White Paper: ChIP-Seq Peak Detection

Introduction

ChIP-Seq is a tool used to analyze protein interactions with DNA, such as the interaction of transcription factors with their DNA binding sites. The protein of interest is cross-linked with its binding site *in vivo*, and the protein-DNA complex is enriched with an antibody specific to the protein of interest. The complexes are filtered out, and the DNA is purified and sequenced. After the sequences have been mapped to the genome, the resulting peaks of sequence overlaps provide insight into the protein binding site locations. This white paper discusses the steps Partek takes to discover and report binding site peaks in ChIP-Seq experiments.

Fragment Length Analysis

The Fragment Length Analysis step calculates the average length of the sequenced DNA fragments in your ChIP samples. If the sample contains paired- end reads, this step will calculate the distribution of fragment lengths between the paired-ends of the fragments. If the data is single-end reads, Partek will use the phase shift between the reads on the forward strand and reads on the reverse strand to estimate the average fragment length; this method is from the Kharchenko, et al. paper (1). Note that the estimation of fragment length for single-end reads can only be done on IP samples and not on the control samples since the control samples will not contain a phase shift of forward and reverse reads. The peak of the distribution is used as the estimated average fragment length.

The Kharchenko method for single end reads is briefly described here. For more details see (1). For each strand *s* of each chromosome *c*, a count vector $n_c^s(x)$ of the number of reads whose 5' end maps to position *x* is formed. Strand cross-correlation for a strand shift of δ is calculated as

$$X(\delta) = \sum_{c \in C} \frac{N_c}{N} P[n_c^+(x), n_c^-(x-\delta)], \; [1]$$

where P[*a*,*b*] is the Pearson correlation coefficient between vectors *a* and *b*, *C* is the set of all chromosomes, N_c is the number of reads mapped to chromosome *c*, and *N* is the total number of reads. Padding of 0 is added to each vector n_c^+ and n_c^- where necessary.

Partek first removes duplicate alignments – alignments that map to the exact same position. As noted in the Kharchenko paper, a jump of cross-correlation is present in some datasets at a strand shift corresponding to the length of the reads. Removing duplicate alignments forces the correlation score at the read length to be 0, which is helpful for datasets that have reads stacked at a particular location, which can occur from mapping errors or PCR biases. Partek tries a range of values for δ from 0 to 500 and finds the δ that gives the maximum correlation coefficient. This distance can be used in the peak detection workflow step below.

Peak Detection Steps

1. Next-generation sequencing data can come in the form of either single end reads or paired end reads. For single end reads, Partek extends the reads in the 3' direction to the average fragment length, creating an *estimated fragment*. The average fragment length can be supplied by the user or estimated using the Fragment Length Analysis step. For paired end reads, both ends of the fragment are sequenced so the *estimated fragment* is the start of the 5' most read to the end of the 3' most read.
2. The midpoint of each *estimated fragment* is calculated, and its location on the genome is recorded.
3. The genome is divided into non-overlapping windows of a user-defined length (default is 100 bp). An aligned read (either single end or paired end) is considered to be within a window if the midpoint of its *estimated fragment* is within the window.
4. The number of midpoints in each window is counted and an empirical distribution of window counts is created.
5. A zero-truncated negative binomial model (ZTNB, explained below) is fit to the distribution, and a peak cut-off is determined based on the FDR calculated from the ZTNB. Windows above this threshold are considered enriched.
6. A window with length the same in step 3 is slid across the genome to find all enriched windows. Overlapping enriched windows are merged into regions and reported.
7. If the data contains control samples, the reported peaks can be filtered by removing peaks based on the statistics associated with them (see Filtering Peaks).

Zero Truncated Negative Binomial Model

A model is fit to the empirical distribution from step 4 of the *Peak Detection Steps*. If the midpoints of non-enriched reads were equally likely to fall anywhere on the genome, the distribution would fit a Poisson model. However, it has been shown in the literature that over-dispersion of read counts occurs in next generation data and a negative binomial model is more appropriate (2-3). Furthermore, mappability biases and insufficient sequencing depth can give rise to excess zeros in the observed counts (4). For these reasons, Partek excludes windows with zero counts and fits a zero-truncated negative binomial (ZTNB) model to the distribution. The ZTNB can be written as

$$f_{ZTNB}(k \mid r,p) = \frac{\Gamma(r+k)}{k!\Gamma(r)} \frac{p^r(1-p)^k}{1-p^r}, \qquad [2]$$

where k is the number of midpoints within a window and p and r are the parameters of the distribution to be estimated from the data. Partek determines the values p and r that will maximize the likelihood equation

$$L = \sum_{k=1}^{n} freq(k) * \log(f_{ZTNB}(k \mid r,p)), \quad [3]$$

where *freq(k)* is the number of windows in the observed data containing exactly k midpoints. The threshold K is chosen so that the ratio of windows with $k \geq K$ midpoints expected by the ZTNB model compared to what is observed from the data is less than or equal to the user-specified false discovery rate (FDR).

Filtering Peaks

Partek returns all peaks above the threshold determined by the ZTNB distribution and specified false discovery rate. However, false peaks can occur due to PCR biases, mapping ambiguities, and non-specific binding. Partek also produces intra- and inter- sample statistics for each peak that can be used to remove peaks in the data that are not real ChIP-Seq peaks. This section describes the p-values given from the Mann-Whitney U Test on a peak and the binomial test between an IP sample and a control.

P-value of Mann-Whitney U test:

This statistic indicates whether or not the forward and reverse reads are well separated. For each peak that is detected, the relative locations of the forward and reverse reads are used to compute a Mann-Whitney U Test. Since read sequences come from the ends of the fragment, there is an expected separation of the forward and reverse reads at enriched regions, which leads to a low Mann-Whitney p-value.

For each region, the forward (F) and reverse (R) alignments are ordered by their left-most position on the reference strand. An example of perfect separation would be FFFFRRRRRR. An example of a mixed ordering would be FFRFRRFFFRFR. The single ranked series is used as input to the Mann-Whitney test to calculate the U statistic and p-value. A low p-value means there is a clear phase shift between the forward and reverse reads, and a high p-value indicates a shuffled distribution of forward and reverse reads.

This measure is good for finding peaks that may arise due to PCR biases. PCR biases usually show forward and reverse reads stacked directly on top of each other, resulting in poor separation of forward and reverse reads and a high (close to 1) p-value.

Binomial P-value and Scaled Fold Change:

These statistics indicate if the number of reads in the ChIP peak is significantly higher than the corresponding region in the control sample. It can be used to filter peaks that occur in both samples, such as in the case of non-specific binding.

The binomial p-value compares two samples: an immuno-precipitated sample (*S*) and a control sample (*C*). First, an overall scaling factor α is calculated between *S* and *C* using non-overlapping windows across the entire genome such that the number of reads in *S* for a particular window is estimated as α times the number of reads in *C* for that window, similar to (5). This is done to normalize the control data set to the sample data set since the number of mapped reads in one set is generally not the same as the number of mapped reads in the other set.

For each detected peak, the number of reads in the region in *S* is compared to *C* using a binomial test. A one-tailed test is conducted on the cumulative distribution

$$F(k,n,P) = \sum_{k=K}^{n} \binom{n}{k} P^{k} (1-P)^{f}, \quad [4]$$

where $P = \frac{\alpha}{1+\alpha}$ is the probably of the read belonging to *S*, *k* is the number of "successes" (reads in window mapped to *S*), *f* is the number of failures (reads in window mapped to *C*), and $n = k + f$ is the total number of reads in the region. A low p-value indicates that the number of reads in the IP sample *S* is significantly higher than the number of reads in the corresponding region of the control sample *C*.

The scaled fold change is calculated as

$$scaled - fold - change = \frac{k + pseudo}{\alpha * f + pseudo}, \qquad [5]$$

where *pseudo* is a pseudo-count to avoid dividing by 0 (*pseudo* = 1). For background windows (those where $k \approx \alpha \times f$), the fold change will be close to 1.

References

(1) P.V. Kharchenko, M.Y. Tolstorukov, & P.J. Park. “Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* 2008: 26.

(2) Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. “An integrated software system for analyzing ChIP-chip and ChIP-seq data” *Nature Biotechnology* 2008, 26:1293-1300.

(3) Christiana Spyrou, Rory Stark, Andy G Lynch, and Simon Tavare. “BayesPeak: Bayesian analysis of ChIP-seq data” *BMC Bioinformatics* 2009, 10:299.

(4) Pei Fen Kuan, Guangjin Pan, James A. Thomson, Ron Stewart, and Sunduz Keles. "A Statistical Framework for the Analysis of ChIP-Seq Data" University of Wisconsin, Department of Statistics, Technical Report (2009). Available at: http://works.bepress.com/sunduz_keles/19

(5) Joel Rozowsky et. al. “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls” *Nature Biotechnology* 2009, 27:66-75.