OXFORD

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

Abstract

Summary: We present a family of **R** packages called **RTCGA** that simplify access to data from the TCGA project. The Cancer Genome Atlas Project (TCGA) is a coordinated effort to accelerate the 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. **RTCGA** packages facilitate access to these datasets, streamline merging characteristics from different platforms and support exploratory statistical analyses and visualizations.

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

Introduction

The Cancer Genome Atlas Data Portal (2017) 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 for 11 thousands patients, over 2.5 PT of data. Compressed *tar.gz* files are available through Firehose Broad GDAC Portal (2017) and recently through Genomic Data Commons Data Portal (2017). One can select a release (monthly snapshots), cancer type (cohort) and data type (e.g. clinical, RNA expression, methylation) and download a text file with raw data.

While working with many cohorts and cancer types we found this approach burdensome. Without easy-to-use API it is harder to reproduce results obtained in past.

- If one requires to download datasets containing e.g. information about genes' expressions for all available cohorts types, then one would have to go through the click-to-download process separately for each cohort. This is inconvenient and time-consuming.
- Some datasets (e.g. clinical) are not in a standard tidy data format, which is: one row for one observation and one column for one variable. Data for some platforms data is transposed (e.g. for expression columns stands for patients) for others data is unstructured (e.g. mutations).

That becomes more onerous when investigating many clinical datasets at once.

- Data governance for many datasets for various cohorts saved in different folders with very long names may be exhausting and uncomfortable for researchers that are not very skilled in data management or data processing.

For reasons listed above we prepared an uniform API to download and pre-process selected datasets along with set of R data packages with pre-processed data. The prepared packages are useful for biostatisticians that work with cancer data along with researchers that work on scalable big data algorithms or lecturers that are are using real world case studies.

Using RTCGA packages

The general architecture of all packages in the **RTCGA** family is presented in the Figure 1 (where functions are presented as yellow boxes and packages are green boxes). All packages listed in this figure are available at Bioconductor. The software package **RTCGA** contains functions that facilitate download of data from particular date, cohort and platform. Other functions benefits data processing, analysis and visualizations. In figures 1b-1c we present example analyses and plots that cover Kaplan-Meier estimates of survival curves and Principal Component Analysis. Detailed instructions on how to apply these and other analyses, and on how to download the selected data are presented at the project's website. Different
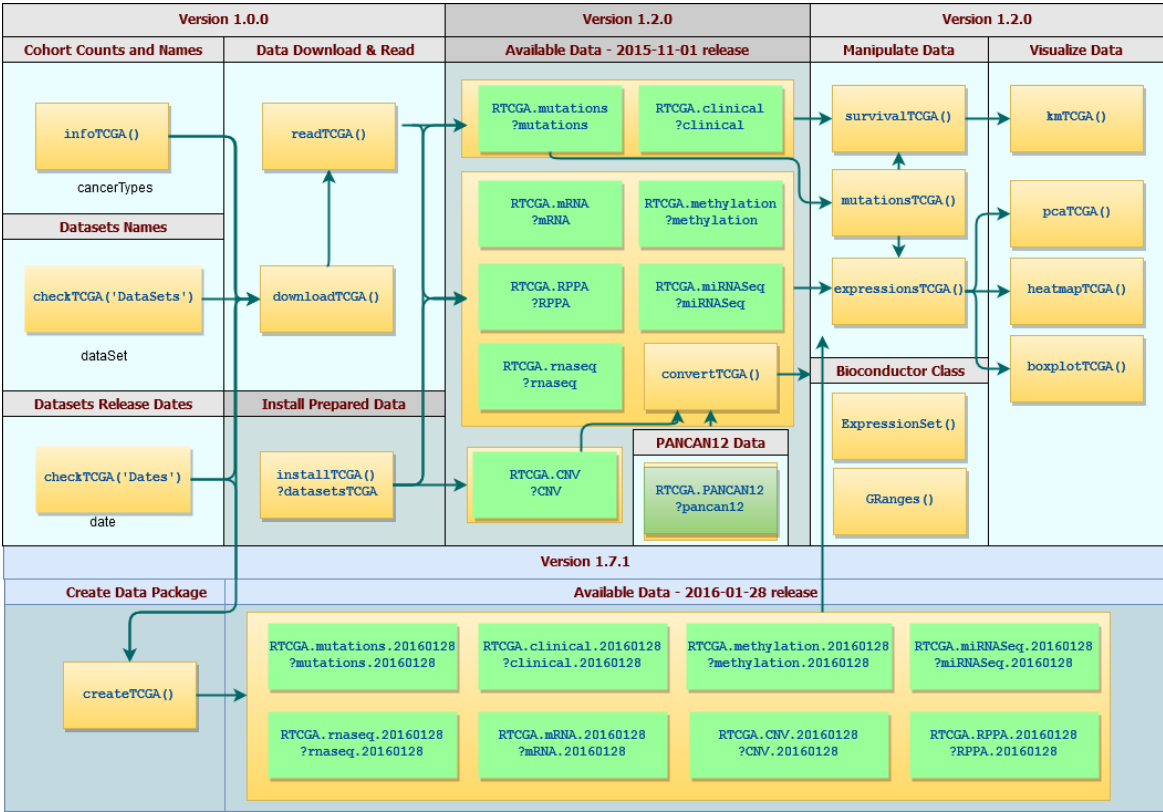
Fig. 1: The architecture of **RTCGA** family of R packages. **RTCGA** is a group of R libraries that were developed over the time.

In the first version (**1.0.0**), the software **RTCGA** package contained of functions designed to *download* TCGA data (`downloadTCGA()`) and verify it's metadata; like *cohort counts and names* (`infoTCGA()`), *datasets names* and their *dates of release* (`checkTCGA()`). *Reading* data into the tidy format was also possible (`readTCGA()`).

In the next version (**1.2.0**), **RTCGA** increased by several number of functions created for *data manipulations and visualizations*; `survivalTCGA()`, `expressionsTCGA()`, `mutationsTCGA()` for gathering the clinical outcomes, expressions data and mutations information, `kmTCGA()` for Kaplan-Meier estimates of survival curves, `pcaTCGA()` for Principal Components Analysis, `heatmapTCGA()` and `boxplotTCGA()` for for heatmaps and boxplots of expressions. This release was also equipped with a *convertTCGA()* for converting **RTCGA** data into the Bioconductor classes and *installTCGA()* for simplfied installation of the whole family of data packages. The latest (for that time: 2015-11-01) datasets from TCGA were packed into 8 data packages; *RTCGA.clinical*, *RTCGA.rnaseq*, *RTCGA.mutations*, *RTCGA.mRNA*, *RTCGA.methylation*, *RTCGA.RPPA*, *RTCGA.miRNAseq*, *RTCGA.CNV*.

The last version (**1.7.1**) presents the createTCGA() function that minimizes the effort of creating a data package from **RTCGA** family. With this function new 8 data packages (with the data from the last possible snapshot: 2016-01-28) were added to the ecosystem;

RTCGA.clinical.20160128, *RTCGA.rnaseq.20160128*, *RTCGA.mutations.20160128*, *RTCGA.mRNA.20160128*, *RTCGA.methylation.20160128*, *RTCGA.RPPA.20160128*, *RTCGA.miRNAseq.20160128*, *RTCGA.CNV.20160128*.

versions of data packages refer to particular snapshots of the data.Is this relevant to the new caption of the figure?

References

Firehose Broad GDAC Portal (2017). https://gdac.broadinstitute.org/.

Genomic Data Commons Data Portal (2017). https://portal.gdc.cancer.gov/.

The Cancer Genome Atlas Data Portal (2017). https://tcga-data.nci.nih.gov/docs/publications/tcga/.