



Project Acronym: MuG

Project title: Multi-Scale Complex Genomics (MuG)

Call: H2020-EINFRA-2015-1

Topic: EINFRA-9-2015

Project Number: 676556

Project Coordinator: Institute for Research in Biomedicine (IRB Barcelona)

Project start date: 1/11/2015

Duration: 36 months

Deliverable 3.4: Preliminary browser-track that implements and connects all the 3D data from a genome or a genomic domain

Lead beneficiary: Institute for Research in Biomedicine (IRB Barcelona)

Dissemination level: PUBLIC

Due date: 30/04/2017

Actual submission date: 03/05/2017

Copyright© 2015-2018 The partners of the MuG Consortium



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676556.

Document history

Version	Contributor(s)	Partner	Date	Comments
0.1	David Castillo	CNAG-CRG	20/04/2017	First Draft
0.2	Mike Goodstadt	CNAG-CRG	26/04/2017	Second Draft
0.3	Marc A. Marti-Renom	CNAG-CRG	29/04/2017	Minor corrections
0.4	Josep Ll. Gelpí	BSC-CNS	02/05/2017	Minor corrections
0.5	Andy Yates, Mark McDowall	EMBL-EBI	3/05/2017	Revision of data management issues
1.0			03/05/2017	Approved by Supervisory Board



Table of Contents

1	INTRODUCTION.....	5
2	3D DATA VISUALIZATION	6
3	CONNECTING 3D WITH 1D AND 2D COMPONENTS	8
4	CONCLUSIONS AND FUTURE PERSPECTIVE	9
5	REFERENCES.....	10



Executive summary

Hi-C experiments capture interaction data of genomes which can be analyzed to reveal structures of chromatin in the nucleus. A number of techniques and tools exist to process the data as matrices and model possible 3D conformations that fit the observations. However, consolidating these spatial understandings with classic linear genomic reads into a productive workspace is a significant challenge. TADkit achieves this within the MuG VRE, enabling researchers to import and synchronize these data to form a comprehensive view from which to gain biological insights. The tool is now integrated in the VRE infrastructure: <http://vre.multiscalegenomics.eu/>



1 INTRODUCTION

Hi-C experiments capture interactions within the genome at distinct genomic coordinates which can be plotted as a matrix indicating the frequency of occurrence[1]. There exists a number of techniques and tools to process the data and compute possible 3D conformations that fit the observations[2]. The MuG VRE uses TADbit, a complete Python library to analyze, model and explore 3C-based data[3]. The probable structures of chromatin within the nucleus revealed by such analysis of these have led to greater understanding of the mechanisms of cell function[4].

This requires experienced and meticulous inspection of matrices to discern patterns indicative of spatial archetypes against which classic linear genomic data can then be aligned. 3D visualization lends itself to addressing the complexities of picturing and reviewing the findings. Yet existing linear genome browsers are ill suited to handle the new matrix or spatial datatypes and the few new tools which attempt this are limited in the data they handle and in ease of integration into existing workflows[5][6].

The challenge of the genome visualization tools is to combine diverse sources of information that represent different dimensions of the genome: 1D genomic features, 2D interactions matrices and 3D structural information. Scientists have the opportunity then to visualize the different dimensional data together, analyze it and come to conclusions otherwise difficult to reach with independent views.

To achieve this in the MuG VRE, the TADbit tool can be used to compute and cluster models that best fit the interaction matrix of a limited region of the genome. TADbit then saves the 3D coordinates into a JSON dataset can subsequently be viewed in TADkit, a unified workspace which connects the 3D data with other data in that genomic domain [Figure 1].

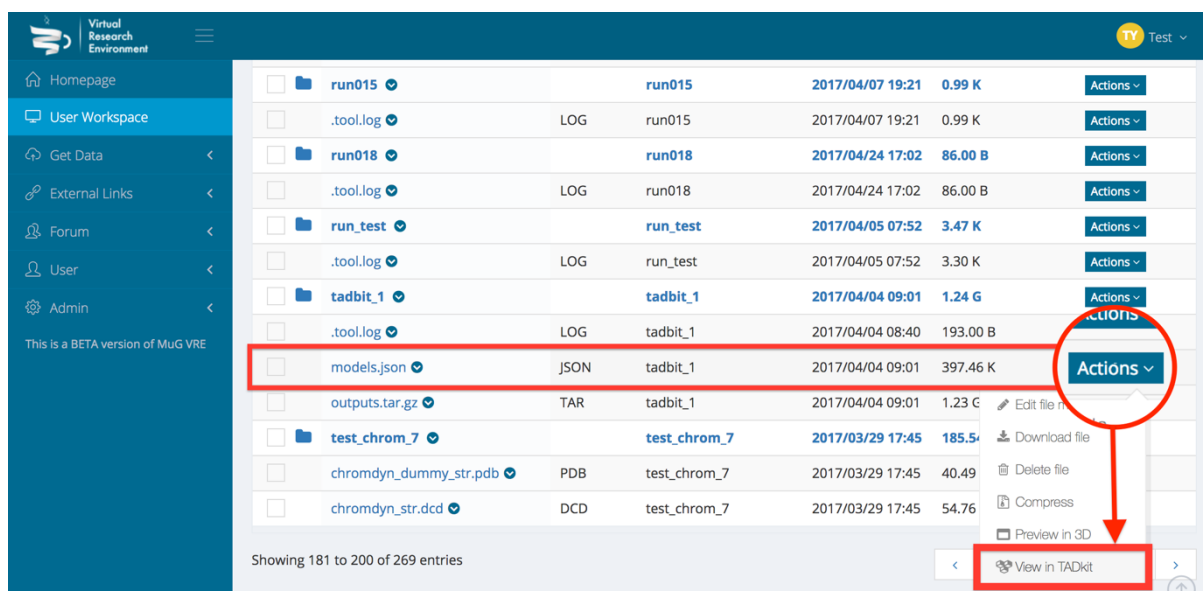


Figure 1: Selecting and Viewing 3D data in the MuG VRE

2 3D Data Visualization

The TADkit workspace displays a panel for each datatype: classic linear genomic tracks below with 2D interaction matrices on the upper right and 3D models in the upper left [Figure 2].

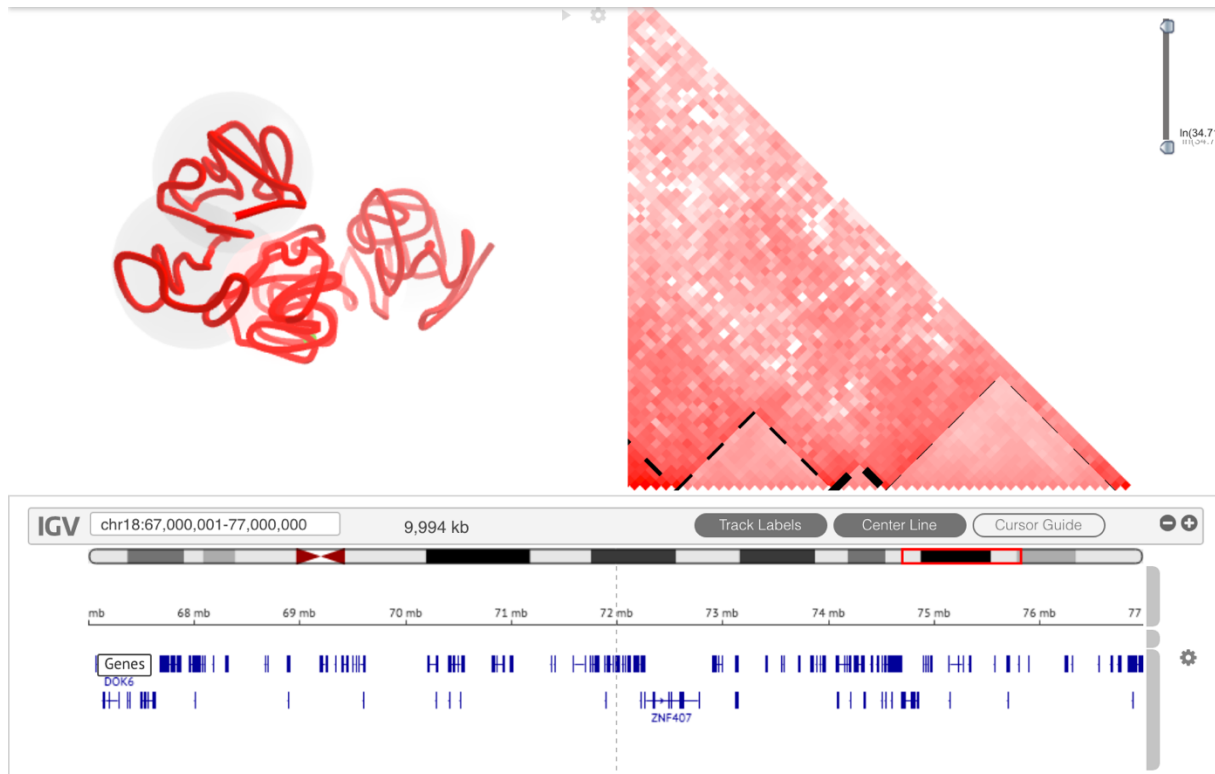


Figure 2: TADkit Layout

The panel for 3D data is an HTML canvas component based on Three.js for the rendering of the 3D structures. Each model in the input file is just a set of consecutive three-dimensional space points which are joined to represent a chromatin strand. Two different representations of the structures have been implemented in TADkit: a tubular strand and a set of consecutive spheres. In the tubular representation points are surrounded by a limited number of segments giving the tubular appearance to the model. Depending on the resolution used to compute the model the strand is given an appropriate thickness to approximately match chromatin width. In the sphere representation spheres are centered on each point of the model [Figure 3].

The scene is completed with illumination, camera and controls allowing the user to animate, rotate and zoom the model for a better visualization experience. The animation automatically rotates the model around its z axis at comfortable speed sparing the user the manual rotation. In the tubular representation, a green ring surrounding the chromatin strand representation has been included to mark the genomic position in the 1D tracks into the 3D structure.

Another important feature is the identification of the Topologically Associating Domains (TADs) as semitransparent spheres in the model. Those spheres delimiting TADs will automatically appear if the source data includes the information about the TAD boundaries. The currently selected TAD is highlighted by stronger coloring of the sphere.



Figure 3: Visualization of Chromatin and TADs derived from the 3D data.

3 CONNECTING 3D with 1D and 2D COMPONENTS

Together with the TADbit JSON in the VRE the user can select amongst different standard track types (BED, GFF, WIG, bigWig, bedGraph, BAM, VCF and SEG) as an input to TADkit. The track files should contain the genomic region of the structure conformations in the JSON.

When TADkit is opened from the VRE User Workspace, the data is automatically connected across the different components. The menu bar indicates current data title and genomic coordinates and provides loading and viewing options. The navigation in the linear genomic tracks is identified as a moving point in the 2D component and as ring in the 3D model [Figure 4a].

The TAD containing the current genomic position is highlighted at the same time in the three components. It is delimited between two dashed lines in the 1D track and highlighted in the 2D interaction matrix and the 3D model [Figure 4b].

A further connection between the 1D and 3D components has been developed. If a track contains colored, scored or valued features, a menu option allows their representation in the model. If the feature includes a color attribute, that color will be overlaid in the chromatin strand through the genomic region occupied by the feature. If the feature includes a score or value attribute; that score or value will be used to create a blue gradient color table used to overlay the strand [Figure 4c].

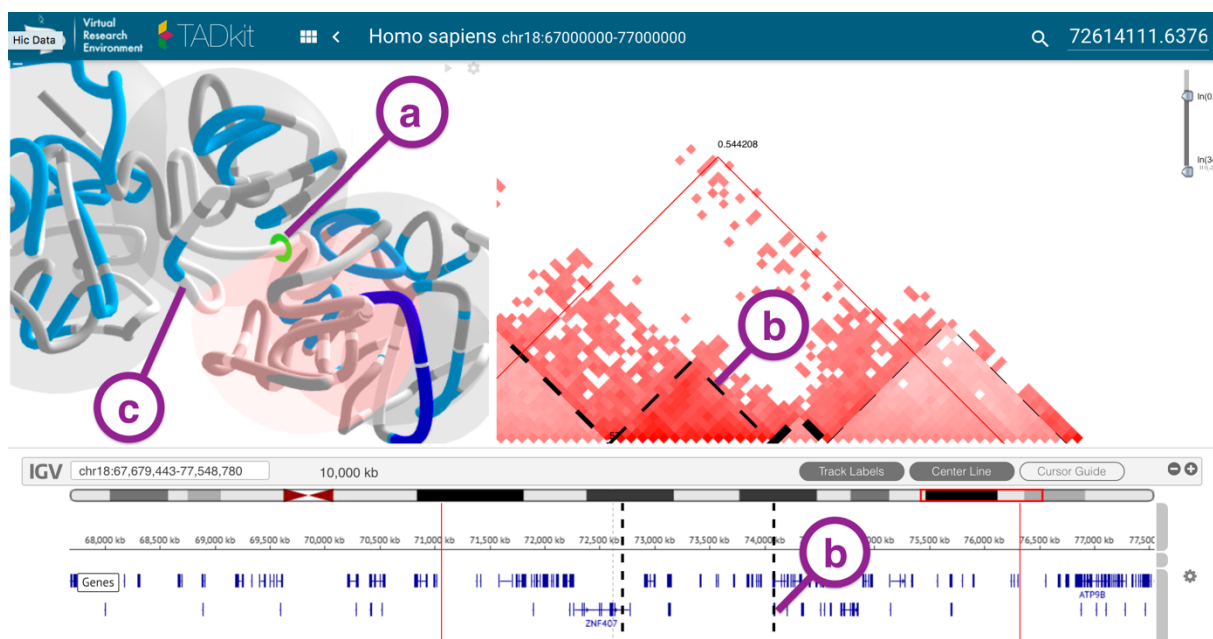


Figure 4: Connections between 3D data and genomic regions.

4 CONCLUSIONS AND FUTURE PERSPECTIVE

The 3D data computed by TADbit has been developed and integrated into TADkit with a 3D visualization component which connects with the 1D browser-track and the 2D interaction matrix of the VRE browser.

The next step in the development path for the VRE includes the integration of internal data sources and other online web services. Data retrieval will include feature tracks, interaction matrices and 3D conformations. Efforts are also focused in allowing the structure representation of genomic regions at very different scales going from full chromosomes to nucleosome scale.

The development of TADkit is an ongoing process that will be enriched by the input of the scientific community and specially the MuG pilot projects. The TADkit code is open source and is available on GitHub: <https://github.com/3DGenomes/TADkit>.



5 REFERENCES

- 1 Galip G & Noble WS (2016) Software tools for visualizing Hi-C data Large scale visualization. , 1–9.
- 2 Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, Baù D & Marti-Renom MA (2015) Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* **589**, 2987–2995.
- 3 Serra F, Baù D, Fillion G & Marti-Renom MA (2016) Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling. .
- 4 Dekker J, Marti-Renom MA & Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403.
- 5 Arndt W, Asbury TM, Zheng WJ, Mitman M & Tang J (2011) Genome3D: A viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-dimensional genome. *2011 IEEE Int. Conf. Bioinforma. Biomed. Work. BIBMW 2011*, 936–938.
- 6 Butyaev A, Mavlyutov R, Blanchette M, Cudré-Mauroux P & Waldispühl J (2015) A low-latency, big database system and browser for storage, querying and visualization of 3D genomic data. *Nucleic Acids Res.* **43**, e103.

