User's Guide for mydata Package

Coraline Qu

2017-01-06

Contents

1 Introduction

1.1 Information of raw data

Our data is downloaded from https://gdc-portal.nci.nih.gov. The original data is a group of text files, each of which consists of ensembl ids and htsequencing counts. These files are numbered and compiled into .rda format for convenience. However, the names of each text files provide essential information and therefore, are stored separately for further use.

1.2 Description of package

Package *mydata* is a data package for users to load different data sets. It is composed of six data sets:

- `luad`, `files_luad`;
- `brca`, `files_brca`;
- `sarc`, `files_sarc`.

These data sets are saved in the subdirectory `~/mydata/data`. "files_luad" and "luad" correspond to TCGA-luad. "files_luad" consists of the original text file names which contain primary information about the data, while "luad" is comprised of the htseq counts matrices for "normal" and "tumor" sample conditions. Similarly, "brca", "files_brca" are data sets with regard to TCGA-brca and "sarc" and "files_sarc" are data sets with regard to TCGA-sarc. It is notable that the sample conditions for TCGA-sarc are "dl(differentiated liposarcoma)" and "lm(leiomyosarcoma)". More information about these data sets will be given in the next chapter.

1.3 Data structure

The basic data structure of six data sets “files_luad”, “files_brca”, “files_sarc”, “luad”, “brca”, “sarc” are given in “Data Structure”:

Table 1: Data Structure

	class	length	element
files_luad	character	114	file names as normal01_luad.txt
files_brca	character	200	file names as normal01_brca.txt
files_sarc	character	162	file names as dl001_sarc01.txt
luad	list	114	60488*2 dataframe
brca	list	200	60488*2 dataframe
sarc	list	162	60488*2 dataframe

2 Standard workflow

This package can be installed from github via `install_github("Coraline66/mydata")`. Note that `install_github()` is a function from package `devtools`.

2.1 Example I

After installation, we can load desired data set into current environment via r code as follows.

```
# load libraries
library(devtools)
library(mydata)

# view all available data sets in package "mydata"
data(package="mydata")$results[, "Item"]
```

```
## [1] "brca"       "files_brca" "files_luad" "files_sarc" "luad"
## [6] "sarc"
```

```
# load data sets "luad" and "files_luad" into global environment
data(luad)
data(files_luad)

# view the first three elements of "files_luad"
files_luad[1:3]
```

```
## [1] "normal01_luad.txt" "normal02_luad.txt" "normal03_luad.txt"
```

```
# view the first three elements of "files_luad" referring to condition "tumor"
files_luad[58:60]
```

```
## [1] "tumor01_luad.txt" "tumor02_luad.txt" "tumor03_luad.txt"
```

```
# view the first three rows of the first and last element of "luad"
length(luad[[1]])
```

```
## [1] 2
```

```
luad[[1]][1:3,]
```

normal_1	normal1
ENSG00000000003.13	6858
ENSG00000000005.5	3
ENSG00000000419.11	2099

```
luad[[length(luad)]][1:3,]
```

tumor_57	tumor57
ENSG00000000003.13	2848
ENSG00000000005.5	2
ENSG00000000419.11	1312

It is easier for users to figure out what the available data sets are in this package by executing `data(package="mydata")`. As we can see from the result above, "luad" is a list of length `114`, each of which contains the htseq counts information of `57` pairs of samples. columns "normal_*" and "tumor_*" represent different genes, while columns "normal*" and "tumor*" represent the corresponding htseq counts. Here "*" gives the serial number of patients, e.g., "normal01_luad.txt" in "files_luad" corresponds to the first element of data set "luad", `luad[[1]]` equivalently.

2.2 Example II

The second example is about how to load data set of TCGA-brca.

```
# load libraries
library(devtools)
library(mydata)

# load data sets "brca" and "files_brca" into global environment
data(brca)
data(files_brca)

# view the first three elements of "files_brca"
files_brca[1:3]
```

```
## [1] "normal001_brca.txt" "normal002_brca.txt" "normal003_brca.txt"
```

```
# view the first three elements of "files_brca" referring to condition "tumor"
files_brca[101:103]
```

```
## [1] "tumor001_brca.txt" "tumor002_brca.txt" "tumor003_brca.txt"
```

```r
# view the first three rows of the first and last element of "brca"
length(brca[[1]])
```

```
## [1] 2
```

```r
brca[[1]][1:3,]
```

normal_1	normal1
ENSG00000000003.13	3616
ENSG00000000005.5	3616
ENSG00000000419.11	1254

```r
brca[[length(brca)]][1:3,]
```

tumor_100	tumor100
ENSG00000000003.13	8543
ENSG00000000005.5	26
ENSG00000000419.11	3919

"files_brca" consists of the original file names of `100` patients from TCGA-brca. Information of `60488` genes and counts is stored in data set "brca". It is a list composed of `114` 60488×2 data frames, odd columns of which give gene serial numbers and even columns of which give corresponding counts, e.g., `brca[[101]]` refers to the element "tumor01_brca.txt" in "files_brca".

2.3 Example III

Here is the r code for loading data set of TCGA-sarc.

```r
# load libraries
library(devtools)
library(mydata)

# load data sets "sarc" and "files_sarc" into global environment
data(sarc)
data(files_sarc)

# view the first three elements of "files_sarc"
files_sarc[1:3]
```

```
## [1] "dl001_sarc01.txt" "dl002_sarc02.txt" "dl003_sarc03.txt"
```

```r
# view the first three elements of "files_sarc" referring to condition "lm"
files_sarc[59:61]
```

```
## [1] "lm059_sarc05.txt" "lm060_sarc06.txt" "lm061_sarc07.txt"
```

```
# view the first three rows of the first and last element of "sarc"
length(sarc[[1]])
```

```
## [1] 2
```

```
sarc[[1]][1:3,]
```

dl_1	dl1
ENSG00000000003.13	631
ENSG00000000005.5	26
ENSG00000000419.11	1492

```
sarc[[length(sarc)]][1:3,]
```

lm_162	lm162
ENSG00000000003.13	531
ENSG00000000005.5	14
ENSG00000000419.11	1087

"files_sarc" consists of the original file names of `162` samples from TCGA-sarc. Note that `58` samples come from liposarcoma and `104` samples come from leiomyosarcoma. Thus, this case differs from the other two above since all samples are from different patients. Information of genes and htseq counts is saved in "sarc", a list consisting of `162` 60488×2 data frames, odd columns of which give gene serial numbers and even columns of which give corresponding counts, e.g., `sarc[[59]]` refers to the element "lm059_sarc05.txt" from data set "files_sarc".

3 Session Info

- R version 3.3.1 (2016-06-21)
- Platform: x86_64-apple-darwin13.4.0 (64-bit)
- Locale:
  - LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C
  - LC_TIME=en_US.UTF-8, LC_COLLATE=en_US.UTF-8
  - LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8
- Base packages: devtools