BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 10 2011, pages 1447–1448
doi:10.1093/bioinformatics/btr156

Databases and ontologies

Advance Access publication March 30, 2011

hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data

Li Chen, George Wu and Hongkai Ji*

Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: hmChIP is a database of genome-wide chromatin immunoprecipitation (ChIP) data in human and mouse. Currently, the database contains 2016 samples from 492 ChIP-seq and ChIP-chip experiments, representing a total of 170 proteins and 11 069 914 protein–DNA interactions. A web server provides interface for database query. Protein–DNA binding intensities can be retrieved from individual samples for user-provided genomic regions. The retrieved intensities can be used to cluster samples and genomic regions to facilitate exploration of combinatorial patterns, cell-type dependencies, and cross-sample variability of protein–DNA interactions.

Availability: http://jilab.biostat.jhsph.edu/database/cgi-bin/hmChIP.pl

Contact: hji@jhsph.edu

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

Received on October 27, 2010; revised on March 3, 2011; accepted on March 16, 2011

1 INTRODUCTION

Chromatin immunoprecipitation (ChIP) followed by genome tiling array hybridization (ChIP-chip) (Ren *et al*., 2000) and ChIP coupled with massively parallel sequencing (ChIP-seq) (Johnson *et al*., 2007) are powerful technologies to study genome-wide protein–DNA interactions. Large amounts of ChIP-chip and ChIP-seq data have been made publicly available in the past few years. These data contain rich information which can be synthesized to make new discoveries or used to boost analysis of new datasets. For example, in order to study the role of a transcription factor (TF) Sox17 in mouse embryonic stem cell (mESC) differentiation, Niakan *et al*. (2010) determined Sox17 binding sites in blastocyst-derived extraembryonic stem cells (XEN) using ChIP-chip. A question of interest is what other TFs can bind to the same *cis*-regulatory elements and potentially interact with Sox17. While it is difficult for a single lab to experimentally test binding of hundreds of mouse TFs to Sox17 binding sites, one can potentially answer the question by analyzing ChIP data in public domains representing diverse proteins and cell types. As another example, c-Myc has been extensively studied by many labs. By analyzing c-Myc binding data from different cellular contexts collected by different labs, one might be able to identify different classes of c-Myc binding sites based on their cell-type dependencies.

A prerequisite for utilizing the public ChIP data is the ability to freely query, retrieve, normalize and compare binding intensities from arbitrary samples and genomic regions. Currently, this is a daunting task for most researchers working on human and mouse. Existing raw data repositories such as The NCBI Gene Expression Omnibus (GEO) (Barrett *et al*., 2007) and Sequence Read Archive (SRA) (Wheeler *et al*., 2008) do not provide tools for interactively exploring the ChIP data. The UCSC genome browser (Kent *et al*., 2002) provides functionalities for visualizing the data, but its ChIP data collection is limited. Although the browser is good at exploring one genomic region at a time, it is incapable of conveniently retrieving, normalizing and comparing data from many genomic regions. The recently developed ChIP-X database (Lachmann *et al*., 2010) has collected TF target gene lists from published ChIP studies, however it does not provide tools for retrieving and comparing binding intensities across samples. hmChIP is developed in this context to meet the pressing need for exploring protein–DNA binding intensities in publicly available ChIP data.

2 DATA COLLECTION, QUERY AND RETRIEVAL

At the time of writing, hmChIP contains 2016 ChIP-chip and ChIP-seq samples collected from GEO, SRA, and the ENCODE at UCSC (Rosenbloom *et al*., 2010). The samples are grouped into 492 experiments which were analyzed using TileProbe (Judy and Ji, 2009) and CisGenome (Ji *et al*., 2008) to generate one peak list per experiment. Each peak list contains DNA-binding locations of one protein in one specific cellular context. In total, the database contains 11 069 914 protein–DNA interactions for 170 proteins in a variety of cell types (Supplementary Tables S1 and S2). For each individual sample, a genome-wide protein–DNA binding intensity profile was generated and stored in the database. These profiles enable one to examine variability across samples, such as variability of biological or technical replicates. Details of data collection and processing can be found in Methods 1–4 in Supplementary Material.

Data in hmChIP can be queried and retrieved through a web server (Fig. 1a). Users can search for available experiments by providing protein names, cell types and/or a list of genomic regions in BED or COD file format (see examples). If a genomic region list is provided, the returned experiments will be rank ordered based on the degree of overlap between the query region list and the peak list of each experiment (Method 5 in Supplementary Material). For each experiment, the query result will list samples in the experiment and provide download links for the associated

*To whom correspondence should be addressed.

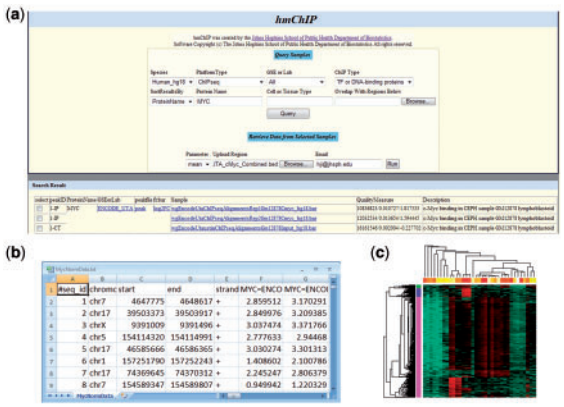

Fig. 1. Illustration of hmChIP. (**a**) The web interface for data query and retrieval. (**b**) Binding intensities are retrieved and returned as text files. (**c**) A hierarchical clustering heat map will be returned as well. Three detailed examples and colored screen shots in larger size are provided in Supplementary Figures S1–S3.

peak list, sample binding intensity profiles and a $\log_2$ fold-change profile of the experiment generated by comparing ChIP and control binding intensities. For each sample, several quality measures are provided, including percentage of the genome covered by peaks, average signal-to-noise ratio and total read count in the sample if the sample is ChIP-seq (Method 3 in Supplementary Material).

From the query results, users can select samples of interest, provide a list of genomic regions and provide an email address. By clicking a 'Run' button, binding intensities from the selected samples will be retrieved for the genomic regions provided, saved into a text file and returned to users through email (Fig. 1b; Method 6 in Supplementary Material). If no email address is provided, the results will be returned through a web page. To facilitate cross-sample comparisons, intensity data from different samples will be normalized, and the normalized data will be returned in a separate text file. A heat map showing the hierarchical clustering of genomic regions and samples based on the normalized intensities will be returned as well (Fig. 1c; Methods 6 and 7 in Supplementary Material). For each genomic region and each peak list with samples selected, hmChIP will return a binary value to indicate the region's binding status in that peak list and a number measuring the $\log_2$ fold-change between ChIP and control binding intensities in the region (Method 6 in Supplementary Material). Data retrieved from hmChIP can be fed into other software tools to carry out further analyses, such as customized clustering using dChip (Li and Wong, 2003), or correlating binding intensities with gene expression data (Ouyang *et al*., 2009).

3 EXAMPLES

Sox17 is involved in the differentiation of mESC. Oct4, Sox2 and Nanog are master regulators to maintain mESC's pluripotency and self-renewal ability. To explore whether Sox17 can interact with these TFs, we queried hmChIP and selected mouse ChIP-chip samples for Sox17 in XEN cells, samples for Oct4, Sox2 and Nanog in mESC and the corresponding control samples (Supplementary Figure S1). We extracted binding intensities from these samples in Sox17 binding regions. Interestingly, the clustering heat map shows that a significant subset of Sox17 binding regions was also bound by Sox2 and Nanog. Next, we queried both ChIP-chip and ChIP-seq data in hmChIP using Sox17 binding regions as input. Through this more unbiased search, we found a number of TFs whose binding sites overlapped with Sox17 binding sites at levels above random expectation, with Sox2 and Nanog ranked the highest (Supplementary Figure S2). Our result suggests that Sox17 may promote differentiation partly by competing with Sox2 and Nanog to bind to the same DNA-binding sites. This analysis has led to follow-up experiments which verified the competition between Sox17 and Nanog in mESC differentiation (Niakan *et al*., 2010).

In another example, we studied c-Myc binding in multiple cancer cell lines using public ChIP-seq data from different labs. The results revealed different classes of c-Myc binding sites based on their cell type dependencies, and clustered samples based on their cell types rather than lab origins (Supplementary Figure S3).

4 DISCUSSION

hmChIP removes a major hurdle for scientists to retrieve and utilize ChIP data in public domains. In future, it will be gradually enhanced to include more data and functionalities (e.g. tools for exploring spatial binding patterns across multiple ChIP samples). Our ultimate goal is to turn it into a toolbox for biologists to efficiently integrate publicly available ChIP-chip and ChIP-seq data to study gene regulation and make novel discoveries.

Funding: National Institute of Health (R01HG005220).

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

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

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

Judy,J.T. and Ji,H. (2009) TileProbe: modeling tiling array probe effects using publicly available data. *Bioinformatics*, **25**, 2369–2375.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Lachmann,A. *et al*. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

Li,C. and Wong,W.H. (2003) DNA-Chip Analyzer (dChip). In Parmigiani,G. *et al*. (eds) *The analysis of gene expression data: methods and software*. Springer, New York, pp. 120–141.

Niakan,K.K. *et al*. (2010) Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. *Genes Dev.*, **24**, 312–326.

Ouyang,Z. *et al*. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **106**, 21521–21526.

Ren,B. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Rosenbloom,K.R. *et al*. (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **38**, D620–D625.

Wheeler,D.L. *et al*. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.