User's Guide for mydata Package

Coraline Qu

2017-01-05

Contents

1 Introduction

1.1 Description

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`.

"files_luad" and "luad" correspond to TCGA-luad. "files_luad" consists of the original text file names which contain all the 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.

2 Standard workflow

2.1 Example I

This package can be installed from github via `install_github("Coraline66/mydata")`. Note that `install_github()` is a function from package `devtools`. After installation, we can load desired data set into current environment via r code as follows.

```r
library(devtools)
library(mydata)
data(package="mydata")
data(luad)
data(files_luad)
files_luad[c(1, length(files_luad)/2)]
```

```
## [1] "normal01_luad.txt" "normal57_luad.txt"
```

```r
files_luad[c(length(files_luad)/2+1, length(files_luad))]
```

```
## [1] "tumor01_luad.txt" "tumor57_luad.txt"
```

```r
dim(luad[[1]])
```

```
## [1] 60488     2
```

```r
luad[[1]][1:2,]
```

normal_1	normal1
ENSG00000000003.13	6858
ENSG00000000005.5	3

```r
luad[[length(luad)]][1:2,]
```

tumor_57	tumor57
ENSG00000000003.13	2848
ENSG00000000005.5	2

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.

```r
library(devtools)
library(mydata)
data(brca)
data(files_brca)
files_brca[c(1, length(files_brca)/2)]
```

```
## [1] "normal001_brca.txt" "normal100_brca.txt"
```

```
files_brca[c(length(files_brca)/2+1, length(files_brca))]
```

```
## [1] "tumor001_brca.txt" "tumor100_brca.txt"
```

```
brca[[1]][1:2,]
```

normal_1	normal1
ENSG00000000003.13	3616
ENSG00000000005.5	3616

```
brca[[length(brca)]][1:2,]
```

tumor_100	tumor100
ENSG00000000003.13	8543
ENSG00000000005.5	26

"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.

```
library(devtools)
library(mydata)
data(sarc)
data(files_sarc)
files_sarc[c(1, 58)]
```

```
## [1] "dl001_sarc01.txt"  "dl058_sarc259.txt"
```

```
files_sarc[c(59, 162)]
```

```
## [1] "lm059_sarc05.txt"  "lm162_sarc256.txt"
```

```
sarc[[1]][1:2,]
```

dl_1	dl1
ENSG00000000003.13	631
ENSG00000000005.5	26

```
sarc[[length(sarc)]][1:2,]
```

lm_162	lm162
ENSG00000000003.13	531
ENSG00000000005.5	14

“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