*W288–W293* *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 16 May 2016*
*doi: 10.1093/nar/gkw437*

# 3D-GNOME: an integrated web service for structural modeling of the 3D genome

**Przemyslaw Szalaj[1,2,3,*,†], Paul J. Michalski[4,*,†], Przemysław Wróblewski[1], Zhonghui Tang[4], Michal Kadlof[1], Giovanni Mazzocco[1], Yijun Ruan[4,5] and Dariusz Plewczynski[1,2,6,*]**

[1]Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland, [2]Center for Bioinformatics and Data Analysis, Medical University of Bialystok, 15-089 Bialystok, Poland, [3]I-BioStat, Hasselt University, 3500 Hasselt, Belgium, [4]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA, [5]Department of Genetics and Genome Sciences, UConn Health, Farmington, CT 06030-6403, USA and [6]Faculty of Pharmacy, Medical University of Warsaw, 02-097 Warsaw, Poland

Received March 02, 2016; Revised May 06, 2016; Accepted May 07, 2016

## ABSTRACT

**Recent advances in high-throughput chromosome conformation capture (3C) technology, such as Hi-C and ChIA-PET, have demonstrated the importance of 3D genome organization in development, cell differentiation and transcriptional regulation. There is now a widespread need for computational tools to generate and analyze 3D structural models from 3C data. Here we introduce our 3D GeNOme Modeling Engine (3D-GNOME), a web service which generates 3D structures from 3C data and provides tools to visually inspect and annotate the resulting structures, in addition to a variety of statistical plots and heatmaps which characterize the selected genomic region. Users submit a bedpe (paired-end BED format) file containing the locations and strengths of long range contact points, and 3D-GNOME simulates the structure and provides a convenient user interface for further analysis. Alternatively, a user may generate structures using published ChIA-PET data for the GM12878 cell line by simply specifying a genomic region of interest. 3D-GNOME is freely available at http://3dgnome.cent.uw.edu.pl/.**

## INTRODUCTION

ChIA-PET (1), Hi-C (2) and related technologies have revealed that the mammalian genome has multiple levels of organization, from large-scale chromosome territories, mega-base sized topologically associated domains (TADs) (3,4) and chromosome contact domains (CCDs) (5), and down to specific CTCF-mediated looping interactions (5,6). Structural chromosome models are a key computational tool for analysis of such data (7,8), as they provide a representation of the data which can be more revealing than 1D looping depictions or 2D heatmaps. In particular, overlaying the 3D structure with additional genomic annotation data, such as histone methylation marks, can reveal spatial clustering of epigenetic factors which is not readily apparent in other representations (9).

Recently, we have shown that a single ChIA-PET experiment reveals information at all relevant genomic resolutions (5). Previous ChIA-PET experiments focused on high-frequency, high-confidence interactions representing true long range interactions, and discarded singletons, which are low frequency, low confidence interactions which in fact constitute the majority of ChIA-PET reads (10–12). However, we showed that the singleton data provides the same information as Hi-C data, namely, low-resolution (∼1 Mb) information about large-scale topological domains. In light of this observation, we developed a structural modeling algorithm which leverages the multiscale nature of ChIA-PET data to produce 3D chromosome models at multiple resolutions (5). Although it was designed for ChIA-PET data, the algorithm works equally well with Hi-C and other genome-wide 3C-like data, provided some distinction is made between weak and strong interactions.

The number of tools available for 3D chromatin modeling is growing rapidly (for a recent review, see (13)), and includes a mix of private (14–16) and publicly available (7,9,17–22) software. However, most of these tools require additional dependencies and/or some computational expertise. For example, ChromSDE (17) requires Matlab; ShRec3D (18) only runs on Linux; TADbit (19) requires the Integrative Modeling Platform (23); BACH (9) requires R, the GNU scientific library, and must be compiled from source; PASTIS (20) requires Python and includes files which must be compiled from source; and MCMC5C (7)

*To whom correspondence should be addressed. Tel: +48 225543654; Fax: +48 225540801; Email: d.plewczynski@cent.uw.edu.pl
Correspondence may also be addressed to Przemysław Szałaj. Email: przemek.szalaj@uhasselt.be
Correspondence may also be addressed to Paul J. Michalski. Email: Paul.Michalski@jax.org
†These authors contributed equally to this paper as first authors.

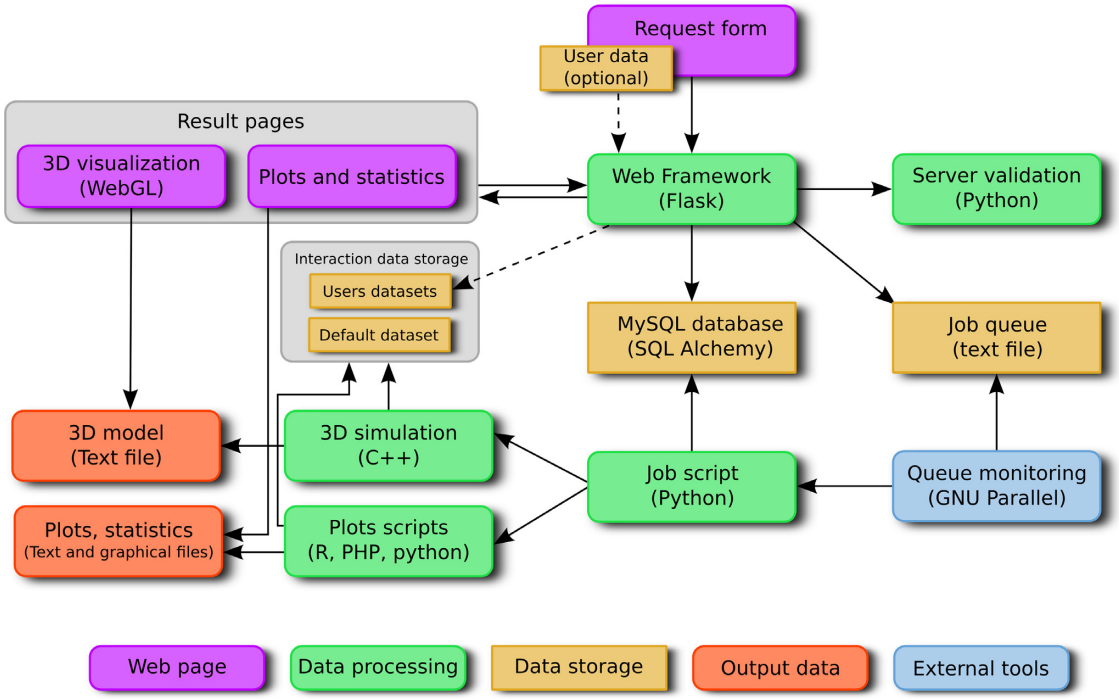



**Figure 1.** Webserver architecture. The central role is played by a Flask-based server which accepts the requests, stores them in the database and adds them to the job queue. GNU Parallel monitors the queue and runs the jobs as soon as there are computational resources available. Each job consists of a number of external scripts executed sequentially.

and MOGEN (22) require Java and must be run from the command line. AutoChrom3D (21) is available as a convenient web application and requires no technical expertise from the user, but it imposes size limits on the reconstructed region dependent on the selected resolution, and its visualization requires Java, which is no longer supported by Google Chrome and generally poses a security risk when run in the browser.

Here we introduce a new, freely available web tool, the 3D GeNOme Modeling Engine (3D-GNOME), which allows a user without any programing experience to generate 3D structures from 3C data with minimal effort, and simply requires any modern web browser to access and use. The simulation program is based on our algorithms published in (5). 3D-GNOME provides a web-based, interactive 3D viewer to visualize and analyze the resulting 3D structure, and includes options for the user to upload genomic annotation data to overlay on the structure. In addition to the 3D structure, 3D-GNOME provides a variety of other analysis tools, including 1D arc representations and 2D heatmap representations of the data.

## IMPLEMENTATION

### Web server

A schematic of the web server architecture is shown in Figure 1. The web server is written in Python using the Flask framework (http://flask.pocoo.org/). Job requests submitted by the user are subject to both the client and server-side validation. Upon validation, the request is saved to the MySQL database and a request with an id of the corresponding database record is added to the job queue. The job queue is simply a text file monitored by GNU Parallel (http://www.gnu.org/software/parallel/), which starts new jobs while managing the available resources (number of threads used, available memory, etc.). This allows us to easily adjust how many simultaneous jobs are run and to distribute processes to several machines. Technically, each job is a Python script containing all the processing steps - parsing the input, running external scripts (written in Python, PHP and R) to calculate statistics and generate plots, and, finally, running the 3D simulations. Structures are viewed using an interactive 3D viewer, which is implemented in WebGL (https://www.khronos.org/webgl/) using the Three.js JavaScript helper library (http://threejs.org/) and the dat.gui library (https://github.com/dataarts/dat.gui) for the user interface.

### imulation

The simulation framework is written in C++. A complete description of the modeling approach is given in (5), and is also available in the technical documentation on the web server. Briefly, we use a multiscale, top-to-bottom modeling approach in which different scales correspond to different resolutions. At each level the chromatin is represented as a beads-and-springs polymer, with beads representing different genomic regions. The assignment of genomic regions to beads is data driven and reflects the underlying biological features that can be identified using interaction clusters—CCDs and interaction anchors—as shown in Figure 2A. We first model the general, low resolution (1–2 Mb) structure using singleton data, and then refine this structure

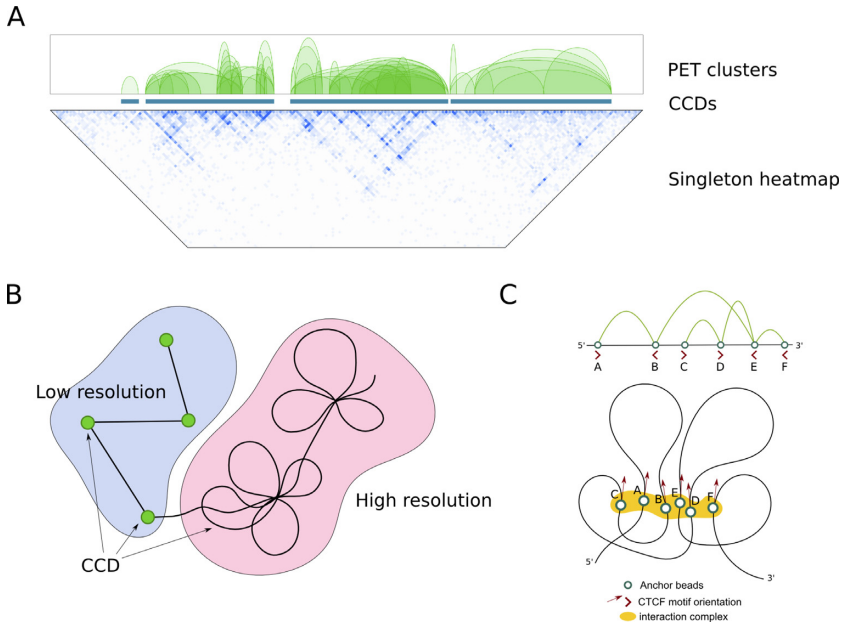



**Figure 2.** Presentation of basic modelling principles used in 3D-GNOME. (**A**) CCDs (marked with blue bars) can be clearly distinguished in both the PET clusters and singleton heatmaps. (**B**) Schematic representation of low (megabase size; left) and high (1–10 kb size; right) resolution structure levels. On the low resolution level each CCD is represented with a single bead. On the high resolution level the interior of CCDs is modeled as an interaction complex and the chromatin loops extending outwards. (**C**) An example of a PET clusters interactions pattern in a single CCD with anchors and CTCF motifs orientations marked (top) and schematic representation of the corresponding structure (bottom).

using PET interactions to achieve a high resolution (1–10 kb) structure (Figure 2B).

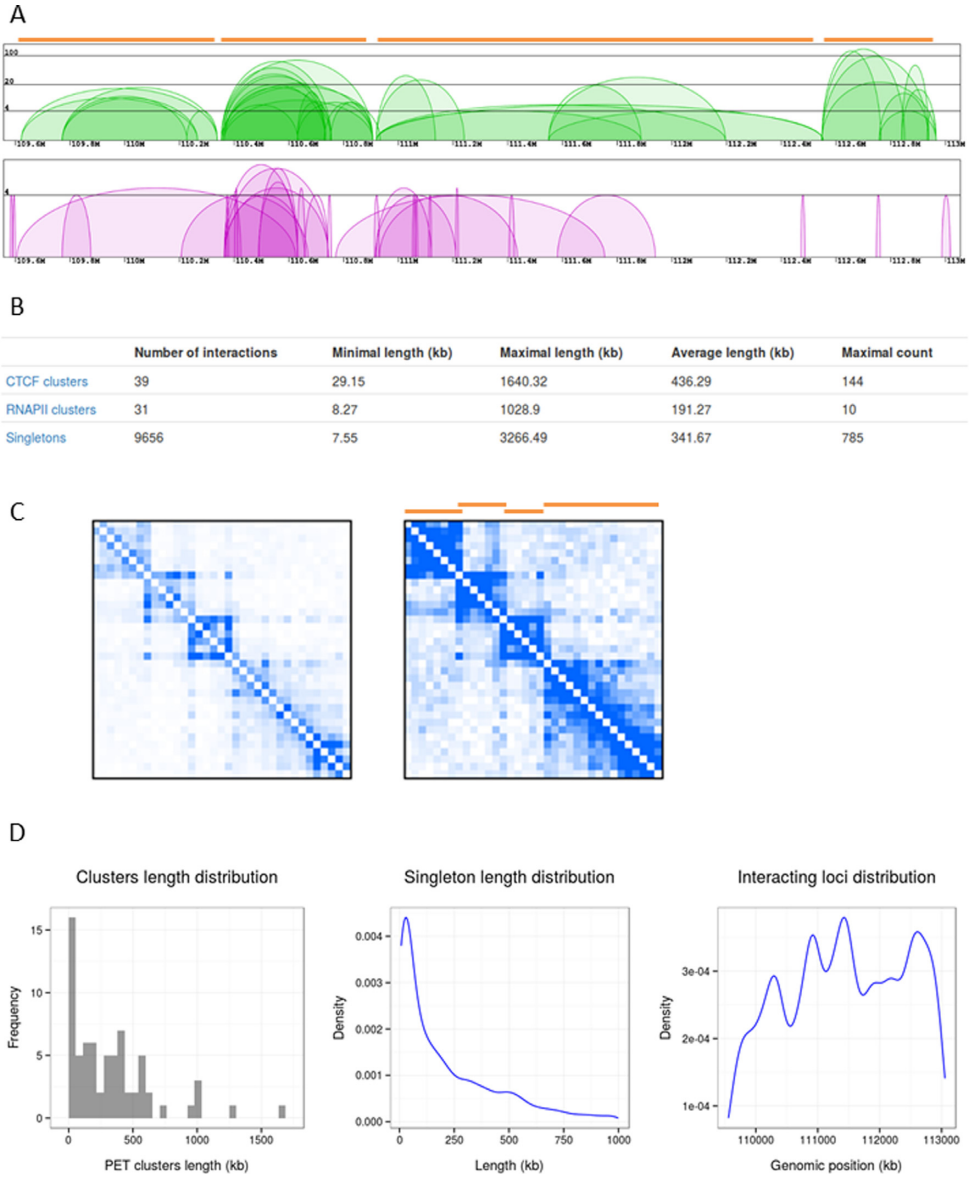
The simulation protocol is similar for all levels. First, an energy function is defined taking into account the data available on this particular level. For the low-resolution level we apply the energy function build using singleton heatmaps. The number of interactions between regions can be used as a proxy of their pairwise physical distances – intuitively, the more interactions between the regions the closer they should be in 3D space. Thus, the interaction frequencies are converted to expected distances between genomic regions. The proper way to convert from interaction frequencies to distances, and even the appropriateness of such a conversion, is the subject of much discussion in the literature. We use a simple inverse power-law with user adjustable parameters, an approach used by most other modeling programs. For the high-resolution level we use PET clusters to position the anchors within an interaction complex, and we include terms to account for typical polymer physics interactions like stretching and bending energies. We do not consider excluded volume interactions, because including such an interaction generally introduced only minor modifications to the structure but dramatically increased the computation time. There are two optional refinements on this level: first, if CTCF motif orientation is available, as it is in the GM12878 line (5,6), then these can be used to orient the interactions (Figure 2C), and secondly, the shape of chromatin loops may be modified using high-resolution singleton heatmaps. A more comprehensive discussion of our modeling assumptions and the functional forms used to describe the various interaction terms can be found in the technical documentation on the website.

## USAGE

### Input

There are two potential use cases. In the first, the user has generated their own dataset from a ChIA-PET or Hi-C library and would like to generate 3D structures. The data should be stored in a tab-delimited, bedpe-like (http://bedtools.readthedocs.org/en/latest/content/general-usage.html) file consisting of seven or eight columns, where the first three columns describe the region on one side of the interaction (chromosome, start and stop positions), the second three columns describe the region on the other side of the interaction (chromosome, start and stop positions), and the seventh column indicates the frequency of that interaction in the dataset. The eighth column is optional and may be used to name the transcription factors pulled down in the experiments. The data should be sorted into two files, the first containing high-frequency, high-confidence 'true' interactions, and the second containing low-frequency, low-confidence singleton interactions. The user is also advised to supply a file with a definition of TADs, as this may lead to a more reasonable structure. If this file is not provided, then a heuristic algorithm will be used to determine TADs automatically. In the second use case, a user may generate 3D structures using our recently published GM12878 ChIA-PET data set (5). This data is stored on our server and requires no additional input file from the user.

In either use case, the user then selects the genomic region they would like to model and, optionally, specifies a name for their model. Additionally, there are a number of simulation parameters, which the user may tune, including the weights of various interactions, or features (like CTCF ori-




| | Number of interactions | Minimal length (kb) | Maximal length (kb) | Average length (kb) | Maximal count |
|---|---|---|---|---|---|
| CTCF clusters | 39 | 29.15 | 1640.32 | 436.29 | 144 |
| RNAPII clusters | 31 | 8.27 | 1028.9 | 191.27 | 10 |
| Singletons | 9656 | 7.55 | 3266.49 | 341.67 | 785 |

**Figure 3.** Example results page for a selected region (chr 4:109556994–113054287). (**A**) Interaction arcs representing the strength of PET interactions. Orange bars on top correspond to disjoint interaction subsets that can be distinguished. (**B**) Various statistics on the PET interactions (separately for CTCF and RNAPII) and singleton interactions. (**C**) The heatmaps showing the raw (left) and normalized (right) singleton data. Orange bars on top of the normalized heatmap represent a possible TAD calling for this region. (**D**) Plots showing the length distribution for PET and singleton interactions and the number of interactions originating from each site.

entation) and the number of algorithm iterations on each genomic scale. These are set to reasonable defaults, which work well for the GM12878 dataset, but may not be appropriate for other species or even other human cell lines. The purpose of each parameter is described, and there is a link to a help document for additional information. Finding a reasonable parameter set for a given data set may involve some trial and error. We tested our modeling approach for four additional cell lines: HEK293T, K562, HeLa and MCF7, confirming that indeed the changes of parameters values and different segmentations of chromatin chain are needed. Nevertheless we were able to prepare successfully three-dimensional models in all cases using our simulation code for those cell lines.

Upon clicking the submit button, the user will be provided with a custom URL to a page where they can find their results. While the simulation is running the page will indicate its status.

## Output

When the simulation finishes the status page will show several plots for data analysis, as shown in Figure 3 for ∼3.5 Mb region on chromosome 4. At the top is a 1D arc representation of the PET interactions, where the x-axis represents genomic position and the arc height represents the measured contact frequency. This representation allows the user to quickly evaluate the number and distribution of in-

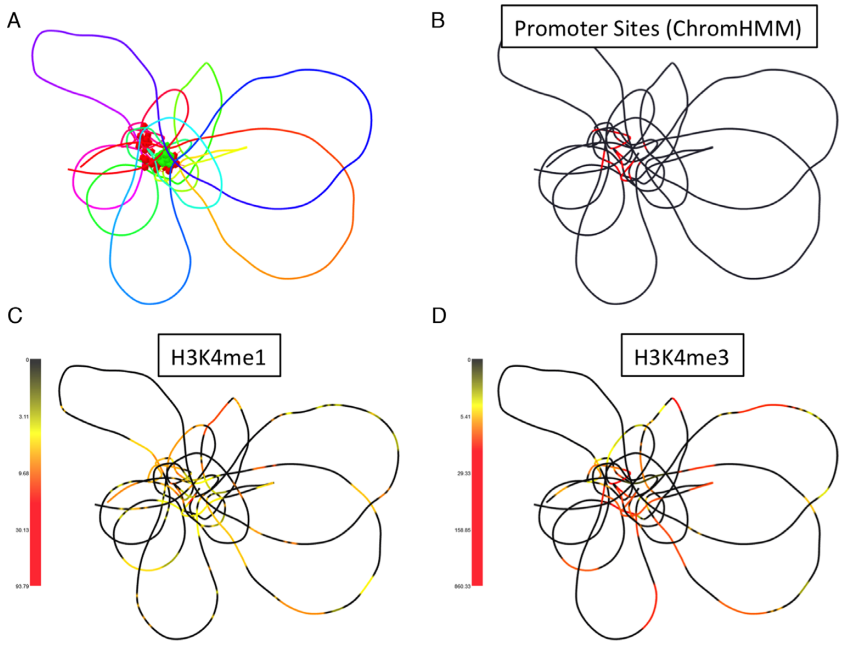



**Figure 4.** 3D structure of chr2:24026977-24629787 region for the GM12878 genome. (**A**) Structure colored according to genomic position. The locations of CTCF (green) and RNA pol II (red) are indicated by spheres. (**B**) Strong and weak promoter sites identified by ChromHMM (ENCODE) are colored red. The promoter sites co-localize within the chromosome cluster. (**C**) Intensity of H3K4me1 histone marks (ENCODE). (**D**) Intensity of H3K4me3 histone marks (ENCODE).

teraction clusters, and can be useful in estimating the accuracy of a 3D model. Next the user is presented with some population-level statistics on the PET interactions and singletons used in the current simulation. The singleton data is represented using conventional 2D heatmaps: one heatmap shows the raw data, and the other shows the normalized heatmap. Lastly, several helpful distributions are plotted: a histogram of PET cluster length, a density plot of singleton interaction distance, and the density of interacting loci in the selected region of interest.

It can be readily seen that RNAPII interactions are much shorter than CTCF ones (with the average length of 191 kb for RNAPII and 436kb for CTCF), and that they are usually found very close to the CTCF sites, which is in concordance with the genome folding model we proposed earlier (5). With some simplification it can be said that CTCF is a major contributor that shapes the genome topology, with one of its functions being bringing together the transcription and regulatory elements. Given the presented interaction plots one could argue that the selected region is comprised of 4 substructures (Figure 3A, marked with orange bars), with no PET interactions joining them. This detailed information conveyed by the interaction arcs plots is complemented by singleton heatmaps which allow easy recognition of TADs. Looking at the heatmaps alone it seems that there are three, four or five TADs, depending on whether we prefer to identify small and dense regions, or larger, but possibly less compacted ones. Interestingly, one of the most apparent splits into TADs (Figure 3C, marked with orange bars) does not entirely align with the regions suggested based on the PET interactions, suggest that inferring 3D structures requires careful examination and interpretation of both types of data.

At the top of the page is a link labeled 'Open 3D view', which will open a page for the interactive 3D viewer with the model pre-loaded. The viewer supports all the usual interactions - translation, rotation, and zoom. A wide variety of options are provided through a dropdown menu on the right. Here we only mention a few of these, and refer the reader to the online tutorial for a full list of options.

Figure 4 shows a model of chr2:24026977-24629787, a region we previously investigated for its complex but functional looping interactions (5). For ChIA-PET libraries, the viewer can display the locations of the DNA binding protein(s) used to create the library, as shown in Figure 4A, for CTCF (green) and RNA pol II (red). Notably, the user may upload a genomic annotation file in the broadPeak (http://genome.ucsc.edu/FAQ/FAQformat.html#format13) format used by ENCODE, which will color the 3D structure according to the intensity of the annotated peak. The color scale may be adjusted in a variety of ways to highlight regions of interest. Figure 4B shows the locations of strong and weak promoters (red), according to ChromHMM (ENCODE). It is immediately obvious that the promoters are colocalized within the cluster. This example demonstrates the utility of 3D modeling; to make such an inference from a 1D or 2D representation would require examining neighbors, next-nearest-neighbors, etc., to elucidate the promoter cluster. Additional genomic annotations are shown in Figure 4C (H3K4me1) and d (H3K4me3). In these cases, there is no apparent colocalization, as the histone marks are rather evenly distributed within the cluster and along the loops.

There are two options for locally saving the structure. The current view can be saved as an image (png), suitable for publication and presentations. Additionally, the user can

download the structure as an STL file, which is a common format for 3D rendering software.

## DISCUSSION

The 3D-GNOME web server provides easy access to our 3D chromatin modeling platform and a wide array of analysis tools, all packaged in a convenient, user-friendly environment. A user with no computational expertise can easily create 3D models from 3C-like data, and we expect the availability of such a resource to greatly expand the opportunities for 3D chromatin analysis. Model generation typically takes a few minutes, and multiple structures can be requested simultaneously. The 1D arcs, 2D heatmaps and 3D structures generated by 3D-GNOME offer complementary representations of the library data. Together, these representations offer ample opportunities for data analysis, hypothesis generation, and testing. We believe that the 3D-GNOME webserver is a valuable tool for researchers that are already interested in the higher order chromatin organization, but are lacking either the experimental data (the first scenario), or advanced simulation software (the second scenario) to infer the 3D structures from their own interaction data.

## ACKNOWLEDGEMENTS

We thank Keith Sheppard and Mei Xiao for their contributions to the web-based viewer, particularly to the server side, and Gosia Popiel for help on preparing Figures. The authors would like to thank the reviewers for their comments that help improve the manuscript and the web server.

## FUNDING

National Leading Research Centre in Bialystok [to P.S., D.P.]; European Union under the European Social Fund [to P.S., D.P.]; Polish National Science Centre [2014/15/B/ST6/05082, 2013/09/B/NZ2/00121 to D.P., G.M., P.W.]; European Cooperation in Science and Technology action [COST BM1405, BM1408 to D.P., G.M., P.W.]; Director Innovation Fund from Jackson Laboratory and NCI [R01 CA186714 to Y.R.]. Y.R. is also supported by the Roux family as a Florine Roux Endowed Chair and Professor in Genomics and Computational Biology. Funding for open access charge: Polish National Science Centre [2014/15/B/ST6/05082].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A. and Mei,P.H. (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.
2. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J. and Dorschner,M.O. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
3. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
4. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J. and Sedat,J. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
5. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J. and Ruszczycki,B. (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
6. 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. and Lander,E.S. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
7. Rousseau,M., Fraser,J., Ferraiuolo,M.A., Dostie,J. and Blanchette,M. (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, **12**, 414.
8. Trieu,T. and Cheng,J. (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res.*, **42**, e52.
9. Hu,M., Deng,K., Qin,Z., Dixon,J., Selvaraj,S., Fang,J., Ren,B. and Liu,J.S. (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.*, **9**, e1002893.
10. Goh,Y., Fullwood,M.J., Poh,H.M., Peh,S.Q., Ong,C.T., Zhang,J., Ruan,X. and Ruan,Y. (2012) Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. *J. Visual. Exp.: JoVE*, **62**, e3770.
11. Li,G., Fullwood,M., Xu,H., Mulawadi,F., Velkov,S., Vega,V., Ariyaratne,P., Mohamed,Y., Ooi,H. and Tennakoon,C. (2010) Software ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, **11**, R22.
12. Zhang,J., Poh,H.M., Peh,S.Q., Sia,Y.Y., Li,G., Mulawadi,F.H., Goh,Y., Fullwood,M.J., Sung,W.-K. and Ruan,X. (2012) ChIA-PET analysis of transcriptional chromatin interactions. *Methods*, **58**, 289–299.
13. Serra,F., Di Stefano,M., Spill,Y.G., Cuartero,Y., Goodstadt,M., Baù,D. and Marti-Renom,M.A. (2015) Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.*, **589**, 2987–2995.
14. Giorgetti,L., Galupa,R., Nora,E.P., Piolot,T., Lam,F., Dekker,J., Tiana,G. and Heard,E. (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, **157**, 950–963.
15. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
16. Meluzzi,D. and Arya,G. (2012) Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Res.*, **41**, 63–75.
17. Zhang,Z., Li,G., Toh,K.-C. and Sung,W.-K. (2013) 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol.*, **20**, 831–846.
18. Lesne,A., Riposo,J., Roger,P., Cournac,A. and Mozziconacci,J. (2014) 3D genome reconstruction from chromosomal contacts. *Nat. Methods*, **11**, 1141–1143.
19. Baù,D. and Marti-Renom,M.A. (2012) Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods*, **58**, 300–306.
20. 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**, i26–i33.
21. Peng,C., Fu,L.-Y., Dong,P.-F., Deng,Z.-L., Li,J.-X., Wang,X.-T. and Zhang,H.-Y. (2013) The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Res.*, **41**, e183.
22. Trieu,T. and Cheng,J. (2015) MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics*, **32**, 1286–1292.
23. Russel,D., Lasker,K., Webb,B., Velázquez-Muriel,J., Tjioe,E., Schneidman-Duhovny,D., Peterson,B. and Sali,A. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.*, **10**, e1001244.