The University Of Sheffield.

Introduction to Data Handling in R (BMS 109)

Instructor: Dr Marta Milo
Demonstrators:

The University Of Sheffield.

Today's Outline

- Introduction of the sessions
- Motivations and your Learning Objectives
- Presentation of the tools
- Concept of script
- Get started with Rstudio

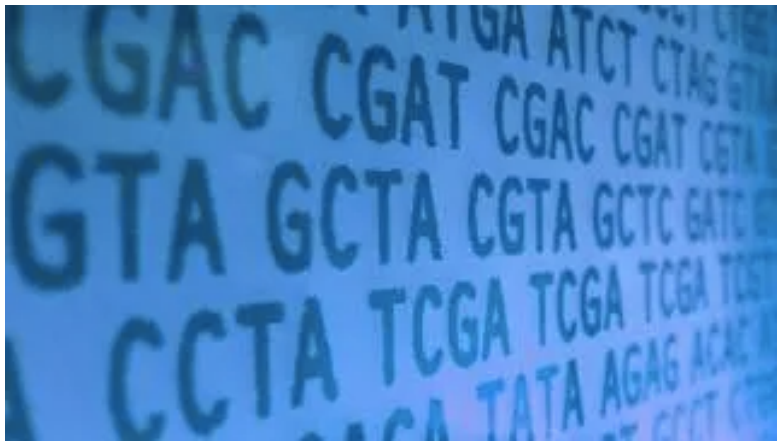
The University Of Sheffield.

What is all about?

These sessions are dedicated to introduce you to data handling using programming languages. We will use R within the R studio Environment.

Biology is today a Data-Rich discipline. Many different type of data are used to characterise and study biological systems and diseases.

Genomics has revolutionised how we study and prevent diseases, creating a paradigm shift in the way we study and think in biology.



The University Of Sheffield.

What are the learning outcomes for these sessions?

- Making you aware of appropriate data handling and presentation techniques for optimal data analysis;
- Acquire skills in presenting, interpreting and visualise data with R programming;
- Be aware of issues related to data handling;
- Be able to import/export data from R enviroment, manipulate data and produce high standard plots using R scripts;
- Enhance your problem solving skills;
- Acquire new transferable skills.

The University Of Sheffield.

How will you be learning?

- Practical lab sessions
- Writing simple scripts for data analysis during practical classes
- Group discussion and forum, interaction in class
- Banging your head on the computer ..
- Giving yourself time to adapt to this new way of thinking…
- Trying by yourself, explore and stretch your imagination…

The University Of Sheffield.

What will you gain from learning data analysis in R?

- New qualifications that will increase your employability
- A new set of transferable skills, like programming and awareness of requirements for data analysis.
- Learning a new terminology and new interdisciplinary skills

The University Of Sheffield.

Sessions Outline

The teaching consists of 1.5 hours of weekly practical lab sessions ending in week 6.

In the lab classes we will use coding to manipulate a plot data. You will follow the instructions and try to solve the exercises. The TAs are here to help.

Lab classes are split in three groups following your BMS109 grouping but material is exact same

Coding requires practice, the more the better….

The University Of Sheffield.

Blackboard(MOLE) and website

Data Handling and Visualisation in R (BMS109)

About Overview Tips and FAQs

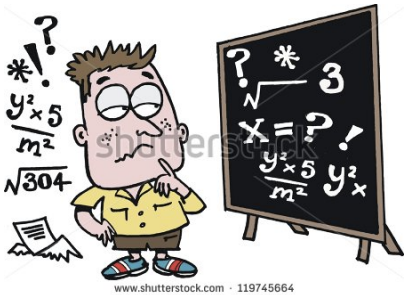
About

This sessions are part of the first year module **BMS109 Introduction to Biomedical Science**, Department of Biomedical Science University of Sheffield. They are designed for first year biology students with little or no knowledge of programming and statistics, to enhance their data handling skils and to provide an introduction to the use of R programming for data analysis.

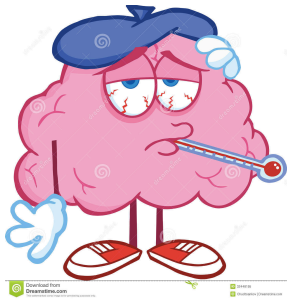
The University Of Sheffield.

The tools we will use

A popular programming language in areas such as bioinformatics, statistics and data analysis.



We will use our brain to create new knowledge



The University Of Sheffield.

What is R, Who uses it and Why

What

- "R" is shorthand for "GNU R”:
- An interactive programming language
- Focus on data analysis: data manipulation, data visualisation, and statistics
- "R" is also shorthand for all the associated applications and environment around this language

Who

- Ordinary useRs like me and you
- Bloggers
- Book authors
- Package developers

Why

Learning to use R will make you more efficient and facilitate use of advanced data analysis tools

The University Of Sheffield.

What is R Studio?

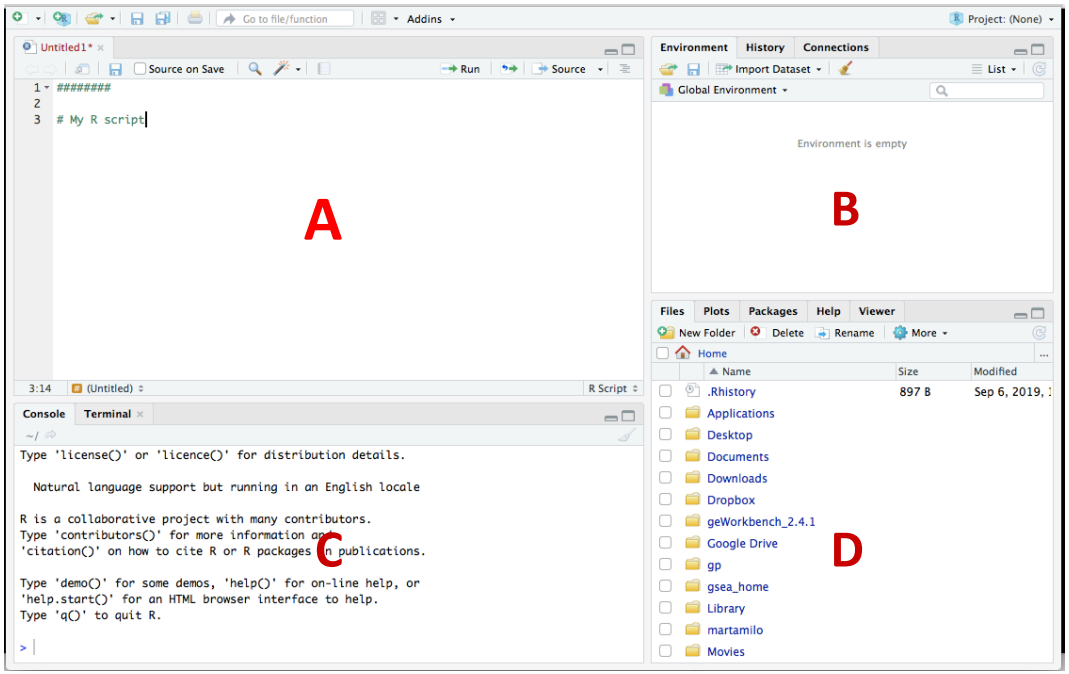
RStudio is an Integrated Development Environment. Makes working with R much easier:

Console to run R
Editor to work with scripts
Package management
Many more features…

R and Rstudio are two different things….

The University Of Sheffield.

The structure of R Studio



A :Script Editor

B :Global Environment and History

C: Console

D :Everything else (e.g. file browser)

We will explore some of the core features of RStudio in this course. We are only going to scratch the surface of what RStudio can do.

The University Of Sheffield.

Getting started

Let's get ready to use R and RStudio. Do the following:

1. Open up RStudio All Programs > Rstudio

2. Maximise the RStudio window (always do this!)

3. Click the Console pane

4. Type 3 + 2 at the prompt >

Hit Enter
What happened?

```
[1] 5
>
```

The University Of Sheffield.

Arithmetic operations

R is like a big calculator. It has all the arithmetic operators you would expect to see:

+: addition
-: subtraction
*: multiplication
/: division
^: exponentiation

Remember:
Exponentiation first, multiplication and division next, and then addition and subtraction. Use the round parenthesis ( ) to change the order of evaluation and to group calculations.

The University Of Sheffield.

Exercises: Basic Arithmetics

Do the following in one step in R:

Add the numbers 2 and 3
Multiple the resulting number by 2
Take the square of the second result.

```
> ((2 + 3) * 2) ^ 2
[1] 100
>
```

Now type the ↑ key. What happened?

The University Of Sheffield.

Exercises: Basic Arithmetics (2)

Combining calculations

Pythagoras's theorem for the length of the hypotenuse of a right-angled triangle has the following formula:

$$LC = \sqrt[2]{LA^2 + LB^2}$$

Use this to equation to calculate the length of the hypotenuse when LA=2 and LB=3. Make sure R does all the work for you.

Hint: Remember that the square root of a number x is equal to ^1/2

```
> (2^2 + 3^2)^(1/2)
[1] 3.605551
>
```

The University Of Sheffield.

Assignment

We usually want to do more with R than just perform a few simple calculations.

We then need a way to store intermediate results. We do this with the **assignment** operator: **<-**.

There will always be two parts to an assignment, sitting either side of the <-:
Right hand side: any valid R expression
Left hand side: the name to assign the result to

```
> x <- 3
> x
[1] 3
>
```

In this example x is called the **name**. We have **assigned** it with the value 3

The University Of Sheffield.

Assignment (2)

```
> x <- 2
```

What happened to the old value of x?

```
> x
 [1] 2
```

Once we have assigned a value a name we can reuse it.
You can use the assignment operator with any kind of object (characters or numbers etc.) —not just numbers.

```
> y <- x * 10^2
```

What do you get?

```
> y
[1] 200
```

The University Of Sheffield.

Assignment (3)

Let's look at the Pythagoras's theorem example again.

Assign the value 2 to **a** and the value 3 to **b**.
Rewrite the calculation to evaluate the length of the hypotenuse as an R expression using **a** and **b.**

Remember: $LC = \sqrt[2]{LA^2 + LB^2}$

```
> a <- 2
> b <- 3
> (a^2 + b^2)^(1/2)
[1] 3.605551
```

The University Of Sheffield.

The Script

We usually need a set of instructions to perform a task.

These instructions follow a logical sequence that is needed to obtain the output.

We store all these instruction is a file called **script,** which is nothing else that a notebook (a txt file) in which you write in a “computer language”.

In R Studio we save them with the extension .R and are called **source files**.

It is very important though that you add natural (human) language in this scripts to help you to describe the different steps of your “**workflow**” (the logical steps) that took you to the output.

The University Of Sheffield.

The Script (cont…)

Suitable script to load script

```
library(Rsubread)
### RNAseq Analysis of NOVOGENE dataset
fastq.files <- list.files(path = "Data/raw_data", pattern = "*fq.gz$", full.names
= TRUE)
#### Alignment ######
# load the Rsubread package
library(Rsubread)
RCh38<-buildindex(basename="GRCh38",reference="GRCh38_latest_genomic.fna",
                  gappedIndex=TRUE,indexSplit=TRUE, memory=4000)
# listing the files in the data folder
fastq.files <- list.files(path = "Data/raw_data", pattern = "*fq.gz$", full.names
= TRUE)

# Printing values to verify that are all loaded.
fastq.files


##### Build the index for the human genome ###
GRCh38<-buildindex(basename="GRCh38",reference="GRCh38_latest_genomic.fna",
                  gappedIndex=TRUE,indexSplit=TRUE, memory=4000)
```

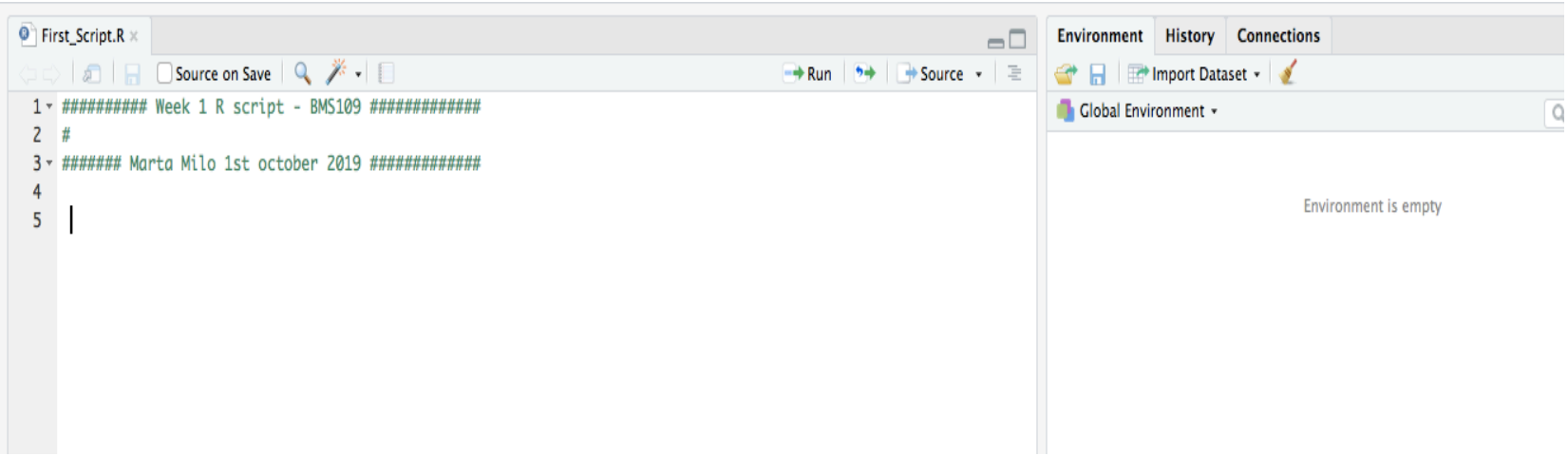
The University Of Sheffield.

Setting up your script

Open a new R script. File ->New File-> R script
Shift+ CTRL+N

Save it, give a name and .R extension. Any other new file will be opened in another tab. Try to open another file.

Add a title and your name with the date of today.



The University Of Sheffield.

Get Started

Exercise: In your newly saved script :

Assign the values "3", "10" and "15" to three different names (**variables**) and perform the following operations storing the results into names (not the same):

- sum all three variables
- take the difference of the first two and divide by the third
- multiply all three together
- take the square value of their sum
- calculate the sum of the vector (all of them) raised to the power of 4
- take the square root (sqrt) of the difference of third and the first

Get Started (cont …)

```
x <-3
y <-10
z <-15
s <-x+y+z
d <-(x-y)/z
m <-x*y*z
v <-s^4
cat("Ratio: ", d, “\n”)
print(paste("Multiplication: ", m))
print(paste("Power of 4: \n", v))
print(sq)
```

Do you see any values produced by the operations?

```
> s
[1] 28
> cat("Ratio: ", d,"\n")
Ratio:  -0.4666667
> print(paste("Multiplication: ", m))
[1] "Multiplication:  450"
> print(paste("Power of 4: \n", v))
[1] "Power of 4: \n 614656"
> print(sq)
[1] 3.464102
```

The University Of Sheffield.

Comment your code

Add some comments in your script to explain your steps

```
# assign values to the variables
x <-3
y <-10
z <-15
```

Save your newly created script and close R studio
Logout from your session on managed desktop