User's Guide for mydata Package

Coraline Qu

2017-01-04

Contents

1 Introduction **1**

1.1 Description 1

2 Standard workflow **1**

2.1 Example I 1

2.2 Example II 2

2.3 Example III 3

3 Session Info **4**

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

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

```
#install.packages("devtools")
library(devtools)
#install_github("Coraline66/mydata")
library(mydata)
```

```
data(package="mydata")
data(luad)
data(files_luad)
dim(files_luad)
```

```
## [1] 114   1
```

```
files_luad[1, 1]
```

```
## [1] normal01_luad.txt
## 114 Levels: normal01_luad.txt normal02_luad.txt ... tumor57_luad.txt
```

```
files_luad[dim(files_luad)[1], 1]
```

```
## [1] tumor57_luad.txt
## 114 Levels: normal01_luad.txt normal02_luad.txt ... tumor57_luad.txt
```

```
dim(luad)
```

```
## [1] 60488   228
```

```
luad[1, 1:4]
```

normal_1	normal1	normal_2	normal2
ENSG00000000003.13	6858	ENSG00000000003.13	2510

```
luad[1, 115:118]
```

tumor_1	tumor1	tumor_2	tumor2
ENSG00000000003.13	3432	ENSG00000000003.13	7734

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 60488×228 data frame and 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., element "normal01_luad.txt" corresponds to the first two columns in data set "luad", `luad[ , 1:2]` equivalently.

2.2 Example II

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

```
library(devtools)
library(mydata)
data(brca)
```

```r
data(files_brca)
dim(files_brca)
```

```
## [1] 200   1
```

```r
files_brca[1, 1]
```

```
## [1] normal001_brca.txt
## 200 Levels: normal001_brca.txt normal002_brca.txt ... tumor100_brca.txt
```

```r
files_brca[dim(files_brca)[1], 1]
```

```
## [1] tumor100_brca.txt
## 200 Levels: normal001_brca.txt normal002_brca.txt ... tumor100_brca.txt
```

```r
dim(brca)
```

```
## [1] 60488   400
```

```r
brca[1, 1:4]
```

normal_1	normal1	normal_2	normal2
ENSG00000000003.13	3616	ENSG00000000003.13	9397

```r
brca[1, 201:204]
```

tumor_1	tumor1	tumor_2	tumor2
ENSG00000000003.13	2679	ENSG00000000003.13	2671

"files_brca" consists of the original file names of `100` patients from TCGA-brca. Information of `60488` genes and counts is stored in a 60488×400 data frame "brca", odd columns of which give gene serial numbers and even columns of which give corresponding counts, e.g., `brca[, 201:202]` 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
library(devtools)
library(mydata)
data(sarc)
data(files_sarc)
dim(files_sarc)
```

```
## [1] 162   1
```

```
files_sarc[1, 1]
```

```
## [1] dl001_sarc01.txt
## 162 Levels: dl001_sarc01.txt dl002_sarc02.txt ... lm162_sarc256.txt
```

```
files_sarc[dim(files_sarc)[1], 1]
```

```
## [1] lm162_sarc256.txt
## 162 Levels: dl001_sarc01.txt dl002_sarc02.txt ... lm162_sarc256.txt
```

```
dim(sarc)
```

```
## [1] 60488   324
```

```
sarc[1, 1:4]
```

dl_1	dl1	dl_2	dl2
ENSG00000000003.13	631	ENSG00000000003.13	555

```
sarc[1, 117:120]
```

lm_59	lm59	lm_60	lm60
ENSG00000000003.13	900	ENSG00000000003.13	433

"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 a 60488×324 data frame "sarc", odd columns of which give gene serial numbers and even columns of which give corresponding counts, e.g., `sarc[, 117:118]` refers to the element "lm059_sarc05.txt" in "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