Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Databases and ontologies

Databases and ontologies

RTCGA - The Family of R Packages Integrating Data from The Cancer Genome Atlas Study

Marcin Kosiński [1,2,*], Witold Chodor [2] and Przemysław Biecek [1,2]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland and
[2] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, City, 02-097 Warsaw, Poland.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: In this article we present a family of **R** packages called **RTCGA** that facilitate access to data from the TCGA project. The Cancer Genome Atlas Project (TCGA) is a coordinated effort to accelerate our understanding of the molecular basis of cancer. It is a source of curated multi-platform data, including RNA-seq, DNA-seq, DNA Methylation, together with clinical data for over 11 thousand patients and 33 cancer types. This rich source of data is accessible in raw format from TCGA Data Portal. The **RTCGA** packages facilitate access to this dataset, merge patients characteristics from different platforms and support typical statistical analyses. These packages will be useful for at least three audiences: biostatisticians that work with cancer data; researchers that are working on large scale algorithms; lecturers that are presenting data analysis method on real data problems.
Availability: RTCGA family of **R** packages is freely available at GitHub http://rtcga.github.io/RTCGA/ and from the Bioconductor project at http://bioconductor.org/packages/RTCGA/ .
Contact: m.p.kosinski@gmail.com

Motivation

The Cancer Genome Atlas Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA Project. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes [1]. Files are available through Firehose Broad GDAC portal [1]. One can select cancer type (cohort) and data type (e.g. clinical, RNA expression, methylation) and download a 'tar.gz' file with compressed data.

While working with many cancer types we may find this approach burdensome:

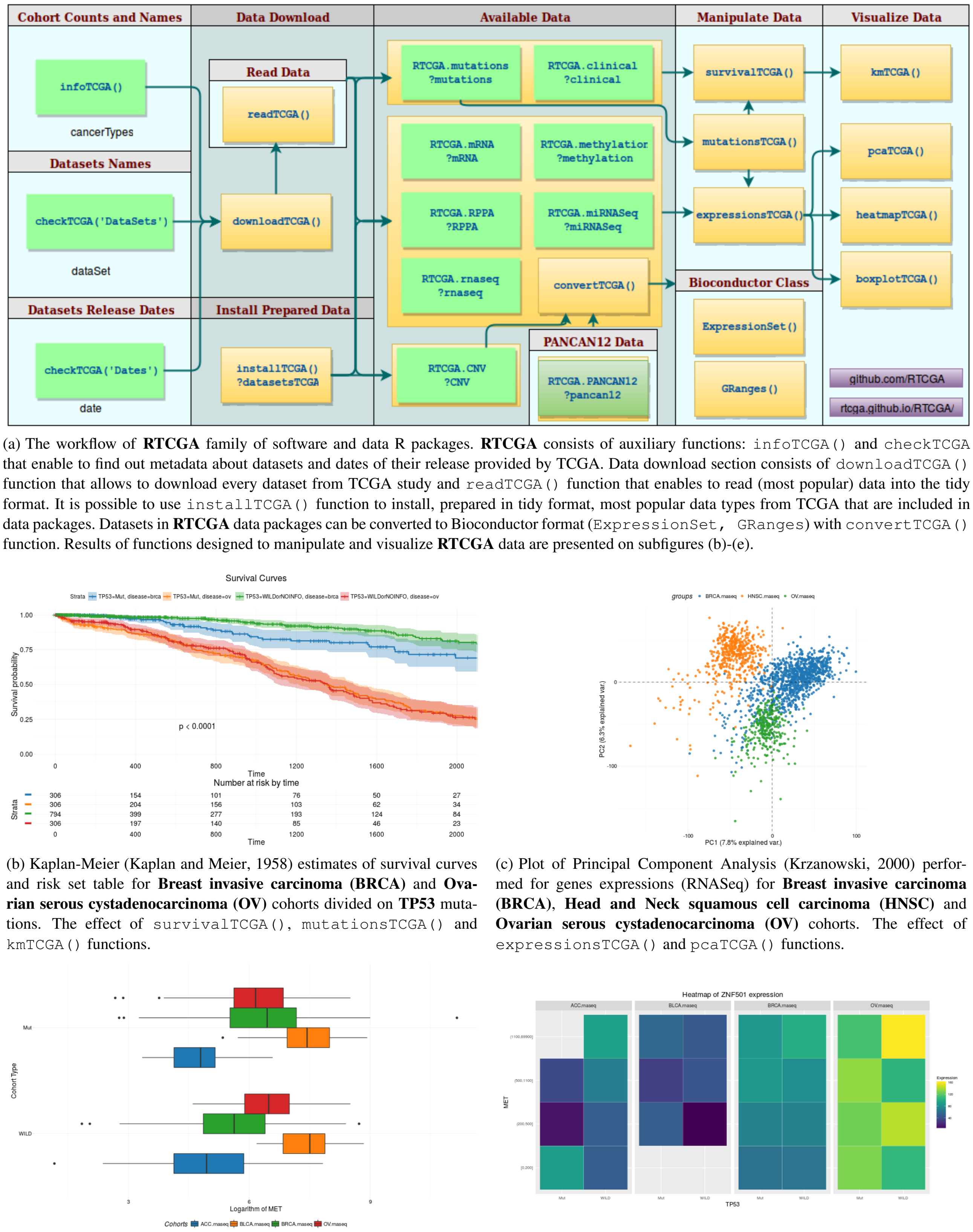
- If one requires to download datasets containing e.g. information about genes' expressions for all available cohorts types (TCGA collected data for more than 30 various cancer types) one would have to go through the click-to-download process many times, which is inconvenient and time-consuming.
- Clinical datasets from TCGA project are not in a standard tidy data format, which is: one row for one observation and one column for one variable. They are transposed, which makes work with that data burdensome. That becomes more onerous when one would like to investigate many clinical datasets.
- Datasets containing information on some data types (e.g. gene's mutations) are not in one easy-to-handle file. Every patient has its own file, which may be an impassable barrier for many potential researchers.
- Data governance for many datasets for various cohorts saved in different folders with strange (default after untarring) names may be exhausting and uncomfortable for researchers that are not very skilled in data management or data processing.

For these reasons we prepared selected datasets from the TCGA project in an easy to handle and process way and embedded them in 9 separate R packages. All packages can be installed from BioConductor.

Architecture

General architecture of RTCGA packages is presented in Figure 1a. The package RTCGA contains functions that facilitate download of data from particular freeze / cohort / platform. Detailed instructions on how to download data is presented in http://rtcga.github.io/RTCGA/Download.html. Other functions in this package facilitate data processing / analyzing and visualizing. In figures 1b-1e we present example analyses and plots that cover Kaplan-Meier survival curves, Principal Component Analysis, comparison of distributions through boxplots or heatmaps. Detailed instructions how to apply these analyses is presented in http://rtcga.github.io/RTCGA/Visualizations.html .

Other packages in this family have names in the format RTCGA.source, where source is a platform like RPPA / mRNA / miRNAseq / etc. They contain copy of data from TCGA for all cohorts from latest freeze.



(a) The workflow of **RTCGA** family of software and data R packages. **RTCGA** consists of auxiliary functions: `infoTCGA()` and `checkTCGA` that enable to find out metadata about datasets and dates of their release provided by TCGA. Data download section consists of `downloadTCGA()` function that allows to download every dataset from TCGA study and `readTCGA()` function that enables to read (most popular) data into the tidy format. It is possible to use `installTCGA()` function to install, prepared in tidy format, most popular data types from TCGA that are included in data packages. Datasets in **RTCGA** data packages can be converted to Bioconductor format (`ExpressionSet, GRanges`) with `convertTCGA()` function. Results of functions designed to manipulate and visualize **RTCGA** data are presented on subfigures (b)-(e).

(b) Kaplan-Meier (Kaplan and Meier, 1958) estimates of survival curves and risk set table for **Breast invasive carcinoma (BRCA)** and **Ovarian serous cystadenocarcinoma (OV)** cohorts divided on **TP53** mutations. The effect of `survivalTCGA()`, `mutationsTCGA()` and `kmTCGA()` functions.

(c) Plot of Principal Component Analysis (Krzanowski, 2000) performed for genes expressions (RNASeq) for **Breast invasive carcinoma (BRCA)**, **Head and Neck squamous cell carcinoma (HNSC)** and **Ovarian serous cystadenocarcinoma (OV)** cohorts. The effect of `expressionsTCGA()` and `pcaTCGA()` functions.

(d) Boxplots (Robert McGill, 1978) of logarithm of **MET** gene expression (RNASeq) for **Adrenocortical carcinoma (ACC)**, **Bladder urothelial carcinoma (BLCA)**, **Breast invasive carcinoma (BRCA)** and **Ovarian serous cystadenocarcinoma (OV)** divided on mutations in gene **TP53**. The effect of `expressionsTCGA()` and `boxplotTCGA()` functions.

(e) Heatmap (Friendly, 1994) presenting medians of **ZNF500** gene for **Adrenocortical carcinoma (ACC)**, **Bladder urothelial carcinoma (BLCA)**, **Breast invasive carcinoma (BRCA)** and **Ovarian serous cystadenocarcinoma (OV)** divided on **MET** gene quantiles. The effect of `expressionsTCGA()` and `heatmapTCGA()` functions.

Fig. 1: The workflow of **RTCGA** family of software and data R packages and effects of functions designed to manipulate and visualize **RTCGA** data.

References

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, **89**(425), 190–200.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

Krzanowski, W. (2000). *Principles of Multivariate Analysis*. Oxford Statistical Science Series. OUP Oxford.

Robert McGill, John W. Tukey, W. A. L. (1978). Variations of box plots. *The American Statistician*, **32**(1), 12–16.