

MATRiX app: Mining and functional Analysis of TRanscriptomics data

Une application dédiée à l'exploration de résultats d'analyses biostatistiques de données transcriptomiques

Problématique

Des données complexes à explorer et à analyser pour les biologistes

- ❶ Critères arbitraires pour la sélection de listes de gènes
- ❷ Compétences spécifiques pour la génération de graphiques/analyses complémentaires



→ Développer une solution pour que les biologistes puissent manipuler leurs jeux de données de manière autonome

Quelle solution d'hébergement?

Utilisation

Matrix

How to import ?

- First click on the browse button to load the data.
- After the pop up has appeared, you will have to select the data files.
- It is also possible to directly drag and drop your data files in the browse screen.
- As soon as your data files are loaded, these files are automatically named as: data_1, data_2, data_3, ...
- data : The experimental design in a 2 column table that associates samples to their respective biological conditions.
- Normalized : The table of the normalized expression values (log2, log10, ...) with genes in rows and samples in columns. The first column must contain the unique gene identifier (for transcripts, genes, ...).
- Results : The table containing the results of differential analysis (FDR change, p-value and FDR) next to a first column with unique gene identifier and a second column with the gene symbol.
- The first step consist to select all the data at once and then confirm the selection by clicking on the same button.
- A green message will then appear to confirm the data loading with a success table.

| Gene | data_1 | data_2 | data_3 | data_4 | data_5 | data_6 | data_7 | data_8 | data_9 | data_10 |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| AT1G01010 | 1.2 | 1.5 | 1.8 | 2.1 | 2.4 | 2.7 | 3.0 | 3.3 | 3.6 | 3.9 |
| AT1G01020 | 1.1 | 1.4 | 1.7 | 2.0 | 2.3 | 2.6 | 2.9 | 3.2 | 3.5 | 3.8 |
| AT1G01030 | 1.3 | 1.6 | 1.9 | 2.2 | 2.5 | 2.8 | 3.1 | 3.4 | 3.7 | 4.0 |
| AT1G01040 | 1.4 | 1.7 | 2.0 | 2.3 | 2.6 | 2.9 | 3.2 | 3.5 | 3.8 | 4.1 |
| AT1G01050 | 1.5 | 1.8 | 2.1 | 2.4 | 2.7 | 3.0 | 3.3 | 3.6 | 3.9 | 4.2 |
| AT1G01060 | 1.6 | 1.9 | 2.2 | 2.5 | 2.8 | 3.1 | 3.4 | 3.7 | 4.0 | 4.3 |
| AT1G01070 | 1.7 | 2.0 | 2.3 | 2.6 | 2.9 | 3.2 | 3.5 | 3.8 | 4.1 | 4.4 |
| AT1G01080 | 1.8 | 2.1 | 2.4 | 2.7 | 3.0 | 3.3 | 3.6 | 3.9 | 4.2 | 4.5 |
| AT1G01090 | 1.9 | 2.2 | 2.5 | 2.8 | 3.1 | 3.4 | 3.7 | 4.0 | 4.3 | 4.6 |
| AT1G01100 | 2.0 | 2.3 | 2.6 | 2.9 | 3.2 | 3.5 | 3.8 | 4.1 | 4.4 | 4.7 |

Démo

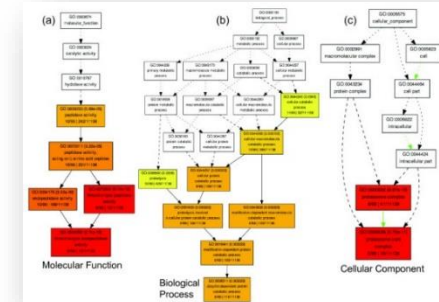
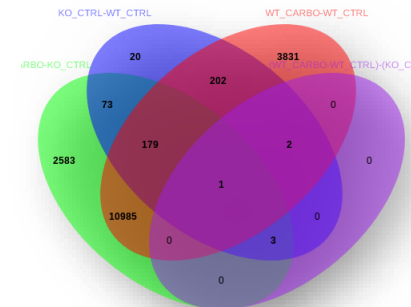
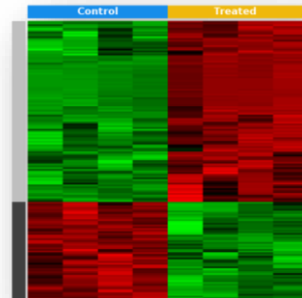
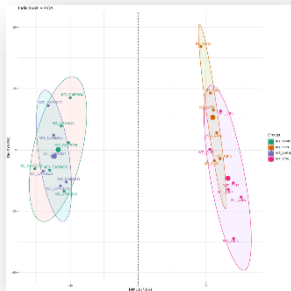
<https://matrix.toulouse.inrae.fr>



Problématique

Des données complexes à explorer et ré-analyser pour les biologistes

- ② Critères arbitraires pour la sélection de liste de gènes
- ② Compétences spécifiques pour la génération de graphiques/ analyses complémentaires



→ Développer une solution pour que les biologistes puissent manipuler leurs jeux de données de manière autonome

Objectifs



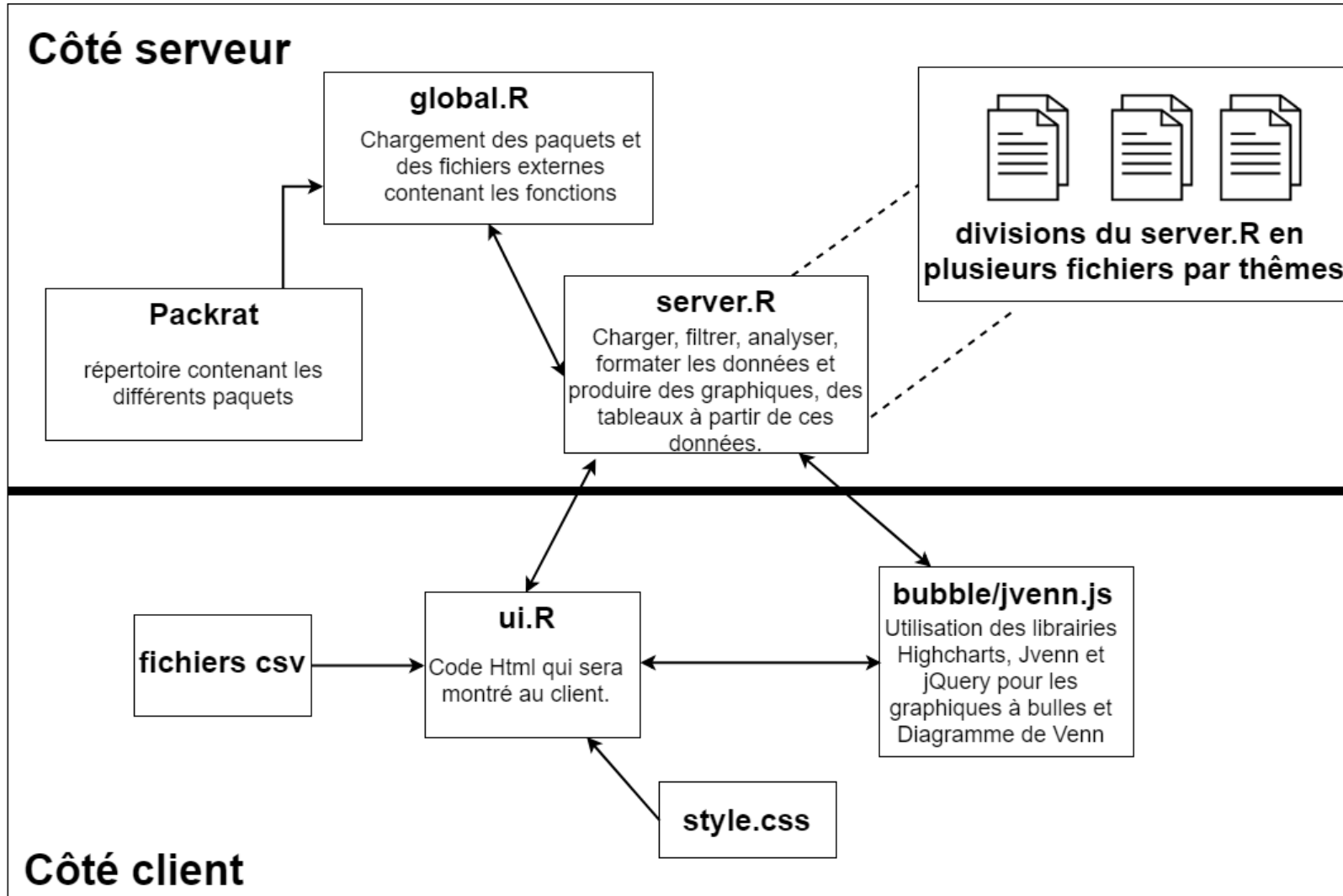
Stage M2 Bioinfo puis CDD Franck Soubès: développement et déploiement de l'application MATRiX

- ✓ **Import datasets** (microarray, RNAseq) with statistical result table
 - Upload from local
 - Load from server
- ✓ **Explore and Analyze data**
 - ✓ PCA
 - ✓ Venn
 - ✓ Heatmap
 - ✓ Functional analysis
- ✓ **Export publishable quality graphics**
 - png
 - eps, svg, pdf
- ✓ **Built into a single web application**
 - **Interactive** interface
 - User friendly
 - **Responsive**
 - **Open source** and open access code (GPL-3)



RShiny : Principes

Architecture



Rshiny: Exemple

UI.R

```
ui <- fluidPage(
  titlePanel("Tabsets"),
  sidebarLayout(
    sidebarPanel(
      radioButtons("dist", "Distribution type:",
        c("Normal" = "norm",
          "Uniform" = "unif",
          "Log-normal" = "lnorm",
          "Exponential" = "exp")),
      sliderInput("n",
        "Number of observations:",
        value = 500,
        min = 1,
        max = 1000)
    ),
    mainPanel(
      tabsetPanel(type = "tabs",
        tabPanel("Plot", plotOutput("plot")),
        tabPanel("Summary", verbatimTextOutput("summary")),
        tabPanel("Table", tableOutput("table"))
      )
    )
  )))
```

Server.R

```
server <- function(input, output) {
  d <- reactive({
    dist <- switch(input$dist,
      norm = rnorm,
      unif = runif,
      lnorm = rlnorm,
      exp = rexp,
      rnorm)

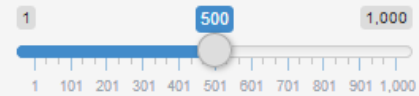
    dist(input$n)
  })
  output$plot <- renderPlot({
    dist <- input$dist
    n <- input$n
    hist(d(),
      main = paste("r", dist, "(", n, ")", sep = ""),
      col = "#75AADB", border = "white")
  })
  output$summary <- renderPrint({
    summary(d())
  })
  output$table <- renderTable({
    d()
  })
}
```

Tabsets

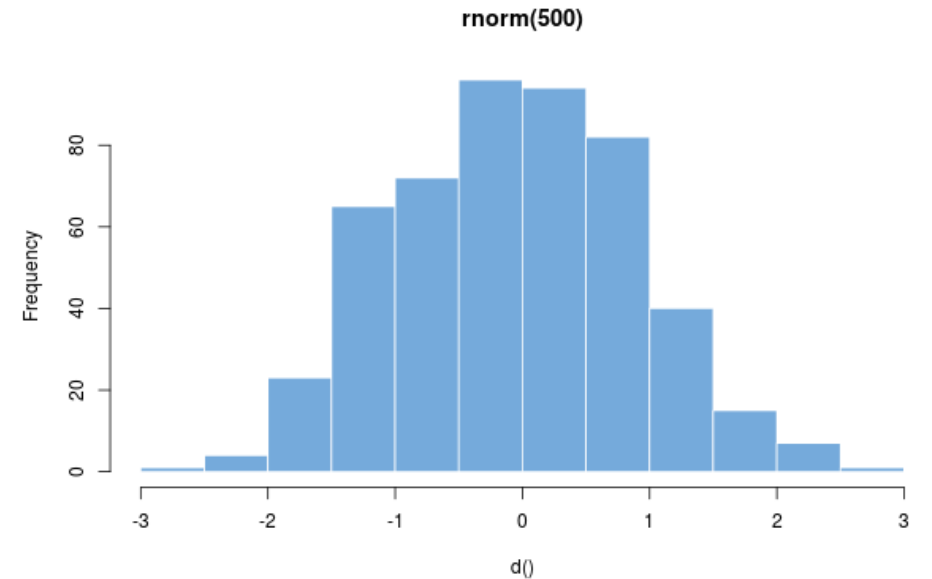
Distribution type:

- Normal
- Uniform
- Log-normal
- Exponential

Number of observations:



Plot Summary Table





MATRIX
≡

Home

Upload Data

How to import ?

- First click on the browse button to load the data
- After the pop up has appeared, you will have to select the data files.
- It is also possible to directly drag and drop your data files in the browse button
- You need three distinct csv files, these files are respectively named xxx_pData, xxx_WorkingSet and xxx_ResTable.
- pData : The experimental design in a 2 column table that associates samples to their respective biological conditions
- WorkingSet : The table of the normalised expression values (log2, cpm, ...) with genes in rows and samples in columns. The first column must contain the unique gene identifier (or transcript, probe, ...).
- ResTable : The table containing the results of differential analysis (fold change, p-value and FDR) next to a first column with unique gene identifier and a second column with the gene symbol.
- The final step consist to select all the data at once and then confirm the selection by clicking on the open button.
- A green message will then appear to confirm the data loading with a summary table.

Table 1: xxx_pData.csv

| X | Grp |
|----|---------|
| S1 | WT_CTRL |
| S3 | WT_COND |
| S4 | KO_CTRL |
| S5 | KO_COND |
| S2 | KO_COND |
| S8 | KO_CTRL |
| S6 | WT_COND |
| S7 | WT_CTRL |
| S9 | WT_COND |

Table 2: xxx_ResTable.csv

| Unique IDs | GeneName | logFC | P.value | adj.P.Val |
|---------------|----------|---------|----------|-----------|
| A_52_P1690 | Dbil5 | 2,611 | 0,0048 | 0,785 |
| A_51_P41424 | C85492 | 1,772 | 0,0048 | 0,773 |
| A_52_P108321 | Ccdc71 | -1,965 | 0,0078 | 0,819 |
| A_55_P1985764 | AV310571 | 0,208 | 0,00058 | 0,714 |
| A_51_P328014 | Cops3 | -0,859 | 0,0048 | 0,604 |
| A_52_P123354 | Tusc2 | -10,254 | 0,000048 | 0,00023 |

Table 3: xxx_WorkingSet.csv

| Unique IDs | S1 | S3 | S4 | S5 | S2 | S8 | S6 |
|---------------|---------|-------|--------|--------|---------|--------|--------|
| A_52_P1690 | 6,512 | 7,511 | 7,007 | 6,276 | 6,760 | 6,276 | 7,064 |
| A_51_P41424 | 9,975 | 5,525 | 10,479 | 9,745 | 10,040 | 9,745 | 9,662 |
| A_52_P108321 | 7,035 | 8,2 | 7,150 | 8,293 | 7,698 | 8,293 | 7,384 |
| A_55_P1985764 | 12,252 | 9,352 | 12,619 | 12,661 | 12,302 | 12,661 | 12,282 |
| A_51_P328014 | 10,8000 | 6,251 | 10,717 | 10,901 | 10,7171 | 10,901 | 10,705 |
| A_52_P123354 | 6,832 | 7,282 | 6,944 | 7,117 | 6,944 | 7,117 | 6,860 |

Upload data

download sample data

User data (.csv format)

Browse... No file selected

Import local example

File select

Decimal

Comma

Point

Unique identifier

ProbeName

Utilisation

Chargement des données (données stockées sur le serveur)



Upload data

[Download example data files](#)

Files configurations:

Feature Identifier **Decimal**

ProbeName
 Comma
 Point

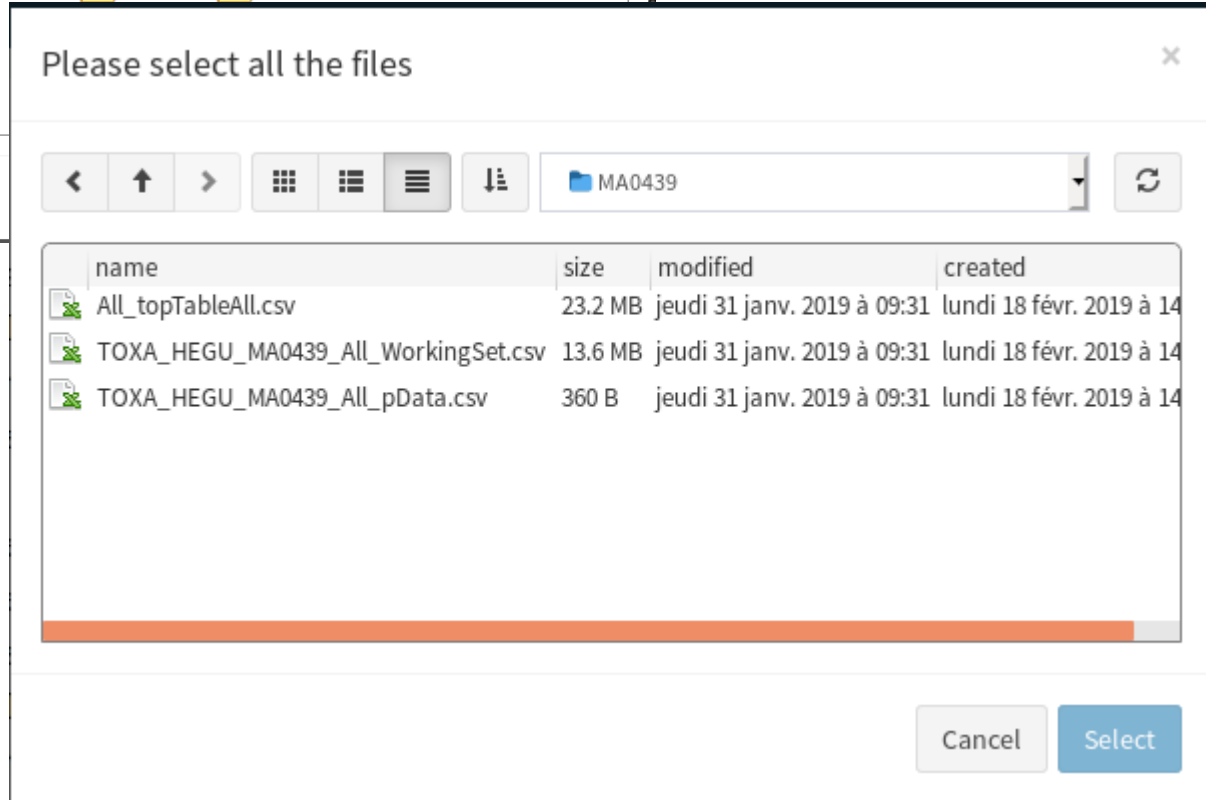
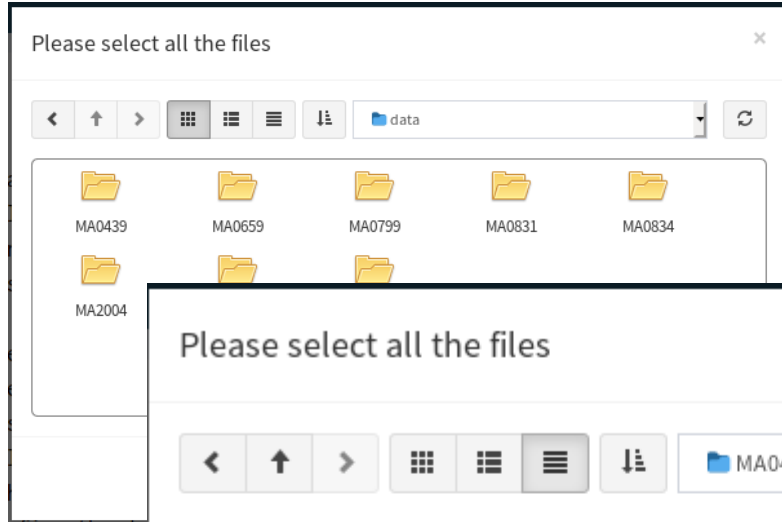
Upload project .csv files from:

1. your local machine

Browse... No file selected

2. your team's project on the server location

File select



Explore Results



MATRIX
☰

Home | Upload Data | Data summary | PCA | Venn diagram | Heatmap

Summary tables
Volcano plot
Stripchart genes
Data summary

Save Volcano plot
png ▾

LXRKO_0 versus WT_0

● Log (base 2) fold-change ● P value & Log (base 2) fold-change

Data summary

Statistical method: adj.p.val (FDR) ▾

VOLCANO plot

Choose a comparison: LXRKO_0-WT_0 ▾ | Choose regulation: both ▾

Fold-Change cutoff:
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6
 | P-value cutoff:
0.01 0.02 0.03 0.04 0.05
0.01 0.02 0.03 0.04 0.05

Advanced Graphical Options

Label size:
1 2 3 4 5 6
1 2 3 4 5 6
 | Point size:
0.5 1 1.7 2.3 2.9
0.5 1 1.7 2.3 2.9
 | Pt transparency:
0.1 0.3 0.5 0.7 0.9
0.1 0.3 0.5 0.7 0.9

Highlight gene(s): fill comma-separated GeneNames: FOXP2,OT,AVPR1a

Highlight a gene family: Cyp | Highlight top X genes: 20 ↕

Explore results



MATRIX

Summary tables
Volcano plot
Stripchart genes
Data summary

Filter the table +

This table shows the normalized values

CSV PDF

| ProbeName | GeneName | KO_CARBO.KO_CARBO1 | KO_CARBO.KO_CARBO2 | KO_CARBO.KO_CARBO3 | KO_CARBO.KO_CARBO4 |
|---------------|----------|--------------------|--------------------|--------------------|--------------------|
| A_55_P2044653 | Cyp2b10 | 15.77 | 14.89 | 15.78 | 15.52 |

Save Stripchart
png

MATRIX

Visualize the Venn diagram
Venn GO enrichment

Download the data
Download venn set
Filter the datatable +

Table showing the probes and genes associated with their respective logFC for the intersection(s) selected

You can directly filter the table by fold change and save the output table

CSV PDF

| ProbeName | GeneName | logFC_WT_CARBO-WT_CTRL |
|----------------|----------------------------|------------------------|
| A_51_P249286 | Rgs16 | 3.65 |
| A_55_P2178578 | Tmprss4 | 2.82 |
| A_55_P2052290 | Psat1 | 2.34 |
| A_55_P2011111 | Cyp4a10 | 2.30 |
| A_66_P112399 | BC029722 | 2.18 |
| A_55_P2032167 | Hfi202b | 2.05 |
| A_55_P1986247 | Fut1 | 1.92 |
| A_55_P2050628 | Cyp4a31 | 1.91 |
| A_52_P608322 | Maff | 1.88 |
| A_51_P254895 | Cyp4a10 | 1.85 |
| A_30_P01019537 | chr1:175653557-175671940_R | 1.83 |

Size of each list

Number of elements: specific (1) or shared by 2, 3, ... lists

Plot top DE genes Save Barplot png

Venn settings

Choose your comparison

WT_CARBO-WT_CTRL

KO_CARBO-KO_CTRL

KO_CARBO-WT_CARBO

KO_CTRL-WT_CTRL

(WT_CARBO-WT_CTRL)-(KO_CARBO-KO_CTRL)

Select all Clear selection

Statistical method: adj.p.val (FDR) Choose your regulation: both

A comma-separated list of **fill** or **hex** colors: green,blue,red,purple,orange,brown

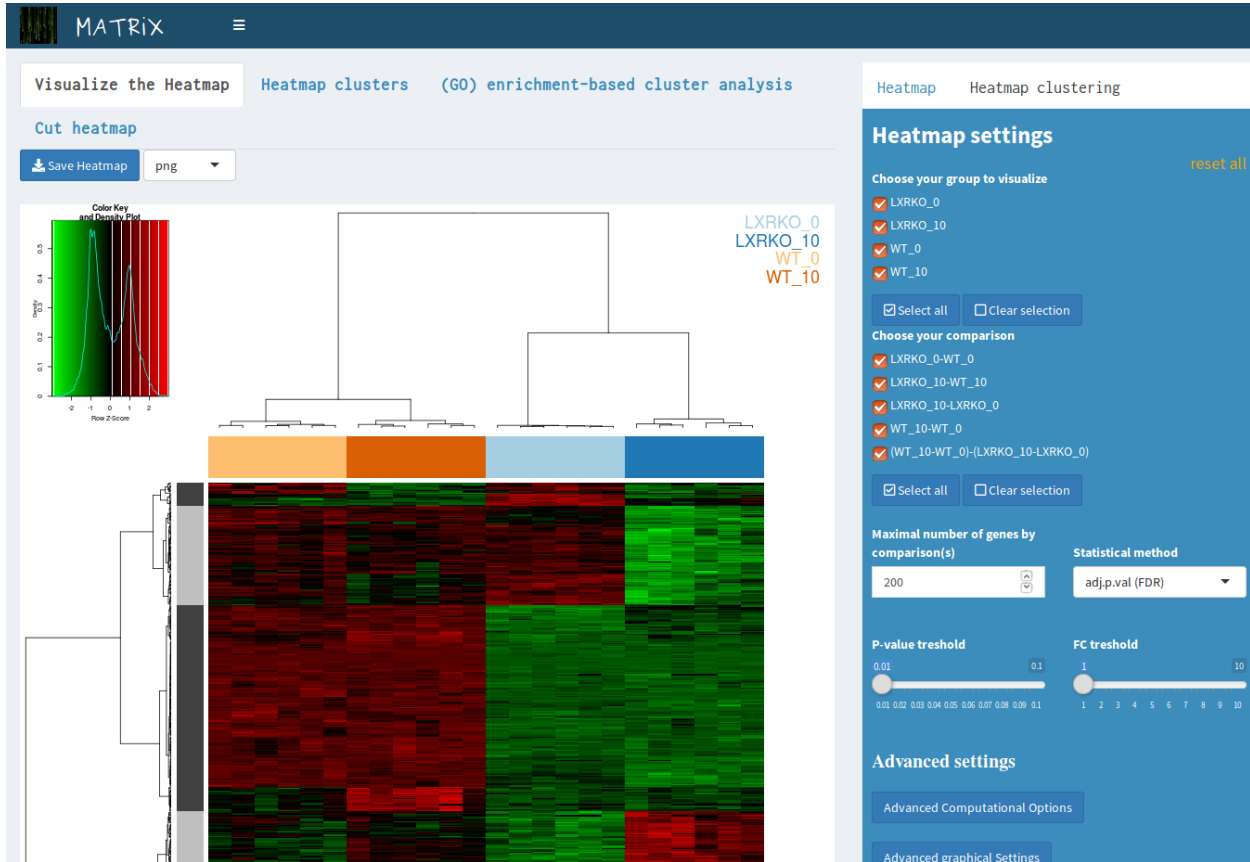
P-value threshold: 0.01 FC threshold: 1

Choose if you want to display probes or genes: probes

Remove the genes that are not annotated Show the logFC for all comparisons



- Heatmap clustering



- Functional Analysis

Functional Analysis

Send genes from a cluster to a web service

Choose a Cluster: 1 Submit to:

Login | Register

Transcription
Pathways
Ontologies
Diseases/Drugs
Cell Types
Misc
Legacy
Crowd

Description HeatmapCluster_1 (49 genes)

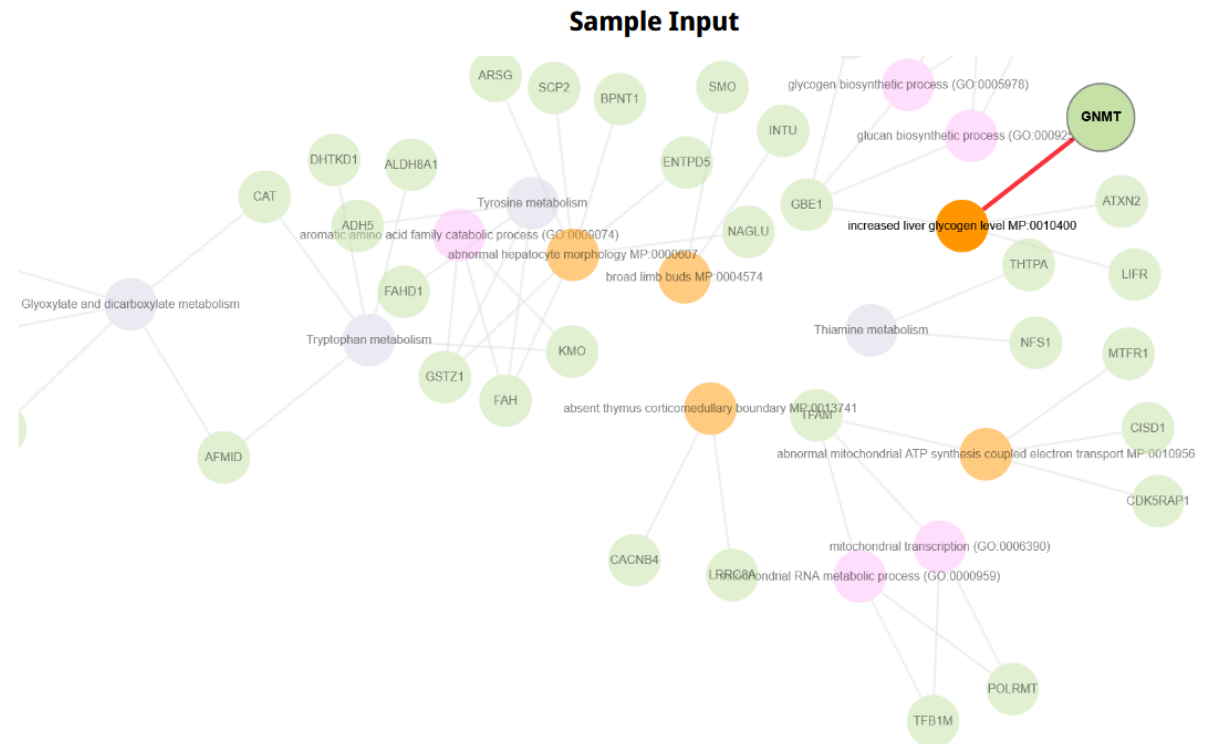
| | | |
|---|--|--|
| <p>GO Biological Process 2023</p> <p>Fatty Acid Elongation (GO:0030497)</p> <p>Very Long-Chain Fatty Acid Biosynthetic Proc</p> <p>Long-Chain fatty-acyl-CoA Biosynthetic Proc</p> <p>Long-Chain fatty-acyl-CoA Metabolic Process</p> <p>Sphingolipid Metabolic Process (GO:0006661)</p> | <p>GO Cellular Component 2023</p> <p>Insulin-Responsive Compartment (GO:0032125)</p> <p>Vesicle Membrane (GO:0012506)</p> <p>Membrane Attack Complex (GO:0005579)</p> <p>Vesicle (GO:0031982)</p> <p>CatSper Complex (GO:0036128)</p> | <p>GO Molecular Function 2023</p> <p>Cytidylyltransferase Activity (GO:0070567)</p> <p>Malate Dehydrogenase Activity (GO:0016611)</p> <p>Telethonin Binding (GO:0031433)</p> <p>Insulin-Like Growth Factor II Binding (GO:0005508)</p> <p>Fatty Acid Elongase Activity (GO:0009922)</p> |
| <p>MGI Mammalian Phenotype Level 4 2021</p> <p>abnormal circulating glucose level MP:00001</p> <p>macrovesicular hepatic steatosis MP:003111</p> <p>pancreatic islet hyperplasia MP:0005491</p> <p>increased insulin secretion MP:0003058</p> <p>decreased cardiac output MP:0003393</p> | <p>Human Phenotype Ontology</p> <p>Abnormality of salivation (HP:0100755)</p> <p>Abnormality of liposaccharide metabolism (HP:0000000)</p> <p>Abnormality of glycolipid metabolism (HP:0000000)</p> <p>Abnormality of glycosphingolipid metabolism (HP:0000000)</p> <p>Recurrent gram-negative bacterial infection (HP:0000000)</p> | <p>Jensen TISSUES</p> <p>Venom duct</p> <p>Primary root</p> <p>Abdominal adipose tissue</p> <p>Cladote</p> <p>Uroepithelial cell line</p> |

<https://maayanlab.cloud/enrich>



Evolutions prévues

- ⌘ **Correction « bugs »**
 - Table sorting
 - Scroll lag
- ⌘ **Améliorations**
 - Send venn list to enricher...
 - Implémenter API string.DB
 - enrichr-KG

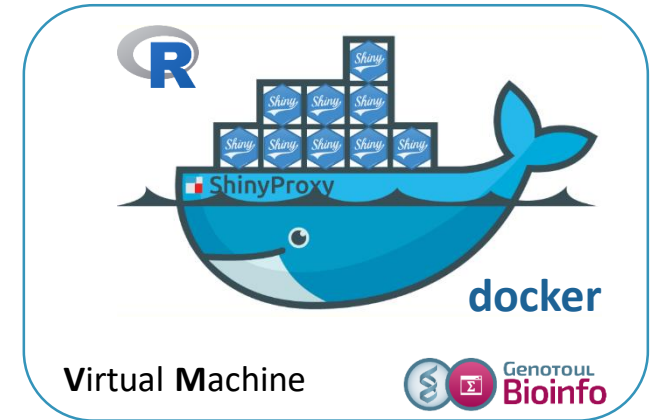
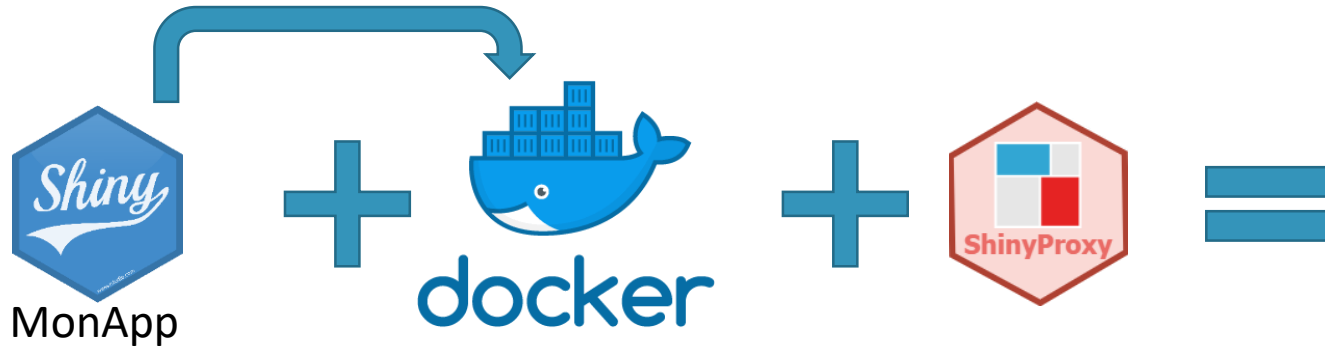


Quelle solution d'hébergement?



Hebergement 1: shinyproxy/ VM genotoul

<https://www.shinyproxy.io/>



https://forgemia.inra.fr/ylippi/MA_Trix_App

Dockerfile
 Ubuntu 20.04 with r-base
 Dependencies packages
 COPY MonApp /root/monapp
 EXPOSE 3838
 Shiny::runApp("/root/monapp")

`$ sudo docker build -t MyDockApp .`

application.yml
 specs:
 - id: monapp
 display-name: Mon application
 container-cmd: ["R", "-e",
 "shiny::runApp('/root/monapp')"]
 container-image: **MyDockApp**

`$ java -jar shinyproxy-3.0.1.jar`



<https://matrix.toulouse.inrae.fr:8080>

Pro:

- configurable

Cons:

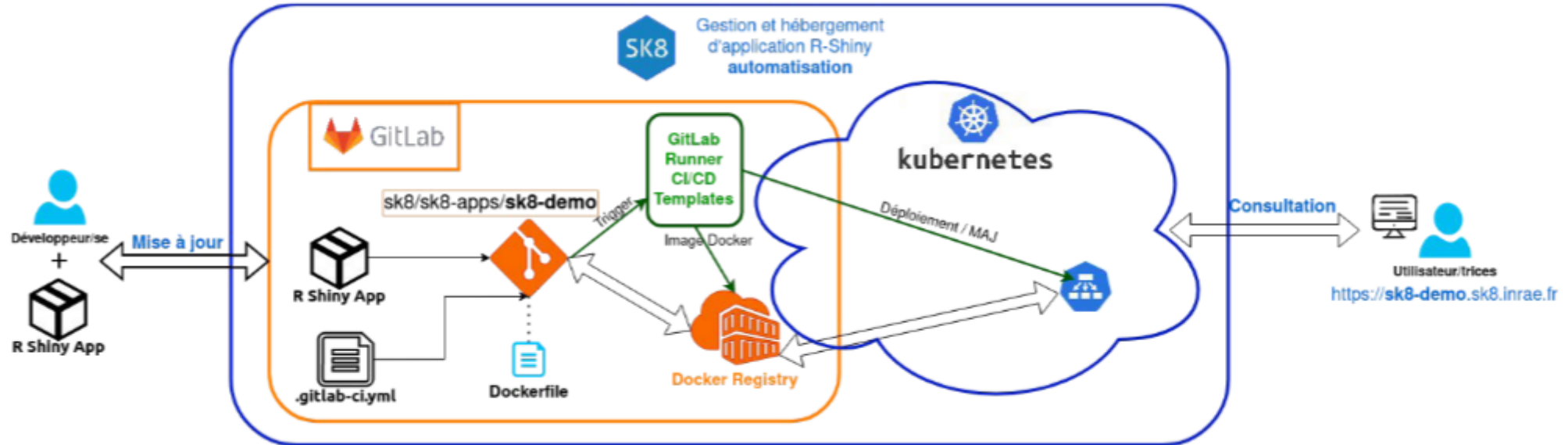
- Configure and maintain host server (VM)
- Pay VM location service

Hebergement 2: kubernetes/ SK8 INRAE

<https://sk8.inrae.fr>

<https://sk8.inrae.fr/>

Hebergement et déploiement à INRAE



Version 1 instance/ multi-users

<https://matrixapp.sk8.inrae.fr/>

Version multi-instances / multi-users

<https://shiny.sk8.inrae.fr/app/get-trix-matrixapp> (shinyProxy)

Pro:

- Déploiement automatisé via gitlab CI/CD
- Cluster kubernetes
- Performance

Cons:

- Réserve développeurs INRAE
- Financement infra à identifier pour solution d'hébergement pérenne

Démo

<https://matrix.toulouse.inrae.fr>

