Session I. R introduction

Armando Reyes-Palomares

Department Biochemistry and Molecular Biology

Complutense University of Madrid

armandorp@ucm.es

Sessions Comp. Genomics Laboratory

1. R Introduction
2. Bioconductor
3. Whole Genome Sequencing (WGS)
4. Transcriptomics (i.e. RNA-seq)
5. Epigenomics & Regulatory Genomics

armandorp@ucm.es

What is R?

- It's a free software environment for statistical computing and graphics.

- Multi-platform (i.e. UNIX, Windows and Mac).

- Collaborative and Open-Source Project which can be downloaded at: https://cran.r-project.org

  ***C**omprehensive **R** **A**rchive **N**etwork (**CRAN**) repository over 7,801 packages (libraries)*

Table of contents

armandorp@ucm.es

Goals

- Acquire basic skills in R to process, manage, analyze and visualize genomic data.
- Understand benefits and limitations of R in data science.
- Write R scripts.

armandorp@ucm.es

What can I do with R?

…What can I do with Excel?

vs

Excel

armandorp@ucm.es

What can I do with R?

- Process, manage and store any kind of data.
  - Interoperates with 3rd party software for Big Data
- Statistical analysis.
  - Descriptive statistics
  - Inference statistics, use models to infer new things from the data.
- Display data using powerful and flexible graphics
  - Save graphics/**plots** into PDF or image files, i.e. any plot of your TFG can be done in R.

armandorp@ucm.es

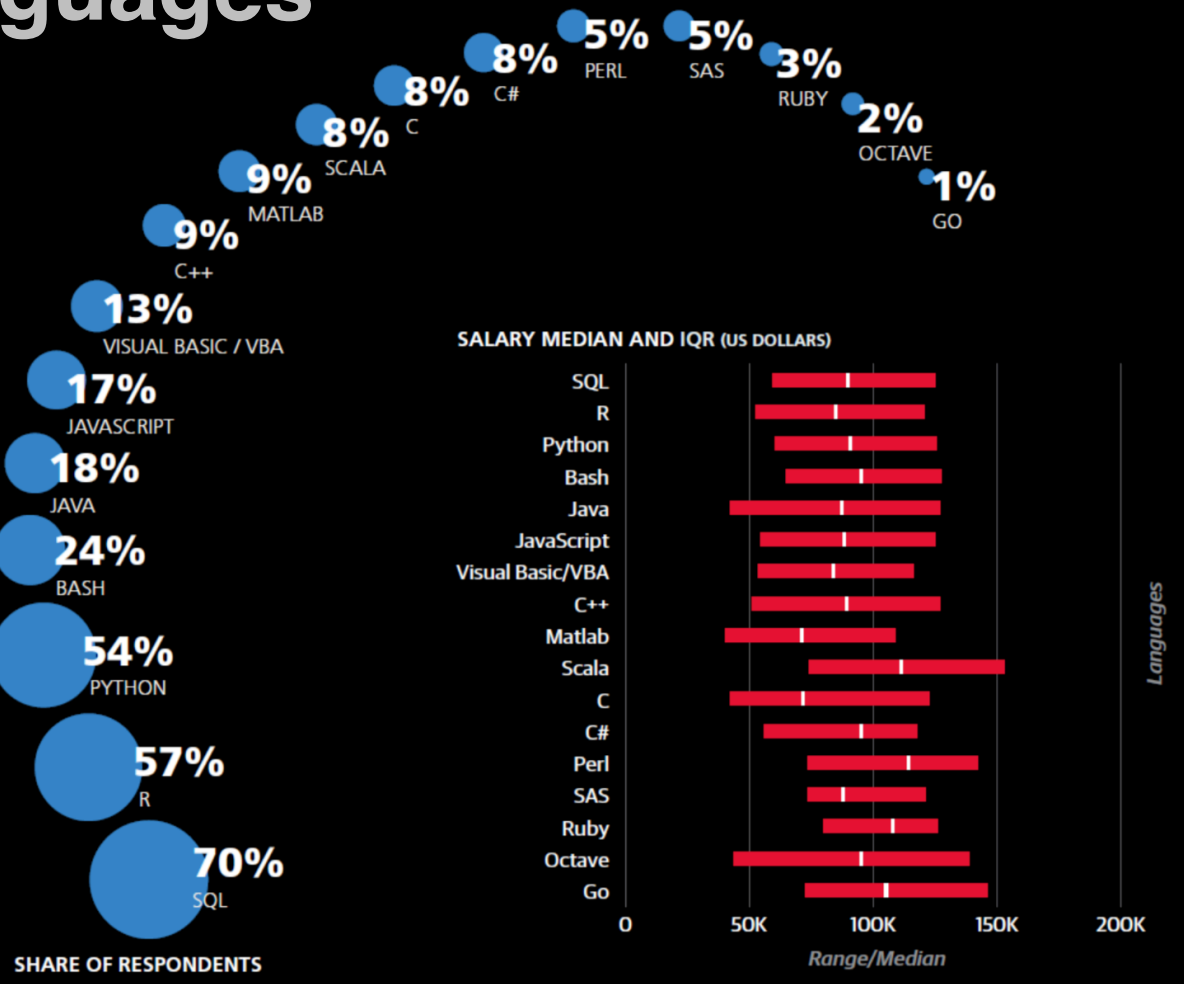
What can I do with R in Bioinformatics?

- **Analyze High-throughput Data** (Genomics, Proteomics, Epigenomics, Metabolomics).
- **Standard connection to updated** data bases of molecular biology, medical genetics, pharmaceutical.
- Make **reproducible** your research analyses.
- Build models to test new hypothesis.
- **Play with data**, development of **integrative approaches** in research.

"From Genome Sequences to Protein-Drug interactions"

armandorp@ucm.es

Science and Job market data science languages



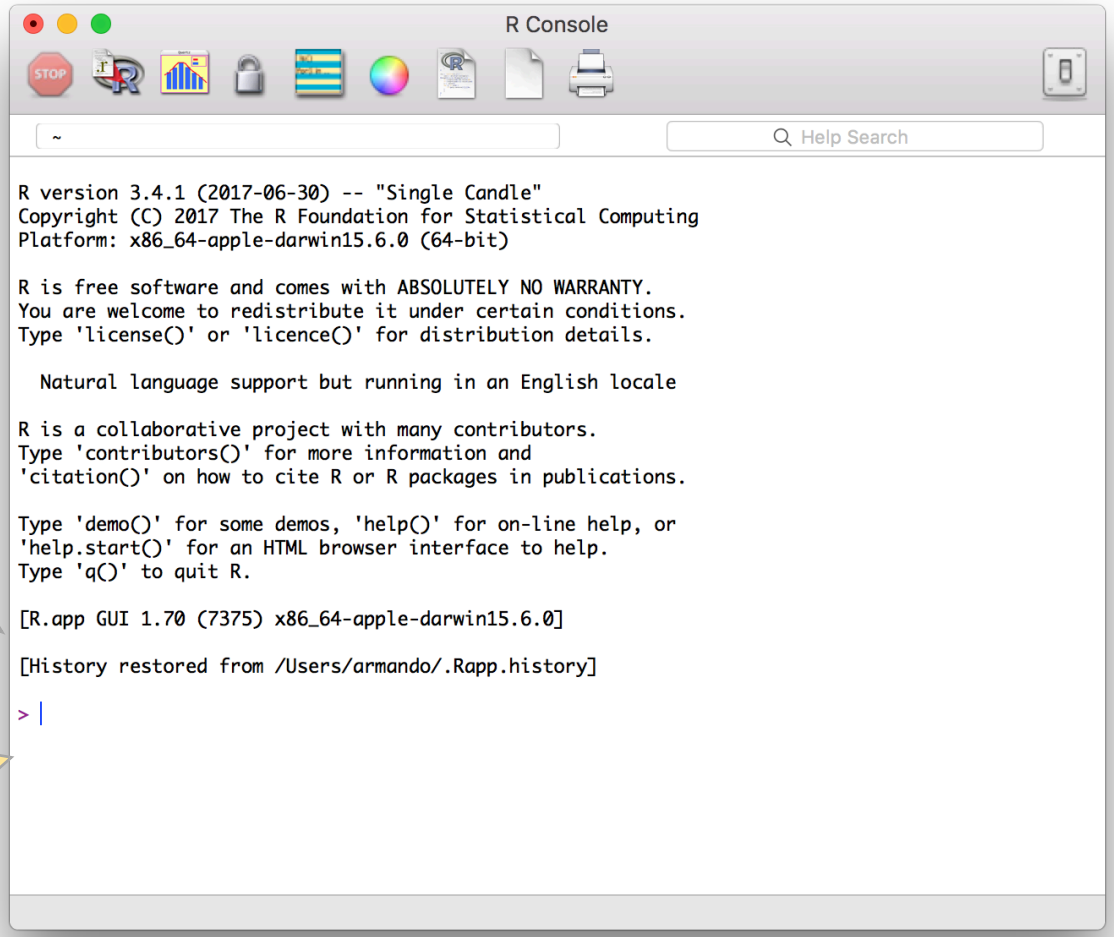
O'Reilly 2016
DATA SCIENCE
SALARY SURVEY

Limitations in R

- Some packages have low quality and are not error free.
- Not memory management, slow and less efficient compared to others languages
- Lack of documentation, but there is no language free of this issue.

armandorp@ucm.es

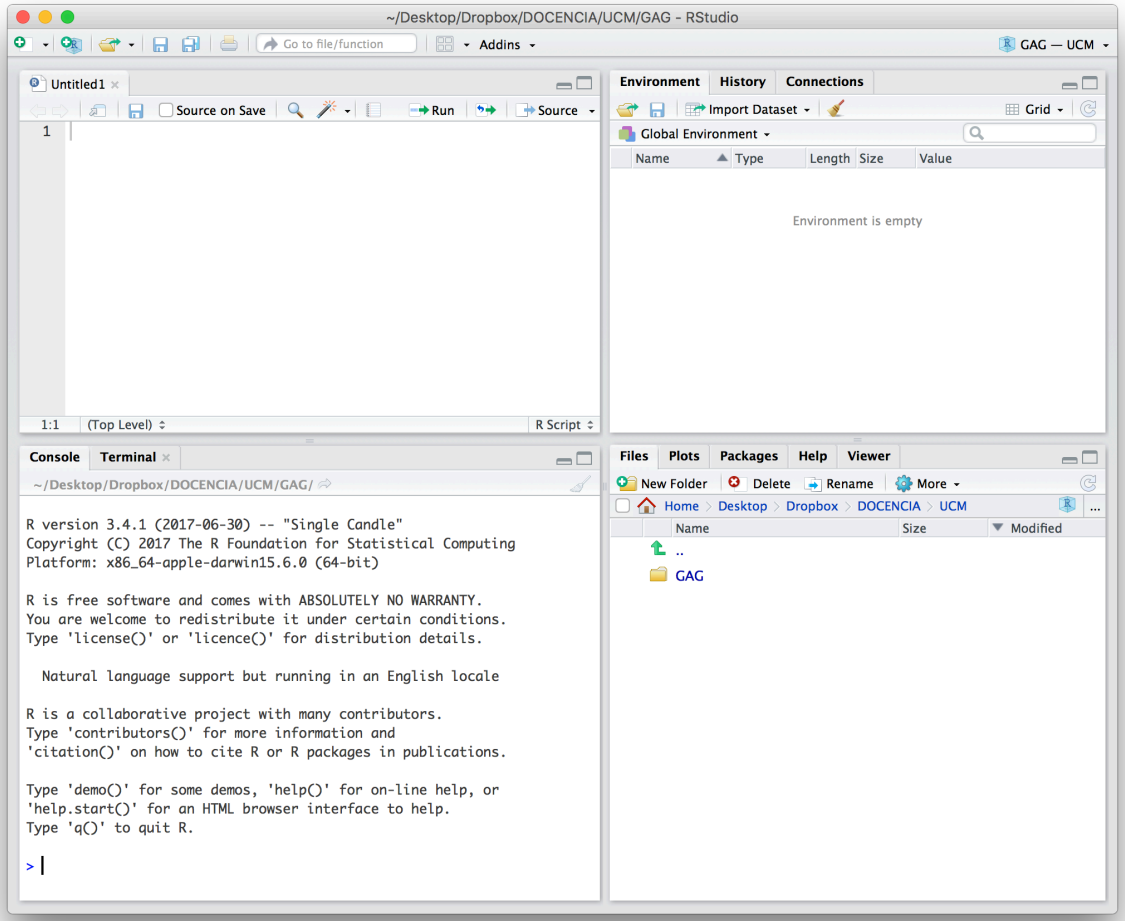
R Console (Graphical User Interfaz)



> means R is waiting for your instructions

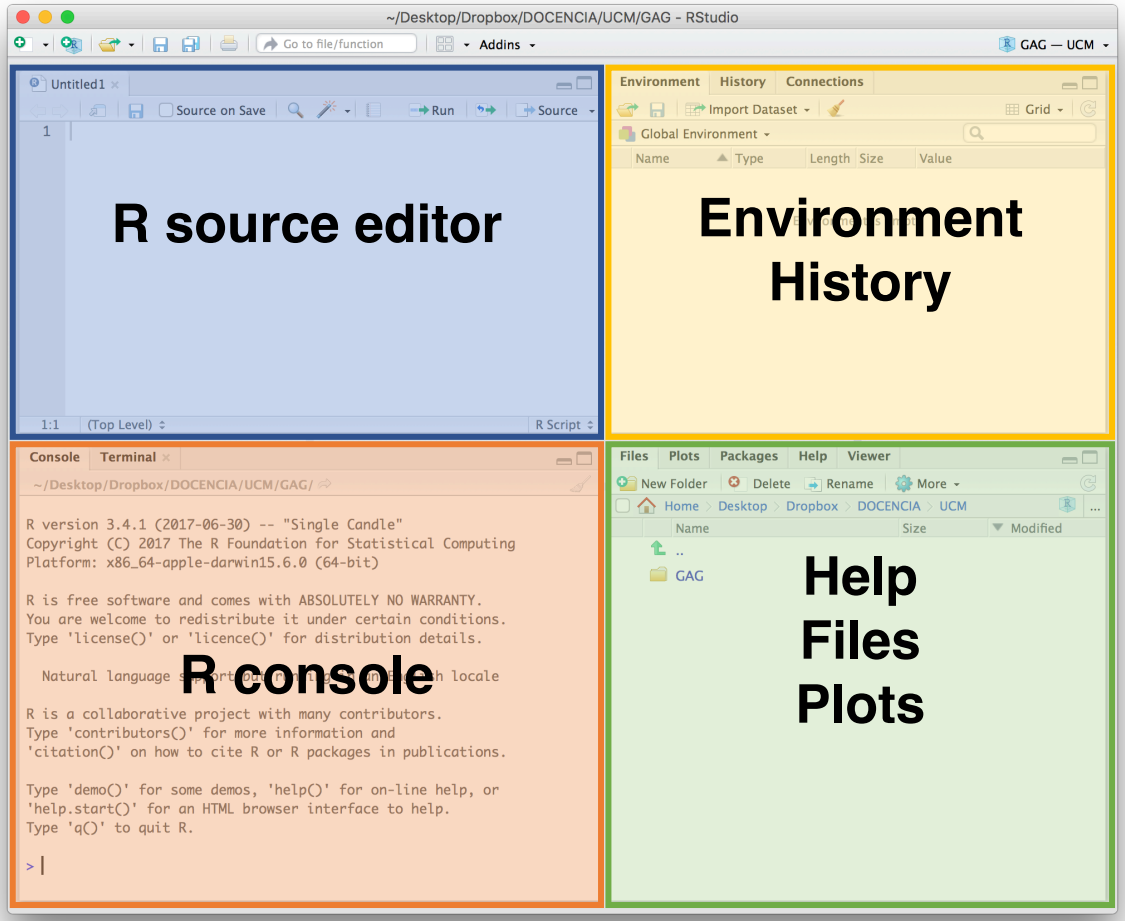
Type:
- **help()**
- **help.start()**
- **q() to quit R**

R IDEs (RStudio, www.rstudio.com)
"Entorno de Desarrollo Integrado"



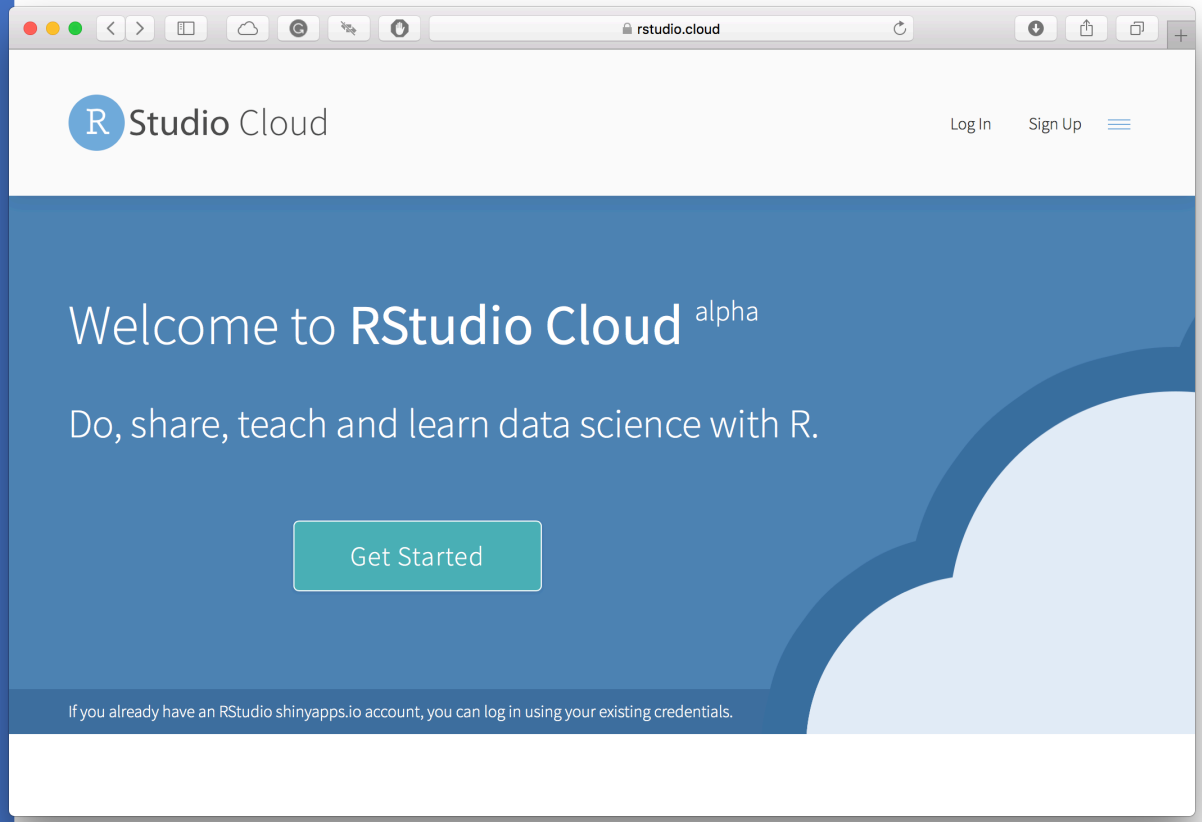
armandorp@ucm.es

R IDEs (Rstudio, www.rstudio.com)



Working with RStudio Cloud

1. **Use UCM account**
2. **Make a project**
3. **Learn how to import files.**



UNIVERSIDAD COMPLUTENSE MADRID

armandorp@ucm.es

R Objects

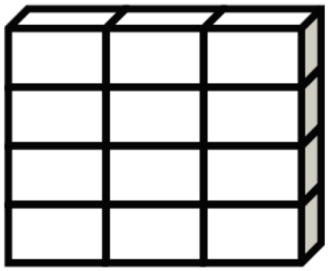
- Vector (a value, a list of values of the same class)
- List
- Matrix
- Data Frame
- Function (commands)
- Array
- Others…

armandorp@ucm.es

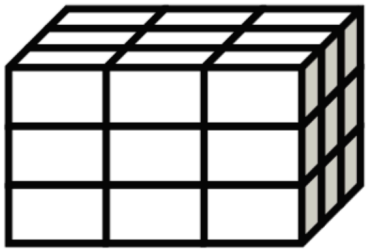
Data structures in R

(a) Vector

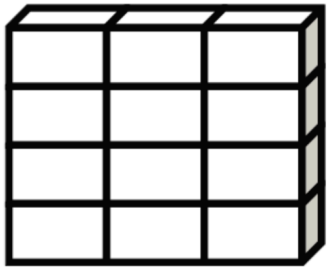
(b) Matrix



(c) Array



(d) Data frame



(e) List

- Vectors
- Arrays
- Data frames
- Lists

armandorp@ucm.es

R Commands (use Cheat Sheet)

- Objects
- Functions (Commands) + Arguments
- Assignment operations
- Expression operations
- Variables; objects where we store data.
- Comments
- Control structures

armandorp@ucm.es

Base R
Cheat Sheet

Getting Help

Accessing the help files

?mean
Get help of a particular function.
help.search('weighted mean')
Search the help files for a word or phrase.
help(package = 'dplyr')
Find help for a package.

More about an object

str(iris)
Get a summary of an object's structure.
class(iris)
Find the class an object belongs to.

Using Packages

install.packages('dplyr')
Download and install a package from CRAN.

library(dplyr)
Load the package into the session, making all its functions available to use.

dplyr::select
Use a particular function from a package.

data(iris)
Load a built-in dataset into the environment.

Working Directory

getwd()
Find the current working directory (where inputs are found and outputs are sent).

setwd('C://file/path')
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors

Creating Vectors

c(2, 4, 6)	2 4 6	Join elements into a vector
2:6	2 3 4 5 6	An integer sequence
seq(2, 3, by=0.5)	2.0 2.5 3.0	A complex sequence
rep(1:2, times=3)	1 2 1 2 1 2	Repeat a vector
rep(1:2, each=3)	1 1 1 2 2 2	Repeat elements of a vector

Vector Functions

sort(x)
Return x sorted.
rev(x)
Return x reversed.
table(x)
See counts of values.
unique(x)
See unique values.

Selecting Vector Elements

By Position

x[4]	The fourth element.
x[-4]	All but the fourth.
x[2:4]	Elements two to four.
x[-(2:4)]	All elements except two to four.
x[c(1, 5)]	Elements one and five.

By Value

x[x == 10]	Elements which are equal to 10.
x[x < 0]	All elements less than zero.
x[x %in% c(1, 2, 5)]	Elements in the set 1, 2, 5.

Named Vectors

x['apple']	Element with name 'apple'.

Programming

For Loop

```
for (variable in sequence){
   Do something
}
```

Example

```
for (i in 1:4){
   j <- i + 10
   print(j)
}
```

While Loop

```
while (condition){
   Do something
}
```

Example

```
while (i < 5){
   print(i)
   i <- i + 1
}
```

If Statements

```
if (condition){
   Do something
} else {
   Do something different
}
```

Example

```
if (i > 3){
   print('Yes')
} else {
   print('No')
}
```

Functions

```
function_name <- function(var){
   Do something
   return(new_variable)
}
```

Example

```
square <- function(x){
   squared <- x*x
   return(squared)
}
```

Reading and Writing Data

Also see the **readr** package.

Input	Ouput	Description
df <- read.table('file.txt')	write.table(df, 'file.txt')	Read and write a delimited text file.
df <- read.csv('file.csv')	write.csv(df, 'file.csv')	Read and write a comma separated value file. This is a special case of read.table/ write.table.
load('file.RData')	save(df, file = 'file.Rdata')	Read and write an R data file, a file type special for R.

Conditions

a == b	Are equal	a > b	Greater than	a >= b	Greater than or equal to	is.na(a)	Is missing
a != b	Not equal	a < b	Less than	a <= b	Less than or equal to	is.null(a)	Is null

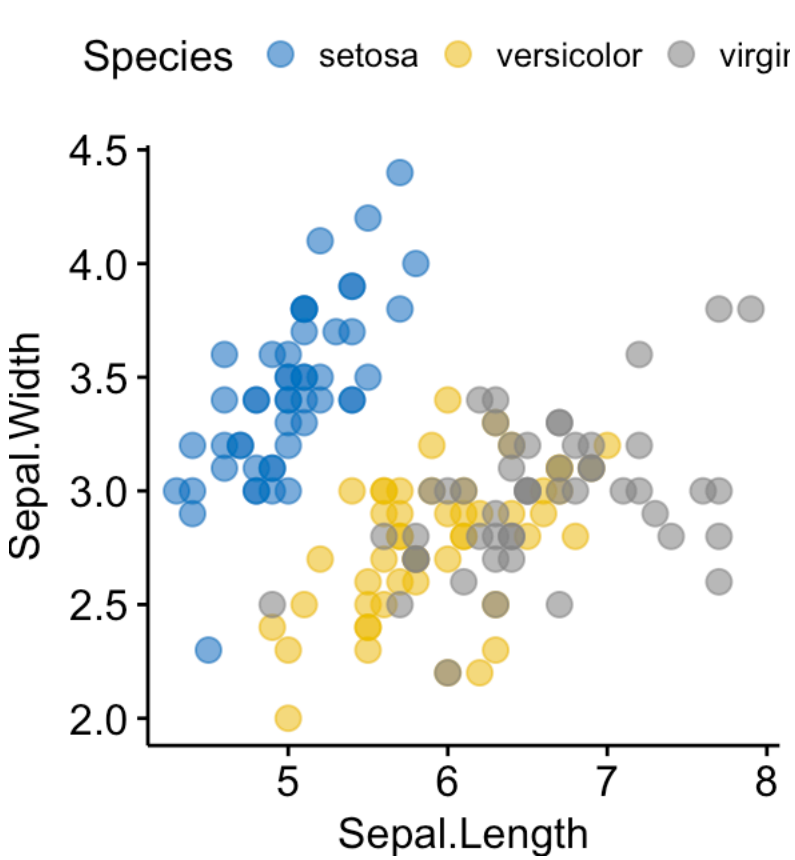
The R Graph Gallery

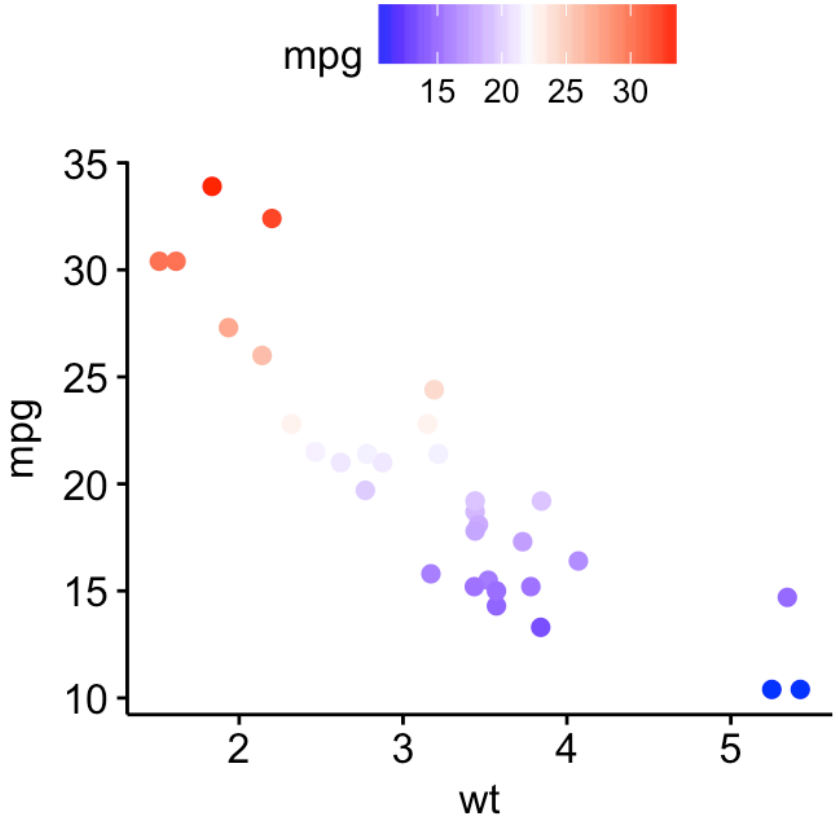
www.r-graph-gallery.com

UNIVERSIDAD COMPLUTENSE MADRID

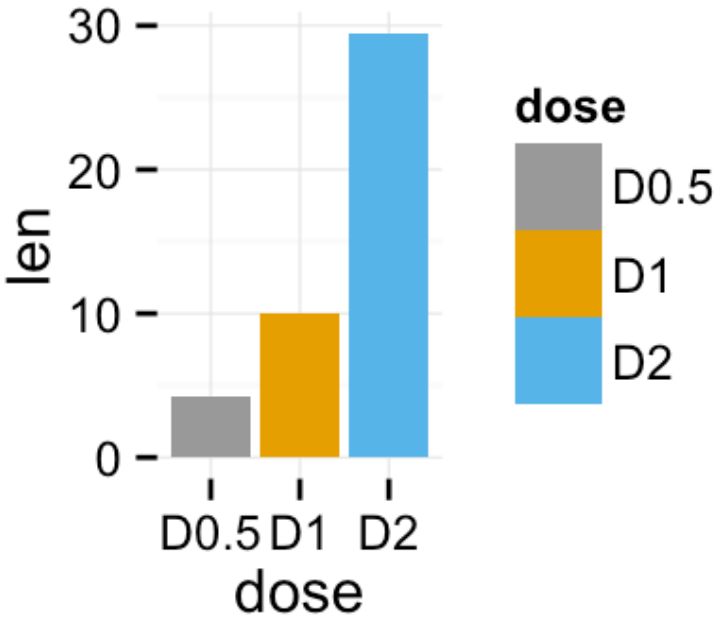
armandorp@ucm.es

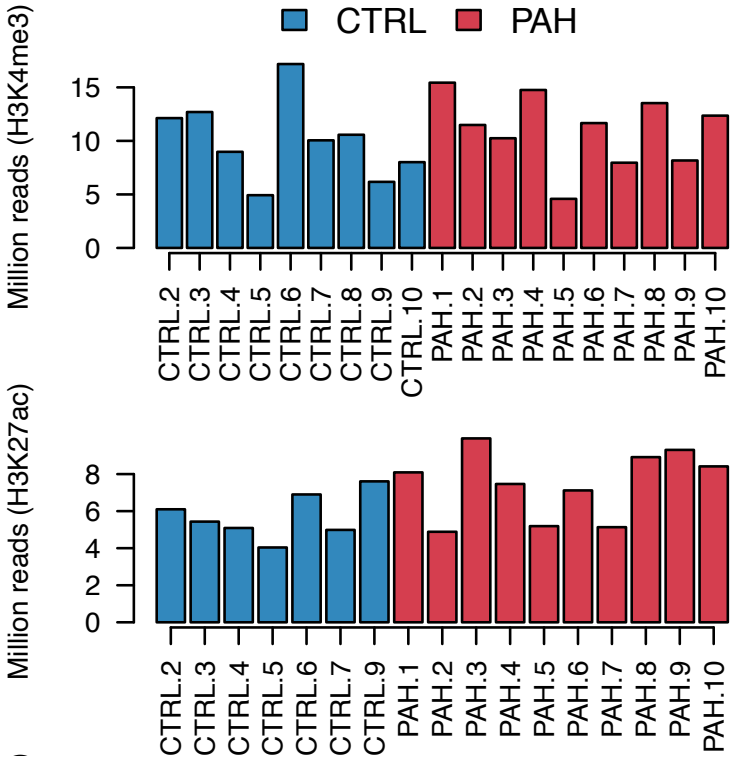
Type of R plots: scatter plot



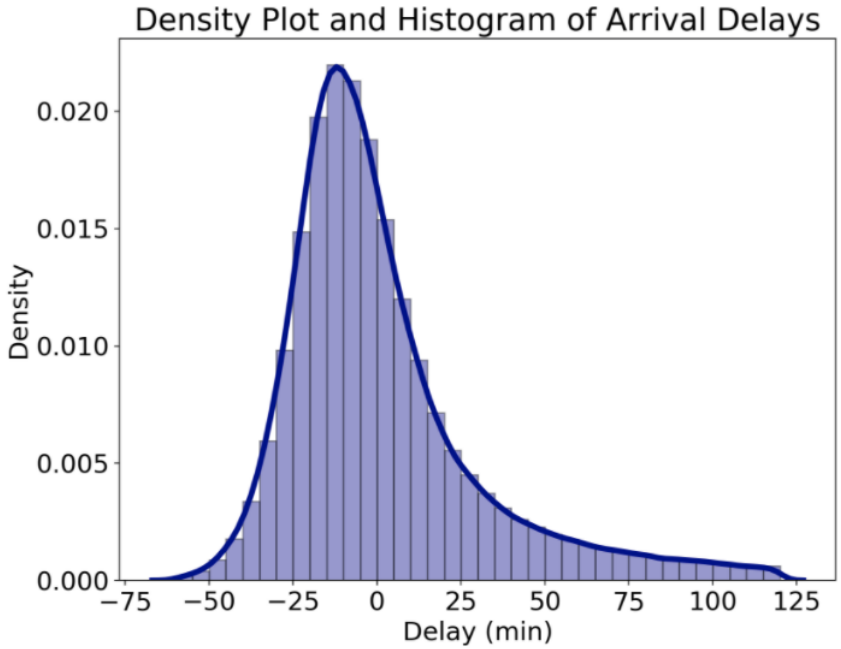


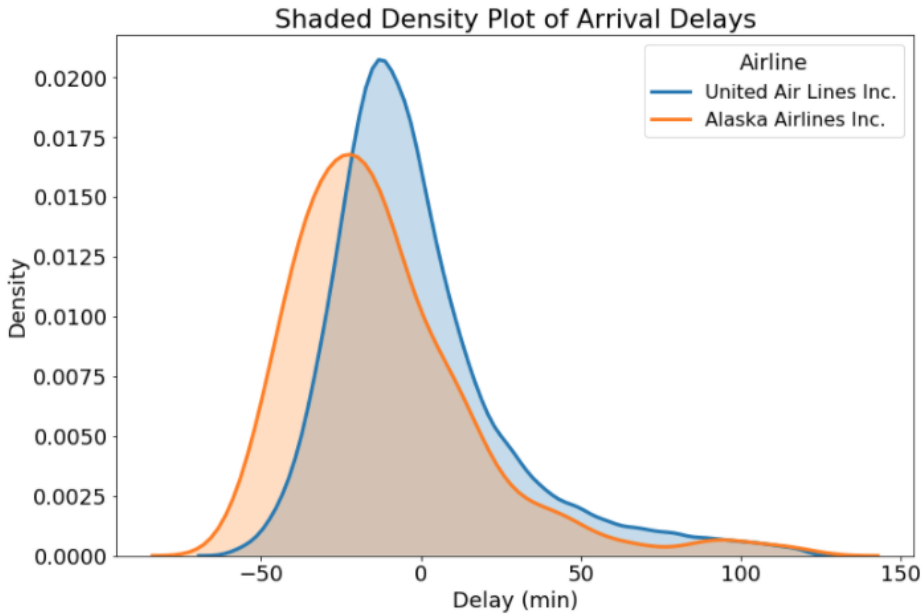
Type of R plots: bar plot





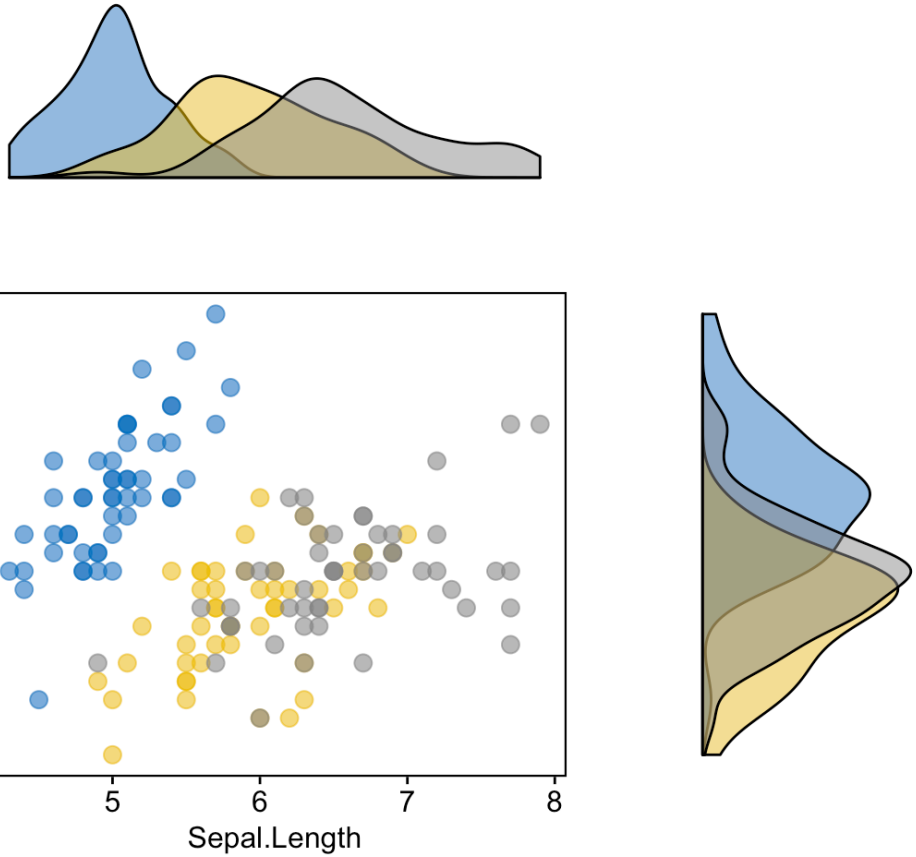
Type of R plots: histograms & density plots



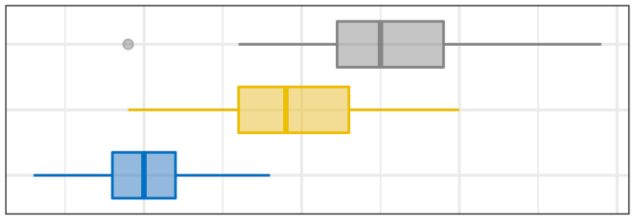


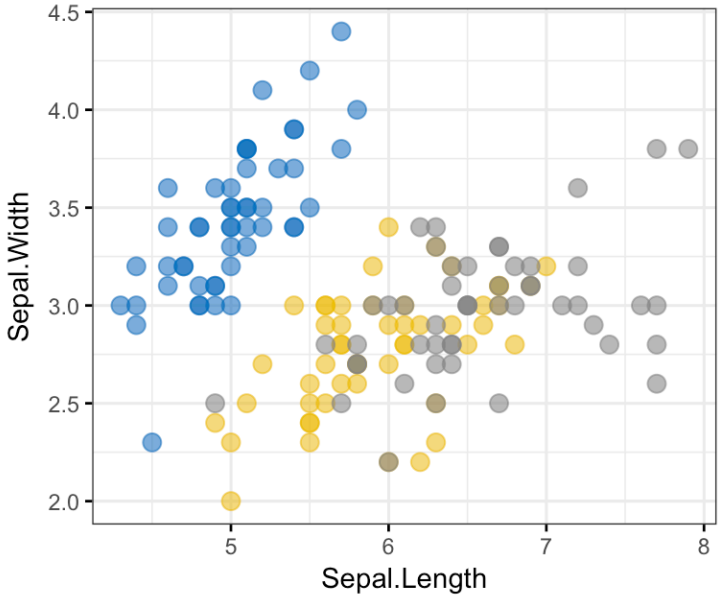
armandorp@ucm.es

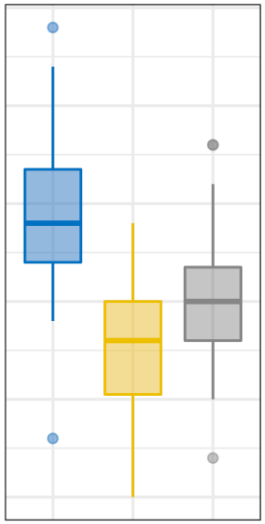
Type of R plots: density plot



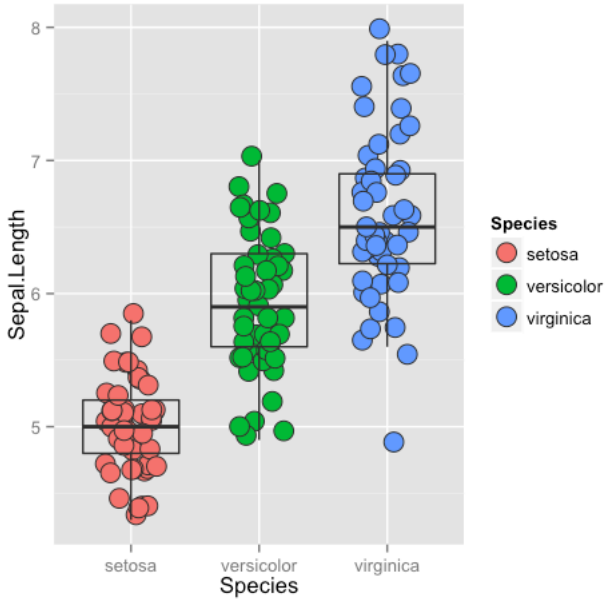
Type of R plots: Boxplot







Strip chart and boxplot



armandorp@ucm.es

GC content

The GC content (percentage) is the number of GC nucleotides divided by the total nucleotides.

$$\frac{G+C}{A+T+G+C} \times 100\%$$

It's relevant for:

- Highly variable across genes or species (Evolution)
- Affects NGS in amplification (Technical bias)
- High GC content around coding regions.

armandorp@ucm.es

Useful references and resources:

Recursos para Prácticas de Genómica Computacional de GAG-UCM

armandorp@ucm.es

2021-09-28

Contenidos de material relacionado con la asignatura de Genomas y Análisis Genómico. Esta web se actualizará con regularidad a lo largo del curso.

Guías, tutoriales y libros de R:

- Introdución a R (inglés)
- R for Data Science Muy recomendable.
- R para Ciencia de Datos en Español
- R Tutor
- R Bloggers
- R Tutorial

Cursos online en Bioinformática

- EBI -Training online

Gráficos en R:

- The R Graph Gallery
- Plotly Gallery

Introducción a Linux:

- Linux para principiantes, 101
- Linux nivel intermedio