BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 4 2014, pages 590–592
doi:10.1093/bioinformatics/btt710

Databases and ontologies
Advance Access publication December 6, 2013

ChromoHub V2: cancer genomics

Muhammad A. Shah[1], Emily L. Denton[1], Lihua Liu[1] and Matthieu Schapira[1,2,*]

[1]Structural Genomics Consortium, and [2]Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON M5G 1L7, Canada

Associate Editor: John Hancock

ABSTRACT

Summary: Cancer genomics data produced by next-generation sequencing support the notion that epigenetic mechanisms play a central role in cancer. We have previously developed Chromohub, an open access online interface where users can map chemical, structural and biological data from public repositories on phylogenetic trees of protein families involved in chromatin mediated-signaling. Here, we describe a cancer genomics interface that was recently added to Chromohub; the frequency of mutation, amplification and change in expression of chromatin factors across large cohorts of cancer patients is regularly extracted from The Cancer Genome Atlas and the International Cancer Genome Consortium and can now be mapped on phylogenetic trees of epigenetic protein families. Explorators of chromatin signaling can now easily navigate the cancer genomics landscape of writers, readers and erasers of histone marks, chromatin remodeling complexes, histones and their chaperones.

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

Contact: matthieu.schapira@utoronto.ca

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

Received on October 15, 2013; revised on November 18, 2013; accepted on December 1, 2013

1 INTRODUCTION

Chromohub is an online interface that allows the epigenetics research community to project biological, structural and chemical data on phylogenetic trees of protein families involved in chromatin-mediated signaling (Liu *et al.*, 2012). The interface is a useful hub for cell biologists to find chemical inhibitors targeting their proteins of interest, medicinal chemists to inspect the structural coverage of specific binding sites or structural biologists to visualize the disease association of phylogenetic neighbors to the construct they crystallized. We previously described how protein families were assembled, phylogenetic trees generated and biological, structural and chemical data extracted from public repositories and mapped on the trees (Liu *et al.*, 2012). We have now added to Chromohub a large section entirely focused on genomic data from cancer patients extracted from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC).

Recent landmark next-generation sequencing campaigns of large cancer patient cohorts have revealed recurrent alterations of genes involved in epigenetic mechanisms (Biankin *et al.*, 2012; Dalgliesh *et al.*, 2010; Ellis *et al.*, 2012; Ho *et al.*, 2013; Jones *et al.*, 2012; Le Gallo *et al.*, 2012; Morin *et al.*, 2011; Pugh *et al.*, 2012; Robinson *et al.*, 2012; Schwartzentruber *et al.*, 2012; Stephens *et al.*, 2012; Varela *et al.*, 2011; Zhang *et al.*, 2012). These results support the notion that chromatin-mediated signaling may be central to cancer initiation and progression (Baylin and Jones, 2011; You and Jones, 2012). The data associated with most of these and other unbiased cancer genomic projects were deposited into TCGA and the ICGC repositories, and made publicly accessible to the scientific community. Chromohub users can now map cancer genomics data on phylogenetic trees of protein families involved in epigenetic mechanisms.

2 METHODS

2.1 Data sources

RNASeq gene expression data, promoter and full genome methylation data and somatic mutation data were downloaded from TCGA's Firehose data run (https://confluence.broadinstitute.org/display/GDAC/Dashboard-Stddata). GISTIC copy number data were downloaded via TCGA's Firehose analyses run (https://confluence.broadinstitute.org/display/GDAC/Dashboard-Analyses). Furthermore, somatic mutation data are also extracted from ICGC's Data Portal (http://dcc.icgc.org/). All data were stored in a MySQL database. A list detailing all datasets as of November 2013 underlying Chromohub's cancer genomics interface is provided in Supplementary Table S1.

2.2 Somatic mutations

Only data derived from patients with both a tumor and a matched normal sample were used. Using an anonymized patient identification code for each patient, the overall number of genes mutated within the patient's genome is stored and is used to filter out genomes that are hypermutated. A protein image is presented showing all mutations matching the set cutoffs; hovering over the mutations shows the amino acid change. When not explicitly specified by TCGA or ICGC, amino acid mutations are derived from genomic location, strand and mutated nucleotide.

2.3 RNASeq gene expression

Only data from patients with matched tumor and normal samples were used. RSEM values are used to quantify messenger RNA (mRNA) expression levels (RNASeq V2 data). A $\log_2$ fold change in gene expression is calculated from RSEM values of tumor and matched normal samples as follows:

$$\log_2\text{fold change} = \text{Log}_2(\text{RSEM[tumor]}/\text{RSEM[matched normal]})$$

Underexpressed genes have negative $\log_2$ values; overexpressed genes have positive $\log_2$ values. A rank is also generated for each gene, which is determined by ordering the frequency of over/underexpression of all genes (with available data using the specified cutoffs).

*To whom correspondence should be addressed.

2.4 Copy number variation

The GISTIC 2.0 algorithm (Mermel *et al*., 2011) is used to produce copy number variation data. This preprocessing step is conducted by TCGA's GDAC Firehose and the results are provided. Using anonimized patient identification codes, for each patient, the overall number of genes with gains/losses within the patient's genome is stored and is used to filter out genomes with a high number of aberrations.

2.5 GISTIC copy number variation versus RNASeq gene expression

Anonymous patient identification numbers, provided by TCGA, were used to determine patients where both GISTIC copy number and RNASeq gene expression data were available. These data were used to find correlations between copy number variation and gene expression levels in tumor samples.

2.6 Promoter methylation in cancer

Promoter methylation data are downloaded exclusively from TCGA's Firehose, but it is derived from two platforms, Human Methylation 27 k (strictly promoter methylation) and Human Methylation 450 k (whole genome methylation). Promoter methylation using the Human Methylation 450 k array was defined as 1000 bp upstream the transcription start site, which was determined for all genes using coordinates from the refGene table from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start).

3 RESULTS

Rather than listing gene-specific links to existing cancer genomics portals, Chromohub provides integrated data focused on chromatin signaling. Users can visualize on phylogenetic trees of protein families involved in epigenetic mechanisms the percent of tumor samples across large patient cohorts where a gene is mutated (compared with a non-tumor sample from the same patient). Highly mutated genomes can be excluded from the analysis by setting a threshold for the maximum number of genes mutated in a sample. The output is grouped by cancer type. As of October 2013, 16 cancer types are represented by cohorts of >100 patients. High or low copy number gains as well as heterozygous and homozygous deletions [corresponding to GISTIC values of 2, 1, −1 and −2, respectively (Mermel *et al*., 2011)] can also be plotted on phylogenetic trees. Statistically relevant data (>100 patients) are available for nine cancer types. Unlike mutation data, copy numbers are compared with those in the reference human genome.

In addition to chromosomal aberrations, changes in transcription profiles are also available: mRNA levels are compared between tumor and non-tumor samples from the same patient and tissue. This provides a bird's eye view of genes that are overexpressed or repressed in specific cancer types for any protein family related to epigenetic mechanisms. Orthogonal data types can be projected on a tree simultaneously. For instance, combining mRNA expression and mutation data, users can rapidly see that the histone methyltransferase MLL3 is mutated in 7% (54 of 776) and repressed in 21% (23 of 107) of breast cancer samples, suggesting that this gene acts as a tumor suppressor.

Change in expression of a given gene is generally not driving cancer initiation or progression, but simply a passenger event (Hanahan and Weinberg, 2011), unless it is directly caused by a chromosomal amplification or deletion (Beroukhim *et al*., 2007; Eifert and Powers, 2012). To identify candidate driver events affecting chromatin factors, Chromohub allows users to automatically highlight genes where overexpression correlates with copy number gains. Using this approach, one can rapidly see that, among genes containing a Tudor domain (which bind methylated lysines and arginines), FXR1 is overexpressed and amplified in 53% (18 of 34) lung squamous cell carcinoma patients.

4 CONCLUSION

Dysregulation of the chromatin signaling platform plays a major role in cancer (Baylin and Jones, 2011; Timp and Feinberg, 2013; You and Jones, 2012); chromosomal aberrations and transcriptional alteration affecting chromatin factors can drive initiation and development of specific cancer types. The new Chromohub interface is a simple tool to navigate the cancer genomics of epigenetic mechanisms.

Funding: The SGC is a registered charity (1097737) that receives funds from AbbVie, Boehringer Ingelheim, the Canada Foundation for Innovation, the Canadian Institutes for Health Research, Genome Canada through the Ontario Genomics Institute [OGI-055], GlaxoSmithKline, Janssen, Eli Lilly Canada, the Novartis Research Foundation, the Ontario Ministry of Economic Development and Innovation, Pfizer, Takeda and the Wellcome Trust [092809/Z/10/Z]

Conflicts of Interest: none declared.

REFERENCES

Baylin,S.B. and Jones,P.A. (2011) A decade of exploring the cancer epigenome–biological and translational implications. *Nat. Rev. Cancer*, **11**, 726–734.

Beroukhim,R. *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. U. S. A.*, **104**, 20007–20012.

Biankin,A.V. *et al*. (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.

Dalgliesh,G.L. *et al*. (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, **463**, 360–363.

Eifert,C. and Powers,R.S. (2012) From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets. *Nat. Rev. Cancer*, **12**, 572–578.

Ellis,M.J. *et al*. (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, **486**, 353–360.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Ho,A.S. *et al*. (2013) The mutational landscape of adenoid cystic carcinoma. *Nat. Genet*., **45**, 791–798.

Jones,D.T. *et al*. (2012) Dissecting the genomic complexity underlying medulloblastoma. *Nature*, **488**, 100–105.

Le Gallo,M. *et al*. (2012) Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat. Genet*., **44**, 1310–1315.

Liu,L. *et al*. (2012) ChromoHub: a data hub for navigators of chromatin-mediated signalling. *Bioinformatics*, **28**, 2205–2206.

Mermel,C.H. *et al*. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*., **12**, R41.

Morin,R.D. *et al*. (2011) Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, **476**, 298–303.

Pugh,T.J. *et al*. (2012) Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*, **488**, 106–110.

Robinson,G. *et al*. (2012) Novel mutations target distinct subgroups of medulloblastoma. *Nature*, **488**, 43–48.

Schwartzentruber,J. *et al*. (2012) Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*, **482**, 226–231.

Stephens,P.J. *et al*. (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature*, **486**, 400–404.

Timp,W. and Feinberg,A.P. (2013) Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat*. *Rev*. *Cancer*, **13**, 497–510.

Varela,I. *et al*. (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, **469**, 539–542.

You,J.S. and Jones,P.A. (2012) Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell*, **22**, 9–20.

Zhang,J. *et al*. (2012) The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*, **481**, 157–163.