Supplementary Methods, Tables and Figures

Calculation of an average fragment length for a ChIP sample

The read enriched area on both strands should be paired together at a target site and show a bimodal pattern. In the first step, BELT sets a very high threshold (0.9999) and finds >=100 well-paired peaks, (if less than 100 pairs are found, the threshold is reduced and the process is repeated until more than 100 are found). Then, BELT calculates the average fragment length at each of these high significant target sites in Formula 1),

$$\bar{l} = \frac{\sum P_r}{n_r} - \frac{\sum P_f}{n_f} + 1 \qquad 1)$$

where P_f and P_r are the read's position on the forward and reverse strands respectively and n_f, n_r represent the number of reads on the forward and reverse strands respectively.

To avoid excluding long fragments that have ends falling out of the enriched area, we extend the area by 200 bp on each side. The whole genome-wide average fragment length is computed $L = \bar{\bar{l}}$ and used to further shift reads, search for possible target site etc.

Decoding the ChIP fragment position

To estimate the fragment positions of the ChIP sample, we shift all reads towards the mid-point of the fragment by $L/2$ whereby each of these resulting points is considered as a representation of one fragment.

Determination of significant enrichment level thresholds

We define a series of bins, varying by default from 100 to 500bp, by evenly dividing the genome and counting the density of reads for each bin. Then, the user's input defines significant enrichment level thresholds based on a percentile ranking statistic method. After sorting the enrichment level for all bins, the threshold is taken at the percentile of the confidence level. By default, five thresholds are defined from 0.95 to 0.999 percentile levels.

Definition of a peak and localization of the target site

A peak is defined as a set of continuous bins that have an average enrichment level higher than the threshold. Our algorithm only allows gaps with one bin in length, since larger gaps might indicate that there are two closely located peaks. After a peak is defined, we calculate the exact target site for that peak using Formula 2) with the assumption that the fragments are symmetrically distributed around a target site.

$$P_m = \frac{\sum (P_f - d) + \sum (P_r - d)}{n_t} \qquad 2)$$

where P_m denotes the exact binding motif, modification site etc, n_t is the number of reads in the region, P_f and P_r are the reads' position on the forward and reverse strand respectively and d is the distance shifted (equals $L/2$).

Data normalization

By default, robust linear regression is applied to normalize data with different sequencing depth since we assume that all experiments were performed under the same controllable conditions. First, enrichment levels of 10k bp bins are counted for each sample. Then a normalization factor is calculated by performing robust linear regression between two samples' bin enrichment levels. Two samples are adjusted to have a comparable enrichment level using the normalization factor.

Calculation of a p-value

BELT performs Fisher exact tests and calculates a *p*-value for each peak if sample comparison is performed. Peaks with a *p*-value less than 0.05 are defined as passed peaks, and others are considered rejected peaks. Both passed and rejected peaks are recorded as output.

Ranking the resultant peaks

Peaks are ranked by a score which measures their "quality/significance" and is empirically defined in Formula 3). This score is also 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, and the shape of a peak.

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

where S_p is the score of a peak; S_a is the average reads count of bins in the peak; m denotes the number of bins in the peak defined as $m = L_p / L_w$, L_p: the length of the peak; and L_w: the width of a bin.

Importantly, the score will increase the weight of a peak's shape and determine its order. Therefore, in general, if two peaks contain same amount of reads, the narrow one is favored, in another case, higher enriched peak is favored among peaks with same width.

Estimation of False Discovery Rate (FDR)

For a percentile rank r and a test statistic Z_k, we want to test a null hypothesis,

$$H_k0: E(P_k) = 0 \qquad 4)$$

where for peaks P_k, k=1,…, n, E is the expected value of the number of false positive peaks among all claimed true n peaks in that level r. In this case, we define this E value as a false discovery rate (FDR).

$$FDR(r) = E\left[\frac{FP(r)}{TP(r)}\right] \qquad 5)$$

where $FP(r)$ is the number of true false positive peaks with level r and $TP(r)$ is the number of peaks claimed as true peaks with level r.

Practically BELT generates simulated datasets to compute FDR, where each dataset includes simulated peaks and background noise reads based on the real ChIP-seq data. The procedure of data generation is described in the following section (Generation of synthetic, simulated background data).

Let N_b denote the number of peaks formed by simulated background data and N_t denote the total number of peaks detected from ChIP data. The FDR can be re-written as

$$FDR = \frac{N_b}{N_t} \cdot 100\% \qquad 6)$$

Generation of synthetic, simulated background data

We define a *signal-noise-ratio* (*SNR*) level as the following Formula 7), which is the basis for the Monte-Carlo simulation of data:

$$SNR = \frac{R_s}{R_o} \qquad 7)$$

Where R ($R = R_o + R_s$) denotes the total number of reads, R_s denotes the number of reads that fall into peaks, and R_o denotes the number of reads that are not in any peak.

A). Simulated peaks (target sites)

(a) Randomly generate *n* target sites (e.g. 10 nt in length) on the genome; (b) For each site, generate a certain number of artificial fragments that mimic the ChIP sample; (c) Randomly define each fragment's length, varying from 100-300 nt; (d) Randomize each fragment's position around the site, all fragments should cover the target site; (e) For single end sequencing record the coordinates of one end of each fragment, for paired end sequencing, both ends are recorded, whereby each end represents one read (the simulation is based on a real sample where the repeated regions have already been removed by the peak finding process; thus, we do not need to exclude the repeats regions in our simulation process).

B). Randomly generate background noise reads

(a) After obtaining the coordinates of the reads in ***A)***, we randomly generate coordinates for background fragments throughout the genome. The number of fragments is equal to *R(1-SNR)* for single end sequencing or *R(1-SNR)*/2 for paired end sequencing. (b) We record the coordinates of the one or two end(s) of the fragments as reads. If aberrant genome flag is on, an amplification factor will be calculated for amplified regions. An amplified region is defined as a large non-pericentromeric non-repetitive genomic region (>=10kb) that has a significant high input enrichment compare to majority of normal genomic region. The number of background noise reads generated in such region is multiplied by the amplification factor.

C). Synthetic dataset

A synthetic dataset is defined as a dataset of a pre-determined number of simulated peaks plus background noise reads; this dataset is a combination of the reads generated in both ***A)*** and ***B)*** and is used for our evaluation purposes.

D). Simulated dataset

A simulated dataset is defined as a dataset of a number of permutated reads within peaks in a corresponding ChIP data plus background noise reads. This dataset is used for the purpose of determining the FDR for each real data.

Percentile cutoff, p-value and FDR for BELT

The percentile cutoff is to distinguish signal enriched regions and back ground noise regions in the experimental data. The p-value is to determine if these signal enriched regions in the experiment data have significantly higher levels of enrichment compared to the counterpart regions in the control data. The peaks that have a p-value greater than 0.05 will be filtered out from the final result and output as rejected peaks. FDR is a measure of the overall confidence level of the identified peaks.

Implementation

Both PELT and BELT use Perl CGI scripts to receive user input and create temporary PHP pages on the server for user jobs. New jobs are polled on average every five minutes (by a Perl script with PELT and a bash shell script in BELT), and sent to processor scripts written in Perl. PELT is written in Perl, and BELT is written in C++. Both run on Linux/Unix platforms. PELT uses a server-side program written in Java to generate the graphics, and BELT uses a client-side Java applet to graphically display the results. GNU plotutils and the GNU scientific library are used for plot creation and spline interpolation. Gene annotation is performed with a Python script.

Gene annotation

Both PELT and BELT use the same annotation method. The midpoint of each peak is calculated, and compared to 5' and 3' of annotated RefSeq Genes. Distances from the nearest gene in both the 5' and 3' lists are calculated, and the gene with the annotation highest in priority is assigned to the peak. Annotations are listed below.

Annotation (in descending order of priority)	Relative distance to an annotated RefSeq gene
5_TSS	-1000 to 1000
3_Core	2000
5_Proximal	10000
3_Proximal	10000
Intragenic	within gene
5_Distal	100000
3_Distal	100000
Gene_Desert	>100000

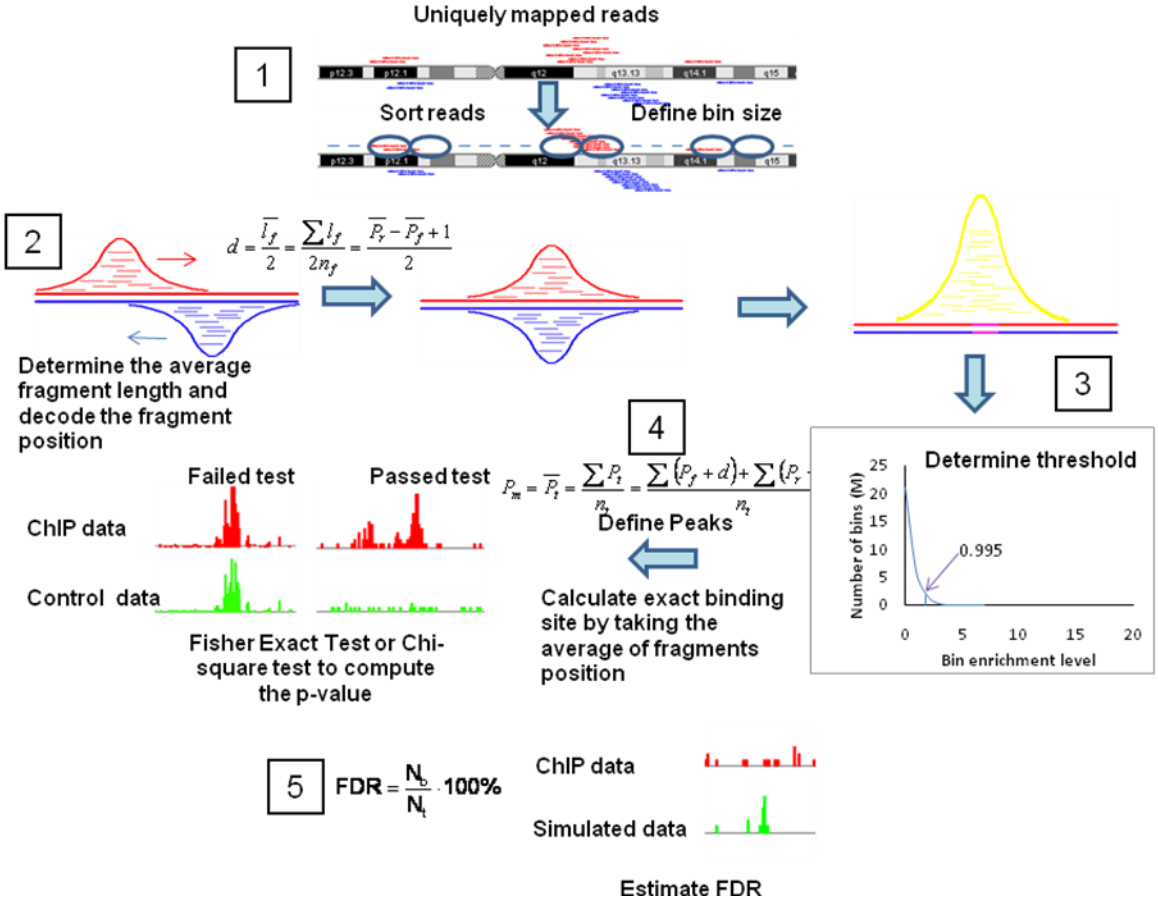

Figure S1. A summary of the BELT algorithm including five steps: 1) Defining a series of bins by evenly dividing the genome varying by default from 100 bp to 500 bp, and counting the density of reads for each bin; 2) Calculating an average fragment length for a ChIP sample by considering the direction of the reads, and decoding the fragment position by shifting reads; 3) Determining significant enrichment threshold levels by a percentile rank statistic method; 4) defining binding regions (peaks) and locating the binding motifs within identified peaks by taking the average of fragments position; 5) Utilizing Monte-Carlo simulation for modeling background based on signal-noise-ratio of ChIP-seq data to estimate false discovery rates. If a control dataset is available, such as IgG, input data, a Fisher exact test or Chi-square test is applied to compute the *p*-value for identified peaks.

A

TF Reads Number	CTCF (2,937,944)		FOXA1 (3,909805)		ER 8,307,676		NRSF 8,813,398	
	No. of Peaks	No. of Overlap	No. of Peaks	No. of Overlap	No. of Peaks	No. of Overlap	No. of Peaks	No. of Overlap
BELT	2,704	2179 (80.6%)	3,104	2,626 (84.6%)	6,687	5,367 (80.3%)	7,151	5,385 (75.3%)
MACS	2,753		3,155		6,619		7,093	
BELT	2,704	2704 (100%)	3,104	2,963 (95.5%)	6,687	6,626 (99.1%)	7,151	7,099 (99.3%)
PeakSeq	2,738		3,219		6,790		7,123	
BELT	2,704	N.A.	3,104	2,511 (80.1%)	6,687	5,875 (87.9%)	7,151	5,285 (74.0%)
QuEST	No control data		3,025		5,878		6,540	
BELT	2,704	2254 (83.4%)	3,104	2,426 (78.2%)	6,687	5,902 (88.3%)	7,151	6431 (89.9%)
SISSR	2,954		2,938		6,298		6,872	

B

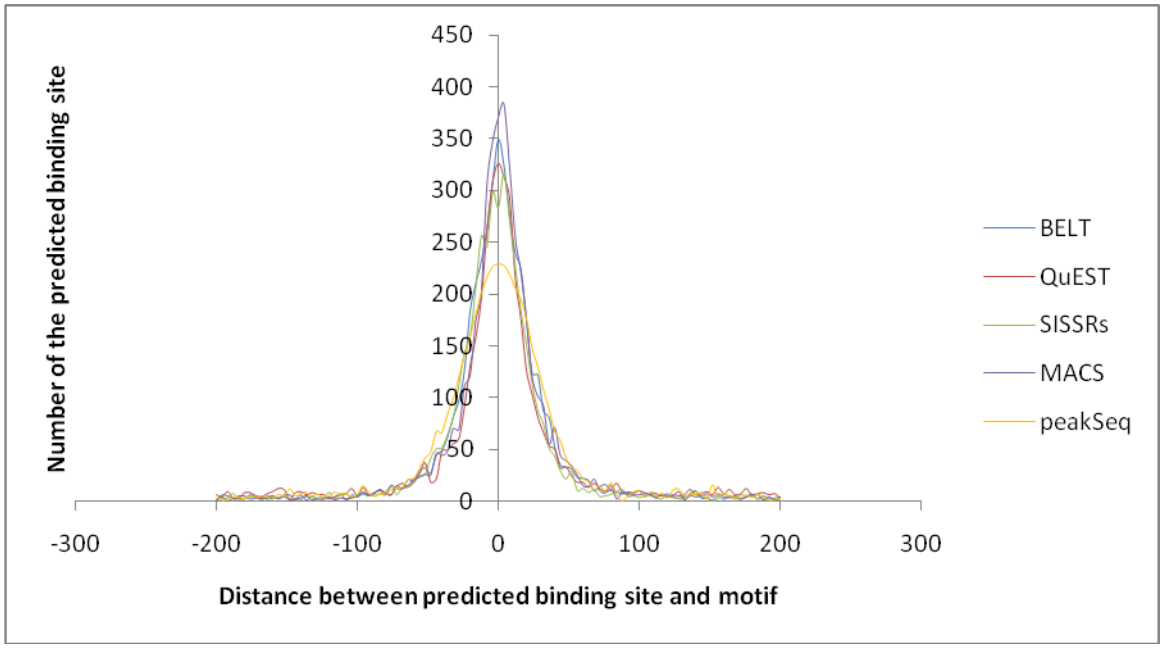

Figure S2. A. In a comparison of BELT and other programs (MACS [Ref 1], QuEST [Ref 2], PeakSeq [Ref 3] and SISSR [Ref 4]), all overlap rates were above 74%. The worst case overlap rate was 74% between BELT and QuEST on NRSF data, and the best overlap rate was 100% between BELT and PeakSeq on CTCF data. Overall, BELT had the highest number of overlapping peaks with PeakSeq (~95.5%-100%) for all four datasets. No comparison was

made between QuEST and BELT on CTCF due to the lack of a control dataset. **B.** Our program showed similar or higher accuracy in terms of motif localization than the other programs tested.

Table S1. Program parameters (default if not specified)

	CTCF	FOXA1	ER	NRSF
MACS	mfold 32 p value 1e-40	mfold 32 p value 1e-18	mfold 32 p value 1e-16	mfold 32 p value 2e-6
PeakSeq	FDR 0.001 p value 0.005	FDR 0.001 p value 0.005	FDR 0.001 p value 0.005	FDR 0.001 p value 0.005
QuEST	None	ChIP tags threshold 27 Region width 850 Other parameters default	ChIP tags threshold 18 Other parameters default	ChIP tags threshold 25 Region width 410 Other parameters default
SISSR	FDR 1e-12 window size 250	FDR 1e-3 e value 10 p value 0.00011	FDR 0.1 e value 10 p value 0.009	FDR 1e-3 window size 134
BELT	percentile 0.999 bin size 350	percentile 0.999 bin size 290	percentile 0.9975 bin size 400	percentile 0.995 bin size 200

Ref 1. Zhang, Y., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS), *Genome Biol*, **9**, R137.

Ref 2. Valouev, A., et al (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*, **5**, 829–834.

Ref 3. Rozowsky et al (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology, **27**, 66-75.

Ref 4. Jothi, R., et al (2008) Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Res*, **36**, 5221–5231.