BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 19 2009, pages 2605–2606
doi:10.1093/bioinformatics/btp479

Genome analysis

CEAS: *cis*-regulatory element annotation system

Hyunjin Shin[1], Tao Liu[1], Arjun K. Manrai[2] and X. Shirley Liu[1,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney St, Boston, MA 02115 and [2]Harvard-MIT Division of Health Sciences and Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA

Received on May 16, 2009; revised on July 8, 2009; accepted on August 3, 2009

Advance Access publication August 18, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We present a tool designed to characterize genome-wide protein–DNA interaction patterns from ChIP-chip and ChIP-Seq data. This stand-alone extension of our web application CEAS (*cis*-regulatory element annotation system) provides summary statistics on ChIP enrichment in important genomic regions such as individual chromosomes, promoters, gene bodies or exons, and infers the genes most likely to be regulated by the binding factor under study. CEAS also enables biologists to visualize the average ChIP enrichment signals over specific genomic regions, particularly allowing observation of continuous and broad ChIP enrichment that might be too subtle to detect from ChIP peaks alone.

Availability: The CEAS Python package is publicly available at http://liulab.dfci.harvard.edu/CEAS.

Contact: shin@jimmy.harvard.edu; xsliu@jimmy.harvard.edu

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

1 INTRODUCTION

In analysis of *cis*-regulatory elements using genome-wide ChIP-chip or ChIP-Seq, it is essential to characterize the ChIP signals and identify potential association of ChIP regions with functionally important genomic regions such as gene promoters or exons. Previously, we developed a web server that analyzes ChIP regions by evaluating GC content and evolutionary conservation, conducting sequence motif search and mapping the regions to their nearest genes (Ji *et al*., 2006). However, additional analysis functions are needed to provide biologists with a more complete perspective. For example, in addition to analyzing the identified ChIP regions of a factor, displaying the average ChIP enrichment signal within/near genes helps biologists better visualize the functional loci of factors, especially for broad histone modifications. However, such analysis functions often require the ability to manipulate large continuous ChIP enrichment signal files (e.g. WIG files of hundreds of mega bytes in size), which are inconvenient to upload to a web server. Therefore, in order to extend our current successful web-based *cis*-regulatory element annotation system (CEAS) (over 35K analysis queries processed in 2008), we present a stand-alone CEAS extension package with more analysis functions of drawing average ChIP signal profiles at genes or user-specified loci from a WIG file and providing summary statistics about the distribution of ChIP regions over important genomic features such as promoters, as well as a report on the association of individual genes with their proximal ChIP regions.

*To whom correspondence should be addressed.

2 DESIGN AND IMPLEMENTATION

CEAS consists of three modules: (i) ChIP region annotation; (ii) gene-centered annotation; and (iii) average signal profiling within and near important genomic features. CEAS requires three inputs: (i) a gene annotation table file such as the UCSC RefSeq file; (ii) a BED file with the ChIP peak calls; and (iii) a WIG file with the continuous ChIP enrichment signal. The package comes with pre-compiled RefSeq tables in sqlite3 files for several genomes (ce4 and ce6 for worm; dm2 and dm3 for fly; mm8 and mm9 for mouse; and hg18 and hg19 for human), and other gene annotation tables can be downloaded from the UCSC genome browser and compiled using our script. The BED file must contain three columns (the chromosome, start and end of every ChIP call). CEAS accepts the fixedStep and variableStep WIG formats when performing average signal profiling. As output, CEAS produces an R file containing the script to generate the graphical results of ChIP region annotation and average signal profiles (or a PDF file with the graphical results if R can be called directly in the same environment as Python), and an XLS file with gene-centered annotation results.

2.1 ChIP region annotation

CEAS estimates the relative enrichment level of ChIP regions in each genomic feature with respect to the whole genome. To do this, it first calculates the percentages of the ChIP regions that reside in the following four categories: (i) promoters; (ii) bidirectional promoters; (iii) downstream regions of genes; and (iv) gene bodies (3′UTRs, 5′UTRs, coding exons and introns). In addition to these categories, the user can add another user-specified extra category (e.g. non-coding regions) as an optional input BED file.

'Promoters' correspond to the upstream regions of the transcription start sites (TSSs) of genes. The user specifies three promoter sizes (1 kb, 2 kb and 3 kb by default) to be used. For instance, if the user sets the promoter sizes to be 1 kb, 3 kb and 10 kb upstream of the TSS, CEAS computes the cumulative percentages of ChIP regions that fall in $\leq$1 kb, $\leq$3 kb and $\leq$10 kb upstream of the TSSs of genes. 'Bidirectional promoters' are promoter regions between divergently transcribed genes whose TSSs are closer in

proximity than user-defined distances (two options, 2.5 kb and 5 kb by default). 'Downstreams' refer to the immediately downstream regions of the transcription termination sites (TTSs) of genes. 'Gene bodies' are divided into 3′ and 5′UTRs, coding exons, and introns. After the percentages of ChIP regions residing within the above categories are obtained, they are compared with the genome background percentages for the same categories and *P*-values are calculated using one-sided binomial test. While algorithms such as CisGenome shows the percentages of ChIP regions in important genic regions such as exons and introns (Ji *et al*., 2008), CEAS provides additional *P*-value of the relative enrichment over genomic background.

In order to summarize the ChIP region annotation, CEAS draws a pie chart displaying the distribution of ChIP regions across the genomic features. If ChIP regions do not fall into any of the categories, they are considered to be 'distal intergenic'.

2.2 Gene-centered annotation

Identifying genes associated with ChIP regions by proximity is important to infer the direct regulatory gene targets of the binding factor under study. CEAS provides the distances to the centers of the nearest ChIP regions upstream and downstream of every gene's TSS. For factors with sharp binding patterns, this might be sufficient for biologists to determine the potential target genes of the binding factor under study, and similar functions are available in other algorithms such as GPAT (Krebs *et al*., 2008). However, in case a broad ChIP peak covers all or part of a gene body, it is useful to know how much of the gene, including its promoter or downstream region, is occupied by the ChIP region. To this end, CEAS divides every gene into three equal fractions as well as extracts exons, and calculates the percentages of the areas covered by ChIP regions. The results are saved as a tab-delimited text file with XLS extension.

2.3 Average signal profiling within/near important genomic features

Since ChIP region and gene-centered annotation operate on discrete ChIP regions identified by a peak-calling algorithm, some subtle binding patterns may fail to be captured, depending on the cut-off used in peak calling. Therefore, CEAS displays the continuous ChIP enrichment signal within and near important genomic features for biologists to visualize the average binding patterns in these regions. CEAS draws the average signals around TSSs and TTSs in a user-defined range (±3 kb from the TSS and TTS by default). In addition, CEAS computes average signals on 'meta-gene', 'meta-concatenated-exon', 'meta-concatenated-intron', 'meta-exons' and 'meta-introns', where the prefix 'meta' indicates that every element is normalized to have the same length. The difference between meta-concatenated-exon and meta-exons is that the first concatenates all exons of a gene (like a meta-cDNA) before calculating the average gene profile, whereas the latter calculates the average exon profile of all exons. CEAS provides an additional function that draws the average ChIP signals of multiple user-specified sub-groups of genes, allowing the user to compare the signals between the gene groups. In addition, we provide a separate script, named 'sitepro', in our CEAS package, which draws the average signal (from a WIG) in a user-provided list of sites (specified in a BED) to visualize the average signal in any arbitrary regions (e.g. transcription factor binding sites).

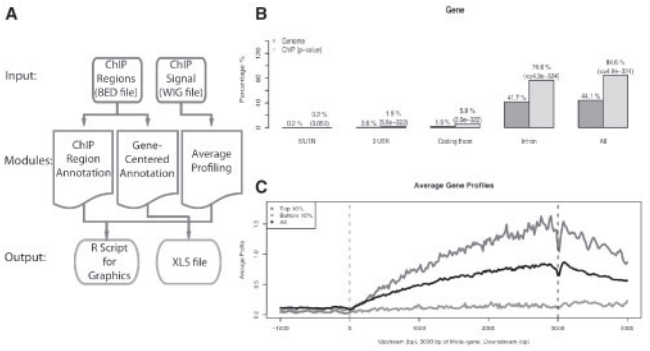

Fig. 1. (**A**) A flow chart explaining the input, modules and output of CEAS. (**B**) An example of ChIP region annotation (gene body) for human CD4T+ H3K36me3 ChIP-Seq. (**C**) An example of average signal profiles over the 3 kb meta-gene of the same data. The top and bottom lines correspond to genes with top 10%, and bottom 10% of expression indexes, respectively. The middle line represents all RefSeq genes.

3 EXAMPLE USAGE

CEAS automatically detects the available input files and runs the corresponding modules (Fig. 1A). Fig. 1B and (C) are the ChIP region annotation on the gene body and the average ChIP enrichment on the meta-gene for human CD4T+ cell H3K36me3 ChIP-Seq (Barski *et al*., 2007), respectively. ChIP regions were called using MACS (Zhang *et al*., 2008) at a *P*-value cut-off of 10^{-5}. H3K36me3 is believed to be a transcription elongation mark, which shows relatively high enrichment (right bars in (B)) with respect to the background (left bars) in gene bodies, particularly in coding exons and introns. This observation is consistent with (C), in which we can see that the ChIP enrichment is low in promoters, but monotonically increases as we move towards the 3' end of the genes (the black line in (C)). In (C), the average ChIP signals of groups of genes with the top 10 % (top line) and bottom 10% (bottom line) of expression indexes are compared with that of all the human RefSeq genes (middle line).

ACKNOWLEDGEMENTS

We thank Josiah Altschuler for his help with building the CEAS web site. We thank Cliff Meyer, Yong Zhang, Zhenhua Wu, Xiangfeng Wang and Housheng He for their helpful feedback and discussion. We are also grateful to Holly Bartel for proofreading this article.

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al*. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Ji,X. *et al*. (2006) CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.*, **34**, W551–W554.

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

Krebs,A. *et al*. (2008) GPAT: retrieval of genomic annotation from large genomic position datasets. *BMC Bioinformatics*, **9**, 533.

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