T2.2 | Tools for data visualization

In bioinformatics

Marta Coronado Zamora
2 October 2024

Keep in touch

Theory lessons

Marta Coronado Zamora	Jose F. Sánchez
✉ marta.coronado@prof.esci.upf.edu	✉ jose.sanchez@prof.esci.upf.edu
🐦 @geneticament	🐦 @JFSanchezBioinf
📍 Institut Botànic de Barcelona (CSIC-CMCNB)	📍 Germans Trias i Pujol Research Institute (IGTP)

Practical lessons

Adrià Auladell
✉ adria.auladell@ibe.upf-csic.es
📍 Institut de Biologia Evolutiva (UPF-CSIC)

Session content

- Biological data
- General statistical graphics used in bioinformatics
- Specialized libraries and software
  - Static
  - Interactive
- Solve doubts

Get started!

Tools for data visualization in bioinformatics

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, …
- **…**

✎ **Exercise | Go to the latest research biology articles from Nature Communications journal and describe the types of visualizations used. How far can you enlarge the previous list?**

💬 Answer:

General statistical graphics

Biological example

Gene expression data from GTEx project:

```
data ← read.table(file = "https://raw.githubusercontent.com/marta-coronado/data/refs/heads/main/expression
                   header = TRUE, sep ="\t",
                   stringsAsFactors = FALSE)
data[1:4,1:4]
```

```
##              gene_id gene_name tissue median_expression
## 1 ENSG00000000003.10    TSPAN6  Brain            5.8635
## 2 ENSG00000000003.10    TSPAN6  Liver           30.7800
## 3 ENSG00000000003.10    TSPAN6 Muscle            2.5905
## 4 ENSG00000000003.10    TSPAN6 Testis           97.0900
```

Biological example

The `data` data frame contains the expression of 52302 genes in 8 tissues.

```
table(data$gene_type, data$tissue)
```

```
##                 
##                   Adipose Brain Liver  Lung Lymphocytes Muscle Stomach Testis
##   antisense          4999  4999  4999  4999        4999   4999    4999   4999
##   lincRNA            7033  7033  7033  7033        7033   7033    7033   7033
##   miRNA              2838  2838  2838  2838        2838   2838    2838   2838
##   misc_RNA           2010  2010  2010  2010        2010   2010    2010   2010
##   protein_coding    19820 19820 19820 19820       19820  19820   19820  19820
##   pseudogene        13705 13705 13705 13705       13705  13705   13705  13705
##   rRNA                520   520   520   520         520    520     520    520
##   snoRNA             1408  1408  1408  1408        1408   1408    1408   1408
##   snRNA              1892  1892  1892  1892        1892   1892    1892   1892
```

Biological example

✎ **Exercise | Use `ggplot2` to answer the following questions:**

- Which type of gene is more expressed on average?

💬 Answer:

- Which tissue has more expressed genes?

💬 Answer:

- Where is HLA-B gene expressed the most? Is its expression low, high or average compared to other genes?

💬 Answer:

- Where are located the genes highly expressed in brain (median_expression >20,000)?

💬 Answer:

Specialised libraries and software

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

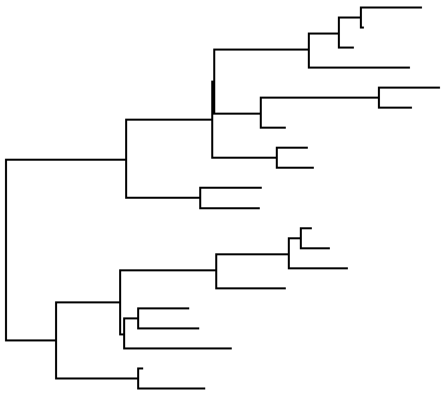
✎ **Exercise | Which `ggplot2` extensions and `htmlwidgets` are designed to cover specific needs of biological data?**

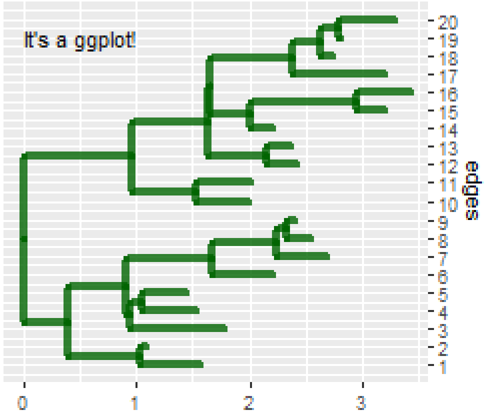
💬 Answer:

Static visualizations: `ggplot2` extensions

Phylogenetic trees: `ggtree`

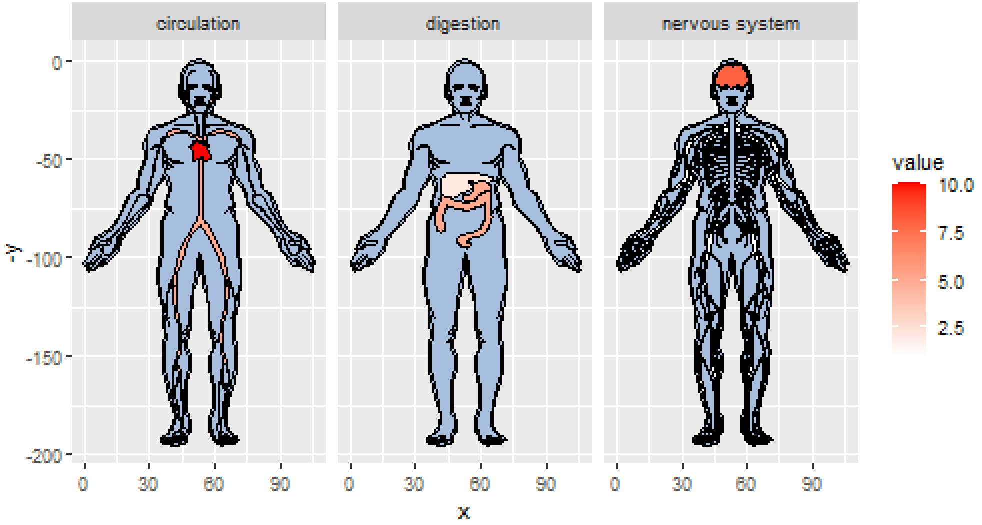
```r
library(ggtree)
set.seed(10); tr <- rtree(20)
ggtree(tr)
ggtree(tr, colour = "darkgreen", alpha = 0.8, size = 1.5)+
    scale_y_continuous(breaks = 1:20, position = "right", name = "edges") +
    annotate(geom = "text", x = 0.5, y = 19, label = "It's a ggplot!") +
    theme_gray()
```





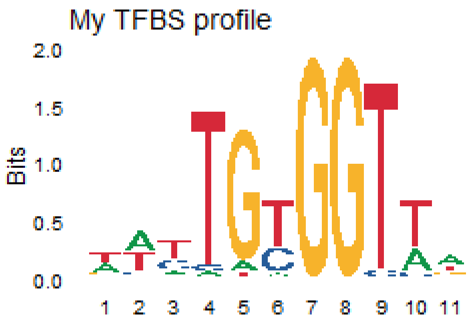
Anatomical structures: `gganatogram`

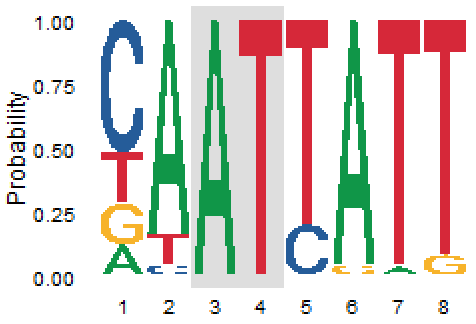
```
library(gganatogram)
gganatogram(data=organ_df, fillOutline='#a6bddb', organism='human',
            sex='male', fill="value") +
    scale_fill_gradient(low = "white", high = "red") +
    facet_wrap(~ type)
```



Sequence logos: `ggseqlogo`

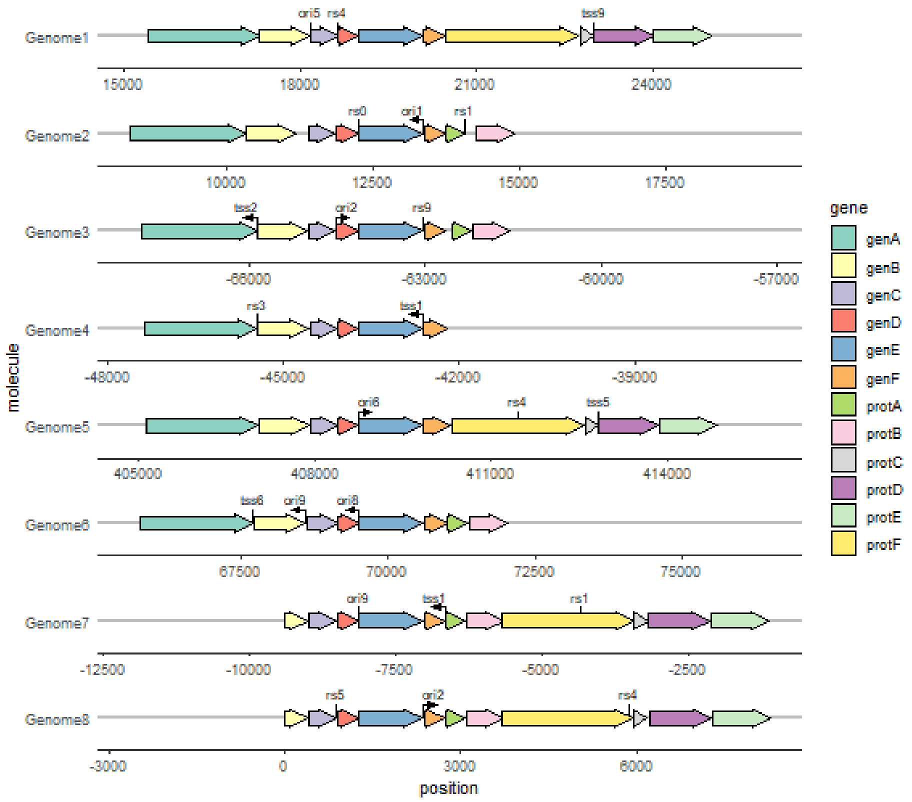
```
library(ggseqlogo)
ggplot() + geom_logo(seqs_dna$MA0002.1) +
    theme_logo() + labs(title = "My TFBS profile")
ggplot() +
    annotate(geom = "rect", xmin = 2.5, xmax = 4.5,
             ymin = -Inf, ymax = Inf, alpha = 0.2) +
    geom_logo(seqs_dna$MA0008.1, method = "probability") +
    theme_logo()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





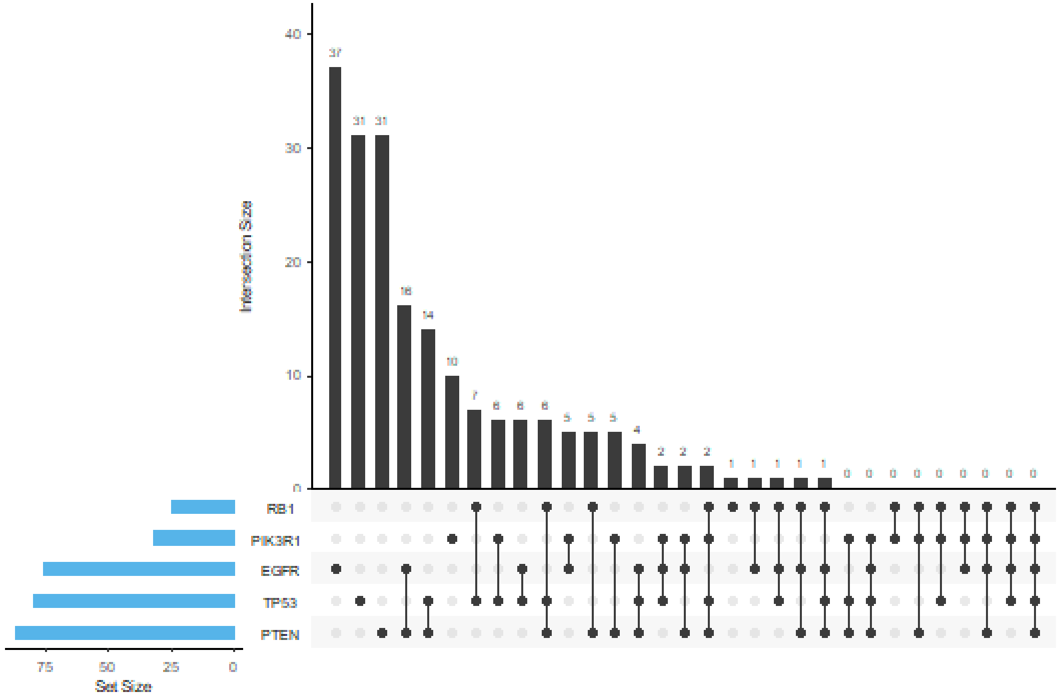
Gene structures: gggenes y gggenomes

```
library(gggenes)
ggplot(example_genes, aes(xmin = start, xmax = e
  geom_feature(
    data = example_features,
    aes(x = position, y = molecule, forward = fo
    ) +
  geom_feature_label(
    data = example_features,
    aes(x = position, y = molecule, label = name
    ) +
  geom_gene_arrow() +
  geom_blank(data = example_dummies) +
  facet_wrap(~ molecule, scales = "free", ncol =
  scale_fill_brewer(palette = "Set3") +
  theme_genes()
```



Show intersections: `UpSetR`

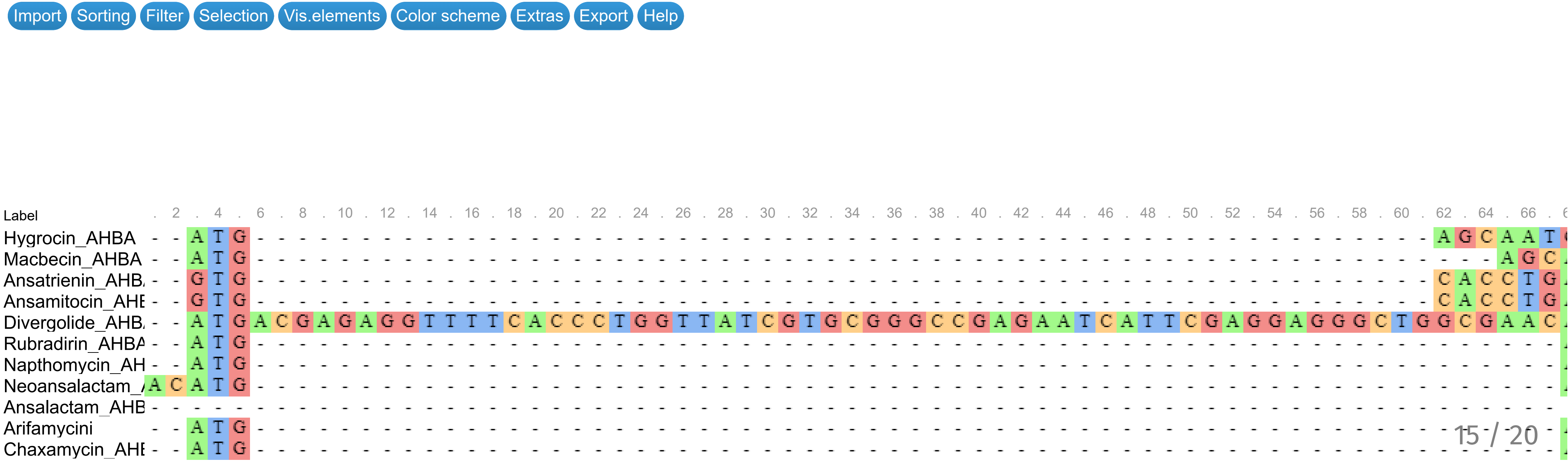
```r
library(UpSetR)

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`

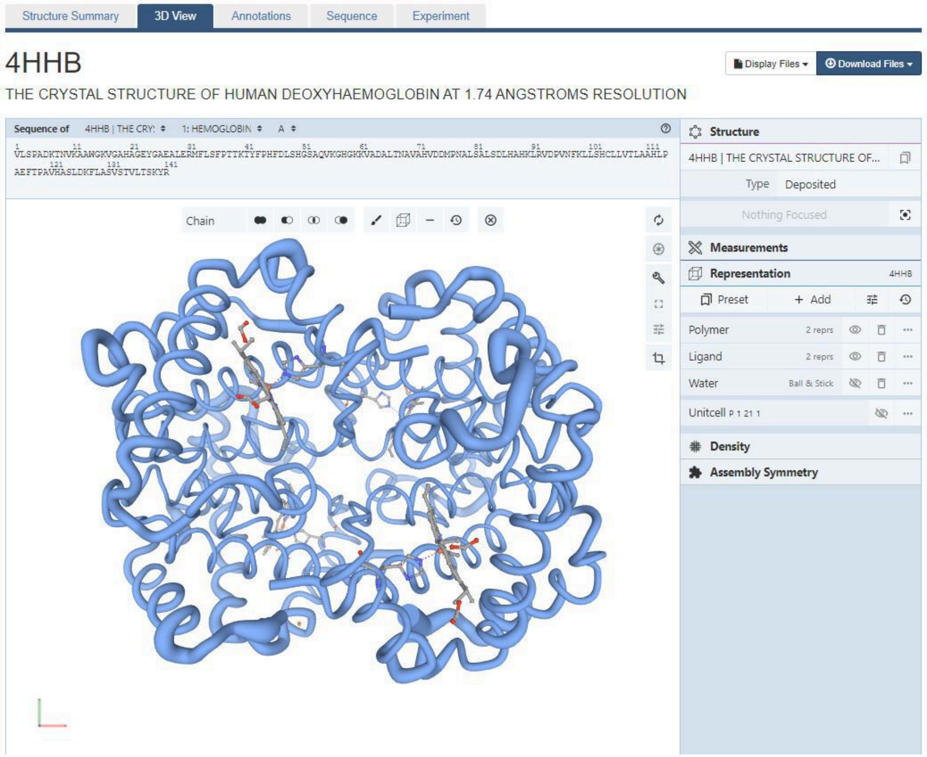
```r
library(msaR)
seqfile ← system.file("sequences","AHBA.aln", package="msaR")
msaR(seqfile)
```



```
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

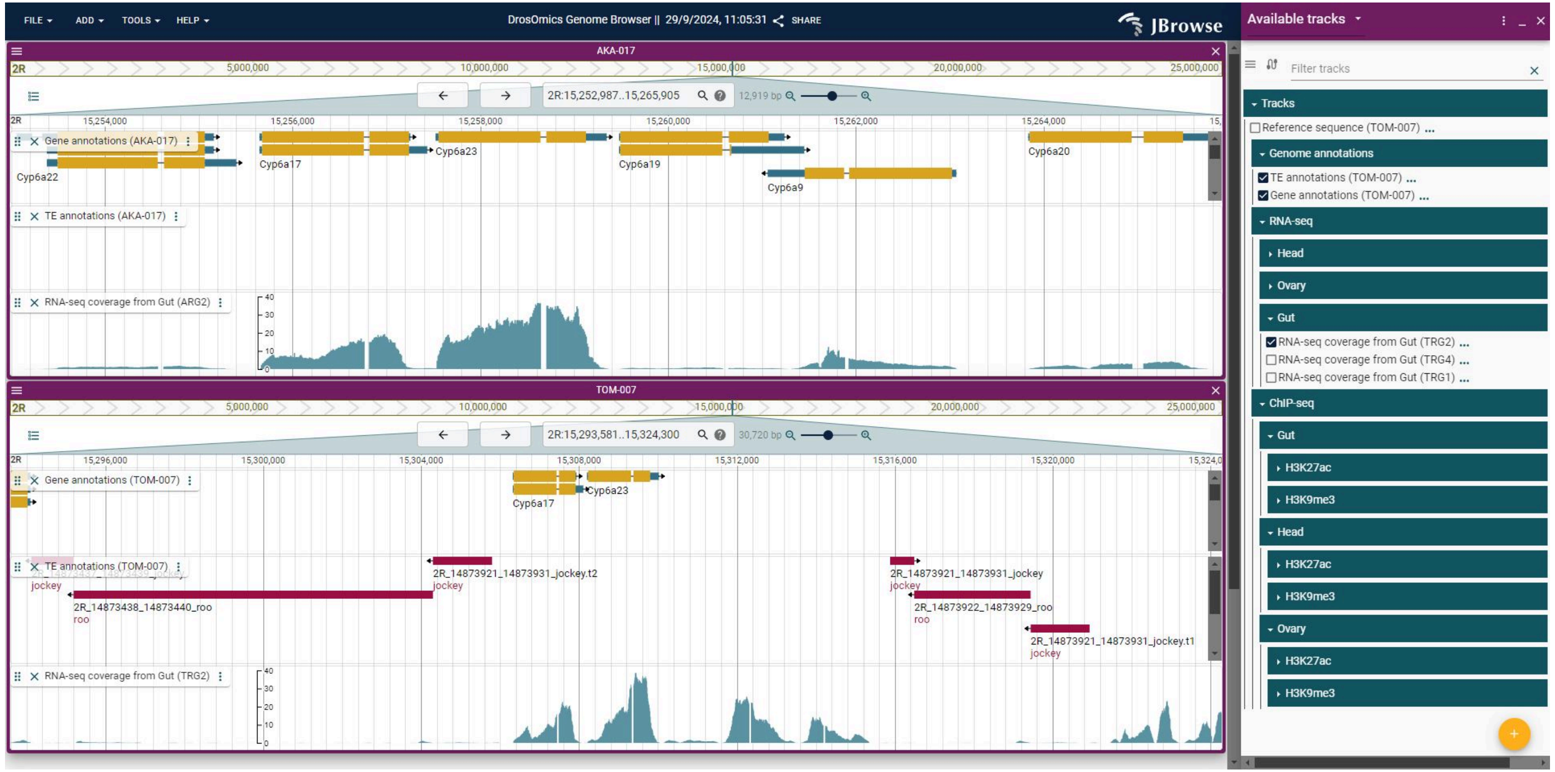
THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS RESOLUTION

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
```

Genome browsers

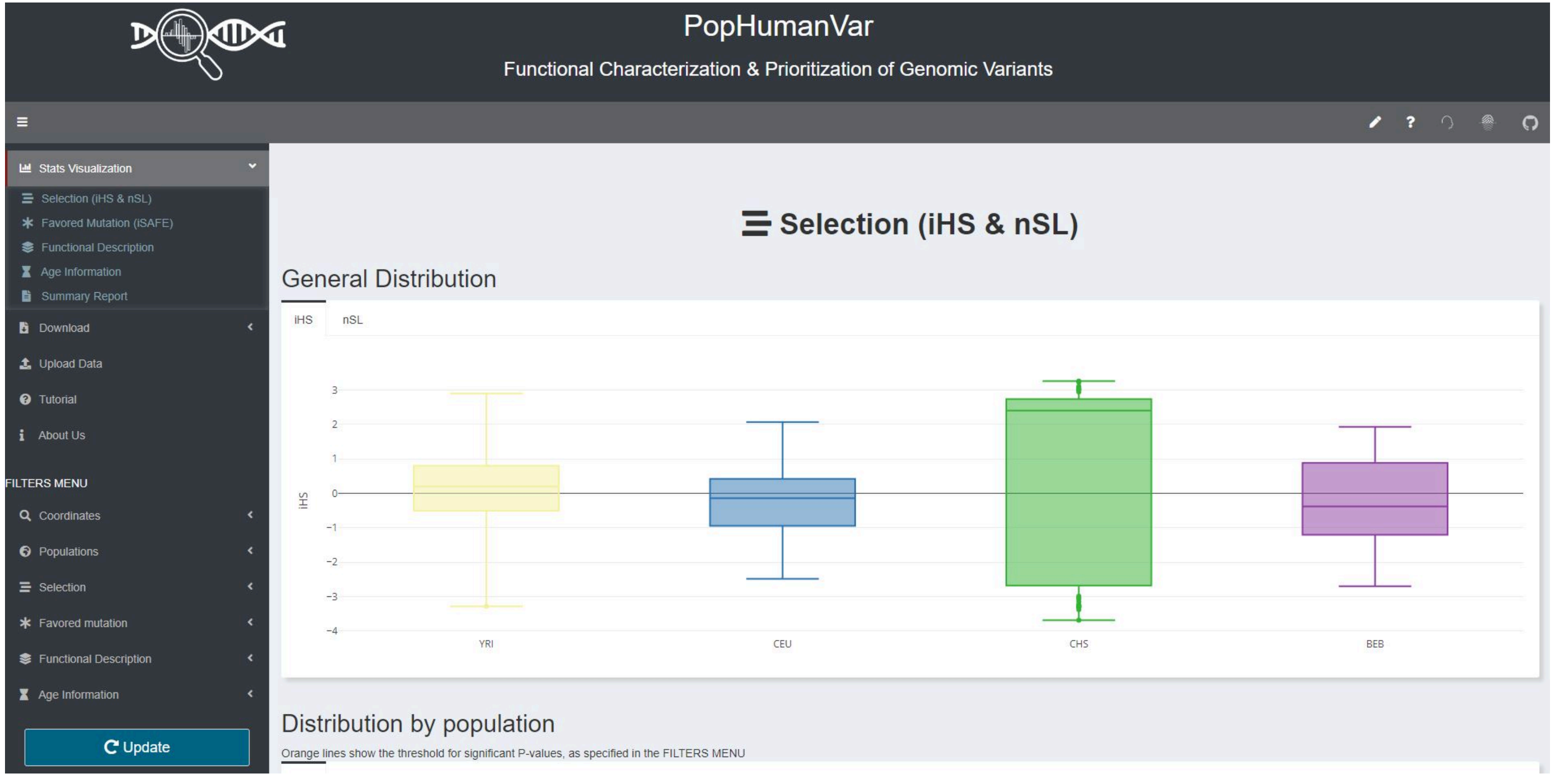
Examples by González lab at Institute of Evolutionary Biology: DrosOmics


FILE
ADD
TOOLS
HELP
DrosOmics Genome Browser || 29/9/2024, 11:05:31
SHARE
JBrowse
AKA-017
2R
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
2R:15,252,987..15,265,905
12,919 bp
Gene annotations (AKA-017)
Cyp6a22
Cyp6a17
Cyp6a23
Cyp6a19
Cyp6a9
Cyp6a20
TE annotations (AKA-017)
RNA-seq coverage from Gut (ARG2)
TOM-007
2R:15,293,581..15,324,300
30,720 bp
Gene annotations (TOM-007)
Cyp6a17
Cyp6a23
TE annotations (TOM-007)
jockey
2R_14873438_14873440_roo
roo
2R_14873921_14873931_jockey.t2
2R_14873921_14873931_jockey
2R_14873922_14873929_roo
2R_14873921_14873931_jockey.t1
RNA-seq coverage from Gut (TRG2)
Available tracks
Filter tracks
Tracks
☐ Reference sequence (TOM-007)
Genome annotations
☑ TE annotations (TOM-007)
☑ Gene annotations (TOM-007)
RNA-seq
Head
Ovary
Gut
☑ RNA-seq coverage from Gut (TRG2)
☐ RNA-seq coverage from Gut (TRG4)
☐ RNA-seq coverage from Gut (TRG1)
ChIP-seq
Gut
H3K27ac
H3K9me3
Head
H3K27ac
H3K9me3
Ovary
H3K27ac
H3K9me3

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
  - Check tools used in similar studies

Upload `T2.2_slides.Rmd` with the completed exercises (text included) to aul@-ESCI