



Clase 5: Introducción de R

Evelin González F.
evelyn.gonzalez@uoh.cl

Organización de las clases: Parte 2

Clase 5: Introducción de R.

Clase 6: Librería Maftools para interpretación de variantes.

Clase 7: Análisis de clustering/PCA y categorización de variantes patogénicas, visualización de los datos.

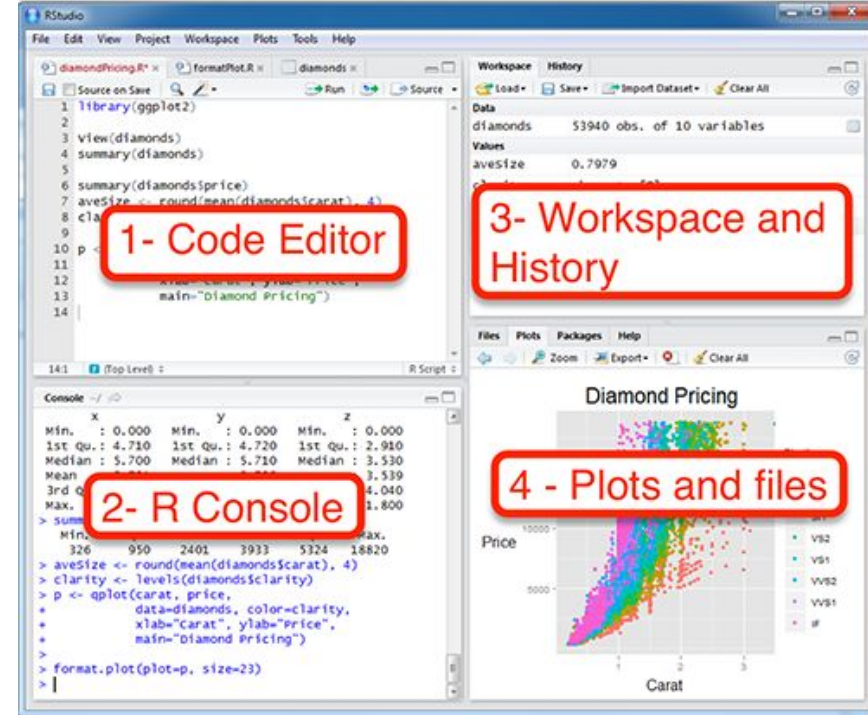
Introducción a R y Rstudio

R es un lenguaje de programación y entorno para el análisis estadístico y la visualización de datos

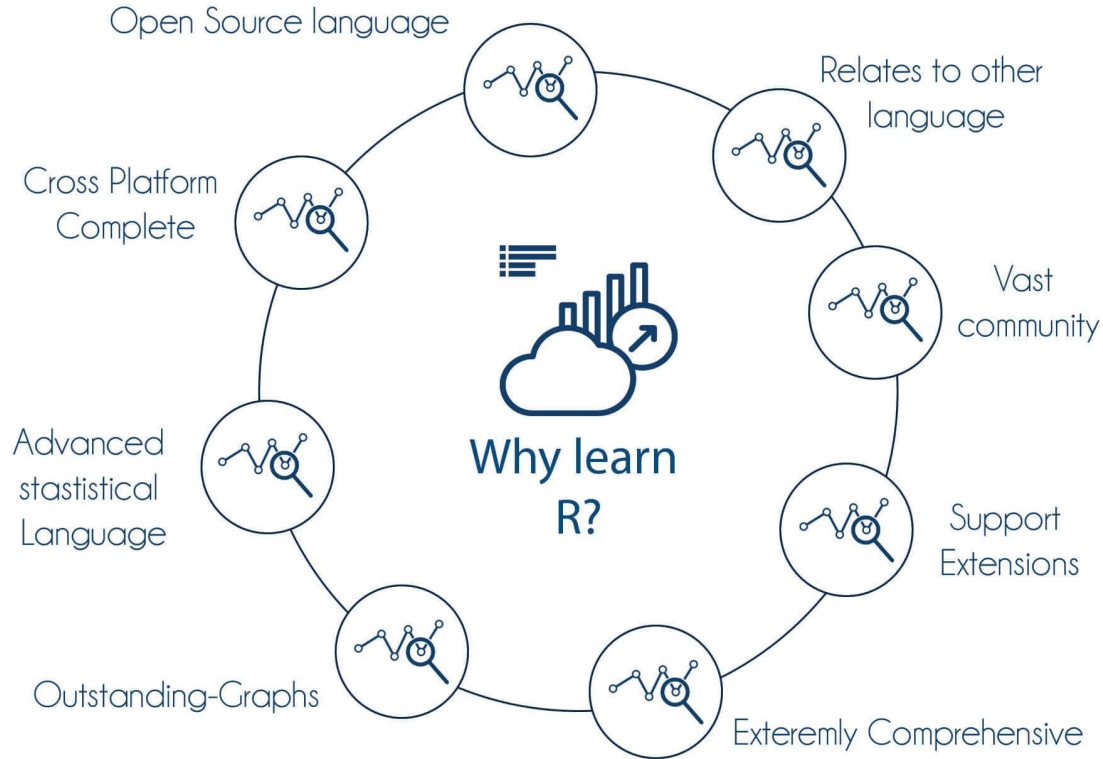
Open Source, multiplataforma y extensible

Por qué usar R?

- Amplio soporte para **estadística**: desde análisis básicos hasta modelado avanzado.
- Potente para **gráficos y visualizaciones**.
- Gran cantidad de paquetes en **CRAN** y **Bioconductor** para aplicaciones específicas.



Por qué aprender R ?



Rstudio

Source Pane

Edit and run scripts (e.g. Rmarkdown templates), and view datasets

Tip: Start new script

Tip: Run script

Environment Pane

Overview of objects (datasets, parameters, lists, etc.) you have imported or created.

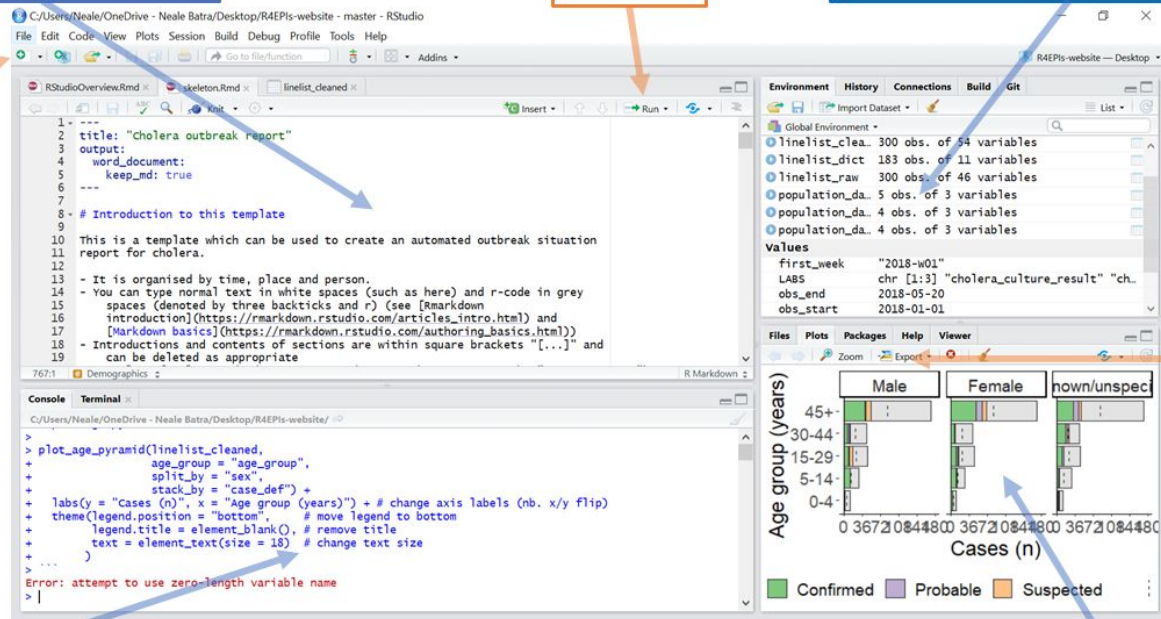
Tip: Zoom and export plots

R Console Pane

R commands run are shown here, and non-graphic output and errors are displayed

Plots, Packages, and Help Pane

Commonly used to view graphics, install packages, and view help

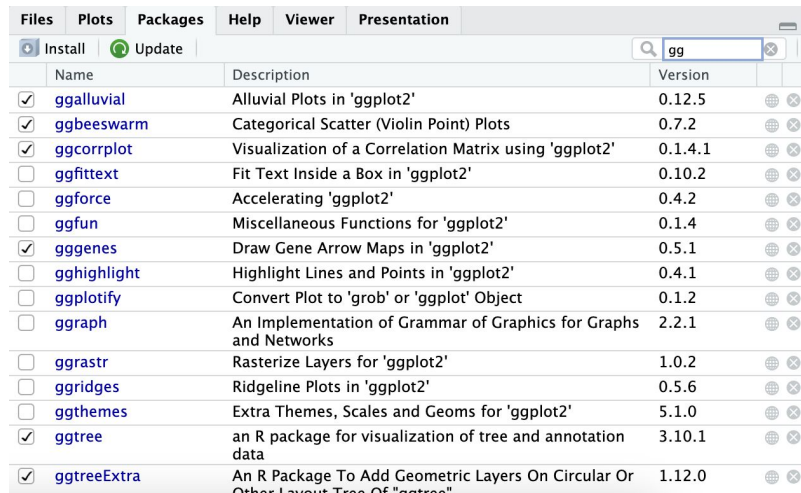


Introducción a R y Rstudio

Paquetes en R

- Instalar (CRAN, devtools)
`install.packages("tidiverse")`
- Import: `library(tidiverse)`
- Recurrentes: `dplyr`, `magnitr`, `ggtree`,
`cowplot`..

```
suppressMessages(library(ape)) #tree basics
suppressMessages(library(treeio)) #join trees
suppressMessages(library(ggtree)) #tree plotting
suppressMessages(library(ggplot2)) #plots
suppressMessages(library(ggtreeExtra)) #geom_fruit
suppressMessages(library(ggnewscale)) #scales in ggplot2
suppressMessages(library(ggstar)) #geom_star for ggplot2
suppressMessages(library(tidyverse)) #strings, dataframes processing
suppressMessages(library(magnitr)) #pipes %>%
suppressMessages(library(tidytree)) #trees as dataframes
suppressMessages(library(RColorBrewer)) #color palettes
suppressWarnings(library(colorspace)) #colors manipulation
suppressWarnings(library(patchwork)) #side-by-side plotting
suppressWarnings(library(openxlsx)) #read excel
```

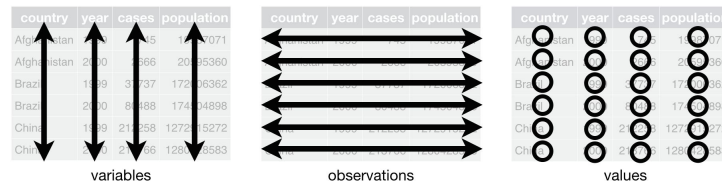


The screenshot shows the RStudio interface with the 'Packages' pane open. A search bar at the top contains the text 'gg'. Below the search bar, a table lists various packages that match the search criteria. The table has columns for 'Name', 'Description', and 'Version'. The packages listed include ggalluvial, ggbeeswarm, ggcorrplot, ggfittext, ggforce, ggfun, gggenes, gghighlight, ggplotify, ggraph, ggrastr, ggribes, ggthemes, ggtree, and ggtreeExtra. The 'ggtree' package is highlighted in blue.

Name	Description	Version
<input checked="" type="checkbox"/> ggalluvial	Alluvial Plots in 'ggplot2'	0.12.5
<input checked="" type="checkbox"/> ggbeeswarm	Categorical Scatter (Violin Point) Plots	0.7.2
<input checked="" type="checkbox"/> ggcorrplot	Visualization of a Correlation Matrix using 'ggplot2'	0.1.4.1
<input type="checkbox"/> ggfittext	Fit Text Inside a Box in 'ggplot2'	0.10.2
<input type="checkbox"/> ggforce	Accelerating 'ggplot2'	0.4.2
<input type="checkbox"/> ggfun	Miscellaneous Functions for 'ggplot2'	0.1.4
<input checked="" type="checkbox"/> gggenes	Draw Gene Arrow Maps in 'ggplot2'	0.5.1
<input type="checkbox"/> gghighlight	Highlight Lines and Points in 'ggplot2'	0.4.1
<input type="checkbox"/> ggplotify	Convert Plot to 'grob' or 'ggplot' Object	0.1.2
<input type="checkbox"/> ggraph	An Implementation of Grammar of Graphics for Graphs and Networks	2.2.1
<input type="checkbox"/> ggrastr	Rasterize Layers for 'ggplot2'	1.0.2
<input type="checkbox"/> ggribes	Ridgeline Plots in 'ggplot2'	0.5.6
<input type="checkbox"/> ggthemes	Extra Themes, Scales and Geoms for 'ggplot2'	5.1.0
<input checked="" type="checkbox"/> ggtree	an R package for visualization of tree and annotation data	3.10.1
<input checked="" type="checkbox"/> ggtreeExtra	An R Package To Add Geometric Layers On Circular Or Other Layout Tree Of "aartree"	1.12.0

Desarrollo en R

Tidydata es esencial



- Inputs: “tibbles” o “data.frames”
- Variables en columnas, observaciones en filas, valores en cada celda
- Estructurar los datos para facilitar su análisis

```
#> # A tibble: 3 x 3  
#>   country    `1999`    `2000`  
#>   <chr>      <int>      <int>  
#> 1 Afghanistan 19987071 20595360  
#> 2 Brazil      172006362 174504898  
#> 3 China       1272915272 1280428583
```



```
#> # A tibble: 6 x 3  
#>   country    year population  
#>   <chr>      <chr>      <int>  
#> 1 Afghanistan 1999      19987071  
#> 2 Afghanistan 2000     20595360  
#> 3 Brazil      1999     172006362  
#> 4 Brazil      2000     174504898  
#> 5 China       1999     1272915272  
#> 6 China       2000     1280428583
```

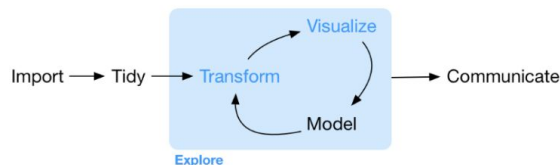
Untidy vs tidy dataframe

Desarrollo en R

Tidyverse

- Data science for tidy data
- Crear un tibble
 - `read.csv()`*
 - `write.table()`*
 - `readr::read_csv()`
 - `readr::write_csv()`
- Package `dplyr` for grammar of data manipulation
- Operaciones y transformaciones

```
tibble(  
  x = 1:5,  
  y = 1,  
  z = x ^ 2 + y  
)  
#> # A tibble: 5 × 3  
#>       x     y     z  
#>   <int> <dbl> <dbl>  
#> 1     1     1     2  
#> 2     2     1     5  
#> 3     3     1    10  
#> 4     4     1    17  
#> 5     5     1    26
```



Comparison operators available are:

- `x == y` – “equal to”
- `x != y` – “not equal to”
- `x < y` – “less than”
- `x > y` – “greater than”
- `x <= y` – “less than or equal to”
- `x >= y` – “greater than or equal to”

More complicated conditions can be constructed using logical operators:

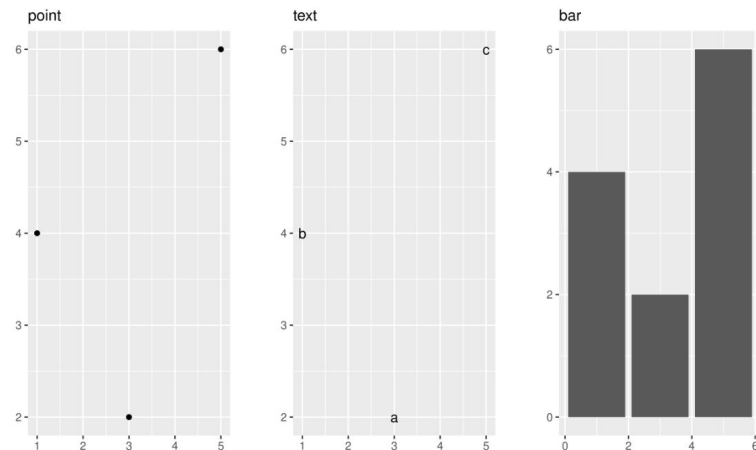
- `a & b` – “and”, true only if both `a` and `b` are true.
- `a | b` – “or”, true if either `a` or `b` or both are true.
- `! a` – “not”, true if `a` is false, and false if `a` is true.

Introducción a ggplot2

Visualización de datos

- Tipos de plots básicos: del set de datos al plot
- `geom_*`: `point()`, `bar()`, `text()`, `line()`..

```
df <- data.frame(  
  x = c(3, 1, 5),  
  y = c(2, 4, 6),  
  label = c("a", "b", "c")  
)  
p <- ggplot(df, aes(x, y, label = label)) +  
  labs(x = NULL, y = NULL) + # Hide axis label  
  theme(plot.title = element_text(size = 12)) # Shrink plot title  
p + geom_point() + ggtitle("point")  
p + geom_text() + ggtitle("text")  
p + geom_bar(stat = "identity") + ggtitle("bar")
```



Introducción a ggplot2

Visualización avanzada

- Formas: `geom_*`
- Aesthetics: `aes()`
 - fill, size, shape
- Colores, apariencia
- label, font size, linewidth/type
- Facetina (split plot)

1 2 3 4 5 6 7 8 9 10 11 12 13

○ △ + × ◇ ▽ ☒ * ⊕ ⊗ ⊛ ⊞ ⊠

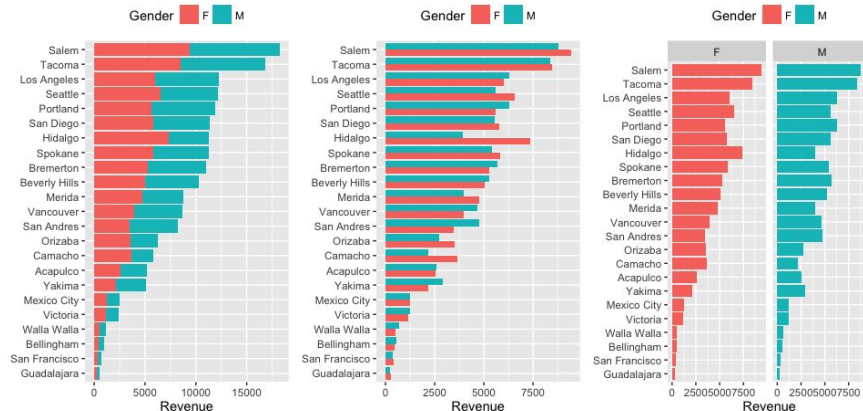
14 15 16 17 18 19 20 21 22 23 24 25

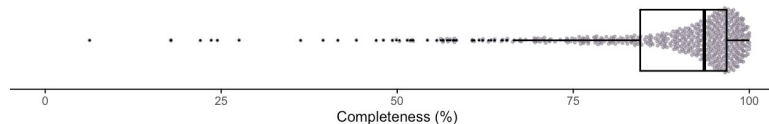
☒ ■ ● ▲ ◆ ● ● ○ □ ◇ △ ▽

```
ggplot(city_gender_rev, aes(City, Revenue, fill = Gender)) +  
  geom_bar(stat = "identity") +  
  coord_flip()
```

```
ggplot(city_gender_rev, aes(City, Revenue, fill = Gender)) +  
  geom_bar(stat = "identity", position = "dodge") +  
  coord_flip()
```

```
ggplot(city_gender_rev, aes(City, Revenue, fill = Gender)) +  
  geom_bar(stat = "identity", position = "dodge") +  
  coord_flip() +  
  facet_wrap(~ Gender)
```

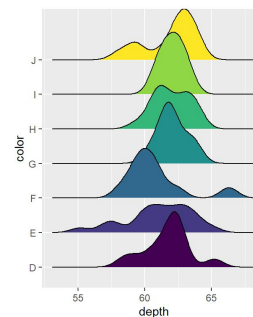




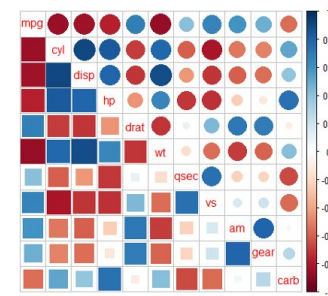
Visualización avanzada

`geom_quasirandom()` + `geom_boxplot()`

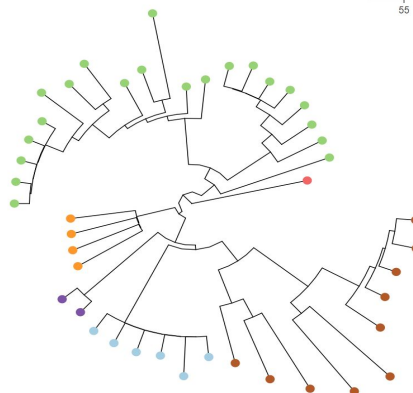
- Bubble plots
- Tree with metadata: `ggtree`, `ggtreeExtra`, `ggnewscale`
- Paquetes avanzados: `ggpubr`, `ggalluvial`, `ggbeeswarm`, `gggenes`, `gghighlight`, `gggridges`, `irlba`, `ggbiplot`, `PCAtools`, `FactoMineR`, `factoextra`, etc..
- Correlation plot (`ggcorrplot`)



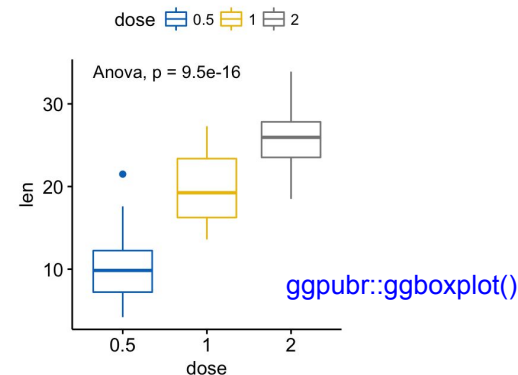
`gggridges()`



`ggcorrplot()`

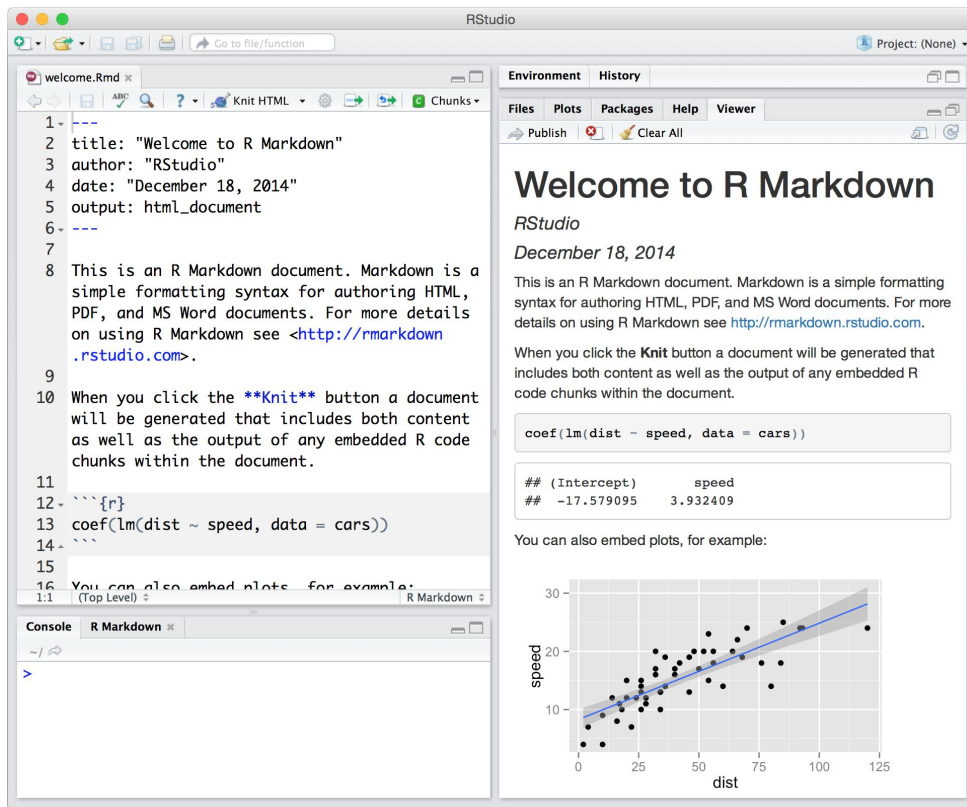


`ggtree()`



`ggpubr::ggboxplot()`

R Markdown



The screenshot shows the RStudio application window. The main editor on the left displays an R Markdown file named 'welcome.Rmd'. The code includes a title, author, date, output format, and several text paragraphs explaining R Markdown. It also contains an R code chunk that calculates the coefficients of a linear model and a text instruction to embed a plot. The right pane shows the rendered HTML output, which includes the title, date, introductory text, the R code output, and a scatter plot of speed vs. distance with a linear regression line. The bottom pane shows the R console with the R prompt.

```
1- ---
2 title: "Welcome to R Markdown"
3 author: "RStudio"
4 date: "December 18, 2014"
5 output: html_document
6- ---
7
8 This is an R Markdown document. Markdown is a
9 simple formatting syntax for authoring HTML,
10 PDF, and MS Word documents. For more details
11 on using R Markdown see <http://rmarkdown
12 .rstudio.com>.
13
14 When you click the Knit button a document
15 will be generated that includes both content
16 as well as the output of any embedded R code
17 chunks within the document.
18
19 You can also embed plots, for example:
```

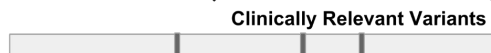
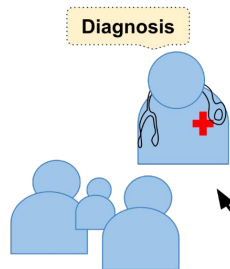
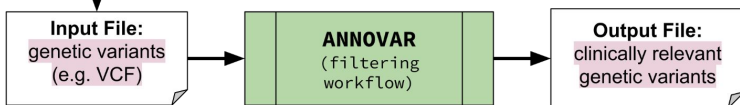
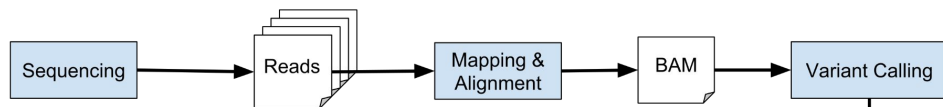
The rendered output on the right includes the title "Welcome to R Markdown", the date "December 18, 2014", and the introductory text. The R code output is displayed as follows:

```
## (Intercept)      speed
## -17.579095      3.932409
```

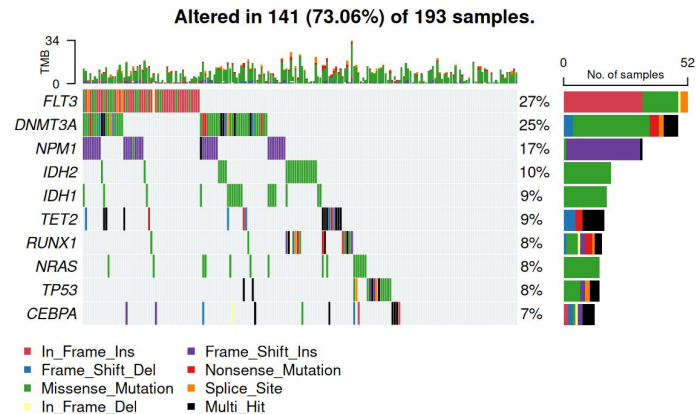
Below the code output, there is a scatter plot showing the relationship between distance (dist) on the x-axis and speed on the y-axis. The plot includes a blue linear regression line and a shaded gray area representing the confidence interval. The x-axis ranges from 0 to 125, and the y-axis ranges from 0 to 30.

- Los documentos RMarkdown son ficheros con extensión `.Rmd`.
- Si acabas utilizando R para hacer análisis estadísticos, los ficheros `.Rmd` te permitirán escribir muy fácilmente informes, tutoriales y transparencias para presentaciones.
- Estos ficheros `.Rmd` son (plenamente) reproducibles.

Maftools: Un paquete de R para resumir, analizar y visualizar archivos MAF



Clinical report



Referencias

- Libro de R de Hadley Wickham
- Tidyverse (<https://www.tidyverse.org>)
- Web: STHDA, datacamp, statology, CRAN
- maftools:

