ESCI upf. International Business

# P3 | Basic tools for data visualization

## In bioinformatics

**Marta Coronado Zamora**
4 October 2023

# Keep in touch

marta.coronado@prof.esci.upf.edu
@geneticament
@marta-coronado
Institut de Biologia Evolutiva (IBE-UPF-CSIC)

# Session content

- Short introduction to bioinformatics visualization
- Exercise: Making sense of the data: common visualizations in bioinformatics (`P3_exercises.Rmd`)
- Group project: finishing parts A and B

# Biological data

- **Quantitative and qualitative data**: scatterplots, barplots, boxplots, heatplots, ...
- **Molecular sequences**: alignments, motifs, genome browsers, ...
- **Species relationships**: trees, networks
- **Molecular pathways and interactions**: cell diagrams, networks, ...
- **Molecular structures**: 3D molecular viewers, ...
- **Anatomical structures**: anatograms, ...
- **...**

# Specialised libraries and software

- Integrated software suites
- Javascript
  - BioJS
- R libraries
  - Specialised repositories bioconductor
  - `ggplot2` extensions
  - `htmlwidgets`, some using BioJS libraries

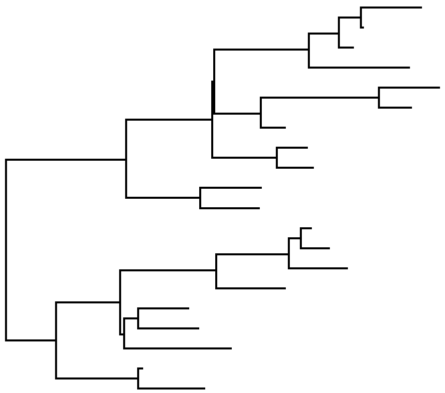
✏ **Exercise | Which `ggplot2` extensions and `htmlwidgets` are designed to cover specific needs of biological data?**

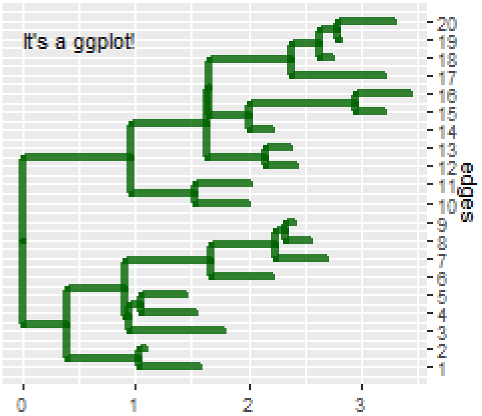
💬 Answer:

# Static visualizations: `ggplot2` extensions

## Phylogenetic trees: `ggtree`

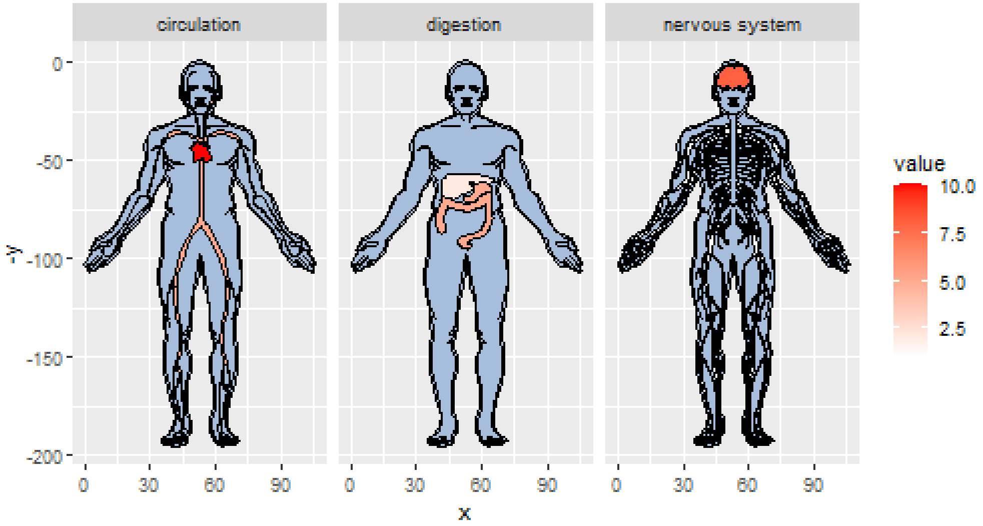
```r
library(ggtree)
set.seed(10); tr ← rtree(20)
ggtree(tr)
ggtree(tr, colour = "darkgreen", alpha = 0.8, size = 1.5)+
    scale_y_continuous(breaks = 1:20, position = "right", name = "edges") +
    annotate(geom = "text", x = 0.5, y = 19, label = "It's a ggplot!") +
    theme_gray()
```





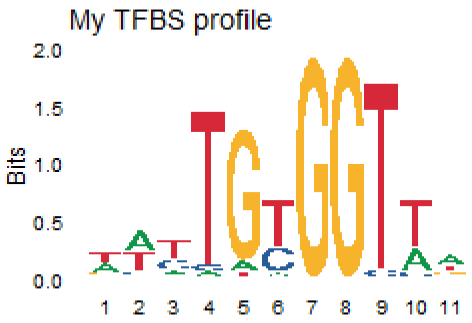
# Anatomical structures: `gganatogram`

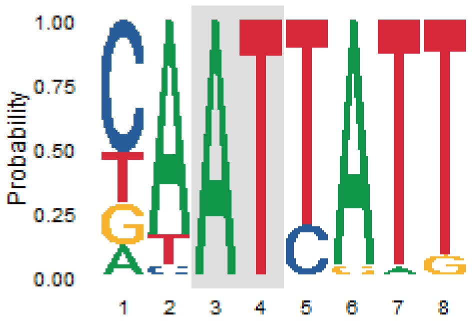
```
library(gganatogram)
gganatogram(data=organ_df, fillOutline='#a6bddb', organism='human',
            sex='male', fill="value") +
    scale_fill_gradient(low = "white", high = "red") +
    facet_wrap(~ type)
```



# Sequence logos: `ggseqlogo`

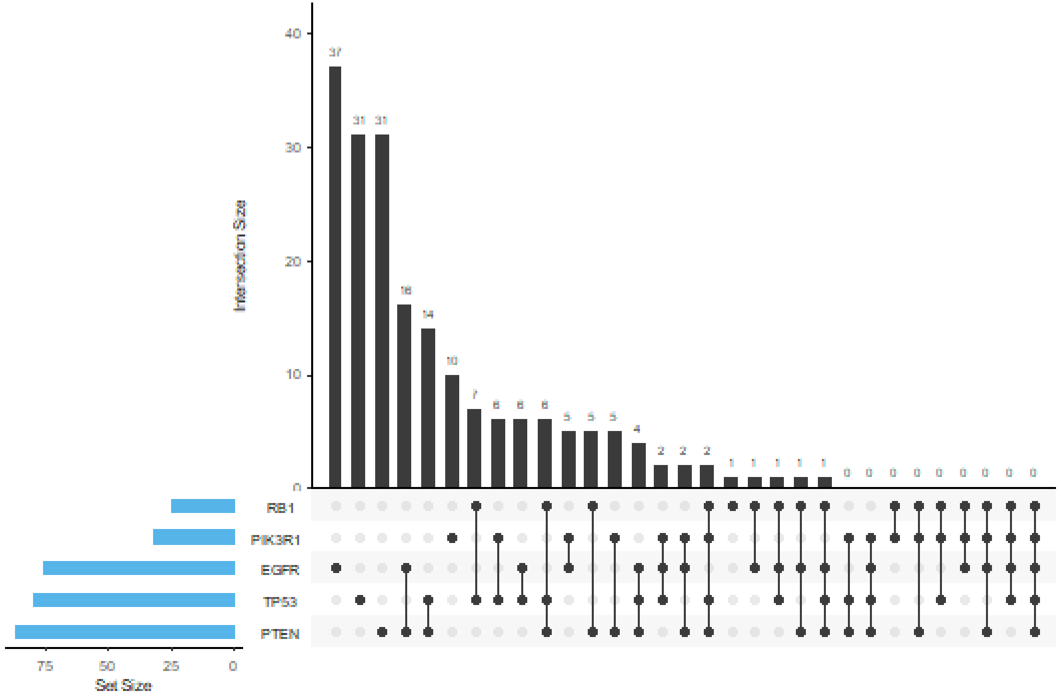
```r
library(ggseqlogo)
ggplot() + geom_logo(seqs_dna$MA0002.1) +
    theme_logo() + labs(title = "My TFBS profile")
ggplot() +
    annotate(geom = "rect", xmin = 2.5, xmax = 4.5,
             ymin = -Inf, ymax = Inf, alpha = 0.2) +
    geom_logo(seqs_dna$MA0008.1, method = "probability") +
    theme_logo()
```





# Show intersections: `UpSetR`

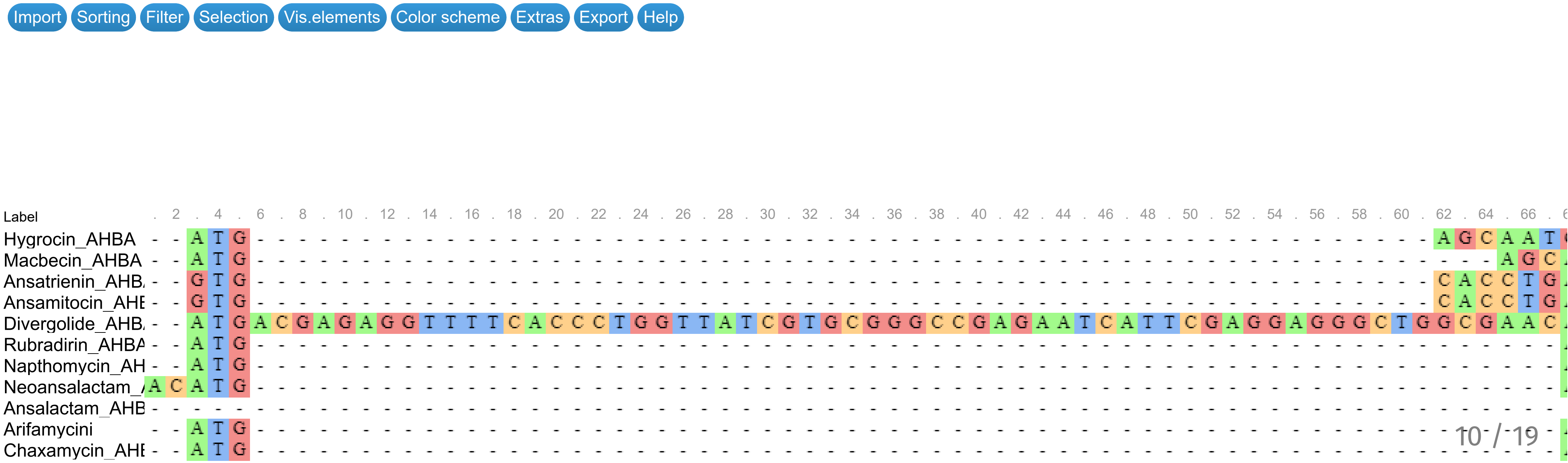
```
library(UpSetR)

mutations <- read.csv( system.file("extdata", "mutations.csv", package = "UpSetR"), header=T, sep = ",")

upset(mutations, sets = c("PTEN", "TP53", "EGFR", "PIK3R1", "RB1"), sets.bar.color = "#56B4E9",
order.by = "freq", empty.intersections = "on")
```



# Interactive visualizations

## Multiple alignment: `msaR`

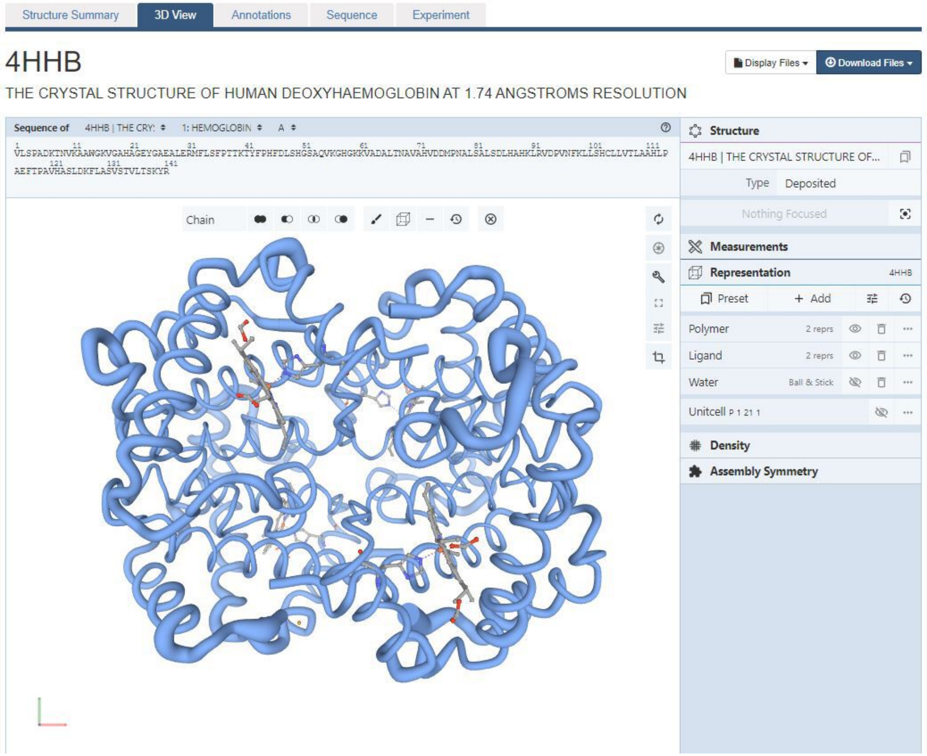
```r
library(msaR)
seqfile <- system.file("sequences","AHBA.aln", package="msaR")
msaR(seqfile)
```



```
Label           . 2 . 4 . 6 . 8 . 10 . 12 . 14 . 16 . 18 . 20 . 22 . 24 . 26 . 28 . 30 . 32 . 34 . 36 . 38 . 40 . 42 . 44 . 46 . 48 . 50 . 52 . 54 . 56 . 58 . 60 . 62 . 64 . 66 .
Hygrocin_AHBA   --ATG--------------------------------------------------------AGCAAT
Macbecin_AHBA   --ATG-----------------------------------------------------------AGC
Ansatrienin_AHB --GTG--------------------------------------------------------CACCTG
Ansamitocin_AHB --GTG--------------------------------------------------------CACCTG
Divergolide_AHB --ATGACGAGAGGTTTTCACCCTGGTTATCGTGCGGGCCGAGAATCATTCGAGGAGGGCTGGCGAAC
Rubradirin_AHBA --ATG--------------------------------------------------------------
Napthomycin_AH  --ATG--------------------------------------------------------------
Neoansalactam_  ACATG--------------------------------------------------------------
Ansalactam_AHB  -------------------------------------------------------------------
Arifamycini      -ATG--------------------------------------------------------------
Chaxamycin_AHB  --ATG--------------------------------------------------------------
```

# Protein structure

Example using NGL: a web application for molecular visualization: display molecules like proteins and DNA/RNA with a variety of representations.



4HHB

THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS RESOLUTION

Sequence of 4HHB | THE CRY: 1: HEMOGLOBIN A

```
1         11        21        31        41        51        61        71        81        91        101       111
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP
          121       131       141
AEFTPAVHASLDKFLASVSTVLTSKYR
```

Structure

4HHB | THE CRYSTAL STRUCTURE OF...

Type: Deposited

Nothing Focused

Measurements

Representation 4HHB

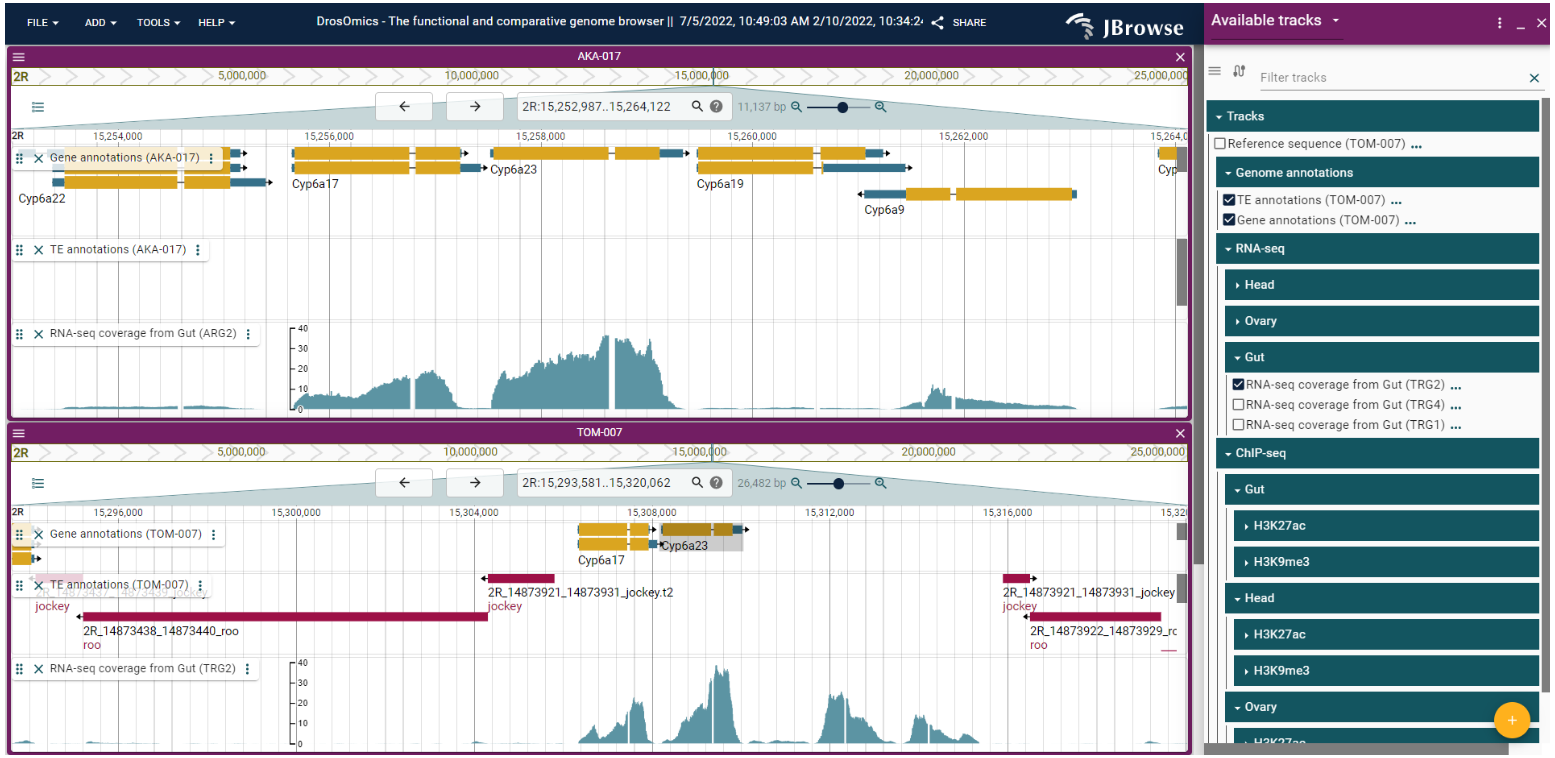
| | | |
|---|---|---|
| Polymer | 2 reprs | |
| Ligand | 2 reprs | |
| Water | Ball & Stick | |
| Unitcell P 1 21 1 | | |

Density

Assembly Symmetry

# Genome browsers

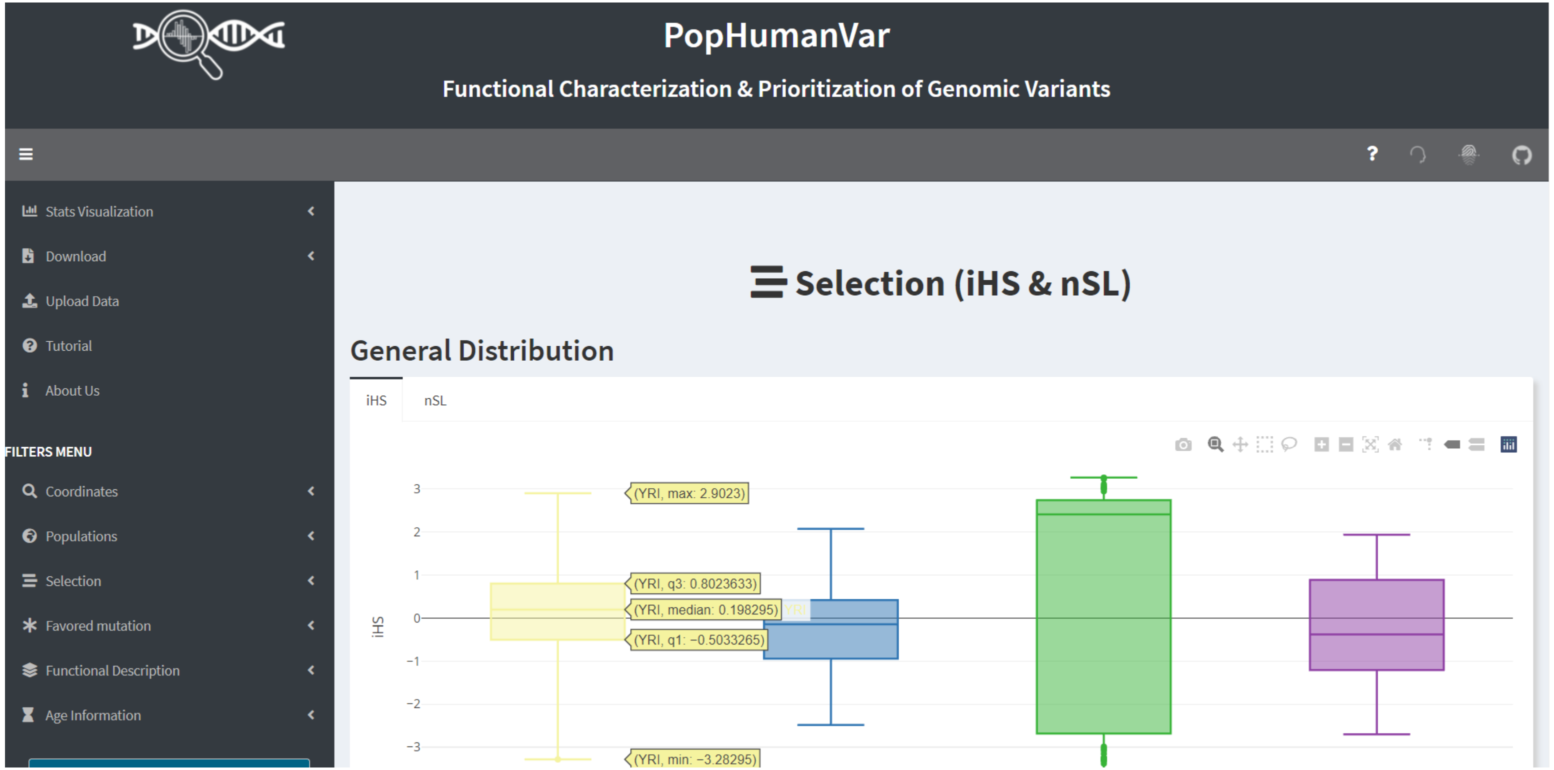
Examples by González lab at Institute of Evolutionary Biology: DrosOmics



Coronado-Zamora M, Salces-Ortiz J, González J. 2023. DrosOmics: A Browser to Explore -omics Variation Across High-Quality Reference Genomes From Natural Populations of *Drosophila melanogaster*. Mol Biol Evol 40:msad075.

# Shiny applications

Example by my group at UAB: PopHumanVar



Colomer-Vilaplana A, Murga-Moreno J, Canalda-Baltrons A, Inserte C, Soto D, Coronado-Zamora M, Barbadilla A, Casillas S. PopHumanVar: an interactive application for the functional characterization and prioritization of adaptive genomic variants in humans. Nucleic Acids Res. 2022;50(D1):D1069-D1076.

# Wrap-up

- Most basic exploratory and communication graphs in Bioinformatics can be achieved with general-purpose statistical graphics tools
- The complexity and characteristics of some biological data requires specialized tools
    - If static requirements, `ggplot2` extensions may help
    - If interactive requirements, `htmlwidgets` may help (next week!)
    - Check tools used in similar studies

# Practice

## Making sense of the data: common visualizations in bioinformatics

- Open the document `P3_exercises.Rmd` in RStudio and complete the exercises.
- Upload the completed document to Aul@-ESCI at the end of the session.

# Project

## Group project

### Parts

- **Part A** | Understand the origin of our data set and the meaning of the variables
- **Part B** | Visually describe our data set
- **Part C** | ❓

# Project

## Group project

### Part A

- Describe your data set:
  - Where and why was the information collected?
  - Which is the meaning of each variable?
  - Do the variables have unit? Which one?
  - Does the data set have a long format?

# Project

## Group project

### Part B

- Write the code to:
  - Read it into R
  - Reshape the data if necessary into long format
  - Check the variable classes and update them if necessary

# Project

## Group project

**Part B**

- Write the code to:
  - Read it into R
  - Reshape the data if necessary into long format
  - Check the variable classes and update them if necessary
- Explore your data using `ggplot2` graphics
  - Represent the **distribution of the variables**: pick one continuous variable and one discrete variable and use histograms or bar graphs to show their distribution
  - **Summarize the data**: use one geom to summarize data (e.g.: `geom_smooth`, boxplots, …) of two variables
- Explain your data with graphics and text
  - Choose the **three graphics** that better describe your data
  - **Customize** and **annotate** them
  - Accompany the figures with your **hypothesis** and/or **interpretation**

Add everything (**tidy**) to the initial `R Markdown` document and **submit the final project: 20 October 2023**.