SCIENTIFIC REPORTS

OPEN

GMOL: An Interactive Tool for 3D Genome Structure Visualization

Jackson Nowotny[1], Avery Wells[1], Oluwatosin Oluwadare[1], Lingfei Xu[1], Renzhi Cao[1], Tuan Trieu[1], Chenfeng He[1] & Jianlin Cheng[1,2,3]

It has been shown that genome spatial structures largely affect both genome activity and DNA function. Knowing this, many researchers are currently attempting to accurately model genome structures. Despite these increased efforts there still exists a shortage of tools dedicated to visualizing the genome. Creating a tool that can accurately visualize the genome can aid researchers by highlighting structural relationships that may not be obvious when examining the sequence information alone. Here we present a desktop application, known as GMOL, designed to effectively visualize genome structures so that researchers may better analyze genomic data. GMOL was developed based upon our multi-scale approach that allows a user to scale between six separate levels within the genome. With GMOL, a user can choose any unit at any scale and scale it up or down to visualize its structure and retrieve corresponding genome sequences. Users can also interactively manipulate and measure the whole genome structure and extract static images and machine-readable data files in PDB format from the multi-scale structure. By using GMOL researchers will be able to better understand and analyze genome structure models and the impact their structural relations have on genome activity and DNA function.

Recent studies have shown that in addition to the genome sequence, the genome's spatial structure also has a tremendous impact on genome activity and DNA function including gene expression and genome stability[1]. Lately, much work has been done in an attempt to decipher such genome structures and several different models have been proposed as probable structures[2–7]. Moreover, several computational methods have been developed to construct realistic 3D structures of genomes or chromosomes from chromosomal conformation capturing data, such as Hi-C, generated by next generation sequencing techniques[2,3,16].

Considering the amount of work in genome structure modeling, a visualization tool that can help researchers visualize and analyze 3D genome structures will undoubtedly benefit genome structure study. Visualization of genome structures is vital to continuous progress in the field because it showcases relationships within the genome that cannot be inferred from sequence information alone. Despite the importance of visualization, not much progress has been made in this area until now. Several tools have been developed for small molecular, such as protein, structure visualization such as Jmol[8], Pymol[9], Chimera[10], etc., but, when it comes to large-scale structures, such as the human genome, there are two limitations that prevent these programs from being effective. First, the Protein Data Bank (PDB) file format is typically used to store molecular structure data for the tools to visualize, however, the standard PDB format was designed for small molecular structures[11] and consequently is not sufficient for storing the vast amounts of data required to visualize genome structures. Second, running tools to load the entire genome structure data so that it is compatible with these programs is a strenuous or impossible task.
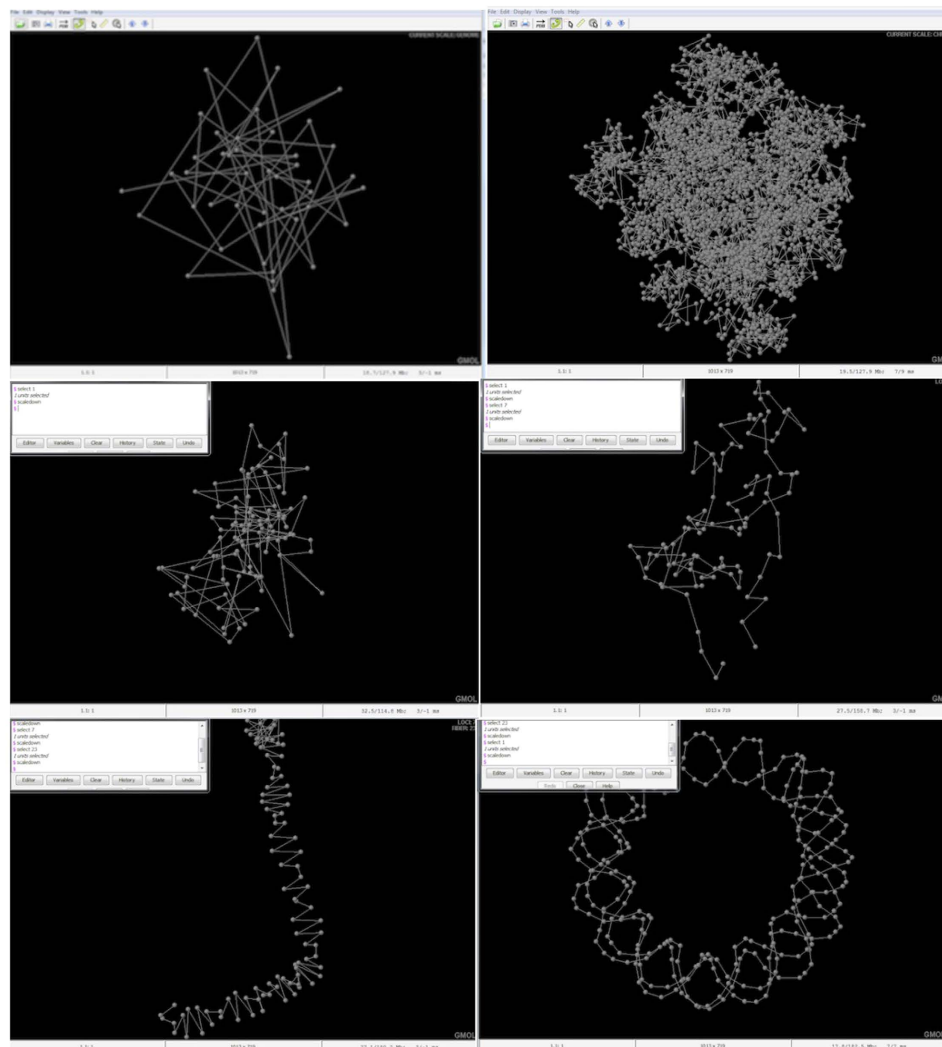
To our knowledge only one tool, Genome3D[12], has been specially designed for genome structure visualization. However, Genome3D lacks advanced functions in a few key areas including selection functions and scales amongst others. Here we introduce another genome structure visualization tool named GMOL that adequately improves upon and provides much needed functions, thus successfully filling the need for a genome visualization tool for researchers. GMOL is available to download on the GMOL sourceforge site along with accompanying sample data and documentation including an installation guide, usage guide, and walkthrough.

Implementation

GMOL was developed from Jmol, an open-source Java application that visualizes chemical structures[8]. The fundamental features of Jmol that are necessary for genome visualization were preserved and are not discussed.

[1]Computer Science Department, University of Missouri, Columbia, MO 65211, USA. [2]Informatics Institute, University of Missouri, Columbia, MO 65211, USA. [3]C.S. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA. Correspondence and requests for materials should be addressed to J.C. (email: chengji@missouri.edu)

Figure 1. Visualization of Human Genome Structure of Different Scales. The structure at each level is visualized in a dynamic fashion such that it can be rotated, translated, colored, and zoomed in and out.

However, GMOL adds and modifies several additional functions to make genome structure visualization possible and sufficient thus differentiating GMOL from Jmol. The added and modified functions that are specific to GMOL including the scaling system, selection system, sequence querying, measuring system, and new file format and are all described below.

To visualize and store genome structures GMOL utilizes a six-scale system. The six scales are (listed from lower resolution to higher resolution or from large scale to small scale): genome scale (Gb), chromosome scale (50–100 Mb), loci scale (Mb), fiber scale (Kb), nucleosome scale (100b) and nucleotide scale (1b). A smaller scale structure is a component of the next larger scale structure, therefore a larger scale is comprised of the combination of the components of the smaller scale. For example, the genome scale visualizes all the chromosomes; the chromosome scale visualizes all the loci, etc. This multi-scale system of genome models is largely inspired by the "fractal globule model"[13]. The multi-scale system can transition between scales through the entire data set or through an individual data point. For example, once at the global scale, the chromosome scale can be scaled to through the entire data set in which all the data points for all chromosomes would be visualized, or the chromosome scale can be scaled to via an individual chromosome in which all of the data points for that particular chromosome would be visualized. This versatility in scaling can be viewed in Fig. 1 which shows the human genome structure visualized at different scales using data from Bancaud *et al.*[13]. Furthermore, Fig. 2 shows screenshots of GMOL at various levels of scales. From left to right, top to bottom, the scales represented are the global scale, chromosome scale of all chromosomes, chromosome scale of a single chromosome, loci scale of a single data point, fiber scale of a single data point, and nucleosome scale of a single data point. Here the data from Asbury *et al.* was used[12].

The backbone of the multi-scale visualization system is toggling between scales. To do this, GMOL utilizes a selection feature that allows the user to select any unit, at any scale, and scale it up to a lower resolution or down to a higher resolution. By scaling up, the user gets an overview of the location of the selected unit, whereas scaling down gives the user the detailed structure of the selected unit. There are multiple ways in which the user can select a unit or units. One way is called index selection, in which the user can select units by using their index

Figure 2. GMOL screenshot of different scales from data of Asbury *et al*.[12]. Screenshots of all the different scales. From left to right, top to bottom, the scales represented are the global scale, chromosome scale of all chromosomes, chromosome scale of a single chromosome, loci scale of a single data point, fiber scale of a single data point, and nucleosome scale of a single data point.

in the current displaying structure. In global scale structure, unit index means chromosome number, while in chromosome/loci/fiber/nucleosome scale unit index means the sequential number of this unit based on genome sequence. Another selection method is via scale information, in which the user selects units by scale information. This method is useful if the user needs to select units within a specified chromosome/loci/fiber/nucleosome in the current displaying structure when the index in unknown. Lastly, the user can select units using genome sequence information. By specifying a genome sequence location, the corresponding units in current displaying structure will be selected.

To allow for the multi-scale system, GMOL is accompanied by a new file format called Genome Scale System (GSS). Currently the standard file format for 3D visualization of biological data is PDB. However, the existing PDB file format is standard for storing protein structures and, therefore, is inadequate for storing genome structure data as genome structure data has a much higher resolution and therefore is much larger. To solve this problem, a new file format, GSS, was designed. Corresponding to our multi-scale system, the GSS format contains the following files (from lower to higher resolution): ".gs.gss" (genome scale), ".cs.gss" (chromosome scale), ".ls.gss" (loci scale), ".fs.gss" (fiber scale), and ".ns.gss" (nucleosome scale). Each file contains a unique set of data in which files of lower resolution store the position of the central point of compartments of the next higher resolution. More specifically, ".gs.gss" files contain the location of the central point of all chromosomes, ".cs.gss" files contain the location of the central point of all loci, ".ls.gss" files contain the location of the central point of all fibers, ".fs.gss" files contain the location of the central point of all nucleosome core particles (NCP), and finally "ns.gss" files contain all the nucleotides in a NCP. Based on this hierarchical organization of the GSS file system, GMOL is able to read and display structures at any resolution according to the user's requirements. GSS file format is easily convertible from PDB format with the appropriate scripts. These scripts and detailed documentation are provided
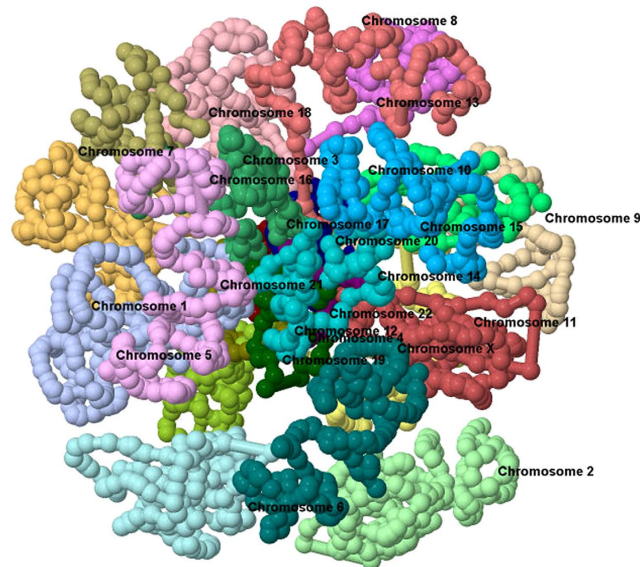
Figure 3. Image of Visualized Genome on Chromosome scale using GMOL. Extracted image of a resulting 3D genome structure visualized in GMOL, in the chromosome scale. Here each chromosome within the genome is highlighted with a different color and labeled for identification. The visualization is from a genome previously modeled[16].

on the GMOL sourceforge site. Various methods and tools are separately available for converting Hi-C data to genome models in PDB format[2,3,16].

In addition to scaling from a selection of units, GMOL can query the selected units into an Ensembl[14] database or a local database to gather genome sequence information about the selection. The integration of JEnsembl[15] with GMOL enables querying of the Ensembl database.

Another feature of GMOL is its measuring capabilities. GMOL allows the user to measure certain selected units in the currently visualized structure. Specifically, GMOL can measure the distance in between any two units in nanometers, and measure the angle formed between any three units in degrees.

Results

Functionality of GMOL. The multi-scale system of GMOL allows the various resolutions of the structure to be viewed with accuracy and precision. In addition, giving each scale its own file type allows for faster viewing and scaling between scales. The multi-scale system also allows for more total data to be represented by giving each scale its own system of data points. This, in turn, creates a more accurate and reliable genome structure.

The selection system grants flexibility and ease in terms of how units are selected. This makes selection easy and simple as certain selection methods are bettered suited for certain ranges of data. In addition, users are free to use different selection methods based on their preference.

Flexibility is also represented in GMOL with regards to querying databases. Since GMOL allows querying to a local database or Ensemble, users are free to choose based on their preference and aren't limited.

The measuring system incorporated into GMOL supplies convenient methods to obtain data regarding the genome structure. Via the interface or console, users can measure distances or angles with respect to selected units with ease.

Finally, the unique file type create for GMOL, GSS, allows GMOL to visualize various scales of structures to its fullest ability as the GSS file system grants a higher resolution of visualization and larger amounts of data to cope with the demands of genome structures of which the PDB file system can't provide.

Visualization Examples of GMOL. Figure 3 shows an extracted image of a resulting 3D genome structure visualized in GMOL. The visualization was done in the chromosome scale so each chromosome is visualized in full and with their respective positions to each other. Here, each chromosome within the genome is highlighted with a different color and labeled for identification. The visualization is from a genome modeled based on Hi-C data[2,16].

Figure 4 shows two screenshots taken of GMOL with a visualized genome open. The interface is shown as well as the measurement tool in use. The visualized models are in the genome scale. One of the chromosomes, represented as point in genome scale, is highlighted. Here the data from Asbury *et al.* was used[12].

Applicability of GMOL through Analyzing Genome Structures. Here, we give an example of how GMOL could be used in a practical situation to analyze the differences between the genome structures of two individuals. As other methods of comparison are certainly useful, analysis of tertiary structures are also beneficial. As previously mentioned, research has shown that genome spatial structures impacts genome activity and DNA function[1]. This means that the variations in genome structures amongst individuals could account for minor
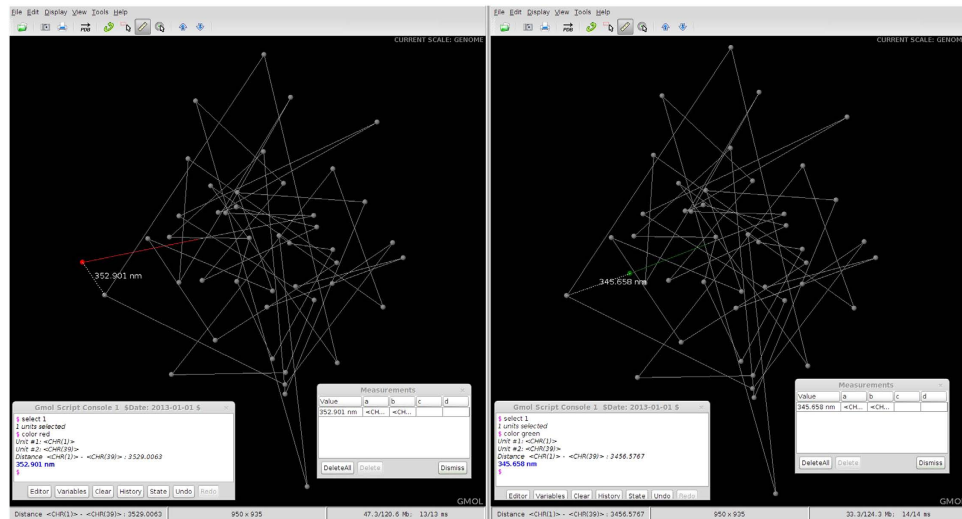
Figure 4. Comparison of Genome Structures at Genome Scale. Side by side screenshots of GMOL visualizing the genome of two models at genome scale with Person A on the left and Person B on the right. The difference in position of Chromosome 1 between the two models is highlighted.
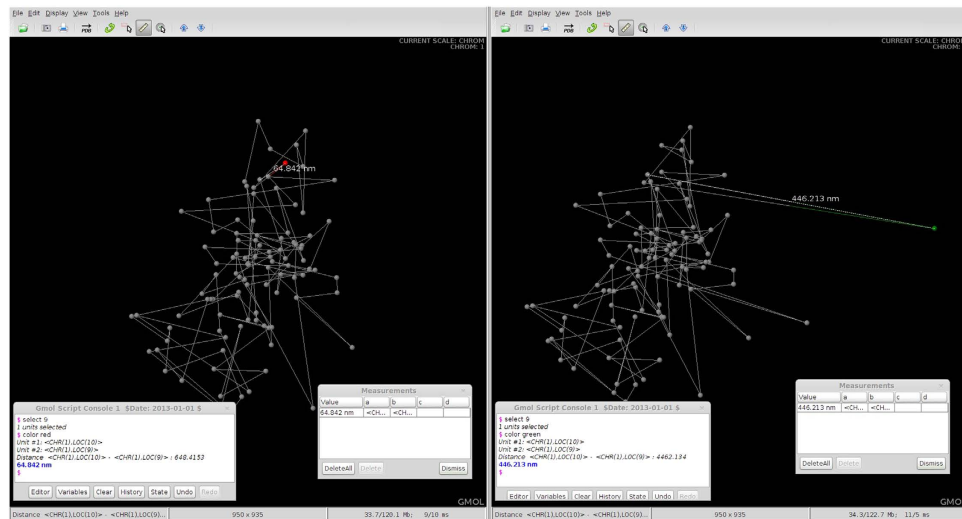
Figure 5. Comparison of Chromosome Structures at Chromosome Scale. Side by side screenshots of GMOL visualizing chromosome 1 of two models at chromosome scale with Person A on the left and Person B on the right. The structural differences within Chromosome 1 are highlighted.

differences such as eye color, but also for more major health concerns such as cancer. By using GMOL to analyze genome structures, researchers/biologists can quickly spot abnormal sections of the genome and easily scale-up or down to get a more detailed view of the areas of concern.

In Fig. 4, the genome of Person A is displayed on the left and the genome of Person B is displayed on the right. As shown, the genome structures are almost identical except for the location of Chromosome 1 (highlighted in red for Person A and green for Person B). Assuming Person A is healthy and Person B has been diagnosed with cancer, this difference in positioning of Chromosome 1 should cause concern. By selecting the chromosome unit and scaling down we can get a more detailed view of the structure of Chromosome 1 for each individual (Fig. 5).

Figure 5 highlights the spatial changes in the structure of Chromosome 1 between Person A and Person B. To view these structural differences within the context of the entire genome at the chromosome scale, we simply scale-up to go back to the genome scale and then scale-down to show the global structure at the chromosome level (Fig. 6).

This sample example demonstrates how one might use GMOL in a practical scenario to analyze the differences between two genome structures. The data in this example were taken from Asbury *et al.*[12] and were from a human CD4+ T cell. For a more comprehensive walkthrough of GMOL's features see the walkthrough on the GMOL website.
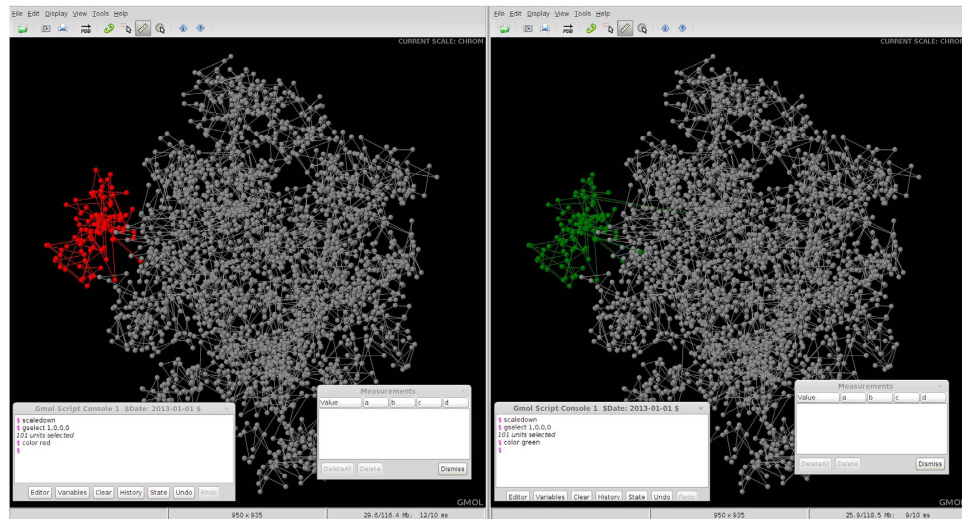
Figure 6. Comparison of Chromosome Structures at Genome Scale. Side by side screenshots of GMOL visualizing chromosomes of two models at the genome scale with Person A on the left and Person B on the right. The structural differences of Chromosome 1 within the context of the entire genome are highlighted.

Functions	GMOL	Genome3D
Select Function	1.select based on index	Select only based on genome location
	2.select based on scale information	
	3.select based on sequence	
Measurement	Supported	Not Supported
Sequence querying	1.query from Ensembl	Only query from local database
	2.query from local database	
Scripts/Commands	Supported	Not Supported
Visualization Scales	Genome	Giant Loop
	Chromosome	Fiber
	Loci	Nucleosome
	Fiber	
	Nucleosome	
	Nucleotide	

Table 1. Comparison of GMOL and Genome3D.

Comparison with Genome3D. GMOL improves upon Genome3D[14] by supplying some important and needed features that are necessary for adequate genome visualization, of which Genome3D lacks. One way this is demonstrated is through the available scales offered, in which Genome3D displays genome structures at three scales: Giant Loop, Fiber and Nucleosome. GMOL's multi-scale system utilizes six scales: genome, chromosome, loci, fiber, nucleosome, and nucleotide. These additional scales allow a more detailed view of genome structures that cannot be achieved with Genome3D. GMOL also implements multiple selection functions (based on index, based on scale information, based on sequence) whereas Genome3D only allows selection based genome location. Having multiple selection functions enables intuitive selection of any portion of the genome structure from any scale. Furthermore, GMOL supports distance and angle measurement functions that Genome3D lacks. With regards to sequence querying, both Genome3D and GMOL support it, but GMOL supports querying from Ensembl and from a local database whereas Genome3D supports querying only from a local database. Finally, GMOL allows the user to write custom scripts and commands so that they may extend the functionality of GMOL to needs specific to their project. Genome3D does not implement this feature.

Ultimately, GMOL and Genome3D perform the same basic functions, however GMOL allows for more detailed analysis of genome spatial structures in several key areas that allows researchers to achieve better answers regarding structural relationships. The main differences between GMOL and Genome3D are summarized in Table 1.

Future Development of New Features. A number of future developments of new features of GMOL are planned for implementation and integration in the near future. One such feature is the integration of additional databases of which to export sequences from GMOL to. Currently, GMOL supports querying sequences to Ensemble and a local database. Additional databases being integrated include UCSC Genome browser[17], ENCODE[18], and Uniprot[19]. Another planned feature to include is a function that allows the selection of two points, of which to then be visualized for comparison. This function will allow an easier and better method to compare two sections of the genome. A third feature planned for future development is the integration of a function to view the sequence of a selected point. Such function would be convenient with regards to getting the sequence of only a selection. A fourth area of future study is the inclusion additional features into GMOL to allow users to study the correlation between genome structural variations and other sources of information such as primary sequence variations, SNPs, transcriptomics, and proteomics data. This will allow for a more versatile approach to analyzing genomes.

Discussion

The recent development and research in genome nature indicates the significance of 3D genome structures as well as genome sequence. The genome's structure provides a key contribution to certain genome activity and DNA functions including gene expression and genome stability[1]. The implications of such findings suggest that much work is needed to figure out 3D structures of genomes. One vital step in the process of studying 3D genome structures is the visualization process that turns the coordinates of a generated structure into a 3D, interactive model.

GMOL is an application designed to sufficiently perform this step of the research process by effectively visualizing genome tertiary structures. GMOL does this through its immense array of features and functions including its multi-scale system that allows visualization of six chronological resolutions. GMOL also successfully fulfills its goals through its multi-selection system, measurement capabilities, options of sequence querying, and new file format system.

Therefore, through GMOL, researchers can better analyze their genomics data. By using GMOL to visualize the genome, researchers may see patterns or other structural relationships that are not evident in their data alone. By utilizing six different scales, GMOL allows for a level of detail that cannot be obtained by any other currently released program. That combined with GMOL's other unique capabilities set it apart in the marketplace for genome visualization software.

References

1. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell*. **128,** 787–800 (2007).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **326,** 289–293 (2009).
3. Dekker, J., Rippe, K., Dekker, M. & Kleckner M. Capturing chromosome conformation. *Science*, **295,** 1306–1311 (2002).
4. Mateos-Langerak, J. *et al.* Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences of the United States of America*. **106,** 3812–3817 (2009).
5. Grosberg, A., Rabin, I., Khavlin S. & Nir, A. Self-similarity in the structure of DNA: why are introns needed? *Biofizika*. **38,** 75–83 (1993).
6. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozzicconacci, J. 3D genome reconstruction from chromosomal contacts. *Nature Methods*. **11,** 1141–1143 (2014).
7. Shavit, Y., Kathryn, F. & Pietro, L. FisHiCal: An R package for iterative FISH-based calibration of Hi-C data. *Bioinformatics*. **30,** 3120–3122 (2014).
8. Jmol Development Team (2013). Jmol: an open-source Java viewer for chemical structures in 3D. URL http://www.jmol.org
9. Schrodinger, L. L. C. (2015). *The PyMOL Molecular Graphics System*. URL http://www.pymol.org
10. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*. **25,** 1605–1612 (2004).
11. Sussman, J. L. *et al.* Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta crystallographica. Section D, Biological crystallography*. **54,** 1078–1084 (2008).
12. Asbury, T. M., Mitman, M., Tang, J. & Zheng, W. J. Genome3D: a viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-dimensional genome. *BMC bioinformatics*. **11,** 444 (2010).
13. Bancaud, A., Lavelle, C., Huet, S. & Elleberg, J. A fractal model for nuclear organization: current evidence and biological implications. *Nucleic acids research*. **40,** 8783–8792 (2012).
14. Flicek, P. *et al.* Ensembl's 10th year. *Nucleic acids research*. **38,** D557–562 (2010).
15. Paterson, T. & Law A. JEnsembl: a version-aware Java API to Ensembl data systems. *Bioinformatics*. **28,** 2724–2731 (2010).
16. Trieu, T. & Cheng, J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research*. **42,** doi: 10.1093/nar/gkt1411 (2014).
17. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res*. **6,** 996–1006 (2002).
18. The ENCODE Consortium. An integrated encyclopaedia of DNA elements in the human genome. *Nature*. **489,** 57–74 (2012).
19. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*. **42,** D191–D198 (2014).

Acknowledgements

Author Contributions

JC conceived the tool. JC, JN, AW, LX, and CH designed the tool. JN, AW, LX, RC, TT, CH, OO implemented the tool. JN, AW, JC, LX, TT, CH, OO wrote and edited the manuscript.

Additional Information

How to cite this article: Nowotny, J. *et al.* GMOL: An Interactive Tool for 3D Genome Structure Visualization. *Sci. Rep.* **6**, 20802; doi: 10.1038/srep20802 (2016).