BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 13 2011, pages 1867–1868
doi:10.1093/bioinformatics/btr279

Genome analysis

Advance Access publication May 5, 2011

BINOCh: binding inference from nucleosome occupancy changes

Clifford A. Meyer[1,*], Housheng H. He[1,2], Myles Brown[2] and X. Shirley Liu[1,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health and [2]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, 02115, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Transcription factor binding events are frequently associated with a pattern of nucleosome occupancy changes in which nucleosomes flanking the binding site increase in occupancy, while those in the vicinity of the binding site itself are displaced. Genome-wide information on enhancer proximal nucleosome occupancy can be readily acquired using ChIP-seq targeting enhancer-related histone modifications such as H3K4me2. Here, we present a software package, BINOCh that allows biologists to use such data to infer the identity of key transcription factors that regulate the response of a cell to a stimulus or determine a program of differentiation.

Availability: The BINOCh open source Python package is freely available at http://liulab.dfci.harvard.edu/BINOCh under the FreeBSD license.

Contact: cliff@jimmy.harvard.edu; xsliu@jimmy.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 14, 2011; revised on April 5, 2011; accepted on April 21, 2011

1 INTRODUCTION

Although transcription factors are known to have a high affinity for specific DNA sequences, binding is often cell type and condition specific, therefore DNA sequence alone is a poor predictor of *in vivo* genome-wide binding locations (Carroll *et al.*, 2006). When the relevant transcription factors are known and suitable antibodies are available, ChIP-chip and ChIP-seq technologies enable us to map their genome-wide binding locations. In many cases, however, the transcription factors governing a regulatory response are not known or antibodies suitable for ChIP-seq are not available. Without transcription factor-specific antibodies, it is nevertheless possible to infer transcription factor binding events based on two observations. First, transcription factor binding sites are often associated with certain types of post-translational histone modifications, in particular histone H3 lysine 4 mono- (H3K4me) and di-methylation (H3K4me2) (Heintzman *et al.*, 2007). Second, transcription factor binding is frequently associated with a pattern of nucleosome occupancy changes in which nucleosomes flanking the binding site increase in occupancy, while those in the vicinity of the binding site itself are displaced (He *et al.*, 2010). Transcription factor (TF) binding events can therefore be inferred by comparing nucleosome resolution H3K4me1/2 ChIP-seq data under treatment and control conditions. We present a software package, BINOCh, to allow biologists to carry out an analysis of nucleosome occupancy data to discover stimulus-induced transcription factor binding.

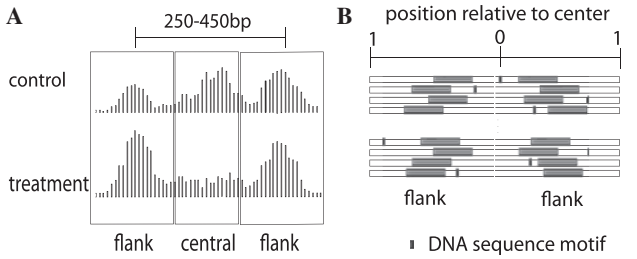

Fig. 1. (**A**) Pattern of nucleosome occupancy changes associated with transcription factor binding in the treatment condition. (**B**) Analysis of the tendency for transcription factor binding to be located toward the midpoint between paired nucleosomes. The midpoint defines the relative position 0, while the edges fix the relative position 1. Within each region the highest scoring DNA sequence element is associated with a relative position between 0 and 1. Sequence motifs detected in the flanking nucleosome positions are excluded from the calculation.

2 METHOD

2.1 Identification and scoring of candidate transcription factor binding regions

The first step in the analysis is to use histone-related ChIP-seq data to detect well-positioned nucleosomes, which we do with the nucleosome positioning from sequencing (NPS) package (Zhang *et al.*, 2008) available at http://liulab.dfci.harvard.edu/NPS. The second step, shown in Figure 1A, is to find pairs of nucleosomes with center to center distances characteristic of nucleosomes flanking TF binding sites (between 250 and 450 bp) and to generate a table summarizing ChIP-seq tag counts in these regions. For every nucleosome pair, we count the number of sequence reads within 100 bp of the center of each positioned 'flanking' nucleosome under both treatment and control conditions, denoted $n_{flank,t}$ and $n_{flank,c}$, respectively. We also count the number of sequence reads mapped to the 'central' region between two flanking regions, denoted $n_{central,t}$ and $n_{central,c}$. To generate this table we provide an easy to use python script, tagtab. We measure nucleosome occupancy change with the nucleosome stabilization–destabilization (NSD) score,

$$s = (\sqrt{n_{flank,t}/N_t} - \sqrt{n_{central,t}/N_t}) - (\sqrt{n_{flank,c}/N_c} - \sqrt{n_{central,c}/N_c})$$

where N_t and N_c are the total number of sequence reads in treatment and control. In the final step, using the python script binoch, we compute NSD scores and search through DNA motif libraries, TRANSFAC (Wingender *et al.*, 2000), JASPAR (Sandelin *et al.*, 2004) and UniPROBE (Berger and

*To whom correspondence should be addressed.

Bulyk, 2009), to detect motifs that are associated with the central regions of high NSD scoring regions.

2.2 Motif position bias toward the midpoint of central regions

Our model for transcription factor binding is one in which transcription factor binding sites are located near the midpoint between paired nucleosomes. To assess a trend of DNA sequence motif occurrence toward this midpoint, we compute the mean relative location of motif hits. Statistical significance is derived using a null model where it is assumed that DNA motifs not associated with binding sites are uniformly distributed relative to the midpoint. We conduct this analysis on 600 bp DNA segments from the N highest NSD-scoring set of paired nucleosomes. Figure 1B shows how each 600 bp segment is derived from a 1 kb sequence centered at the midpoint between a nucleosome pair from which 200 bp centered on each of the pair of well-positioned nucleosomes is excluded. All DNA subsequences within these sequences are scored by a known motif position weight matrix using a low-order Markov model to account for the genomic background sequence composition. For each nucleosome pair i, we obtain two values: s_i, the maximum motif score for that pair and x_i, a value between 0 and 1 representing the location of the motif position relative to the midpoint. Sorting paired regions by motif score from high to low, we compute a list of z scores, $z_j = \sum_{i=1}^{j}(x_i - 0.5)/\sqrt{j/12}$, that represent the positional bias of a motif toward the centers of these regions. We calculate the P-value for a motif from the minimum z-score, making an adjustment for N.

2.3 Motif enrichment in high-scoring NSD regions

The motif centrality statistic, while telling us that a motif is occurring in the DNA sequence between paired nucleosomes does not allow us to assess whether this motif is enriched in the high NSD-scoring regions. To make this assessment, BINOCh computes a P-value motif enrichment statistic where the frequency with which a motif is found within the set of high NSD-scoring regions is compared with its occurrence frequency within a control set of regions. We select the control regions as those having NSD scores closest to the median of the NSD-score distribution. The analysis is carried out on the 200 bp of DNA sequence centered on the midpoint between paired nucleosomes. Motifs that are significant by both enrichment and centrality measures are likely to be bona fide transcription factor motifs important to the system.

3 CASE STUDY

We applied this approach to nucleosome resolution H3K4me2 ChIP-seq data from the prostate cancer cell line LNCaP under the control condition and in response to stimulation by the androgen receptor (AR) agonist 5a-dihydrotestosterone (DHT) (He *et al*., 2010). The BINOCh position analysis of this data produces a ranked list of motifs with their associated P-values (Supplementary Table 1), showing the AR motif to be the most strongly associated with the midpoint of high NSD scoring nucleosome pairs; in this analysis TRANSFAC progesterone receptor motif which is the same as the AR motif obtains the most significant P-value ($<$1e-8). The FoxA1 motif is also biased toward the midpoint ($P<$1e-6). This observation is biologically meaningful as FoxA1 is thought to be a 'pioneer factor' facilitating the binding of AR. The BINOCh motif enrichment analysis (Supplementary Table 2) also finds a TRANSFAC AR motif to be the most significant ($P<$1e-48). See http://liulab.dfci.harvard.edu/BINOCh to download the data and further results of this analysis.

ACKNOWLEDGEMENTS

We thank Jun Song, Tao Liu, Hyunjin Gene Shin and Bo Jiang for contributing elements of this software.

Funding: National Institutes of Health (R01 HG4069).

Conflict of Interest: none declared.

REFERENCES

Berger,M.F. *et al*. (2009)Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc*., **4**, 393–411.

Carroll,J.S. *et al*. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet*., **38**,1289–1297.

He,H.H. *et al.* (2010) Nucleosome dynamics defines transcriptional enhancers. *Nat. Genet*., **42**, 343–347.

Heintzman,N.D. *et al*. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet*., **39**, 311–318.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Wingender,E. *et al*. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

Zhang,Y. *et al*. (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, **9**, 537.