Bioinformatics, 32(9), 2016, 1433–1435
doi: 10.1093/bioinformatics/btv757
Advance Access Publication Date: 31 December 2015
Applications Note

Databases and ontologies

ChromothripsisDB: a curated database of chromothripsis

Jian Yang, Gaofeng Deng and Haoyang Cai*

Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610064, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on 27 September 2015; revised on 4 December 2015; accepted on 21 December 2015

Abstract

Summary: Chromothripsis is a single catastrophic event that can lead to massive genomic rearrangements confined to one or a few chromosomes. It provides an alternative paradigm in cancer development and changes the conventional view that cancer develops in a stepwise progression. The mechanisms underlying chromothripsis and their specific impact on tumorigenesis are still poorly understood, and further examination of a large number of identified chromothripsis samples is needed. Unfortunately, this data are difficult to access, as they are scattered across multiple publications, come in different formats and descriptions, or are hidden in figures and supplementary materials. To improve access to this data and promote meta-analysis, we developed ChromothripsisDB, a manually curated database containing a unified description of all published chromothripsis cases and relevant genomic aberrations. Currently, 423 chromothripsis samples representing 107 research articles are included in our database. ChromothripsisDB represents an extraordinary resource for mining the existing knowledge of chromothripsis, and will facilitate the identification of mechanisms involved in this phenomenon.

Availability and implementation: ChromothripsisDB is freely available at http://cgma.scu.edu.cn/ChromothripsisDB.

Contact: haoyang.cai@scu.edu.cn

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

1 Introduction

Chromothripsis is a single catastrophic event of massive chromosomal rearrangements that has been described in a wide variety of diseases, especially in cancer (Stephens *et al.*, 2011). In contrast to the widely accepted model of gradual accumulation of genetic alterations during cancer development, chromothripsis is supposed to be a newly discovered paradigm in tumorigenesis (Stephens *et al.*, 2011). In brief, one or several chromosomes are shattered into many different sized fragments through a catastrophic event. Some of the fragments are then randomly stitched back together to generate a derivative chromosome, and the others are lost. This process may generate DNA amplifications, deletions and chromosomal fusion events that underpin chromosome evolution (Liu *et al.*, 2011; Rausch *et al.*, 2012). Chromothripsis has received a growing amount of attention in recent years since it may reveal a new way that cancer genomes can evolve. However, the mechanisms leading to this event are not yet fully characterized. A meta-analysis based on a large number of chromothripsis samples is needed to uncover more about the phenomenon, which would substantially enhance our understanding of tumor initiation and progression (Cai *et al.*, 2014; Kim *et al.*, 2013). Unfortunately, due to the overall low incidence of chromothripsis (occurs in only 2–3% of all cancers) (Stephens *et al.*, 2011), most studies reported a relatively small number of chromothripsis cases. Moreover, the relevant data acquisition is a cumbersome process, since they are scattered in different parts of various publications and are heterogeneous in their contents and formats.

To address this problem, we created ChromothripsisDB, which is a manually curated database that catalogs all published chromothripsis articles and samples. In the current release, ChromothripsisDB includes data and results from 107 publications covering 423 cases. DNA copy number aberrations (CNAs), structural variations (SVs) and clinical data were curated and normalized into standardized formats. To the best of the authors' knowledge, ChromothripsisDB is the first database dedicated to processing and visualizing chromothripsis data. It can facilitate future research to identify key features of this event and elucidate the underlying molecular mechanisms.

2 Methods

Our literature searches up to November 25, 2015, yielded 812 publications, which contained the keyword 'chromothripsis' in the title, abstract or full text. All full papers were screened for eligibility and classified into three categories according to their contents: Review and Opinion, Methodology, and Research. The Review and Opinion category included literature reviews, comments and reports of innovations. The publications categorized as methodology were articles concerning new theories, algorithms or softwares for chromothripsis identification. The research articles containing experimental data were the primary data source for ChromothripsisDB. The following three criteria were applied to ensure selection of eligible samples: (i) the sample had to be explicitly declared as chromothripsis by the authors; (ii) the clinical diagnosis and pathology of the sample must be clearly recognized; (iii) the sample had to be assigned a specific case ID or can be traced back to the original publication. For each case, CNAs, SVs, relevant genes, genome assembly version and clinical information were extracted from the main text, tables, figures and supplementary materials, if available.

3 Results

ChromothripsisDB contains data from all published studies that screening for chromothripsis events (Supplementary Table S1). Currently, it includes 423 cases manually curated from 107 research articles on two species: human and mouse. These data represent 52 disease types, of which ∼90% are cancers. Besides malignant tumors, congenital abnormalities also account for a notable proportion of the database (Supplementary Fig. S1). The CNA profiles and SVs of chromothripsis cases were identified by next generation sequencing, array comparative genomic hybridization or single nucleotide polymorphism arrays.

3.1 Data browsing

The data-browsing interface allows researchers to browse the database by studies. Clicking on the article title leads to a page that includes a table with detailed information of the study and a visualization of chromothripsis cases. For research articles, the table also provides hyperlinks to download the curated data and access raw experimental data (if available). Moreover, we extracted and normalized the definition of chromothripsis used in publications. Since this information is given in various descriptions of different publications, we used three features to represent chromothripsis criteria: (1) the minimum number of close-by breakpoints; (2) the number of copy number states; (3) the random joining of chromosome fragments. The detailed description of data normalization can be found in Supplementary Material. For each chromothripsis case, the published sample ID was used as the case identifier in our database, which allows users to easily trace it back to the original publication. If this information was not available, a unique identifier consisting of the PubMed ID and a case number was assigned. The relevant information of each case was extracted from the publication, including pulverized chromosome IDs, detection technology and platform, and so on. The 'affected genes' field contains a hyperlink to the gene list page, which contains reported fusion genes and rearranged genes. This page allows users to investigate each gene in more detail and has hyperlinks to NCBI Gene and Entrez Gene. Both the pulverized chromosome IDs and the affected genes were recorded for each individual case and consistent with the authors' original interpretations. Furthermore, ChromothripsisDB provides a graphical summary of genomic aberrations in an idiogram of chromosomes (Fig. 1A), and SVs and CNAs are illustrated using a Circos plot (Fig. 1B).

3.2 Database querying

We designed a search page for researchers to easily explore and analyze data of interest. Users can interactively query the database for specific disease types or affected genes. The 'disease search' interface allows users to search by species, detection technology and disease type. All fields support multiple selections, and the field content changes dynamically according to the user's selection. A query from the search page returns a summary table and a visualization of the results. The table provides an overview of the query inputs and outputs, and includes a link to download the tabular data. The figures below the table represent statistical analysis of the results. The bar plot displays the distribution of pulverized chromosomes across the genome (Fig. 1C). The difference of chromosomal pulverization patterns may suggest the possibility of different mechanisms underlying chromothripsis phenomenon in various cancer types. For example, we searched the database for acute lymphoblastic leukemia and chronic lymphocytic leukemia, respectively. We observed a high prevalence of chromosome 21 pulverization in acute lymphoblastic leukemia, while chronic lymphocytic leukemia showed a high frequency of chromosome 9 and 13 shattering (Supplementary Fig. S2). The differential distribution of pulverized chromosomes is an indicator for a disease-related selection of specific genomic rearrangements, and may reveal different cancer associated genes and oncogenesis processes in these two leukemias. More examples can be found in Supplementary Material. The pie chart shows the percentage distribution of samples

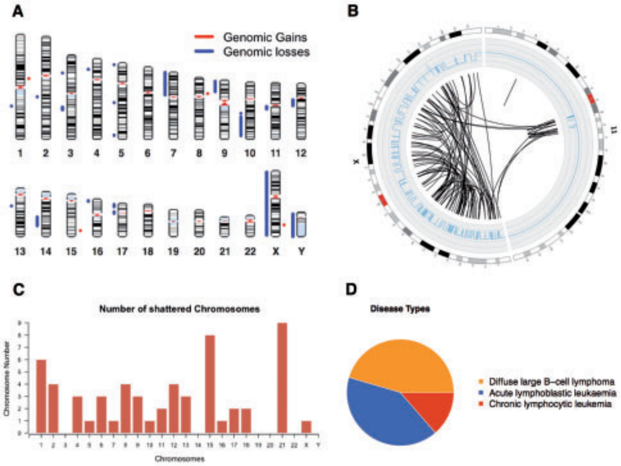

Fig. 1. The data visualization of ChromothripsisDB. (**A**) Summary of DNA copy number alterations in a tumor sample. Each line on either side of chromosomal ideogram represents chromosomal aberration. (**B**) Circos plot representing CNAs, inter and intra-chromosomal rearrangements of a chromothripsis case. (**C**) An example of the distribution of pulverized chromosomes across the genome. (**D**) An example of the percentage distribution of the queried samples

among the queried disease types (Fig. 1D). In addition, the result page contains the summarized information and visualization of each chromothripsis case. In the 'gene search' interface, the users can query the database for common affected genes between two disease types, or for specific genes in particular diseases. Three Boolean search operators (AND, OR, NOT) are available in the drop-down options box. The resulting view will list the qualified genes and related annotations.

4 Conclusion

ChromothripsisDB is the first repository providing convenient public access to chromothripsis data. We curated and integrated hundreds of chromothripsis samples from the published literature into the database. We will continue to update our database every 2 months with newly published data. It represents an invaluable resource for the community and will assist in the meta-analysis of chromothripsis phenomenon.

Funding

National Natural Science Foundation of China (grant number 31571314).

Conflict of Interest: none declared.

References

Cai,H. *et al*. (2014) Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens. *BMC Genomics*, **15**, 82.

Kim,T.M. *et al*. (2013) Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res*., **23**, 217–227.

Liu,P. *et al*. (2011) Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, **146**, 889–903.

Rausch,T. *et al*. (2012) Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, **148**, 59–71.

Stephens,P.J. *et al*. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.