



Databases and ontologies

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

Marcin Kosiński^{1,2,*}, Witold Chodor² and Przemysław Biecek^{1,2}

¹Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland and

²Faculty 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: The following article presents RTCGA: a family of R packages with data from The Cancer Genome Atlas Project (TCGA) study. TCGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing [1]. We converted selected datasets from this study into few separate packages that are hosted on CRAN. These R packages make selected datasets easier to access and manage. Datasets in RTCGA packages are large and cover complex relations between clinical outcomes and genetic background. These packages will be useful for at least three audiences: biostatisticians that work with cancer data; researchers that are working on large scale algorithms; teachers that are presenting data analysis method on real data problems

Availability: RTCGA family of R packages is freely available from the Bioconductor project at <http://bioconductor.org/packages/RTCGA/>.

Contact: m.p.kosinski@gmail.com

1 Motivation

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes [1].

TCGA data is 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.

When working with many cancer types we 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, what for many potential researchers may be an impassable barrier.
- 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 embed them in 9 separate R packages. All packages can be installed from BioConductor.

2 Examples

Acknowledgements

Text Text Text Text Text Text Text. ? might want to know about text
For permissions, please e-mail: journals.permissions@oup.com

Funding

This work has been supported by the... Text Text Text Text.

References