BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1640–1642
doi:10.1093/bioinformatics/btu081

Databases and ontologies

Advance Access publication February 12, 2014

The 3DGD: a database of genome 3D structure

Chao Li[1,2,†], Xiao Dong[1,2,†], Haiwei Fan[3,†], Chuan Wang[3,*], Guohui Ding[1,4,*] and Yixue Li[1,4,*]

[1]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, P. R. China, [2]University of Chinese Academy of Sciences, 19A Yuquan Road, Beijing 100049, [3]National Center for Protein Science, Shanghai 333 Haike Road, Pudong District, Shanghai 201210 and [4]Shanghai Center for Bioinformation Technology, 1278 Keyuan Road, Shanghai 201203, P. R. China

Associate Editor: Michael Brudno

ABSTRACT

Summary: The studies of chromatin 3D structure help us to understand its formation and function. Techniques combining chromosome conformation capture and next generation sequencing can capture chromatin structure information and has been applied to several different species and cell lines. We built 3DGD (3D Genome Database), a database that currently collected Hi-C data on four species, for easy accessing and visualization of chromatin 3D structure data. With the integration of other omics data such as genome-wide protein–DNA-binding data, this data source would be useful for researchers interested in chromatin structure and its biological functions.

Availability and implementation: The 3DGD v1.1, data browser, downloadable files and documentation are available at: http://3dgd.biosino.org/.

Contact: cwang@sibs.ac.cn; gwding@sibs.ac.cn; yxli@sibs.ac.cn

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

Received on September 4, 2013; revised on January 16, 2014; accepted on February 3, 2014

1 INTRODUCTION

The 3D conformation of chromosome in nucleus is relevant with several important biological processes such as gene expression, regulation and replication (Dong *et al.*, 2010; Handoko *et al.*, 2011; Liu *et al.*, 2013). Understanding these relationships would provide a new perspective to the studies of chromatin (Li and Reinberg, 2011). In recent years, the development of chromatin conformation capture techniques, such as Hi-C, allows us to identify the chromatin interactions across the whole genome unbiasedly (Belton *et al.*, 2012; Lieberman-Aiden *et al.*, 2009). The technique has been applied to several species such as human, mouse, *Drosophila* and yeast (Duan *et al.*, 2010; Lieberman-Aiden *et al.*, 2009; Sexton *et al.*, 2012; Zhang *et al.*, 2012). These datasets have been shown useful in studying the property and function of chromosome organization (Dong *et al.*, 2010, 2013).

However, there is still a lack of an integrative data source of these kind of data for easy accessing and visualization. Besides, integrating additional information such as genome annotation and protein binding information could be useful for studying chromatin structure's function (Botta *et al.*, 2010; Dong *et al.*, in submission; Li and Heermann, 2013). Recent studies have shown that some important DNA-binding proteins are extremely relevant to the chromosome organization and its function. For example, CCCTC-binding factor (CTCF) is of importance in the formation of chromatin structure (Botta *et al.*, 2010; Handoko *et al.*, 2011). Combining DNA-binding affinities of these proteins together with Hi-C data may help us to understand the formation and the functions of the chromatin structure.

Here, we present 3D Genome Database (3DGD), a novel database that collected published 3D structure of genome (currently mainly based on Hi-C experiment) of four different species. We organized and displayed the interaction data in an easy way for accessing, while integrating genome annotation data and some important DNA-binding protein information. To our knowledge, 3DGD is the first database collecting and displaying chromatin 3D structure data.

2 METHODS AND USAGE

The original Hi-C datasets are collected from four published papers (Supplementary Table S1) (Duan *et al.*, 2010; Lieberman-Aiden *et al.*, 2009; Sexton *et al.*, 2012; Zhang *et al.*, 2012). Previous studies pointed out the potential bias in the original Hi-C dataset and described four normalization methods to reduce bias (Cournac *et al.*, 2012; Hu *et al.*, 2012, 2013; Imakaev *et al.*, 2012; Yaffe and Tanay, 2011). We applied the four normalization methods to the original datasets of human, mouse and *Drosophila* with default procedures and parameters. Owing to the different experimental protocols that may introduce additional circularization bias, only the SCN method was applied to yeast data. These datasets are converted to a unified format as follows. Chromosomes are segmented into several bins and the bin size is the minimal resolution of the original dataset. Interaction matrixes or other file formats in the original dataset are then unified formatted as the interaction value between the bins. Then, we built a browser to display interaction data between any regions of the genome. Interactions between chromosome regions are computed as the weighted sum of the interaction value (raw or normalized Hi-C observed value) of different bins in that region; the weight is the ratio between the length of the region and the minimal bin size. Thus, the interaction value I of Regions R1 and R2 is defined as follows:

$$I(R_1, R_2) = \sum_i w_i \sum_j w_j I(R_{1,i}, R_{2,j}) \quad (1)$$

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Where i and j are the number of bins in R1 and R2 separately, and w is the ratio between the length of ith bin overlapped with the selected region and the dataset's bin size.

This weighted sum approach assumes that the chromosome interaction distributes uniformly along one bin. We use human GM06990 dataset to test this assumption. The published datasets contain interaction matrix of 100 kb and 1 Mb. We randomly chose 1000 regions of 1 Mb, which consists of 10 100-kb bins, and test whether the interaction value between this region and the other bin is uniformly distributed along the 10 100-kb bins with χ^2 goodness-of-fit test. Only a small percentage (4.7%) of the interactions in one bin is not uniformly distributed ($P < 0.05$), indicating the method works well for most of the occasions. For those non-uniformly distributed interaction pairs (Supplementary Table S3), bias would be introduced if the search region contains segment smaller than that of the minimal bin. For each dataset, we calculated the value generated by weighted sum approach and compared them to the value calculated by sum of 100-kb data. The median difference is 24.2%. To avoid this potential bias, we suggest the search region should be set as the integer times of the bin size of each dataset, which is listed in a table in the Browse page.

Protein binding datasets are then integrated into the browser. The proteins we selected are annotated or reported relevant to the chromosome structure formation or function. Data mainly comprised CHIP-seq data downloaded from GEO or ENCODE project (Supplementary Table S2). Besides the pre-selected proteins, users can also upload their own protein binding data to be visualized in our browser. The binding intensities of the proteins are averaged to the chromosome bins. Thus, we can display the interaction number and the protein binding intensity at the same time in one browser.

Data browsing and downloading can be done via 3DGD Web site. The original and normalized Hi-C datasets can be downloaded from the download page. Browse page is used to visualize Hi-C and other data. One can view the interactions between the two chromosome regions, the query region and the target region. Query and target regions can be input as either a chromosome region or a gene symbol. The query region is treated as a single part, and the interactions between the bins in the target region and the query region itself are displayed. For example, select Human Gm06990 from the sidebar, then select one region as a query region by chromosome position (e.g. chr8:2000000–5000000), then specify the target chromosome region (e.g. chr8:1000000–6000000) and press SEARCH to show the interaction graph. To show protein binding information, select one protein (e.g. CTCF) or upload a protein binding data and press SEARCH, and the protein binding intensity in the target region will be displayed (Fig. 1). All these data in the current selected region will be displayed in a UCSC genome browser by clicking 'Display in UCSC genome browser' button.

Fig. 1. An example of the data search and display. This figure shows the Hi-C interaction value, the genome annotation and the CTCF binding in the specific region. Bar plot shows the Hi-C interaction value between the query region and the bins in the target region. Blue line shows the CTCF binding intensity in bins in target region. Genes in that region are shown below

3 CONCLUSION

The 3DGD collects Hi-C data of four species and integrates them with other omics data such as genome-wide protein binding information. As the studies of the chromatin structure accumulated, more structure information can be added into this data source and more reported relevant omics data can be integrated as well. Researchers from wide research areas can benefit from this data source to study the fundamental role of chromatin structure functions in biological processes.

Funding: State key basic research program (973) (2011CB910204, 2010CB529206, 2010CB912702, 2011CBA00801), Research Program of CAS (KSCX2-EW-R-04, KSCX2-YW-R-190, 2011KIP204), National Natural Science Foundation of China (30900272, 31070752), Chinese Ministry for Science and Technology Grant (2008BAI64B01), Chinese High-Tech R&D Program [(863)-2009AA02Z304] and SA-SIBS Scholarship Programme.

Conflict of Interest: none declared.

REFERENCES

Belton,J.M. *et al.* (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods.*, **58**, 268–276.

Botta,M. *et al.* (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.*, **6**, 426.

Cournac,A. *et al.* (2012) Normalization of a chromosomal contact map. *BMC Genomics.*, **13**, 436.

Dong,X.A. *et al.* (2010) Human transcriptional interactome of chromatin contribute to gene co-expression. *BMC Genomics.*, **11**, 704.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

Handoko,L. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.

Hu,M. *et al.* (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics.*, **28**, 3131–3133.

Hu,M. *et al.* (2013) Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. *Quant. Biol.*, **1**, 156–174.

Imakaev,M. *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization, *Nat. Methods*, **9**, 999–1003.

Li,G. and Reinberg,D. (2011) Chromatin higher-order structures and gene regulation. *Curr. Opin. Genet. Dev.*, **21**, 175–186.

Li,S. and Heermann,D.W. (2013) Transcriptional regulatory network shapes the genome structure of *Saccharomyces cerevisiae*. *Nucleus*, **4**, 216–228.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.*, **326**, 289–293.

Liu,L. *et al.* (2013) DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.*, **4**, 1502.

Sexton,T. *et al.* (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.

Zhang,Y. *et al.* (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.