

RESEARCH HIGHLIGHT

DNA-protein interactions in high definition

Eric M Mendenhall^{1,2,3} and Bradley E Bernstein^{*1,2,3}

Abstract

An elegant, genome-wide approach to define the precise DNA sequences bound by transcription factors has been developed by Rhee and Pugh.

Cis-acting DNA sequence elements are crucial to the proper regulation of transcription. The identification and functional characterization of these elements, often DNA sequence motifs, has progressed unevenly. For example, although the yeast Gal4 motif is well characterized and widely used in transgenic applications, up to half of yeast transcription factors (TFs) have opaque target sequences. ChIP-exo, a new method developed by Rhee and Pugh [1], seeks to assay DNA-protein binding at high resolution in order to fill in the *cis*-regulatory motif blanks.

The elusive functional motifs

To identify motifs in an unbiased, high-throughput manner, experimentalists have used *in vitro* methods such as SELEX [2] and protein-binding microarrays [3]. Such studies have accumulated a large catalogue of position weight matrices for DNA-binding motifs that is available in databases, including JASPER and TRANSFAC.

Whereas well-described motifs in *Escherichia coli* comprise as many as 24 bp, motifs in *Drosophila* and other metazoans tend to be much shorter, often comprising just 6 or 8 bp [4]. When analyzed in the large genomes of eukaryotes, such short motifs are commonplace. Cells must employ mechanisms to highlight the functional motifs or mask the decoy motifs *in vivo*. One major mechanism, identified through DNase hypersensitivity experiments, is accessibility of motifs in open DNA regions [5]. These nucleosome-free regions are also flanked by distinct chromatin signatures. In any cell type, a small fraction of the genome, including promoters and enhancers, adopts such an open chromatin configuration.

Thus, the majority of motifs are found in inaccessible regions with higher nucleosome occupancy and chromatin compaction. Notably, the combination of DNase hypersensitivity assays with high-throughput sequencing can enable high-resolution footprinting of bound sequence motifs, albeit without direct identification of the interacting DNA-binding protein [6,7]. The DNase approach can be complemented by direct mapping of TF-binding sites through combination of chromatin immunoprecipitation with deep sequencing (ChIP-Seq).

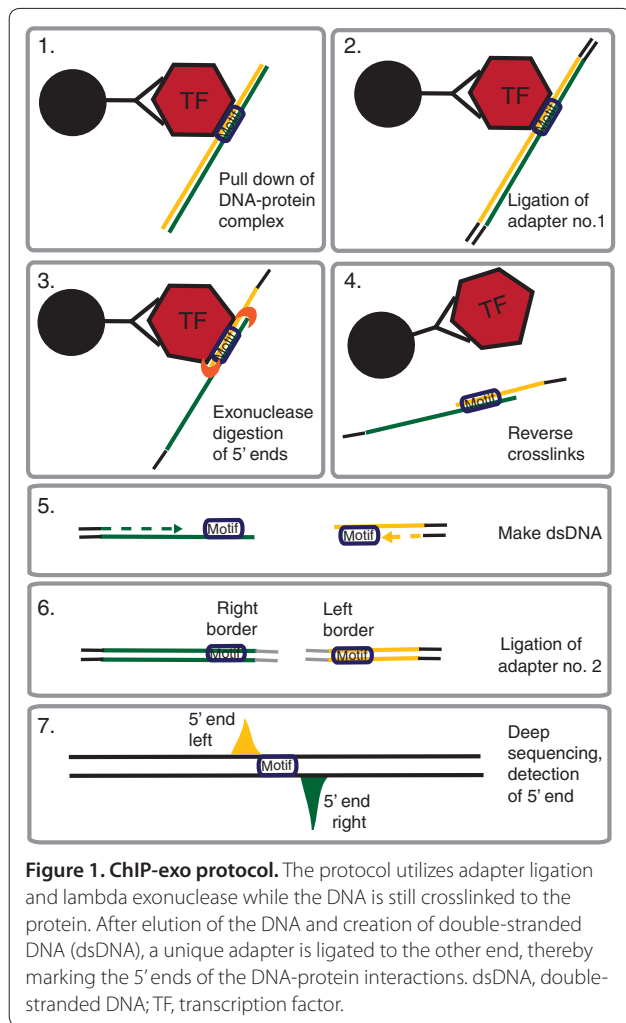
Despite biochemical and molecular evidence linking TFs to specific motifs, a high proportion of binding sites returned in systematic TF-mapping studies have lacked the consensus motifs. Although the correspondence between binding and motif presence improved with the transition from array-based approaches to ChIP-seq, it remained that half or more of all detected sites could not be linked to a cognate motif [8]. Computational efforts to define alternative or degenerate motifs have been successful [9], but have been hindered by the modest resolution of ChIP-Seq. Additional filters, such as conservation across species, have helped narrow down the candidate regions, but can only infer the relevant TF or cell type. These analyses are likely to be limited by biological complexity, such as DNA methylation status of the motif, competitive or collaborative binding of multiple TFs, DNA looping, low occupancy sites, and stability or half-life of DNA-protein interactions. Taken together, these limitations point to a critical need for additional experimental methods capable of defining, at base pair resolution, protein-DNA interaction *in vivo*.

ChIP-exo improves resolution

In a recent issue of *Cell*, Rhee and Pugh [1] present a new method called ChIP-exo to improve identification of DNA motifs by significantly narrowing down the region of protein binding. The technique combines aspects of TF ChIP and DNase footprinting for base pair resolution of binding sites. Specifically, the authors utilized lambda exonuclease, which degrades DNA in a 5' to 3' direction. When lambda exonuclease is added near the end of a ChIP protocol, it removes only one strand of the DNA until it is blocked by the DNA-protein crosslink (Figure 1). After digestion, the authors proceed to reverse the crosslink, and ligate a second adapter, which gives directionality

*Correspondence: Bernstein.Bradley@mgh.harvard.edu

¹Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
Full list of author information is available at the end of the article



to the DNA fragments. The clever method produces a defined boundary of DNA-protein interaction while still leaving enough intact DNA in the 3' direction for aligning sequencing reads, thus enabling the authors to delineate the extent of the bound region with base pair resolution.

Rhee and Pugh used ChIP-exo on four proteins in yeast and one protein in human cells. The single base pair resolution of ChIP-exo allowed for new insights into DNA-protein interactions. For example, they found yeast Reb1 is bound exactly 95 bp upstream of transcription start sites, and often coincides with a second Reb1 binding event of lower occupancy 40 bp away. As these two binding events are relatively close, they would be indistinguishable from a single binding event in ChIP-Seq. Additionally, the authors investigated the binding of Phd1, a transcriptional activator with an ambiguous motif. They found Phd1 binds two distinct motifs, and a third degenerate motif, possibly explaining its motif ambiguity. To the extent that ChIP-exo sequencing reads represent protein occupancy, all three motifs had the

same occupancy of Phd1 binding, indicating protein binding is not always decreased at degenerate motifs.

Rhee and Pugh also investigated Rap1, a unique DNA binding protein that binds at both telomeres and ribosomal genes. The telomere-bound sites had a GT-rich 27 bp motif, while non-telomere binding had three versions of a 12 to 13 bp motif. The motifs had significant heterogeneity that the authors suggest is compensated for by the length of the motif. Interestingly, they note that Rap1 is bound to regions previously identified as containing nucleosomes. They propose that lambda exonuclease chews through DNA that is wrapped about histones, yet stops at the Rap1 DNA crosslink. The authors do not speculate on why the nucleosomal DNA is relatively more susceptible to exonuclease digestion, but nonetheless the TF-binding event is detected by ChIP-exo.

Modular motifs

Lastly, the authors used the new method to map CTCF in human HeLa cells. A minority of CTCF is bound at promoters; it is not a TF in the traditional sense, but likely plays more diverse roles in genome folding, insulation and boundary formation. It has been shown *in vitro* that CTCF utilizes combinations of its 11 zinc-finger domains to bind to different DNA sequences [10]. Using the new method, the authors identified six related motifs, composed of combinations of four modules. CTCF occupancy increased with the increasing number of modules present, and only half of CTCF binding occurred at the motif most similar to the canonical CTCF motif, which contains modules no. 2 and no. 3.

The data of Rhee and Pugh provide insights into TF binding at low occupancy sites, which they argue plays a biologically functional role, as opposed to random TF scanning or noisy binding events. They found that low TF occupancy correlates with either single nucleotide substitutions in the underlying motif (as seen with Reb1), or a decrease in the number of modules within a motif (as seen with CTCF). However, Rhee and Pugh note that low occupancy sites have similar characteristics to high occupancy sites, such as distance to transcription start sites. As ChIP-exo is done on additional TFs, particularly in mammalian models, it will be interesting to see if this holds as a common theme, or if more diverse binding modes will be found. Additionally, the current data do not address the mechanisms for low occupancy binding, such as a protein binding for shorter lengths of time due to reduced binding energy, competitive binding of other proteins, or rapid removal by transcriptional machinery.

In conclusion, ChIP-exo presents a valuable complement to ChIP-Seq, DNase footprinting and chromatin mapping in pursuit of a more precise understanding of the biochemical interactions that mediate context-specific genome function.

Abbreviations

bp, base pair; TF, transcription factor.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA. ²Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, MA 02114, USA. ³Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Published: 27 January 2012

References

1. Rhee HS, Pugh BF: **Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.** *Cell* 2011, **147**:1408-1419.
2. Cui Y, Wang Q, Stormo GD, Calvo JM: **A consensus sequence for binding of Lrp to DNA.** *J Bacteriol* 1995, **177**:4872-4880.
3. Bulyk ML, Gentalen E, Lockhart DJ, Church GM: **Quantifying DNA-protein interactions by double-stranded DNA arrays.** *Nat Biotechnol* 1999, **17**:573-577.
4. Bird A: **The dinucleotide CG as a genomic signalling module.** *J Mol Biol* 2011, **409**:47-53.
5. Weintraub H, Groudine M: **Chromosomal subunits in active genes have an altered conformation.** *Science* 1976, **193**:848-856.
6. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA: **Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting.** *Nat Methods* 2009, **6**:283-289.
7. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS: **High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells.** *Genome Res* 2011, **21**:456-464.
8. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED: **Comparative epigenomic analysis of murine and human adipogenesis.** *Cell* 2010, **143**:156-169.
9. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nat Biotechnol* 2008, **26**:1293-1300.
10. Ohlsson R, Renkawitz R, Lobanenkov V: **CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease.** *Trends Genet* 2001, **17**:520-527.

doi:10.1186/gb-2012-13-1-139

Cite this article as: Mendenhall EM, Bernstein BE: DNA-protein interactions in high definition. *Genome Biology* 2012, **13**:139.