

THE ANALYSIS OF CHIP-SEQ DATA

Wenxiu Ma^{*} and Wing Hung Wong^{†,‡}

Contents

1. Introduction	52
2. Planning of ChIP-Seq Experiments	53
2.1. Choices of sequencing platforms	53
2.2. Sequencing statistics and quality control	55
2.3. Saturation	56
2.4. Negative controls	57
2.5. Biological replicates	58
3. Processing and Analyzing ChIP-Seq Datasets	58
3.1. Step 1. Map the reads back to the reference genome	59
3.2. Step 2. Background estimation	60
3.3. Step 3. Peak calling	62
3.4. Step 4. Gene assignment and peak annotation	66
3.5. Step 5. <i>De novo</i> motif analysis	67
4. Discussion	68
Acknowledgment	70
References	71

Abstract

Chromatin immunoprecipitation coupled with ultra-high-throughput parallel DNA sequencing (ChIP-seq) is an effective technology for the investigation of genome-wide protein–DNA interactions. Examples of applications include the studies of RNA polymerases transcription, transcriptional regulation, and histone modifications. The technology provides accurate and high-resolution mapping of the protein–DNA binding loci that are important in the understanding of many processes in development and diseases. Since the introduction of ChIP-seq experiments in 2007, many statistical and computational methods have been developed to support the analysis of the massive datasets from these experiments. However, because of the complex, multistaged analysis workflow, it is still difficult for an experimental investigator to conduct the analysis of his or her own ChIP-seq data. In this chapter, we review the basic

^{*} Department of Computer Science, Stanford University, Stanford, California, USA

[†] Department of Statistics, Stanford University, Stanford, California, USA

[‡] Department of Health Research and Policy, Stanford University, Stanford, California, USA

design of ChIP-seq experiments and provide an in-depth tutorial on how to prepare, to preprocess, and to analyze ChIP-seq datasets. The tutorial is based on a revised version of our software package CisGenome, which was designed to encompass most standard tasks in ChIP-seq data analysis. Relevant statistical and computational issues will be highlighted, discussed, and illustrated by means of real data examples.

1. INTRODUCTION

Chromatin Immunoprecipitation (ChIP) coupled with oligonucleotide hybridization genome tiling array (ChIP-chip) (Carroll *et al.*, 2006; Cawley *et al.*, 2004; Kapranov *et al.*, 2002) and with ultra-high-throughput sequencing (ChIP-seq) (Chen *et al.*, 2008; Johnson *et al.*, 2007; Robertson *et al.*, 2007; Wederell *et al.*, 2008) have been widely used to study transcription factor (TF) regulation in the entire genome. In these experiments, cells are treated with formaldehyde to crosslink DNA-associated protein factors such as TFs or histones to the DNA. The DNA molecules are then randomly sheared to sub-kilobase sized double-strand DNA fragments. TF-bound fragments are targeted by a specific antibody and collected during the immunoprecipitation (IP) process. After the crosslinks between TF and DNA have been reversed, the DNA fragments with length falling within a certain range are selected and amplified. In ChIP-chip experiments, which were popular until recently, these fragments are hybridized to a microarray with millions of 25–75mer probes tiling the whole genome. In contrast, in the ChIP-seq protocol, oligonucleotide linkers or adapters are ligated to the both ends of the ChIP fragments to produce the ChIP-seq library which are then sequenced by the next-generation sequencing machine in a massively parallel manner.

Although the ChIP preparation step is essentially the same in both the ChIP-chip and the ChIP-seq platforms, the subsequent steps of the two approaches are quite different. The fluorescence intensity of each probe is captured and digitalized in the ChIP-chip, whereas raw nucleotide short tags (aka *reads*) are sequenced base-by-base in the ChIP-seq. Moreover, the noise sources are distinct: the major noise in microarray experiments results from the probe affinity effect and the cross-hybridization effect, whereas linker/adaptor contamination, background noise, image processing error, and others all contribute to the ChIP-seq error profile.

Several studies have compared results from ChIP-chip experiments and ChIP-seq experiments on the same TF (Euskirchen *et al.*, 2007; Ji *et al.*, 2008; Robertson *et al.*, 2007). Their analysis revealed that higher sensitivity and sharper resolution of protein–DNA bindings are achieved using ChIP-seq. As sequencing cost continues to decrease rapidly, the ChIP-seq

technique is expected to become more and more dominant in the study of transcriptional regulatory pathways and networks. However, because ChIP-seq datasets are massive and complex, their analysis requires advanced statistical methods, efficient computational algorithms, and user-friendly software for processing and visualization. After a brief discussion of some design issues related to ChIP-seq experiments, we will examine the pipeline of ChIP-seq data analysis step by step (Fig. 3.1). We will illustrate the analysis by using the software CisGenome (Ji *et al.*, 2008) to analyze two datasets (PolII ChIP-seq and STAT1 ChIP-seq) (Rozowsky *et al.*, 2009) (Table 3.1) that were produced as part of the ENCODE project (The ENCODE Project Consortium, 2007).



2. PLANNING OF CHIP-SEQ EXPERIMENTS

The planning of a ChIP-seq experiment involves the consideration of many practical issues: Which sequencing platform to use? How many short reads need to be sequenced? Is a control sample necessary? How to choose and design the control experiment? How many biological replicates are recommended for each ChIP-seq experiment? It is useful to have a brief discussion of these issues before our treatment of data analysis.

2.1. Choices of sequencing platforms

Several commercial ultra-high-throughput sequencing platforms have emerged on the market since 2005. Popular brand names include Illumina Solexa, ABI SOLiD, and Roche 454. Smith *et al.* have compared the accuracy and efficiency of the above three platforms in the study of a mutant strain of *Pichia stipitis* (Smith *et al.*, 2008). The authors found that all three next-generation sequencing platforms successfully identified nucleotide variations between the reference genome and the mutant strains given sufficient coverage. They concluded all three are suitable for accurate and high-throughput sequencing studies. However, there are differences in their error profiles: the primary sequencing error for Illumina Solexa and ABI SOLiD is base substitution, whereas Roche 454 has difficulty in sequencing stretches of repetitive identical bases (*homopolymers*) (Margulies *et al.*, 2005), thus leading to a higher rate of insertion and deletion (indel) errors. In *de novo* assembly study without a reference sequence, longer read platform such as Roche 454 is favored. For the purpose of ChIP-seq experiment, it is common to use either Illumina Solexa or ABI SOLiD because of their ability to deliver a much higher number of sequence reads in parallel.

Table 3.1 Overview of ChIP-seq datasets

Sample name	Replicate	Number of Illumina lanes	Number of raw reads	Number of mappable reads	Number of nonredundant reads	% of nonredundant reads out of mappable reads
IFN- γ STAT1 ChIP	1	4	48,138,968	10,619,323	9,850,217	92.76
	2	2	31,896,881	17,446,502	16,253,194	93.16
	Total	6	80,035,849	28,065,825	25,818,743	91.99
IFN- γ input control	1	6	50,515,792	24,851,515	22,538,851	90.69
PolII ChIP	1	2	15,768,600	8,699,929	7,769,678	89.62
	2	5	21,301,982	11,628,339	10,171,478	87.47
	3	4	20,789,343	10,601,602	9,221,336	86.98
	Total	11	57,859,925	30,899,870	24,822,736	80.33
Input control	1	13	60,452,858	30,827,818	24,602,505	79.81

The sequencing datasets of STAT1 and PolII are obtained from the public gene expression omnibus (GEO) database (GSE12783 series), which is generated as part of the ENCODE project. One dataset contains STAT1 ChIP-seq data on interferon (IFN)- γ simulated human HeLa cells and the total DNA input control on the IFN- γ human HeLa cells; the other dataset contains PolII ChIP-seq data on the unstimulated human HeLa cells and the total DNA input control of the unstimulated cells. There are two biological replicates for the IFN- γ STAT1 ChIP-seq data and three biological replicates for the PolII data. The first 25 bp of the 27–32 bp raw reads are mapped back to the human genome assembly (hg18/NCBI Build 36) obtained from the UCSC Genome Browser. *Mappable reads* are those that map to a unique location in the genome (with up to two mismatches allowed). *Nonredundant reads* are mappable reads that occur only once in the dataset.

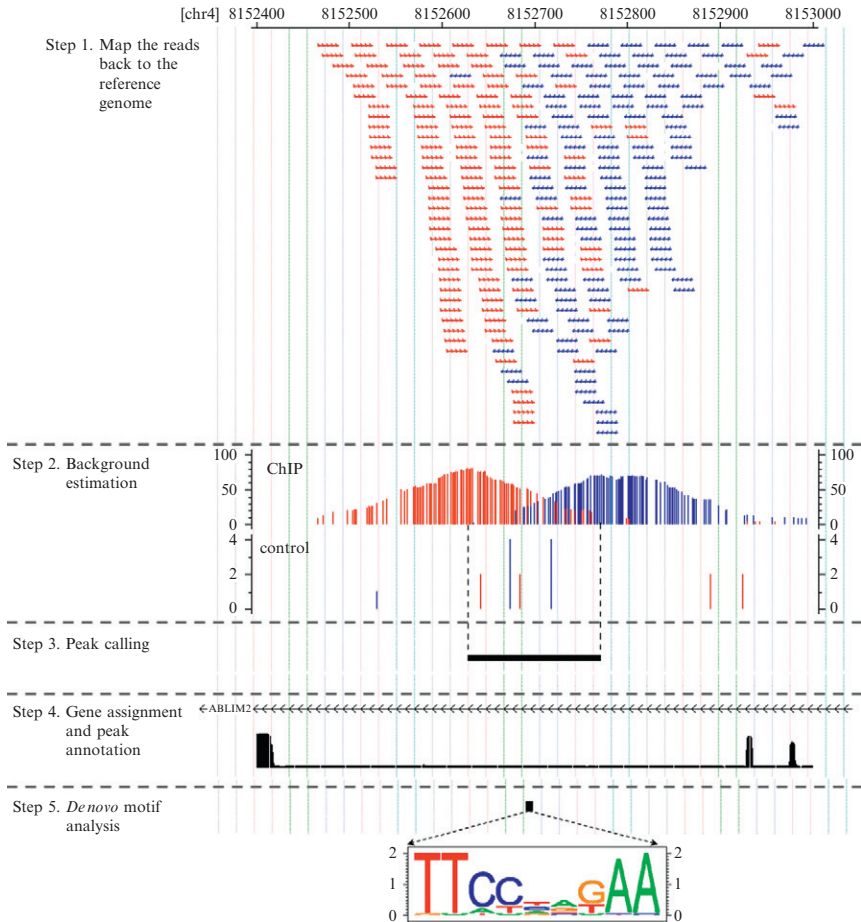


Figure 3.1 Work flow of ChIP-seq data processing and analyzing pipeline.

2.2. Sequencing statistics and quality control

For sequence-specific TF-binding localization and pattern discovery, single-end 25–35 bp reads are commonly used in ChIP-seq studies. We define a *mappable read* as a read that maps (aka aligns) to a unique location in the genome (with up to two mismatches allowed); *nonredundant reads* are the mappable reads that occur only once in the dataset. The goal of a ChIP-seq experiment is to gain an adequate number of mappable reads aggregated at the target regions. At current Solexa Illumina capacity, a single sequencing lane yields tens of millions of short reads and approximately half of them can be uniquely aligned back to the reference genome. In mammalian genomes, a coverage of 10 million reads typically provides clear binding signals at a

large fraction of the binding sites. For a more compact genome such as *Drosophila melanogaster*, one can attain higher signal intensity at the same sequencing depth.

Before any large-scale production run of a ChIP-seq experiment, which may consume valuable biological samples and may require large amounts of reagents and machine time, it is useful to first conduct a pilot experiment run on a single lane (a sequencing machine typically has multiple lanes that can be used in parallel in each run). As a very rough guideline, a successful pilot experiment should meet the following four requirements. First, the number of raw sequencing reads should meet the expected depth within the same sequencing batch. Low yield in one particular lane implies the failure of the antibody or the library preparation. Second, the percentage of uniquely mappable reads achieves at least one-third of the total reads for mammalian genomes. Otherwise, contamination of the library should be suspected and investigated. Third, the percentage of nonredundant reads (two reads are *redundant* if they give identical sequence), should be greater than 50% of the total mappable reads. The nonredundant rate is a powerful measurement of the quality of a sequencing experiment as well as an informative estimation of the saturation status. As the sequencing depth increases, the nonredundant rate will decrease. However, at the first pilot lane of the experiment, we do not expect the saturation to occur so soon, unless the antibody failed to pull down desirable number of protein–DNA complexes. Finally, visual examination should reveal instances of clearly defined peaks with the expected form (see [Section 3.2](#)). The pilot run will provide valuable data for quality control, and will save the experimenter both time and money, should there be a need to troubleshoot.

2.3. Saturation

If the pilot experiment is satisfactory, the next step is to generate more reads in production runs. The *saturation* point of the sequencing depth is defined as the minimum number of reads which would enable the detection of all true protein–DNA binding loci. We suggest an evaluation procedure similar to the one used in Robertson’s paper ([Robertson *et al.*, 2007](#)) to test whether a ChIP-seq dataset is saturated. First, we run the peak calling program on the full dataset. Then, we randomly sample one half of the reads and call peaks from one of the half sets using the sample peak calling parameters. The same set of controls will be used, if applicable.

The peak numbers and quality between results from the full set and from the half set are then compared. Peaks are binned at different false discovery rates (FDRs): for instance, 0.01, 0.05, 0.1, and 0.2; three statistics are calculated for each bin of peaks: (1) the motif site enrichment within the peak regions; (2) the peak conservation score; and (3) the conserved motif

site enrichment. If these statistics improve when we double the size of the reads, it demonstrates that the data have not yet reached the saturation point. In this case, obtaining more reads will definitely be helpful.

2.4. Negative controls

In addition to sequence reads from positive ChIP samples, it is recommended that sequence reads be generated also from negative control samples. Use of negative controls can significantly increase the sensitivity and specificity of the peak detection (Ji *et al.*, 2008; Rozowsky *et al.*, 2009). Negative controls that are commonly employed in ChIP-seq experiments can be classified into three types. The first type is total DNA input control, where non-IP'ed DNA is sheared, size-selected, and sequenced. The second type is Mock IP control, where we use a nonspecific antibody, for example, the immunoglobulin G (IgG) antibody on the same cells. The third type includes all specially designed controls. For instance, if the ChIP is performed on simulated cells, then using the same antibody on unstimulated cells is a good negative control. Also, in some studies, when the antibody of the TF of interest is not available or the antibody affinity is not strong enough, we might use FLAG or other epitope tags in the IP step. In this case, utilizing the FLAG antibody on un-FLAGed cells will provide a good negative control.

Specifically, negative controls are important for several reasons. First, it provides a background distribution to aid the FDR estimation. In this viewpoint, the input control is thought to be a better control than the mock IP because the reads of the input control will have a more balanced distribution throughout the genome. Whereas the reads of the mock IP control constitute numerous repetitive reads sequenced from the DNA fragments that the antibody pulled off, therefore, leaving fewer randomly distributed reads for background estimation. Furthermore, negative controls provide a means for us to indentify the genomic regions that are expected to have more reads for reasons largely unrelated to the binding of the TF of interest. For example, data from Illumina Solexa sequencer may exhibit a bias toward the GC-rich sequences (Dohm *et al.*, 2008). In addition, Rozowsky *et al.* discovered that input controls have small peaks in transcribed regions especially near transcription start sites (TSSs), because the chromatin tends to be more open at these regions (Rozowsky *et al.*, 2009). Finally, specially designed negative controls will aid in the detection of abnormal associations between the antibody and the DNA sequences. In the situation of the FLAG antibody, the negative control that uses the FLAG antibody on un-FLAGed cells is the best control for detecting potential bindings between the FLAG antibody and the DNA sequences.

2.5. Biological replicates

Although ChIP-seq data are believed to be much less noisy as compared to ChIP-chip data, it is still important to use multiple biological samples whenever it is possible. Having biological replicates helps to reduce sample-specific or sequence-specific biases, which can be caused by a variety of reasons, such as antibody affinities, sonication and amplification variations, library contaminations, and sequencing errors.

If the variability in the replicates is unacceptably large, for example, if the set of peaks detected are largely inconsistent, then one may need to improve the experimental protocol and repeat the experiments to obtain acceptable data. Sometimes, if there are enough replicates and only one of them is inconsistent with the others, then it may be reasonable to proceed with the analysis after removing the outlier sample. In any case, after we have obtained replicate samples with largely consistent results, then we are still faced with the question of how to combine the information in the replicates in the subsequent steps of the analysis.

In this tutorial, we handle replicates by a simple procedure, which calls peaks from individual biological replicate separately and then intersects them to obtain the final common peak regions. We choose this intersecting approach because we found it to be a safe and effective way to detect robust and reproducible binding events when there are noticeable differences among the biological replicates. An alternative and common practice is to pool all reads from multiple biological replicates together and call peak regions from pooled data. For example, Rozowsky *et al.* sampled the same number of reads from each replicate and combined them to proceed to further analysis (Rozowsky *et al.*, 2009). However, the sampling method or the linear scaling method might not be effective in some dataset, because the ratio of signal intensities between biological replicates is not always equal to the ratio of sequencing depths (Fig. 3.2). Thus, a more rigorous statistical method to address the multiple sample normalization and the multiple sample consistency testing problem is desired.

3. PROCESSING AND ANALYZING CHIP-SEQ DATASETS

In this section, we demonstrate step by step how to use a revised version of the CisGenome software suite (Ji *et al.*, 2008) to process and analyze two published ChIP-seq datasets (for STAT1 and PolII) (Fig. 3.1). In this revised version, we designed and implemented an improved peak calling procedure based on the use of an iterative background count estimation technique. The new peak caller is described in more details in Section 3.3

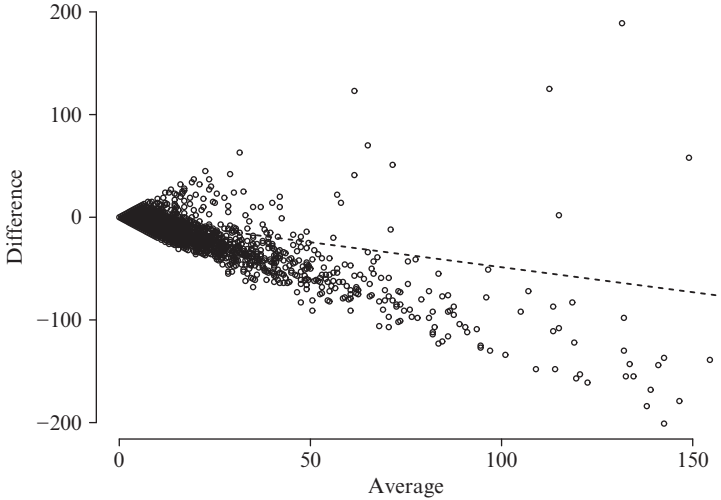


Figure 3.2 *M–A plot of two STAT1 biological replicates.* The entire human genome is divided into 100 bp nonoverlapping windows and the number of reads from the two STAT1 biological replicates in each window is counted separately. The M–A plot is drawn for only the window counts on chromosome 1. X-axis: the average number of reads between the two STAT1 biological replicates in each window. Y-axis: the difference between the number of reads from the first STAT1 biological replicate and the number of reads from the second STAT1 replicate in each window. The black-dashed line represents the relation between the sequencing depths of the two replicates. This figure demonstrated that the differences between the two biological samples are more significant than the differences between the sequencing depths.

below. Although our discussion is illustrated by the CisGenome system, there are other software tools available which may be used to accomplish similar analyses (Boyle *et al.*, 2008; Fejes *et al.*, 2008; Johnson *et al.*, 2007; Jothi *et al.*, 2008; Kharchenko *et al.*, 2008; Rozowsky *et al.*, 2009; Tuteja *et al.*, 2009; Valouev *et al.*, 2008; Zhang *et al.*, 2008).

3.1. Step 1. Map the reads back to the reference genome

Almost always, the first step in a ChIP-seq data analysis is the mapping of reads to a reference genome. In this step our goal is to identify, for each short read in the dataset, all the locations in a reference genome that show perfect or near perfect (say with no more than two mismatches in a 25-bp read) matches to the read (Fig. 3.1, Step 1).

There are quite a few programs available to map the short reads back to the reference genome (Jiang and Wong, 2008; Langmead *et al.*, 2009;

Li *et al.*, 2008, 2009). Mapping is a straightforward task because there is only one correct result, given the raw reads, the reference genome assembly, and the number of mismatches allowed. Therefore, it does not matter which short-read mapping program is used. The differences among these software tools lay primarily on the algorithm designs and computational efficiencies. A couple of them have add-on features: Bowtie (Langmead *et al.*, 2009) is one of the fastest short-read mapping program; Maq (Li *et al.*, 2008) can leverage on the reads quality scores; and SeqMap (Jiang and Wong, 2008) considers insertions and deletions (indels).

However, there is a tradeoff between the length of the reads to be used in the mapping and the yield of uniquely mappable reads. We usually do not use the full length of the reads in the mapping step, primarily because the sequencing error at each base increases rapidly near the end of the read. For example, in the STAT1 ChIP-seq data, 27–28 bp reads were sequenced. The error rate at positions 26–28 is much higher than that in positions 10–12 in a typical sequencing lane (Fig. 3.3). For ChIP-seq, it is standard to use the first 25 bases of the raw reads to map back to the reference genome assembly. Considering the fact that the human genome is highly repetitive (same for the other mammalian genomes), theoretically about 75% of the human genome can be uniquely mapped using 25 bp reads within two mismatches (McKernan *et al.*, 2009). A mappability profile that counts the redundancy of a read beginning at each nucleotide position on the genome has been calculated to improve peak detection accuracy and specificity (Rozowsky *et al.*, 2009).

The mapping results and statistics of the PolII and STAT1 dataset are listed in Table 3.1. The original reads have varying length 27–35 bp. Here, we remapped the first 25 bases of each read to the human genome assembly (hg18/NCBI Build 36) obtained from the UCSC Genome Browser Database (Rhead *et al.*, 2010; The Genome Sequencing Consortium, 2001) using the SeqMap software (Jiang and Wong, 2008). From the mapping statistics, we can deduce that this dataset is a deeply sequenced, comprehensive, and not highly redundant.

3.2. Step 2. Background estimation

Based on the ChIP-seq protocol and technique, ideally, all reads should be sequenced from the ends of the ChIP fragments that were bound by the target TF. However, in any ChIP-seq datasets, a considerable fraction of the reads may not have originated from these ChIP fragments. For instance, the antibody might target proteins other than the one studied, therefore capturing nonspecific fragments. Other factors that may induce such extraneous reads include library contamination, PCR amplification selection,

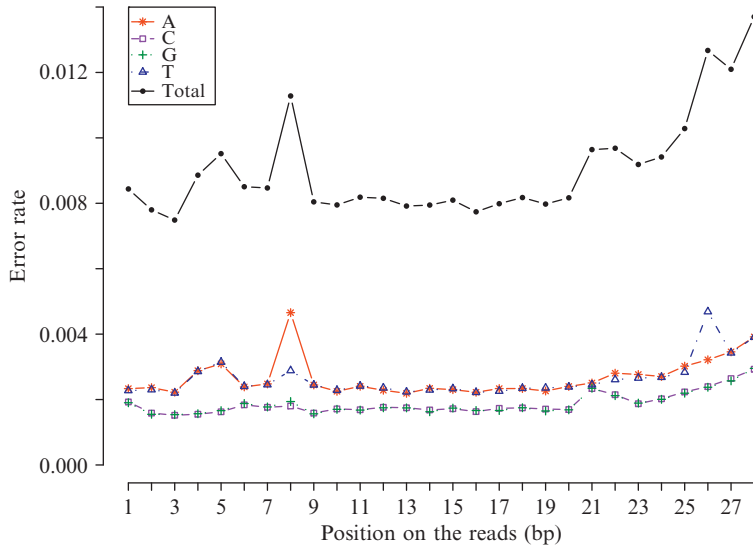


Figure 3.3 Sequencing error rate comparing to the reference genome. The 28-bp reads from land B, replicate 1 of the STAT1 ChIP-seq dataset are mapped back to the human genome (hg18/NCBI Build 36). For each uniquely mappable (up to two mismatches) read, we count each position on the read which is different from the reference genome. X-axis: nucleotide position on the short reads (1–28 bases); Y-axis: the error rate comparing to the total number of uniquely mappable reads in that lane. Asterisks on solid line: the error rate for nucleotide A (the reference genome has A in that position but the sequencing read has C/G/T); boxes on dashed line: C; crosses on dotted line: G; triangles on dot-dashed line: T; and finally, points on solid line: the addition of these four types of errors.

linker/adaptor contamination, and image processing errors. What is more, because of sequencing errors, a read that originated in one location of the genome may be uniquely mapped to a different location of the genome that has sequence similarity with the original source.

We can regard all these reads that are unrelated to the binding events of interest, as “background reads” in our ChIP-seq experiment. It is not easy to ascertain whether each read results from a true binding event in the cells or from background. However, we can attempt to estimate the rate of occurrence of the background reads. Knowledge on the background rate is important for the assessment of the statistical significance of the binding regions detected by the peak caller (see Section 3.3). For each ChIP-seq sample, we define the background rate to be the ratio of the number of background reads to the total number of reads in the sample. We call a read a *true signal read* if it falls into the called peak regions (from the actual

binding fragments). Otherwise, we call it a *background read*. Since the estimation of the background rate and the detection of the peak regions are dependent on each other, we propose an iterative method to solve this problem (see Section 3.3).

In other words, the peak calling problem is a signal-over-noise detection problem. We can take advantage of classical methods and measuring standards in the signal-over-noise problem. At the global scale, we use the background rate to estimate the ratio of the ChIP signal intensity to the landscape intensity; in each called peak region, we take the fold change of the ChIP signal intensity to the control signal intensity as the local estimate of the signal-to-noise ratio.

3.3. Step 3. Peak calling

In this step, we discuss the most critical task in the ChIP-seq data analysis pipeline. This is to identify the ChIP signal enriched genomic regions. In other words, where did the TF bind? This process is referred to as *peak calling* because the count of reads from the same strand of DNA (Watson or Crick) in a TF-bound fragment should show a peak near the binding location (Fig. 3.1, Step 2).

If a protein factor has a sharply focused binding site, in a successful experiment, one should be able to observe the bi-horned peaks nice bell-shaped peaks will be shaped at both the Watson strand and the Crick strand. This is because a fragment is always sequenced from its ends toward its midpoint. A Watson read represents the 5'-end of a ChIP fragment, whereas a Crick read represents the 3'-end. Therefore, the Watson peaks (as the left red peak in Fig. 3.1) and the Crick peaks (as the right blue peak in Fig. 3.1) are located in the opposite sides of the TF-binding site (TFBS). Thus the two peaks may be used to define a candidate binding region.

In contrast, if the ChIP-seq experiment fails because of the weak affinity of the antibody, extremely high, block-shaped peaks at repetitive regions or generally flat signal across the entire genome would be observed.

As depicted in Fig. 3.1, the mapped reads and the signal profile (defined below) can be visualized in a genome browser. Here, we use the CisGenome Browser (Jiang *et al.*, 2010) that is integrated with the CisGenome software, but other genome browsers, such as the UCSC Genome Browser (Kent *et al.*, 2002) and the Affymetrix Integrated Genome Browser (Nicol *et al.*, 2009) can also be used after we export the suitable files from CisGenome. To obtain the signal profile, we use a fixed window size w and count the number of the Watson and the Crick reads that fall into each nonoverlapping window along the entire genome. Window size $w = 100$ is recommended for sequence-specific TF-binding ChIP-seq data (Ji *et al.*, 2008).

In the signal profile track, we can see that the Watson reads and the Crick reads are clearly separated. Some programs shift the Watson and Crick reads toward their mid-points, and then detect peak regions by combining

the shifted reads (Ji *et al.*, 2008; Kharchenko *et al.*, 2008; Valouev *et al.*, 2008). Kharchenko *et al.* used the cross-correlation magnitude to find the optimal shifting distance, while Ji *et al.* used half of the average peak length to shift. An alternative strategy, which we have used to identify the candidate region marked by the black line in the third step of Fig. 3.1, is to call peaks from the Watson strand and from the Crick strand separately, and then regard the region bracketed by the two strand-specific peak locations as the candidate TF-binding region. One advantage of this strategy is that we can easily check the balance between the numbers of reads in the coupled peaks. Moreover, since the length of the ChIP fragments varies greatly among peak regions, using a fixed global shifting distance may not be sufficient.

The first ChIP-seq peak caller, implemented in (Johnson *et al.*, 2007), was an intuitive, *ad hoc* method. It arbitrarily decides a genome-wide cutoff of signal intensity, and defines peaks as the regions above the predetermined cutoff level. A limitation of this method is that it does not provide significance and ranking of the detected peak regions. Similar to Johnson's approach, there are numerous other tag-aggregation methods in which the ranking of the peaks is based solely on the number of tags assembled in each peak region. The rationale for this ranking rule is based on the assumption that the height of a peak is a linear function of the proportion of cells that have the TF bound in the peak locus. Thus, the most significant binding will produce the highest peak. Nevertheless, this assumption is not always true. Certain chromatin regions are open and, therefore, easily fragmented, leading to stronger peaks in or around the transcribed gene neighborhood (Rozowsky *et al.*, 2009). In addition, if the TF binds to multiple locations close to each other, the multiple peak signal strength will add together and shape a broader, stronger, and continuous peak.

To improve the above fixed-cutoff method, a more advanced peak calling program may first estimate the background distribution and then use it to help assess statistical significance of the peaks.

Commonly used distributions for the ChIP-seq background counts include the Poisson distribution (Ji *et al.*, 2008; Zhang *et al.*, 2008) and the negative binomial distribution (Ji *et al.*, 2008). Comparing to the Poisson distribution, the negative binomial distribution is a better fit to the ChIP-seq background reads distribution (Ji *et al.*, 2008).

When a negative control sample is available, the same peak caller can be performed on both the ChIP sample and the control sample. A simple way to filter out false peaks in the control samples is to require a minimum fold change of the ChIP signal to the control signal (Johnson *et al.*, 2007; Valouev *et al.*, 2008). A more statistically rigorous approach to this two-sample problem is implemented in the CisGenome peak caller (Ji *et al.*, 2008). For each read in a given genomic window, we regard it as a *success* if it is from the ChIP sample; a *failure* if from the negative control sample. Thus, given k_{1i} is the number of ChIP reads in the window and k_{2i} is the

number of control reads, the number of ChIP reads in that window follows a conditional binomial model, that is, $k_{1i}|n_i \sim \text{binomial}(n_i, p_0)$, where $n_i = k_{1i} + k_{2i}$ and p_0 is the probability of seeing a ChIP read in that window.

In CisGenome, Ji *et al.* divided the entire genome into nonoverlapping windows of w -bp width and counted the number of ChIP reads and the number of negative control reads in each window. Then they took the ratio of the number of windows that contain only one ChIP read and the number of windows that contain only one read (either a ChIP read or a control read) as the estimation of p_0 . We have found that this estimation method sometimes failed when a ChIP-seq sample was highly redundant (either because the sample is over-saturated or because the antibody failed) and had very few windows containing only one read, thus leading to a biased estimation of p_0 .

To deal with this problem, we have recently developed an improved version of this conditional binomial approach. The main idea is to estimate the expected success rate p_0 in an iterative manner. Assuming that background reads are uniformly distributed, r_0 is the ratio of the probability of seeing a ChIP read to the probability of seeing a control read at any genomic position, that is: $r_0 = (\text{number of background reads in a ChIP sample})/(\text{number of background reads in control})$. Similarly, it is assumed that the number of ChIP reads within a sliding window of w -bp follows a conditional binomial model, that is, $k_{1i}|n_i \sim \text{binomial}(n_i, p_0)$, where $p_0 = r_0/(1 + r_0)$. The program starts with r_0 equals to the total number of reads in the ChIP sample divided by the total number of reads in the control sample ($r_0 = \text{total number of ChIP reads}/\text{total number of control reads}$). The program identifies peaks using the initial estimation of r_0 , and then filters out the ChIP and control reads that fall into peak regions. Once the peak regions have been filtered out, r_0 is reestimated and iterations continue until r_0 converges.

This peak caller program is applied to the Watson and the Crick reads separately. After all of the peaks have been identified, the Watson and Crick strand peaks are combined. Only those peaks containing a balanced number of Watson and Crick reads are paired. The peak boundaries are set as the modes of the coupled peaks. The fold change between the ChIP signal and the control signal is also calculated for each peak region. An example of output file of the top 20 STAT1 peaks is displayed in [Table 3.2](#).

In addition to acquiring a set of peak regions, we are also interested in the significances of the peak regions. In CisGenome, we calculate the FDR based on the read distribution in both the ChIP and control samples. To be more specific, the FDR of each w -bp window with k_{1i} ChIP reads and n_i total reads, is the ratio of the expected number of windows that have equal to or more than k_{1i} ChIP reads out of n_i total reads given p_0 , divided by the observed number of such windows (see Methods in [Ji, *et al.*, 2008](#)). The better the background estimation fits the data, the more accurate is the FDR estimation.

In our case study, two-sample peak calling is performed on each biological replicate of the interferon- γ (IFN- γ) STAT1 ChIP versus

Table 3.2 Top 20 STAT1 peaks

Rank	Chromosome	Start	End	Length	Peak height	Number of ChIP reads	Number of control reads	Fold change of ChIP/control
1	chr20	48342552	48342733	182	1328.25	5698	183	31.05
2	chr2	191593263	191593466	204	1204	4606	84	54.84
3	chr14	23700077	23700317	241	1124	5198	102	50.96
4	chr6	30565062	30565269	208	1077.75	3720	65	56.8
5	chr15	42808242	42808406	165	985.25	4447	104	42.76
6	chr5	131854389	131854584	196	956.25	3506	95	36.72
7	chr5	131860571	131860746	176	848.25	6768	220	30.76
8	chr16	55580792	55580988	197	815.25	2897	112	25.75
9	chr16	18845403	18845625	223	771.5	2283	46	49.64
10	chr12	107546406	107546594	189	741.25	3047	87	34.83
11	chr19	10242625	10242835	211	685.25	2505	77	32.33
12	chr17	37794087	37794446	360	662.25	3649	115	31.59
13	chr1	148801113	148801409	297	658.5	3301	109	30.15
14	chr12	47532339	47532527	189	656	3339	121	27.48
15	chr5	43076129	43076301	173	613.75	2995	133	22.44
16	chr16	10830348	10830537	190	608.75	2415	71	33.78
17	chr17	55218675	55218884	210	583.5	2373	49	48.43
18	chr3	126327407	126327607	201	582.25	2258	55	40.68
19	chr7	101493956	101494129	174	565.25	2273	74	30.72
20	chr16	28451190	28451421	232	562.25	2741	58	46.86

The peak regions are ranked by the peak height, which is the average of the maximum numbers of reads in a 100-bp window on the Watson and Crick strands. Alternatively, we can rank the peak regions by the last column, which is the fold change of ChIP signal intensity to the control signal intensity. The peak start is the mode of the Watson peak; the peak end is the mode of the Crick peak. The number of ChIP (or control) reads is the average count in the coupled Watson and Crick peaks.

IFN- γ input control comparison and also each biological replicate of the PolII ChIP versus input control comparison. Peak numbers, the percentage of ChIP reads fell into peak regions, and the ChIP/input signal fold change enrichment in the called peak regions (normalized by the background ratio between the ChIP and the input control samples) at the FDR 0.01 cutoff are provided as output (Table 3.3). For sequence-specific transcriptional factor ChIP-seq data sequenced at a depth of 10 million mappable reads in mammalian genomes, we expect to see at least 3% of ChIP reads originating from the binding peaks (varies from factors and samples) and a ChIP/input fold change above 5.0. After peak detection on each biological replicate, we intersect peak regions of individual replicate to get the common regions as the final list. Finally, we have 3347 reproducible STAT1 peaks and 9087 PolII peaks. Because of the noticeable differences among the biological replicates, using this stringent intersection approach, we get 54–86% fewer peaks than the published peaks in the Rozowsky *et al.*'s paper.

3.4. Step 4. Gene assignment and peak annotation

After we obtain a list of peak coordinates, it is important to study the biological implications of the protein–DNA bindings. Certain questions have always been asked: what are the genomic annotations and the functions of these peak regions?

Because many cis-regulatory elements are close to TSSs of their targets, by default CisGenome associates each peak to its nearest gene, either upstream or downstream. In our example dataset, the PolII peaks are closer

Table 3.3 Summary of peak calling results of STAT1 and PolII

	Replicate 1	Replicate 2	
<i>IFN-γ STAT1 versus IFN-γ input control peak calling</i>			
Number of peaks	3,822	14,644	
Percentage of ChIP reads fell into peak regions	4.04	11.54	
ChIP/input (normalized) signal ratio in peak regions	11.09	9.48	
	Replicate 1	Replicate 2	Replicate 3
<i>PolII versus input control peak calling</i>			
Number of peaks	16,893	21,585	22,296
Percentage of ChIP reads fell into peak regions	27.14	48.71	51.15
ChIP/input (normalized) signal ratio in peak regions	15.51	20.00	22.10

For each called peak list, we calculated the percentage of ChIP sample reads fell into the called peak regions and the ChIP/input signal fold change in the peak regions (normalized by the background read ratios in both samples).

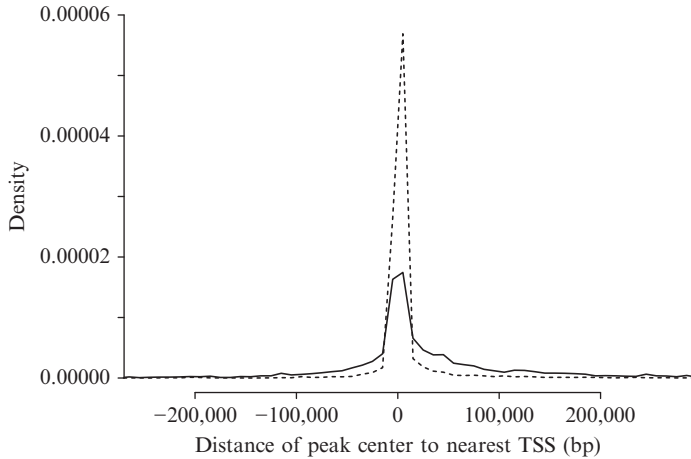


Figure 3.4 *Density of peak locations relative to nearest genes.* X-axis: distance from the peak center to the nearest transcription start site (TSS), where 0 is the position of the TSS. Negative numbers: 5' of TSS. Positive numbers: 3' of TSS. Only distance from -250 to 250 kbp is displayed in this figure. Y-axis: probability of seeing a peak within each bin of 10 kbp (percentage of peaks in each bin/bin size). Solid line: 3347 STAT1 peaks. Dashed line: 9087 PolII peaks.

to the nearest TSSs than the STAT1 peaks (Fig. 3.4). The STAT1 peaks are more enriched in the intron regions and the intergenic regions (Table 3.4), which is consistent with its known cis-regulatory function.

In CisGenome, we can also associate peaks with all genes in their neighborhood. This function is more informative and helpful in screening potential cis-regulatory targets, especially when transcription profiling data under the same cellular condition are available. As a different approach, Ouyang *et al.* (2009) computed a weighted average of peak signals detected near a gene and regard it as a quantitative score for the strength of the association between the TF and the gene. The authors showed that such scores can be used to build good predictive models of the absolute gene expression in mouse embryonic stem cells.

In the CisGenome Browser, the genetic landscape around each peak can be displayed at any resolution. Other genome features such as phylogenetic conservation can also be added to the visualization (Fig. 3.1, Step 4).

3.5. Step 5. *De novo* motif analysis

Another important task in the analysis of the predicted peak regions is *de novo* motif discovery. In some studies, the exact sequence to which the TF binds is known, or even better, a set of validated binding sites is available. However, if this information is not available, we will need to recover the binding motifs from the peak sequences as well as from their

Table 3.4 Genomic location of peak regions comparing to the nearest genes

Peak location	3347 STAT1 peaks		9087 PolII peaks	
	#	%	#	%
IntraGenic	1430	42.72	5505	60.58
5'-UTR	143	4.27	2288	25.18
3'-UTR	32	0.96	167	1.84
CDS	40	1.20	936	10.30
Intron	1215	36.30	2114	23.26
Exon	215	6.42	3391	37.32
InterGenic	1917	57.28	3582	39.42
Upstream	1237	36.96	2849	31.35
Downstream	680	20.32	733	8.07

IntraGenic: transcribed region of a gene; 5'-UTR: 5'-untranslated region; 3'-UTR: 3'-untranslated region; CDS: coding sequence; InterGenic: outside the transcribed gene regions.

orthologous sequences. CisGenome has incorporated a Gibbs sampling module (Lawrence *et al.*, 1993; Liu, 1994) which can recover enriched motifs from the sequences of the peak regions.



In CiGenome Browser, we can visualize the top motif logos. The degree of consistency between the known or published motif and the *de novo* discovered motif can be used to assess the success of the experiment. Motif occupancy and enrichment in peak regions and motif conservation scores offer additional means for assessments.

Three independent runs of Gibbs sampler are performed on the FDR 0.01 STAT1 peak regions. Top enriched *de novo* motifs include the canonical STAT1 motif and the activating protein 1 (AP-1) motif (Table 3.5). There are 2544 (76.01%) of the 3347 STAT1 ChIP-Seq peaks that contained one or more STAT1 *de novo* recovered motif sites within the peak boundaries. The STAT1 motif sites are close to the peak center (Fig. 3.5A). About 33.49% of these STAT1-containing peaks have conserved STAT1 motif sites which are located within the top 10% conserved genomic regions (conservation scores of the 44-vertebrate alignment phastCons scores for the human hg18 genome are obtained from the UCSC Genome Browser Database (Rhead *et al.*, 2010; Siepel *et al.*, 2005)). On average, the conservation scores for the motif sites are significantly higher than their neighborhood regions (Fig. 3.5B).

4. DISCUSSION

In summary, we have provided a systematic discussion of issues related to the analysis of ChIP-seq data. We demonstrated how several key steps, including data exploration and visualization, peak calling, genomic

Table 3.5 Enrichments of *de novo* discovered motifs

Motif name	Motif logo	Enrichment in 3347 STAT1 peaks		
		r_1	r_2	r_3
STAT1	 <p>The motif logo for STAT1 shows a sequence of 9 nucleotides: T, T, C, C, G, A, A. The first two positions (TT) are highly conserved, indicated by tall red bars. The third and fourth positions (CC) are also conserved, shown with blue bars. The fifth position (G) has a shorter orange bar, and the sixth and seventh positions (AA) have tall green bars. Below the sequence, there are small colored bars representing enrichment at each position.</p>	13.57	18.63	24.35
AP-1	 <p>The motif logo for AP-1 shows a sequence of 6 nucleotides: T, G, A, T, C, A. The first position (T) has a tall red bar, the second (G) has a tall orange bar, and the third (A) has a tall green bar. The fourth position (T) has a tall red bar, the fifth (C) has a blue bar, and the sixth (A) has a tall green bar. Below the sequence, there are small colored bars representing enrichment at each position.</p>	4.33	6.10	7.96

Three motif enrichment ratios are calculated as described in [Ji *et al.*'s \(2006\) paper](#): r_1 = percentage of peak regions containing the indicated motif(s) versus percentage of matched control regions with the same motif(s); r_2 = percentage of phylogenetically conserved peak regions containing the motif(s) versus percentage of phylogenetically conserved matched control regions with the same motif(s); r_3 = percentage of regions containing the phylogenetically conserved motif(s) versus percentage of matched control regions with the phylogenetically conserved same motif(s). Matched control regions are randomly selected such that the distance between the control region and the closest transcription start site (TSS) has the same probability distribution as the distance between the peaks region and the TSS. For TF ChIP-seq peaks, we expect the r_1 , r_2 , and r_3 to be simultaneously greater than 5.0 for the primary binding factor, and to be simultaneously greater than 2.0 for other collaborating binding factors.

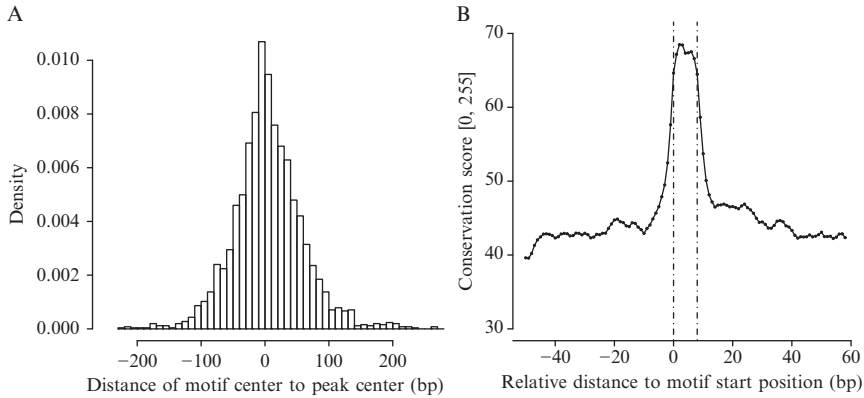


Figure 3.5 Motif analysis on the *STAT1* peak regions. (A) Histogram of *STAT1* *de novo* motif resolution. *X*-axis: distance from the center of *STAT1* recovered motifs to the peak center. *Y*-axis: density (percentage of motifs) per bin. Bin size: 10 bp. *STAT1* motifs are most often located close to the peak center. (B) Average conservation scores for motif sites and flanking positions. *X*-axis: distance to the motif start position, where 0 is the first base of the motif site. Negative numbers: 5' of the motif sites. Positive numbers: 3' of the motif sites. The flanking region of 50 bp on either side is displayed in this figure. *Y*-axis: converted UCSC phastCon scores of 44 vertebrate alignments to the hg18 human genome. The converted conservation score is ranging from 0 (least conserved) to 255 (most conserved). Black solid line: 3218 *STAT1* sites within peak regions. Black dashed line: *STAT1* motif sites boundaries.

annotation, and downstream motif analyses, can be accomplished by a user-friendly software package CisGenome. We rely on other specialized software kits for the low-level analyses, such as base calling, image processing, error filtering, and so on.

The example datasets include TF ChIP-seq and polymerases transcription ChIP-seq. Finally, our analysis pipeline can be extended to analyze histone modification ChIP-seq dataset. For such applications, some minor modifications on the peak calling algorithms have to be made, including enlarging window size for data exploration and background estimation, and shifting/coupling strategies on the Watson and Crick strands.

The CisGenome software is available at <http://www.biostat.jhsph.edu/~hji/cisgenome/>.

ACKNOWLEDGMENT

This chapter is based on research supported by NIH Grants R01HG004634 and R01HG003903.

REFERENCES

- 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.
- Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoutte, J., Brodsky, A. S., Keeton, E. K., Fertuck, K. C., Hall, G. F., Wang, Q., Bekiranov, S., et al. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289–1297.
- Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., 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.
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117.
- Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105.
- Euskirchen, G. M., Rozowsky, J. S., Wei, C. L., Lee, W. H., Zhang, Z. D., Hartman, S., Emanuelson, O., Stolc, V., Weissman, S., Gerstein, M. B., Ruan, Y., and Snyder, M. (2007). Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. *Genome Res.* **17**, 898–909.
- Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. (2008). FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730.
- 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**(21), e146.
- Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M., and Wong, W. H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300.
- Jiang, H., and Wong, W. H. (2008). SeqMap: Mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396.
- Jiang, H., Wang, F., Dyer, N. P., and Wong, W. H. (2010). CisGenome Browser: A flexible tool for genomic data visualization. *Bioinformatics* **26**, 1781–1782.
- 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.
- 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.
- Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* **12**, 996–1006.
- 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.
- 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.

- Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.
- 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.
- Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., and Wang, J. (2009). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967.
- Liu, J. S. (1994). The collapsed Gibbs sampler with applications to a gene regulation problem. *J. Am. Stat. Assoc.* **89**, 958–966.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
- McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., *et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541.
- Nicol, J. W., Helt, G. A., Blanchard, S. G., Jr., Raja, A., and Loraine, A. E. (2009). The Integrated Genome Browser: Free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730–2731.
- Ouyang, Z., Zhou, Q., and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **106**, 21521–21526.
- Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., *et al.* (2010). The UCSC Genome Browser database: Update 2010. *Nucleic Acids Res.* **38**, D613–D619.
- Robertson, G., Hirst, M., Bainbridge, M.,ilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657.
- Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75.
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050.
- Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P., Stewart, D. A., Zhang, L., *et al.* (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* **18**, 1638–1642.
- The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816.
- The Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Tuteja, G., White, P., Schug, J., and Kaestner, K. H. (2009). Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res.* **37**, e113.
- Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglu, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834.

- Wederell, E. D., Bilenky, M., Cullum, R., Thiessen, N., Dagpinar, M., Delaney, A., Varhol, R., Zhao, Y., Zeng, T., Bernier, B., Ingham, M., Hirst, M., *et al.* (2008). Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* **36**, 4549–4564.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoutte, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.