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

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 1a. 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

A) Data Acquisition

Functions for exploring, downloading and unpacking data for selected snapshot and cohort directly from TCGA data portal.

`RTCGA::infoTCGA`
`RTCGA::checkTCGA`
`RTCGA::readTCGA`
`RTCGA::downloadTCGA`
`RTCGA::instalTCGA`

In order to simplify access to TCGA data selected snapshots and cohorts are available as R data packages. Names of packages available in Bioconductor are listed on right side. These packages contain data from the 2015-11-01 release.

`RTCGA.clinical`
`RTCGA.mutations`
`RTCGA.CNV`
`RTCGA.PANCAN12`
`RTCGA.methylation`
`RTCGA.rnaseq`
`RTCGA.mRNA`
`RTCGA.miRNAseq`
`RTCGA.RPPA`

B) Data Preprocessing

Functions for merging different patients characteristics, converging TCGA data into objects of the class GRanges or ExpressionSet.

`RTCGA::convertTCGA`
`RTCGA::convertPANCAN12`
`RTCGA::expressionsTCGA`
`RTCGA::mutationsTCGA`

C) Data Exploration / Visualisation

Functions for exploratory modeling and data visualization produce elegant and highly customizable plots with the ggplot2 package.

`RTCGA::survivalTCGA`
`RTCGA::kmTCGA`
`RTCGA::boxplotTCGA`
`RTCGA::heatmapTCGA`
`RTCGA::pcaTCGA`

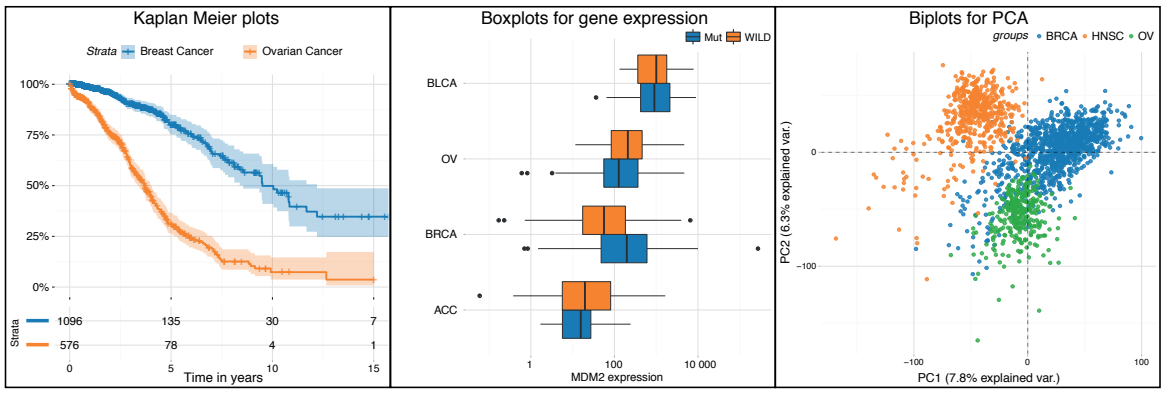

Fig. 1: The architecture of **RTCGA** family of R packages. (A) The **RTCGA** package contains functions for exploration of TCGA metadata for all available releases; function that allows to download every dataset from TCGA study and functions that enables to read data into the tidy format. In addition to this package a set of data packages are available in Bioconductor. (B) Data.frames with data from specific platform can be converted to Bioconductor format (`ExpressionSet, GRanges`) or merged with data from different platforms. (C) The **RTCGA** package also contains functions for exploratory analyses of TCGA data along with some popular visualisations. In this example following statistics are presented: Kaplan-Meier estimates of survival curves and risk set table for **Breast invasive carcinoma (BRCA)** and **Ovarian serous cystadenocarcinoma (OV)** cohorts; Boxplot for samples with mutated of wild-type versions of MDM2 gene; Principal Component Analysis performed for genes expressions (RNASeq) for **Breast invasive carcinoma (BRCA)**, **Head and Neck squamous cell carcinoma (HNSC)** and **Ovarian serous cystadenocarcinoma (OV)** cohorts

.

download the selected data are presented at the project's website. Different versions of data packages refer to particular snapshots of the data.

Acknowledgements
This work was supported by NCN Grant 2016/21/B/ST6/02176.

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/.