Supplementary Information

BINOCh: Binding Inference from Nucleosome Occupancy Changes

Clifford A. Meyer, Housheng H. He, Myles Brown and X. Shirley Liu

Supplementary Methods

Nucleosome Stabilization-Destabilization (NSD) Score Calculation

The aim of this analysis is to identify transcription factor binding events at specific genomic loci that occur under *treatment* conditions and not under *control* conditions. Such transcription factor binding events are often associated with a pattern of nucleosome occupancy changes in which two well-positioned nucleosomes *flanking* either side of the transcription factor binding site display higher occupancy after the binding event than before. A weakly positioned *central* nucleosome in the vicinity of the binding site and located between the *flanking* nucleosome shows lower occupancy on transcription factor binding. The nucleosome stabilization-destabilization (NSD) score allows nucleosome resolution histone modification ChIP-seq to be used to quantify this effect. Nucleosome pairs having center-to-center distances of 250-450 bp are selected as candidate flanking nucleosomes. For every such 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 $n_{central,c}$. The NSD score is defined as,

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

where N_t and N_c are the total number of sequence reads in treatment and control paired nucleosome regions. The NSD score compares ChIP-seq tag count data between treatment and control. If the number of counts under the different conditions were unequal a change in the absolute count in one condition would not be directly comparable with the count in the other. For this reason it is appropriate to scale by the factors N_t and N_c. From the Poisson statistics of count data we know that the variance of a Poisson random variable scales with the mean. To mitigate the trend of high variability for higher counts we use the square root variance-stabilizing transformation.

Centrality Statistic Calculation

This section describes the statistic used to assess the bias of a set of DNA motif locations towards the midpoint of a set of genomic intervals. The key assumption is that any motif not associated with transcription factor binding in these intervals will be observed at random positions relative to the midpoints of these intervals.

1. The list of N DNA sequences is scanned with a known position weight matrix. For each sequence, i, in this list let the highest motif match score be s_i and the position of this match within this sequence be x_i. If the best match score occurs at the midpoint of the sequence the position of this score within the sequence, x_i is 0. If the match occurs at either end of the sequence x_i is 1.
2. Each sequence in this list is now associated with a pair of numbers (s_i,x_i) for $i = 1,\ldots,N$. The list of pairs (s_i, x_i) is sorted by motif match score from highest score to lowest.

3. As a null hypothesis it is assumed that if a motif is not associated with transcription factor binding the position of the motif will be randomly distributed across the DNA sequences. In other words, each x_i is a random variable drawn from a uniform 0-1 distribution. Starting at the top of the list, for each entry *j,* the following z-score is calculated based on all hits up to that position in the list:

$$z_j = \sum_{i=1}^{j} \left(x_i - 0.5\right) \Big/ \sqrt{j/12}$$

At position *j* in the list the mean position of hits is $\frac{\sum_{i=1}^{j} x_i}{j}$ and the expected mean position is 0.5. The variance, *j/12 ,* is derived for the j-fold convolution of uniform distributions.

4. The p-value is based on the minimum z-score, $\min_{j \in L,\dots,N} z_j$. In practice we use a normal distribution to approximate the null distribution that is the j-fold convolution of uniform distributions. The normal approximation does not hold for small values of *j*, therefore, we exclude z-scores for *j* less than a lower bound, *L = 20.* Because the test statistic is based on the minimum of multiple z-scores an adjustment needs to be made to account for this z-score optimization. The p-value, based on the standard normal distribution, is therefore calculated from the following adjusted z-score,

$$z^* = \frac{\min_{j=L,..,N} z_j - (\alpha_0 + \alpha_1 \log(\log(N))}{\beta_0 + \beta_1 \log(\log(N))}$$

where the parameters $\alpha_0, \alpha_1, \beta_0$ and β_1 have been calculated using simulated data from the null distribution.

Supplementary Tables

Supplementary Table 1 shows the 10 TRANSFAC motifs with the highest p-values using the position bias criterion. All motifs closely resemble either the AR motif or the FoxA1 (HNF3alpha) motif.

TRANSFAC ID	Symbol	Consensus	p-value
M00957	PR	.......G.AC....TGTTCT....	1.40e-09
M00956	AR	......GG.AC....TGTTCT....	1.14e-08
M00447	AR	AGTAC.T.WTGTTCT	2.50e-08
M00953	AR	......GG.ACA..GTGTTCT....	5.12e-08
M01012	HNF3	.....TGTTTR.......	1.75e-07
M00481	AR	GG.ACA...TGT.CT	2.66e-07
M00791	HNF3	....ACAAACA..	3.38e-06
M00954	PR	.......G.A.....TGTTCT....	4.42e-06
M00724	HNF3alpha	TGTTTGTTTT	8.38e-06
M00292	Freac-4	CTTAAGTAAACA....	5.56e-05

Supplementary Table 1. 10 most significant motifs detected by BINOCh position bias analysis of Nucleosome Stabilization-Destabilization scores in DHT stimulated LNCaP cell line.

Supplementary Table 2 shows the 10 TRANSFAC motifs with the highest p-values using the enrichment criterion. The top 9 motifs closely resemble the AR motif while the 10th resembles the FoxA1 (HNF3alpha) motif.

TRANSFAC ID	Symbol	Consensus	p-value
M00481	AR	GG.ACA...TGT.CT	4.21e-48
M00956	AR	......GG.AC....TGTTCT....	6.39e-45
M00953	AR	......GG.ACA..GTGTTCT....	1.69e-42
M00957	PR	.......G.AC....TGTTCT....	1.38e-33
M00192	GR	...........TGT.CT..	2.83e-31
M00955	GR	.......G..C....TGTTCT...	4.82e-29
M00921	GR	..TGT.CT	1.67e-27
M00954	PR	.......G.A.....TGTTCT....	3.90e-26
M00960	PR	...AGAACA.	1.49e-23
M00290	Freac-2	.....GTAAACA....	2.77e-14

Supplementary Table 2. 10 most significant motifs detected by BINOCh enrichment analysis of Nucleosome Stabilization-Destabilization scores in DHT stimulated LNCaP cell line.