Supplementary Methods

The algorithm for ChIP-seq BELT peak calling program

Given the unique features of ChIP-seq, which identifies only the ends of a proportion of the DNA fragments in the ChIP sample, there might be numerous systematic biases such as errors during sequencing and alignment process and the degree of repetitiveness; therefore, we have developed a false-discovery-rates defined **B**in-based **E**nrichment **T**hreshold **L**evel (**BELT**) program which employs percentile rank scores and utilizes Monte-Carlo simulation to generate background data to identify the target loci from in vivo binding data performed by the ChIP-seq.

The algorithmic strategy including a pre-processing method is comprised of the following five steps:

Step 1. Develop a pre-processing method to process millions of raw reads (short sequences varying from 27bp to 45bp), map these reads onto a reference genome to get unique mapped reads, and then compile and sort those unique mapped reads based on chromosome numbers.

Step 2. Develop a bin-based method to score regions for a sum of the unique mapped reads to a GFF or BED formatted files. The formatted GFF and BED files can be visualized on a UCSC Genome Browser or Integrated Genome Browser.

Step 3. Apply a percentile rank statistic method to determine each level of percentile scores for the top percentile from Top 0.1% to 10% at the lowest level.

Step 4. Generate a background model of binding peaks by applying Monte-Carlo simulation and identify the number of the binding sites for each percentile rank.

Step 5. Estimate the false discovery rate (FDR) using the formula **2** to measure the significance of identified targets.

For a percentile rank *d* for a test statistic $\boldsymbol{Z}_k$, we want to test the null hypothesis,

$$\mathbf{H}_k0\colon \boldsymbol{E}(\mathbf{P}_k) = 0 \qquad \mathbf{1)}$$

where for peaks $\mathbf{P}_k$, *k*=1,…,**B**, ***E*** is the expected value of the number of false positive peaks among all claimed true **B** peaks in that level *d*. In this case, we defined this **E** value as a false discovery rate (**FDR**).

$$\mathbf{FDR}(d) = \mathbf{E}\left[\frac{\mathbf{FP}(d)}{\mathbf{TP}(d)}\right] \qquad \mathbf{2)}$$

where **FP**(*d*) is the number of the true false positive peaks in the level *d*, **TP**(*d*) is the number of peaks claimed as true peaks in the level *d*.

The procedure to generate a background data

The background data is a certain number of randomly distributed reads in the human genome. The basic idea for generating background data is to apply our **BELT** program to these data to determine the number of peaks called from randomly distributed data.

A detailed procedure of generating background data:

1. The genome reference sequences were downloaded from UCSC genome website (http://genome.ucsc.edu/) to our local server and were used for the reference sequences for generating the background data.
2. The number of reads and length of each read in the background data are exactly the same as those in the input data which have been pre-determined.
3. Randomly pick chromosomes for read generating points weighted by chromosome length (in base pairs). Generate an array to indicate how many reads should be on each of the chromosomes in human genome.
4. Randomly distribute certain number of read generating points on each chromosome according to the array.
5. The reads were generated according to read generating points, and output to the background data file.

The method for scoring a called peak

A score for a called peak by our **BELT** program is empirically defined in formula **3** and is used to rank the peaks in a particular percentile. We take several factors into account, the length of a peak, the average score of bins, the "peakedness" of a peak.

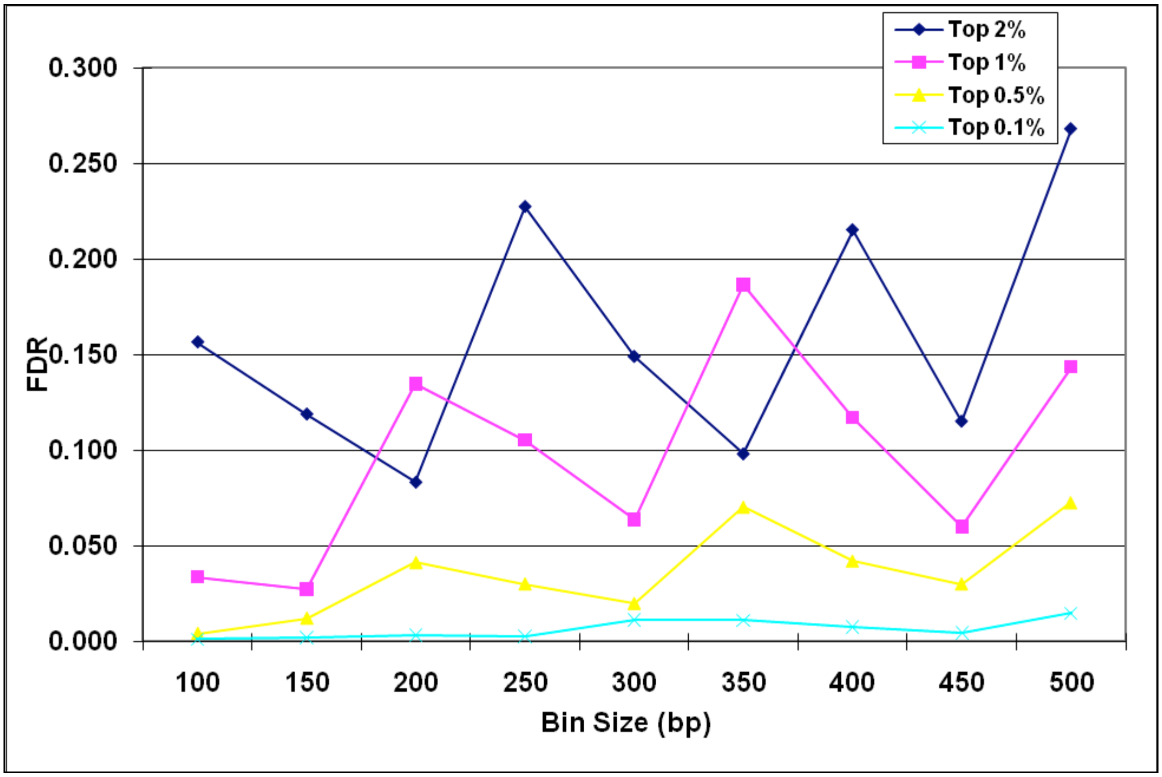
$$Sp = Log_2\left(\sqrt{n} \cdot S_a\right) \qquad \textbf{3)}$$

where S_p: the score of a called peak; S_a: the average score of each bin in the peak; n: the number of bins in the peak defined as $n = L_p / L_w$, L_p: the length of the peak; L_w: the width of a bin.

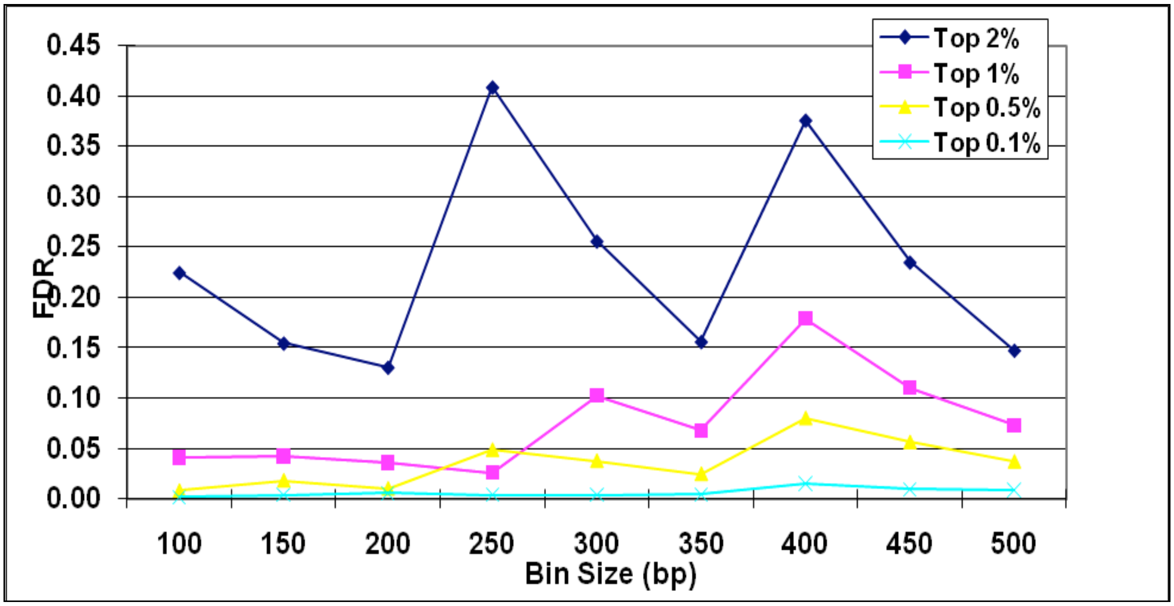
Importantly, by multiplying the average score and the number of bins in the peak, it will increase the weight of a peak's shape and determine the order of peaks. For example, supposed there are three peaks, **Peak 1**: $\mathbf{L_w}$ 150, $\mathbf{L_p}$ 500, $\mathbf{S_a}$ 5, **Peak 2**: $\mathbf{L_w}$ 150, $\mathbf{L_p}$ 300, $\mathbf{S_a}$ 5, **Peak 3**: $\mathbf{L_w}$ 150, $\mathbf{L_p}$ 500, $\mathbf{S_a}$ 3; If only considering $\mathbf{S_a}$, **Peaks 1** and **2** would be same rank (score of 5); If simply multiplying $\mathbf{L_p}$ x $\mathbf{S_a}$ two factors, then **Peaks 2** and **3** would be same rank (score of 1500); However, based on our formula **3**, the ranking order are **Peak 1** ($\mathbf{S_p}$ =3.19), **Peak 2** ($\mathbf{S_p}$ = 2.82), **Peak 3** ($\mathbf{S_p}$ = 2.45). Therefore, in general, peaks with higher average score and width shape are ranked the highest, followed by peaks with higher average score and narrow shape, then peaks with lower average scores and width shape, peaks with lower average score and narrow shape are ranked the lowest.

To determine the parameters (bin-size and percentile level) used for identification of ZNF263 binding sites

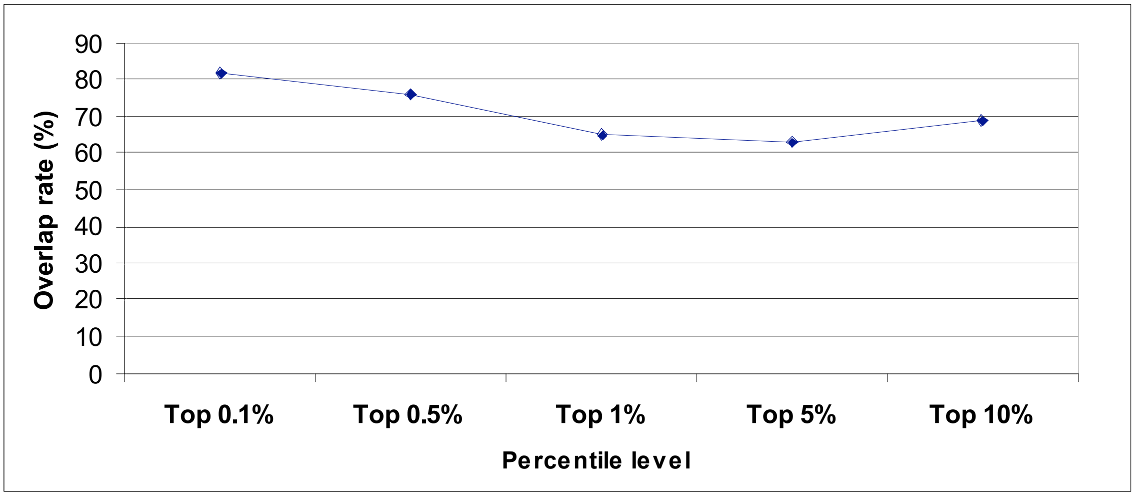
We tested a set of different combinations of various bin sizes from 100 to 500 bp with a 50 bp increment and several percentile levels from the Top 0.1% to 10% on ZNF263 ChIP-seq data in order to get a set of optimized parameters. Shown in **Suppl. Methods Figure SM1A** for Rep A and **Suppl. Methods Figure SM1B** for Rep B, the FDR rates for both replicates at both Top 0.1% and 0.5% levels are smaller than 0.1 for all bin-sizes, where they can be used as parameters for determining the significant set of binding sites. We then calculate the overlapping number of ZNF263 targets from both datasets (**Suppl. Methods Figure SM2**), and found that the overlapping percentage is decreasing along with increased percentile levels. A set of binding sites identified in the Top 0.5% level of replicate A and B were overlapped by 76%, therefore, reported as a significant set of ZNF263 binding sites for all other analysis.



Supplementary Methods Figure SM1A A plot of FDR *vs.* Bin Size for Rep A showing that both the Top 0.1% and 0.5% levels have an FDR smaller than 0.1 for all bin-sizes.



Supplementary Methods Figure SM1B A plot of FDR *vs.* Bin Size for Rep B showing that both the Top 0.1% and 0.5% levels have an FDR smaller than 0.1 for all bin-sizes.



Supplementary Methods Figure SM2 A plot of Overlapping rate *vs*. Percentile level showing that the overlapping percentage decreases along with increased percentile levels.