TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data.

Tiago Chedraoui Silva[1,2], Antonio Colaprico[3,4], Catharina Olsen[3,4], Gianluca Bontempi[3,4], Michele Ceccarelli[5,6], Benjamin P. Berman[2], Houtan Noushmehr[1,7*]

[1]Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil
[2]Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[3]Interuniversity Institute of Bioinformatics in Brussels (IB)2, Brussels, Belgium
[4]Machine Learning Group, ULB, Brussels, Belgium
[5]Bioinformatics Laboratory, BIOGEM, Ariano Irpino, Avellino, Italy
[6]Department of Science and Technology, University of Sannio, Benevento, Italy
[7]Department of Neurosurgery, Henry Ford Hospital, Detroit, MI, USA

*To whom correspondence should be addressed.

ABSTRACT

Summary: The GDC (Genomic Data Commons) data portal provides users with data from cancer genomics studies. Recently, we developed the R/Bioconductor TCGAbiolinks package, which allows users to search, download and prepare cancer genomics data for integrative data analysis. The use of this package requires users to have advanced knowledge of R thus limiting the number of users. To overcome this obstacle and improve the accessibility of the package by a wider range of users, we developed a graphical user interface (GUI) available through the R/Bioconductor package TCGAbiolinksGUI. Availability: The TCGAbiolinksGUI package is freely available within the Bioconductor project at http://bioconductor.org/packages/TCGAbiolinksGUI/.And a demo version is available at https://tcgabiolinksgui.shinyapps.io/demo/ and a docker image is available at https://hub.docker.com/r/tiagochst/tcgabiolinksgui/.

Contact: hnoushm1@hfhs.org

Supplementary information: Supplementary information is available at http://bit.do/TCGAbiolinksDocs

INTRODUCTION

The National Cancer Institute’s (NCI) Genomic Data Commons (GDC), a data sharing platform that promotes precision medicine in oncology, provides a rich resource of molecular and clinical data of almost 13,000 tumor patient samples across 38 different cancer types and subtypes. The platform includes data from The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET). The data, which is publicly available, have been utilized by researchers to make novel discoveries and/or validate important findings. To enhance these findings, several important bioinformatics tools to harness genomics cancer data were developed, many of them belonging to the Bioconductor project (Gentleman et al., 2004). Among those tools is our tool TCGAbiolinks (Colaprico et al., 2016), which was developed to facilitate the analysis of TCGA data by incorporating the query, download and processing steps within the Bioconductor project (Gentleman et al., 2004). This tool allows users to integrate TCGA data with Bioconductor packages thus harnessing a wealth of statistical methodologies for biologically derived data.

In addition, it provides integrative methodologies to perform several important downstream analyses, such as DNA methylation and Gene expression integration. A full detailed comparison between TCGAbiolinks and other bioinformatics tools to analyze TCGA data was previously detailed in our report in which we highlight key advantages of using TCGAbiolinks (Colaprico et al., 2016). Although TCGAbiolinks is a suitable R package for most data analysts with a strong knowledge and familiarity with R specifically those who can comfortably write strings of common R commands, we developed TCGAbiolinksGUI to enable user access to the methodologies offered in TCGAbiolinks and to give users the flexibility of point-and-click style analysis without the need to enter specific arguments. TCGAbiolinksGUI takes in all the important features of TCGAbiolinks and offers a graphics user interface (GUI) thereby eliminating any need to familiarize TCGAbiolinks' key functions and arguments. Tutorials via online documents and YouTube video instructions will assist end-users in taking full advantage of TCGAbiolinks. Here we present TCGAbiolinksGUI an R/Bioconductor package which uses the R web application framework shiny (Chang et al. (2016) to provide a GUI to process, query, download, and perform integrative analyses of TCGA data.

DESCRIPTION

The user interface of TCGAbiolinksGUI has been divided into three main types of menus. The first type defines the acquisition of GDC data. The second defines the analysis steps which were subdivided according to the molecular data types. And the third one is dedicated to harnessing integrative analyses. We present below a brief description of each menu and their features that can be accessed through a side panel (see figure 1):

- **GDC Data:** Provides a guided approach to search for published subtype information, clinical and molecular data. In addition, it downloads and processes the molecular data into an R object that can be used for further analysis.
- **Clinical analysis:** Performs survival analysis to quantify and test survival differences between two or more groups of patients.
- **Epigenetic analysis:** Performs a Differential Methylation Regions (DMR) analysis, visualizes the results through both volcano and heatmap plots, and visualizes the mean DNA methylation level by groups.
- **Transcriptomic analysis:** Performs a Differential Expression Analysis (DEA), and visualizes the results through both volcano and heatmap plots. For the genes found as upregulated or downregulated an enrichment analysis can be performed and pathway data can be integrated (Luo and Brouwer, 2013).
- **Genomic analysis:** Visualize the mutations from MAF files (mutation annotation information file) through an oncoprint plot (Gu et al., 2016).
- **Integrative analysis:** Integrate the DMR and DEA results through a starburst plot. Also, using the DNA methylation data and the gene expression data the R/Bioconductor

ELMER package can be used to discover functionally relevant genomic regions associated with cancer (Yao et al., 2015).

We designed the GUI of TCGAbiolinksto control parameters of the analysis and visualization functions and to export the results from the analysis as a csv spreadsheet. Moreover, we provide a guided tutorial for users via a vignette document which details each step and menu function (see supplementary data).To further simplify the usability of our tool, we provide a docker image compatible with most popular operating system. This file allows users to run TCGAbiolinksGUI without the need to install associated dependencies or configure system files, common steps required to run R installations and load R/Bioconductor packages such as TCGAbiolinksGUI. We expect TCGAbiolinksGUI to be used by clinicians, experimental biologists and novice to inexperienced computational biologists with limited R experience, while our initial package (TCGAbiolinks) would be most suited for advanced bioinformaticians.

ACKNOWLEDGMENTS

We are grateful to the OMICs lab and the GDC team for suggestions in the design of TCGAbiolinksGUI interface. We are also grateful for Susan MacPhee for critical review of the manuscript and vignettes.

FUNDING

This work has been supported by a grant from Henry Ford Hospital (H.N.) and by the São Paulo Research Foundation (FAPESP) (2016/01389-7 to T.C.S. & H.N. and 2015/07925-5 to H.N.) the BridgeIRIS project, funded by INNOVIRIS, Region de Bruxelles Capitale, Brussels, Belgium, and by GENomic profiling of Gastrointestinal Inflammatory-Sensitive CANcers (GENGISCAN), Belgian FNRS PDR (T100914F to A.C., C.O. & G.B.). T.C.S. and B.P.B. were supported by the NCI Informatics Technology for Cancer Research program, NIH/NCI grant 1U01CA184826.

REFERENCES

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2016). shiny: Web Application Framework for R. *R package versio*n 0.14.

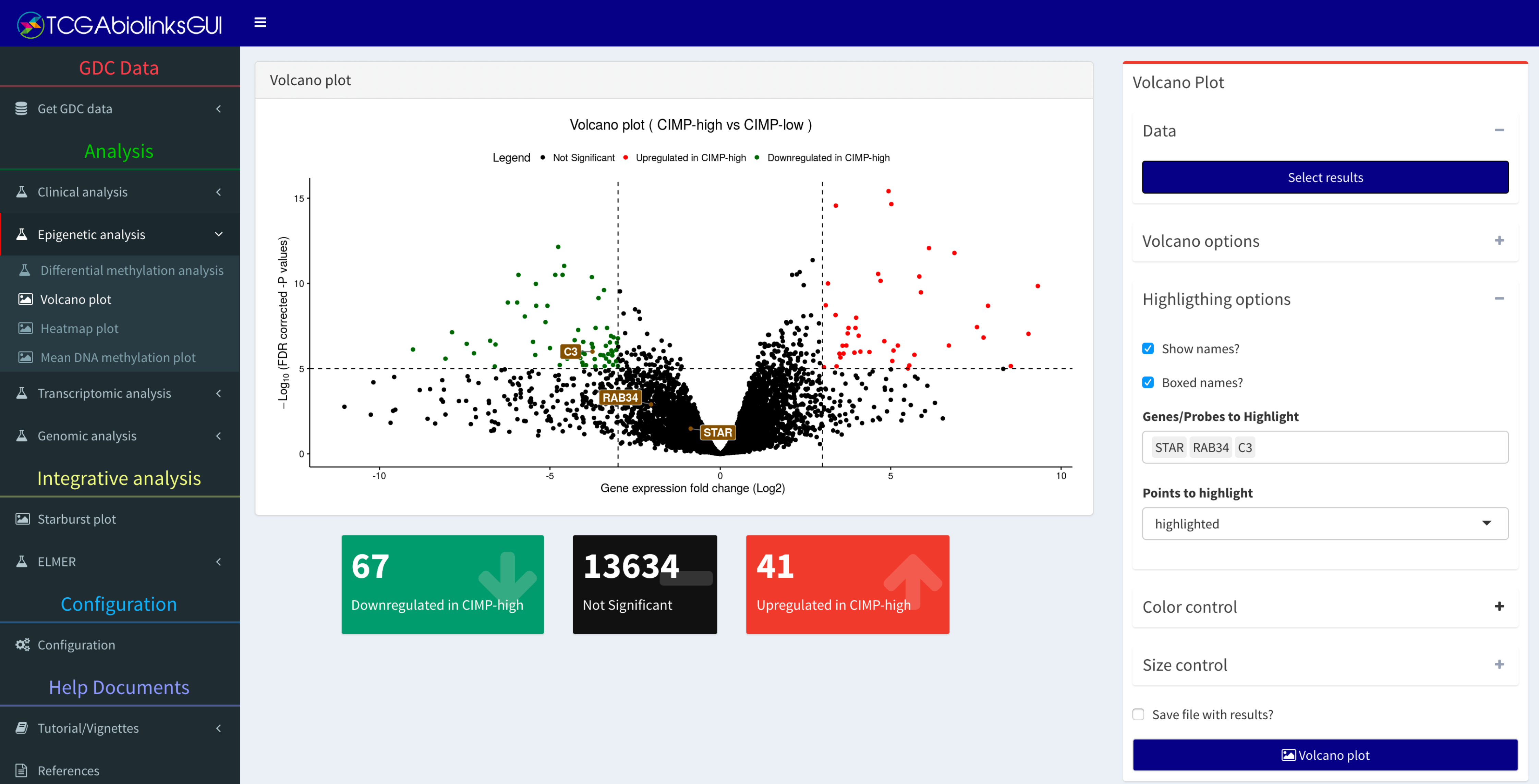
Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., and Noushmehr, H. (2016). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, 44(8), e71.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), R80.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.

Luo, W. and Brouwer, C. (2013). Pathview: an r/bioconductor package for pathway- based data integration and visualization. *Bioinformatics*, 29(14), 1830–1831.

Yao, L., Shen, H., Laird, P., Farnham, P., and Berman, B. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome biology*, 16(1), 105–105.



67
Downregulated in CIMP-high

13634
Not Significant

41
Upregulated in CIMP-high

☑ Show names?

☑ Boxed names?

Genes/Probes to Highlight
STAR RAB34 C3

Points to highlight
highlighted

□ Save file with results?