R crash course: A quick introduction to R

Alex Sanchez, Miriam Mota, Ricardo Gonzalo and Santiago Perez-Hoyos

Statistics and Bioinformatics Unit. Vall d'Hebron Institut de Recerca

Readme

- License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-nc-sa/4.0/
- You are free to:
  - **Share** : copy and redistribute the material
  - **Adapt** : rebuild and transform the material
- Under the following conditions:
  - **Attribution** : You must give appropriate credit, provide a link to the license, and indicate if changes were made.
  - **NonCommercial** : You may not use this work for commercial purposes.
  - **Share Alike** : If you remix, transform, or build upon this work, you must distribute your contributions under the same license to this one

Section 1

A Crash Course in R

Outline

- Why R
  - R basics
  - How does one work with R and Rstudio
- Getting Started
  - A primer of data import
  - Variables and data types
  - Functions, Packages and more stuff
- Working with data
  - Selecting, Filtering and ordering datasets
  - A primer of statistics and plots
  - R Notebooks and RMarkdown

Motivation

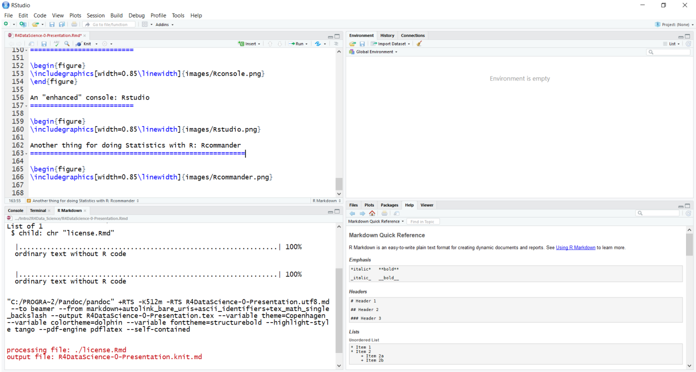
- We (you) all work with data, most of the time and often we need to do “things” with those data.
  - I have three lists of genes and I would like to see which genes they have in common (or which ones appear only in one list).
  - We have received the data from that lab but I only want to work with a subset of the samples.
  - Is it possible to repeat that plot changing the line colors, the font size etc?
  - I have some scripts tu re-run an analysis but I don’t know how to start
- These, and many other things can be done with a basic knowledge of R.

What is R?

- R is a *language and environment* for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible.
- It can be used fro simple tasks to highly complex reproducible projects.
- It compiles and runs on a wide variety of UNIX platforms and similar systems Windows and MacOS.

How is R used

- Different ways to use R, but the best trade-off simplicity-efficiency is provided by Rstudio



```
==========================

\begin{figure}
\includegraphics[width=0.85\linewidth]{images/Rconsole.png}
\end{figure}

An "enhanced" console: Rstudio
==========================

\begin{figure}
\includegraphics[width=0.85\linewidth]{images/Rstudio.png}

Another thing for doing Statistics with R: Rcommander
=====================================================

\begin{figure}
\includegraphics[width=0.85\linewidth]{images/Rcommander.png}
```

```
List of 1
 $ child: chr "license.Rmd"

  |.................................................................| 100%
  ordinary text without R code

  |.................................................................| 100%
  ordinary text without R code

"C:/PROGRA~2/Pandoc/pandoc" +RTS -K512m -RTS R4DataScience-0-Presentation.utf8.md
 --to beamer --from markdown+autolink_bare_uris+ascii_identifiers+tex_math_single
_backslash --output R4DataScience-0-Presentation.tex --variable theme=Copenhagen
--variable colortheme=dolphin --variable fonttheme=structurebold --highlight-styl
e tango --pdf-engine pdflatex --self-contained


processing file: ./license.Rmd
output file: R4DataScience-0-Presentation.knit.md
```

Markdown Quick Reference

R Markdown is an easy-to-write plain text format for creating dynamic documents and reports. See Using R Markdown to learn more.

Emphasis

```
*italic*   **bold**

_italic_   __bold__
```

Headers

```
# Header 1

## Header 2

### Header 3
```

Lists

Unordered List

```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

Exercise

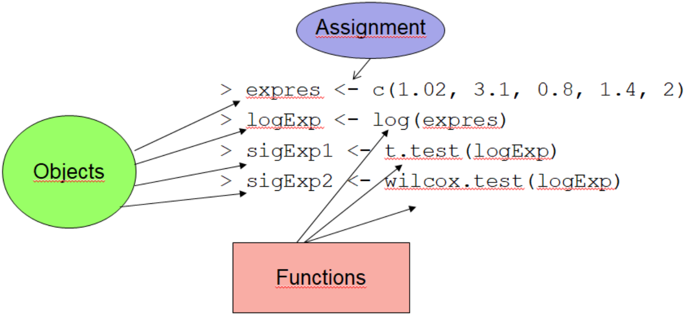
- Get to know R. Visit the R-project page and see what can be found there.
- If you haven't done it before, download and install R and Rstudio in your computer
- Open R studio. Look at the panels and figgure out what can we do at each window.

Section 2

Using R

Commands, Objects and Functions

- Shortly, using R consists of
  - Working with *objects* using *commands* and *functions*



```
> expres <- c(1.02, 3.1, 0.8, 1.4, 2)
> logExp <- log(expres)
> sigExp1 <- t.test(logExp)
> sigExp2 <- wilcox.test(logExp)
```

Variables and data types

- Data managed in R . . .
    - is stored as *variables*
- Variables can be of distinct types
    - Numerical
        - numeric (13.7)
        - int (3)
    - Character
        - "R is cute"
    - Factors
        - A,B,C,D
        - WT, Mut

R packages

- R can be used for many different types of data processing and analysis from distinct fields, besides statistics such as Ecology, Omics Sciences, Psychology etc.
- All these capabilities are not present from the begining because most of them will never be used by most users.
- Instead, thay can be added when needed by
  1. installing and
  2. loading the appropriate packages.

Installing and loading packages

We want to analyze some data using cox proportional hazards model.

```r
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
```

```
Error in coxph(Surv(time, status) ~ sex, data = lung)
: could not find function "coxph"
```

We need to install and load the package before we can use it.

```r
install.packages("survival")
library(survival)
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
```

Bioconductor

- Packages analyse all kinds of Genomic data (>800)
- Compulsory documentation (*vignettes*) for each package
- 6-month release cycle
- Course Materials
- Example data and workflows
- Common, re-usable framework and functionality
- Available Support
    - Often you will be able to interact with the package maintainers / developers and other power-users of the project software

The tidyverse

- The tidyverse is an opinionated collection of R packages designed for data science.
- All packages share an underlying design philosophy, grammar, and data structures.
- The complete tidyverse collection can be installed with:

```
install.packages("tidyverse")
```

- https://www.tidyverse.org/

Section 3

Getting data into R

Importing data with Rstudio

- The easiest way to get data into R is to click on the Ìmport Datasets button.
- Alternatively R code can be written using functions from Base R or the tidyverse
  - Base R functions start with read.: read.table, read.csv
  - tidyverse functions start with read_: read_delim, read_csv or read_excel

Reading Excel or csv files

- Files can be read from any location, let it be a physical support or a web site.
- To read files from disk be sure to indicate their location.
- Alternatively the default working directory can be set to the folder where the file is located.

Reading Excel or csv files (continued)

- Assume files `TIO2+PTYR-human-MSS+MSIvsPD.XLSX` has been downloaded to your working directory
- Start setting the default directory to the folder where you have saved the file.
    - `Session --> Set Working directory --> To source file location...`
- Import the `TIO2+PTYR-human-MSS+MSIvsPD.XLSX` with the default options
- Code generated for reading the files can be reused any time changing the file name if needed.

```
# Read Excel file
library(readxl)
otherData <- read_excel("otherFiles")
```

Interlude: Summarizing data

- Once a dataset is available it is easy to “have a look at it”

```
head(phosphoprotData)
str(phosphoprotData)
summary (phosphoprotData)
```

Section 4

Dynamic output with Rmarkdown

Reproducible research with R notebooks

- R and Rstudio are strongly involved in promoting reproducibility and reproducible research.
- This is implemented in **R notebooks**
- A notebook combines
  - Natural language text, e.g. describing what we are doing in our own words.
  - R code with the instructions needed to do the data management or the analysis.
  - The output of the analysis

Creating Notebooks

- A notebook can be created in Rstudio with
  - `File --> New File --> R Notebook`
- The notebook contains example text and code so it is straightforward to adapt it to your analysis.
- To produce an html file with text, code and output:
  - Press the button “Preview”
  - Or Select “Knitr to Html”

Section 5

Resources and exercises

Introductory materials

The web is full of all types of materials about R

Below there are a couple of brief introductions:

- A short introduction to R
- Getting started with R

Exercise

- Select a dataset with which you wish to work along the course.
- Read it into R
  - How many variables are there in it
  - What are their types
- Try to summarize it briefly
- Create an R notebook to encapsulate all your steps and share it with somebody.