GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data

Oluwatosin Oluwadare[1], Max Highsmith[1], and Jianlin Cheng[1, 2*]

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

[2]Informatics Institute, University of Missouri, Columbia, MO 65211, USA

* To whom correspondence should be addressed. Tel: +1 5738827306; Fax: +1 5738828318

Email: chengji@missouri.edu

ABSTRACT

Advances in the study of chromosome conformation capture (3C) technologies, such as Hi-C technique - capable of capturing chromosomal interactions in a genome-wide scale - have led to the development of three-dimensional (3D) chromosome and genome structure reconstruction methods from Hi-C data. The 3D genome structure is important because it plays a role in a variety of important biological activities such as DNA replication, gene regulation, genome interaction, and gene expression. In recent years, numerous Hi-C datasets have been generated, and likewise, a number of genome structure construction algorithms have been developed. However, until now, there has been no freely available repository for 3D chromosome structures. In this work, we outline the construction of a novel Genome Structure Database (GSDB) to create a comprehensive repository that contains 3D structures for Hi-C datasets constructed by a variety of 3D structure reconstruction tools. GSDB contains over 50,000 structures constructed by 12 state-of-the-art chromosome and genome structure prediction methods for publicly used Hi-C datasets with varying resolution. The database is useful for the community to study the function of genome from a 3D perspective. GSDB is accessible at http://sysbio.rnet.missouri.edu/3dgenome/GSDB

INTRODUCTION

The three-dimensional (3D) organization of the genome plays a significant role in many diverse biological functions and processes including gene expression [1], regulation [2,3] and transcriptional regulation [4]. Several studies of the architecture of the genome in the cell have linked genome structure to the mechanism of these functions; hence, it is essential to understand the spatial arrangement within the cell nucleus in order to fully elucidate this relation [5-7]. Early studies of the structure of the genome have relied on the use of microscopy techniques such as fluorescence in situ hybridization (FISH), a technique that employs fluorescence probes to detect the presence of a specific chromosome region and the proximity between two regions in a genome sequence [8-10]. Other microscopy methods developed to study the genome organization include stimulated emission depletion (STED) [11], stochastic optical reconstruction microscopy (STORM) [12], and photo-activated localization microscopy (PALM or FPALM) [13,14]. While these techniques have proven very useful in providing insights into the organization of the genome for DNA fragments or chromatin regions, they are limited and unsuitable for an overall view of the genome-wide inter-and intra- chromosomal relationship study of the genome within the cell nucleus [15].

In order to capture these inter- and intra- chromosomal interactions, a variety of next-generation, high-throughput sequencing technologies have emerged including: 3C [16], 4C [17], 5C [18], Hi-C [19], TCC [20] and ChIA-PET [21,22]. Out of all these techniques, the Hi-C technique has seen a particularly high usage because of its ability to comprehensively map the chromatin interactions at a genome wide scale.

A Hi-C experiment results in the generation of an interaction frequency (IF) matrix for chromosomal regions (loci) within a chromosome or between any two chromosomes in a population of cells [19,23-25]. With the advancement of the Hi-C research, sophisticated tools such as GenomeFlow [23], Juicer [26], and HiC-Pro [27] have been developed to generate IF matrices from raw sequence pair reads data [28]. Some methods represent the contact matrix in a sparse 3-column format where columns 1-2 denote the interacting loci and column 3 denotes the number of interactions (or contacts) between the corresponding loci in a Hi-C dataset [24,29,30].

Many methods have been developed for chromosome 3D structure reconstruction from chromosome conformation capture (3C) such as the Hi-C data. Generally, these data-driven methods can be grouped into three classes [31] based on how the IF is used for 3D structure construction: distance-based, contact-based and probability-based. First, distance-based methods implement the 3D structure construction through a two-step process. These methods convert the IF matrix to a distance matrix between loci based on an inverse relation observed from FISH 3D distance data [19]. An optimization function is thereafter used to infer a 3D structure from an initial random structure with the objective of satisfying the distances in the distance matrix as much as possible [24,29,32- 40]. Second, contact-based methods consider each chromosomal contact as a restraint and apply an optimization algorithm to ensure that the number of contacts in the input contact matrix is satisfied in the 3D structure [30,41-43]. Third, probability-based methods define a probability measure over the IF, by constructing the structure inference problem as a maximum likelihood problem and thereafter using a sampling e.g. Markov chain Monte Carlo (MCMC) or optimization algorithm to solve the prediction problem [25,44-46]. Despite the significant progress in the methodological development in 3D chromosome and genome structure modeling and availability of a lot of Hi-C datasets, there is still no public database to store 3D chromosome and genome models for the biological community to use.

Here, we present Genome Structure Database (GSDB), a novel database that contains the chromosome/genome 3D structural models of publicly and commonly used Hi-C datasets reconstructed by twelve state-of-the-art 3D structure reconstruction algorithms at various Hi-C data resolution ranging from 25KB – 10MB. The database is organized such that users can view the structures online and download the 3D structures constructed for each dataset by all the reconstruction methods. Our database is the first of its kind to provide a repository of 3D structures and the evaluation results for 3D structures constructed from many Hi-C datasets by different Hi-C data reconstruction methods all in one place.

MATERIALS AND METHODS

DATASETS

Our Hi-C data is pulled from a variety of sources which we list here. Some datasets were downloaded from the Gene Expression Omnibus (GEO) database, including the Hi-C contact matrices datasets (GEO accession Number: GSE63525) of cell line GM12878 from Rao *et* al. [47], normalized interaction matrices for each of the four cell types - mouse ES cell, mouse cortex, human ES cell (H1), and IMR90 fibroblasts – (GEO accession Number: GSE35156) [48,49], and the Hi-C contact matrices datasets (GEO Accession Number: GSE18199) of karyotypically normal human lymphoblastic cell line (GM06990, K562)[19]. All other Hi-C datasets were obtained from the ENCODE project repository [50], and the GEO accession Number and the ENCODE ID for each dataset are available on the GSDB website. Currently, GSDB contains over 50,000 structural models of various resolutions reconstructed from 32 unique Hi-C datasets by 12 state-of-the-art 3D genome/chromosome modeling methods. More Hi-C datasets will be used to build 3D models as they are available.

NORMALIZATION

Hi-C data normalization is an important process in 3D structure reconstruction from Hi-C data, because the raw contact count matrix obtained from 3C experiments may contain numerous systematic biases, such as GC content, length of restriction fragments, and other technical biases that could influence the 3D structure reconstruction [51-55]. Consequently, all the contact matrices were normalized prior to applying the 3D structure reconstruction algorithms. GM12878 cell line datasets were normalized using the Knight–Ruiz normalization (KR) method [53, 47], and the normalized interaction matrices downloaded from Dixon et al. [48] were normalized using Yaffe and Tanay normalization method [55]. The Vanilla Coverage (VC) technique [47] was used as the default technique to normalize all the other Hi-C datasets.

DATABASE IMPLEMENTATION

The GSDB website interface was implemented using HTML, PHP and JavaScript, and the database was implemented in MySQL (https://www.mysql.com/). The online 3D structure visualization was done through 3Dmol viewer, a molecular visualization JavaScript library [56].

3D MODELING ALGORITHMS INCLUDED

We used twelve existing algorithms for the 3D structure construction. We selected a mixture of distance-based, contact-based, and probability-based algorithms [31]. We first describe the distance-based algorithms. LorDG [24] uses a nonlinear Lorentzian function as the objective function with the main objective of maximizing the satisfaction of realistic restraints rather than outliers. LorDG uses a gradient ascent algorithm to optimize the objective function. 3DMax [29] used a maximum likelihood approach to infer the 3D structures of a chromosome from Hi-C data. A log-likelihood was defined over the objective function which was maximized through a stochastic gradient ascent algorithm with per-parameter learning rate [57]. Chromosome3D [32] uses distance geometry simulated annealing (*DGSA*) to construct chromosome 3D structure by translating the distance to positions of the points representing loci. Chromosome3D adopts the Crystallography & NMR System (CNS) suite [58] which has been rigorously tested for protein structure construction for the 3D genome structure prediction from Hi-C data. HSA [34] introduced an algorithm capable of taking multiple contact matrices as input to improve performance. HSA can generate same structure irrespective of the restriction enzyme used in the Hi-C experiment. miniMDS [37] proposed an algorithm to model Hi-C data by partitioning the contact matrix first into segments and building the 3D structure bottom-up from each segment which are eventually aggregated to form a final 3D structure. ChromSDE [38] (Chromosome Semi-Definite Embedding) framed the 3D structure reconstruction problem as a semi-definite programming problem. Shrec3D [39] formulated the 3D structure reconstruction problem *as* a graph problem and attempts to find the shortest-path distance between two nodes on the graph. The length of a link is determined as the inverse contact frequency between its end nodes. Each fragment is regarded as the nodes connected by a link. The represented 3D structure for a Hi-C data is one in which distance between the nodes is the shortest. InfoMod3DGen [40] converts the IF to a distance matrix and used an expectation-maximization (EM) based algorithm to infer the 3D structure.

In the contact-based category, we used MOGEN [30] and GEM [42] for the 3D structure reconstruction. MOGEN [30] does not require the conversion of IF to distances and is suitable for large-scale genome structure modeling. GEM [42] considers both Hi-C data and conformational energy derived from knowledge about biophysical models for 3D structure modeling. It used a manifold learning framework, which is aimed at extracting information embedded within a high-dimensional space, in this case the Hi-C data.

Lastly, in the probability-based category, Pastis [25] defined a probabilistic model of IF and casted the 3D inference problem as a maximum likelihood problem. It defined a Poisson model to fit contact data and used an optimization algorithm to solve it. SIMDA3D [46] used a Bayesian approach to infer 3D structures of chromosomes from single cell Hi-C data.

COMPUTATIONAL MODEL RECONSTRUTION

The GSDB chromosome structure generation was done on three server machines: a x86_64 bit Redhat-Linux server consisting of multi-core Intel(R) Xeon(R) CPU E7-L8867 @ 2.13GHz with 120 GB RAM, x86_64 bit Redhat-Linux server consisting of multi-core Intel(R) Xeon(R) CPU E5649 @ 2.53GHz with 11GB RAM, x86_64 bit Redhat-Linux server consisting of multi-core AMD Opteron(tm) Processor 4284 @ 3.0GHz with 62GB RAM, and a high-performance computing cluster (Lewis) with Linux. Using a high-performance computing (HPC) cluster machine, we allocated 10 cores, 80G of memory, with a time limit of 2 days for each chromosome structure reconstruction task per algorithm. Structures not constructed within 48 hours were terminated.

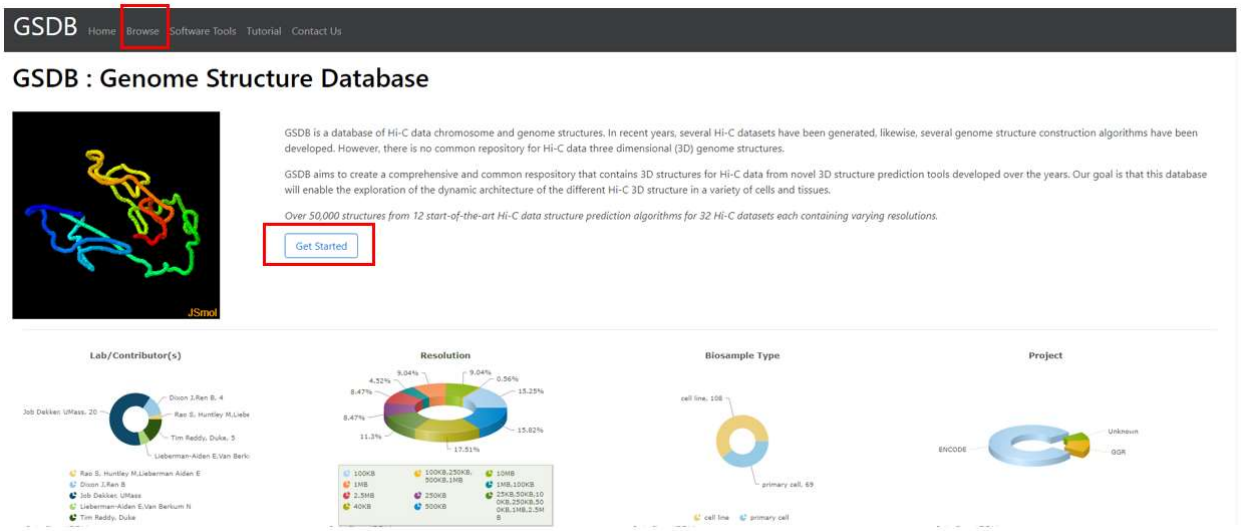

Figure 1: Highlights the two ways to access the database from the homepage. Clicking on the “Browse” menu in the Navigation tab or on the “Get started” button on the home page will load the Database search window.

DATABASE CONTENT AND USAGE

All the 3D structures in the GSDB have been pre-generated, so that the 3D structure visualization is faster and can be easily downloaded. The steps to navigating the database have been separated into five sections as follows:

i. Browse the database (Figure 1) — Click on “Browse” menu in the navigation bar to load the full list of the Hi-C datasets. Alternatively, users can click on the “Get Started” button on the homepage.

ii. Search the database — The GSDB provides two ways to search for a Hi-C data and its corresponding 3D models:

   a. GSDB provides a summary of the information provided in the database through a Summary Pane. By clicking on a property/item in the summary, the user can search the database for all the Hi-C data containing this property and their corresponding 3D structural models. (Figure 2)

b. Users can search the database by typing the keywords about the filename, title of Hi-C data, Hi-C data resolution, project that Hi-C data was generated from (e.g. ENCODE), project ID, and the GEO accession No in the “Search Pane” (Figure 2).

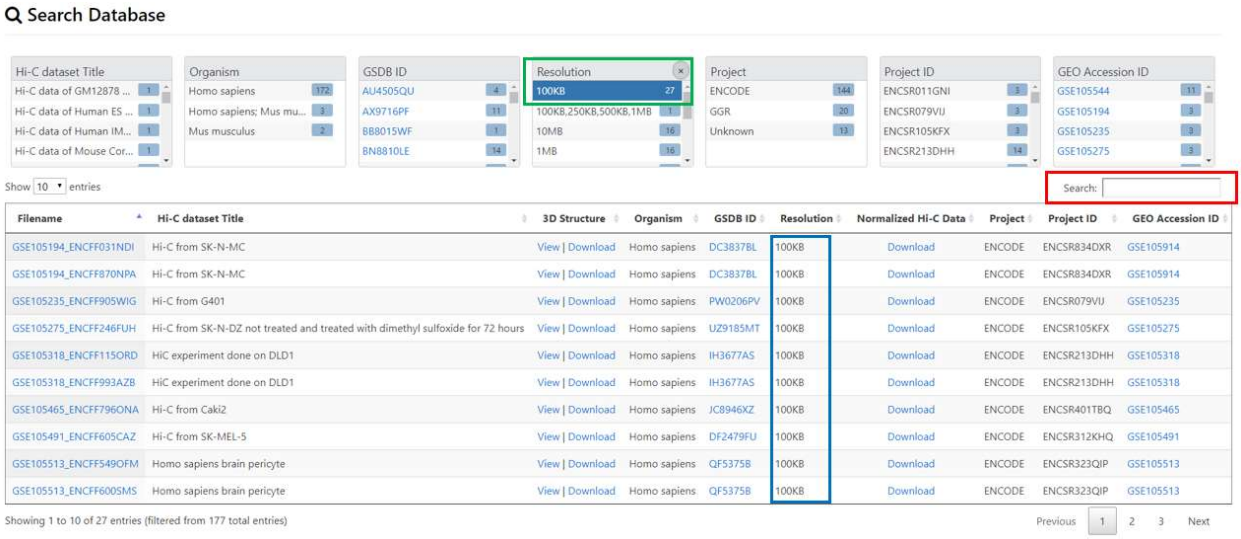

Figure 2: Data search and display. An example of data search using the two approaches for searching. First, search by clicking on an item on the “Summary Pane” highlighted in green. The figure shows when the user clicks on Resolution 100kb, all the datasets with 100kb resolutions are listed. Second, the user can search by typing the key word in the “Search Pane” highlighted in red.

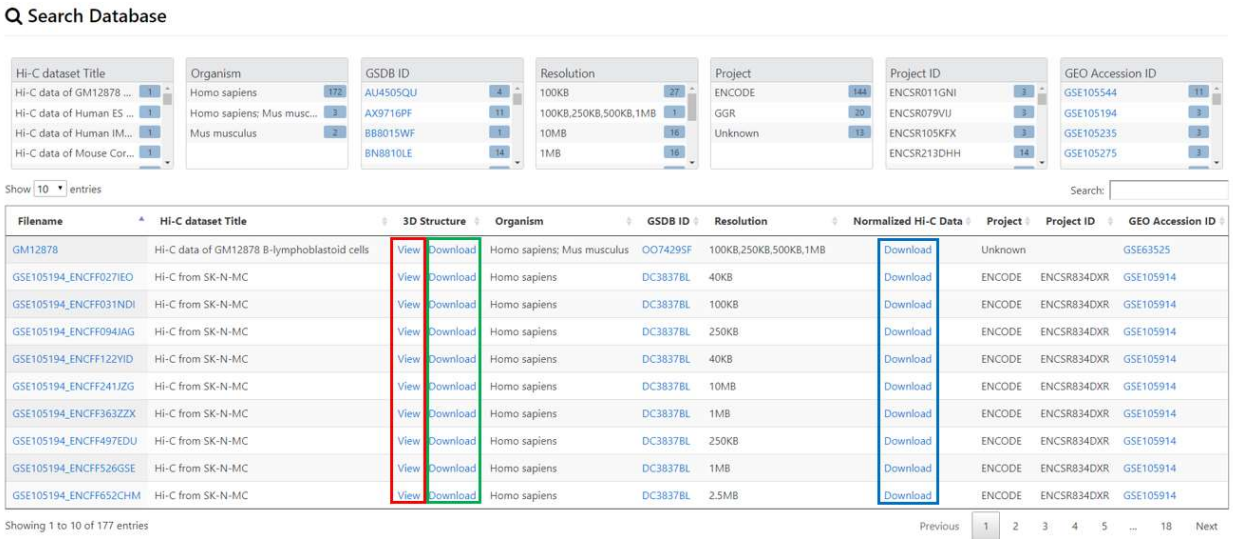

Figure 3: Displaying the database search window. In the “3D Structure” column, highlighted in red is the “View” link to display the 3D structure for a Hi-C data. Highlighted in green is the “Download” link to download the 3D structures constructed by the different algorithms for the Hi-C data. Pressing on the “Download” link will download the 3D structures for all the algorithms for a Hi-C data. In the “Normalized Hi-C Data” column, the

"Download" link is highlighted in blue. Pressing on the "Download" link will download the Normalized Hi-C data used for 3D structure construction.

iii. 3D structure visualization and download — To view the details and structures for a Hi-C data, click on the "View" link in the "3D Structure Column" (Figure 3). The data information and visualization tab will be displayed (Figure 4). To show the 3D structure, select the algorithm, dataset, chromosome, and press "Display this Structure" button. The structure will be displayed on the viewer. The modeling parameters and the reconstruction quality (e.g. the Spearman's correlation between reconstructed distances and expected distances) are reported in the box under the viewer. To compare two structures at the same time, press the "Display Multiple Structures" button. Two structures will be displayed side by side with two distinct options for selecting each visualization's 3D structure algorithm and dataset (Figure 5). To view a heatmap of the 2D contact matrix used to reconstruct the 3D structure, click the "View Contact Heatmap" button. The heat map can be configured with a variety of helper visualization functions as well as color settings to customize visualization (Figure 6). Users can download the 3D structures by clicking on the "Download" link in the "3D Structure Column" (Figure 3). The normalized Hi-C data used for the 3D structure generation for all the algorithms can also be downloaded by clicking on the "Download" link in the "Normalized Hi-C Data" column (Figure 3).

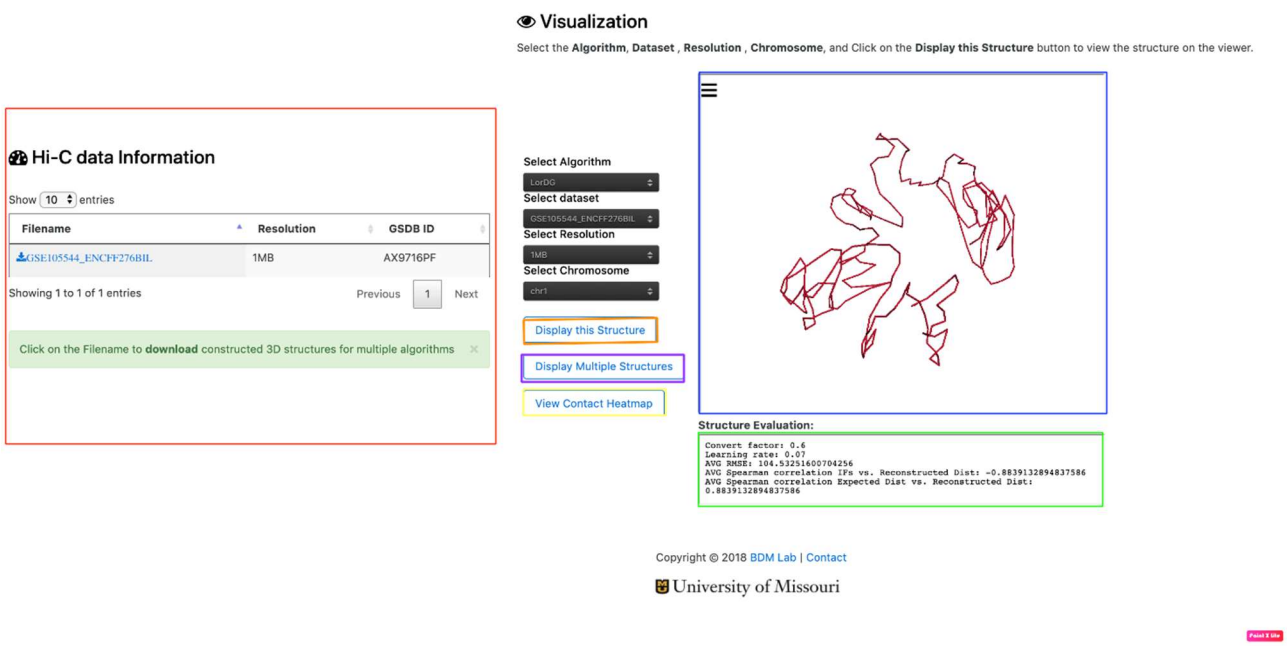

Figure 4: Data visualization. The figure shows the output displayed when a user clicks on the "View link" for the GM12878 dataset. The red highlighted section shows the information about the Resolution(s) available for the Hi-C data. The blue highlighted section displays the structure available for the Hi-C data. The green highlighted section shows the evaluation result available for the Hi-C data. It displays the Spearman Correlation between the output structure and the input Hi-C data, and other evaluation result obtained. To evaluate each 3D structure, we compute the distance Spearman's correlation coefficient (dSCC) between reconstructed distances and distances obtained from the Hi-C datasets. The value of dSCC is in the range of -1 to +1, where a higher value is better. For distance-based methods, we report the conversion factor (α) used for the IF to distance conversion. For LorDG and 3DMax, which use gradient ascent optimization algorithm, we report the learning rate used for the optimization process. The parameters used by each method to generate 3D structures are available on GSDB GitHub page.

Figure 5: Multiple structure visualization. The figures shows the output displayed when a user clicks the "Display Multiple Structures" button. The multiple structure view permits the comparison of structures using different 3D structure algorithms or different Hi-C contact matrices.

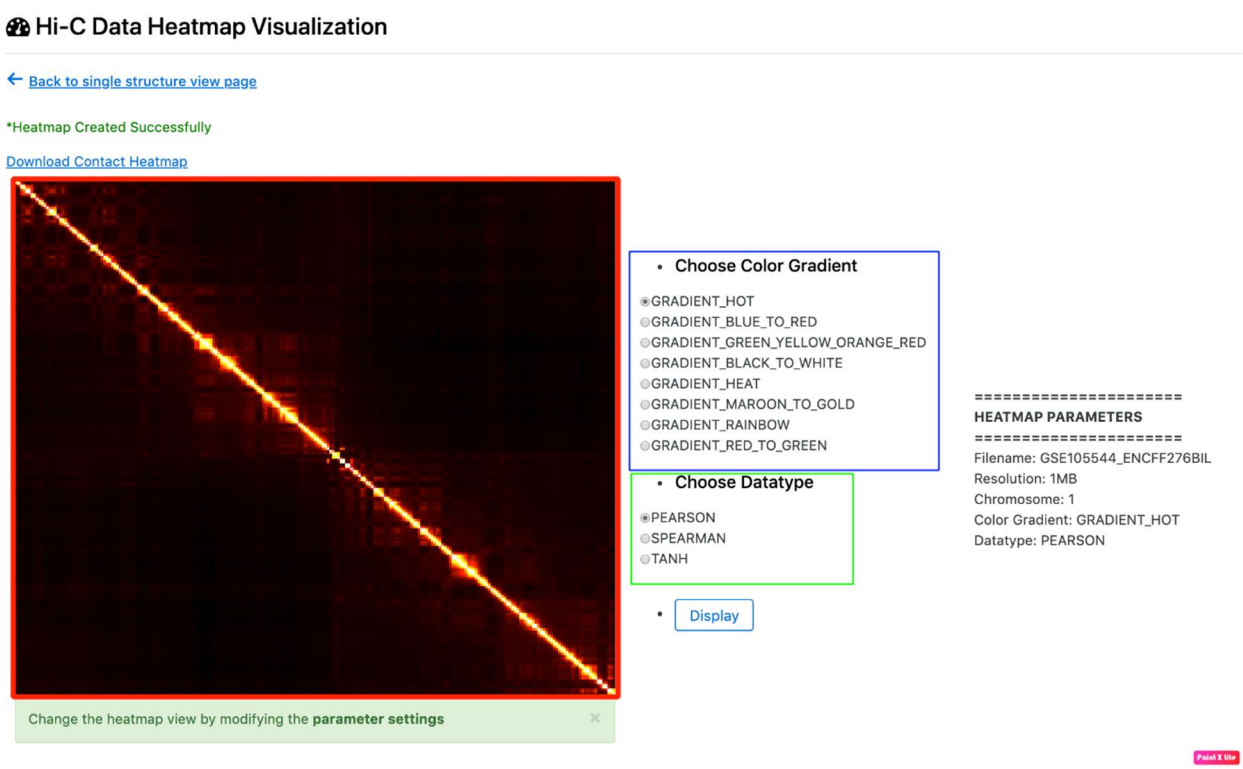

Figure 6: Heat map visualization. The figure shows the output displayed when a user clicks the yellow outlined "View Contact Heatmap" button shown in Figure 4. The figure highlighted in red indicates heat map visualization of

the selected 2D chromosomal contact map. The radio buttons outlined in blue display options for the heat map color. The radio buttons outlined in green indicate functions that may be applied to the raw contact matrix prior to heat map construction so as to improve visualization.

iv. Evaluation of Structure -- The GSDB contains an evaluation module which permits users to evaluate their own 3D models by comparing model distances to the expected distances of an IF matrix or another 3D model (Figure 7). Upon uploading two PDB files or a PDB file and an IF matrix file and clicking on "Compare" button, users are provided with a collection of evaluation scores including: Spearman Correlation, Pearson Correlation and Root Mean Squared Distance (RMSD).

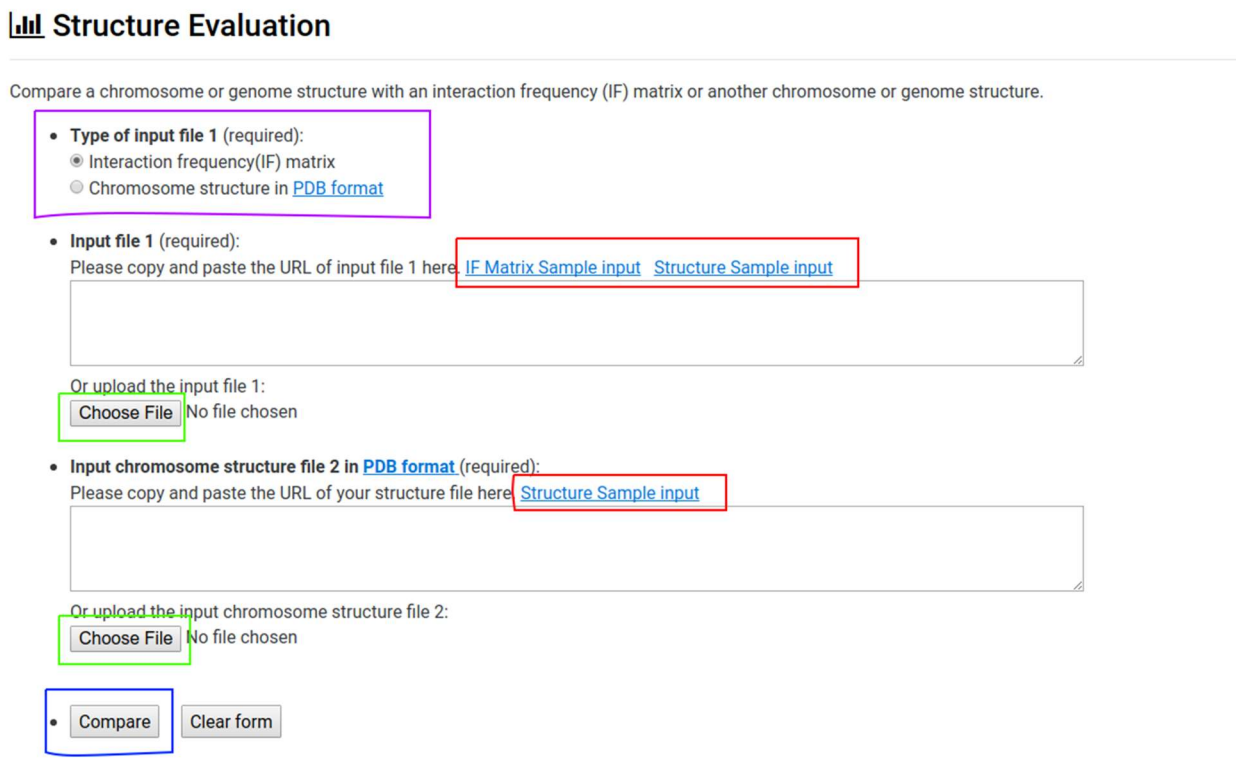

Figure 7: Evaluation. The figure shows the window displayed if a user selects the "Evaluation" tab. The purple box displays the radio buttons which determine whether a comparison will involve 2 structures stored in the Protein Data Bank (PDB) format or a structure in the PDB format and an IF matrix. The green boxes indicate buttons for selecting the files to be compared. The red box denotes links to sample data for testing comparison. The purple box indicates the evaluation button, which will submit the comparison job.

DISCUSSION AND FUTURE DEVELOPMENT

The GSDB contains 3D structures generated from different Hi-C structure reconstruction algorithms for Hi-C data collected from multiple sources. To the best of our knowledge, it is the first repository for 3D structures generated from multiple Hi-C reconstruction algorithms. Currently, our database contains over 50,000 structures reconstructed for 32 Hi-C datasets by 12 modeling algorithms. The normalized Hi-C dataset used and 3D structures generated from all the algorithms are available to be downloaded. This database will enable the fast and easy exploration of the dynamic

architecture of the different Hi-C 3D structure in a variety of cells to improve our understanding of the structural organization of various organisms' chromosome and genome 3D structures. In addition, we envision that it will be helpful to researchers and scientist to keep track of the performance of the existing approaches for 3D structure construction, and also lead to the development of novel methods that outperform existing approaches. Future directions of the GSDB will include the integration of more algorithms and latest Hi-C datasets generated as the research in 3D structure construction expands.

AVAILABILITY

GSDB database is freely available at the URL http://sysbio.rnet.missouri.edu/3dgenome/GSDB. Scripts and the parameters used for the 3D structure generation for each algorithm are available at https://github.com/BDM-Lab/GSDB

ACKNOWLEDGEMENTS

The computation for this work was performed on the high-performance computing infrastructure provided by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia.

FUNDING

This work was supported by the National Science Foundation (NSF) CAREER award (grant no: DBI1149224) to JC.

CONFLICT OF INTEREST

None declared

REFERENCES

1. de Laat, W., and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. *Chromosome Research*, *11*(5), 447-459.
2. Dekker, J. (2008). Gene regulation in the third dimension. *Science*, *319*(5871), 1793-1794.
3. Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, *14*(6), 390.
4. Miele, A., and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. *Molecular biosystems*, *4*(11), 1046-1057.
5. de Wit, E., and De Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes and development*, *26*(1), 11-24.
6. Zou, C., Zhang, Y., and Ouyang, Z. (2016). HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome biology*, *17*(1), 40.
7. Park, J., and Lin, S. (2016). Impact of data resolution on three-dimensional structure inference methods. *BMC bioinformatics*, *17*(1), 70.
8. Amann, R., and Fuchs, B. M. (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology*, *6*(5), 339.
9. Langer-Safer, P. R., Levine, M., and Ward, D. C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences*, *79*(14), 4381-4385.

10. Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, *2*(4), 292.

11. Westphal, V., Rizzoli, S. O., Lauterbach, M. A., Kamin, D., Jahn, R., and Hell, S. W. (2008). Video-rate far-field optical nanoscopy dissects synaptic vesicle movement. *Science*, *320*(5873), 246-249.

12. Rust, M. J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods*, *3*(10), 793.

13. Betzig E, Patterson GH, Sougrat R, Lindwasser OW, Olenych S, Bonifacino JS, Davidson MW, Lippincott-Schwartz J and Hess HF. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, *313*(5793), 1642-1645.

14. Huang, B., Babcock, H., and Zhuang, X. (2010). Breaking the diffraction barrier: super-resolution imaging of cells. *Cell*, *143*(7), 1047-1058.

15. Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, Dostie J and Bickmore WA. (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes and development*, *28*(24), 2778-2791.

16. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, *295*(5558), 1306-1311.

17. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, De Wit E, Van Steensel B and De Laat W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics*, *38*(11), 1348.

18. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C and Green R.D. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, *16*(10), 1299-1309.

19. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. and Sandstrom, R., (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, *326*(5950), 289-293.

20. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, *30*(1), 90.

21. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. and Chew, E.G. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, *462*(7269), 58.

22. Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y.B., Ooi, H.S., Tennakoon, C. and Wei, C.L (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, *11*(2), R22.

23. Trieu, T., Oluwadare, O., Wopata, J., and Cheng, J. (2018). GenomeFlow: a comprehensive graphical tool for modeling and analyzing 3D genome structure. *Bioinformatics*.

24. Trieu, T., and Cheng, J. (2016). 3D genome structure modeling by Lorentzian objective function. *Nucleic acids research*, *45*(3), 1049-1058.

25. Varoquaux, N., Ay, F., Noble, W. S., and Vert, J. P. (2014). A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, *30*(12), i26-i33

26. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., and Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, *3*(1), 95-98.

27. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J. and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology*, *16*(1), 259.

28. Ay, F., and Noble, W. S. (2015). Analysis methods for studying the 3D architecture of the genome. *Genome biology*, *16*(1), 183.

29. Oluwadare, O., Zhang, Y., and Cheng, J. (2018). A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC genomics*, *19*(1), 161.

30. Trieu, T., and Cheng, J. (2015). MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics*, *32*(9), 1286-1292.

31. Oluwadare O., Highsmith M., Cheng J. An overview of methods for reconstructing 3D chromosome and genome structures from Hi-C data. *Biological Procedures Online*. Accepted, 2019.

32. Adhikari, B., Trieu, T., and Cheng, J. (2016). Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC genomics*, *17*(1), 886.

33. Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M., and Dostie, J. (2009). Chromatin conformation signatures of cellular differentiation. *Genome biology*, *10*(4), R37.

34. Zou, C., Zhang, Y., and Ouyang, Z. (2016). HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome biology*, *17*(1), 40.

35. Hua, K. J., and Ma, B. G. (2018). EVR: Reconstruction of Bacterial Chromosome 3D Structure Using Error-Vector Resultant Algorithm. *bioRxiv*, 401513.

36. Szalaj, P., Michalski, P.J., Wróblewski, P., Tang, Z., Kadlof, M., Mazzocco, G., Ruan, Y. and Plewczynski, D. (2016). 3D-GNOME: an integrated web service for structural modeling of the 3D genome. *Nucleic acids research*, *44*(W1), W288-W293.

37. Rieber, L., and Mahony, S. (2017). miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics*, *33*(14), i261-i266.

38. Zhang, Z., Li, G., Toh, K. C., and Sung, W. K. (2013, April). Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. In *Annual international conference on research in computational molecular biology* (pp. 317-332). Springer, Berlin, Heidelberg.

39. Lesne, A., Riposo, J., Roger, P., Cournac, A., and Mozziconacci, J. (2014). 3D genome reconstruction from chromosomal contacts. *Nature methods*, *11*(11), 1141.

40. Wang, S., Xu, J., and Zeng, J. (2015). Inferential modeling of 3D chromatin structure. *Nucleic acids research*, *43*(8), e54-e54.

41. Nowotny, J., Ahmed, S., Xu, L., Oluwadare, O., Chen, H., Hensley, N., Trieu, T., Cao, R. and Cheng, J. (2015). Iterative reconstruction of three-dimensional models of human chromosomes from chromosomal contact data. *BMC bioinformatics*, *16*(1), 338.

42. Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., ... and Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic acids research*, *46*(8), e50-e50.

43. Paulsen, J., Sekelja, M., Oldenburg, A.R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sørensen, A.L., Vigouroux, C., Buendia, B. and Collas, P., . (2017). Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome biology*, *18*(1), 21.

44. Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, and Liu JS. (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, *9*(1), e1002893.

45. Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X.J., Le Gros, M.A. and Larabell, C.A (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, *113*(12), E1663-E1672.

46. Rosenthal, M., Bryner, D., Huffer, F., Evans, S., Srivastava, A., and Neretti, N. (2018). Bayesian Estimation of 3D Chromosomal Structure from Single Cell Hi-C Data. *BioRxiv*, 316265.

47. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665-1680.

48. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376.[

49. GSE35156, Normalized Hi-C data. http://chromosome.sdsc.edu/mouse/hi-c/download.html. Accessed 10 Apr 2019.

50. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57.

51. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, *9*(10), 999.

52. Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, *28*(23), 3131-3133.

53. Knight, P. A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, *33*(3), 1029-1047.

54. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., and Mozziconacci, J. (2012). Normalization of a chromosomal contact map. *BMC genomics*, *13*(1), 436.

55. Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, *43*(11), 1059.

56. Rego, N., and Koes, D. (2014). 3Dmol. js: molecular visualization with WebGL. *Bioinformatics*, *31*(8), 1322-1324.

57. Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121-2159.

58. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ. (1998). Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, *54*(5), 905-921.