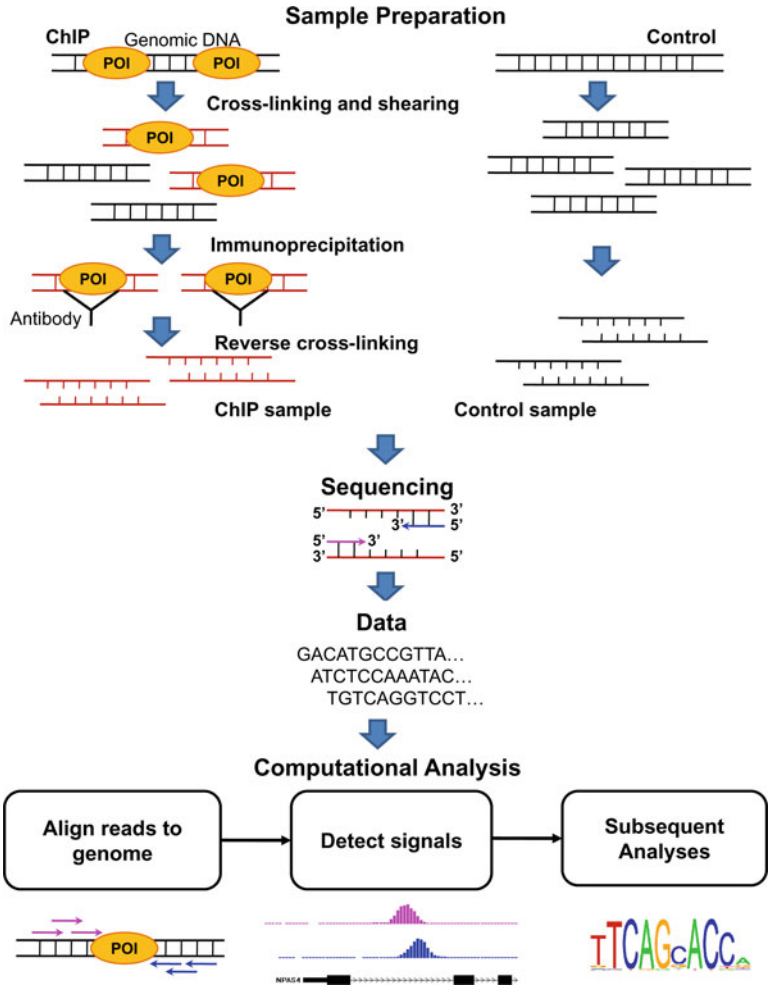

Fig. 9.1. Workflow for ChIP-seq.

platforms (5). This produces tens of millions of sequence tags, also known as sequence reads. By computationally mapping these reads to a reference genome and looking for genomic regions where ChIP reads are enriched, genomic loci with protein–DNA interactions can be identified (**Fig.** 9.1). Currently, this technology is widely used to study transcription factor binding sites (TFBS) (1, 2) and chromatin modifications (3, 4). The genome-wide transcription factor binding site and chromatin state data produced by ChIP-seq provide invaluable information for studying gene regulation.

An earlier technology to map protein–DNA interactions in genomes is ChIP-chip (6, 7), which uses chromatin immunoprecipitation to enrich protein-bound DNAs and hybridizes the enriched DNA fragments to genome tiling arrays. Compared to ChIP-chip, ChIP-seq has several advantages (8). First, ChIP-seq

does not rely on array hybridization. As a result, it does not suffer from the biases and noise caused by cross-hybridization, the varying GC content of probe sequences and other issues related to hybridization chemistry, although ChIP-seq may have its own biases that are not well understood currently. Second, ChIP-chip measures enrichment by intensities of hybridization which may saturate at high signal, whereas ChIP-seq measures enrichment by tag counts which can handle signals in a much broader dynamic range. Third, protein–DNA interactions detected by ChIP-chip are restricted to genomic regions for which probes are available. Repetitive regions in the genome usually are excluded from the array design. In contrast, ChIP-seq can be used to study protein–DNA interactions in any part of the genome as long as reads can be unambiguously aligned to places where they are originally produced. For this reason, ChIP-seq is able to offer much less biased genome coverage. Fourth, for mapping TFBS, ChIP-seq is able to locate binding sites at 50–100 base pair (bp) resolution. This represents a significantly improved precision compared to the 300–1,000 bp resolution provided by ChIP-chip. Other advantages of ChIP-seq include requirement of less input materials and ability to provide extra information to study allele-specific protein binding. Thanks to these advantages, as the cost of high-throughput sequencing continues to decrease, ChIP-seq has the potential to become the dominant technology for creating genome-wide maps of protein–DNA interactions.

ChIP-seq creates unprecedented amounts of data. Extracting information from the data is not trivial. Typically, the analysis is a multiple step procedure (**Fig.** 9.1). First, raw sequence reads are mapped to the reference genome. Next, genomic regions in which ChIP reads are enriched are identified and the statistical significance of the predicted genomic regions is evaluated. Regions that satisfy certain significance criteria are reported. Subsequently, the reported regions are analyzed in various ways to help scientists understand their functional implications. These include adding gene annotations, finding or mapping transcription factor binding motifs, and correlating the protein–DNA interactions with gene expression information. The purpose of this chapter is to briefly review some basic characteristics of ChIP-seq data and introduce a computational procedure to analyze the data. We will mainly focus on describing a method to identify protein–DNA interactions and estimate the false discovery rates (FDR). Tools to perform subsequent analyses will be discussed briefly.

1.1. Types of ChIP-seq Experiments

We focus on two types of ChIP-seq experiments, namely the "one-sample experiment" and the "two-sample experiment." A two-sample experiment involves sequencing both a ChIP sample and a negative control sample. In contrast, a one-sample

experiment only involves sequencing a ChIP sample. Readers are referred to **Note 1** for a discussion on how to analyze experiments that have technical or biological replicates.

Compared to the two-sample experiment, the one-sample design is more cost effective. However, the negative control sample in the two-sample experiment allows one to build a better model to describe locus-dependent background noise, which can significantly reduce the number of false positives and false negatives in the subsequent data analyses (9, 10).

1.2. Models for Background Noise

In both one-sample and two-sample experiments, protein–DNA interactions can be identified by searching for enrichment of ChIP reads. A key component of ChIP-seq data analysis is to understand what level of enrichment is required to distinguish signals from noise.

1.2.1. Background Model for One-Sample Experiments

First consider a one-sample experiment. Assume that the length of the genome is L bps and the sample has N uniquely mapped reads in total. Consider a w bp window in the genome, and let n be the number of reads mapped to the window. Studies of negative control samples show that if the window does not contain any protein–DNA interaction of interest, n can be approximately modeled by a negative binomial distribution $NB(\alpha, \beta)$ (9). In other words, $\Pr(n=k) = \binom{k+\alpha-1}{\alpha-1}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{\beta}{\beta+1}\right)^{k}$. Here all background windows in the genome have the same values of α and β. Based on this result, one approach to characterize the background noise is to find appropriate parameter values of α and β using the observed data. When estimating α and β, one should keep in mind that the data (i.e., the ChIP sample) usually consist of a mixture of background windows and windows that contain signals; however, α and β are parameters to describe background noise only. An algorithm that estimates the background parameters α and β from a mixture of signal and noise windows will be described in **Section** 3.2.1.

Another natural way to model the read count of a background window is to assume that n follows a Poisson distribution with a rate parameter λ (i.e., $\Pr(n=k) = \lambda^k e^{-\lambda}/k!$). Recent studies show that the Poisson distribution with a fixed rate λ does not perform well to characterize the background variability in real data (9–11). For example, in **Table** 9.1, a negative control sample from a ChIP-seq experiment in mouse embryonic stem cells (12) is analyzed by both the Poisson background model and the negative binomial model. The genome is divided into 100 bp long non-overlapping windows and the number of uniquely mapped reads in each window is counted. The negative control sample contains no protein–DNA interactions of interest;

Table 9.1
Comparison of the Poisson and negative binomial background model

Read count	Observed frequency	Expected by Poisson	Expected by NB
0	0.792664	0.792664	0.792230
1	0.164843	0.164843	0.164753
2	0.034140	0.017140	0.034122
3	0.006587	0.001188	0.007057
4	0.001320	0.000062	0.001459
5	0.000288	0.000003	0.000301
6	0.000075	0.000000	0.000062
7	0.000023	0.000000	0.000013
...	...	...	...

hence all windows represent background noise. The second column of the table shows the observed frequency that a window contains k reads. The third and fourth columns show frequencies expected by the Poisson and negative binomial models, respectively. This table clearly shows that the Poisson model is not able to describe the heavy tail of the empirical read count distribution and the negative binomial model performs much better.

Using a fixed rate Poisson model assumes that background reads are generated at the same rate for all loci in the genome or, in other words, background reads are distributed uniformly across the genome. **Table** 9.1 illustrates that this assumption does not fit well with the real data. In the negative binomial model, it is implicitly assumed that the background reads are generated by Poisson distributions with different rates at different loci, and as a result, the background reads are not uniformly distributed across the genome. In order to see this, we note that a negative binomial distribution can be related to a Poisson distribution via a hierarchical model. Let us divide the genome into w bp long non-overlapping windows and assume that different windows generate reads independently. Let λ_i be the rate to generate reads in the ith window, n_i be the number of reads in window i, and assume that $n_i|\lambda_i \sim$ Poisson (λ_i). If we allow λ_i to vary across the genome but assume that λ_i's are random samples drawn independently from a locus-independent gamma distribution Gamma(α, β) (the probability density function for Gamma(α, β) is $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$), then the marginal distribution of n_i of a background window, $\Pr(n_i = k|\alpha, \beta) = \int \Pr(n_i = k|\lambda_i) f(\lambda_i|\alpha, \beta) d\lambda_i$, has the same probability density function as that of the $NB(\alpha, \beta)$.

The hypothesis that read sampling rates vary across the genome is supported by analyses of independent samples from

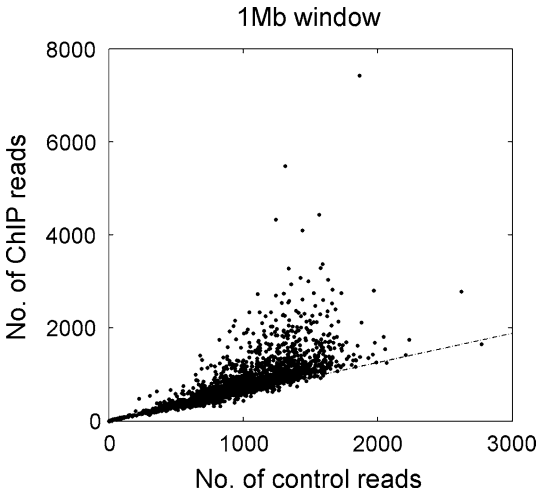

Fig. 9.2. Correlation of read numbers at the same genomic loci between a ChIP sample and a control sample. The samples are obtained from a ChIP-seq experiment that maps the NRSF TFBSs (1). The human genome is divided into non-overlapping windows, each window containing 1 million base pairs. For each window, ChIP and control reads are counted and plotted as a *dot*.

the same experiment (10). As an example, **Fig.** 9.2 shows a scatter plot that compares the window read counts between a ChIP sample and a matching negative control sample in an experiment involving transcriptional repressor NRSF (1). The plot has a positive slope and the counts from the two samples in the same genomic window are clearly correlated. This indicates that the rate for generating reads is locus dependent and is not a constant across the genome. Unfortunately, in a one-sample experiment, background reads in a particular window cannot be separated from reads that represent biological signals in the same window. For this reason, the locus-dependent Poisson rate cannot be estimated without making additional assumptions. The negative binomial model makes the assumption that the background rates λ_is follow a common gamma distribution. By making this assumption, information from all windows can be used to infer the common parameters α and β, which are then used to describe the background for each individual window. This is the underlying rationale for using a negative binomial distribution as the background model (*see* **Note 2** for an alternative solution).

1.2.2. Background Model for Two-Sample Experiments

Now consider a two-sample experiment that involves a control sample in addition to a ChIP sample. Assume that the ChIP sample has N uniquely mapped reads in total and the control sample has M uniquely mapped reads. For a w bp window indexed by i, let n_i be the number of ChIP reads mapped to the window, and m_i be the number of control reads. In the previous section, it

has been shown that the read counts in background windows can be viewed as Poisson random variables with varying rates across the genome (which results in negative binomial marginal distributions). In light of this observation, one can assume that $n_i \sim \text{Poisson}(\mu_i)$ and $m_i \sim \text{Poisson}(\lambda_i)$, where μ_i and λ_i are rates at which reads are produced in window i in the ChIP and control samples, respectively, and we allow μ_i and λ_i to have different values at different loci in the genome. For each genomic window, μ_i can be decomposed into two parts $\mu_i = \mu_{i1} + \mu_{i0}$, where μ_{i0} is the rate at which background reads are generated and μ_{i1} is the rate to generate reads corresponding to signals. Often, it is reasonable to assume that the background rates in the ChIP and control samples, μ_{i0} and λ_i, are equal up to a proportionality constant, i.e., $\mu_{i0} = c\lambda_i$. The proportionality constant c reflects the observation that the total numbers of reads in the ChIP and control samples are usually not the same. Under the assumption that $\mu_{i0} = c\lambda_i$, information from the negative control sample can be used to describe the background read sampling rate in the ChIP sample. As a result, the assumption used in the one-sample analysis that background read sampling rates from different genomic windows follow a common probability distribution is no longer required.

For a window that does not contain any protein–DNA interactions, $\mu_i = \mu_{i0} = c\lambda_i$. It is known that the sum of two independent Poisson random variables $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ follows a Poisson distribution, $\text{Poisson}(\lambda_1 + \lambda_2)$, and conditional on the sum, X, follows a binomial distribution. In other words, $X \,|\, X + Y = n \sim \text{Bin}(n,\ p)$, where $p = \lambda_1/(\lambda_1+\lambda_2)$ (i.e., $\Pr(X = k \mid X + Y = n) = \binom{n}{k} p^k (1-p)^{n-k}$).

Using these results, the number of ChIP reads in a background window conditional on the total number of reads in that window should follow a binomial distribution, i.e., $n_i \,|\, m_i + n_i \sim \text{Bin}(m_i + n_i,\ p_0)$, where $p_0 = c/(1 + c)$ represents the expected proportion of ChIP reads in a background window. If p_0 is known, the enrichment of ChIP reads in any window can be evaluated. This evaluation does not require the knowledge of the actual values of the background sampling rates, λ_i.

In order to estimate p_0, one should keep in mind that the ratio $N/(M + N)$ based on the total read numbers in the two samples is a biased estimate. This is because the ChIP sample contains both background reads and reads that represent signals, whereas p_0 is related only to the background. If we divide the genome into w bp long non-overlapping windows (indexed by i) and assume that read numbers in different windows follow independent Poisson distributions, then $N \sim \text{Poisson}(\sum_i \mu_{i0} + \sum_i \mu_{i1})$ and $M \sim \text{Poisson}(\sum_i \lambda_i)$. As a

result, $N \mid M+N \sim \text{Bin}(M+N,\ q)$, where $q = (c+d)/(1+c+d) \neq c/(1+c)$ and $d = \sum_i \mu_{i1} \big/ \sum_i \lambda_i$. It can be shown that given λ_i, μ_{i1}, and c, the expectation of $N/(M+N)$ is q which is not equal to p_0. An algorithm that estimates p_0 and uses the binomial distribution to evaluate the enrichment of ChIP reads will be described in **Section** 3.2.2. An alternative approach to evaluate background variability for two-sample experiments is discussed in **Note 3**.

1.3. Normalization

The proportionality constant $c = p_0/(1-p_0)$ in the two-sample analysis can be viewed as a way to normalize the read counts of two different samples. This normalizing constant can be used to compute the fold enrichment of ChIP reads, which is defined by (9) as the ratio $(n_i+1)/(cm_i+1)$. Here m_i and n_i are read

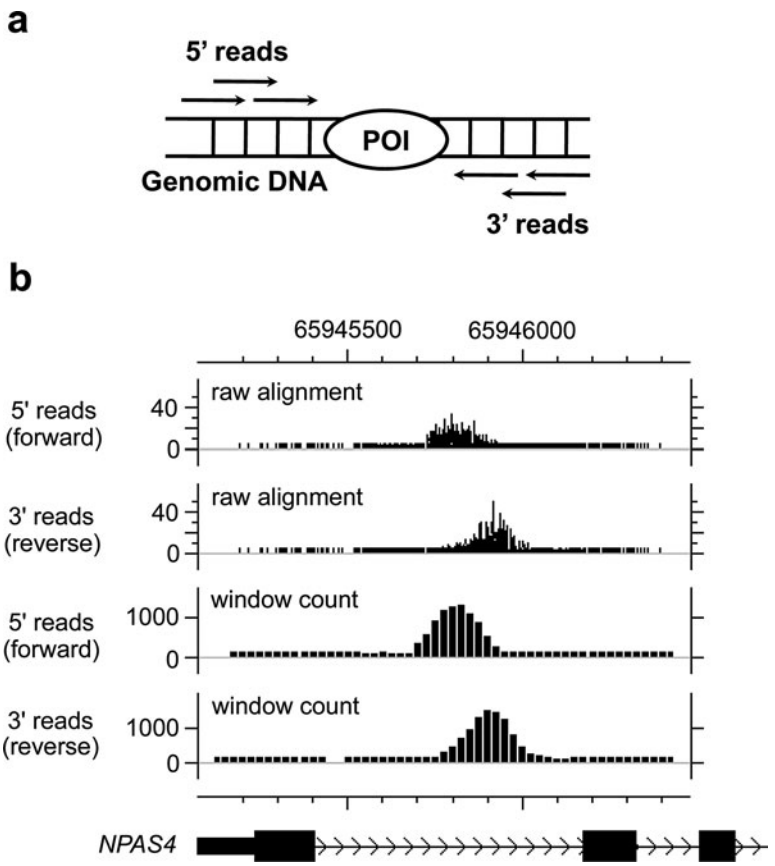

Fig. 9.3. Peak shape for a TFBS. **a** Reads are generated from both ends of DNA fragments. **b** 5′ reads are aligned to the forward strand of the reference genome, and 3′ reads are aligned to the reverse complement strand. These two types of reads form two separate peaks. The binding site is located between the modes of the peaks. From *top* to *bottom*, the four signal tracks are the number of 5′ reads aligned to each genomic position, number of 3′ reads aligned to each position, 5′ read count in a 100 bp sliding window, and 3′ read count in a 100 bp sliding window. The read counts in sliding windows form *smooth curves*. The modes of the curves define boundaries of binding sites.

numbers in the control and ChIP samples in a window indexed by i and a regularization constant one is added to both the numerator and the denominator to avoid dividing by zero.

1.4. Peak Shape

In most current high-throughput sequencing platforms, sequence reads are produced from both ends of DNA fragments. Surrounding a TFBS on the chromosomal map, reads that are aligned to the forward strand of the genome will form a peak upstream of the binding site, and reads that are aligned to the reverse complement strand will form a peak downstream of the binding site (13, 14) (**Fig.** 9.3). This forms a characteristic peak shape that contains useful information for distinguishing bona fide binding sites from false positives. Predicted TFBSs without this bimodal peak shape are often false positives and should be eliminated from the final results. The bimodal shape is also useful for making high-resolution binding site predictions. The bona fide binding site should fit in between the modes of the two peaks. Using this information, a TFBS can usually be mapped to a 50∼100 bp long region (9, 11, 14–16).

2. Software

The methods described in this chapter for building background models and detecting protein–DNA interactions from mapped sequence reads are implemented in the open-source software CisGenome which is available at http://www.biostat.jhsph.edu/~hji/cisgenome (9). CisGenome provides a user-friendly graphic interface and it can also be used to perform various types of subsequent analyses. Sequence reads can be mapped to a reference genome using one of the following software tools: Eland provided by Illumina, Inc., Bowtie at http://bowtie.cbcb.umd.edu (17), MAQ at http://maq.sourceforge.net/ (18), SeqMap at http://biogibbs.stanford.edu/~jiangh/SeqMap/ (19), Corona Lite provided by the Life Technologies (http://solidsoftwaretools.com/gf/project/corona/), and SHRiMP at http://compbio.cs.toronto.edu/shrimp/ (20).

3. Methods

In this section, we describe a procedure to detect protein–DNA interactions from ChIP-seq data. Alternative methods are discussed in **Note 4**.

3.1. Align Sequence Reads

The first step of data analysis is to align sequence reads to a reference genome. A number of software tools have been developed to support fast mapping of millions of short-sequence tags to complex genomes. Examples include Eland (Cox, unpublished), Bowtie (17), MAQ (18), SeqMap (19), and SHRiMP (20). For data generated by the Life Technologies' SOLiD platform, alignment needs to be performed in color-space using tools such as Corona Lite (unpublished) and SHRiMP (20). From now on, we assume that all sequence reads are mapped, and reads that are uniquely aligned to the genome are retained for subsequent analyses.

3.2. Building Background Models

Using the mapped reads, build a background model using CisGenome (9).

3.2.1. Background Model for Analyzing One-Sample Experiments

Divide the genome into non-overlapping windows. The window size w should be chosen to roughly match the expected length of enrichment signals. For TFBS analysis, the window size w is typically set to 100 bp (*see* **Note 5** for more discussions). The entire set of windows can be viewed as a mixture of windows that represent background noise and windows that contain protein–DNA interactions of interest. Let π_0 denote the proportion of background windows. π_0 is unknown and needs to be estimated from the data.

For each window, count the number of reads that are uniquely aligned to the window. Let n_i be the number of reads within the ith window. It is assumed that for background and non-background windows, n_i follows two different probability distributions for which density functions are $f_0(n)$ and $f_1(n)$, respectively. Under this assumption, the data generating distribution for n_i can be described by a mixture distribution $g(n) = \pi_0 f_0(n) + (1 - \pi_0) f_1(n)$. Use the empirical distribution of n_i, i.e., the observed frequencies that $n_i = n(n = 0,\ 1,\ 2, \ldots)$, to estimate $g(n)$.

Based on the discussions in **Section** 1.2.1, the background distribution $f_0(n)$ can be modeled by a negative binomial distribution $NB(\alpha,\ \beta)$. In order to estimate α and β, we assume that windows with small number of reads are mostly background. Under this assumption, the background parameters α and β can be estimated using windows with no more than two reads. For a random variable n that follows negative binomial distribution $NB(\alpha,\ \beta)$, define $r_1 = \Pr(n = 1)/\Pr(n = 0)$ and $r_2 = \Pr(n = 2)/\Pr(n = 1)$. Since $r_1 = \alpha/(\beta + 1)$ and $r_2 = (\alpha + 1)/[2(\beta + 1)]$, we have $\alpha = r_1/(2r_2 - r_1)$ and $\beta = 1/(2r_2 - r_2) - 1$. Therefore, to estimate α and β count the number of windows that contain k reads and denote it as u_k. Use u_1/u_0 to estimate r_1 and use u_2/u_1 to esti-

mate r_2. Plug the estimated values of r_1 and r_2 into $r_1/(2r_2 - r_1)$ and $1/(2r_2 - r_1) - 1$ to obtain the estimates of α and β.

In order to estimate π_0, we assume that most windows with no mapped read represent background noise. Under this assumption, $g(0) \approx \pi_0 f_0(0)$ and $\pi_0 \approx g(0)/f_0(0)$. Therefore, π_0 can be estimated by $u_0 \Big/ [(\sum_k u_k)\hat{f}_0(0)]$. Finally, using the estimated π_0, $f_0(.)$ and $g(.)$, one can estimate the local false discovery rate (local FDR) for any w bp window as follows: $lfdr$ (window i) $= \pi_0 f_0(n_i)/g(n_i)$. Here, n_i is the observed read count for window i.

3.2.2. Background Model for Analyzing Two-Sample Experiments

Divide the genome into w bp long non-overlapping windows. For each window, count the number of reads that are uniquely aligned to the window. For window i, let n_i and m_i denote the number of reads in the ChIP and control samples, respectively, and let $t_i = n_i + m_i$ be the total read count.

Using windows for which t_i is small (we usually use windows that contain only one mapped read, i.e., indices i for which $t_i = 1$), estimate the expected proportion of ChIP reads in background windows as $\hat{p}_0 = \sum_i n_i \Big/ \sum_i (n_i + m_i)$. This implicitly assumes that windows with small read counts mainly represent background. Estimate the normalizing constant $\hat{c} = \hat{p}_0 \Big/ (1 - \hat{p}_0)$.

Next, group windows based on their total read counts t_i. For each group of windows for which $t_i = t (t = 0, 1, 2, \ldots)$, compute the observed frequency that $n_i = n (n = 0, 1, \ldots, t)$. Derive the function $g_{\text{obs}}(n \mid t) =$ {number of windows for which $t_i = t$ and $n_i = n$} / {number of windows for which $t_i = t$}. Define $f_{\text{Bin}}(n \mid t, p_0) = \Pr(X = n)$ where $X \sim \text{Bin}(t, p_0)$. For a window that contains t reads among which n are ChIP reads, estimate the local FDR as $f_{\text{Bin}}(n \mid t, \hat{p}_0)/g_{\text{obs}}(n \mid t)$. When t becomes big, there will be fewer windows available for estimating $g_{\text{obs}}(n \mid t)$. In order to get robust local FDR estimates, if there are fewer than 100 independent windows for a particular t, we suggest extrapolating the local FDR estimates from windows with smaller total read counts. In other words, find the biggest $t' < t$ that has more than 100 windows. For a window that contains t reads and n ChIP reads, the local FDR is estimated as $f_{\text{Bin}}(n' \mid t', \hat{p}_0)/g_{\text{obs}}(n' \mid t')$, where $n' = \lfloor t' n / t \rfloor$ and $\lfloor x \rfloor$ represents the maximal integer that is not bigger than x.

3.3. Detect Protein–DNA Interactions

Using CisGenome (9), scan the reference genome using a w bp long-sliding window. Compute the local FDR for each window. For analyzing a one-sample experiment, use the estimated background model described in **Section** 3.2.1. For analyzing a two-sample experiment, use the procedure described in **Section** 3.2.2. For the two-sample analysis, also compute a fold enrichment for each window: $(n_i + 1)/(\hat{c} m_i + 1)$. Here n_i is the number of ChIP

reads in the window, m_i is the number of control reads, and $\hat{c}$ is the normalizing constant estimated using the method in **Section** 3.2.2.

Select all windows with local FDR smaller than a given cutoff (usually $\leq 10\%$). Merge overlapping windows into a single region. Report all regions obtained after merging. During the process in which windows are merged, use the smallest local FDR among the overlapping windows as the local FDR for the merged region. For the two-sample analysis, use the biggest fold enrichment among all the overlapping windows as the fold change of the merged region.

3.4. Improve Predictions of Transcription Factor Binding Sites

If the purpose of the ChIP-seq experiment is to locate TFBSs, the reported regions should be further processed using CisGenome as follows to improve the results.

3.4.1. Determine the Binding Site Boundary

Use a w bp sliding window to scan each reported region. For each window, count reads in the ChIP sample that are aligned to the forward strand of the genome and those that are aligned to the reverse complement strand. This creates two smooth curves of read counts (**Fig.** 9.3). Identify the locations where the two curves achieve their maxima (i.e., the modes of the curves) and use these locations to define boundaries of binding sites.

3.4.2. Adjust for DNA Fragment Length

For each reported region, compute the distance between the modes of the peaks on the forward and reverse complement strands. Compute the median of all distances and denote it as L. Shift all reads toward the center of the DNA fragments by $L/2$ base pairs. Reads aligned to the forward strand of the genome are shifted toward 3′ of the reference genome and reads aligned to the reverse complement strand are shifted toward 5′ of the reference genome. Using the shifted reads, perform the analyses described in **Sections** 3.2 and 3.3 again. For the reported regions, determine the binding site boundaries using unshifted reads as described in **Section** 3.4.1.

3.5. Subsequent Analyses

Having identified protein-binding regions, they can be analyzed in different ways to study the biological implications. Here we suggest a few common analyses, most of which can be carried out using CisGenome (9). First, compute frequencies that reported regions occur in intragenic and intergenic regions, exons, introns, promoter regions, and other structural features of genes and compute the average level of conservation across species for each region. These two analyses may provide information on functional contexts and importance of the reported regions. Second, extract genes in the neighborhood of the reported regions as a gene set and perform Gene Set Enrichment analysis

(http://www.broadinstitute.org/gsea/) (21) and Gene Ontology analysis (http://www.geneontology.org/GO.tools.shtml). These analyses may provide information on functional categories or pathways that are involved in the biological system in question. Third, perform de novo motif discovery or map the known motifs to the reported transcription factor binding regions and their flanking regions. Identify motifs that are enriched in the binding regions compared to control genomic regions using CisGenome. These analyses may identify motifs that are recognized by the transcription factor in question. They may also suggest collaborating factors. In addition, the motif analysis provides a way to verify that the reported TFBSs are bona fide signals. For example, if the ChIP-seq experiment studies a transcription factor and the binding motif of the transcription factor is known, then the motif is expected to be enriched in the reported binding regions. If this is not the case, it may indicate problems in the ChIP-seq experiment or data analyses. Last but not least, it is always a good idea to visualize the ChIP-seq data along with other structural and functional annotations of the genome. Both the CisGenome Browser and the Genome Browser at UCSC (http://genome.ucsc.edu/) (22) can be used to interactively visualize the data. Interesting patterns may emerge by simply eye balling the data. These patterns may create new hypotheses and suggest future research directions.

4. Notes

1. Analysis of experiments with replicate samples. The methods introduced in this chapter are developed for analyzing experiments that contain a single replicate. If an experiment contains more than one replicates, the analysis can be carried out in two steps. First, merge the replicate data into a combined ChIP sample and a combined control sample (there will be no control sample in a one-sample experiment). The combined sample can then be analyzed using the methods described in **Section** 3. Second, for the reported peaks, extract read counts from individual replicate samples. Normalize the read counts by multiplying the raw read numbers with the normalizing constants obtained using the approach described in **Section** 3.2.2. The normalized read counts can then be analyzed using existing methods developed for detecting differentially expressed genes in microarray experiments (e.g., limma (23)) to remove regions for which the observed ChIP enrichment over the controls can be explained by the random variability among replicates. Suppose that the normalized read counts are saved in a

tab-delimited text file named "data.txt," the R commands below show how limma can be used to perform the analysis in the second step.

```
> library(affy)
> library(limma)
> exprs <- as.matrix(read.table("data.txt",
  header =TRUE,
sep="\t", row.names=1, as.is=TRUE))
> exprs <- log2(exprs)
> eset<-new("ExpressionSet", exprs=exprs)
> design<-cbind(Base=1, ChIP=c(1,1,1,0,0,0)) ##
  3 ChIP vs.
3 controls
> fit<-lmFit(eset,design)
> fit<-eBayes(fit)
```

2. An alternative approach to estimate background in a one-sample experiment. Zhang et al. (15) proposed another approach to estimate the background Poisson rate. To estimate the rate λ_i for a genomic window (usually dozens of base pairs in length), this approach considers a few larger windows (usually 5 and 10 kb in a one-sample analysis) surrounding the window in question. λ_i is estimated using read occurrence rates derived from these larger windows. The underlying assumption of this method is that small windows (with a few dozens of base pairs) close to each other have similar background read sampling rate and reads in the larger surrounding windows are mostly background reads. This is usually a reasonable assumption for analyzing TFBSs. However, it may not hold true in data which contain broad signals or where signals occur at high frequency in the genome. When the assumption is true, this method may provide higher statistical power for detecting signals.
3. An alternative approach to estimate background in a two-sample experiment. Statistical significance of the observed enrichment in the ChIP-control comparison can also be assessed by swapping the sample labels (15). In other words, one treats the ChIP sample as the control and treats the control sample as the ChIP. One then applies the same peak detection procedure to detect "signals" in the label-swapped data. Any "signals" reported in this analysis should represent noise. The false discovery rate for a given enrichment level in the original analysis can be estimated by the ratio {number of regions reported in the label-swapped data}/ {number of regions reported in the original data}. This approach requires that the two samples have about the same number of background reads in order to produce correct FDR estimates. If two samples have different number of reads, a

random subset of reads is usually drawn from the larger sample to create a subsample that has roughly the same number of reads as the other sample. Because this procedure excludes some data from the analysis, it may sacrifice some statistical power. This procedure attempts to match the total number of reads between the two samples, which is not equivalent to matching the number of background reads. In light of discussions in **Section** 1.2.2, this may introduce bias into the FDR estimates. Compared to this approach, the approach described in **Section** 3.2.2 does not require the two samples that have the same read numbers. However, since it depends on assumptions about the underlying data generating distribution, it may produce biased estimates as well if the model assumptions do not hold true in the data.

4. Alternative approaches to detect peaks from ChIP-seq data. Several other methods have been developed for detecting "enrichment peaks" from ChIP-seq data. QuEST (14) (*see* also **Chapter 10**) uses a kernel density estimation approach to build density profiles for forward and reverse reads separately. It then combines the two profiles to detect peaks. FDR is estimated by dividing the control sample into two halves and comparing the two subsets of the control. This requires one to have twice as many reads in the control sample as in the ChIP sample. SISSRs (16) detects points in the genome where the net difference between the forward and reverse read counts in a sliding window switches from positive to negative. It then detects statistically significant binding sites by using a constant rate Poisson model to evaluate the enrichment of the total read counts in the windows surrounding the detected switching points. *MACS* (15) uses a sliding window to scan the genome, and uses a locally estimated Poisson rate to detect enrichment peaks, as discussed in **Note 3**. Other methods include FindPeaks (24), USeq (25), PeakSeq (10), and a ChIP-seq processing pipeline developed by Kharchenko et al. (11). Currently, relative performance of various methods has not been benchmarked. However, for locating TFBSs, all these methods provide similar spatial resolution (a few dozens of base pairs) and the difference among them is subtle compared to the difference between ChIP-chip and ChIP-seq.
5. The choice of window size. The choice of window size w represents a trade-off between sensitivity and specificity. When independent information is available, it may be used to guide the choice of w. For example, in an experiment that locates TFBSs with known motif(s), one can map the motif to the reported binding regions and compute the motif occurrence rates (i.e., the number of motif sites per 1 kb).

The motif occurrence rate is a measure of signal-to-noise ratio. It decreases when the window size becomes too small or too big (9). Motif occurrence rates for regions reported using different window sizes can be compared and the window size that maximizes the rate can be selected to generate the final analysis results. If the transcription factor binding motif is not known before the study, one may first perform de novo motif discovery and use the method described in (26) to identify the motif. It has been shown that the approach described in (26) can correctly identify binding motifs for most genome-wide ChIP studies that involve transcription factors recognizing sequence-specific binding patterns. If one is not able to get the motif information but gene expression data are available, the window size may also be chosen based on what fractions of binding regions are associated with a particular gene expression pattern of interest for different choices of window sizes.

References

1. 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.
2. Robertson, G., Hirst, M., Bainbridge, M. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4, 651–657.
3. Mikkelsen, T.S., Ku, M., Jaffe, D.B. et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.
4. Barski, A., Cuddapah, S., Cui, K. et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837.
5. Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26, 1135–1145.
6. Ren, B., Robert, F., Wyrick, J.J. et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* 290, 2306–2309.
7. Cawley, S., Bekiranov, S., Ng, H.H. et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499–509.
8. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10, 669–680.
9. Ji, H., Jiang, H., Ma, W. et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26, 1293–1300.
10. Rozowsky, J., Euskirchen, G., Auerbach, R.K. et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27, 66–75.
11. Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26, 1351–1359.
12. Marson, A., Levine, S.S., Cole, M.F. et al. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134, 521–533.
13. Schmid, C.D., and Bucher, P. (2007) ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell* 131, 831–832.
14. Valouev, A., Johnson, D.S., Sundquist, A. et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nat Methods* 5, 829–834.
15. Zhang, Y., Liu, T., Meyer, C.A. et al. (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol* 9, R137.
16. Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucleic Acids Res* 36, 5221–5231.
17. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.

18. Li, H., Ruan, J., and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18, 1851–1858.
19. Jiang, H., and Wong, W.H. (2008) SeqMap : mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24, 2395–2396.
20. Rumble, S.M., Lacroute, P., Dalca, A.V. et al. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* 5, e1000386.
21. Subramanian, A., Tamayo, P., Mootha, V.K. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102, 15545–15550.
22. Kent, W.J., Sugnet, C.W., Furey, T.S. et al. (2002) The human genome browser at UCSC. *Genome Res* 12, 996–1006.
23. Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3, 1–9 (Article 3).
24. Fejes, A.P., Robertson, G., Bilenky, M. et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 24, 1729–1730.
25. Nix, D.A., Courdy, S.J., and Boucher, K.M. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-seq peaks. *BMC Bioinformatics* 9, 523.
26. Ji, H., Vokes, S.A., and Wong, W.H. (2006) A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. *Nucleic Acids Res* 34, e146.