

SOFTWARE TOOL ARTICLE

TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data [version 1; peer review: 1 approved, 1 approved with reservations]

Tiago Chedraoui Silva^{1,2}, Antonio Colaprico^{3,4}, Catharina Olsen^{3,4},
 Tathiane M Malta^{1,5}, Gianluca Bontempi^{3,4}, Michele Ceccarelli^{6,7},
 Benjamin P Berman^{2,8}, Houtan Noushmehr  ^{1,5}

¹Department of Genetics, Ribeirao Preto Medical School, University of Sao Paulo, Ribeirao Preto, Brazil

²Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA

³Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium

⁴Machine Learning Group (MLG), University Libre de Bruxelles, Brussels, Belgium

⁵Department of Neurosurgery, Henry Ford Hospital, Detroit, Detroit, MI, USA

⁶Department of Science and Technology, University of Sannio, Benevento, Italy

⁷Bioinformatics Laboratory, BioGeM, Ariano Irpino, Italy

⁸Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, USA

v1 **First published:** 10 Apr 2018, 7:439 (<https://doi.org/10.12688/f1000research.14197.1>)
Latest published: 10 Apr 2018, 7:439 (<https://doi.org/10.12688/f1000research.14197.1>)

Abstract

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) using Shiny available through the package *TCGAbiolinksGUI*. The *TCGAbiolinksGUI* package is freely available within the Bioconductor project at <http://bioconductor.org/packages/TCGAbiolinksGUI/>. Links to the GitHub repository, a demo version of the tool, a docker image and PDF/video tutorials are available from the *TCGAbiolinksGUI* site.

Keywords

TCGA, cancer, genomics, epigenomics, bioinformatics

Open Peer Review

Reviewer Status  

	Invited Reviewers	
	1	2
version 1		
published 10 Apr 2018	report	report

- Zuguang Gu**, German Cancer Research Center (DKFZ), Heidelberg, Germany
- Michael Lawrence**, Genentech, South San Francisco, USA

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **RPackage** gateway.

This article is included in the **International Society**



for **Computational Biology Community Journal**
gateway.



This article is included in the **Bioconductor**
gateway.

Corresponding authors: Tiago Chedraoui Silva (tiagochst@gmail.com), Houtan Noushmehr (houtana@gmail.com)

Author roles: **Silva TC:** Data Curation, Formal Analysis, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Colaprico A:** Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Olsen C:** Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Malta TM:** Data Curation, Formal Analysis; **Bontempi G:** Funding Acquisition, Resources; **Ceccarelli M:** Data Curation, Methodology, Resources, Visualization; **Berman BP:** Formal Analysis, Funding Acquisition, Methodology, Resources, Software, Visualization; **Noushmehr H:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: 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 and Genomic Data Analysis Network NIH/NCI grant 1U24CA210969

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Silva TC *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Silva TC, Colaprico A, Olsen C *et al.* **TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2018, 7:439 (<https://doi.org/10.12688/f1000research.14197.1>)

First published: 10 Apr 2018, 7:439 (<https://doi.org/10.12688/f1000research.14197.1>)

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. As of 2018, almost 13,000 tumor patient samples across 38 different cancer types and subtypes are freely available for download and analysis. Currently, the platform includes data from The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), with the expectation that many other cancer genomic repositories to be incorporated into GDC over the next few years. The publicly available data have been utilized by researchers for novel discoveries and/or validate important findings related to tumorigenesis, improvements in treatment diagnosis and refinement of tumor classifications. To enhance these findings, several important bioinformatics tools to harness genomics cancer data were developed, many of them belonging to the Bioconductor project¹.

TCGAbiolinks², an R/Bioconductor package, was developed to facilitate the analysis of cancer genomics data by incorporating the query, download and processing steps directly from GDC. This tool allows users to advance their data analysis of cancer genomics by harnessing additional Bioconductor packages thereby allowing users access to a wealth of statistical methodologies. In addition, it can perform integrative data analysis across different types of experimental data types, such as DNA methylation and Gene expression data. A detailed comparison between TCGAbiolinks and other bioinformatics tools to analyze cancer genomics data was previously described². 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 be familiar with TCGAbiolinks' key functions and arguments. In addition, we added new functions to import users' own raw data for further integrative analysis with GDC data. Tutorials via online documents and YouTube video instructions are available from the website to assist end-users in taking full advantage of TCGAbiolinks.

Here we present TCGAbiolinksGUI, an R/Bioconductor package which uses the R web application framework Shiny³ to provide a GUI to process, query, download, and perform integrative analyses of GDC data.

Implementation

Infrastructure

The TCGAbiolinksGUI was created using Shiny, a Web Application Framework for R. TCGAbiolinksGUI incorporates several packages, that provide advanced features to enhance Shiny apps, such as shinyjs to add JavaScript actions⁴, shinydashboard to add dashboards⁵ and shinyFiles⁶ to provide access to the server file system. The following R/Bioconductor packages are used as back-ends for the data retrieval and analysis:

- TCGAbiolinks² which allows to search, download and prepare data from the NCI's Genomic Data Commons (GDC) data portal into an R object and perform several downstream analysis;
- ELMER (Enhancer Linking by Methylation/Expression Relationship)^{7,8} which identifies DNA methylation changes in distal regulatory regions and correlate these signatures with the expression of nearby genes to identify transcriptional targets associated with cancer;
- ComplexHeatmap⁹ to visualize data as oncoprint and heatmaps;
- pathview¹⁰ which offers pathway based data integration and visualization;
- maftools¹¹ to analyze, visualize and summarize genomics MAF (Mutation Annotation Format) files.

Graphical user interface design

The user interface has been divided into three main GUI menus. The first menu defines the acquisition of GDC data. The second, the 'Analysis' menu, is subdivided according to the molecular data types. And the third is dedicated to harnessing integrative analyses. Each menu is described below (see [Figure 1](#)):

- **Data:** Provides a guided approach to search for published molecular subtype information, clinical and molecular data available in GDC. In addition, it downloads and processes the molecular data into an R object that can be used for further analysis. For raw DNA methylation data obtained in the form of Intensity Data (IDAT) files, we provide a pipeline using the R/Bioconductor [minfi package](#) to prepare the data for subsequent bioinformatics analysis¹² performing a background and dye-bias correction with the

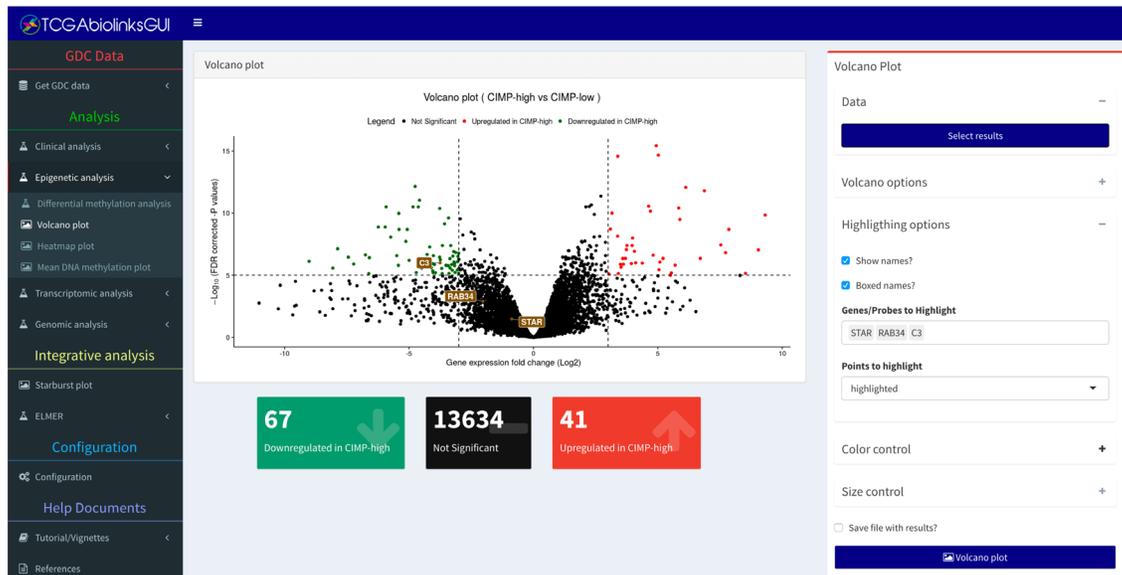


Figure 1. The volcano plot menu of TCGAbiolinksGUI. The panel on the left shows the menu divided into different analyses, the panel on the right shows the controls available for the selected menu. In the center is a volcano plot window from the analysis menu. It is possible to control the colors, to change cut-offs, to export results into a CSV document and to export the plot.

`preprocessnoob` function followed by a detection P-value quality masking (sample-specific)¹³ and probes overlapping repeats or single nucleotide polymorphisms masking (non-sample specific)¹⁴ (Figure 3).

- **Clinical analysis:** Performs survival analysis to quantify and test survival differences between two or more groups of patients and draws survival curves with the 'number at risk' table, the cumulative number of events table and the cumulative number of censored subjects table using the R/CRAN package `survminer`¹⁵.
- **Epigenetic analysis:** Performs a differentially methylated regions (DMR) analysis, visualizes the results through both volcano and heatmap plots, and visualizes the mean DNA methylation level by groups or subtypes. For certain tumor types like Glioma, we have added a function to classify non TCGA derived DNA methylation data into one of the 7 published epigenomic subtypes¹⁶ using a RandomForest (RF) trained model derived from DNA methylation signatures available from the [Cancer Genome Atlas](#) (Figure 4). Description of how the RF models were created can be found in [TCGAbiolinksGUI.data vignette](#).
- **Transcriptomic analysis:** Performs a differential expression analysis (DEA), and visualizes the results as either volcano or heatmap plots. Pathway analysis can be performed on a list of differentially expressed genes¹⁰.
- **Genomic analysis:** Visualize and summarize the mutations from MAF (Mutation Annotation Format) files through summary plots and oncoplots using the R/Bioconductor `maftools` package^{9,11} (Figure 2 and Figure 6).
- **Integrative analysis:** Integrate the DMR and DEA results through a starburst plot. Integrate clinical and mutation data by way of a Kaplan-Meier survival analysis for groups of mutated samples vs non-mutated for a given gene (Figure 5). DNA methylation and gene expression data can be further analyzed using the R/Bioconductor `ELMER` package to discover functionally relevant genomic regions associated with cancer^{7,8}.

Documentation

We provide a guided tutorial for users via an [online vignette document](#) which details each step and menu function. [Printable PDF](#) and YouTube video instructions (http://bit.ly/TCGAbiolinksGUI_videoTutorials) are provided to help users utilize TCGAbiolinksGUI. A demonstration version of the tool is available at [TCGAbiolinksGUI](#). To help improve and expand our tool over time, users are encouraged to report and file bug reports or feature requests via our [GitHub repository](#).



Figure 2. Visualizing mutation summary. This maftools plot shows a summary of the MAF file. Highlighting the most mutated genes, SNV class, and variant classification distributions within a tumor type.

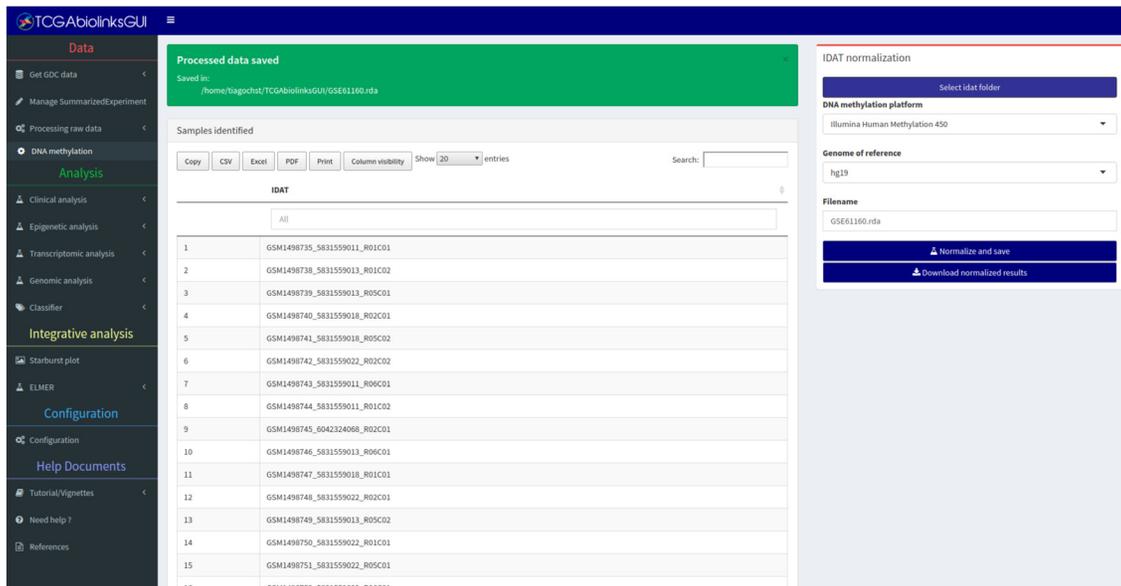


Figure 3. IDAT normalization. Table lists all files which will be processed. Data retrieved from GEO (accession GSE61160).

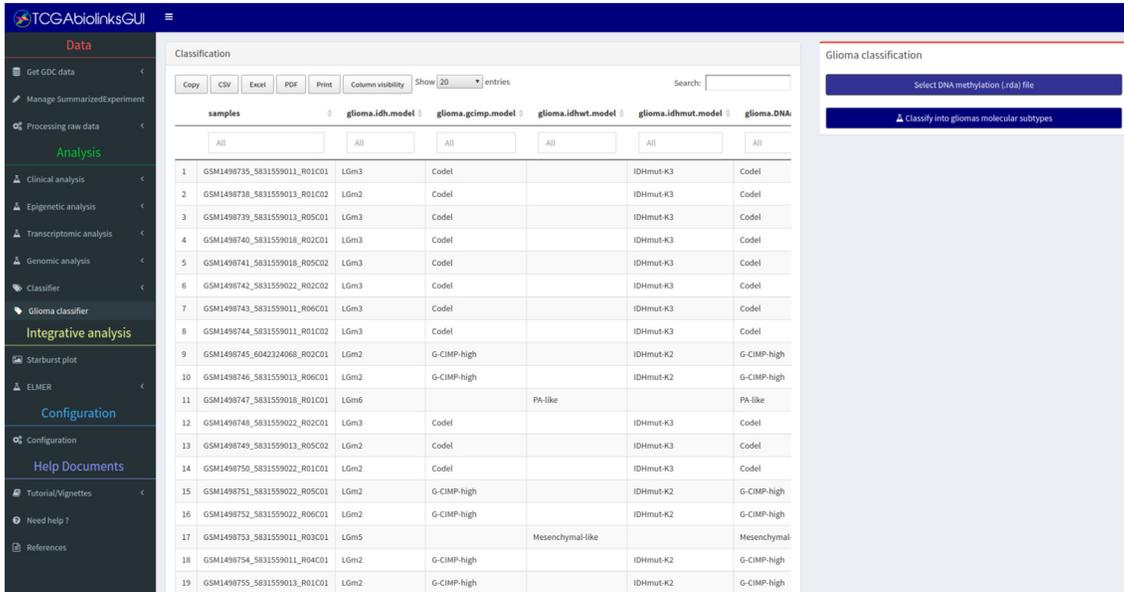


Figure 4. Glioma classifier. Predicting glioma epigenomic molecular subtypes based on DNA methylation using data from GEO (accession [GSE61160](#)).

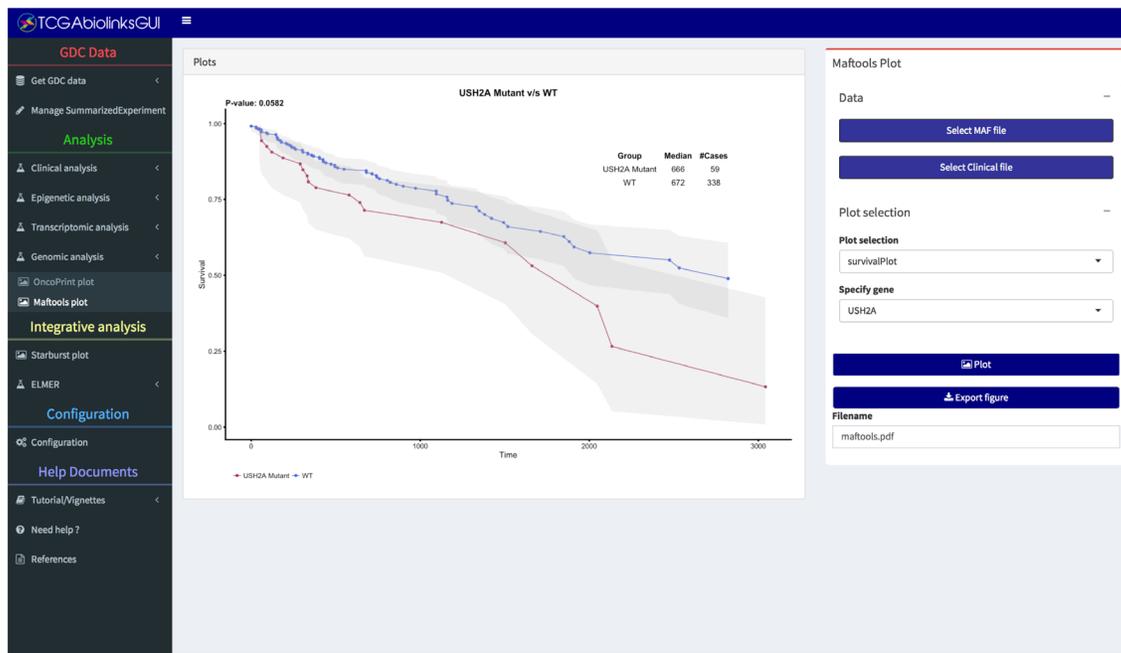


Figure 5. Integrating clinical data and mutation data. Performing Kaplan-Meier survival analysis for groups of samples with a mutation in gene USH2A vs WT.

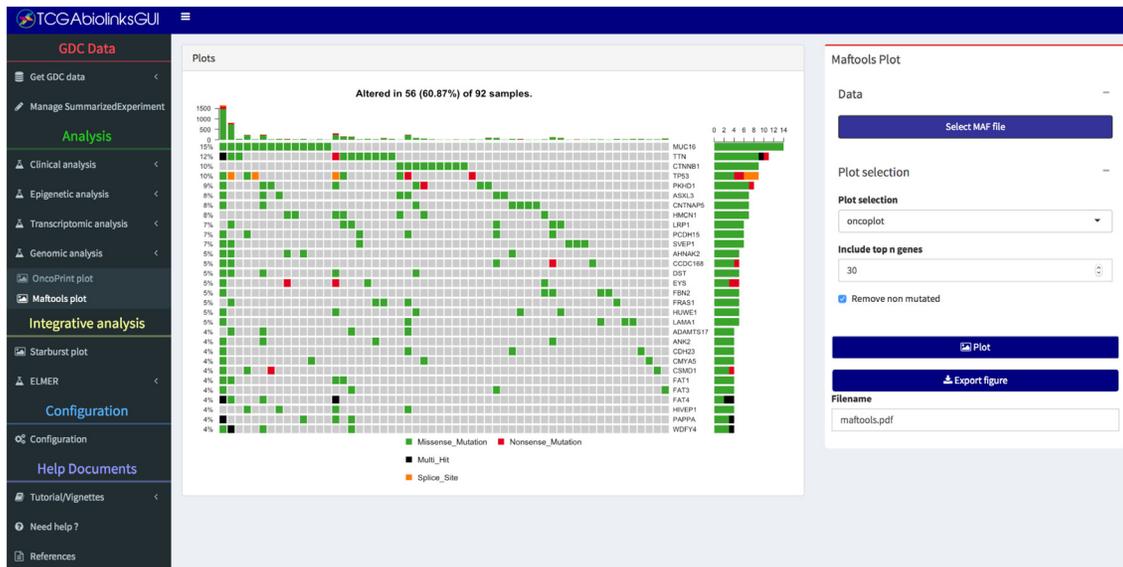


Figure 6. Visualizing mutation as an oncoplot. Each column represents a sample and each row a different gene. The top barplot has the frequency of mutations for each patient, while the right barplot has the frequency of mutations in each gene. By default, samples ordered by the most mutated genes.

Docker container

We recognize that one of the possible drawbacks of using our tool is the arduous process of installing the R/Bioconductor environment and all of the required dependencies. One solution would be to host the tool on a server, however the high demand for space, computational processing, and access for multiple users would make this approach financially challenging; instead, we encourage users to use it on their own computers. However, to further simplify the usability and accessibility of our tool, we provide a Docker image file that contains the complete R/Bioconductor environment configured to use TCGAbiolinksGUI. This file is compatible with most popular operating systems and it is available [online](#). This image can be easily downloaded and deployed through the [Kitematic tool](#), a simple application for managing Docker containers for Mac, Linux, and Windows. A detailed documentation of how to obtain and use the docker image via the Kitematic application is available at [TCGAbiolinksGUI help document](#). The Docker image will be updated to coincide with any regular updates of TCGAbiolinksGUI. We believe this will provide several types of access for end-users interested in analyzing cancer genomics data stored at GDC.

Operation

The package can run on any platform with a R version ≥ 3.4 or higher and Bioconductor version ≥ 3.6 and higher.

Results and discussion

To provide end-users with insights and application of TCGAbiolinksGUI; 1) we compare TCGAbiolinksGUI with other published bioinformatics tools and 2) we provide a use-case that allows users a step-bystep guide to analyzing their own cancer molecular data alongside TCGA data.

Comparison of alternative software

Web tools used for cancer data analysis might be classified into two broad groups. The first group only provides an interface to existing software analysis tools. [The Galaxy project](#), which is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research, is an example of such a tool that belongs to this group. The other group is composed of exploratory tools mainly focused on the visualization of processed data and pre-computed results. The [cBioPortal project](#)^{17,18}, by providing several visualizations for mining the TCGA data, is an example of a tool that falls within this category.

If one were to classify TCGAbiolinksGUI, it would belong to the first group. Compared to the Galaxy project, TCGAbiolinksGUI offers an open platform which improves the accessibility of R/Bioconductor packages,

allowing users an advantage to integrate their features with existing Bioconductor packages. Unlike the Galaxy project, which requires the interface elements to be structured through XML files¹⁹, TCGAbiolinksGUI can resolve this simply because it was built within the R/Shiny framework. cBioPortal and TCGAbiolinksGUI provide users with access to raw and processed data, however TCGAbiolinksGUI allows users to perform in-depth integrative analysis, a functionality which cBioPortal currently lacks. For example, if a user is interested in defining differentially expressed genes or DNA methylation events between two populations of tumors (i.e. FOXA1 mutants and FOXA1 wildtypes), by using cBioPortal, a user would have to download the gene expression, DNA methylation and mutation data, define the samples per group, and import the data into their favorite statistical tool to identify their list of differentially expressed or methylated genes. Whereas, with TCGAbiolinksGUI, we developed the platform so that the user can define the sample groups based on their mutation spectrum, perform supervised analysis that can then define differentially expressed or methylated gene list and this can be directly ported into pathway analysis. In addition, if the user is interested in observing survival differences between the groups, this can also be done within TCGAbiolinksGUI and thereby reducing the need to exit a specific data platform or having to transform the downloaded data to fit some other statistical platform to achieve the same goals.

Use case

In order to illustrate an integrative analysis using TCGAbiolinksGUI, we provide a use case available at <https://bioinformaticsfmrp.github.io/Bioc2017.TCGAbiolinks.ELMER/index.html>. This use case highlights a step by step guide for one to perform an integrative analysis using TCGA-LUSC (Lung Squamous Cell Carcinoma) data retrieved directly from GDC server²⁰.

Conclusion

TCGAbiolinksGUI was developed to provide a user-friendly interface of our TCGAbiolinks package. TCGAbiolinksGUI is designed specifically for the least experienced R user to import GDC data and perform R/Bioconductor analysis as well as for the most experienced R user, who could execute several of the R/Bioconductor functions without the need to write several lines of R code. For the R/Bioconductor developers, the package has an extensible design feature that allows users 1) to add new features by modifying a few lines of the main code, 2) to add a file with user interface elements on the client side, and 3) add a file with their control on the server side.

Also, TCGAbiolinksGUI supports the most updated R/Bioconductor data structures (i.e. SummarizedExperiment and MultiAssayExperiment) which allow handling data and metadata into one single object and validates several integrity requirements. Thereby, TCGAbiolinksGUI package allows data handling to be as efficient as possible and thereby limits and avoids user errors in data manipulation such as sample removal that involves also metadata deletion.

Finally, several efforts to understand genomic and epigenomic alterations associated with tumor development has been made over the last few years, which presents several bioinformatics challenges, such as data retrieval and integration with clinical data and other molecular data types. By creating a graphical interface to tools like TCGAbiolinks whose relevance is seen in various articles²¹⁻²³, this package will allow end-users to facilitate the mining of cancer data deposited in GDC, in hopes to aid in analyzing and discovering new functional genomic elements and potential therapeutic targets for cancer.

Data and software availability

TCGAbiolinksGUI is a platform independent R package ($R \geq 3.4$) available at: <https://doi.org/10.18129/B9.bioc.TCGAbiolinksGUI>.

Source code *TCGAbiolinksGUI* is available at: <https://github.com/BioinformaticsFMRP/TCGAbiolinksGUI>.

License: GNU General Public License version 3 (GNU GPL3)

Complementary data required to execute the package is available at: <https://github.com/BioinformaticsFMRP/TCGAbiolinksGUI.data> or at <https://doi.org/doi:10.18129/B9.bioc.TCGAbiolinksGUI.data>²⁴.

Software documentation is available at: <https://bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinksGUI/inst/doc/index.html>

Detailed steps of the use case are available at: <https://bioinformaticsfmrp.github.io/Bioc2017.TCGAbiolinks.ELMER/index.html>.

Software installation

To install the stable version from the Bioconductor repository <http://bioconductor.org/packages/TCGAbiolinksGUI/> please use the following code.

```
source("https://bioconductor.org/biocLite.R")
biocLite("TCGAbiolinksGUI", dependencies = TRUE)
```

And to install the development version of the package via GitHub:

```
source("https://bioconductor.org/biocLite.R")
deps <- c("devtools")
for(pkg in deps) if (!pkg %in% installed.packages()) biocLite(pkg, dependencies = TRUE)
devtools::install_github("tiagochst/ELMER.data")
devtools::install_github("tiagochst/ELMER")
devtools::install_github("BioinformaticsFMRP/TCGAbiolinksGUI.data", ref = "R_3.4")
devtools::install_github("BioinformaticsFMRP/TCGAbiolinksGUI")
```

This installation process has been tested on a Debian 9.1 machine (the following libraries had to be installed: *libpng-dev* and *libmariadb-client-lgpl-dev* (command: `sudo apt-get install libpng-dev libmariadbclient-lgpl-dev`).

Also, due to the number of libraries loaded we had to increase the maximum number of DLL R can load, for more information please check the vignette section "Increasing loaded DLL".

```
> sessionInfo()
R version 3.4.1 (2017-06-30)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 9 (stretch)

Matrix products: default
BLAS/LAPACK: /usr/lib/libopenblas-r0.2.19.so

locale:
 [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8          LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8      LC_MESSAGES=C
 [7] LC_PAPER=en_US.UTF-8         LC_NAME=C
 [9] LC_ADDRESS=C                  LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods   base

other attached packages:
[1] TCGAbiolinksGUI.data_0.99.4 shinydashboard_0.6.1
[3] BiocInstaller_1.28.0

loaded via a namespace (and not attached):
 [1] R.utils_2.6.0          tidyselect_0.2.3
 [3] RSQLite_2.0            AnnotationDbi_1.40.0
 [5] htmlwidgets_0.9        grid_3.4.1
 [7] BiocParallel_1.12.0    ELMER_2.2.7
 [9] devtools_1.13.4        DESeq_1.30.0
[11] munsell_0.4.3          codetools_0.2-15
[13] withr_2.1.1            colorspace_1.3-2
[15] Biobase_2.38.0         knitr_1.18
```

[17]	rstudioapi_0.7	stats4_3.4.1
[19]	robustbase_0.92-8	dimRed_0.1.0
[21]	git2r_0.20.0	GenomeInfoDbData_1.0.0
[23]	mnormt_1.5-5	hwriter_1.3.2
[25]	KMSurv_0.1-5	bit64_0.9-7
[27]	downloader_0.4	ipred_0.9-6
[29]	biovizBase_1.26.0	ggthemes_3.4.0
[31]	EDASeq_2.12.0	ELMER.data_2.2.2
[33]	R6_2.2.2	doParallel_1.0.11
[35]	GenomeInfoDb_1.14.0	locfit_1.5-9.1
[37]	DRR_0.0.2	AnnotationFilter_1.2.0
[39]	bitops_1.0-6	reshape_0.8.7
[41]	DelayedArray_0.4.1	assertthat_0.2.0
[43]	scales_0.5.0	nnet_7.3-12
[45]	gtable_0.2.0	ddalpha_1.3.1
[47]	sva_3.26.0	ensemldb_2.2.0
[49]	timeDate_3042.101	rlang_0.1.6
[51]	CVST_0.2-1	genefilter_1.60.0
[53]	cmprsk_2.2-7	RcppRoll_0.2.2
[55]	GlobalOptions_0.0.12	splines_3.4.1
[57]	rtracklayer_1.38.2	lazyeval_0.2.1
[59]	ModelMetrics_1.1.0	acepack_1.4.1
[61]	dichromat_2.0-0	selectr_0.3-1
[63]	broom_0.4.3	checkmate_1.8.5
[65]	yaml_2.1.16	reshape2_1.4.3
[67]	GenomicFeatures_1.30.0	backports_1.1.2
[69]	httpuv_1.3.5	Hmisc_4.1-1
[71]	RMySQL_0.10.13	caret_6.0-78
[73]	lava_1.5.1	tools_3.4.1
[75]	psych_1.7.8	ggplot2_2.2.1
[77]	RColorBrewer_1.1-2	BiocGenerics_0.24.0
[79]	MultiAssayExperiment_1.4.4	Rcpp_0.12.14
[81]	plyr_1.8.4	base64enc_0.1-3
[83]	progress_1.1.2	zlibbioc_1.24.0
[85]	purrr_0.2.4	RCurl_1.95-4.9
[87]	prettyunits_1.0.2	ggpubr_0.1.6
[89]	rpart_4.1-11	GetoptLong_0.1.6
[91]	sfsmisc_1.1-1	S4Vectors_0.16.0
[93]	zoo_1.8-0	SummarizedExperiment_1.8.1
[95]	ggrepel_0.7.0	cluster_2.0.6
[97]	magrittr_1.5	data.table_1.10.4-3
[99]	circlize_0.4.3	survminer_0.4.1
[101]	ProtGenerics_1.10.0	matrixStats_0.52.2
[103]	aroma.light_3.8.0	hms_0.4.0
[105]	mime_0.5	xtable_1.8-2
[107]	XML_3.98-1.9	IRanges_2.12.0
[109]	gridExtra_2.3	shape_1.4.3
[111]	compiler_3.4.1	biomaRt_2.34.1
[113]	tibble_1.4.1	R.oo_1.21.0
[115]	htmltools_0.3.6	mgcv_1.8-22
[117]	Formula_1.2-2	tidyr_0.7.2
[119]	geneplotter_1.56.0	lubridate_1.7.1
[121]	DBI_0.7	matlab_1.0.2
[123]	ComplexHeatmap_1.17.1	MASS_7.3-48
[125]	ShortRead_1.36.0	Matrix_1.2-12
[127]	readr_1.1.1	R.methodsS3_1.7.1
[129]	gower_0.1.2	parallel_3.4.1
[131]	Gviz_1.22.2	bindr_0.1

[133]	GenomicRanges_1.30.1	pkgconfig_2.0.1
[135]	km.ci_0.5-2	GenomicAlignments_1.14.1
[137]	foreign_0.8-69	plotly_4.7.1
[139]	recipes_0.1.1	xml2_1.1.1
[141]	foreach_1.4.4	annotate_1.56.1
[143]	XVector_0.18.0	prodlim_1.6.1
[145]	rvest_0.3.2	stringr_1.2.0
[147]	VariantAnnotation_1.24.2	digest_0.6.13
[149]	ConsensusClusterPlus_1.42.0	Biostrings_2.46.0
[151]	TCGAbiolinks_2.6.9	survMisc_0.5.4
[153]	htmlTable_1.11.1	edgeR_3.20.5
[155]	kernlab_0.9-25	curl_3.1
[157]	shiny_1.0.5	Rsamtools_1.30.0
[159]	rjson_0.2.15	nlme_3.1-131
[161]	jsonlite_1.5	bindrcpp_0.2
[163]	viridisLite_0.2.0	limma_3.34.5
[165]	BSgenome_1.46.0	pillar_1.0.1
[167]	lattice_0.20-35	DEoptimR_1.0-8
[169]	httr_1.3.1	survival_2.41-3
[171]	interactiveDisplayBase_1.16.0	glue_1.2.0
[173]	iterators_1.0.9	bit_1.1-12
[175]	class_7.3-14	stringi_1.1.6
[177]	blob_1.1.0	AnnotationHub_2.10.1
[179]	latticeExtra_0.6-28	memoise_1.1.0
[181]	dplyr_0.7.4	

Competing interests

No competing interests were disclosed.

Grant information

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 and Genomic Data Analysis Network NIH/NCI grant 1U24CA210969.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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.

References

- Gentleman RC, Carey VJ, Bates DM, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol.* 2004; 5(10): R80.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Colaprico A, Silva TC, Olsen C, *et al.*: **TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data.** *Nucleic Acids Res.* 2016; 44(8): e71.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chang W, Cheng J, Allaire JJ, *et al.*: **shiny: Web Application Framework for R.** R package version 0.14, 2016.
[Reference Source](#)
- Attali D: **shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds.** R package version 0.9.1, 2017.
[Reference Source](#)
- Chang W, Ribeiro BB: **shinydashboard: Create Dashboards with 'Shiny'.** R package version 0.6.1, 2017.
[Reference Source](#)
- Lin PT: **shinyFiles: A Server-Side File System Viewer for Shiny.** R package version 0.6.2, 2016.
[Reference Source](#)
- Yao L, Shen H, Laird PW, *et al.*: **Inferring regulatory element landscapes and transcription factor networks from cancer**

- methylomes.** *Genome Biol.* 2015; **16**(1): 105.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Silva TC, Coetzee SG, Yao L, *et al.*: **Enhancer linking by methylation/expression relationships with the r package elmer version 2.** *bioRxiv.* 2017.
[Reference Source](#)
 9. Gu Z, Eils R, Schlesner M: **Complex heatmaps reveal patterns and correlations in multidimensional genomic data.** *Bioinformatics.* 2016; **32**(18): 2847–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
 10. Weijun L, Brouwer C: **Pathview: an r/bioconductor package for pathway-based data integration and visualization.** *Bioinformatics.* 2013; **29**(14): 1830–1831.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 11. Mayakonda A, Koeffler PH: **Maftools: Efficient analysis, visualization and summarization of maf files from large-scale cohort based cancer studies.** *bioRxiv.* 2016.
[Publisher Full Text](#)
 12. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al.*: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of infinium DNA methylation microarrays.** *Bioinformatics.* 2014; **30**(10): 1363–1369.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 13. Morris TJ, Beck S: **Analysis pipelines and packages for infinium humanmethylation450 beadchip (450k) data.** *Methods.* 2015; **72**: 3–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 14. Zhou W, Laird PW, Shen H: **Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes.** *Nucleic Acids Res.* 2017; **45**(4): e22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 15. Kassambara A, Kosinski M: **survminer: Drawing Survival Curves using 'ggplot2'.** R package version 0.4.0, 2017.
[Reference Source](#)
 16. Ceccarelli M, Barthel FP, Malta TM, *et al.*: **Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.** *Cell.* 2016; **164**(3): 550–563.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 17. Gao J, Aksoy BA, Dogrusoz U, *et al.*: **Integrative analysis of complex cancer genomics and clinical profiles using the cBioportal.** *Sci Signal.* 2013; **6**(269): pl1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 18. Cerami E, Gao J, Dogrusoz U, *et al.*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov.* 2012; **2**(5): 401–404.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 19. Turaga N, Freeberg MA, Baker D, *et al.*: **A guide and best practices for r/bioconductor tool integration in galaxy [version 1; referees: 1 approved, 1 approved with reservations].** *F1000Res.* 2016; **5**: 2757.
[Publisher Full Text](#)
 20. Grossman RL, Heath AP, Ferretti V, *et al.*: **Toward a Shared Vision for Cancer Genomic Data.** *N Engl J Med.* 2016; **375**(12): 1109–1112.
[PubMed Abstract](#) | [Publisher Full Text](#)
 21. Broutier L, Mastrogianni G, Versteegen MM, *et al.*: **Human primary liver cancer-derived organoid cultures for disease modeling and drug screening.** *Nat Med.* 2017; **23**(12): 1424–1435.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 22. Ghassemi S, Vejdovszky K, Sahin E, *et al.*: **Fgf5 is expressed in melanoma and enhances malignancy in vitro and in vivo.** *Oncotarget.* 2017; **8**(50): 87750–87762.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 23. Letellier E, Schmitz M, Ginolhac A, *et al.*: **Loss of myosin vb in colorectal cancer is a strong prognostic factor for disease recurrence.** *Br J Cancer.* 2017; **117**(11): 1689–1701.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 24. Silva TC, Colaprico A, Olsen C, *et al.*: **TCGAbiolinksGUI: A Graphical User Interface to analyze cancer molecular and clinical data.** *bioRxiv.* Bioinformatics - Submitted for review. 2017.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 25 April 2018

<https://doi.org/10.5256/f1000research.15443.r33000>

© 2018 Lawrence M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Michael Lawrence

Department of Bioinformatics, Genentech, South San Francisco, CA, USA

The major concern is that this seems like a GUI that is doing a lot of different things and thus should be more modular. The authors emphasize the connection to GDC, but many of the features are applicable to any similar dataset. Ideally, each panel would be its own Shiny module (or maybe even a separate package entirely), and expert users and app developers could select specific modules and integrate them with custom modules to create new applications. There is mention of extensibility in the text. Does that rely on the Shiny module system? It would be helpful for the reader to know that if true. The authors actually compare their tool to Galaxy, a highly modular system. GUIs have a tendency to be monolithic; we should resist that in Bioconductor. This package has 181 total dependencies!

Another point is that GUIs for exploratory data analysis should not only be useful for novices in a programming language. There are times when a GUI is more convenient than programming, even for an expert programmer. A GUI provides an alternative interface to the command line, thus opening the underlying functionality to other use cases, whether a bench biologist desperate for a way to see the data, or a computational biologist who wants to quickly explore the data visually while implementing a more sophisticated analysis.

Why not include the use case / workflow in the publication itself? It's useful to have an archive of that.

Some minor points:

- The abstract mentions a website but does not link to it (until maybe later)
- The use of the word "advanced" to describe the R users of TCGABiolinks is ambiguous. What exactly do you mean by "advanced"? Later on, advanced appears to be mean anyone who can write simple R code. Maybe just drop the ambiguous adjective?
- "Gene expression" do not capitalize "Gene"
- The phrase "specifically those who can comfortably write strings of common R commands" is awkward, especially since R does not really have "commands". Maybe say something like: "Although TCGABiolinks is accessible to data analysts who are familiar with R programming, ...", although I would phrase it more positively, rather than as deficiency of TCGABiolinks, which is just playing the role that it is meant to play in a larger framework.

- "graphics user interface" should be "graphical user interface"
- "Web Application Framework" - no need to capitalize

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 23 April 2018

<https://doi.org/10.5256/f1000research.15443.r33002>

© 2018 Gu Z. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Zuguang Gu

Division of Theoretical Bioinformatics (B080), German Cancer Research Center (DKFZ), Heidelberg, Germany

Silva et al. developed a new package TCGAbiolinksGUI which provides an easy way to query, download and analyze TCGA datasets. It would be a useful tool for bioinformaticians also non-bioinformaticians to process and get use of the massive TCGA datasets.

Besides taking care of downloading the TCGA datasets, TCGAbiolinksGUI package also provides basic functions/plots to process mutation/expression/methylation datasets. Additionally, TCGAbiolinksGUI performs integrative analysis of methylation and gene expression and does motif finding on the inferred regulatory network.

The TCGAbiolinksGUI has a very modern UI design, quite sophisticated programming and very friendly user interface. The structure of the analysis workflow is very clear. Besides that, it also has very detailed documentations even with videos. For the functionalities TCGAbiolinksGUI has already implemented, basically they are nice and I don't have major issues on it.

For some reason, I only tested TCGAbiolinksGUI with version 1.4.7 (the release version on Bioc at the time of reviewing this paper) and R version 3.4.4. I had some errors when testing some of the functionalities for which I think it should due to the lower version I was using and I would expect they should work OK with the development version.

Following are my minor comments:

1. There are some low-level errors which cause TCGAbiolinksGUI() function crashed and the webpage closed. E.g. when I tried to use "network inference" or "maftools plots", it gave error "no minet function/no read.maf function". I would guess it's mainly due to I was using the old version of this package. But it would be nice to capture these low-level errors also and print them in the web interface without stopping `TCGAbiolinksGUI()`. Also in DEA analysis, if group column is not set, `TCGAbiolinksGUI()` stops.
2. In many analysis where "group column" is needed, there are so many "clinical information fields" in the drop-down list. Is it possible to remove some of them? e.g. sample Ids or columns with too many missing values, or these numeric columns which I think they would never be used in group comparisons.
3. When a file is selected, there is no information on the web interface to tell users whether the file is selected or which file is selected.
4. For DEA analysis, if group levels which are used for comparison are forgot to provides, the error information is not informative ("Each group should have at least one sample")
5. When I do enrichment analysis for differentially expressed genes, I directly used the file generated at the "DEA analysis" step. However, it gave the error "no Gene_symbol column", but I checked there does have a "Gene_symbol" column (which is the first column in the file). I guess it might due to I was using the old version of the package.
6. When making heatmaps for different genes or DMRs, it is possible to put the grouping information which was used for comparison as default column annotation? I think it won't be too difficult because the group column and group levels are encoded in the file name of DEA or DMR file.
7. Is it possible to export the R code of making each plot (or performing each analysis)? Users can use these R scripts as template and customize later. E.g. since there is no option to configure the annotation colors, with the R script, users can adjust this part by themselves later.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, next generation sequencing, R packages development, visualization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research