PROSPECT
Journal of Cellular Biochemistry 107:11–18 (2009)

Journal of Cellular Biochemistry

Genomic Location Analysis by ChIP-Seq

Artem Barski* and Keji Zhao*
Laboratory of Molecular Immunology, National Heart, Lung and Blood Institute, NIH, Bethesda, MD 20892

ABSTRACT

The interaction of a multitude of transcription factors and other chromatin proteins with the genome can influence gene expression and subsequently cell differentiation and function. Thus systematic identification of binding targets of transcription factors is key to unraveling gene regulation networks. The recent development of ChIP-Seq has revolutionized mapping of DNA–protein interactions. Now protein binding can be mapped in a truly genome-wide manner with extremely high resolution. This review discusses ChIP-Seq technology, its possible pitfalls, data analysis and several early applications. J. Cell. Biochem. 107: 11–18, 2009.

KEY WORDS: ChIP-Seq; CHROMATIN; TRANSCRIPTION FACTOR

Regulation of gene expression by transcription factors is one of the major mechanisms for controlling cell proliferation, differentiation and function. Thus, understanding which transcription factors (TFs) operate in differentiating cells, how they are regulated at the molecular level and which genes they regulate is key to unraveling mysteries of the development of organisms.

There are several approaches to identification of protein–DNA contacts in vivo. These include computational sequence analysis, various kinds of microscopy as well as biochemical approaches.

METHODS USED FOR THE DISCOVERY OF TRANSCRIPTION FACTOR TARGET GENES

The simplest approach to discovery of transcription factor binding sites in the genome is computer analysis of genomic sequence [reviewed in Wasserman and Sandelin, 2004]. However binding site consensus sequences or motifs are often short and can be met in the genomic sequence too often, leading to low statistical significance of the matches and high false positive rate. Most sites identified in this way would likely not bind the transcription factor of interest in living cells. At the same time even low significance site might bind a protein in vivo.

Microscopy on extended chromosome fibers [Sims et al., 2006] also can be used to find genomic binding sites for proteins of interest. In living cells, microscopy and fluorescence recovery after photobleaching was used to follow binding of nuclear receptor to tandem array MMTV promoters in real time [Voss et al., 2006].While these methods allow for the analysis of colocalization of various transcription factors and produce beautiful images, the resolution of these methods is low and is more suited to identification of extended binding islands rather than individual binding events.

Two biochemical methods that were used for binding site identification in vivo are DAM-ID and ChIP. For DAM-ID protein of interest is fused to *E. coli* DNA Adenine Methyltransferase (DAM) and expressed in target cell. Dam preferentially methylates adenosines in the areas of DNA which are bound by the protein. These can be identified by methylation sensitive restriction digest and microarray analysis [Greil et al., 2006]. The advantage of DAM-ID is that it does not need specific antibodies. However, it has low resolution and requires overexpression of a protein of interest which might result in false positives.

Chromatin immunoprecipitation (ChIP) (Fig. 1A) is a method that allows analysis of direct interaction of proteins with DNA in vivo with high resolution. First, cells are lysed and chromatin is fragmented. Fragmentation can be achieved either by digestion with micrococcal nuclease (MNase) of native chromatin or sonication if cells were crosslinked.

In general, native, MNase-digested chromatin is used for ChIP of histones that stably bind DNA, whereas for more mobile non-histone proteins it is necessary to crosslink chromatin using formaldehyde. In case of histone ChIP, MNase digestion is convenient, since it can result in mononucleosome resolution and at the same time provide information for nucleosome positioning [Schmid and

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Intramural Research Program of the NIH; Grant sponsor: National Heart, Lung and Blood Institute.

*Correspondence to: Dr. Artem Barski and Dr. Keji Zhao, Laboratory of Molecular Immunology, National Heart, Lung and Blood Institute, NIH, Bldg 10, Room 7B05, 9000 Rockville Pike, Bethesda, MD 20892.
E-mail: barskia@nhlbi.nih.gov; zhaok@nhlbi.nih.gov

Received 24 December 2008; Accepted 30 December 2008 • DOI 10.1002/jcb.22077 • 2009 Wiley-Liss, Inc.
Published online 27 January 2009 in Wiley InterScience (www.interscience.wiley.com).

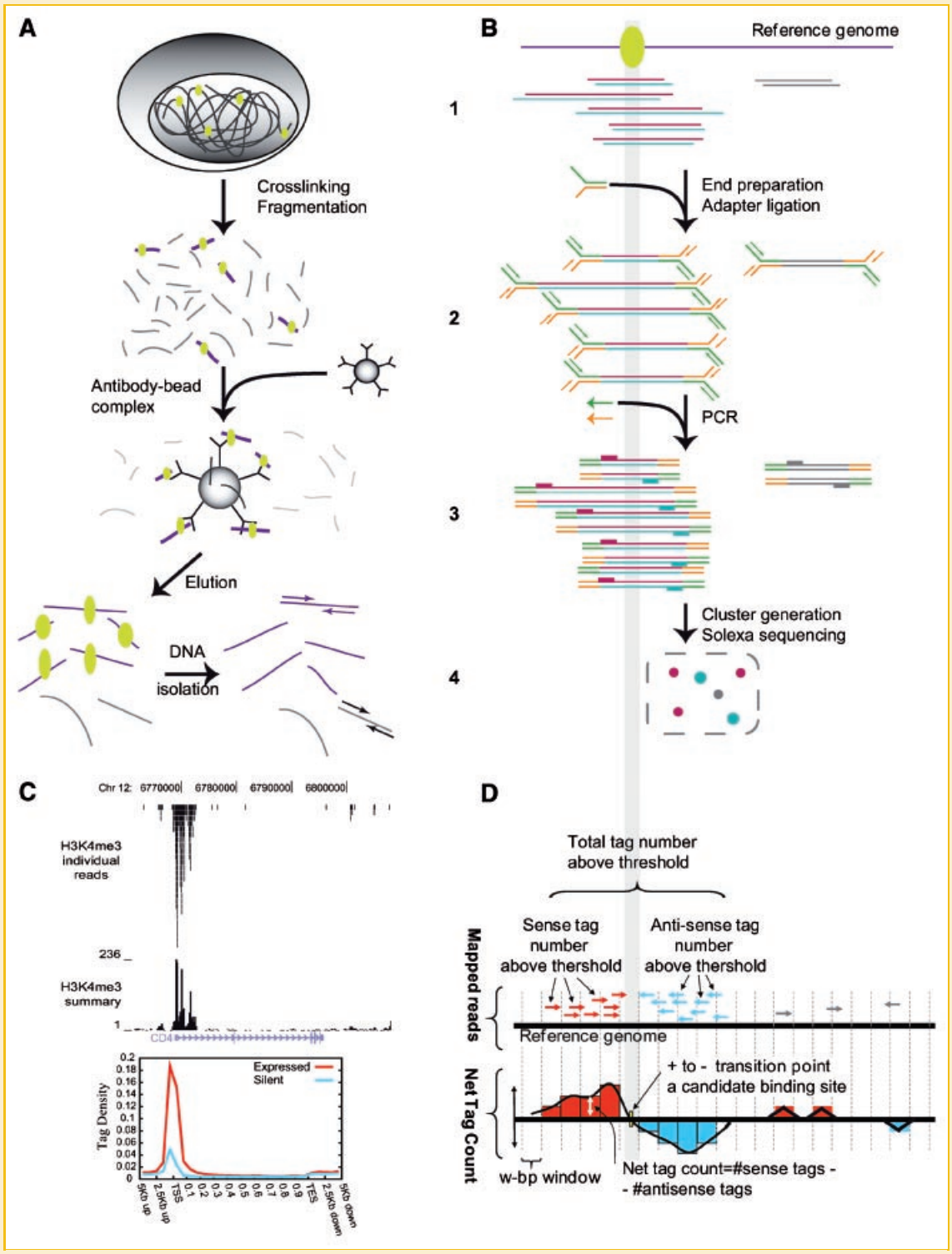

Fig. 1. A: Chromatin immunoprecipitation. 1. DNA is bound by a protein of interest in the nucleus. 2. Cells are lysed and DNA is fragmented. 3. Fragments bound by protein of interest (purple) are bound by antibody bead complexes and precipitated. Some DNA fragments can be precipitated nonspecifically(gray). 4. Protein–DNA complexes are eluted and DNA is purified. Relative abundance of specific and non-specific fragments is analyzed by qPCR. B: ChIP-Seq library construction. 1. Specific (colored) non-specific (gray) immunoprecipitated fragments are shown mapped to genome. 2. DNA termini are polished, phosphorylated, A is added and adapters are ligated. 3. Library is PCR amplified. 4. DNA fragments are hybridized to flowcell, clusters are synthesized and sequenced. C: Typical view of ChIP-Seq results. Results of H3K4me3 ChIP-Seq in $CD4^+$ T cells [Barski et al., 2007a] in the vicinity of CD4 gene are shown. Top track: Individual tags. Bottom track: Summary view bars show number of tags in 200 bp windows. Graph below shows average H3K4me3 tag density profile in the gene body (TSS) for expressed and silent genes. D: SISSRs algorithm. Net tag count is calculated as a number of sense strand tags minus number of antisense tags in small windows. The point where net tag count crosses 0 with a negative derivative is considered a binding site if sense, antisense and total tag numbers for the site are above certain threshold. See Jothi et al. [2008] for details.

Bucher, 2007; Barski et al., 2007b; Schones et al., 2008]. Chromatin is immunoprecipitated (IPed) using antibodies against factor of interest and protein A or G bound magnetic beads. Unfortunately some DNA fragments can be trapped and nonspecifically precipitated along with actual binding targets. After immunoprecipitation, crosslinks (if any) are reversed and DNA is purified. The ChIP DNA can be analyzed by PCR using gene-specific primers. Concentration of DNA fragments containing a region of interest are compared to those of known targets and unrelated sequences. The enrichment of a specific DNA region can vary dramatically depending on the protein factor, quality of antibody, and region of interest. Typically, ChIP experiments yield 5–50-fold enrichment of target regions, depending on the site, antibody, protein of interest and experimental conditions. This means that there are 5–50 times more fragments containing *real* binding sites compared to control DNA regions. In its conventional variant this method allows to test if a candidate promoter is bound by a transcription factor in living cells [see Hecht and Grunstein, 1999 for review]. However, conventional ChIP does not allow screening for novel targets in an unbiased manner.

Attempts have been made to identify binding sites in an unbiased manner based on ChIP. Low throughput approaches included combination of ChIP and EMSA [Bigler and Eisenman, 1994], direct cloning of IPed DNA fragments [Weinmann et al., 2001] and analyzing them using ChIP Display approach [Barski and Frenkel, 2004]. High throughput was achieved with the use of microarrays (ChIP-chip). In this approach IPed DNA was hybridized to microarrays and signal was compared to that obtained from input DNA or control ChIP. Initially, this technique was used in yeast [Ren et al., 2000; Iyer et al., 2001]. Higher eukaryotic genomes are larger and more repetitive, which presented a challenge to the technology. Weinmann et al. [2002] used CpG island microarray to identify E2F1 targets. Promoter arrays where ~1 kb pieces of promoters were spotted on microarrays were also used [Ren et al., 2002]. While these arrays allowed identification of genes regulated by TF they still lacked resolution and were biased towards certain genomic areas. To overcome these limitations, tiling microarrays containing short regularly spaced DNA fragments covering large genomic areas were developed. Cawley et al. [2004] used Affymetrix tiling microarrays to find targets of p53, c-myc and Sp1 on chromosomes 21 and 22. Nimblegen later produced custom synthesized microarrays offering greater flexibility. Kim et al. [2005] used these arrays to characterize all active promoters in the human genome. Applications of ChIP-chip are discussed in greater detail in another review of this series.

Directly sequencing IPed DNA fragments instead of hybridizing them to microarrays seems a simple idea, but in the past it was easier said than done. Getting meaningful results requires large amounts of sequencing. This led to development of approaches based on sequencing concatenated short pieces of IPed fragments in a way similar to SAGE, which allowed researchers to bring the amount of sequencing within the realm of possibility. Several versions of ChIP-SAGE were developed: GMAT was used to identify islands of histone acetylation in yeast [Roh et al., 2004] and human T cells [Roh et al., 2005, 2006]; SACO was used to find targets of CREB [Impey et al., 2004], beta-catenin [Yochum et al., 2007a], and TFIIB [Yochum et al., 2007b] binding sites, and ChIP-PET was applied to Oct4 and Nanog [Loh et al., 2006] and p53 [Wei et al., 2006]. ChIP-PET encompasses sequencing both ends of an IPed fragment and allows easier identification of a binding site. However, the cost of sequencing was still prohibitive and only the most well heeled laboratories could undertake it. The development of the next-generation massively parallel sequencing by 454 (now Roche), Solexa (now Illumina) and ABI (now Life Technologies) revolutionized the field. ChIP-Seq, as the combination of ChIP with the massively parallel sequencing was named, not only decreased the cost and allowed replicates, now identification of binding sites genome-wide can be accomplished with greater sensitivity and resolution than ever before. First applications of ChIP-Seq included localization of histone modifications in human T cells [Barski et al., 2007a; Wang et al., 2008] and in mouse ES cells [Mikkelsen et al., 2007], and mapping binding sites of the insulator binding protein CTCF and RNA polymerase II [Barski et al., 2007a], transcription factors Stat1 [Robertson et al., 2007] and NRSF [Johnson et al., 2007]. Methods for analyzing genome-wide epigenetic modifications have been reviewed previously [Schones and Zhao, 2008]. In the next sections we will discuss preparation of ChIP-Seq libraries, analysis of ChIP-Seq data and results obtained in early ChIP-Seq experiments conducted in our and other laboratories.

ChIP-Seq LIBRARY CONSTRUCTION AND SEQUENCING

To construct a ChIP-Seq library (Fig. 1B), the ends of IPed DNA fragments are blunted using DNA Polymerase I and phosphorylated using T4 kinase. This is followed by the addition of A (Adenine) using either Taq or Klenow exo-polymerase. An adapter of unusual structure, that allows for the automatic creation of an asymmetric PCR template, is ligated to both ends of the DNA fragment. The PCR amplified library is hybridized to a flowcell and clusters are prepared in a PCR-like process. Following linearization and blocking, clusters are sequenced in a sequencing-by-synthesis process.

Having a good quality immunoprecipitating antibody is the most important factor for ChIP-Seq success. There is an almost linear reverse relationship between the amount of sequencing required to detect a peak and enrichment. Assuming uniform distribution of background amount of sequencing S required to call a peak with *P*-value of 10^{-4} can be estimated from this relationship (see Supplementary Note for the derivation and assumptions):

$$S\frac{w}{G}(E-1.76)\frac{1}{(n(E-1)+1)} \geq 4.76$$

where G is the non-repetitive genome size, w is the window size, E is the enrichment (ratio of average tag number in a window containing a binding site to that in an unenriched window), and n is the fraction of total windows that have binding sites.

Assuming that only a small fraction of genome is enriched ($n \ll 1$), the condition for successful ChIP-Seq becomes:

$$S\frac{w}{G}(E-1.76) \geq 4.76$$

Of course this relationship is very crude and does not take many considerations into account [see Zhang et al., 2008b or Feng et al.,

2008 for discussion of more advanced statistical models], but it allows us to make several interesting conclusions:

- If enrichment is twice worse the amount of sequencing should be increased slightly more than twice in order to be able to call the same peaks.
- For higher resolution (smaller window size, w) more sequencing should be done.
- If there are many peaks (high n) more sequencing needs to be done.
- Minimum enrichment required for calling peaks at 1 tag per window average coverage is approximately 6.5.
- And the obvious: larger genome needs more sequencing.

In our work we used only Illumina sequencing technology and thus we are not able to comment on sample preparation for ABI Solid and other sequencers. Technical details of library preparations for Solexa sequencing were recently discussed in an excellent review from the Sanger sequencing center [Quail et al., 2008]. We will discuss only tips specifically applicable to ChIP-Seq.

Usually, the amount of immunoprecipitated DNA is quite low. This often results in problems during the adapter ligation step: adapters can become self-ligated. If not completely removed prior to PCR, these adapter-dimers will be amplified and form clusters on the flowcell. As a result, the number of aligned reads will be strongly decreased. Thus it is important to decrease the amount of adapters used in a ligation reaction. Illumina recommends using 0.1 μl of adapter for regular ChIPs. For histone ChIPs amount of IPed DNA is usually higher than for non-histone ChIPs thus more adapter can be used.

Given the small amount of DNA, cross sample contamination can be a problem. Especially dangerous is contamination of pre-PCR sample with post PCR library. To avoid this danger, one can use self-contained precast gels, for example, E-gels (Invitrogen) in library preparation. Even when using E-gels, pre-PCR and post-PCR samples should not be run on the same gel.

When performing ChIP-Seq for smaller genomes such as yeast, one lane often produces an excessive amount of data. To decrease the cost of experiment, it is possible to use bar coding. For this one can use adapters that have one or more specific nucleotides at the 3′ end. For example, adapters used for library 1 might have AG at the 3′ end and adapters for library 2 might have CT. After library preparation these two libraries can be mixed and sequenced in the same lane using the standard primer. After sequencing, those reads that start with AG will be used in analysis of library 1 and those with CT for library 2. This way experiment cost can be halved. An alternative way of multiplexing has been recently offered by Illumina. Barcodes can be integrated in the middle of an adapter and read after hybridization of a second sequencing primer on a machine with paired-end module.

Paired end sequencing also offers important opportunities. While more expensive, it will likely allow for easier data analysis especially in respect to TFBS identification and for nucleosome positioning studies. In addition, it will allow researchers to analyze areas of the genome that could not be analyzed using single read libraries: one will be able to map fragments overlapping with repetitive regions if one end of the fragment belongs to repetitive and another to non-repetitive sequence. This will be especially beneficial for analyzing reads in the vicinity of shorter repeats, such as tRNA derived repeats.

DATA ANALYSIS

Basic analysis of ChIP-Seq results is relatively straightforward, but requires significant computational and data storage capabilities. Typically, it can be conducted using the Illumina Analysis Pipeline within 1–2 days on a computer with at least eight processors and large memory. The pipeline measures cluster intensities from images, conducts base-calling and aligns the sequences to a genome. There are several alternatives to the programs used in the standard pipeline. Alternative base-calling can be done by Alta-Cyclic [Erlich et al., 2008] or Rolexa [Rougemont et al., 2008], which claim to produce more alignable reads than the Illumina pipeline. There are many more alternatives to the standard ELAND [Cox, in preparation] for alignment of the short reads to a genome. They differ in speed, maximum read length, number of allowed mismatches, ability to align with gaps, ability to use base quality scores produced by sequencer, and treatment of repetitive reads. The list includes RMAP [Smith et al., 2008] and MAQ (maq.sourceforge.net) that are slower but can use base quality scores and BowTie (bowtie-bio.sourceforge.net) that is claimed to be several times faster than ELAND. Also available are RMAP, Novoalign (www.novocraft.com), Seqmap [Jiang and Wong, 2008], SOAP [Li et al., 2008], ZOOM [Lin et al., 2008], and many others.

Alignment results can then be converted into browser extensible data (BED) or other browser readable format and displayed in a genome browser. ''Summary'' files containing the number of tags in windows of certain size instead of individual tags can be more convenient for data visualization (Fig. 1C). In order to observe sharper peaks top (bottom) strand tags can be shifted half of fragment length right (left) when making these files. Read numbers can be averaged across some genomic elements for example, all expressed genes to obtain average tag density profiles. In order to display the data one can use, for example, the UCSC genome browser on the Internet or as a local mirror. An example of a UCSC browser showing individual reads (top) and summary display (middle) for H3K4me3 ChIP-Seq in the vicinity of CD4 gene in T cells is shown in Figure 1C. Average H3K4me3 tag density profile for expressed and silent genes is shown on the bottom of the figure. Alternative browsers can be used: for example Argo [Engels et al., 2006] is a convenient browser for smaller genomes which can be run on a desktop machine without complicated setup. Other browsers include Ensembl, CisGenome [Ji et al., 2008], Apollo [Misra and Harris, 2006], Gene-Track [Albert et al., 2008], EagleView [Huang and Marth, 2008], and Gbrowse [Stein et al., 2002].

More advanced analysis is far less standardized. Usually researchers are interested in finding genomic regions that have significantly high number of tags. For this purpose genome can be divided into windows/bins of certain size. Random distribution of

tags into bins can be described by Poisson distribution and tag number threshold can be calculated for the desired *P*-value, window size and number of tags. More advanced background models can also be used [Feng et al., 2008; Zhang et al., 2008b].

Early peak finders used in Robertson et al. [2007] and Johnson et al. [2007] just searched for the regions of genome containing high tag number. While this produces general area of the peak, it does not give exact binding site location. Use of ChIP-Seq tag direction can deliver higher resolution. This information can be used in several ways. After peak identification peaks can be trimmed to narrow on binding site [FindPeaks; Fejes et al., 2008]. Alternatively peaks of (+) and (−) strand tags can be identified separately and binding site can be called between them [CisGenome; Ji et al., 2008, it also includes a user-friendly browser]. In the QUEST software [Valouev et al., 2008], the tags belonging to a top/bottom strand were shifted 1/2 of fragment size right or left, respectively, resulting in sharper peaks. A similar approach is used in MACS [Zhang et al., 2008a], which can also account for local enrichment biases. Another solution was realized in the SISSRs program [Jothi et al., 2008]. Since binding site has to be inside each immunoprecipitated fragment, top strand tags can be only to the left and bottom strand tags can be only to the right of the binding site (Fig. 1D). To find sites, a net scoring function is calculated as a number of top strand tags minus number of bottom strand tags in small windows. The binding site is called where the scoring function crosses zero. SISSRs offers precise identification of binding sites locations, high sensitivity and resolution of closely located binding sites. Another advantage is that for SISSRs to correctly map binding sites, the peak does not have to be symmetrical. With SISSRs, the standard deviation of distance between predicted binding site and actual consensus position can be as low as 13 bp for high enrichment CTCF sites [Jothi et al., 2008]. The smaller the size of immunoprecipitated fragments, the lower SD can be achieved. Use of paired-end sequencing will likely make finding sites easier and will allow for better resolution of closely located binding sites.

There are several sources of bias in the ChIP-Seq data: One is bias in library construction which results in under-representation of AT-rich regions with low melting temperature [see discussion in Quail et al., 2008]. Another source is PCR, in which high GC content areas are not well amplified. For chromatin prepared using micrococcal nuclease preparation of chromatin bias can result from MNase sequence preferences: for example, it does not like to cut between G and C bases. Furthermore, microrepetitivness of genome might result in ''dead zones''—positions in the genome to which tags cannot be mapped uniquely and which might appear unenriched, improperly.

Two more notes of caution: Firstly, some regions of the genome tend to be non-specifically enriched in ChIP-Seq. This can be a result of preferred fragmentation of ''open'' chromatin, non-specific immunoprecipitation, PCR bias or incorrect mapping of repeat derived sequence. The latter regions are likely to be represented by tags on only one strand. To avoid identifying these areas as binding sites it might be important to use an IgG or input DNA control library. Such a library can be sequenced once and used as a control for all future experiments performed with the same organism and in similar fragmentation conditions. Most of the programs referenced above can use such library as background. Secondly, while ChIP-Seq allows identification of enriched regions it does not mean that TF actually binds DNA there: enrichment can be achieved via protein–protein interactions and looping of DNA.

APPLICATIONS OF ChIP-Seq

The ChIP-Seq approach can be used for mapping binding sites of any DNA binding protein [Schones and Zhao, 2008]. What kind of results can be obtained with ChIP-Seq? Our laboratory used it to map most of histone modifications [Barski et al., 2007a; Wang et al., 2008] and nucleosome positioning [Schones et al., 2008] in $CD4^+$ T cells. Unlike earlier methods, ChIP-Seq produces data with nucleosome level or better resolution and genome-wide coverage. These data has allowed us to make a number of interesting findings. For example, we found that some modifications such as H3K27me3 and H3K9me3 occupy large domains, whereas others including H3K4me1/2/3 and H3K9ac are located in small loci several nucleosomes long (usually at promoters and/or enhancers). H3K4me1/2/3 and other ''positive'' marks usually mark active promoters, but some silent promoters also have these modifications. H3K36me3 and H3K79me2 both occupy gene bodies, but H3K79 peaks near the promoter and decays into the gene body, H3K36me3 is absent at the promoter but increases after the TSS. Most of known DNA breakpoints found in T cell cancer are located within areas marked with H3K4me3 (''open chromatin''), whereas non-T cell cancer breakpoints show much lower association with the localization of this modification in T cells, suggesting that open chromatin is more fragile. Some of these results could have been obtained using ChIP-chip but the cost for analysis of almost 40 modifications would likely be prohibitive. Furthermore, nucleosome level resolution obtained in these studies would not be possible with older methods.

Mikkelsen et al. [2007] used ChIP-Seq to analyze genome-wide distribution of several histone trimethylations in mouse embryonic stem cells, neural progenitor cells and embryonic fibroblasts. Thanks to the genome-wide coverage of ChIP-Seq, the authors were able to produce data confirming their hypothesis regarding the role of bivalent domains (regions where H3K4me3 and H3K27me3 marks co-localize) in plasticity and lineage commitment [Bernstein et al., 2007]. Our analysis of various histone modifications using ChIP-Seq during differentiation of human hematopoietic stem cells into erythrocyte precursors suggests that the direction of resolution of the bivalent modifications upon differentiation is associated with histone modification patterns at the stem cell stage [Cui et al., 2009]. We have also observed co-existence of H3K4me3 and H3K27me3 in T cells [Roh et al., 2006], where they may function to maintain plasticity and reversibility of differentiated T helper cell subsets [Wei et al., 2009].

Interestingly, the sequencing strategy allows to distinguish alleles using differences at SNPs and to analyze chromatin modifications at imprinting control regions [Mikkelsen et al., 2007]. This would have been more difficult to achieve with ChIP-chip due to the possibility of cross-hybridization.

Of course, ChIP-Seq was also used to map binding sites of transcription factors and other DNA-associated proteins. Robertson

et al. [2007] analyzed STAT1 binding in HeLa cells and compared results with a previous ChIP-chip study. ChIP-Seq confirmed ~70% of sites detected by ChIp-chip, but was much more sensitive detecting almost 4 times as many sites and authors were quite conservative: our re-analysis of data using SISSRs software resulted in further 77% increase in the number of detected STAT1 sites.

Johnson et al. [2007] studied binding of NRSF(REST) repressor in Jurkat cells. NRSF is one of the TFs with large binding sites (21 bp consensus), which allows computational identification of its targets with high significance. ChIP-Seq showed that most of computationally predicted sites indeed bound NRSF. Unexpectedly, however, NRSF also bound sites that lacked the consensus sequence. Motif finding showed that in some cases having lower homology "half-site" was sufficient for NRSF binding. Authors note that motif finding was greatly facilitated by high resolution and low false-positive rate of ChIP-Seq. Ontology analysis of experimentally determined NRSF transcriptional network produced more significant results compared to network produced by purely computational means. As an example of global coverage of ChIP-Seq results authors showed an NRSF regulatory network related to pancreatic beta cell development.

Combining ChIP-Seq data for various DNA associated proteins in the same and in different cell types allowed us to make interesting conclusions. For example, by combining data for CTCF and H3K27me3 in three different cell types we were able to establish the role of CTCF in chromatin domain barrier formation [Cuddapah et al., 2009].

CONCLUSIONS AND PERSPECTIVES

Currently the ChIP-Seq protocol is well developed and relatively easy to perform: ChIP has been performed for more than a decade and construction of ChIP-Seq library is no rocket science either. Post-sequencing, one run of Illumina Genome Analyzer produces hundreds of gigabytes of data per run and basic pipeline analysis on our server takes more than a day. Nonetheless, even though performing ChIP-Seq does require significant computational and data storage resources, basic analysis is relatively standard. The difficult part of ChIP-Seq experiment is more advanced data analysis. It has now been made easier thanks to peak calling tools described above, but use of these tools requires more computer expertise than molecular biologists typically possess. Thus development of user-friendly tools will likely make ChIP-Seq results more usable. Still simple creation of the long lists of targets is no longer a great achievement in itself: when embarking on a ChIP-Seq experiment, it is important to understand how this huge amount of data will be used. Creative processing of ChIP-Seq data requires close collaboration between biologists and bioinformaticians.

Discovery of all binding sites of transcription factor in a given cell type is a step to understanding the transcriptional network that regulates gene expression and function of this cell. Further, since number of tags found at a specific target site reflects strength of interaction, these data can be used to determine which target sites have higher affinity and will be occupied first upon the start of transcription factor expression. Target data for a number of transcription factors combined will enable the construction of predictive models of transcription control, gene expression and interaction of transcription factors and co-factor proteins.

Many researchers will likely use ChIP-Seq results to examine just several individual binding sites. When comparing binding between different samples it is relatively easy to compare localization of binding. Much more care should be taken when comparing binding levels at the same site between different samples: enrichment in a given ChIP experiment depends on many intractable parameters, likely including a phase of a moon. Thus, change in tag number at certain sites might result not only from "real" binding changes but also from experimental irregularities. On a positive side ChIP-Seq offers an internal control in the form of average tag density profiles for the control sites.

One of the drawbacks of the ChIP approach is the inability to distinguish between a binding event happening in the whole cell population and an event happening in only a few cells at a time. To this end, it is important to learn to reduce the number of cells required for ChIP-Seq. The ability to perform ChIP-Seq on a single cell level will provide answers to many interesting questions.

ChIP is not the only method that went genome-wide thanks to high-throughput sequencing. Now almost any method used for analysis of protein–DNA interactions at a single locus can be combined with sequencing for genome-wide coverage. In fact the new Seq techniques can be outright easier and retire older, often radioactivity based techniques. A prime example is DNase footprinting [Boyle et al., 2008]. This method is used to identify regions of open chromatin, which are considered candidate areas for enhancers, promoters and other genomic elements. In the past, DNase treatment was followed by indirect end-labeling or LM-PCR and gel separation or Southern blotting. Now for genome-wide coverage one can simply sequence the ends of DNase-cut fragments. This approach can be applied to other types of footprinting including, for example, $KMnO_4$ or DMS footprinting. Chromosome Conformation Capture (3C) [reviewed by Simonis et al., 2007] and its variants are also likely to benefit from next generation sequencing. 3C-like methods are used to identify areas of DNA that are brought together in the nucleus, for example as a result of looping. Deciphering long-range inter- and intrachromosomal interactions will help to understand how far-away enhancers can contribute to gene regulation.

Another application is mapping of nucleosome positions using footprinting with micrococcal nuclease. We used this approach to map positions of nucleosomes in human T cells [Schones et al., 2008]. This approach allowed us to discover specific nucleosomal organization of open promoters: unlike closed promoters, the first several nucleosomes in open promoters are positioned at highly specific distances from the transcription start sites. The analysis of nucleosome positioning will likely be much easier with the use of paired-end sequencing.

Currently number of tags produced by a single lane of Illumina GAII is sufficient for mapping most of narrowly distributed chromatin modifications, such as H3K4me1/2/3, but is not sufficient for mapping H3K27me3, which occupies larger share of genome or nucleosomes. Improvements in current sequencing technologies and development of new ones by Helicos [Harris et al., 2008], Pacific

Biosciences [Eid et al., 2009] will make mapping of such modifications easier. Longer reads, which are being promised by Illumina and ABI, will make a larger share of the genome mappable. Further, single molecule sequencing techniques promoted by several companies will eliminate the need for pre-sequencing PCR thus solving problems of biased coverage. Lastly, introduction of ''open source'' Polonator will likely drive down prices of instrumentation and reagents making ChIP-Seq more affordable.

High-throughput sequencing also revolutionized other areas of research in transcription and epigenetics. DNA methylation now can be studied using bisulfite sequencing [Cokus et al., 2008; Meissner et al., 2008]. Expression can be studied using SAGE-like Digital Gene Expression or RNA-Seq [Mortazavi et al., 2008; Nagalakshmi et al., 2008]. RNA-Seq can also provide data on alternative splicing and alternative promoter usage. Direct sequencing can be used to discover novel small RNAs [Stark et al., 2007] and to measure their concentration. Altogether these novel approaches will help to decipher secrets of gene regulation in cell function and differentiation.

ACKNOWLEDGMENTS

The authors thank Dustin Schones, Iouri Chepelev, and Raja Jothi for discussions and critically reading the manuscript. The research in the authors' laboratory was supported by the Intramural Research Program of the NIH, National Heart, Lung and Blood Institute.

REFERENCES

Albert I, Wachi S, Jiang C, Pugh BF. 2008. GeneTrack—A genomic data processing and visualization framework. Bioinformatics 24:1305–1306.

Barski A, Frenkel B. 2004. ChIP Display: Novel method for identification of genomic targets of transcription factors. Nucleic Acids Res 32:e104.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007a. High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007b. Response: Mapping nucleosome positions using ChIP-Seq data. Cell 131:832–833.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. Cell 128:669–681.

Bigler J, Eisenman RN. 1994. Isolation of a thyroid hormone-responsive gene by immunoprecipitation of thyroid hormone receptor-DNA complexes. Mol Cell Biol 14:7621–7632.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. Cell 132:311–322.

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. 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.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452:215–219.

Cox AJ. in preparation. Ultra high throughput alignment of short sequence tags.

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19: 24–32.

Cui K, Zang C, Roh T-Y, Schones DE, Childs RW, Peng W, Zhao K. 2009. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell 4:80–93.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133–138.

Engels R, Yu T, Burge C, Mesirov JP, DeCaprio D, Galagan JE. 2006. Combo: A whole genome comparative browser. Bioinformatics 22:1782–1783.

Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ., 2008. Alta-Cyclic: A self-optimizing base caller for next-generation sequencing. Nat Methods 5:679–682.

Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. 2008. FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24:1729–1730.

Feng W, Liu Y, Wu J, Nephew KP, Huang TH, Li L. 2008. A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. BMC Genomics 9 (Suppl 2): S23.

Greil F, Moorman C, van Steensel B. 2006. DamID: Mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. Methods Enzymol 410:342–359.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. 2008. Single-molecule DNA sequencing of a viral genome. Science 320:106–109.

Hecht A, Grunstein M. 1999. Mapping DNA interaction sites of chromosomal proteins using immunoprecipitation and polymerase chain reaction. Methods Enzymol 304:399–414.

Huang W, Marth G. 2008. EagleView: A genome assembly viewer for next-generation sequencing technologies. Genome Res 18:1538–1543.

Impey S, McCorkle SR, Cha-Molstad H, Dwyer JM, Yochum GS, Boss JM, McWeeney S, Dunn JJ, Mandel G, Goodman RH. 2004. Defining the CREB regulon: A genome-wide analysis of transcription factor regulatory regions. Cell 119:1041–1054.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409:533–538.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26:1293–1300.

Jiang H, Wong WH. 2008. SeqMap: Mapping massive amount of oligonucleotides to the genome. Bioinformatics 24:2395–2396.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein–DNA interactions. Science 316:1497–1502.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008. Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36:5221–5231.

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. 2005. A high-resolution map of active promoters in the human genome. Nature 436:876–880.

Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: Short oligonucleotide alignment program. Bioinformatics 24:713–714.

Lin H, Zhang Z, Zhang MQ, Ma B, Li M. 2008. ZOOM! Zillions of oligos mapped. Bioinformatics 24:2431–2437.

Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CW, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38:431–440.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454:766–770.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560.

Misra S, Harris N. 2006. Using Apollo to browse and edit genome annotations. In: Baxevanis AD, editor. Curr Protoc Bioinformatics. New York: Wiley, Unit 9 5.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. Nat Methods 5:1005–1010.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306–2309.

Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev 16:245–256.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657.

Roh TY, Ngau WC, Cui K, Landsman D, Zhao K. 2004. High-resolution genome-wide mapping of histone modifications. Nat Biotechnol 22:1013–1016.

Roh TY, Cuddapah S, Zhao K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev 19:542–552.

Roh TY, Cuddapah S, Cui K, Zhao K. 2006. The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci USA 103: 15782–15787.

Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F. 2008. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 9:431.

Schmid CD, Bucher P. 2007. ChIP-Seq data reveal nucleosome architecture of human promoters. Cell 131:831–832; author reply 832-3.

Schones DE, Zhao K. 2008. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet 9:179–191.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898.

Simonis M, Kooren J, de Laat W. 2007. An evaluation of 3C-based methods to capture DNA interactions. Nat Methods 4:895–901.

Sims JK, Houston SI, Magazinnik T, Rice JC. 2006. A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. J Biol Chem 281:12760–12766.

Smith AD, Xuan Z, Zhang MQ. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128.

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M. 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. Genome Res 17:1865–1879.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. 2002. The generic genome browser: A building block for a model organism system database. Genome Res 12:1599–1610.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5:829–834.

Voss TC, John S, Hager GL. 2006. Single-cell analysis of glucocorticoid receptor action reveals that stochastic post-chromatin association mechanisms regulate ligand-specific transcription. Mol Endocrinol 20:2641–2655.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40:897–903.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5:276–287.

Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y. 2006. A global map of p53 transcription-factor binding sites in the human genome. Cell 124:207–219.

Wei G, Wei L, Zhu J, Zang C, Hu-Li J, Yao Z, Cui K, Kanno Y, Roh TY, Watford WT, Schones DE, Peng W, Sun HW, Paul WE, O'Shea JJ, Zhao K. 2009. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. Immunity 30:155–167.

Weinmann AS, Bartley SM, Zhang T, Zhang MQ, Farnham PJ. 2001. Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol Cell Biol 21:6820–6832.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 16:235–244.

Yochum GS, McWeeney S, Rajaraman V, Cleland R, Peters S, Goodman RH. 2007a. Serial analysis of chromatin occupancy identifies beta-catenin target genes in colorectal carcinoma cells. Proc Natl Acad Sci USA 104:3324–3329.

Yochum GS, Rajaraman V, Cleland R, McWeeney S. 2007b. Localization of TFIIB binding regions using serial analysis of chromatin occupancy. BMC Mol Biol 8:102.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. 2008a. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9:R137.

Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M. 2008b. Modeling ChIP sequencing in silico with applications. PLoS Comput Biol 4:e1000158.