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 3.7: A first prototype of analysis tools for mining of the data provided by the browser

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

Dissemination level: PUBLIC

Due date: 31/10/2017

Actual submission date: 18/12/2017

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	06/12/2017	First Draft
1.0			18/12/2017	Approved by Supervisory Board

Table of Contents

Executive summary

TADkit provides synchronized visualization of 1D genomic information, 2D matrix interactions and 3D structural models to the Virtual Research Environment (VRE) in a single and unified environment. The convergence of these different types of information turns TADkit into a convenient point of source for a data mining engine. The combination of the data accessible by TADkit opens new opportunities to provide scientist with enriched figures and reports.

This deliverable takes the first steps to provide TADkit with an analysis module for data mining. This first prototype includes two statistics interrelating sources of information with different dimensionality.

1 INTRODUCTION

The mining tools module required the implementation of two additional layers in TADkit: a middle layer responsible of data sorting and preparation and the final user interface.

The middle layer communicates with the previously existing core engine that gathers, interprets and filters the data that is finally combined in the main view of the browser. This layer collects and prepares the information from the different dimensional sources and combine it according to the type of report and the filters specified in the user interface.

The user interface has been included as a new “Data mining” tab in the Information Panel [***Figure***]. The new tab contains access buttons to the two first analysis tools included in this prototype. Each tool is composed by a set of filters and a container where the result of the data query is displayed. For the implementation of the data container, the development team uses the existing and widely adopted angular module “ui-grid”. The module provides useful out-of-the-box functionalities like the sorting and filtering of the data in the container. The use of existing modules gives the opportunity to developers to concentrate in the production of the combined reports and the middle layer.

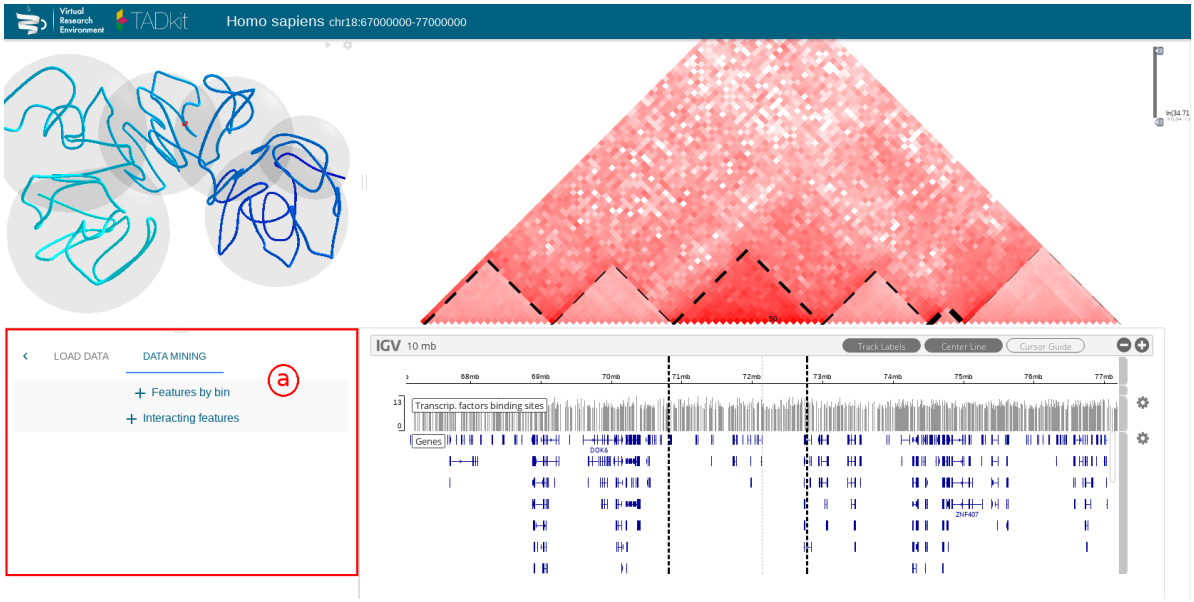

Figure 1: (a) Data mining tab.

2 Analysis reports

2.1 Features by bin

The Features by bin report combines information from tracks in the 1D panel with the resolution of the current dataset [***Figure***]. The user can select any of the tracks loaded in TADkit to retrieve the features present per bin of genomic size equal to the resolution. Each bin of the interaction matrix and bead of the 3D model correspond to a genomic size equal to the resolution. Therefore, the features by bin relates lineal genomic position of features in 1D with the beads in which they lay in 3D.

Diverse information of the features is queried from the track: id, name, genomic location start and end, strand direction and value. As mentioned in the introduction there exists the possibility of sorting and filtering the data using the header of the container.

The filtered and sorted information table can be exported to a Tab-Separated Values (TSV) file ready to be eventually used in other external analysis.

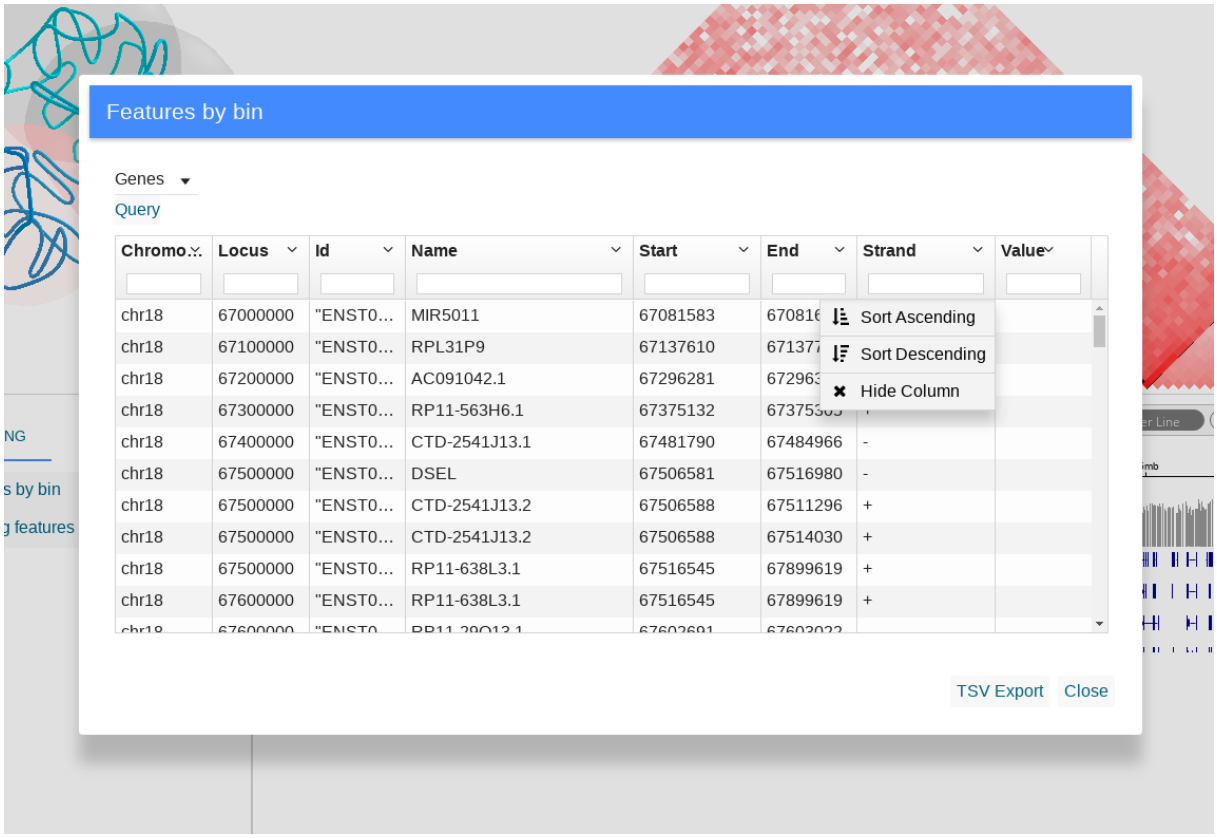

Figure 2: Features by bin report

2.2 Interacting features

The Interacting features report relates information of the interaction matrix in the 2D panel with information of the tracks in the 1D panel [***Figure***]. The user can query for features located in different loci whose interaction frequency is higher than a certain threshold. The features could be in the same track or between two tracks, for instance a promoter vs enhancer tracks. Due to the inherent high interaction frequencies of close neighbor loci, another filter has been implemented to exclude loci that are closer than a defined genomic distance.

This analysis tool addresses the question of relating functional features that might be interacting by proximity in 3D space in the genome.

As in the previous report the filtered and sorted information table can be exported to a Tab-Separated Values (TSV) file.

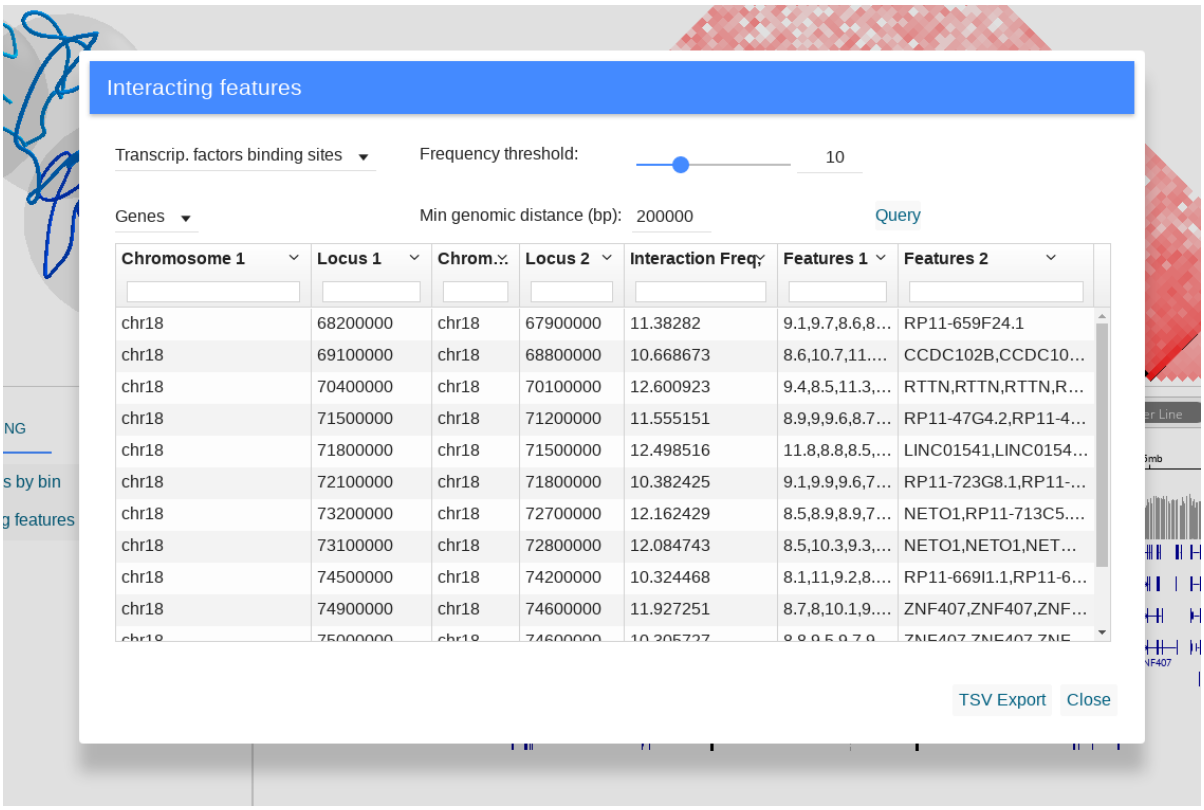

Figure 3: Interacting features report

3 Future perspective

The first prototype of the analysis tools includes two reports that combine data sources with different dimensionality to provide richer information of the genomic region into study. The VRE should benefit from TADkit as a point of convergence of these different types of information sources and provide new types of reports based on the needs and ideas of users and pilot projects.

The potential of the data mining tools will be substantially enhanced with the connection of TADkit with VRE data sources accessible via web services. At that point TADkit will no longer be limited to the information of the initially loaded tracks and datasets but will be able to access a wider range of online information.

The abovementioned additional features and enhancements will be reported in subsequent deliverables (WP4) and as an update of WP3 results in the 2nd periodic report due at the end of the project.