Cell

Resource

Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution

Ho Sung Rhee[1] and B. Franklin Pugh[1,*]
[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA
*Correspondence: bfp2@psu.edu
DOI 10.1016/j.cell.2011.11.013

SUMMARY

Chromatin immunoprecipitation (ChIP-chip and ChIP-seq) assays identify where proteins bind throughout a genome. However, DNA contamination and DNA fragmentation heterogeneity produce false positives (erroneous calls) and imprecision in mapping. Consequently, stringent data filtering produces false negatives (missed calls). Here we describe ChIP-exo, where an exonuclease trims ChIP DNA to a precise distance from the crosslinking site. Bound locations are detectable as peak pairs by deep sequencing. Contaminating DNA is degraded or fails to form complementary peak pairs. With the single bp accuracy provided by ChIP-exo, we show an unprecedented view into genome-wide binding of the yeast transcription factors Reb1, Gal4, Phd1, Rap1, and human CTCF. Each of these factors was chosen to address potential limitations of ChIP-exo. We found that binding sites become unambiguous and reveal diverse tendencies governing in vivo DNA-binding specificity that include sequence variants, functionally distinct motifs, motif clustering, secondary interactions, and combinatorial modules within a compound motif.

INTRODUCTION

Proteins bind to specific DNA sequences to regulate genes. A fundamental and long-sought goal in understanding how these interactions have evolved and their mechanism of regulation is the precise determination of where they are bound in a genome. Chromatin immunoprecipitation (ChIP) is the most widely used method to identify genomic binding locations of sequence-specific regulatory proteins (Solomon and Varshavsky, 1985). In the ChIP assay, proteins are crosslinked to their DNA-binding sites in vivo and then immunopurified from fragmented chromatin. Subsequently, the bound DNA is identified genome-wide by microarray hybridization (ChIP-chip) or deep sequencing (ChIP-seq) (Albert et al., 2007; Johnson et al., 2007; Ren et al., 2000).

Because unbound DNA contaminates the immunoprecipitate, ChIP only provides a set of statistically enriched high-occupancy binding regions, rather than a complete and precise set of bound locations (Peng et al., 2007; Rozowsky et al., 2009; Tuteja et al., 2009). A sizeable fraction of this DNA may represent false positives (erroneous calls), and many other lower-affinity sites may be missed (false negatives). Moreover, size heterogeneity of randomly sheared ChIP DNA technically limits mapping resolution, and thus cannot distinguish binding among clusters of neighboring sites.

Motif searches are insufficient to identify all in vivo binding locations for a protein because proteins recognize a wide variety of related sequences, of which only a small fraction are bound (Badis et al., 2009; Walter and Biggin, 1996). Consequently, although a consensus target motif may be extracted from data as a whole, a large fraction of putatively bound locations either lack an obvious motif or contain multiple degenerate versions of the motif (Cawley et al., 2004; Yang et al., 2006) and thus cannot be definitively assigned to a particular recognition sequence.

Protein-binding microarrays have proven to be powerful in defining a DNA-binding domain's intrinsic specificity in vitro (Badis et al., 2009). However, in vivo, such specificity may be altered, prevented, or constrained in the context of the thousands of other proteins that constitute the nuclear milieu. Digital genomic footprinting can detect highly occupied binding sites at high resolution (Hesselberth et al., 2009), but identifying the source of protected genomic footprints requires a priori knowledge of which protein binds to the identified sequence. Problematically, different proteins may bind to the same sequence. Importantly, low-occupancy binding is widespread in genomes (Li et al., 2008), but its physiological importance and distinction from noise have been difficult to discern by any assay thus far.

Here, we develop ChIP-exo, to precisely map a comprehensive set of protein-binding locations genome-wide in any organism and to greatly diminish both erroneous and missed calls associated with mapping. Importantly, ChIP-exo achieves near single-base resolution. The resulting maps provide a striking display of genome-wide site utilization that vividly delineates the variation in sequence recognition specificity and the underlying principles that drive specificity in vivo. From these binding events, potential mechanisms of site evolution, chromatin interplay, and genome-wide network regulation become clearer.

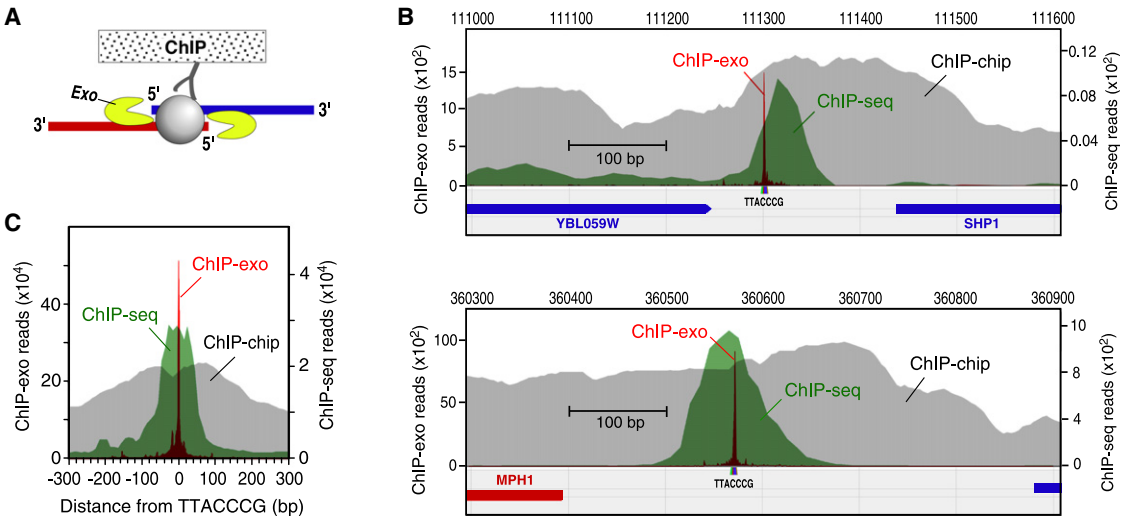

Figure 1. Single Base-Pair Resolution of ChIP-exo

(A) Illustration of the ChIP-exo method. ChIP DNA is treated with a 5′ to 3′ exonuclease while still present within the immunoprecipitate. The 5′ ends of the digested DNA are concentrated at a fixed distance from the sites of crosslinking and are detected by deep sequencing (see also Figure S1).

(B) Comparison of ChIP-exo to ChIP-chip and ChIP-seq for Reb1 at specific loci. The gray, green, and magenta filled plots, respectively, show the distribution of raw signals, measured by ChIP-chip using Affymetrix microarrays having 5 bp probe spacing (Venters and Pugh, 2009), ChIP-seq, and ChIP-exo. Sequencing tags on each strand were shifted toward the 3′ direction by 14 bp so as maximize opposite-strand overlap.

(C) Aggregated raw Reb1 signal distribution around all 791 instances of TTACCCG in the yeast genome. The ChIP-seq and ChIP-exo datasets included 2,938,677, and 2,920,571 uniquely aligned tags, respectively.

See also Figure S1 and Table S1.

RESULTS

ChIP-exo Design

We considered the possibility that a protein covalently crosslinked to DNA would block strand-specific 5′-3′ degradation by lambda (λ) exonuclease (Figure 1A), thereby creating a homogeneous 5′ border at a fixed distance from the bound protein. DNA sequences 3′ to the exonuclease block remain intact and are sufficiently long to uniquely map to a reference genome, after identification by deep sequencing (Figure S1A available online). Uncrosslinked nonspecific DNA is largely eliminated by exonuclease treatment, as evidenced by the repeated failure to generate a ChIP-exo library from a negative control BY4741 strain.

ChIP-exo Improves Genome-wide Mapping Accuracy and Sensitivity

We initially focused on the yeast Reb1 protein, which has a clear DNA recognition site (TTACCCG) that can be used for independent validation (Badis et al., 2008; Harbison et al., 2004). Reb1 is involved in many aspects of transcriptional regulation by all three yeast RNA polymerases and promotes formation of nucleosome-free regions (NFRs) (Hartley and Madhani, 2009; Raisner et al., 2005). It is also found at telomeres. We compared ChIP-exo to ChIP-chip and standard sonication-based ChIP-seq.

The unfiltered ChIP-exo signal was highly focused across the genome at TTACCCG sequences (Figures 1B and 1C). ChIP-chip and ChIP-seq displayed broader signals. When converted to peak-pair calls (described below), ChIP-exo displayed a standard deviation (SD) of 0.3 bp (Figure S1B), which indicates that ChIP-exo of Reb1 has single-base accuracy. In comparison, ChIP-seq displayed more than 90-fold greater mapping variability (SD = 24 bp). ChIP-exo also displayed lower raw background. The raw signal-to-noise ranged from 300- to 2800-fold (Table S1). Subsequent employment of noise filters produced a comprehensive set of bound locations. In contrast, ChIP-chip and ChIP-seq had 7- and 80-fold raw signal-to-noise, respectively. ChIP-exo retained its quantitative properties, in that occupancy levels correlated with those from ChIP-seq (Figure S1C), and peak-pair intensities correlated (Figure 2A).

Reb1 Has Multiple Highly Organized Secondary Interactions at Promoters

The 5′ ends of ChIP-exo tags (as well as peaks) located on one strand were largely at a fixed distance (~27 bp) from another tag or peak on the other strand, corresponding to the two exonuclease barriers formed by Reb1 (Figures 2A, and S2A, and S2B). A total of 1,776 Reb1 peak pairs were identified (Data S1). Importantly, these peak pairs were not preselected based upon the presence of any DNA sequence motif, although a motif was present in nearly all cases.

Of the peak pairs, 60% (1,058/1,776) were classified as primary locations, and 40% (718/1,776) as secondary. Secondary locations were defined as less-occupied locations within 100 bp of a more-occupied location. Thus, most Reb1 locations were found in clusters. Nearly all (92%) primary locations contained the TTACCCG Reb1 recognition site or a single-nucleotide variant centered between its borders (Figures 2A, 2B, and S2C). Increased deviations from TTACCCG