High-throughput methods of regulatory element discovery

Michael E. Hudson and Michael Snyder
Yale University, New Haven, CT, USA

With the number of organisms whose genomes have been sequenced, a vast amount of information concerning the genetic structure of an organism’s genome has been collected. However, effective experiment means to study how this information is accessed have only recently been developed. In this review, three basic methods for identifying regions of protein-DNA interaction will be introduced. The first two, chromatin immunoprecipitation (ChIP)-chip and ChIP-PET (for paired-end ditag), rely on the enrichment provided by chromosomal immunoprecipitation to interrogate the genomic sequence for the interaction sites of a protein of interest. In contrast, protein microarrays allow the identification of DNA binding protein that interacts with a DNA sequence of interest. These complementary methods of exploring protein-DNA interactions will increase our fundamental knowledge of how the information contained within the genome sequence is accessed and processed.

Introduction

Recent technological advances have allowed the application of high-throughput methods to the identification of transcription factor binding regions on a genomic scale. The combination of chromatin immunoprecipitation (ChIP) with microarray and other high-throughput technologies has enabled the identification of protein-DNA interactions on an unprecedented scale. With this ability, the identification of all regions of DNA where a given transcription factor binds has become a theoretical possibility.

With the sequencing of the genomic DNA of a variety of organisms, including yeast and humans, the question of how the expression of the genetic information contained within the genome is regulated becomes very important and pertinent. Recent efforts to investigate the interactions of transcription factor proteins with DNA on a genomic scale have underlined the complexity of the regulatory networks.

Several approaches can be used to identify the sites of protein-DNA interaction. If the consensus sequence that directs the binding of a protein to a specific region of DNA is known, this information can be used in a bioinformatics approach to identify other potential binding regions. If the binding site requirements are not known, techniques such as systematic evolution of ligands by exponential enrichment can be used to elucidate this information (1). Electrophoretic mobility shift assays can also be used to determine binding specificity (2). While these approaches can be used to identify potential target sites of DNA binding proteins, their utility is limited for a variety of reasons. Technically, such approaches are labor-intensive and provide very low-throughput. From the standpoint of biology, many additional factors are likely to be involved in the interaction of a protein with a specific region of DNA, including other associated proteins, the local chromatin structure, and interactions with other, often far removed regions of chromatin.

Previous work in our laboratory, and in others’, has led to the development of a number of techniques that allow the high-throughput identification and analysis of DNA binding by a variety of DNA binding proteins. In both mammalian (3,4) and yeast (5,6) systems, high-throughput microarray-based analysis has been used to study the binding of transcription factor proteins to potential regulatory DNA elements. These analyses demonstrate the ability of these high-throughput methods to confirm and expand our knowledge of previously identified protein-DNA interactions, as well as to identify previously unknown interactions. At the present time, many of these techniques are being used by the ENCODE Project Consortium, a group of laboratories whose aim is to identify all functional elements within the human genome (7).

ChIP-on-Chip: Microarray Analysis of Protein-DNA Interactions

The most dramatic improvement in the throughput of identifying transcription factor-DNA interactions came with the combination of microarray analysis with the ChIP technique. Initial uses of ChIP required prior knowledge of a potential binding target DNA sequence to identify binding of protein to DNA regions of interest. By the combining microarray analysis with ChIP, the identification of protein binding sites was possible on a genomic scale, limited only to the DNA features present on the microarray.

The standard protocol for ChIP-on-Chip (or ChIP-chip as it is sometimes called) is shown in Figure 1. As with the basic ChIP method, a DNA binding protein is covalently attached to its target DNA using formaldehyde cross-linking. After shearing of the chromatin, the protein-DNA adduct is immunoprecipitated from nuclear extracts using an antibody specific for the protein in question or for an epitope sequence appended to the end of the protein. Commonly
used in this manner are the hemagglutinin or myc tag sequences. After immunoprecipitation, the enriched DNA sample is commonly, but not always, amplified using a variety of methods. The final step is to fluorescently label the enriched DNA and the reference DNA. In the case of yeast ChIP, the reference sample is usually composed of DNA from the mock immunoprecipitation of a strain where the target transcription factor is untagged. In mammalian ChIP, the reference is usually composed of the input sheared chromatin or mock immunoprecipitations using normal, nonspecific sera from mice or rabbits.

Once the DNA samples have been amplified and labeled, the control and enriched samples are hybridized to various microarrays. Each type of microarray has advantages and disadvantages and is defined by three characteristics. These are coverage, density, and resolution. Coverage refers to the proportion of the genomic sequence that is physically present on the array. At present time, repetitive DNA sequences or highly homologous regions are not represented on arrays. As well, microarrays known as intergenic arrays contain sequences that intervene between the individual genes present within the genome sequence. Microarray density refers to the individual strands of DNA that are present on the arrays. The individual elements that make up the array, known as features, can be made up of PCR-amplified DNA sequences, commonly used in the construction of intergenic microarrays or oligonucleotides of varying lengths. The term resolution refers to the length of the intervening genomic sequences between the individual DNA sequences present on the array, when superimposed on the genome. Arrays with high resolution have smaller intervening sequences not represented on the array.

The two most commonly used microarray formats used for the analysis of ChIP-chip samples are PCR product arrays and oligonucleotide arrays. In PCR arrays, each of the DNA features on the array is amplified using PCR, and these products are mechanically spotted onto glass slides. Due to the limited packing density that can be achieved, this form of microarray is limited to 20,000 to 30,000 features.

Modern oligonucleotide arrays are generated using a fundamentally different technique. In contrast to PCR arrays, the oligonucleotides are “grown” in situ through a variety of techniques, for instance using photolithography in a manner having many parallels to the microchip fabrication process. The individual oligonucleotides are then extended in place using synthetic oligonucleotide chemistry. By synthesizing each feature in situ, the density of the array can be much greater than that of mechanically spotted arrays; where each high density oligonucleotide array can have over 1 million features. This dramatic increase in feature density makes these oligonucleotide arrays an obvious choice when identifying transcription factor binding sites on a genome-wide scale.

One of the major limiting factors in ChIP-chip experiments is the expression of the tagged DNA binding protein or the availability of an antibody sufficiently specific for the protein of interest. One method that overcomes this limitation is the direct probing of double-stranded microarrays using purified protein. This method allows direct quantitation of the kinetic properties of proteins binding to the DNA strand. A modification of the ChIP-chip protocol, termed DIP-chip, for DNA immunoprecipitation with microarray detection, has been developed in order to address these potential limitations. Developed by Clarke and colleagues, purified protein is mixed with naked genomic DNA fragments in vitro. The DNA binding protein of interest is expressed as a fusion with a protein such as maltose binding protein, allowing affinity purification of the protein fusion. Protein-bound DNA is isolated through affinity purification, and the enriched DNA is amplified. Subsequent analysis of protein-DNA interaction is performed in a similar manner as standard ChIP-chip.

The benefit of DIP-chip is that the concentration of the protein available to interact with DNA is known, and the conditions in the interaction reaction can be modified. As well, the presence of the affinity fusion partner relieves the need for well-characterized antibodies with high sensitivity and specificity. However, while DIP-chip does allow identification of protein binding sites, it does not allow the characterization of changes in protein-DNA interaction that occur in response to various stimuli or cellular processes. The in vivo
The nature of standard ChIP-chip allows not only the determination of protein binding sites within the DNA sequence, but allows one to observe changes in protein-DNA interaction during cellular processes and in response to stimuli (10).

### High-Throughput Sequencing Analysis of Protein-DNA Interactions

Recent advances in DNA sequencing have allowed the application of high-throughput DNA sequencing methods to the analysis of protein-DNA interactions. Two similar methods, termed ChIP-STAGE (for sequence tag analysis of genomic enrichment) and ChIP-PET (for paired-end ditag) have been developed and are used to map regions of protein-DNA interaction in both yeast and human cells.

Developed by Iyer and colleagues, ChIP-STAGE was initially used to analyze targets of the TATA-box binding protein in yeast and the transcription factor E2F4 in human cells (11). Described by Ruan and colleagues in their analysis of p53 genomic binding sites, ChIP-PET combines chromatin immunoprecipitation with the paired-end ditag strategy of high-throughput sequencing (12). While DNA sequencing of ChIP-enriched DNA has been previously described, it is ultimately limited in its effectiveness by the underlying rate of DNA sequencing (12–14). While immunoprecipitated DNA is enriched for sequences bound by the specific DNA binding protein, a significant amount of contaminating DNA will be present. Traditional cloning and sequencing methods are limited in their ability to extract the enriched DNA fragments from the background noise. However, by increasing the number of fragments sequenced, thereby increasing the amount of DNA sampled, the enriched DNA fragments can be identified. This approach was recently used to explore the regulatory networks of the transcription factors Nanog and Oct4 and their involvement in the regulation of stem cell pluripotency (15).

In the ChIP-PET protocol (Figure 2), chromatin bound by DNA binding proteins is enriched as described for the ChIP-chip protocol. In contrast to ChIP-chip, the enriched DNA sequences are randomly cloned into a plasmid vector where the ChIP DNA fragment is flanked by restriction enzyme sites. These plasmids are cut at these sites with the type II restriction enzyme Mmel, which cleaves 18 bp into the cloned genomic DNA fragment. The single-stranded ends of the plasmid are blunt-ended and ligated to each other to form short 36-bp ditag signature sequences within the plasmid, combining 18 bp from the 5’ end and 18 bp from the 3’ end of the ChIP DNA fragment. Together, the two halves of the ditag signature sequence demarcate the full length of the ChIP fragment originally present. The ditags derived from all the enriched ChIP DNA fragments are then isolated, concatenated together, and cloned into a second plasmid, creating the ChIP-PET library. Concatenation of the short sequence tags allows more efficient sequencing of the PET library and greatly increases the sequencing depth achieved, with 10–15 PETs analyzed per sequencing read. This library is then extensively sequenced, and the sequences resulting from the enriched ChIP DNA fragments...
are mapped back to the genome. By identifying overlapping PET sequences, the sites of protein-DNA interaction can be determined. Previous methods with limited depth of sequencing were ineffective at extracting true protein-DNA interactions from the underlying nonspecific noise. However, ChIP-PET, through the use of small sequence pairs and high-throughput sequencing, increases the depth of sequencing and increases the signal of true binding sites. Nonspecific fragments will randomly generate PETs, essentially producing single ditags, while true binding sites will produce clusters of multiple overlapping PET sequences, the overlapping regions of which denote the protein binding region.

The use of sequencing versus microarray analysis to identify enriched DNA sequences removes an important limitation inherent to ChIP-chip. As mentioned above, ChIP-chip-based analysis is limited to areas of the genome that are represented on the microarray used. ChIP-PET overcomes this limitation by potentially identifying all enriched sequences; therefore it is potentially a more open and less biased method.

However, this requirement for a substantial amount of sequencing (approximately 40,000 sequencing reads) leads to the major limitation of ChIP-PET, which is cost. While microarray-based analysis of DNA-protein interaction is relatively inexpensive, the magnitude of sequencing required to perform ChIP-PET effectively may limit its availability to most researchers. The development and adoption of high-throughput sequencing methods may alleviate this restriction.

**Analysis of DNA-Protein Interaction Using Protein Microarrays**

A novel approach to the identification of protein-DNA interactions is the direct application of labeled DNA fragments to protein microarrays. In this application, DNA containing the binding sequence of interest is labeled using a fluorescent moiety and incubated with a microarray composed of purified proteins. An early study introduced the potential of using purified human proteins arrayed on a solid support to investigate protein-DNA interactions. In a more recent study, oligonucleotides were used to probe microarrays containing approximately 300 known and potential transcription factors from the yeast _Saccharomyces cerevisiae_. In this study, protein microarrays were probed with fluorescently labeled DNA probes containing concatemers of protein-DNA binding sequences.

The results from this work highlighted the usefulness of this approach, with a number of known protein-DNA interactions observed. As well, a number of previously unknown protein-DNA interactions were observed. In addition, probing of protein microarrays with labeled DNA probes resulted in the identification of a DNA binding and transcriptional regulatory role for the metabolic enzyme Arg5,6, a protein with no previously recognized DNA binding activity.

One major benefit of using protein microarrays to analyze the binding of DNA sequences by proteins is the ability to easily increase the resolution at which one is analyzing the binding interactions. For instance, in addition to identifying regions of DNA where proteins bind, these interactions can be further explored by making modifications in the sequence of the DNA probe. These modifications allow the determination of the consensus sequence that encodes the binding site, determination of the spacing requirements between the components of the consensus sequence, and finally how each portion of the consensus sequence is involved in the physical interactions that occur between the nucleotides and the interacting protein.

Finally, it is possible to expand this method to develop protein microarrays that contain the approximately 1000–1500 mammalian transcription factors. While ChIP-based methods allow identification of DNA interaction events for one DNA binding protein, protein microarray-based analysis of protein-DNA interactions allows a reverse analysis, in which the binding of one DNA sequence of interest by multiple proteins is assayed. By combining these two fundamentally different approaches to addressing the same experimental question, an accurate and robust map of protein-DNA interactions can be drawn.

**Acknowledgments**

The authors would like to thank G. Euskirchen for critically reading this manuscript and providing helpful comments.

**References**

circuitry at the G1/S transition in *Saccharomyces cerevisiae*. Genes Dev. 16:3017-3033.


