

Introduction to R

Workshop on Data Visualization in R

Lokesh Mano • 12-Oct-2022

NBIS, SciLifeLab

Contents

- Course and webpage
- Overview of R
- Data formats
- Data frames
- Important functions
- Tips

Quick checkups



If you don't recognize the correlation you see in the figure above, I would highly recommend you to read the following paper ;)

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* 21, 231 (2020).

- Coffe breaks (15 minutes is fine?)
- Webpage structure
- Plots from drop-down
- Times mentioned in schedule are **super** arbitrary

R

- Derived from a statistical programming language called S
- You can write your own functions
- Powerful and flexible.
- Available for all platforms
- **GUI** with Rstudio
- **RMarkdown**: Embedding codes and results together



Data Formats

- Wide format

| | Sample_1 | Sample_2 | Sample_3 | Sample_4 |
|------------------|----------|----------|----------|----------|
| ENSG000000000003 | 321 | 303 | 204 | 492 |
| ENSG000000000005 | 0 | 0 | 0 | 0 |
| ENSG000000000419 | 696 | 660 | 472 | 951 |
| ENSG000000000457 | 59 | 54 | 44 | 109 |
| ENSG000000000460 | 399 | 405 | 236 | 445 |
| ENSG000000000938 | 0 | 0 | 0 | 0 |

- familiarity
- conveniency
- you see more data

Data Formats

- Long format

| Sample_ID | Gene | count |
|-----------|-----------------|-------|
| Sample_1 | ENSG00000000003 | 321 |
| Sample_1 | ENSG00000000005 | 0 |
| Sample_1 | ENSG00000000419 | 696 |
| Sample_1 | ENSG00000000457 | 59 |
| Sample_1 | ENSG00000000460 | 399 |
| Sample_1 | ENSG00000000938 | 0 |

| Sample_ID | Sample_Name | Time | Replicate | Cell | Gene | count |
|-----------|-------------|------|-----------|------|-----------------|-------|
| Sample_1 | t0_A | t0 | A | A431 | ENSG00000000003 | 321 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG00000000005 | 0 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG00000000419 | 696 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG00000000457 | 59 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG00000000460 | 399 |

Data Formats

- Long format

| Sample_ID | Sample_Name | Time | Replicate | Cell | Gene | count |
|-----------|-------------|------|-----------|------|------------------|-------|
| Sample_1 | t0_A | t0 | A | A431 | ENSG000000000003 | 321 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG000000000005 | 0 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG000000000419 | 696 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG000000000457 | 59 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG000000000460 | 399 |
| Sample_1 | t0_A | t0 | A | A431 | ENSG000000000938 | 0 |

- easier to add data to the existing
- Most databases store and maintain in long-formats due to its efficiency
- R tools like **ggplot** require data in long format.

Data Frames

- Let us take a quick look into `data.frame` in `R`:



- imported files are usually in `data.frame`
- Structured matrix with `row.names` and `colnames`
- Probably most used `data.type` in Biology!

Vectors

```
n <- c(2,3,4,2,1,2,4,5,10,11,8,9)
print(n)
```

```
## [1] 2 3 4 2 1 2 4 5 10 11 8 9
```

```
z <- n +3
print(z)
```

```
## [1] 5 6 7 5 4 5 7 8 13 14 11 12
```

```
z <- n +3
mean(z)
```

```
## [1] 8.083333
```

```
s <- c("I", "love", "Batman")
print(s)
```

```
## [1] "I"      "love"   "Batman"
```

Vector types

- `int` stands for *integers*
- `dbl` stands for *doubles* or real numbers
- `chr` stands for *character* vectors or strings
- `dtm` stands for *date and time*,
- `lgl` stands for *logical* with just TRUE or FALSE
- `fctr` stands for *factors* which R uses to state categorical variables.
- `date` stands for *dates*

You can find what kind of vectors you have or imported by using the function `class()`



Thank you. Questions?

R version 4.1.3 (2022-03-10)

Platform: x86_64-pc-linux-gnu (64-bit)

OS: Ubuntu 18.04.6 LTS

Built on: 📅 12-Oct-2022 at 🕒 12:28:38

2022 • SciLifeLab • NBIS