<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>CR2Cancer: a database for chromatin regulators in human cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>RU, B; SUN, J; TONG, Y; Wong, CN; Chandra, A; Tang, ATS; Chow, LKY; Wun, WL; Levitskaya, Z; Zhang, J</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Nucleic Acids Research, 2018, v. 46 n. D1, p. D918-D924</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2018</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/249561">http://hdl.handle.net/10722/249561</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
CR2Cancer: a database for chromatin regulators in human cancer

Beibei Ru, Jianlong Sun, Yin Tong, Ching Ngar Wong, Aditi Chandra, Acacia Tsz So Tang, Larry Ka Yue Chow, Wai Lam Wun, Zarina Levitskaya and Jiangwen Zhang*

School of Biological Sciences, The University of Hong Kong, Hong Kong 999077, China

Received August 14, 2017; Revised September 16, 2017; Editorial Decision September 19, 2017; Accepted September 28, 2017

ABSTRACT
Chromatin regulators (CRs) can dynamically modulate chromatin architecture to epigenetically regulate gene expression in response to intrinsic and extrinsic signalling cues. Somatic alterations or mis-expression of CRs might reprogram the epigenomic landscape of chromatin, which in turn lead to a wide range of common diseases, notably cancer. Here, we present CR2Cancer, a comprehensive annotation and visualization database for CRs in human cancer constructed by high throughput data analysis and literature mining. We collected and integrated genomic, transcriptomic, proteomic, clinical and functional information for over 400 CRs across multiple cancer types. We also built diverse types of CR-associated relations, including cancer type dependent (CR-target and miRNA-CR) and independent (protein-protein interaction and drug-target) ones. Furthermore, we manually curated around 6000 items of aberrant molecular alterations and interactions of CRs in cancer development from 5007 publications. CR2Cancer provides a user-friendly web interface to conveniently browse, search and download data of interest. We believe that this database would become a valuable resource for cancer epigenetics investigation and potential clinical application. CR2Cancer is freely available at http://cis.hku.hk/CR2Cancer.

INTRODUCTION
Chromatin biology has been the focus of intense research due to its significant implication in development and pathogenesis (1). Chromatin is a macromolecular complex of DNA and histone proteins forming repeating units of nucleosome, which consists of 147 bp of DNA wrapped around an octamer of histones (two copies of H2A, H2B, H3 and H4) (2). The interplay of epigenetic mechanisms regulates the architecture of chromatin to control DNA-templated processes such as transcription, replication and repair (3,4). These epigenetic marks mainly include DNA methylation, covalent histone modifications (e.g. acetylation and methylation) and nucleosome placement, which constitute a dynamically changing and complex epigenomic landscape. Chromatin regulators (CRs), a class of enzymes with specialized function domains, can recognize, shape and maintain the epigenetic state in a cell context-dependent fashion (5). For instance, DNA methyltransferases (DNMTs) transfer methyl groups to 5’ position of cytosines in CpG dinucleotides while ten-eleven translocation (TET) enzymes catalyze demethylation (6). Histone acetyltransferases (HATs) and deacetylases (HDACs) can deposit and remove acetyl groups to lysine amino acids on histone N-terminal tails, respectively (7). Chromatin remodeling complexes are capable of either moving, evicting or restructuring nucleosomes in an ATP-dependent manner (8).

Genomic alterations or dysregulation of CRs have now been identified in a wide range of cancer types and are increasingly regarded as novel therapeutic targets (9–11). Since CRs can regulate local or global epigenetic patterns to influence multiple target genes simultaneously, aberrant CRs offer cancer cells a highly efficient mechanism to rewire transcriptional regulatory circuits. Recent cancer genome sequencing studies showed that CR genes are among the most recurrently mutated gene sets (e.g. DNMT3A, EP300, CREBBP, ARID1A and SMARCB1), and 25%-30% of the identified driver mutations have an effect on genes encoding CRs (12,13). For example, the histone methyltransferase EZH2 is mutated in 22% of diffuse large B cell lymphoma cases, which directly promotes the trimethylation of histone H3 at K27 mark (H3K27me3) to repress the expression of genes required for cell cycle exit and cell differentiation (14,15). In addition, the reversible nature of epigenetic modifications provides plausible prevention and treatment options for cancer therapy through targeting CRs (16,17). As the first epigenetic therapies approved in 2005, two DNMT1 inhibitors (5-azacytidine and 5-aza-2’-deoxycytidine) can induce global hypomethylation for the treatment of myelodysplastic syndrome (18,19). Subsequently, more than 6 molecules inhibiting the activity of HDACs have also been granted for clinical use in sev-

*To whom correspondence should be addressed. Tel: +852 2299 0341; Fax: +852 2559 9114; Email: jzhang1@hku.hk

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
eral cancers (20). Apart from these approved drugs, a large number of novel molecules targeting various CRs, such as EZH2, EP300, SIRT1 and LSD1, are currently under development or in clinical trials, which provide an exciting outlook for cancer epigenetic therapies (16,17).

It is therefore clear that comprehensive collection of CRs together with related annotation information would contribute greatly to the study of cancer development and treatment. Currently, there are three major databases developed for CRs and histone modifications. Histone collects 55 histone proteins, 106 distinct sites of their modifications and 152 histone-modifying enzymes in human (21). WERAM stores experimentally identified and computationally predicted histone regulators as well as site-specific regulator-histone relations in eukaryotes (22). CR Cistrome integrates ChIP-Seq data of CRs and histone modifications to specifically explore their relationships in human and mouse (23). These data resources mainly focus on the collection of basic information and ChIP-Seq data of CRs, and include less knowledge about DNA methylation and chromatin remodeling proteins. Although they are crucial for deciphering the mechanism and function of CRs, a systematic and specialized source for CRs in human cancer is needed urgently.

Over the past decade, large-scale cancer genomic projects, such as The Cancer Genome Atlas (TCGA) (24) and the Cancer Cell Line Encyclopedia (CCLE) (25), have produced multi-dimensional omics data for a number of cancer types. This definitely provides biologists an opportunity to uncover the molecular mechanisms of regulation and interaction in pan-cancer level (26–28). In addition, low-throughput experiments of tumor-associated CRs are rapidly emerging, but hidden in thousands of scattered publications. In this study, we curated, assembled and integrated these data as well as other public resources to build a database CR2Cancer with user-friendly interface for easy data access. Through this database, users can explore the functional, genomic, transcriptomic, proteomic, clinical, biological network and literature-reported information for CRs across multiple cancer types. We believe that CR2Cancer represents a valuable resource that provides unique and useful knowledge of CRs for investigators in cancer epigenetics research community.

**DATABASE OVERVIEW**

CR2Cancer contains over 400 CRs that are annotated with eleven categories of knowledge (Figure 1). (i) The function category provides information on protein domains, function description and classification as well as substrates and products of chemical reactions that CRs catalyze. In addition, this category shows the Gene Ontology terms and pathways including KEGG (29) and Reactome (30). (ii) The mutation category offers mutation landscape of primary tumor tissues and cancer cell lines in single and pan-cancer level. For each primary tumor type, effect of semantic mutation on protein activity is examined and presented in this category. In addition, we also collected various kinds of mutations from COSMIC (31), and curated mutated CRs from publications. (iii) The post-translational modification (PTM) category allows users to scan the positions, modifications and upstream enzymes of PTMs extracted from UniProt, and explore whether PTMs are affected by mutations. (iv) The RNA expression category shows expression pattern in primary tumor tissues, cancer cell lines, and human normal tissues. For primary tumor types with > 10 corresponding normal samples, the results of differential expression analysis are also shown here. Moreover, this category includes dysregulated patterns of CRs in cancer retrieved by literature mining. (v) The copy number category offers correlation analysis between RNA expression and semantic copy number alteration (SCNA) across various cancer types. In addition, users can examine the percentage of samples for three classes of copy number change, i.e., gain, loss and neutral. (vi) Similar to copy number category, the methylation category offers correlation analysis between RNA expression and promoter methylation level as well as differential methylation status between tumor and normal samples. (vii) The proteomics category is based on reverse phase protein array (RPPA) data for primary tumor tissues and antibody staining level for human normal tissues. (viii) The clinical category shows association tests between RNA expression and clinical features, including subtype, overall survival, stage and grade. (ix) The target category presents transcriptional targets of CRs inferred by reverse engineering method and ChIP-Seq data analysis. (x) The drug category offers drug- and gene-centric networks derived from DrugBank (32) to highlight the druggable nature of CRs. (xi) The interaction category includes three parts: protein-protein interaction (PPI) network with co-occurrence and mutual exclusive mutation analysis, miRNA-target regulatory connections with expression correlation, and multiple types of interactions reported in literatures.

**DATABASE CONSTRUCTION**

**Data collection**

A comprehensive list of 429 CRs functioning as DNA modifiers, histone-modifying enzymes or chromatin remodelers, was compiled from three recent papers (33–35). Among them, 167 (38.9%) CRs have been gathered in three cancer gene databases, i.e. 50 in ONGene (36), 72 in TSGene (37) and 100 in NCG (38). We annotated all CRs with full name, aliases, chromosome location and external IDs. In addition, the site-specific substrates and products of CRs were acquired from EpiFactors (39). The somatic muta-
tion, RNA-Seq, DNA copy number, methylation, RPPA and clinical data of primary tumor tissues were downloaded from TCGA; the mutation profiles and gene expression microarray data of cancer cell lines from CCLE; the RNA-Seq and antibody staining data of human normal tissues from The Human Protein Atlas (HPA) (40).

**Literature mining**

A text mining pipeline was developed to extract cancer-associated CRs from public peer review papers. First, we used ‘cancer’, ‘tumor’ and ‘carcinoma’ as keywords to search PubMed database and downloaded all retrieved abstracts (update to 1 August 2017). Second, these abstracts were split into single sentences using GENIA Sentence Splitter for the input of next step (41). Third, we located gene names in the sentences by using BANNER (42) and normalized the identified gene names based on a dictionary comprised of official names and aliases of CRs. Finally, the CR-mentioned sentences were manually curated to extract aberrant molecular changes (e.g. mutation and dysregulation) of CRs and their interactions with other biological molecules in specific cancer type. The cancer names in our results were normalized to the ‘Preferred Name’ registered in NCI Thesaurus. In current release, CR2Cancer recruits around 4700 aberrant molecular changes and 1300 interactions for 358 CRs from 5007 publications.

**Mutation landscape creation**

We collected semantic mutation data of 8,788 primary tumor samples in 26 cancer types and 967 cancer cell lines in 24 cancer types. Lollipop mutation diagrams of single nucleotide variation (SNV, e.g. missense, nonsense and silent) were produced in single and pan-cancer level using Lollipop software (43). The mutational distribution of CRs across cancer types were calculated and visualized in pie charts for primary tumor tissues and cancer cell lines, respectively. Moreover, we also retrieved over 120,000 semantic mutations for 412 CR genes from COSMIC v81 data sets.

**RNA expression data preparation**

For RNA-Seq data of primary cancer types, we normalized raw counts matrix using Voom (44) after filtering out genes that have over one count per million mapped reads in less than 5% samples. Differential expression analysis was carried out using limma package (45) for 15 cancer types with >10 corresponding normal samples (Fold Change > 1.5 and adjusted P-value < 0.05). In addition, the expression level of CRs was visualized across diverse cancer/tissue types, and a total of 1228 images about gene expression were included in the CR2Cancer database.

**Correlation between RNA expression and copy number, methylation or clinical features**

In each cohort, the percentage of patients with copy number loss, neutral and gain is subject to −1/2, 0 and +1/2 copy number threshold based on GISTIC output (46). Pearson correlations were used to assess relationships between RNA expression and copy number value. The methylation level of CRs was represented by the probe in the promotor region with the most negative correlation with expression, and Student’s t-test was used to examine the differential methylation status between tumor and normal samples. In addition, Spearman correlations were used to assess relationships between expression and methylation, stage or grade. The log-rank test was used to examine whether RNA expression significantly correlated with the overall patient survival times. Kruskal-Wallis test was used to determine if there are significant differences among multiple tumor subtypes on expression.

**Identification of transcriptional targets**

We identified transcriptional targets of CRs based on two strategies, i.e., reverse engineering and ChIP-Seq data analysis. Using gene expression profiles in cancer, ARACNe (47) was run with default parameters to prioritize cancer-specific CR-target relations based on mutual information and data processing inequality, which can reduce the number of indirect connections. In addition, there now exist several resources for high quality processed ChIP-Seq data of human cell lines and tissues (23,48,49). CistromeDB is by far the most comprehensive one for ChIP-Seq and chromatin accessibility data analyzed through the standard analysis pipeline (49). We downloaded 1359 ChIP-Seq BED files of 139 CRs from CistromeDB, and used BETA (50) to infer the putative target genes.

**Effect of semantic mutations on protein activity**

The protein activity of CRs was calculated by using VIPER (51) from gene expression profile data and ARACNe-inferred targets. Considering those CRs with at least 20 nonsilent mutated samples in single cancer type (291 pairs, e.g. ARID1A in gastric cancer), we compared the protein activity between samples with and without mutations to evaluate the functional effect of mutations (Wilcoxon test).

**Interaction network**

Around 280,000 protein-protein interactions involved in CRs and their first neighbours were extracted from Pathway Commons v8 (52). In addition, we carried out co-occurrence and mutual exclusive mutation analysis (Fisher’s exact test) for each edge based on cancer genome sequencing data. The miRNA-CR regulatory relations were integrated from several experiment validation or sequence-based prediction databases such as RAID (53), mir2Disease (54), miRTarBase (55), miRDB (56), miRWalk (57) and TargetSCAN (58). We then calculated Pearson correlations between the expression levels of miRNA and CR, and sorted the 12,000 miRNA-CR pairs by the frequency of cancer types in which they are negatively correlated (P-value < 0.05).

**USER INTERFACE**

CR2Cancer provides a friendly interface for users to browse, search and download data. In the ‘Browse’ page,
Figure 2. Browse page for ARID1A in CR2Cancer. A user can access different categorized annotations by clicking the tabs at the top of the page.

the user can browse CRs by six different panels: ‘Cancer Type’, ‘Function’, ‘Mutation Rate’, ‘Differential Expression’, ‘ChIP-Seq Data’ and ‘Targeted Drug’. The ‘Cancer Type’ shows 26 and 24 cancer types from primary tumor tissues and cancer cell lines, respectively. By clicking one cancer type, a summary table of all CRs in this cancer can be presented. The ‘Function’ panel offers detailed function annotations of CRs, and users can click the buttons at the top of the panel to filter them by DNA modification, histone modification and chromatin remodelling. In the ‘Mutation Rate’ page, users can obtain the mutation rate of CRs in 8,788 primary tumor samples and in 967 cancer cell lines. We observed that five CRs, KMT2C, ARID1A, CREBBP, TRRAP and TAF1L, are common in the top 10 mutated CR genes of tumor tissues and cell lines. The ‘Differential Expression’ page summaries differential expression analysis of CRs in 15 primary cancer types with more than ten normal samples. We found that 327 (76.2%) CRs are differentially expressed in at least one cancer type, and most of them show consistent (up- or down-) dysregulated patterns across diverse cancers. Finally, users can browse 139 CRs with ChIP-Seq profiles and 39 CRs regarded as drug targets in ‘ChIP-Seq Data’ and ‘Targeted Drug’ panels, respectively. In above pages, to click on the hyperlinks of gene symbols can pinpoint to the entry page of CR, which includes 11 categories of knowledge in different tabs (Figure 2). For each tab, the summary table contains basic information of CR and relevant contents in this tab, which is linked to responding parts for more details. In the ‘Search’ page, CR2Cancer can be queried in an easy-to-use manner with a number of search tools. Users can retrieve CRs of interest in simple and batch format by gene symbol, name, En-
CRs in a regular basis. Furthermore, genome-wide maps of DNA methylation, histone modifications and chromatin accessibility will be integrated in our database to provide insights of gene regulation through complex processes that involve functional elements, transcription factors and chromatin regulators (62). Finally, we plan to add drug pharmacological data from high throughput screening studies for drug discovery and repurposing.

ACKNOWLEDGEMENTS

The authors are grateful to the editor and the anonymous reviewers for their valuable suggestions and comments, which has led to the improvement of this paper.

FUNDING

Research Grants Council of Hong Kong, China [HKU 17127014 General Research Fund (GRF), HKU T12-710/16R Theme-based Research Scheme (TBR5)]. Funding for open access charge: University of Hong Kong Seed Funding Programme for Basic Research.

Conflict of interest statement. None declared.

REFERENCES


