BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 3 2011, pages 428–430
doi:10.1093/bioinformatics/btq669

Gene expression

Advance Access publication December 7, 2010

W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data

Xun Lan[1], Russell Bonneville[1], Jeff Apostolos[1], Wangcheng Wu[2] and Victor X Jin[1,*]

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43221 and [2]Department of Informatics, University of Washington, Seattle, WA 33333, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: ChIP-based technology is becoming the leading technology to globally profile thousands of transcription factors and elucidate the transcriptional regulation mechanisms in living cells. It has evolved rapidly in recent years, from hybridization with spotted or tiling microarray (ChIP-chip), to pair-end tag sequencing (ChIP-PET), to current massively parallel sequencing (ChIP-seq). Although there are many tools available for identifying binding sites (peaks) for ChIP-chip and ChIP-seq, few of them are available as easy-accessible online web tools for processing both ChIP-chip and ChIP-seq data for the ChIP-based user community. As such, we have developed a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Our web tool W-ChIPeaks employed a probe-based (or bin-based) enrichment threshold to define peaks and applied statistical methods to control false discovery rate for identified peaks. The web tool includes two different web interfaces: PELT for ChIP-chip, BELT for ChIP-seq, where both were tested on previously published experimental data. The novel features of our tool include a comprehensive output for identified peaks with GFF, BED, bedGraph and .wig formats, annotated genes to which these peaks are related, a graphical interpretation and visualization of the results via a user-friendly web interface.

Availability: http://motif.bmi.ohio-state.edu/W-ChIPeaks/.

Contact: victor.jin@osumc.edu

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

Received on August 23, 2010; revised on November 18, 2010; accepted on December 1, 2010

1 INTRODUCTION

ChIP-based technology is becoming the leading technology to globally profile thousands of transcription factors and elucidate the transcriptional regulation mechanisms in living cells (Farnham, 2009). It has evolved rapidly in recent years, from hybridization with spotted or tiling microarray (ChIP-chip) (Kim *et al.*, 2005), to pair-end tag sequencing (ChIP-PET) (Loh *et al.*, 2006), to current massively parallel sequencing (ChIP-seq) (Johnson *et al.,* 2007). Despite the fact that microarray-based chromatin immunepreciptation (ChIP) method, ChIP-chip, is gradually being replaced by the emerging sequencing-based method such as ChIP–seq, both methods are currently being used in many laboratories as a major tool to survey transcription factor binding patterns, study various histone modifications in an unbiased manner.

Currently available tools for the ChIP-chip data are exemplified and comprehensively compared in the Spike-In data (Johnson *et al.,* 2008). ChIP-seq technology and related computational tools are also reviewed in Park (2009). CisGenome provided an integrated analyzing software system for both technologies (Ji *et al.,* 2008). While we appreciate the accuracy and efficiency of these tools, few of them are available as easy-accessible online web tools for processing both ChIP-chip and ChIP-seq data for the ChIP-based user community. As such, we have developed a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Our web tool W-ChIPeaks employed a probe-based (or bin-based) enrichment threshold to define peaks and applied statistical methods to control false discovery rate for identified peaks. The web tool includes two different web interfaces: probe-based enrichment threshold level (PELT) for ChIP-chip and BELT (bin-based enrichment threshold level) for ChIP-seq, where both were tested on previously published experimental data.

2 METHODS

2.1 Overview

The utility and layout of the W-ChIPeaks is demonstrated in Figure 1. W-ChIPeaks provides a web-based interface with three main features: identification of peaks with GFF, BED, bedGraph and .wig formats, annotated genes to which these peaks are related, annotated genes to which these peaks are related, a graphical interpretation and visualization of the results via a user-friendly web interface. The link of results will be emailed to the address given in the contact information. For two or three ChIP-chip datasets, a plot of overlapping comparison between datasets at different threshold levels is also provided. Usage of W-ChIPeaks web service is simple and does not require any knowledge of the underlying software.

2.2 Input

For ChIP-chip, there are three required inputs from the user: GFF files from NimbleGen or Agilent Array (allow eight sets in maximum), the selection of array types and genomes, and e-mail contact information; For ChIP-seq, there are a few options and one required inputs from the user including Eland and extended Eland for Illumia GAII, bowtie alignment output, BED, GFF, SAM, or BAM format of aligned reads, and e-mail contact information.

2.3 Algorithms and statistical methods

2.3.1 PELT We employed a probe-based enrichment threshold to define peaks and a permutation-based statistical method to control false discovery rate for identified peaks. Suppose for a sample with **N** probes ($i=1,\ldots,N$)

*To whom correspondence should be addressed.

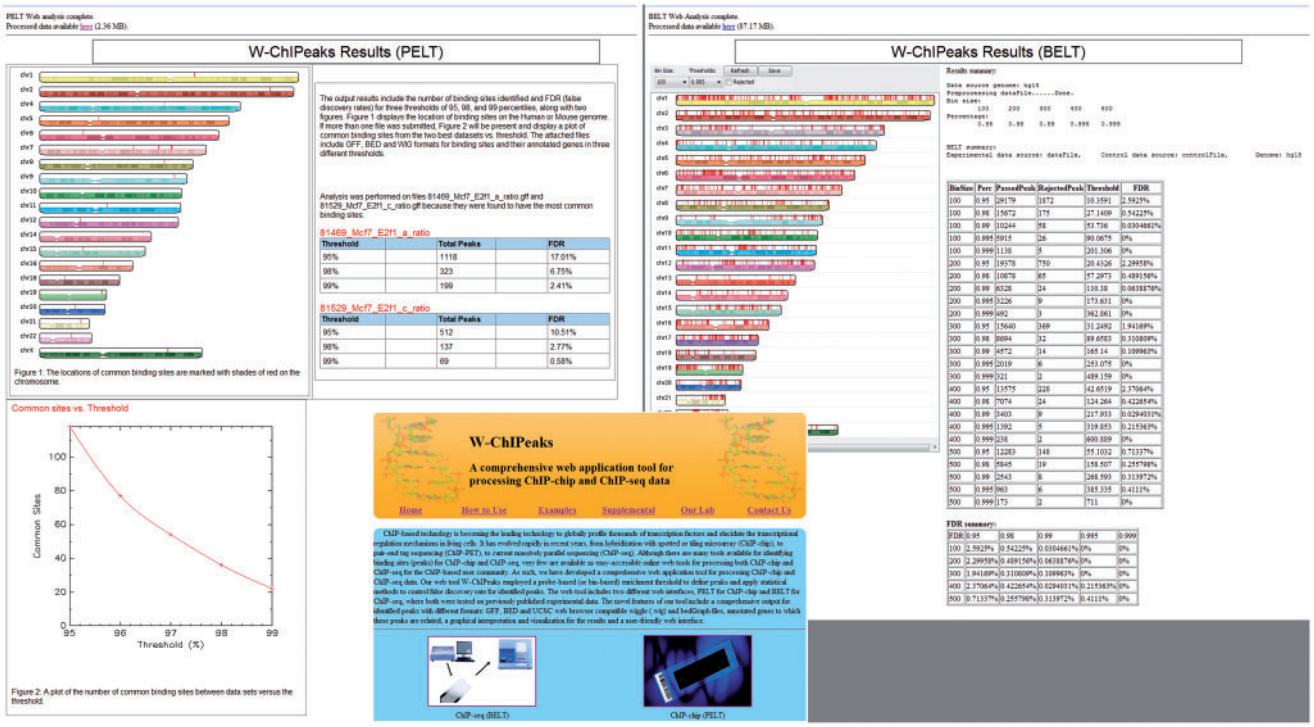

Fig. 1. The utility and layout of the W-ChIPeaks.

on each array, after normalization of each array, a probe i on the array j has an intensity: $\mathbf{I}_{ij}$. For any particular peak $\mathbf{P}_k$ among $\mathbf{B}$ peaks, it is first defined by a percentile level d based on a distribution of the probes, in which the mean value of the peak intensity consisting of at least three probes in a row has to be greater than the value of that percentile level d (for example, the top 1, 2, or 5% of all probes on the array). We applied the permutation-based approach to estimate the false discovery rate. We permutated each array, and found the number of peaks at a percentile level d, and then repeated the permutation process 1000 times, finally averaged the number of peaks from these 1000 permutations. The number of peaks without permutation at level d is considered as TP(d), and the average of the number of peaks after permutations at the same level d is considered as FP(d). The FDR(d) [FDR(d) = FP(d)/TP(d)] was then obtained at that level d.

2.3.2 BELT We employed a bin-based enrichment threshold to define peaks and a Monte-Carlo simulation statistical method to control false discovery rate for identified peaks. The BELT algorithm includes four steps: (i) define a series of bin size by evenly dividing the genome varying from 100 bp to 500 bp, and counting the density of reads for each bin; (ii) calculate an average length of ChIP fragments by considering the direction of the reads, decoding the binding site position by shifting the reads (Zhang *et al*., 2008); (iii) determine significant enrichment threshold levels by a percentile rank statistic method and (iv) Estimate false discovery rates by utilizing Monte Carlo simulation for modeling background based on signal-noise-ratio of ChIP-seq data. (Supplementary Methods and Supplementary Figure S1).

Scoring called peaks and estimation of FDR: A score for a called peak by BELT is empirically defined in Supplementary Methods formula (3) and is used to rank the peaks. A *FDR* is estimated using Supplementary Methods formulas (5) and (6).

Comparison with other ChIP-seq programs: The performance of BELT was compared to four publicly available ChIP-seq programs: MACS, QuEST, PeakSeq and SISSRs on four published datasets: CTCF, FOXA1, ER and NRSF. The results of the number of overlapping peaks between BELT and other programs showed that all of the overlap rates are over 74% (Supplementary Figure S2A). A plot of the relative distance from the predicted binding motif to the real motif showed our program has a similar or higher accuracy than the other programs (Supplementary Figure S2B, Supplementary Table S1).

2.4 Implementation

W-ChIPeaks was implemented with PHP, Perl, Java and C++.

2.5 Output

W-ChIPeaks has a comprehensive output for identified peaks with different formats: GFF, BED, bedGraph and .wig files, annotated genes to which these peaks are related, a graphical interpretation and visualization for the results. For two or three ChIP-chip datasets, a plot of overlapping comparison between datasets at different threshold levels is also provided.

2.6 Sample test

The W-ChIPeaks was tested with different published datasets from the ChIP-chip and ChIP-seq experiments. The array platform for ChIP-chip data is from NimbleGen or Agilent Array Platform. Some of such datasets include E2F1 (Jin *et al.*, 2006), N-MYC (Cotterman *et al.*, 2008), ZNF263 (Frietze *et al.*, 2010), PolII, H3K4me3 in K562 cell line (ENCODE consortium), H3K9me2, H3Ac (Bapat *et al*., 2010) and results are available online at: http://motif.bmi.ohio-state.edu/W-ChIPeaks/examples.shtml.

ACKNOWLEDGEMENTS

This study was partly supported by funds UL1RR025755 from the National Center for Research Resources, NIH, and from the Department of Biomedical Informatics, The Ohio State University.

Conflict of Interest: none declared.

REFERENCES

Bapat,S.A. *et al.* (2010) Multivalent epigenetic marks confer microenvironment-responsive epigenetic plasticity to ovarian cancer cells. *Epigenetics*, **5**, 716–729.

Cotterman,R. *et al.* (2008) N-Myc regulates a widespread euchromatic program in the human genome partially independent of its role as a classical transcription factor. *Cancer Res.*, **68**, 9654–9662.

Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nature Rev. Genet.*, **10**, 605–616.

Frietze,S. *et al.* (2010) Genomic targets of the KRAN and SCAN domain-containing zinc figer protein ZNF263. *J. Biol. Chem.*, **285**, 1393–1403.

Ji *et al*. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**,1293–1300.

Jin,V.X. *et al.* (2006) A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data–a case study using E2F1. *Genome Res.*, **16**, 1585–1595.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Johnson,D.S. *et al.* (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.*, **18**, 393–403.

Kim,T.H. *et al.* (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.

Loh,Y.-H. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, **38**, 431–440.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.,* **10**, 669–680.

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