RTCGA.data - The Family Of R Packages Containing TCGA Data

by Marcin Kosinski, Przemyslaw Biecek

Abstract The following article presents RTCGA.data: a family of R packages containing The Cancer Genome Atlas Project (TCGA) data. The Cancer Genome Atlas (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 provide TCGA data in few separate packages that are hosted on one GitHub repository, what made those luxurious data easier to possess and manage. We hope providing researchers with comprehensive catalogs of the key genomic changes in many major types and subtypes of cancer will support advances in developing more effective ways to diagnose, treat and prevent cancer.

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). The key is to understand genomics to improve cancer care.

Data origin

Data from TCGA are available through Firehose Broad GDAC portal (2). One can select cohort (assigned to the cancer type) and data type, i.e. clinical data, and download a `tar.gz` file with compressed data.

The main disadvantages of such data download methodology are:

- The downloaded files are compressed `tar.gz` files and not everyone manages to unpack such files.
- If one requires to download datasets containing i.e. 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 click-to-download process many times, which is inconvienent 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 what makes work with those data burdensome. That becomes more onerous when one would like to investigate many clinical datasets.
- Datasets containing information on gene's mutations are not in one easy-to-handle file. Every patient has it's own file, what for many potential researchers may be a impassable barrier. Just think about BRCA cohort (breast cancer) with more that 1000 various patients and more than 1200 files with patients' genes mutations information.

RTCGA.data family data

For reasons described in previous section we prepared selected datasets from TCGA project in an easy to handle and process way and embed them in 5 separate R packages. All packages can be installed from GitHub by evaluating the following code:

```
if (!require(devtools)) {
    install.packages("devtools")
    require(devtools)
}
install_github("mi2-warsaw/RTCGA.data",
                subdir = paste0("RTCGA.",
                                c("clinical", "rnaseq", "mutations", "cnv", "PANCAN12")
                               )
 )
```

One package, i.e. `RTCGA.clinical` can be installed with the command

```
if (!require(devtools)) {
    install.packages("devtools")
    require(devtools)
}
install_github("mi2-warsaw/RTCGA.data",
                subdir = "RTCGA.clinical")
```

If you are using Windows, make sure you have rtools [3] installed on your computer, before evaluating aboved commands.

RTCGA.data family contains 5 packages:

- `RTCGA.clinical` package containing clinical datasets from TCGA. Each cohort contains one dataset prepared in a tidy format. Each row, marked with patients' barcode, corresponds to one patient.
- `RTCGA.rnaseq` package containing genes' expressions datasets from TCGA. Each cohort contains one dataset with over 20 thousand of columns corresponding to genes' expression. Rows correspond to patients, that can be matched with patient's barcode.
- `RTCGA.mutations` package containint genes' mutations datsets from TCGA. Each cohort contains one dataset with extra column specifying patient's barcode which enables to distinguish which rows correspond to which patient.
- `RTCGA.cnv` package explanation needed.
- `RTCGA.PANCAN12` package explanation needed.

More detailed information about datasets included in RTCGA.data family are shown in Table **??**

How to work with RTCGA.data family

Patient's barcode as a key to merge data

Applications examples

`[1] http://cancergenome.nih.gov/`

`[2] http://gdac.broadinstitute.org/`

`[3] http://cran.r-project.org/bin/windows/Rtools/`

`\bibliography{RJreferences}`

Marcin Kosinski
Warsaw University of Technology
Faculty of Mathematics and Information Science
Koszykowa 75, 00-662 Warsaw, Poland
`M.P.Kosinski@gmail.com`

Przemyslaw Biecek
Warsaw University
line 1
line 2
`Przemyslaw.Biecek@gmail.com`

Disease Name	Cohort	*.clinical	*.mutations	*.rnaseq	*.cnv	*.PANC
Adrenocortical carcinoma	ACC					
Bladder urothelial carcinoma	BLCA					
Breast invasive carcinoma	BRCA					
Cervical and endocervical cancers	CESC					
Cholangiocarcinoma	CHOL					
Colon adenocarcinoma	COAD					
Colorectal adenocarcinoma	COADREAD					
Lymphoid Neoplasm Diffuse	DLBC					
Esophageal carcinoma	ESCA					
FFPE Pilot Phase II	FPPP					
Glioblastoma multiforme	GBM					
Glioma	GBMLGG					
Head and Neck squamous cell carcinoma	HNSC					
Kidney Chromophobe	KICH					
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN					
Kidney renal clear cell carcinoma	KIRC					
Kidney renal papillary cell carcinoma	KIRP					
Acute Myeloid Leukemia	LAML					
Brain Lower Grade Glioma	LGG					
Liver hepatocellular carcinoma	LIHC					
Lung adenocarcinoma	LUAD					
Lung squamous cell carcinoma	LUSC					
Mesothelioma	MESO					
Ovarian serous cystadenocarcinoma	OV					
Pancreatic adenocarcinoma	PAAD					
Pheochromocytoma and Paraganglioma	PCPG					
Prostate adenocarcinoma	PRAD					
Rectum adenocarcinoma	READ					
Sarcoma	SARC					
Skin Cutaneous Melanoma	SKCM					
Stomach adenocarcinoma	STAD					
Stomach and Esophageal carcinoma	STES					
Testicular Germ Cell Tumors	TGCT					
Thyroid carcinoma	THCA					
Thymoma	THYM					
Uterine Corpus Endometrial Carcinoma	UCEC					
Uterine Carcinosarcoma	UCS					
Uveal Melanoma	UVM					