BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 10 2012, pages 1411–1412
doi:10.1093/bioinformatics/bts157

Databases and ontologies

Advance Access publication April 11, 2012

CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human

Bo Qin[1,†], Meng Zhou[2,†], Ying Ge[1,†], Len Taing[3,4], Tao Liu[3,4], Qian Wang[1], Su Wang[1], Junsheng Chen[1], Lingling Shen[5], Xikun Duan[1], Sheng'en Hu[1], Wei Li[6], Henry Long[3], Yong Zhang[1,*] and X. Shirley Liu[3,4,*]

[1]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, 200092, China, [2]Department of Biological Sciences, Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA 90089, [3]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, [4]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 450 Brookline Ave, Boston, MA, 02115, USA, [5]Department of Information Technology, China Novartis Institutes for BioMedical Research Co.,Ltd., Shanghai 201203, China and [6]Division of Biostatistics, Dan L Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Transcription and chromatin regulators, and histone modifications play essential roles in gene expression regulation. We have created CistromeMap as a web server to provide a comprehensive knowledgebase of all of the publicly available ChIP-Seq and DNase-Seq data in mouse and human. We have also manually curated metadata to ensure annotation consistency, and developed a user-friendly display matrix for quick navigation and retrieval of data for specific factors, cells and papers. Finally, we provide users with summary statistics of ChIP-Seq and DNase-Seq studies.

Availability: Freely available on the web at http://cistrome.dfci.harvard.edu/pc/

Contact: yzhang@tongji.edu.cn; xsliu@jimmy.harvard.edu

Received on January 12, 2012; revised on March 14, 2012; accepted on March 29, 2012

1 INTRODUCTION

Transcription regulators, chromatin regulators and histone modifications play key roles in eukaryotic gene transcription regulation in normal physiology and diseases (Kouzarides, 2007; Pan *et al.*, 2010). Here transcription regulators include transcription factors, transcription initiation complexes, mediator complex and transcription co-activators/repressors. Chromatin regulators refer to chromatin-associated factors such as chromatin remodelers, histone modifying enzymes and other factors associated with specific chromosome or nuclear compartments. Mapping the genome-wide *in vivo* locations of these factors and marks is important to understand the mechanisms of transcriptional and epigenetic regulation. With the development of high-throughput sequencing, ChIP-Seq and DNase-Seq have become popular techniques to capture the genome-wide *in vivo* locations of regulatory factors, histone modifications and chromatin accessible regions.

Several data depositories contain the majority of the publicly available ChIP-Seq and DNase-Seq data. They include SRA (for raw data) and GEO (for processed data) at NCBI, ENA at EBI, ENCODE data at UCSC Genome Browser. In addition, some published datasets are hosted on the authors' websites. Most of these data are accompanied with submitter-supplied metadata for each experiment (Barrett *et al.*, 2009; Barrett *et al.*, 2011; Fingerman *et al.*, 2011). However, there are many inconsistencies in the metadata annotation, such as factor naming, cell types and disease states discrepancies. These make data search and retrieval inefficient and create challenges for public data reuse.

Several resources, such as nuclear receptor Cistrome (Lanz *et al.*, 2006; Tang *et al.*, 2011), NCBI Epigenomics (Fingerman *et al.*, 2011) and hmChIP (Chen *et al.*, 2011), have provided annotated metadata for some ChIP-Seq studies. However, they focused on specific subset of ChIP-Seq experiments, and some of their metadata annotations are also limited. There is a great need to efficiently query all publicly available ChIP-Seq and DNase-Seq data in mouse and human, as well as an infrastructure for regular update. With a well-curated knowledgebase and controlled vocabulary for metadata annotation, users could very quickly find all the ChIP-Seq and DNase-Seq data for a particular factor, all the factors that have been profiled in a particular cell or from a particular study.

In this study, we created the CistromeMap knowledgebase to fill this gap. It contains manually curated annotations from all published studies and publicly available data on ChIP-Seq and DNase-Seq experiments in mouse and human as of September 2011. CistromeMap also has a user-friendly web interface allowing researchers to quickly find and navigate the data. The resource not only saves experimental labs the time and money from unnecessarily duplicating an experiment, but also allows them to analyze their own data in the context of other related public datasets. It also enables computational labs to more efficiently conduct integrative analysis of all the publicly available data.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

2 METHODS

We collected published or public ChIP-Seq and DNase-Seq datasets for mouse and human experiments from public repositories (i.e. GEO, SRA and ENA), the ENCODE and Epigenome Roadmap projects, and the remaining from authors' website. In total, CistromeMap contains 6474 ChIP-Seq and DNaes-Seq samples. We systematically annotated the following metadata for each sample: cell line/population, cell type, tissue origin, strain (for mouse), disease state, factor name, PubMed ID (for published data), data source, reference and last author. Based on the original literature and online information about official gene symbol, cell line and tissue origin, we created our own ontology for better annotation and organization.

The CistromeMap user interface (UI) contains three view tabs. The first is a factor view. Users first select the factor(s) of interest, and CistromeMap searches the database for all the cells where a ChIP-Seq has been conducted for the factor and displays the results in a data matrix. When users click on specific entries in the data matrix, the CistromeMap UI displays a summary view of the ChIP-Seq factor in a specific cell, such as reference (link to PubMed entry for published data), lab, data source (link to data download), species, cell information and disease state. The second tab is a cell view, where users can get answers to questions such as 'what factors have been profiled in HeLa?' The results are displayed in similar format as the factor view. The third tab is a paper view, where a user could look at all the ChIP-Seq or DNase-Seq papers by author, title, publication date and journal. When a user clicks on a particular paper, it also displays information such as author list, abstract, data source, species and factors profiled. To further support browse function, we developed a built-in text search engine to help users refine the contents in view tabs. Help document describing all the functions is available online.

3 RESULTS

In summary, CistromeMap is a comprehensive and consistently annotated knowledgebase of all published or public ChIP-Seq and DNase-Seq data in mouse and human. In total, there are 2711 ChIP-Seq datasets for transcription and chromatin regulators, 2355 for histone modifications and variants, 412 DNase-Seq and 996 control datasets. Among transcription and chromatin regulators, POLR2A, CTCF, ESR1, RELA and EP300 are the most often profiled ChIP-Seq factors. For histone marks, H3K4me3, H3K27me3, H3K4me1, H3K36me3 and H3K9me3 ChIP-Seq are the most common, which together accounts for over 70% of all of the histone ChIP-Seq data. In addition, Bernstein, B.E., Stamatoyannopoulos, J.A., Snyder, M., Ren, B., Myers, R.M. and Zhao, K. laboratories have generated the highest number of ChIP-Seq or DNase-Seq data for the community. Interestingly, *Nature Biotechnology* and *Nature* are the top two journals where the largest number of mouse and human ChIP-Seq or DNase-Seq studies were published. More details of the above statistics are available at http://cistrome.dfci.harvard.edu/pc/dcstats/, and will be automatically updated as more ChIP-Seq and DNase-Seq data become available.

As the number of published and public ChIP-Seq and DNase-Seq increases exponentially, we will continue to improve and maintain the CistromeMap. In addition, we provide ways for the community to help us update CistromeMap: users can input the PubMed ID or GEO ID accession to inform us of newly available data, so we can quickly update the CistromeMap knowledgebase. Based on the ID, CistromeMap has a function to automatically retrieve the related metadata from PubMed or GEO. After manual curation, the new data will be available for release on CistromeMap. With the joint effort from our laboratory and the community, we believe that CistromeMap will be a valuable resource for the greater scientific community.

ACKNOWLEDGEMENTS

We thank Martha Bulyk for her initial suggestion of the project. We also thank Myles Brown, Kornelia Polyak, Ramesh Shivdasani, Prakash Rao, Qixuan Wang, Juan Wang, Jing Liu, Xueqiu Lin, Haiyang Zheng and Xiaofeng Wang for their help in data collection and curation, as well as Jie Wang for her technical help on the user interface.

Funding: National Basic Research (973) Program of China [2010CB944904], National Natural Science Foundation of China [31028011] and National Institutes of Health [HG4069].

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.

Chen,L. *et al.* (2011) hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, **27**, 1447–1448.

Fingerman,I.M. *et al.* (2011) NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res.*, **39**, D908–D912.

Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

Lanz,R.B. *et al.* (2006) Nuclear Receptor Signaling Atlas (www.nursa.org): hyperlinking the nuclear receptor signaling community. *Nucleic Acids Res.*, **34**, D221–D226.

Pan,Y. *et al.* (2010) Mechanisms of transcription factor selectivity. *Trends Genet.*, **26**, 75–83.

Tang,Q. *et al.* (2011) A comprehensive view of nuclear receptor cancer cistromes, *Cancer Res.*, **71**, 6940–6947.