

Data Management, Programming and Graphics with the R language

Alex Sanchez, Miriam Mota, Ricardo Gonzalo and Mireia Ferrer

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.

Introduction

Outline

- Introduction
 - Who are we (“we”=teachers & students)
 - Why are we here (Why learn R?)
 - - Objectives and competences
 - - Course contents
- How will we proceed: Methodology
- HW Data Science approach to using R
- References & Resources
- A first contact with R & Rstudio

Who are we (1): The Statistics and Bioinformatics Unit

www.ueb.vhir.org

Welcome to VHIR's Statistics and Bioinformatics Unit

Who we are

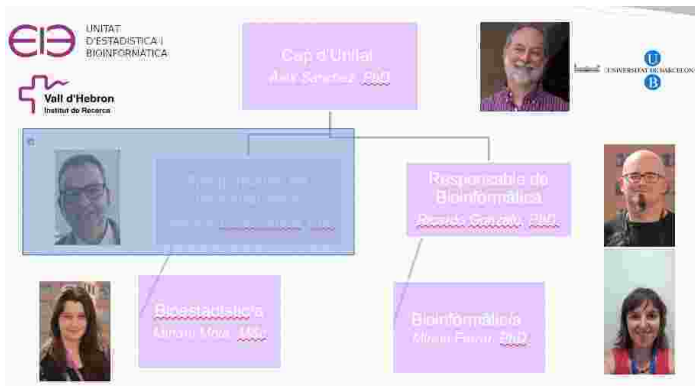
The Statistics and Bioinformatics Unit (UEB-USMB) is a service unit from the Scientific Support Area of the Vall d'Hebron Research Institute (VHIR - www.vhir.org).

The UEB was created in 2006 within the Research Institute of the Hospital Vall d'Hebron in order to promote the use and development of modern statistical and bioinformatics resources on research performed in its environment.



Nowadays, the Statistics and Bioinformatics Unit includes the former Support Unit in Methodology for Biomedical Research (USMB) and is part of the Scientific and Technical Support Area of the Vall d'Hebron Research Institute. It has the mission to provide expert advice, services and training for clinical and biomedical research.

Who are we (2): Teachers



Who are we (3): The GRBio Research group

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Site Map - Contact - Log In

Grup de Recerca en **Bioestadística** | **Bioinformàtica**

Home About us Research Activites GRBIO Seminars K&T transfer Contact News Photo Gallery

GRBIO seminars, Home

$X_i + \beta_j$

$h(t) = h_0$

Welcome to the GRBIO website!

Our research group has expertise in **Biostatistics** and **Bioinformatics**; mainly: Survival Analysis, Clinical Trials and Biostatistical Methods for Integrative Analysis of Omics Data. Visit our web to see our activities, publications and statistical tools.

Applications for PhD studies are welcomed.

News

GRBIO: Concessió de l'ajut per donar suport a les activitats del grup de recerca (SGR 2017-2019) Sep 29, 2018

Proposal consuee: ISCB 2019 Sep 28, 2018

JOB: Setymne Health! Sep 12, 2018

Twitter

Tweets by @GRBIO_BCN

GRBIO Announced

BCAM @BCAMBitbo

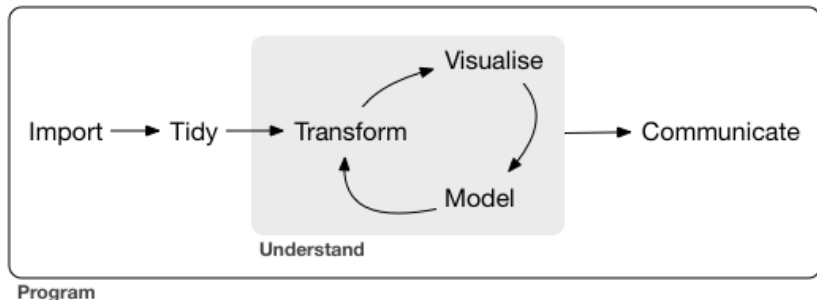
Zorionak @ZorionakIT 🏆🏆🏆

Oct 3, 2018

Why learn R

- Most people in most jobs have to *manage* information in their every day work.
- “Managing” may mean different things such as:
 - *retrieving*
 - *manipulating*
 - *visualizing*
 - *analyzing*
 - *reporting*
- R is a powerful tool that can be used to facilitate, improve or automate tasks such as those described above.

Hadley Wickam's approach to learning and applying Data Science



Your turn

- Provide examples of informations you may wish to manage
- Describe briefly
 - what this information is about
 - how it is stored
 - what you may wish to do with it
 - Transformations
 - Computations
 - Reports

How we will work

- Mastering R requires as many other disciplines
 - ❶ Time
 - ❷ Study, and
 - ❸ Practice.
- Our lectures will have the following structure (all but the first)
 - 1st part: Discuss the work you have done during the week
 - 2nd part: We introduce a few new ideas
 - 3rd part: Practice exercises and start working on the case study suggested/your data.

A first contact with R, Rstudio and the tidyverse

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 compiles and runs on a wide variety of UNIX platforms and similar systems Windows and MacOS.

R PRO's (why you are here!)

- The system is
 - free (as in *free beer*)
 - It's platform independent
 - It is constantly improving (2 new versions/year)
- It is a statistical tool
 - Implements almost every statistical method that exists
 - Great graphics (Examples)
 - Simple reporting tools
 - Also state-of-the-art in Bioinformatics through the Bioconductor Project.
- Programming language
 - Easy to automate repetitive tasks (Example_1.1)
 - Possibility to create user friendly web interfaces with a moderate effort. (Examples)

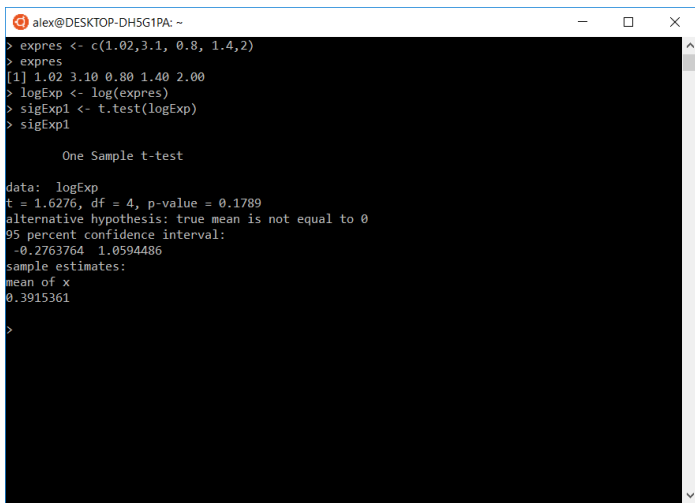
R CON's

- R is mainly used issuing commands from a console
 - less user friendly than almost any other statistical tool you may know.
- Constantly having new versions may affect our projects
- Not necessarily the best language nor suitable for every existing task

Using R

- Traditionally R was used from an Operating System console (“Terminal”)
- This is an intimidating approach for many users
- A variety of options exist to decrease the learning curve.
 - Use a supportive development environment such as **Rstudio**
 - Use an interface to Statistical tools, such as **Rcommander** or **::DeduceR**** allowing to concentrate on Statistics, not in commands.

A raw R console in linux



```
alex@DESKTOP-DH5G1PA: ~  
> expres <- c(1.02,3.1, 0.8, 1.4,2)  
> expres  
[1] 1.02 3.10 0.80 1.40 2.00  
> logExp <- log(expres)  
> sigExp1 <- t.test(logExp)  
> sigExp1  
  
      One Sample t-test  
  
data:  logExp  
t = 1.6276, df = 4, p-value = 0.1789  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 -0.2763764  1.0594486  
sample estimates:  
mean of x  
0.3915361  
  
>
```

An “enhanced” console: Rstudio

The screenshot shows the RStudio IDE. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar below the menu has icons for creating a new file, opening a file, saving, running, and other standard IDE functions.

The main editor window displays an R script with the following content:

```

150 | .....|
151 | .....|
152 | \begin{figure}
153 | \includegraphics[width=0.85\linewidth]{images/Rconsole.png}
154 | \end{figure}
155 | .....|
156 | An "enhanced" console: Rstudio
157 | .....|
158 | .....|
159 | \begin{figure}
160 | \includegraphics[width=0.85\linewidth]{images/Rstudio.png}
161 | .....|
162 | Another thing for doing Statistics with R: Rcomander
163 | .....|
164 | .....|
165 | \begin{figure}
166 | \includegraphics[width=0.85\linewidth]{images/Rcomander.png}
167 | .....|
168 | .....|

```

The console window at the bottom left shows the output of the script:

```

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 -KS12m -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

```

The sidebar on the right contains three panels: Environment (showing a global environment), History (showing a list of commands), and Help (showing the Markdown Quick Reference).

Another thing for doing Statistics with R: Rcommander

