CHAPTER FOUR

# ChIP-Seq: Using High-Throughput DNA Sequencing for Genome-Wide Identification of Transcription Factor Binding Sites

Philippe Lefrançois, Wei Zheng, *and* Michael Snyder

## Contents



## Abstract

Much of eukaryotic gene regulation is mediated by binding of transcription factors near or within their target genes. Transcription factor binding sites (TFBS) are often identified globally using chromatin immunoprecipitation (ChIP) in which specific protein–DNA interactions are isolated using an antibody against the factor of interest. Coupling ChIP with high-throughput DNA sequencing allows identification of TFBS in a direct, unbiased fashion; this

Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, USA

*Methods in Enzymology,* Volume 470
ISSN 0076-6879, DOI: 10.1016/S0076-6879(10)70004-5
© 2010 Elsevier Inc.
All rights reserved.

technique is termed ChIP-Sequencing (ChIP-Seq). In this chapter, we describe the yeast ChIP-Seq procedure, including the protocols for ChIP, input DNA preparation, and Illumina DNA sequencing library preparation. Descriptions of Illumina sequencing and data processing and analysis are also included. The use of multiplex short-read sequencing (i.e., barcoding) enables the analysis of many ChIP samples simultaneously, which is especially valuable for organisms with small genomes such as yeast.

## 1. Introduction

The number of completely sequenced genomes has increased dramatically with improvements in DNA sequencing technologies, and the determination of coding sequences both *in vivo* and *in silico* has identified novel genes within these newly sequenced genomes (Aparicio *et al.*, 2002). First, understanding gene regulation requires more than just knowing their genomic sequence. Second, one must identify the repertoire of transcription factors present. Coding sequences of transcription factors are often conserved in the course of evolution (Borneman *et al.*, 2006), allowing their discovery in many cases by comparison to homologous transcription factors from closely related organisms (Frazer *et al.*, 2004). Third, it is crucial to establish a list of regulated genes (target genes) for each transcription factor. Computational searches for particular transcription factor DNA binding motifs upstream of putative target genes have been useful tools to obtain such a list, although these predictions require experimental validation *in vivo* (Tompa *et al.*, 2005). Moreover, the presence of a consensus binding motif is not always directly linked to transcription factor binding, as many perfect motifs are not bound by a transcription factor whereas some imperfect motifs are bound under the same environmental conditions (Borneman *et al.*, 2007; Martone *et al.*, 2003). Fourth, transcription factors can regulate genes subjugated to multiple cellular environments and stresses (Harbison *et al.*, 2004). Global characterization of binding sites of a single transcription factor, therefore, demands multiple experiments in different conditions as well as profiling across the whole genome, making such studies labor-intensive. Large international consortiums, such as ENCODE in humans (Birney *et al.*, 2007) and modENCODE in *Drosophila melanogaster* and *Caenorhabditis elegans* (Celniker *et al.*, 2009), aim to characterize every functional DNA element across the whole genome, and the study of transcription factor binding represents a major part of these efforts.

In *Saccharomyces cerevisiae*, there are approximately 200–300 described transcription factors (TFs) among the ~6000 predicted ORFs (Costanzo *et al.*, 2000). Direct analysis of transcription factor binding upstream of target genes was performed initially using DNase footprinting (Axelrod

and Majors, 1989) and/or PCR quantification of DNA associated with an immunoprecipitated transcription factor, a procedure called chromatin immunoprecipitation (ChIP) (Kuo and Allis, 1999; Orlando *et al.*, 1997). These methods could only analyze a few promoters at a time, making the comprehensive discovery of unexpected, novel TF-bound DNA elements unrealistic. The development of DNA microarrays technology has provided the field of gene regulation with a powerful tool for genome-wide characterization of transcription factor binding. This technique, termed ChIP-chip, relies on the immunoprecipitation of a transcription factor of interest with its associated DNA, followed by hybridization to a DNA microarray (Horak and Snyder, 2002). In addition, C-terminal protein tagging with an exogenous well-defined epitope (e.g., Myc, HA) circumvents the need for raising native antibodies against every transcription factor (Janke *et al.*, 2004; Longtine *et al.*, 1998). The advantages of epitope tagging include the use of commercially available antibodies, the ability to tag multiple DNA-binding proteins in a high-throughput fashion and a lower occurrence of nonspecific immunoprecipitation and cross-reaction of chromatin, ultimately resulting in decreased noise.

Novel high-throughput sequencing technologies, such as 454/Roche, Solexa/Illumina and ABI/SOLiD, have revolutionized genomic studies by allowing for large-scale sequence analysis through the generation of millions of short sequencing reads. For example, new transcripts and splice variants have been discovered in multiple organisms using RNA-Seq (Lister *et al.*, 2008; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008). Transcription factor binding studies have also benefited from ultra-throughput sequencing via the development of ChIP-Sequencing (ChIP-Seq) (Johnson *et al.*, 2007; Robertson *et al.*, 2008). Instead of hybridizing the ChIP DNA sample to a microarray, each sample is processed directly into a DNA library for sequencing and analyzed separately after sequencing. The improved sensitivity and reduced background of ChIP-Seq is replacing the array-based ChIP-chip in mammalian studies aiming to characterize transcription factor binding. Typically, two to four times more transcription factor binding sites (TFBS) are determined using ChIP-Seq in comparison with ChIP-chip; the accuracy and resolution of the data are higher as well (Robertson *et al.*, 2007). ChIP-Seq studies have been used to characterize transcription factor binding during cell growth and a stress response (Johnson *et al.*, 2007; Robertson *et al.*, 2007), enabled the establishment of a regulatory network (Chen *et al.*, 2008) as well as helped to determine epigenetic changes (Marks *et al.*, 2009). ChIP-Seq is widely used by the ENCODE and modENCODE consortia for mapping TFBS in humans, *C. elegans* and *D. melanogaster*. In humans, it has been used to examine nucleosome positioning (Schones *et al.*, 2008) and, in yeast, our group has characterized the distribution of three DNA-binding proteins, Cse4p, RNA polymerase II, and Ste12p, using this procedure (Lefrancois *et al.*, 2009).

Recently, efforts have been made to develop a multiplexing scheme for Illumina sequencing, allowing many DNA samples to be sequenced simultaneously (Craig *et al.*, 2008; Cronn *et al.*, 2008; Lefrancois *et al.*, 2009). As a typical flowcell lane currently yields approximately 8 or more million uniquely mapped sequence reads, the number of mapped reads far exceeds the minimal number required for mapping binding sites in yeast, flies, and worms (Lefrancois *et al.*, 2009; Zhong *et al.*, 2010). We have therefore developed a barcoded ChIP-Seq strategy that enables the accurate sequencing and analysis of multiple yeast ChIP samples in the same flowcell lane (Lefrancois *et al.*, 2009). Data generated in this fashion have identified binding sites for Ste12p and RNA PolII, and novel noncentromeric binding sites for Cse4p. We have also characterized the distribution of a reference sample, for example, input DNA. Input DNA, consisting of nonimmunoprecipitated, sonicated, cross-linked DNA, has great importance in ChIP-Seq studies as ChIP DNA samples are normally scored against it for TFBS identification (Auerbach *et al.*, 2009; Rozowsky *et al.*, 2009). The following protocols describe yeast ChIP-Seq, from ChIP to sequence data analysis. We also include our modifications to Illumina sequencing library preparation for generation of barcoded DNA libraries or standard, nonbarcoded DNA libraries.

Computationally, high-throughput sequencing involves handling and analysis of terabytes of sequencing data. Illumina sequencing is a four-color sequencing-by-synthesis approach where incorporation of a reversible terminator nucleotide generates a fluorescence signal detected by a high-sensitivity camera for A, C, G, and T during each cycle. The fluorescent dye is cleaved and the next base is incorporated. Typically, preliminary sequence data analyses are performed using built-in software supplied with the instrument. Fluorescent images of DNA clusters are first analyzed with a module called Firecrest to map cluster location while base-calling is performed with Bustard, which determines the probability of a given nucleotide using fluorescence intensities from the images. Finally, Gerald rapidly aligns 32 bases from the sequence reads to the reference genome using an algorithm called Eland, typically allowing for a maximum of two mismatches. These selected parameters effectively map sequence reads back to the deeply sequenced yeast reference genome. For ChIP-Seq, determination of binding sites from the sequence data is a challenge that has been tackled by different groups with various algorithms (Fejes *et al.*, 2008; Ji *et al.*, 2008; Johnson *et al.*, 2007; Jothi *et al.*, 2008; Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Xu *et al.*, 2008; Zhang *et al.*, 2008). ChIP-Seq analysis and the algorithms applied will be described in detail after the protocol section. Conceptually, sequencing reads (or tags) are compiled and genomic regions with an increased number of sequence tags compared to the tags from a control sample are considered as putative TFBS. Next, statistical filtering criteria are used to determine if these

putative sites represent true binding sites. After obtaining a preliminary set of TFBS, further bioinformatic analyses are necessary to further analyze the data. These may include analysis of the location of binding sites relative to nearby potential target genes, comparison with gene expression information and gene ontology (GO) analyses of potential targets.

## 2. Protocols

### 2.1. Chromatin immunoprecipitation

DNA–protein complexes formed *in vivo* can be reversibly cross-linked through the application of formaldehyde, and specific DNA–protein interactions are isolated from covalently bound populations using an antibody specific to the transcription factor of interest. Figure 4.1 from Horak and Snyder (2002) summarizes the principal steps of ChIP. We suggest, when possible, tagging the transcription factor of interest with a Myc or HA epitope and performing the immunoprecipitation using commercial antibodies against this epitope; these antibodies generally give little background. As an experimental control, it is possible to IP an untagged version of the same strain and to follow the same protocol. The DNAs from the tagged and untagged strain can be used for qPCR enrichment analysis of selected binding sites prior to proceeding toward sequencing library generation. We adapted this protocol from Aparicio *et al.* (2004, 2005).

(1) Grow 500 ml of yeast cells to exponential mid-log phase ($OD_{600} = 0.6–1.0$). We suggest performing ChIP experiments in biological triplicates.
(2) Treat cells with 14 ml 37% formaldehyde for 15 min, with occasional swirling every 5 min. This allows cross-linking of protein–DNA complexes.
(3) Quench cross-linking reaction by adding 27 ml of 2.5 *M* glycine for 10 min, with occasional swirling every 5 min.
(4) Collect cells by filtration and wash cells twice with 100 ml of sterile Milli-Q (Millipore, Billerica, MA) water. Rinse the filter with $2 \times 20$ ml of sterile Milli-Q water to collect cells in a 50 ml Falcon tube. Spin down cells at 4000 rpm for 10 min and discard supernatant. Resuspend the cells in 1 ml water and divide them equally in two 2 ml screw-cap tubes. Repeat this step. Spin down cells at top speed for 3 min, remove the supernatant and put on ice. Measure cell weight. Add 1 ml zirconium beads. One can continue forward to cell lysis or freeze cells at $-70$ °C for long-term storage.
(5) Resuspend cells in lysis/IP buffer (50 m*M* Hepes/KOH [pH 7.5], 140 m*M* NaCl, 1 m*M* EDTA, 1% Triton X-100, and 0.1% sodium

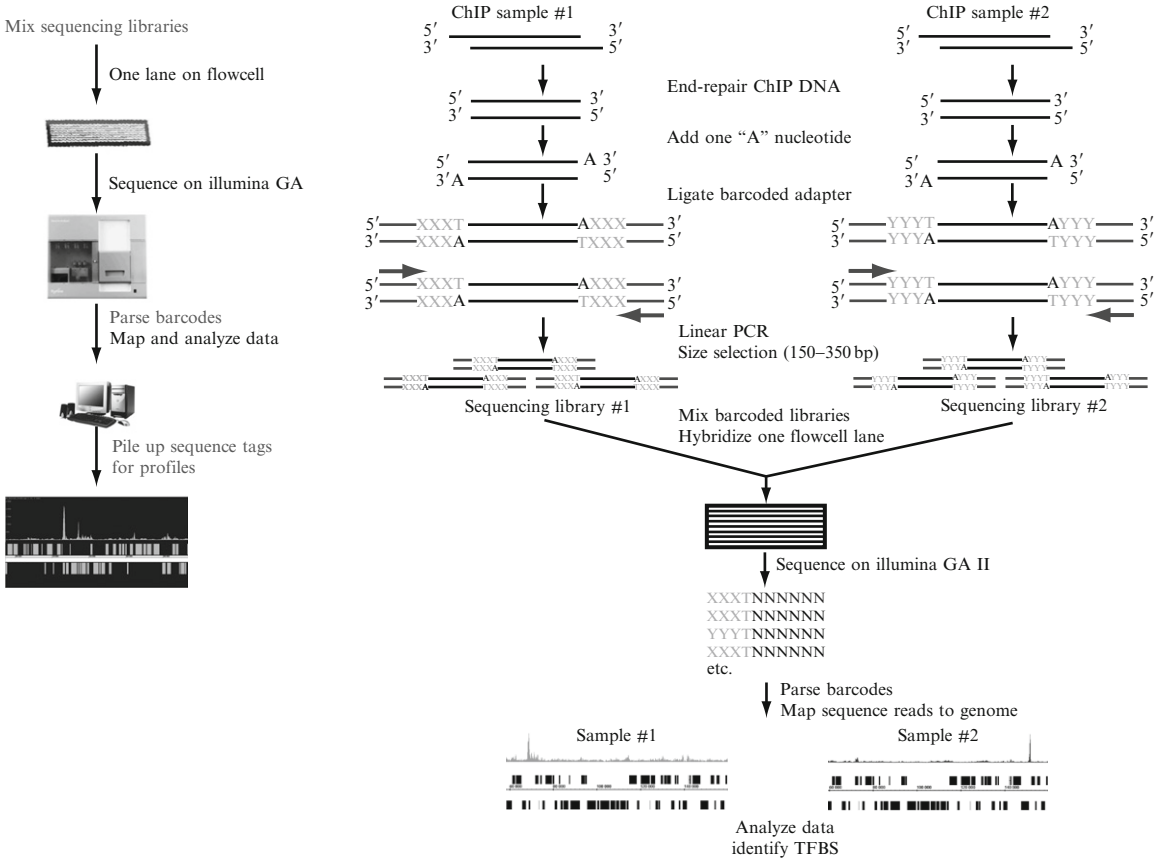


**Figure 4.1** Example of a multiplex ChIP-Seq workflow highlighting the principal steps of Illumina sequencing library generation. XXXT and YYYT represent different index sequences. Nonbarcoded ChIP-Seq can be performed by substituting the barcoded adapters with standard Illumina genomic DNA adapters.

deoxycholate) with 1 m*M* PMSF (Fluka, Buchs, Switzerland) and protease inhibitors (one tablet of Roche Complete protease inhibitor cocktail/50 ml lysis/IP buffer) and lyse them with zirconium beads using a FastPrep Machine (MP Biomedical, Irvine, CA; five times 60 s at a speed of 6.0 m/s).

(6) Recover lysates in a 5-ml snap-cap tube from one 2-ml screw-cap tube by centrifugation at 1500 rpm for 3 min, add 0.5 ml of lysis/IP buffer to the microfuge tube and centrifuge again. Pool the lysate from the other 2 ml screw-cap tube the same way. Add 1 ml of lysis/IP buffer prior to sonication.

(7) Sonicate cell lysates using a Branson Digital 450 Sonifier (Branson, Danbury, CT) to shear DNA. Each sample was sonicated five times 30 s with amplitude of 50%. Between each round of sonications, samples were put on ice for 2 min. The sonicated lysates should be clarified twice, first by a first centrifugation in a Sorvall centrifuge for 5 min at 3000 rpm and then by a second centrifugation in Eppendorf microfuge at 14,000 rpm for 10 min.

(8) Save 250 $\mu$l of clarified, sonicated lysate prior to immunoprecipitation to generate input DNA for Illumina sequencing (see next protocol).

(9) Add 2 ml of lysis/IP buffer to each sample, the total volume should be around 6 ml.

(10) Prewash the entire bottle of antibody-coupled beads using lysis/IP buffer. Remove the beads with a broadened 1 ml pipette and transfer to a 15 ml Falcon tube. Wash three times the bottle with 1 ml of fresh lysis/IP buffer to collect all the beads in the 15 ml tube. Vortex briefly and spin 2 min at 2000 rpm in a 4 °C centrifuge. Remove supernatant. Repeat three times with 4–5 ml fresh lysis/IP buffer. Resuspend the beads in an equal volume of lysis/IP buffer (1 ml). For Myc- or HA-tagged strains, we use Sigma EZview anti-Myc affinity gel (Sigma, St. Louis, MO) and Sigma EZview anti-HA affinity gel (Sigma). One antibody bottle can be used for 12 samples.

(11) Add 150–300 $\mu$l of prewashed beads to each sample. Immunoprecipitate overnight (12–16 h) on a rocker in the cold room.

(12) After incubation, fill Falcon tube with fresh lysis/IP buffer, pellet antibody beads by spinning 5 min at 3000 rpm in a cold centrifuge and discard supernatant.

(13) Wash the immunoprecipitated samples with 10 ml of appropriate buffer for 5–10 min on a rocker in the cold room. Between washes, spin down the beads in cold centrifuge at 2000 rpm for 2 min. Wash twice with lysis/IP buffer, once with IP/500 m*M* NaCl buffer (18 ml 5 *M* NaCl added to 232 ml of lysis/IP buffer), twice with IP wash buffer (10 m*M* Tris–HCl, 0.25 *M* LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, and 1 m*M* EDTA), and once with 1× TE (50 m*M* Tris–HCl, 10 m*M* EDTA, pH 8.0). Following the last wash in TE, keep some buffer to transfer beads to a 1.5-ml tube.

**(14)** Transfer beads from the 15 ml Falcon tube to a 1.5-ml microcentrifuge tube using a broadened 1 ml pipette. Transfer the remaining beads with an additional 0.5 ml of 1× TE. Spin down the beads for 3 min at 14,000 rpm and remove all TE.

**(15)** Elute the immunoprecipitate from the beads by adding 100–150 μl 1× TE/1% SDS and incubating for 15 min at 65 °C. After 10 min, mix samples briefly. Pellet the beads at 14,000 rpm for 1 min and transfer eluate to a new 1.5 ml tube. Add 150–200 μl of 1× TE/0.67% SDS to the beads and incubate for 10 min at 65 °C. Pellet the beads and pool with previous eluate. Spin down at top speed for 2 min the pooled eluates to remove all beads, as their presence will reduce cross-linking reversal efficiency and therefore ChIP DNA recovery. Transfer eluates to a 2-ml screw-cap tube, avoiding the last 10 μl and the beads at the bottom of the tube.

**(16)** Reverse protein–DNA cross-links by incubating at 65 °C overnight or for 6–8 h.

**(17)** Treat samples with proteinase K to remove all proteins from samples. Dilute 20 mg/ml proteinase K (Ambion, Austin, TX) 50-fold in 1× TE. Add 250 μl of diluted proteinase K solution per sample. Incubate between 37 and 50 °C for 2–4 h.

**(18)** Precipitate DNA with ethanol. Add 3 μl of 20 mg/ml glycogen, 2–5 μl of pellet paint (Novagen, San Diego, CA), 45 μl of 5 *M* LiCl, and 1 ml of 100% ethanol. Mix thoroughly and incubate at −20 °C overnight or at least several hours. Put samples 1 h at −70 °C. Spin in a cold centrifuge for 20 min at top speed and remove supernatant. The DNA pellet should be slightly pink due to pellet paint. Wash with 1 ml 70% ethanol for 5 min, spin in a cold centrifuge for 10 min at top speed and remove supernatant. Air dry for 10 min. Resuspend in 100 μl 1× TE.

**(19)** Purify DNA using MinElute PCR purification kit (Qiagen, Valencia, CA). We recommend processing the ChIP DNA sample in two MinElute spin columns. Elution is done in 21 μl EB per column and the two eluates from the same sample are pooled. Samples are stored in a −20 °C freezer.

This procedure typically yields 100–300 ng of ChIP DNA. DNA concentrations can be measured using a Nanodrop spectrophotometer (Thermo Scientific, Waltham, MA) or PicoGreen dsDNA quantification assay (Invitrogen, Carlsbad, CA). qPCR analysis should be performed prior to generation of sequencing library. We typically compare ChIP samples from a tagged strain (experimental sample) versus an untagged strain (control sample) for enrichment in the experimental sample at three known binding sites and at a genomic locus where the transcription factor is not expected to bind (negative control). We have found that ChIP efficiency is

the most critical step for success of the entire procedure. Here we suggest steps in the protocol for quality control as well as parameters that can be modified:

(a) *Formaldehyde cross-linking*: Changing the concentration of formaldehyde and the duration of cross-linking can modify the extent of cross-linking. Too much cross-linking can mask the HA or Myc epitopes on tagged transcription factor while too little cross-linking will decrease the immunoprecipitation of the associated DNA.

(b) *Cell lysis*: Five 1-min burst of FastPrep machine typically lyse over 95% of cells. Breaking cells using a paint shaker for 30 min yields about 40% of lysed cells.

(c) *Sonication*: Chromatin should be sheared to a median size of 450–500 base pairs (bp), as measured by gel electrophoresis in a 2% agarose gel. After step 7, take 250 $\mu$l of clarified lysate and add an equal volume of 1× TE/1% SDS. Follow the aforementioned protocol from steps 16 to 18, without purification through a MinElute spin column. Then load on a 2% agarose gel for electrophoresis. Ideally, a smear between 100 and 1000 bp should be present, with a median size of 450–500 bp (stronger smear intensity).

(d) *Antibody*: Prior to performing a ChIP experiment with a new tagged strain, a Western blot should be performed to confirm the correct insertion of the epitope. Antibody quantities can also be optimized by preliminary IP experiments with various amounts of antibody.

## 2.2. Input DNA preparation

Input DNA serves as an important reference sample for ChIP-Seq experiments (Lefrancois *et al.*, 2009; Robertson *et al.*, 2007; Rozowsky *et al.*, 2009). It is used during the scoring process where TFBS are determined based on the sequence reads obtained from a ChIP sample in comparison to input DNA. Input DNA consists of sonicated cross-linked chromatin that is processed in parallel to a ChIP sample, but lacking the immunoprecipitation step. Recent reports have suggested that input DNA represents breaks in chromatin regions of increased accessibility (Auerbach *et al.*, 2009; Teytelman *et al.*, 2009). However, there is currently a debate whether input DNA, normal IgG or, in the case of yeast, untagged strain should be the control sample for scoring ChIP-Seq data. Here we present our laboratory's protocol for isolation of input DNA, which starts at step 8 of the previous ChIP protocol.

(20) Combine 250 $\mu$l of 1× TE/1% SDS to the reserved 250 $\mu$l of clarified sonicated lysate (from step 8, ChIP protocol) in a 2-ml screw-cap tube.

(21) Reverse cross-links overnight by incubating at 65 °C.

(22) Treat samples with proteinase K as described (step 17, ChIP protocol).
(23) Extract input DNA three times with phenol:chloroform:isoamyl alcohol (25:24:1) (Fluka) followed by a single extraction with chloroform alone. In each case, keep the upper aqueous phase.
(24) Precipitate DNA with ethanol by adding 50 $\mu$l of 5 *M* LiCl and 1 ml of 100% ethanol to the upper aqueous phase from the last chloroform extraction. Enhance precipitation by transferring at $-20$ °C for 1 h. Centrifuge samples at top speed for 20 min and discard supernatant. Wash with 1 ml of 70% ethanol in the cold room for 5 min, spin down DNA at top speed for 10 min, discard ethanol and air dry for 10 min. Resuspend DNA in 1× TE (pH 8.0).
(25) RNase-treat the input DNA sample. Add 2 $\mu$l of 10 mg/ml DNase-free RNase A (Roche, Indianapolis, IN) and incubate for 30 min at 37 °C.
(26) Purify DNA using a MinElute PCR purification column (Qiagen). Elution is done in 21 $\mu$l of EB.

The amount of input DNA recovered using this procedure is much greater than that of a ChIP sample. We recommend the use of one-fifth of each input DNA sample for Illumina sequencing library preparation. If the upper phase seems unclear and the interphase is still very cloudy after the three phenol:chloroform:isoamyl alcohol extractions, an additional extraction should be performed. The phase-lock gel system (5 Prime, Gaithersburg, MD) can be used to perform safer extractions with higher recovery of DNA due to the organic phase and the interphase being sequestered physically at the bottom. This facilitates the removal of the upper, aqueous phase containing DNA.

## 2.3. Illumina sequencing DNA library generation

ChIP samples must be converted into DNA libraries for sequencing. Protocols differ depending of the sequencing platform used; 454/Roche, Solexa/Illumina, SOLiD/ABI, and Helicos each use different strategies to create a library representing the population of short DNA fragments selected by ChIP. Analysis of TFBS by sequencing technologies does not require very long sequencing reads; large numbers of short reads (e.g., 35 bp) are sufficient for mapping binding sites in most organisms. Therefore, Illumina/Solexa and ABI/SOLiD have been favored over Roche/454 because they both generate millions of very short reads (about 35 bases/read) whereas Roche/454 generate less reads but of longer length (200–300 bases/read). Currently, most ChIP-Seq studies have been performed on the Illumina platform and a few have used SOLiD. Here we describe our procedure to generate standard, nonbarcoded Illumina libraries. We have optimized the ChIP-Seq protocol used in mammalian cell lines experiments

(Robertson *et al.*, 2007) to the yeast context, which follows the manufacturer's guidelines. During Illumina library generation, oligonucleotide adapters are introduced at the ends of the small ChIP DNA fragments that were bound previously by the transcription factor of interest. These adapters allow hybridization of the sample to a flowcell containing a lawn of primers which is used for subsequent cluster generation and sequencing-by-synthesis.

During Illumina library preparation, the sheared ChIP DNA is end-repaired. A single adenosine base ("A") is added to the 3′ end of both strand followed by annealing and ligation to the double-stranded adapter containing a "T" overhang. A short PCR amplification (15–17 cycles) with primers annealing to the adapter sequence is performed to generate a population of adapter-ChIP DNA fragments termed the library. Size selection on a 2% agarose gel allows isolation of the amplified DNA library between 150 and 350 bp. This is the optimal range of fragment size for hybridization to the flowcell and cluster generation according to Illumina's recommendations.

According to bioinformatic simulations based on the yeast genome, only 260,000 uniquely mapped reads would be sufficient to determine at least 95% of the TFBS from a typical punctual TF if these binding sites are enriched at least fivefold in the ChIP sample (Lefrancois *et al.*, 2009). This is very low when compared to the human genome, where 12 *M* mapped reads is usually used (Rozowsky *et al.*, 2009). A single Illumina flowcell lane generates about 8 *M* mapped reads so multiple yeast ChIP-Seq samples can be sequenced simultaneously using multiplex Illumina sequencing (Lefrancois *et al.*, 2009). As shown in Fig. 4.1, to generate barcoded Illumina libraries, one can substitute Illumina's genomic DNA adapters for custom-made adapters that contain the adapter sequence from Illumina genomic DNA adapter followed by a nucleotide tag of at least two bases (called the barcode or index; we usually use three bases) and terminated by a single "T" for annealing and ligation to the end-repaired DNA containing an "A" overhang (Craig *et al.*, 2008; Cronn *et al.*, 2008; Lefrancois *et al.*, 2009). Standard Illumina genomic DNA PCR primers are used and the rest of the procedure is intact. ABI/SOLiD has established an indexing strategy since the commercial launch of their platform.

**(27)** Perform gel electrophoresis on a 2% agarose gel with at least 100 ng of ChIP DNA from step 19 (or input DNA) and size select the DNA smear between 100 and 700 bp. For ChIP, we usually use between 15 and 35 $\mu$l of MinElute-purified DNA from step 19. For input, due to its higher DNA concentration, one can apply a lower volume of MinElute-purified DNA from step 26 on the gel (5–10 $\mu$l) or gel-purify the same volume as for ChIP but use only 20–25% of the gel-purified input DNA for the next steps. A 100-bp DNA ladder

should be included during gel electrophoresis. Samples should not migrate too much on the agarose gel to allow for isolation of the 100–700 bp smear in a relatively small gel volume. Samples are run typically ~20 min at 100–110 V. The Qiagen QIAquick gel extraction kit is used (Qiagen) and elution is done in 34 μl EB. Although this size selection step is optional, we recommend it for exclusion of very short fragments and longer fragments which are not suitable for Illumina sequencing. For input DNA, the intensity of the smear should be high while for ChIP DNA, the smear should be visible although much fainter.

**(28)** End-repair DNA for 45 min at room temperature using End-It DNA end-repair kit (Epicentre, Madison, WI). DNA fragments are blunted by end-repair and all 5′ ends are phosphorylated. DNA is purified after end-repair using a QIAquick PCR purification column (Qiagen) and eluted in 34 μl EB.

**(29)** Add a single adenosine nucleotide (''A'') to the 3′ blunted ends of end-repaired DNA fragments (in 34 μl EB). Perform a reaction on eluted sample from step 28 with 10 μl 1 m*M* dATP, 5 μl 10× NEB buffer 2 and 1 μl Klenow fragment (3′ → 5′ exo minus) (NEB, Ipswich, MA). Mix all components in a PCR plate and cover with a sealing microfilm. Reaction is performed at 37 °C for 30 min in a PCR machine, without the use of a heated lid. Aliquots of 1 m*M* dATP should be prepared from a 100-m*M* dATP stock solution (Invitrogen) and frozen at −20 °C. Freeze–thaw should be avoided. The low concentration of dATP permits the single addition of an ''A.'' A MinElute PCR purification column (Qiagen) is used to purify the reaction and DNA is eluted in 10 μl EB.

**(30)** Ligate Illumina genomic DNA adapters (Illumina, San Diego, CA) or barcoded adapters to the sample for 15 min at room temperature. Mix 10 μl of sample from step 29, 1 μl of diluted oligonucleotide adapters, 1.5 μl of LigaFast T4 DNA Ligase (3 units/μl; Promega, Madison, WI), and 12.5 μl of Rapid Ligation Buffer (Promega). The dilution of Illumina genomic DNA adapters depends of the nature of the sample. For input DNA, Illumina nonbarcoded genomic DNA adapters are diluted 1:20 with Gibco RNase-free, DNase-free water (Invitrogen); for ChIP DNA, Illumina adapters are diluted 1:40. After the 15 min reaction, ligation products are purified with a MinElute PCR purification column and eluted in 10 μl EB. These adapters contain an unpaired ''T'' overhang which anneals to the 3′ ''A'' on the sample DNA. Barcoded adapters must have been annealed before being added to the end-repaired DNA. The concentration of diluted barcoded adapters for ligation to input DNA or ChIP DNA samples should mimic that of standard genomic DNA adapters. Barcoded adapter design and annealing will be described in the next section.

(31) Perform a gel electrophoresis on a 2% agarose gel and size select the DNA smear between 150 and 500 bp. We have found that this gel purification prior to PCR amplification has increased the quality of the sequencing libraries. More importantly, it decreases the intensity and occurrence of adapter–adapter dimerization after the PCR amplification. Adapter–adapter dimers amplify preferentially during the following PCR step and appear as a compact bright band around 100–120 bp. This compact band can totally or partially replace the normal smear indicative of a successful library. For this reason, at this step, DNA fragments below 150 bp should be excluded. We recommend using a 2% agarose E-Gel to separate adequately samples during loading and migration. Load 20 μl of a 1:10 diluted Track-It 50 bp DNA ladder (Invitrogen). Add 3 μl of a 1:10 diluted Track-It Cyan/Orange loading buffer (Invitrogen) to each sample. Samples should be separated by at least two empty wells. Load 20 μl of Gibco RNase-free, DNase-free water to all empty wells. Perform gel electrophoresis for 20 min. Recover DNA using the QIAquick gel extraction kit (Qiagen) and elute ligated samples in 28 μl EB. At this step, input DNA libraries should be visible but rather faint while ChIP DNA libraries are fainter than input DNA ones and even sometimes cannot be seen. The lack of a visible ChIP DNA smear at this step does not prevent generation of successful and high-quality libraries.

(32) Amplify the sequencing library by PCR using Illumina genomic DNA primers 1.1 and 2.1. In a PCR plate, mix 28 μl of eluted DNA sample from step 32, 1 μl of 1:1 diluted Illumina genomic DNA primer 1.1, 1 μl of 1:1 diluted Illumina genomic DNA primer 2.1, and 30 μl of Phusion Master Mix with HF Buffer (NEB). Use the following PCR settings with a heated lid: denaturation at 98 °C for 30 s, 17 cycles of amplification (10 s at 98 °C, 30 s at 65 °C, and 30 s at 72 °C), an extra amplification at 72 °C for 5 min and a cool down to 4 °C. Remove enzymes and buffer using a MinElute PCR purification column (Qiagen) and elute in 10 μl EB.

(33) Size select the Illumina sequencing library between 150 and 350 bp by gel electrophoresis on a 2% agarose gel. These size specifications meet the manufacturer's guidelines for cluster generation, optimal at a median fragment size of about 230 bp. The use of a 2% agarose E-Gel is preferable. Loading of samples and ladder are identical to step 31. Run gel electrophoresis for 20 min. A picture of the final library on the gel should be taken. At this step, a medium-to-high intensity smear over 150 bp and under 500 bp should be easy to visualize, suggesting the sequencing library preparation was successful. If there is a faint well-defined band at around 100–120 bp, extreme care should be taken during gel excision to avoid completely this adapter–adapter dimer band. The presence of adapter–adapter dimers

during sequencing will greatly decrease the overall mappability of sequencing reads. The number of uniquely mapping reads could then be very low. Gel extraction is done using a MinElute gel extraction kit (Qiagen) and elution is done in 20–25 $\mu$l EB.

**(34)** Measure DNA concentration and the $Abs_{260\ nm/280\ nm}$ ratio using a Nanodrop spectrophotometer (Thermo Scientific). Good quality libraries have an A260/280 ratio between 1.7 and 2.0. Lower values indicate poor quality. The minimal DNA concentration to proceed toward Illumina sequencing is 5.0 ng/$\mu$l. Libraries with lower DNA concentrations should be discarded. We typically obtain DNA concentrations over 8.0 ng/$\mu$l for ChIP DNA libraries and over 15.0 ng/$\mu$l for input DNA libraries.

**(35)** Store Illumina sequencing libraries at $-70\ ^{\circ}C$ until they are processed for sequencing.

Samples are now ready for the sequencing step of the ChIP-Seq procedure. They are compatible with Illumina Genome Analyzer and Genome Analyzer II. Generation of barcoded libraries follows an identical procedure except barcoded adapters are added at step 30 instead of Illumina genomic DNA adapters. Prior to the sequencing of barcoded DNA libraries, they must be mixed together in an equimolar ratio using DNA concentrations obtained from Nanodrop. A more precise method to measure DNA concentrations such as the PicoGreen dsDNA quantification assay (Invitrogen) could also be used, although we have obtained good barcode representation with Nanodrop concentrations (less than twofold difference in the number of mapped reads between the least abundant and the most abundant barcoded sample). Here are a few considerations for Illumina sequencing DNA library generation:

**(e)** *ChIP efficiency*: An insufficient amount of starting DNA material (in this case, ChIP DNA) is the most important cause of failure in library preparation as noted by the complete absence of a smear on the final agarose gel in step 33, the presence of a single intense adapter–adapter dimer band at 100–120 bp or the cooccurrence of a strong adapter–adapter dimer band and of a very faint library smear. Scaling up the ChIP protocol is a solution to generate more DNA as well as the use of tagged strains and/or of ChIP-grade antibodies.

**(f)** *Adapter dilution*: It is crucial to dilute barcoded adapters to the working concentration of the diluted standard Illumina genomic DNA adapters. If the concentration of barcoded adapters is too high, it may favor the ligation of adapter to other adapters, resulting in the formation of a strong adapter–adapter dimer band during the final gel extraction (step 33) or in the absence of a DNA smear indicative of a successful library. Optimization of concentrations should be first performed on input DNA. Similarly, if problems occur using Illumina genomic DNA

adapters, optimization should also be performed on input DNA and then on ChIP DNA.

**(g)** *PCR amplification*: PCR amplification using Illumina genomic DNA primers 1.1 and 2.1 should stay in the linear range to avoid overrepresentation of some genomic areas among the sequencing library. Sequencing reads would then be very high for these overrepresented regions, creating a bias during data analysis. The manufacturer recommends no more than 18 PCR cycles. One can perform less cycles. The common range for PCR amplification of Illumina DNA libraries lies between 13 and 18 cycles.

## 2.4. Barcode design and adapter annealing

This section applies specifically to barcoded ChIP-Seq on an Illumina platform. Multiplex sequencing-by-synthesis has been accomplished through the introduction of indexed (or barcoded) adapters (Craig *et al.*, 2008; Cronn *et al.*, 2008; Lefrancois *et al.*, 2009). These strategies have allowed multiplex sequencing and analysis of HapMap loci from different individuals (Craig *et al.*, 2008), chloroplast genomes from different species (Cronn *et al.*, 2008), and yeast ChIP samples (Lefrancois *et al.*, 2009), without the introduction of barcode-induced errors or artifacts. In all cases, a barcode was introduced after the Illumina adapter sequence required for PCR amplification and hybridization to Illumina's flowcell. The resulting sequencing reads first contain the index followed by the sequenced DNA sample. The barcode must contain a final ''T'' for pairing and ligation to the end-repaired DNA with an ''A'' overhang. We have used four indexes for barcoded ChIP-Seq: ACGT, CATT, GTAT, and TGCT. We have created a three base index where no barcode contained the same base at each position mainly for two reasons. First, these barcodes have a balanced nucleotide composition in compliance with manufacturer's guidelines. Second, one- or two-base sequencing errors would not result in a barcode being assigned to an erroneous sample as the remaining index base would not match another sample. In our work, the barcode must be intact in all sequencing reads assigned to a sample. With the new Illumina Genome Analyzer II, the increase in read length and the decrease in error rates at later sequenced bases permit generation of longer barcodes. This, coupled to an expected increased number of sequencing reads, could significantly increase the level of multiplexing in yeast ChIP-Seq studies as well as ChIP-Seq studies in small genome organisms. Here we present the procedure for oligonucleotide design and annealing to generate barcoded adapters.

**(36)** Synthesize oligonucleotides at a 0.05 $\mu$mol scale with HPLC purification from MWG/Operon (Eurofins MWG Operon, Huntsville, AL). Oligonucleotide sequences are given in Table 4.1. Note that the

**Table 4.1** Oligonucleotide sequences for barcoded ChIP-Seq

| Barcode | Forward/reverse | Sequence (5′ → 3′)[a] |
|---|---|---|
| ACGT | Forward[b] | ACACTCTTTCCCTACACGACGCTC TTCCGATCTACGT |
| | Reverse[c] | CGTAGATCGGAAGAGCTCGTATG CCGTCTTCTGCTTG |
| CATT | Forward[b] | ACACTCTTTCCCTACACGACGCTC TTCCGATCTCATT |
| | Reverse[c] | ATGAGATCGGAAGAGCTCGTATGC CGTCTTCTGCTTG |
| GTAT | Forward[b] | ACACTCTTTCCCTACACGACGCTC TTCCGATCTGTAT |
| | Reverse[c] | TACAGATCGGAAGAGCTCGTATGC CGTCTTCTGCTTG |
| TGCT | Forward[b] | ACACTCTTTCCCTACACGACGCTC TTCCGATCTTGCT |
| | Reverse[c] | GCAAGATCGGAAGAGCTCGTATGC CGTCTTCTGCTTG |

[a] From Lefrançois *et al.* (2009).
[b] No modification.
[c] 5′ ends are phosphorylated.

forward primer contains the index and the final "T" at the 3′ end while the reverse primer is phosphorylated at the 5′ end and the reverse-complement index sequence is found at the 5′ end.

**(37)** Resuspend each primer in annealing buffer (10 m*M* Tris [pH 7.5], 50 m*M* NaCl, 1 m*M* EDTA) to 200 $\mu M$.

**(38)** Mix the forward and reverse primers for each index pair in equal volumes to a final concentration of 100 $\mu M$.

**(39)** Heat to denature in a wet heat block at 95 °C for 5 min.

**(40)** Remove heat block to room temperature and let primers cool down during 45 min to promote annealing.

**(41)** Keep on ice for a few minutes and store barcoded adapters at −20 °C.

**(42)** Dilute barcoded adapters with Gibco RNase-free, DNase-free water (Invitrogen) to the working concentrations of Illumina genomic DNA adapters for generation of input DNA and ChIP DNA libraries (previous protocol). Annealed indexed adapters have different concentrations from each other and differ in the dilutions to obtain the adequate working concentrations. As an example, with our barcoded adapters given in Table 4.1, we have diluted all four adapters 1:30 to generate barcoded input DNA libraries while we have diluted differently adapters for barcoded ChIP DNA libraries: 1:750 for ACGT, 1:450 for CATT, 1:500 for GTAT, and 1:330 for TGCT.

## 2.5. Illumina sequencing

We follow manufacturer's protocols and guidelines. Detailed protocols for operating the cluster station and sequencing using Genome Analyzer II are available from Illumina's web site. Here we will only briefly describe the various steps of Illumina sequencing. First step is cluster generation on Illumina's cluster station. The cluster station uses microfluidics to physically attach DNA from the sequencing library (step 34) onto one lane of an Illumina flowcell. An Illumina flowcell contains eight lanes and each lane has a lawn of primers with a sequence corresponding to the complement of the Illumina adapter sequence. Samples are denatured and each single-stranded ChIP DNA fragment is connected to the lawn primer via one adapter end. A solid-phase bridge amplification replicates the template DNA fragment from the paired adapter–primer. After denaturation of this double-stranded bridge of DNA, the initial template DNA is washed away and the flowcell-attached replica of the template can undergo successive rounds of bridge amplification to generate a cluster. A cluster contains about 1000 copies from an identical initial template. There are typically 100–120,000 clusters on a single tile, with 100 tiles per flowcell lane. This can give rise to 10–12 *M* reads per lane. DNA loaded on the flowcell should be at a concentration between 3 and 5 p*M* and optimal library size should be between 150 and 350 bp. If the DNA library is smaller, too many clusters of smaller size will be present due to the lesser reach of bridge amplification, giving rise to a fewer number of clusters passing quality metrics and a lower number of mapped reads. On the other hand, if the smear size of the library is bigger, fewer clusters of bigger size will be generated due to the greater reach of bridge amplification, resulting to a decreased number of clusters and mapped reads. Just prior to sequencing, a sequencing primer is annealed. The flowcell is then transferred from the cluster station to the Genome Analyzer II for sequencing of DNA clusters. Illumina employs a four-color sequencing-by-synthesis method. Fluorescently labeled reversible terminator ddNTPs are added simultaneously and one base is incorporated per cluster. Laser excitation and fluorescence allows the detection of the first base. The fluorescent dye is cleaved and the first base is unblocked to add the second base using the same reagents. This process is continued for 34–36 cycles by sequencing one base at a time. Each read starts with the template DNA sequence or, in the case of barcoded ChIP-Seq, with the 4-bp barcode. The following section focuses on sequencing reads analysis.

# 3. Sequencing Data Management

Illumina uses a massively parallel sequencing-by-synthesis approach. A typical run on the Illumina instrument lasts 2–3 days, and generates at least 1 terabyte of data, which poses a big challenge on data storage system

and data transfer method. Analyzing these raw image data to get biologically meaningful sequences is also a computationally intensive task. We use Genome Analysis Pipeline (GAP) software (Illumina) to analyze the sequencing data. The minimum system requirement for running this software is a dual-processor, dual-core computer. If multiprocessor facilities are available, the data analysis time can be greatly reduced by parallelization. The outputs produced by GAP are stored in a hierarchical directory structure called the ''run folder.'' Currently our run-folder resides on the Yale biomedical high-performance computing cluster, which consists of 170 Dell PowerEdge 1955 nodes, and each node contains two dual core 3.0 GHz EM64T Intel CPUs and 16 GB RAM.

It is very important to have an IT infrastructure with sufficient computation capacity, data storage and transfer abilities to support Illumina Genome Analyzer. Depending on the scale of sequencing runs, a laboratory can also consider commercially available Laboratory Information Management System (LIMS), such as WikiLIMS (BioTeam).

## 4. Genome Analysis Pipeline

Since detailed instructions of installing and running the GAP are available from Illumina, we will only briefly introduce the functionality of pipeline modules related to ChIP-Seq data analysis. Users can refer to the GAP documentation for more details. The documentation files can be obtained with Genome Analyzer machine setup, or browsed through publicly accessible domains. The link to the documentation files at Yale University is http://sysg1.cs.yale.edu:3443/pDir/GAP-1.1.0-docs/.

There are three main modules in the GAP. The first module Firecrest is an image-analysis module. The images are generated from sequencing-by-synthesis at hundreds of thousands of clusters. At each cluster, the sequencing machine records four images of added nucleotides (A, G, C, or T) at each synthesis cycle. Firecrest analyzes images captured by the sequencing machine, and remaps cluster positions. In an updated version, Illumina introduced the Integrated Primary Analysis and Reporting (IPAR) Software, which processes images and performs quality control in real time. IPAR removes the need of storing raw images. The second module, Bustard, performs base-calling. From the four images captured for A, G, C, and T at each round of synthesis, Bustard calculates the occurrence probability of a certain nucleotide at each cluster, and after 30–34 cycles of synthesis, it concatenates a chain of nucleotides with highest occurrence probabilities into a short sequence tag with length equal to the number of synthesis cycles. It has also some built-in quality control mechanism to determine the confidence of its base-calling. The third module, Gerald, aligns short sequence reads to a reference genome. The alignment software (Eland) in the Gerald module

runs very fast and accurately aligns sequence tags of less than 32 bp to a reference genome. Several other open-source programs can also align a large number of short sequences, such as SOAP (Zhang *et al.*, 2008) and MAQ (Li *et al.*, 2008). SOAP has a unique feature of aligning sequences across small gaps in the genome, which is helpful in sequencing transcriptomes. MAQ has a dedicated module to call SNPs and *de novo* genome assembly.

## 5. Examining Data Quality and Parsing Barcoded Data

Before we can use the ChIP-Seq data to answer biological questions, we need to make sure the data quality is of sufficient quality. There are several summary statistics to examine after a sequencing run: % Error (multiplexing runs usually have higher % error than nonbarcoded runs, around 5%), % Phasing ($<1\%$), total reads (GAII can reach 12–14 million) and cluster density ($\sim 100{,}000$). Some other statistics to verify alignment percentage of short sequences to the genome are Total No Match, % No Match, Total QC Fail, % QC Fail, R0 Multiple Match, R1 Multiple Match, R2 Multiple Match, Total Multiple Match, % Multiple Match, U0 Unique Match, U1 Unique Match, U2 Unique Match, Total Unique Match, and % Unique Match.

For multiplexing runs, users need to parse the Eland query file by matching the first several nucleotides with the barcode sequences, remove the barcode sequences from the sequencing reads, and rerun Eland separately for each parsed data set. In our case, the barcodes are GTAT, CATT, ACGT, TGCT (Lefrancois *et al.*, 2009). Users do not need to check alignment statistics for the entire lane because the alignment uses sequence tags with barcodes at the 5′ end. Instead, users should check these statistics for parsed data with removed barcode sequence. A typical run of yeast multiplexing sequencing with four barcodes has $\sim 15\%$ Total Multiple Match, and $\sim 60\%$ Total Unique Match.

Some sample Perl scripts to perform the barcode parsing and Eland rerunning tasks can be found at http://pantheon.yale.edu/~wz4/Homepage.html. Since the scripts call the Eland program, they need to be run on the same server as the GAP resides, and modified according to user-specific directory structures. If automatic barcode parsing and Eland alignment are desired, please consult with IT support to integrate these functions into the GAP.

## 6. Visualization in Genome Browser

After aligning short sequencing tags onto a reference genome, we can load the data into a Genome Browser to directly visualize how the short tags are distributed across the genome, and whether there is any enrichment near

the regions of interest. There are many versions of Web-based genome browsers available for the yeast community, including Gbrowse, UCSC Genome Browser. One can upload data to the server in a format that can be recognized by the genome browser, and visualize the signal track along with other annotations on the host webpage of the specific genome browser. Here we will provide a step-by-step guide to visualize ChIP-Seq data on a local machine using Integrated Genome Browser (IGB) developed by Affymetrix. First, download and launch IGB following the instructions at: http://www.affymetrix.com/partners_programs/programs/developer/tools/download_igb.affx.

To load *S. cerevisiae* annotations, click File → Access DAS/1 Servers; in the pop-up window named "DAS/1 Feature Loader," choose "UCSC" in "DAS Server" pull-down menu, "sacCer1" in "Data Source" pull-down menu, and "1 (or any other chromosome you want to see)" in "Sequence" pull-down menu, and check any interested annotations in the "Available Annotations" window.

To load ChIP-Seq data into IGB, one needs to transform the Eland results into a format that can be recognized by IGB. A sample Perl script for this purpose can be found at http://pantheon.yale.edu/~wz4/Homepage.html. To run this script, simply type the following command in a command line shell:

perl create_sgr_file.pl "eland_result_folder"

The "eland_result_folder" is the folder containing Eland results files parsed into chromosomes, with names "eland_results_chr*.txt."

This script transforms Eland results file into .sgr format, which is compatible with Affymetrix's IGB. The format for each line of an sgr file is:

Chromosome Start_Position Score

where "Score" is the number of overlapping ChIP fragments from the current Start_Position to 1 bp upstream of the next Start_Position. Before counting overlapping ChIP fragments for each genomic position, create_sgr_file.pl also extends sequencing tags in the 3′ direction to 200 bp since Illumina sequencing tags only represent one end of ChIP fragments, and the average ChIP fragment size in the sequencing library is 200 bp.

## 6.1. Low-level analysis

The Genome Browser can help scientists visualize and roughly determine sequence-tag enriched regions. To precisely identify TFBS across the genome, we need more rigorous peak-scoring algorithm. Since the emergence of ChIP-Seq technology, a bunch of peak-scoring algorithms have been developed for mammalian genomes, and most of them can also be used to analyze yeast ChIP-Seq data. Here we briefly describe the basic flowchart

of peak-scoring algorithms, and summarize the major features of several popular peak-scoring algorithms in Table 4.2. A peak-scoring algorithm usually compares sequencing data from a ChIP experiment with simulated background, or sequencing data from a control experiment (nontagged strain, IgG ChIP, or input DNA). To determine regions with enriched sequence-tag distribution, the scoring algorithm normalizes the ChIP-Seq data and control sequencing data to the same scale, and then uses proper statistical tests (e.g., Binomial, Poisson, Normal) to compare distributions of sequence tags in the two data sets. Significance level is adjusted to control the false discovery rate.

Because the Illumina sequencing platform only reads 30–32 bp from one end of the ChIP DNA fragment, the most enriched regions (peak centers) of sequencing tags may not overlap exactly with the most enriched regions (peak centers) of ChIP DNA fragments, the latter ones corresponding to potential binding sites meaningful to biologists. Several different methods were proposed to convert sequencing tag position to peak center position in published peak-scoring algorithms. The most straightforward way is to extend the sequencing tags to the length of original ChIP DNA fragment, which is about 200 bp due to size selection on agarose gel in sequencing library construction (Rozowsky *et al.*, 2009; Xu *et al.*, 2008). More sophisticated methods include estimating the length of original ChIP DNA fragments with triangle or bell shaped distribution centered at $\sim$200 bp (Fejes *et al.*, 2008), or separating sequencing reads aligned onto Watson and Crick strands, and using the distances between peak center on Watson strand and peak center on Crick strand to estimate the length of original ChIP DNA fragments (Ji *et al.*, 2008; Jothi *et al.*, 2008; Zhang *et al.*, 2008).

The biggest distinction among existing peak-scoring algorithms is the method of extracting background from ChIP-Seq data. Due to the high cost of ChIP-Seq procedure, earlier peak-scoring algorithms often considered one-sample analysis, which compares ChIP data with a null background generated from random permutation or estimated from a Poisson model. One-parameter Poisson model (Feng *et al.*, 2008; Marson *et al.*, 2008; Robertson *et al.*, 2007) has been widely used in these peak-scoring algorithms. Another popular method to estimate background is Monte Carlo sampling (Bhinge *et al.*, 2007; Chen *et al.*, 2008; Fejes *et al.*, 2008; Johnson *et al.*, 2007; Mikkelsen *et al.*, 2007; Robertson *et al.*, 2007; Zhang *et al.*, 2008). Later studies found out that Poisson model with a fixed $\lambda$ is not good enough to describe nonrandom fluctuations as observed in the input control. To alleviate this problem, CisGenome (Ji *et al.*, 2008) used Negative Binomial instead of Poisson to model the background. MACS (Zhang *et al.*, 2008) used dynamic Poisson parameters. Both studies recognized that the random sampling process had different sampling rates at different positions in the genome, and tried to capture the nature of changing parameters in the underlying Poisson model. Nonetheless, it becomes clear that two-sample

**Table 4.2** Comparison of popular ChIP-Seq peak-scoring algorithms

| Algorithm | Tag2 peak | Model | 1 or 2 sample/ scaling | Unique feature and other notes | References |
|---|---|---|---|---|---|
| ChIP-Seq PeakFinder | Use 25-nt tags directly | min_reads > 13 (default 20 in program), max-gap <75, minratio > 5 | 1 or 2, NA | qPCR to find threshold, MEME motif finding | Johnson *et al.* (2007) |
| FindPeaks3.1 | Three methods: extend fixed length, triangle with user-specified average length, adaptive | FDR using effective genome size and tag number (Monte Carlo background) | 1 | Directional: fragment after a peak are removed as "noise," subpeak: user-specified valley, peak trimming | Fejes *et al.* (2008) |
| CisGenome | Boundary refinement, single-strand filtering (detect peaks for +/− data separately) | Sliding window passing user-specified cutoff, one-sample FDR negative binomial, two-sample FDR conditional binomial | 1 or 2, use window read counts $<n$, conditional binomial $(n, c/1 + c)$ | Negative binomial model underestimate FDR when window read count is high | Ji *et al.* (2008) |

| | | | | | |
|---|---|---|---|---|---|
| SISSRs | Sep+/−, estimate average fragment length F from data | Net tag counts in overlapping (10 bp) sliding window (20 bp). Use sense–antisense read counts transition point as binding site, user-selected cutoff E, R | 1 or 2, NA | One-sample FDR Poisson, two-sample no FDR, use control sample to estimate sensitivity: specificity:empirical *p*-values | Jothi *et al.* (2008) |
| QuEST | Sep+/−, peak shift estimated from position shift between +/− data and most significant peaks | 21 bp sliding window, cutoff determined by eFDR, difference between neighboring windows $<0.9H$ of higher peak window, control H < threshold or ChIP/control ratio > threshold | 2, extract same number of control data as experimental data, and use the other half of control data to estimate empirical FDR | Gaussian KDE, bandwidth 30, ±3b, FDR for two-sample, at least three user-specified cutoff | Valouev *et al.* (2008) |
| U-Seq | Sep+/−, estimate mean fragment length, and shift read position | Overlapping sliding window defined by read position and max size 350 bp | 2, trim to equal number of reads | Spike in simulation in input control data compare sum, diff, normdiff, and binomial. Normdiff outperforms | Nix *et al.* (2008) |

(*continued*)

**Table 4.2** (*continued*)

| Algorithm | Tag2 peak | Model | 1 or 2 sample/ scaling | Unique feature and other notes | References |
|---|---|---|---|---|---|
| ChipDiff | Shifting tag by 100 bp | 1. Putative enrich region by normalized window read counts<br>2. fold change or HMM (better) | 2, normalize to total reads | Used in histone methylation detection | Xu *et al.* (2008) |
| MACS | Given bandwidth and fold enrichment, sample 1000 high-quality peaks, sep+/−, calculate $d$ between modes, shift tag position by $d/2$ | Use 2D windows, Poisson with dynamic parameter to model local background, and calculate $p$-value, eFDR calculated by swapping control/ChIP | 1 or 2, linearly scale with total read counts, remove duplicate reads from the same position | eFDR definition different from others, motif occurrence within 50 bp of peak center, average distance from peak center to motif are better than competitors | Zhang *et al.* (2008) |
| PeakSeq | Extend 200 bp | Binomial ($n$, 1/2) after scaling | 2, Simple Linear Regression using window counts | Two-pass comparison, first to random shuffled background, second to input control data | Rozowsky *et al.* (2009) |

analysis is superior because in certain genomic regions, the sequencing tag distribution in an input control experiment shows nonrandom enrichment. Sometimes the same enrichment pattern is also observed in ChIP experiment (Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Zhang *et al.*, 2008). Such enrichment is not likely to be caused by TFBS; instead it probably represents fluctuations due to systematic biases. There are many sources of systematic biases. Known sources include technical reasons such as the method of DNA fragmentation, biased amplification in PCR, error in the sequencing and/or the alignment processes; biological reasons such as the degree of genome repetitiveness, open chromatin structure; statistical reasons such as the dependency among observations from neighboring positions on a chromosome. In both one- and two-sample analyses, it is assumed that the number of sequencing tags observed in a small window of the genome comes from random sampling process. Binomial or Poisson model are often used in two-sample analyses to compare the number of reads in windows of two samples (Feng *et al.*, 2008; Ji *et al.*, 2008; Jothi *et al.*, 2008; Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Zhang *et al.*, 2008).

## 6.2. High-level analysis

Once TFBS are identified from the ChIP-Seq data, one can carry out more high-level analysis to answer biological questions, such as motif analysis, association of TFBS with neighboring genes, and comparison with ChIP-chip data. One can also study the positions of TFBS relative to genome annotation features, such as intragenic versus intergenic binding and binding in 5′ or 3′ untranslated regions. These analyses are all implemented in an integrative open-source software, CisGenome (Ji *et al.*, 2008). It has many functions varying from low-level analysis to high-level analysis, and its graphic interface under Windows OS is user-friendly for bench scientists. It can be downloaded from the following web site: http://www.biostat.jhsph.edu/~hji/cisgenome/.

## 6.3. Troubleshooting

If the sequencing run yields many reads, but the percentage of matched reads after alignment is low, a possible explanation for this phenomenon is sample overloading. If too much DNA is loaded onto the flowcell, there will not be enough separation between neighboring clusters and base-calling error rates will be high. By checking the summary statistics ''cluster density'' and ''% Error,'' one can find deviations from optimal values, and adjust sample concentration accordingly. If summary statistics for sequencing runs are adequate, one still needs to search for technical problems in ChIP and library construction procedures.

A common problem in IGB visualization is that ChIP-Seq data is shown in a separate window from other annotations. The reason is that Eland result files sometimes have different chromosome naming system (e.g., chr01, chr02, ..., chrmt) from that in IGB (e.g., chr1, chr2, ..., chrM). In this case, one needs to rename all the chromosomes in the IGB system to visualize ChIP-Seq data along with other annotations in the same window. If there is not enough memory to load data into IGB, one can load data for one chromosome at a time.

## 7. Conclusion and Future Directions

ChIP-Seq has emerged as a highly sensitive and cost-effective method for genome-wide mapping of TFBS at a high resolution. Barcoded ChIP-Seq enables multiplex short-read sequencing and offers a higher throughput and lower cost per sample. An ongoing debate in the ChIP-Seq field concerns the nature of the control DNA used for scoring ChIP-Seq experiments. Although most groups use input DNA, it is still unsettled whether input DNA, normal IgG DNA or, in the case of yeast, ChIP DNA from an untagged strain is the preferable reference sample for ChIP-Seq. With the read length, read quality, and read quantity improvements of high-throughput DNA sequencing technologies, it will be possible to obtain an increased total of reads with longer sequence lengths. For yeast ChIP-Seq, this will allow an increased multiplex capability. Computational challenges of data handling and long-term data storage require high-performance computing clusters and are much more complex than ChIP-chip analyses that could be performed by most users. The protocols developed for yeast, such as barcoded ChIP-Seq, can be readily extended to lower eukaryotes and eventually to higher eukaryotes with the advent of higher capacity DNA sequencers.

## ACKNOWLEDGMENTS

Past and present members of the M. Snyder laboratory have helped to develop these protocols. We are particularly grateful to Jennifer Li-Pook-Than for insightful comments on this manuscript, Ghia M. Euskirchen for pioneer ChIP-Seq protocol development and Christopher M. Yellman for yeast ChIP protocol optimization in our laboratory. This work has been supported by NIH grants. P. Lefrançois has been supported by a master's fellowship from FQRNT and a doctoral fellowship from NSERC during this work.

## REFERENCES

Aparicio, S., *et al.* (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297,** 1301–1310.

Aparicio, O., *et al.* (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. *Curr. Protoc. Cell Biol.* Chapter 17, Unit 17 7.

Aparicio, O., *et al.* (2005). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. *Curr. Protoc. Mol. Biol.* Chapter 21, Unit 21 3.

Auerbach, R. K., *et al.* (2009). Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. USA*. Epub ahead of print.

Axelrod, J. D., and Majors, J. (1989). An improved method for photofootprinting yeast genes *in vivo* using Taq polymerase. *Nucleic Acids Res.* **17,** 171–183.

Bhinge, A., *et al.* (2007). Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res.* **17,** 910–916.

Birney, E., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816.

Borneman, A. R., *et al.* (2006). Target hub proteins serve as master regulators of development in yeast. *Genes Dev.* **20,** 435–448.

Borneman, A. R., *et al.* (2007). Divergence of transcription factor binding sites across related yeast species. *Science* **317,** 815–819.

Celniker, S. E., *et al.* (2009). Unlocking the secrets of the genome. *Nature* **459,** 927–930.

Chen, X., *et al.* (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133,** 1106–1117.

Costanzo, M. C., *et al.* (2000). The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28,** 73–76.

Craig, D. W., *et al.* (2008). Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5,** 887–893.

Cronn, R., *et al.* (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36,** e122.

Fejes, A. P., *et al.* (2008). FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24,** 1729–1730.

Feng, W., *et al.* (2008). A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. *BMC Genomics* **9,** S23.

Frazer, K. A., *et al.* (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **32,** W273–W279.

Harbison, C. T., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104.

Horak, C. E., and Snyder, M. (2002). ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350,** 469–483.

Janke, C., *et al.* (2004). A versatile toolbox for PCR-based tagging of yeast genes: New fluorescent proteins, more markers and promoter substitution cassettes. *Yeast* **21,** 947–962.

Ji, H., *et al.* (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26,** 1293–1300.

Johnson, D. S., *et al.* (2007). Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316,** 1497–1502.

Jothi, R., *et al.* (2008). Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36,** 5221–5231.

Kuo, M. H., and Allis, C. D. (1999). *In vivo* cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods* **19,** 425–433.

Lefrancois, P., *et al.* (2009). Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics* **10,** 37.
Li, H., *et al.* (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858.
Lister, R., *et al.* (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133,** 523–536.
Longtine, M. S., *et al.* (1998). Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14,** 953–961.
Marks, H., *et al.* (2009). High-resolution analysis of epigenetic changes associated with X inactivation. *Genome Res.* **19,** 1361–1373.
Marson, A., *et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–533.
Martone, R., *et al.* (2003). Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* **100,** 12247–12252.
Mikkelsen, T. S., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560.
Mortazavi, A., *et al.* (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628.
Nagalakshmi, U., *et al.* (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349.
Nix, D. A., *et al.* (2008). Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9,** 523.
Orlando, V., *et al.* (1997). Analysis of chromatin structure by *in vivo* formaldehyde cross-linking. *Methods* **11,** 205–214.
Robertson, G., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4,** 651–657.
Robertson, A. G., *et al.* (2008). Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* **18,** 1906–1917.
Rozowsky, J., *et al.* (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27,** 66–75.
Schones, D. E., *et al.* (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898.
Teytelman, L., *et al.* (2009). Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One* **4,** e6700.
Tompa, M., *et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23,** 137–144.
Valouev, A., *et al.* (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5,** 829–834.
Wilhelm, B. T., *et al.* (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453,** 1239–1243.
Xu, H., *et al.* (2008). An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* **24,** 2344–2349.
Zhang, Y., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137.
Zhong, M., *et al.* (2010). Genome-wide identification of binding sites defines distinct functions for *C. elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet.* (In press).