BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 16 2012, pages 2205–2206
doi:10.1093/bioinformatics/bts340

Databases and ontologies
Advance Access publication June 19, 2012

ChromoHub: a data hub for navigators of chromatin-mediated signalling

Lihua Liu[1,†], Xi Ting Zhen[1,†], Emily Denton[1], Brian D. Marsden[2] and Matthieu Schapira[1,3,*]

[1]Structural Genomics Consortium, University of Toronto,Toronto, ON M5G1L7, Canada, [2]Structural Genomics Consortium, Nuffield Department of Clinical Medicine, Oxford University, Headington, OX3 7DQ, UK and [3]Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON, M5S 1A8, Canada

Associate Editor: Martin Bishop

ABSTRACT

Summary: The rapidly increasing research activity focused on chromatin-mediated regulation of epigenetic mechanisms is generating waves of data on writers, readers and erasers of the histone code, such as protein methyltransferases, bromodomains or histone deacetylases. To make these data easily accessible to communities of research scientists coming from diverse horizons, we have created ChromoHub, an online resource where users can map on phylogenetic trees disease associations, protein structures, chemical inhibitors, histone substrates, chromosomal aberrations and other types of data extracted from public repositories and the published literature. The interface can be used to define the structural or chemical coverage of a protein family, highlight domain architectures, interrogate disease relevance or zoom in on specific genes for more detailed information. This open-access resource should serve as a hub for cell biologists, medicinal chemists, structural biologists and other navigators that explore the biology of chromatin signalling.

Availability: http://www.thesgc.org/chromohub/.

Contact: matthieu.schapira@utoronto.ca

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

Received on May 2, 2012; revised on June 6, 2012; accepted on June 7, 2012

1 INTRODUCTION

Chromatin-mediated control of gene expression and cell fate is regulated in part by distinct combinations of post-translational modifications on histone proteins, predominantly methylation or acetylation of lysine or arginine side chains at the N-terminal tails of histones (Fierz and Muir, 2012; Kouzarides, 2007; Strahl and Allis, 2000). Alteration of this histone code can lead to diseases states, and chemical inhibition of proteins that write, read or erase histone marks represents a promising avenue to restore to normal level disease-associated gene expression (Arrowsmith *et al*., 2012). Following the clinical validation of this strategy with histone deacetylase (HDAC) inhibitors (Prince *et al*., 2009), writers, readers and erasers of histone marks constitute emerging therapeutic target classes for a variety of disease conditions. Consequently, a dramatic increase in research activity has been observed in the field. The rapidly growing body of heterogeneous data generated by diverse communities of scientists is in part accessible through public repositories or the published literature. But retrieving the data is time-consuming at best for the non-specialist: will a cell biologist know where to find what small molecule inhibitors are available for his protein of interest, and what their IC_{50}'s are? Will a medicinal chemist easily retrieve all compounds co-crystallized with her protein target? Will a biochemist know which somatic aberration, buried in vast cancer genomics databases, is linked to the protein he is characterizing?

We have built a database that integrates such heterogeneous data types extracted from multiple repositories and the published literature. A simple yet powerful web interface allows research scientists coming from diverse horizons who are interested in writers, readers and erasers of the histone code to interrogate this database through phylogenetic representations of each protein family. The interface is freely available and should promote cross-pollination between diverse communities of scientists interested in epigenetic signalling.

2 METHODS

2.1 Assembling protein families

Human protein families were defined by the presence of specific domains involved in writing, reading and erasing histone marks (Arrowsmith *et al*., 2012; Kouzarides, 2007): protein methyltransferase (PMT) and histone acetyltransferase (HAT) domains for writers, lysine demethylase (KDM) and HDAC domains for erasers and bromodomains (BRD) for readers of acetylated lysines. Different domains are known to bind methyl-lysines (Tudor, MBT, Chromo, PWWP, PHD and BAH), and each defined a subfamily of its own (Kuo *et al*., 2012; Taverna *et al*., 2007). The human protein reference database (Keshava Prasad *et al*., 2009), the PFAM (Punta *et al*., 2012) and SMART databases (Schultz *et al*., 2000) were queried to retrieve all human genes containing at least one of these domains. Duplicates were removed and missing genes clearly documented in the published literature were added manually.

2.2 Generating phylogenetic trees

For each protein family, two phylogenetic trees were produced. The first was based on a ClustalW (Larkin *et al*., 2007) multiple sequence alignment of the default UniProt protein variant of each human gene. The second was based on a multiple sequence alignment of the domain after which the family was named (a domain-based tree was not generated for HATs as the catalytic domain is not always clearly defined for this family). In this case, a seed

[*]To whom correspondence should be addressed.

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

sequence alignment was derived from available protein structures by aligning residues that were superimposed in the three-dimensional space in ICM (Molsoft, San Diego). Additional sequences were appended by aligning them to the closest seed sequence in ICM. A PHP script plotted a phylogenetic tree from the Newick string of the multiple sequence alignment and automatically defined X,Y coordinates next to each leaf of the tree for metadata mapping (Supplementary Methods). We verified that this methodology produced a phylogeny in agreement with trees previously published in the literature (Filippakopoulos *et al*., 2012; Richon *et al*., 2011). A larger version of the PMT family was reported that includes numerous putative arginine methyltransferases; these were not included as the authors of that work stated that they did not want to imply that these proteins are protein arginine methyltransferases *per se* (Richon *et al*., 2011).

2.3 Metadata source

Data related to the biology, structural and chemical coverage of each gene were extracted from diverse repositories and stored in MySQL. Function summary, sub-cellular location and polymorphisms were retrieved from UniProt records. Tissue expression data were collected from the GNF's BioGPS (Wu *et al*., 2009). Cancer-associated chromosomal aberrations were extracted from the Mitelman database (http://cgap.nci.nih.gov/Chromosomes/Mitelman) and the Sanger Institute's cancer gene census (http://www.sanger.ac.uk/genetics/CGP/Census/). Protein interactions were from the String database (Szklarczyk *et al*., 2011). Structural coverage was produced by querying the Protein Databank (http://www.rcsb.org/pdb) with Blast. Protein domain architecture was defined by querying the PFAM database with HMMER (*e*-value cutoff of 0.01) (Sonnhammer *et al*., 1998). NIH funding was extracted from NIH's RePORT (http://projectreporter.nih.gov/reporter.cfm) and published literature from Pubmed. NCBI's built-in links between Pubmed records and genes were used to retrieve articles associated to human, mouse or rat orthologues of the gene of interest, and keywords embedded in Pubmed's MeSH terms served to associate Pubmed records with diseases. Histone substrates and chemical inhibitors were manually extracted from the literature and all records were linked to their respective Pubmed or patent reference. All chemical inhibitors from BindingDb can also be mapped on the trees (Liu *et al*., 2007). Pubmed records, disease association, funding and structure coverage are updated automatically on a weekly basis. Other data are updated manually.

3 RESULTS

The online user interface is based on phylogenetic representations of protein families involved in writing, reading and erasing histone post-translational modifications. Users can choose between phylogenetic classification derived from multiple alignments of full-length sequences or sequences of the domain after which the family was named. Thumbnails of phylogenetic trees for each protein family can be clicked to display larger images. Once a tree is selected, the sequence alignment used to generate the tree can be downloaded. Checkboxes can be selected to map a diverse array of data on the tree of interest. Information on the data source is provided in a window that pops-up when hovering over a [i] icon next to the checkbox. Once a checkbox is selected, associated symbols are shown next to each protein for which data is available. More information is then accessible by hovering over or clicking on the symbol of interest.

Users can easily navigate the functional, structural and chemical landscape of each protein family. They can display functional summaries for each gene on the trees, list structures in the Protein Databank covering each gene and map them on linear representations of the protein where PFAM domains are highlighted, display small molecule co-crystallized with any protein or retrieve chemical inhibitors reported in the published or patent literature; they can see the number of entries in Pubmed for each gene and inspect disease associations automatically inferred from Pubmed records; users can easily access chromosomal aberrations linked to cancer, tissue expression data, sub-cellular location or histone substrates. Images can be saved on the desktop and embedded in presentations. Newcomers in the field can search for potential collaborators by looking for research laboratories with active funding on their gene of interest.

4 CONCLUSION

The explosion of research activity on epigenetic signalling and recent technological breakthroughs in genome-scale biology are providing a wealth of data related to writers, readers and erasers of histone marks. The open-access resource that we have developed should help research scientists involved in chromatin biology rapidly find data that inform their research.

Funding: The Structural Genomics Consortium is a registered charity (1097737) that receives funds from Canadian Institutes of Health Research, Eli Lilly Canada, Genome Canada, GlaxoSmithKline, the Ontario Ministry of Economic Development and Innovation and the Novartis Research Foundation, Pfizer, Abbott, Takeda and the Wellcome Trust.

Conflict of Interest: None declared.

REFERENCES

Arrowsmith,C.H. *et al*. (2012) Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discov.*, **11**, 384–400.

Fierz,B. and Muir,T.W. (2012) Chromatin as an expansive canvas for chemical biology. *Nat. Chem. Biol.*, **8**, 417–427.

Filippakopoulos,P. *et al*. (2012) Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell*, **149**, 214–231.

Keshava Prasad,T.S. *et al*. (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

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

Kuo,A.J. *et al*. (2012) The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier–Gorlin syndrome. *Nature*, **484**, 115–119.

Larkin,M.A. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Liu,T. *et al*. (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities *Nucleic Acids Res*, **35**, D198–D201.

Prince,H.M. *et al*. (2009) Clinical studies of histone deacetylase inhibitors. *Clin. Cancer Res.*, **15**, 3958–3969.

Punta,M. *et al*. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Richon,V.M. *et al*. (2011) Chemogenetic analysis of human protein methyltransferases. *Chem. Biol. Drug Des.*, **78**, 199–210.

Schultz,J. *et al*. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.

Sonnhammer,E. *et al*. (1998) PFAM: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res.*, **26**, 320–322.

Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

Szklarczyk,D. *et al*. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.

Taverna,S.D. *et al*. (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.*, **14**, 1025–1040.

Wu,C. *et al*. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.