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

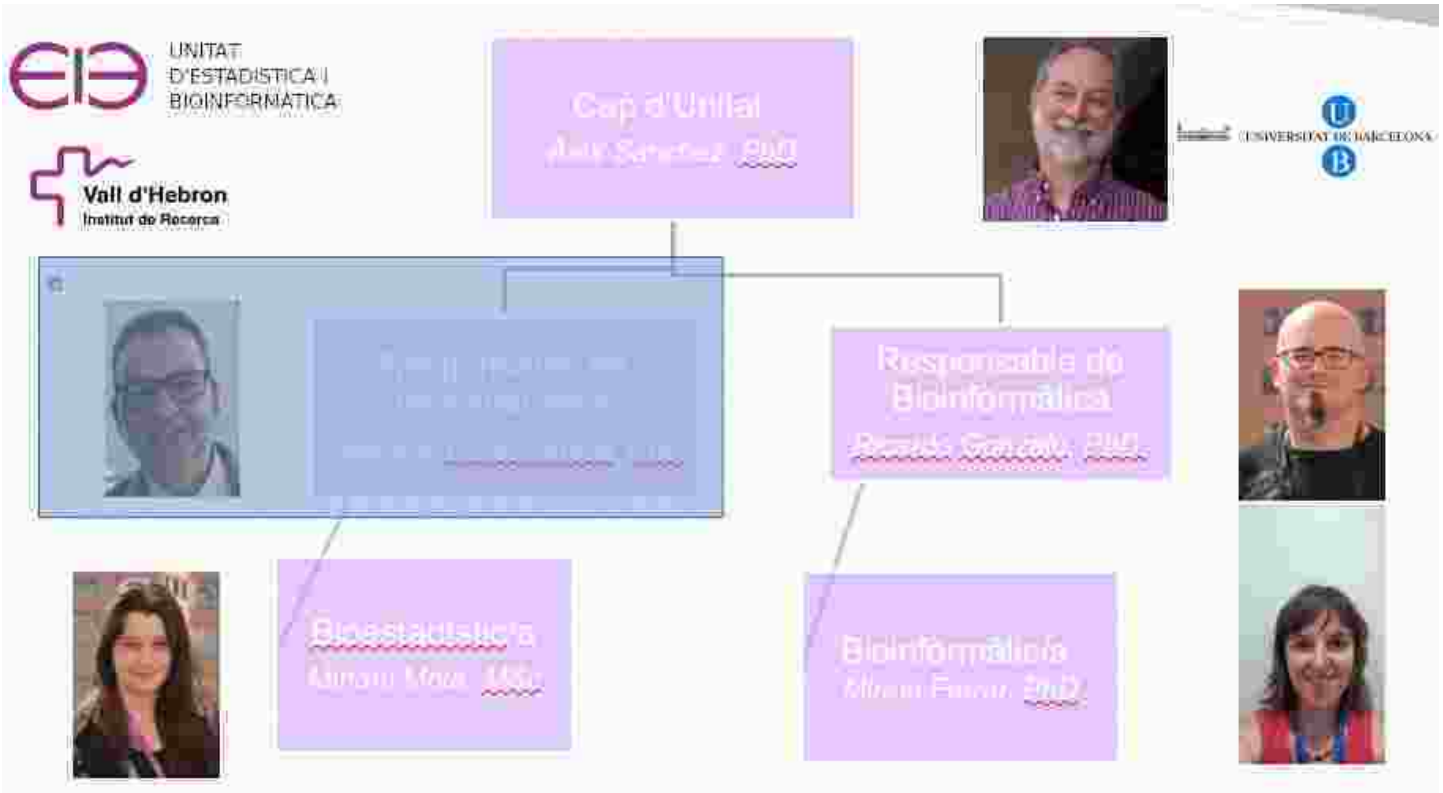
Who we are

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

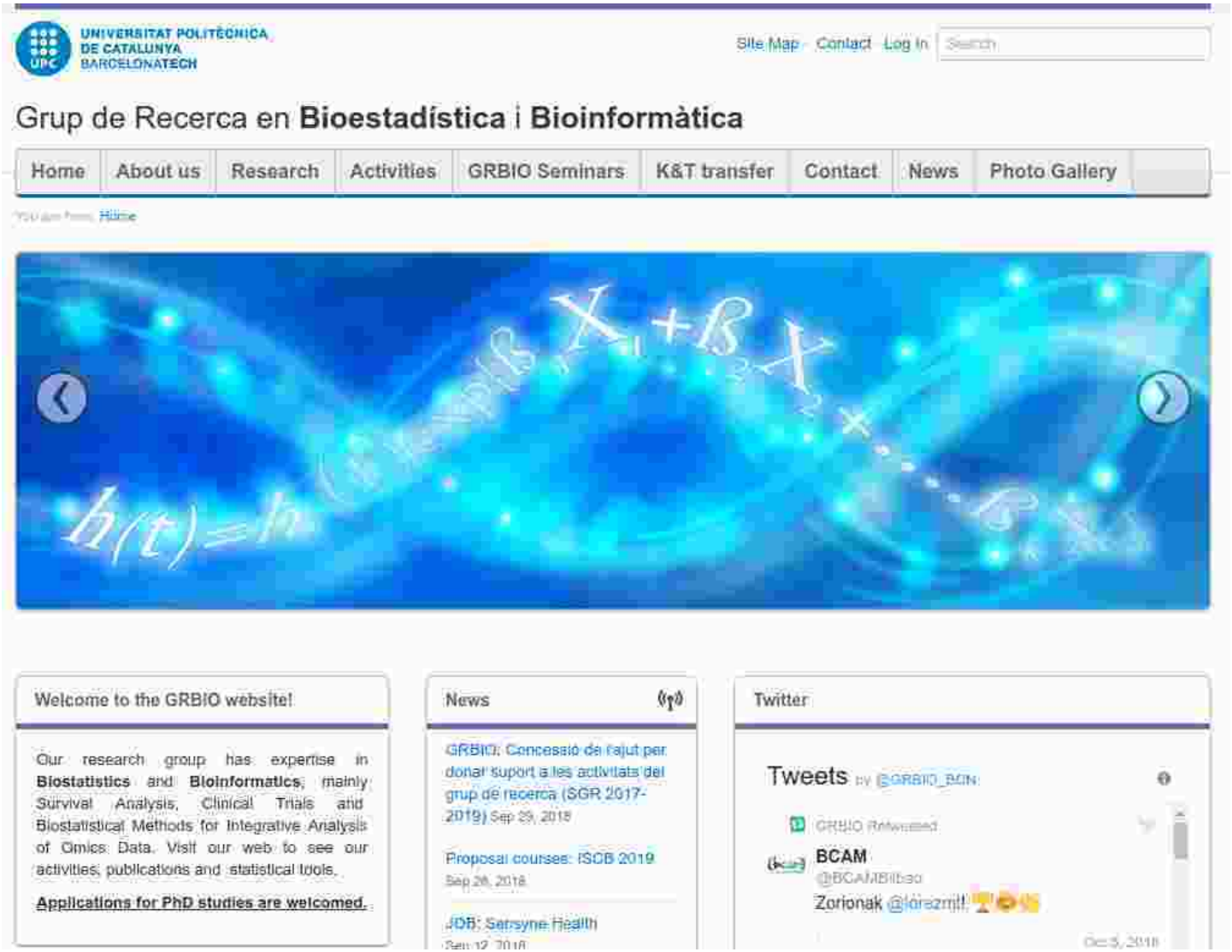
The UEB was created in 2005 within the Research Institute of the Hospital Vall d'Hebron in order to promote the use and development of modern statistical and bioinformatics resources on research performed in its environment.

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

Who are we (2): Teachers



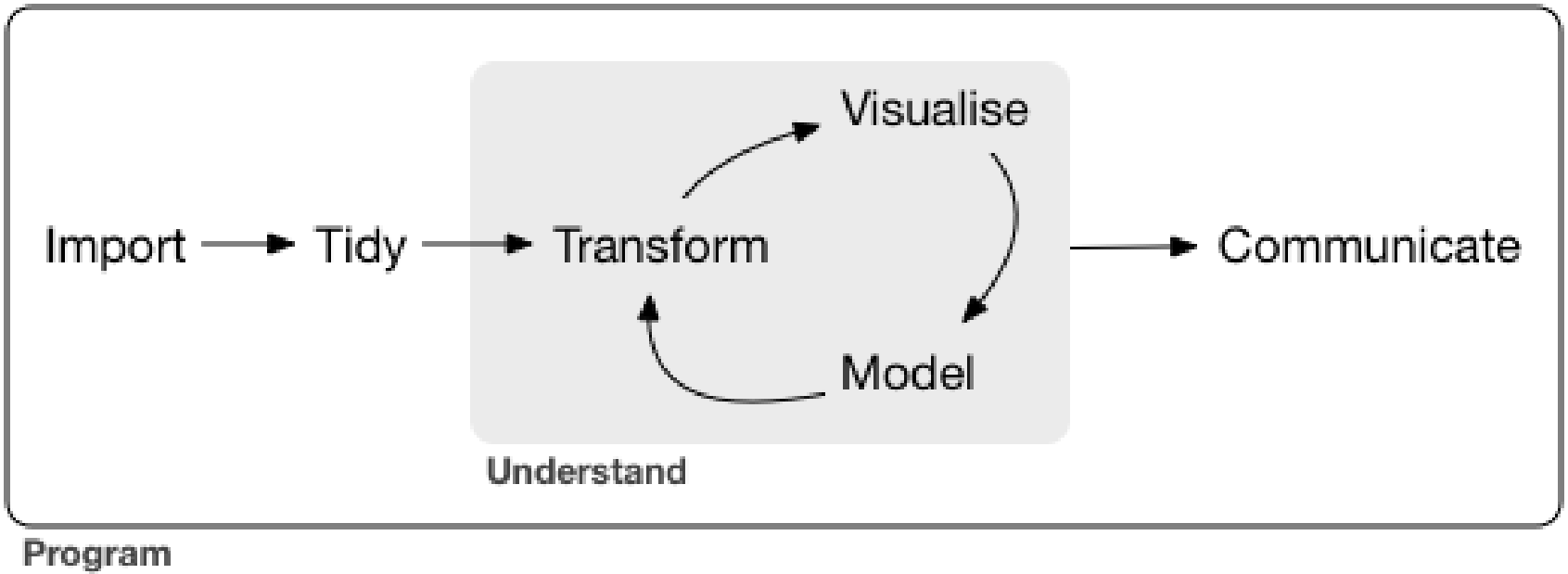
Who are we (3): The GRBio Research group



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 Sciendce



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
  (i) Time
  (ii) Study, and
  (iii) 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)

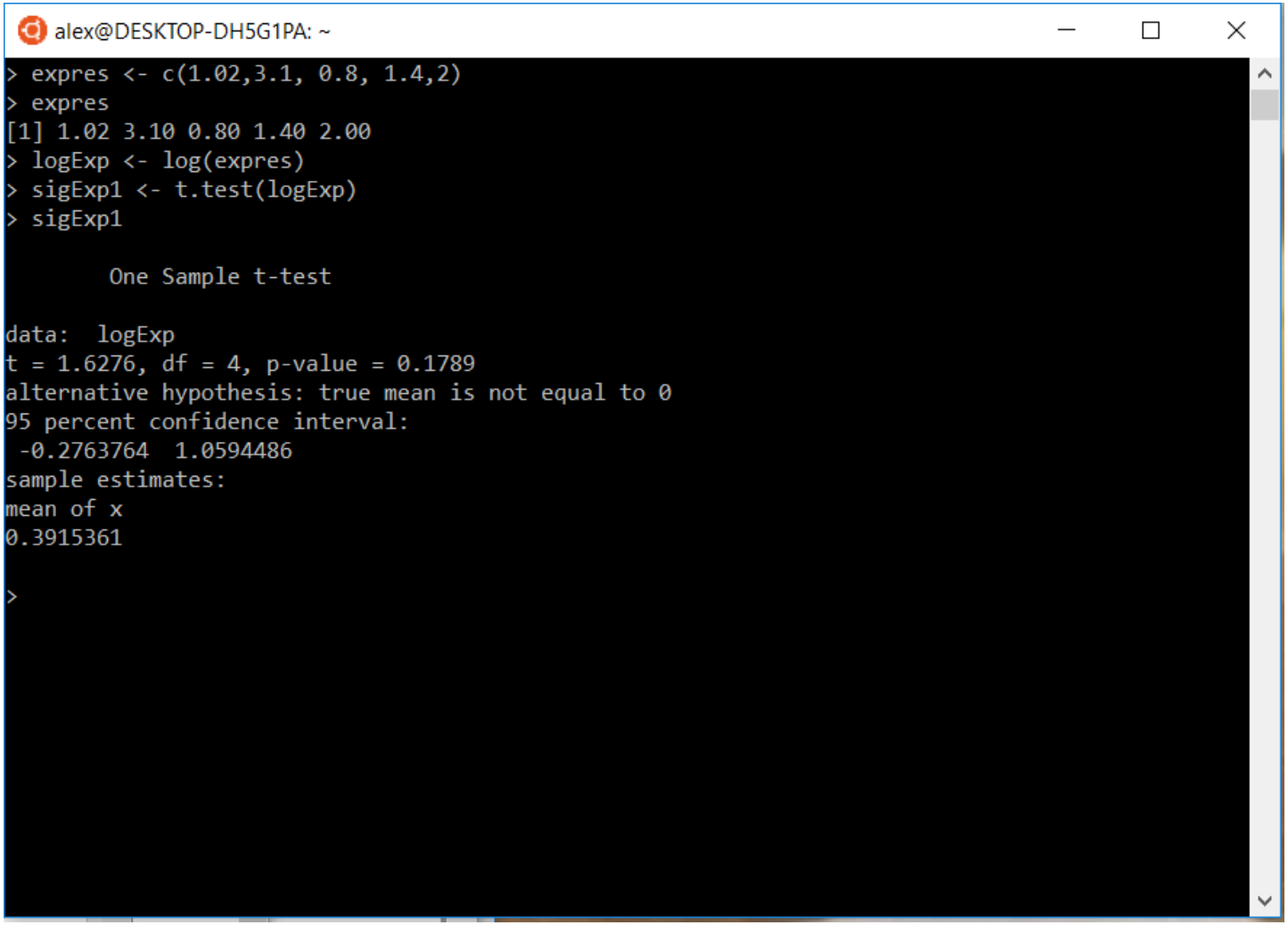
R CON's

- R is mainly used issuing commands from a console
  - less user friendly that 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

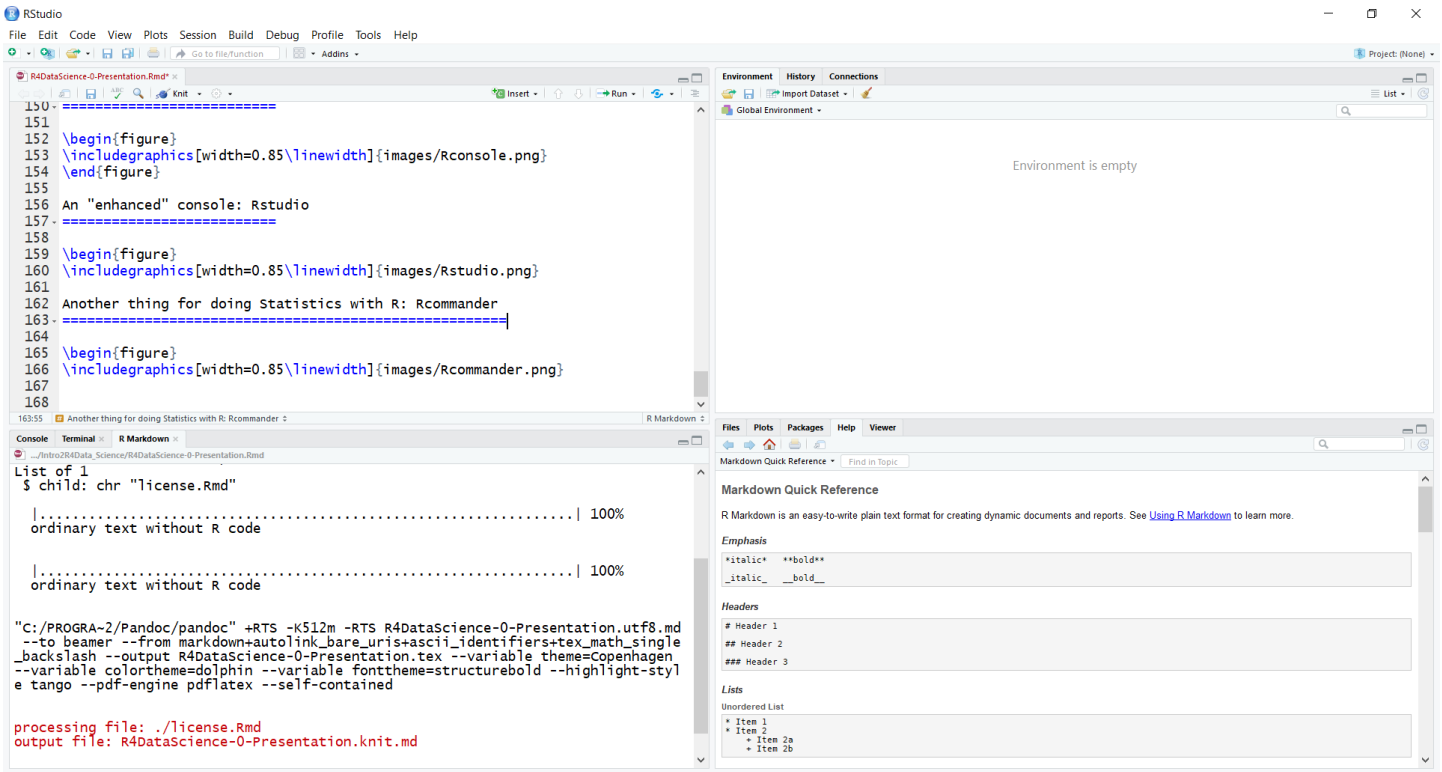
Using R

- 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 concetrate an Statistics, not in commands.

A raw R console in linux



An "enhanced" console: Rstudio



Another thing for doing Statistics with R: Rcommander