ChromoHub V2: cancer genomics
Muhammad A. Shah1, Emily L. Denton1, Lihua Liu1 and Matthieu Schapira1,2,*
1Structural Genomics Consortium, and 2Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON MSG 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/Dashborad-Sddata). 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 log2 fold change in gene expression is calculated from RSEM values of tumor and matched normal samples as follows:

\[
\log_2 \text{fold change} = \log_2 \left( \frac{\text{RSEM}_{\text{tumor}}}{\text{RSEM}_{\text{normal}}} \right)
\]

Underexpressed genes have negative log2 values; overexpressed genes have positive log2 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 anonymized 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 27k (strictly promoter methylation) and Human Methylation 450k (whole genome methylation). Promoter methylation using the Human Methylation 450k array was defined as 1000bp 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. Low or high 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 over-expressed 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, 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