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.

```
#> Warning: package 'knitr' was built under R version 3.2.1
```

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). 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.
- Data governance for many datasets for various cohorts saved in different folders with strange (default after untarring) names may be extremely exhausting and uncomfortable for researchers that are not very skilled in data management or data processing.

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. Clinical data format is explained here https://wiki.nci.nih.gov/display/TCGA/Clinical+Data+Overview
- `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. Genes' expressions data format is explained here https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2
- `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. Mutations' data format is explained here https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification.
- `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	Cases	clinical	cnv
1	Adrenocortical carcinoma	ACC	92	92x1046	21052x6
2	Bladder urothelial carcinoma	BLCA	412	393x1978	105795x6
3	Breast invasive carcinoma	BRCA	1098	1080x3464	284510x6
4	Cervical and endocervical cancers	CESC	307	304x1556	59450x6
5	Cholangiocarcinoma	CHOL	36	36x794	7570x6
6	Colon adenocarcinoma	COAD	460	453x2935	91166x6
7	Colorectal adenocarcinoma	COADREAD	631	624x3241	126931x6
8	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	47x693	9343x6
9	Esophageal carcinoma	ESCA	185	174x1115	60803x6
10	FFPE Pilot Phase II	FPPP	38	38x3277	
11	Glioblastoma multiforme	GBM	613	592x4935	146852x6
12	Glioma	GBMLGG	1129	1067x5158	226643x6
13	Head and Neck squamous cell carcinoma	HNSC	528	522x1625	110289x6
14	Kidney Chromophobe	KICH	113	110x854	10164x6
15	Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	914x2554	142122x6
16	Kidney renal clear cell carcinoma	KIRC	537	533x2474	85044x6
17	Kidney renal papillary cell carcinoma	KIRP	323	271x1746	46914x6
18	Acute Myeloid Leukemia	LAML	200	200x1148	28324x6
19	Brain Lower Grade Glioma	LGG	516	475x1940	79791x6
20	Liver hepatocellular carcinoma	LIHC	377	363x1441	93328x6
21	Lung adenocarcinoma	LUAD	585	521x2765	122927x6
22	Lung squamous cell carcinoma	LUSC	504	495x2487	134864x6
23	Mesothelioma	MESO	87	77x855	18335x6
24	Ovarian serous cystadenocarcinoma	OV	602	591x3305	261680x6
25	Pancreatic adenocarcinoma	PAAD	185	174x1148	34808x6
26	Pheochromocytoma and Paraganglioma	PCPG	179	179x1102	31256x6
27	Prostate adenocarcinoma	PRAD	499		117345x6
28	Rectum adenocarcinoma	READ	171	171x2492	35765x6
29	Sarcoma	SARC	260		106617x6
30	Skin Cutaneous Melanoma	SKCM	470	447x1750	108084x6
31	Stomach adenocarcinoma	STAD	443	438x1583	118389x6
32	Stomach and Esophageal carcinoma	STES	628	612x1719	
33	Testicular Germ Cell Tumors	TGCT	150	133x924	24952x6
34	Thyroid carcinoma	THCA	503	501x1556	55377x6
35	Thymoma	THYM	124	122x771	15571x6
36	Uterine Corpus Endometrial Carcinoma	UCEC	560	537x2038	127430x6
37	Uterine Carcinosarcoma	UCS	57	57x876	19298x6
38	Uveal Melanoma	UVM	80	80x541	12973x6