Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W551–W554***
doi:10.1093/nar/gkl322

CEAS: *cis*-regulatory element annotation system

Xuwo Ji, Wei Li[1], Jun Song[1], Liping Wei* and X. Shirley Liu[1,]*

Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing, People's Republic China 100871 and [1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA

Received February 14, 2006; Revised March 11, 2006; Accepted April 13, 2006

ABSTRACT

The recent availability of high-density human genome tiling arrays enables biologists to conduct ChIP–chip experiments to locate the *in vivo*-binding sites of transcription factors in the human genome and explore the regulatory mechanisms. Once genomic regions enriched by transcription factor ChIP–chip are located, genome-scale downstream analyses are crucial but difficult for biologists without strong bioinformatics support. We designed and implemented the first web server to streamline the ChIP–chip downstream analyses. Given genome-scale ChIP regions, the *cis*-regulatory element annotation system (CEAS) retrieves repeat-masked genomic sequences, calculates GC content, plots evolutionary conservation, maps nearby genes and identifies enriched transcription factor-binding motifs. Biologists can utilize CEAS to retrieve useful information for ChIP–chip validation, assemble important knowledge to include in their publication and generate novel hypotheses (e.g. transcription factor cooperative partner) for further study. CEAS helps the adoption of ChIP–chip in mammalian systems and provides insights towards a more comprehensive understanding of transcriptional regulatory mechanisms. The URL of the server is http://ceas.cbi.pku.edu.cn.

INTRODUCTION

Chromatin immunoprecipitation coupled with DNA microarrays (ChIP–chip) has become a popular technique to identify genome-wide *in vivo* protein–DNA interactions. With the recent availability of commercial human genome tiling microarrays, many laboratories are starting to combine these two technologies to detect *cis*-regulatory elements in the human genome.

Despite the importance of ChIP–chip, there is still a shortage of convenient tools developed to streamline the downstream analyses with the capability of processing genome-scale ChIP regions. So far all the ChIP–chip papers in mammalian systems are published as a direct result of powerful bioinformatics support (1–6), which may not be available for smaller labs. Therefore, web servers that can perform comprehensive analyses of hundreds or thousands of ChIP regions are not only valuable to biologists, but also useful for promoting the adoption of this powerful technology.

We present a comprehensive *cis*-regulatory element annotation system (CEAS) web server that integrates useful tools for sequence analysis and annotation of ChIP regions in the human genome. CEAS results not only help biologists analyze and validate their ChIP regions, but also can be directly included in their manuscript or Supplementary Data.

WEB APPLICATION

CEAS is composed of three parts: (i) a front-end web-based user interface for input data submission, input data validation and job scheduling; (ii) an annotation engine for sequence analysis and annotations; and (iii) a reporting system for output generation and Email notification to the user.

User input

CEAS accepts an input file with ChIP regions in either UCSC BEDformat (http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED) or Sanger GFF format (http://www.sanger.ac.uk/Software/formats/GFF/). The standard BED files have three required fields: *chrom* for the chromosome name, *chromStart* for the starting position of a ChIP region on the chromosome and *chromEnd* for the ending position of the ChIP region on the chromosome. The chromosome coordinates of the ChIP regions should follow the human genome assembly version Build 35 (Hg17). Coordinates

*To whom correspondence should be addressed. Tel +86 10 6276 4970; Fax +86 10 6275 2438; Email: weilp@mail.cbi.pku.edu.cn
*Correspondence may also be addressed to X. Shirley Liu. Tel +1 617 632-2472; Fax +1 617 632-2444; Email: xsliu@jimmy.harvard.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors based on earlier genome assembly can be converted to Hg17 using the Batch Coordinate Conversion at UCSC genome browser (7). A unique identifier for every ChIP region, ordinarily an optional fouth field in BED files, is also required by CEAS.

Because sequence analysis and annotation for genome-scale ChIP regions are time consuming, CEAS requires the user to supply an Email address. After submission, the server will put each job submission on queue and Email the user once the computation is finished. Alternatively, if the user inputs ‘guest’ instead of an Email address, the server will return a confirmation page which will be redirected to a result page when the annotation is finished. The output files will be stored on the server for 3 days to ensure that the user has enough time to browse and download the results.

Sequence retrieval

Although several websites can retrieve repeat-masked sequence for a particular genomic region, none can handle hundreds to thousands of ChIP regions simultaneously. Furthermore, current retrieval websites mask only RepeatMasker repeats (http://repeatmasker.org) and tandem repeats with period of 12 or less (8). Tandem repeats with period of >12 could greatly affect the qPCR primer design for ChIP region validation and sequence motif finding within the ChIP regions. CEAS automatically retrieves the genomic sequences of all the ChIP regions with all RepeatMasker repeats and all tandem repeats masked, and presents them in FASTA format for user download.

Conservation plot

Comparative genomics has been widely used to identify *cis*-regulatory elements in higher eukaryotes (9), and thus biologists are often interested in knowing the level of conservation of the ChIP regions. CEAS uses the high-quality phastCons (10) information from the UCSC GoldenPath genome resource, which assigns a conservation score based on a phylogenetic hidden Markov model to virtually every nucleotide in the human genome. CEAS generates a thumbnail phastCons conservation plot for each ChIP region, allowing biologists to skim through hundreds of ChIP regions in a single pdf file. In addition, the server extends both ends of each ChIP region to 3 kb, calculates an average phastCons score for each position and generates an average conservation plot. This final conservation plot can give biologists an idea of how conserved their ChIP regions are (in the middle of the plot) compared to the genomic background (at both ends of the plot).

Nearby gene mapping

For each ChIP region, CEAS reports the nearest RefSeq genes in both upstream and downstream directions on both strands unless no gene is found within 300 kb. When a ChIP region lies within a gene, CEAS reports whether it is in the 5′-untranslated region (5′-UTR), 3′-UTR, a coding exon or an intron. For each ChIP region, CEAS provides its length, GC content and a link to UCSC genome browser. The server also gives a summary statistic for GC content and gene mapping of all the ChIP regions, including the percentages of ChIP regions that reside in proximal promoters (1 kb upstream from RefSeq 5′ start), immediate downstream (1 kb from RefSeq 3′ end), 5′-UTRs, 3′-UTRs, coding exons, introns and enhancers (>1 kb from RefSeq). This rough estimate of the ChIP region distribution helps biologists understand the specific binding behavior of their transcription factor.

Motif finding and enrichment analysis

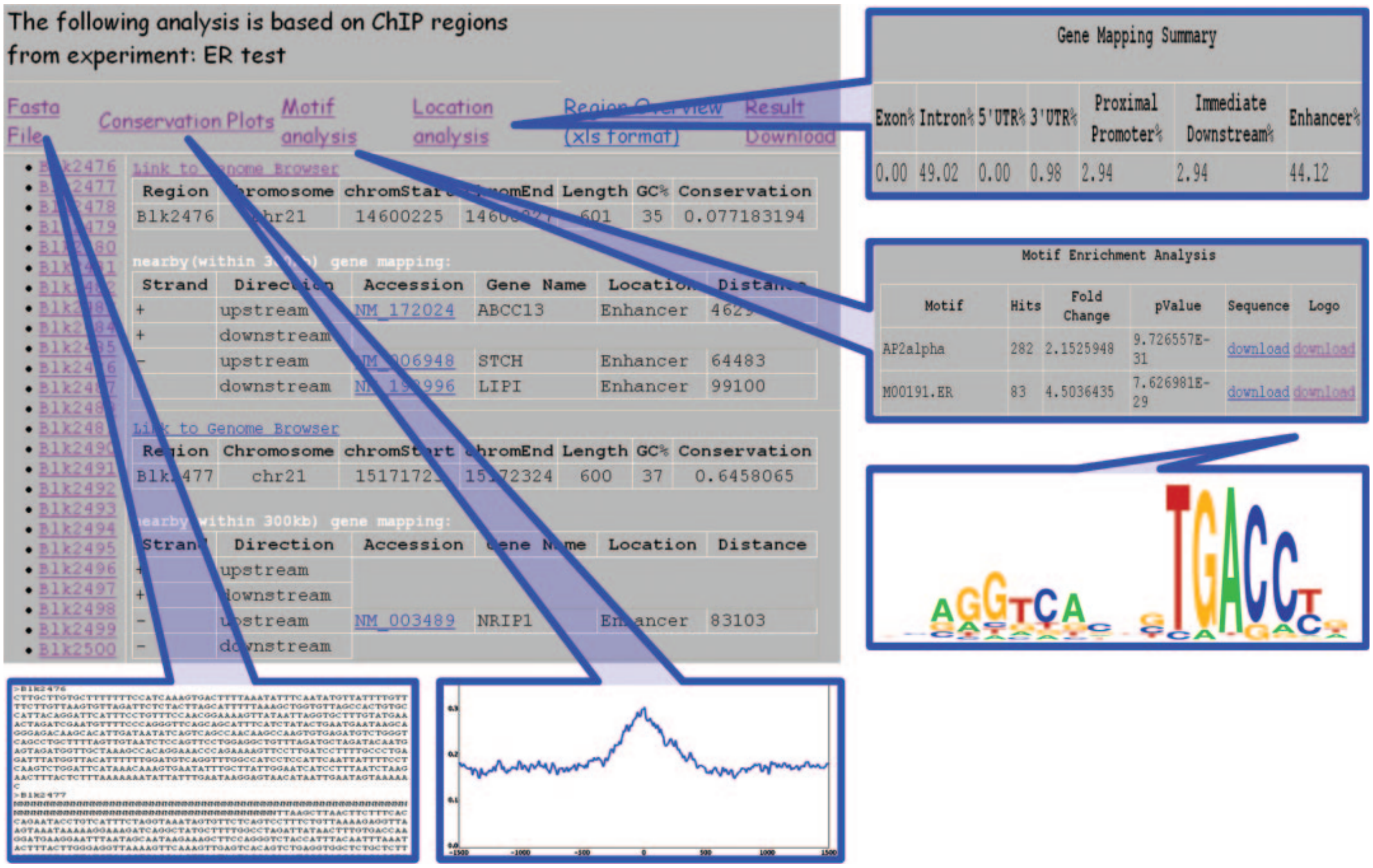
CEAS finds enriched sequence motifs in the ChIP regions that are putatively bound by the ChIP–chip transcription factor and its cooperative-binding partners. The current best *de novo* motif finding methods for ChIP–chip includes MEME (11), AlignACE (12), Mascan (13) and their combinations (14). For known motif-scanning methods, the best is TRANSFAC (15) or JASPAR (16) motif scan. Since the latter is less time consuming and can be pre-computed, we decided to use it. CEAS pre-collected all the motif matrices in the TRANSFAC (15) and JASPAR (16) databases, and filtered out motifs from microbial genomes or constructed with <10 sites to get ~800 well-characterized eukaryotic motifs. For each motif, CEAS pre-computed and stored all its hits (with information on chromosome, position, strand and score) in the fully repeat-masked human genomic sequence. The score of a particular *w*-mer hit to a motif of width *w* is calculated as follows:

$$score(w - mer) = \log\left[\frac{pb(w - mer\ from\ motif)}{pb(w - mer\ from\ Markov\ background)}\right],$$

where the background is the 9th order nucleotide Markov dependency estimated from the human genomic sequence. A score cutoff of *Max* (5,0.9 × *Motif Relative Entropy*) is used to call a motif a hit. The relative entropy of a motif of width w is calculated as $\sum_{i=A,C,G,T} \sum_{j=1}^{w} m_{ij} \log(m_{ij}/p_i)$, where m_{ij} is the probability of seeing nucleotide i at position j, and p_i is the probability of i in the human genome. Given user’s ChIP regions, CEAS counts the number of hits for every motif both within the ChIP region and in the whole genome. To be comprehensive, CEAS chooses a relative less stringent criteria of >1.5-fold change and binomial test *P*-value <1E−5 to report motifs enriched in the ChIP regions. Reported motifs are ranked by their *P*-values so biologists could refine the motif list with a more stringent cutoff. With each reported motif, CEAS provides its fold change, *P*-value, hit sequence in the ChIP regions and sequence logo (17).

Example output

Without other jobs pending on the queue, it takes CEAS ~20 min to process an input with 1000 ChIP regions each of length ~600 bp. Once the computation is finished, CEAS notifies the user by Email with a link to the result page. The result Html page reports each of the CEAS analysis results in different sections for user to view and download (Figure 1).



The following analysis is based on ChIP regions
from experiment: ER test

- Blk2476
- Blk2477
- Blk2478
- Blk2479
- Blk2480
- Blk2481
- Blk2482
- Blk2483
- Blk2484
- Blk2485
- Blk2486
- Blk2487
- Blk2488
- Blk2489
- Blk2490
- Blk2491
- Blk2492
- Blk2493
- Blk2494
- Blk2495
- Blk2496
- Blk2497
- Blk2498
- Blk2499
- Blk2500

Link to Genome Browser

Region	Chromosome	chromStart	chromEnd	Length	GC%	Conservation
Blk2476	chr21	14600225	1460[illegible]	601	35	0.077183194

nearby(within 300kb) gene mapping:

Strand	Direction	Accession	Gene Name	Location	Distance
+	upstream	NM_172024	ABCC13	Enhancer	462[illegible]
+	downstream				
-	upstream	NM_006948	STCH	Enhancer	64483
	downstream	NM_198996	LIPI	Enhancer	99100

Link to Genome Browser

Region	Chromosome	chromStart	chromEnd	Length	GC%	Conservation
Blk2477	chr21	1517172[illegible]	15172324	600	37	0.6458065

nearby(within 300kb) gene mapping:

Strand	Direction	Accession	Gene Name	Location	Distance
+	upstream				
+	downstream				
-	upstream	NM_003489	NRIP1	Enhancer	83103
-	downstream				

Gene Mapping Summary

Exon%	Intron%	5'UTR%	3'UTR%	Proximal Promoter%	Immediate Downstream%	Enhancer%
0.00	49.02	0.00	0.98	2.94	2.94	44.12

Motif Enrichment Analysis

Motif	Hits	Fold Change	pValue	Sequence	Logo
AP2alpha	282	2.1525948	9.726557E-31	download	download
M00191.ER	83	4.5036435	7.626981E-29	download	download

```
>Blk2476
CTTGCTTGTGCTTTTTTTTCCATCAAAGTGACTTTTAAATATTTCAATATGTTATTTTGTT
TTCTTGTTAAGTGTTAGATTCTCTACTTAGCATTTTTAAAGCTGGTGTTAGCCACTGTGC
CATTACAGGATTCATTTCCTGTTTCCAACGGAAAAGTTATAATTAGGTGCTTTGTATGAA
ACTAGATCGAATGTTTTCCCAGGGTTCAGCAGCATTTCATCTATACTGAATGAATAAGCA
GGGAGACAAGCACATTGATAATATCAGTCAGCCAACAAGCCAAGTGTGAGATGTCTGGGT
CAGCCTGCTTTTAGTTGTAATCTCCAGTTCCTGGAGGCTGTTTAGATGCTAGATACAATG
AGTAGATGGTTGCTAAAGCCACAGGAAACCCAGAAAAGTTCCTTGATCCTTTTGCCCTGA
GATTTATGGTTACATTTTTTGGATGTCAGGTTTGGCCATCCTCCATTCAATTATTTTCCT
CAAGTCTGGATTCATAAACAAAGTGAATATTTGCTTATTGGAATCATCCTTTAATCTAAG
AACTTTACTCTTTAAAAAAAATATTATTTGAATAAGGAGTAACATAATTGAATAGTAAAAA
C
>Blk2477
[illegible]
[illegible]TTAAGCTTAACTTCTTTCAC
CAGAATACCTGTCATTTCTAGGTAAATAGTGTTCTCAGTCCTTTCTGTTAAAAGAGGTTA
AGTAAATAAAAAGGAAAGATCAGGCTATGCTTTTGGCCTAGATTATAACTTTGTGACCAA
GGATGAAGGAATTTAATAGCAATAAGAAAGCTTCCAGGGTCTACCATTTACAATTTAAAT
ACTTTACTTGGGAGGTTAAAAGTTCAAAGTTGAGTCACAGTCTGAGGTGGCTCTGCTCTT
```

Figure 1. CEAS sample output. The top window contains links to each of the analysis results. Excerpts from the result sections are shown in the blue callouts in counter-clockwise order as genomic sequence of the ChIP regions in FASTA format, average conservation plot of the ChIP regions, sequence logo of an enriched motif, motif site list with fold change and *P*-values and summary of nearby gene mapping of all the ChIP regions.

DISCUSSION

CEAS is the first web server that allows high-throughput and comprehensive downstream analyses of human ChIP–chip data. The sequence retrieval function helps biologists design qPCR primers for validation and perform motif finding. The conservation plot function explores the functional conservation of the ChIP–chip transcription factor which could potentially be used to refine motif search. The nearby gene mapping function predicts the genes regulated by the transcript factor-bound regions. The motif finding function predicts the putative binding motif of the ChIP–chip transcription factor, which further validates the ChIP regions. It also predicts the cooperative-binding partners of the transcription factor. Many of the CEAS results can be directly incorporated in the user's ChIP–chip manuscript or Supplementary Data.

ChIP–chip on genome tiling array is still in its infancy. We are very lucky to work with the pioneers in this field, and foresee the necessary analysis tools that other ChIP–chip laboratories would need. As tiling arrays of other eukaryotic genomes become available and more biologists adopt the ChIP–chip technology, we envision CEAS to include more organisms with more and friendlier functionalities such as qPCR primer design for each ChIP region, motif scan for user provided motifs or *de novo* motif discovery.

ACKNOWLEDGEMENTS

The authors would like to thank Cliff A. Meyer for the helpful suggestions, and the members of the laboratories of Myles Brown, Kevin Struhl and Pamela Silver for their insights. The project was supported by grants from the Claudia Adams Barr Program in Innovative Basic Cancer Research (X.S.L), China Ministry of Education 'Grant for Talent of the New Century', China Ministry of Science and Technology 863 Program, China National Science Foundation (30550004), and China Program of Introducing Talents of Discipline to Universities (B06001). Funding to pay the Open Access publication charges for this article was provided by China Program of Introducing Talents of Discipline to Universities (B06001).

Conflict of interest statement. None declared.

REFERENCES

1. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al*. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
2. Carroll,J.S., Liu,X.S., Brodsky,A.S., Li,W., Meyer,C.A., Szary,A.J., Eeckhoute,J., Shao,W., Hestermann,E.V., Geistlinger,T.R. *et al*. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

3. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
4. Euskirchen,G., Royce,T.E., Bertone,P., Martone,R., Rinn,J.L., Nelson,F.K., Sayward,F., Luscombe,N.M., Miller,P., Gerstein,M. *et al.* (2004) CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.*, **24**, 3804–3814.
5. Kim,T.H., Barrera,L.O., Zheng,M., Qu,C., Singer,M.A., Richmond,T.A., Wu,Y., Green,R.D. and Ren,B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.
6. Bernstein,B.E., Kamal,M., Lindblad-Toh,K., Bekiranov,S., Bailey,D.K., Huebert,D.J., McMahon,S., Karlsson,E.K., Kulbokas,E.J.,III, Gingeras,T.R. *et al.* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169–181.
7. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
8. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
9. Liu,Y., Wei,L., Batzoglou,S., Brutlag,D.L., Liu,J.S. and Liu,X.S. (2004) A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res.*, **32**, W204–W207.
10. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
11. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
12. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
13. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
14. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
15. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, **31**, 374–378.
16. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, **32**, D91–D94.
17. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188–1190.