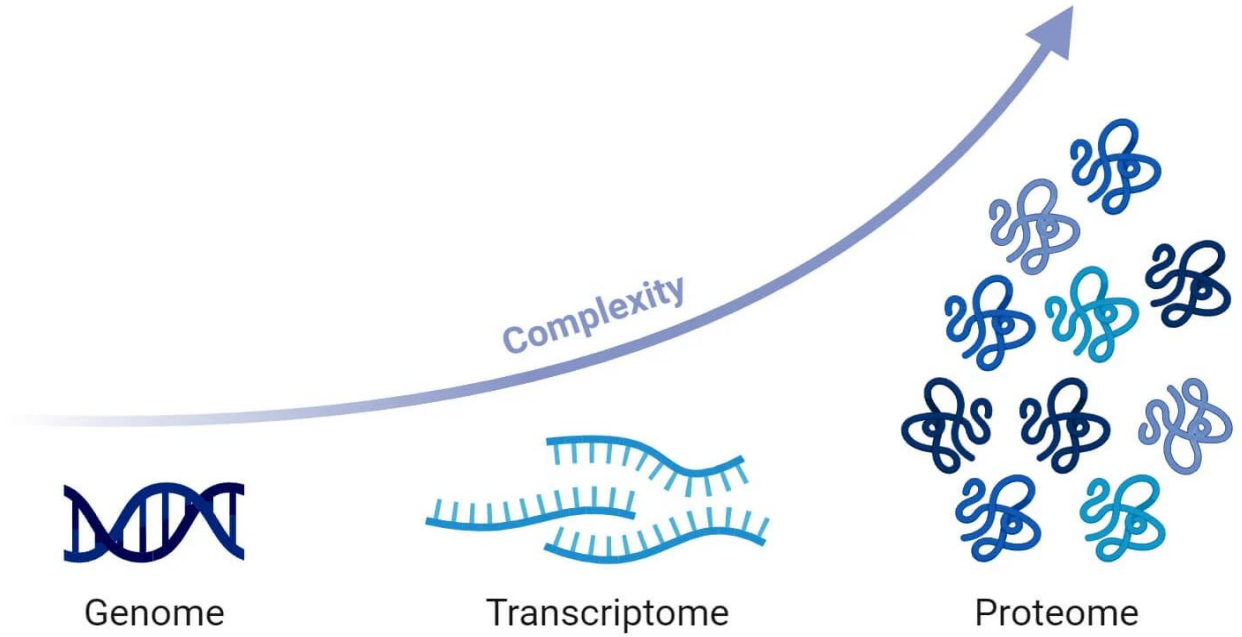
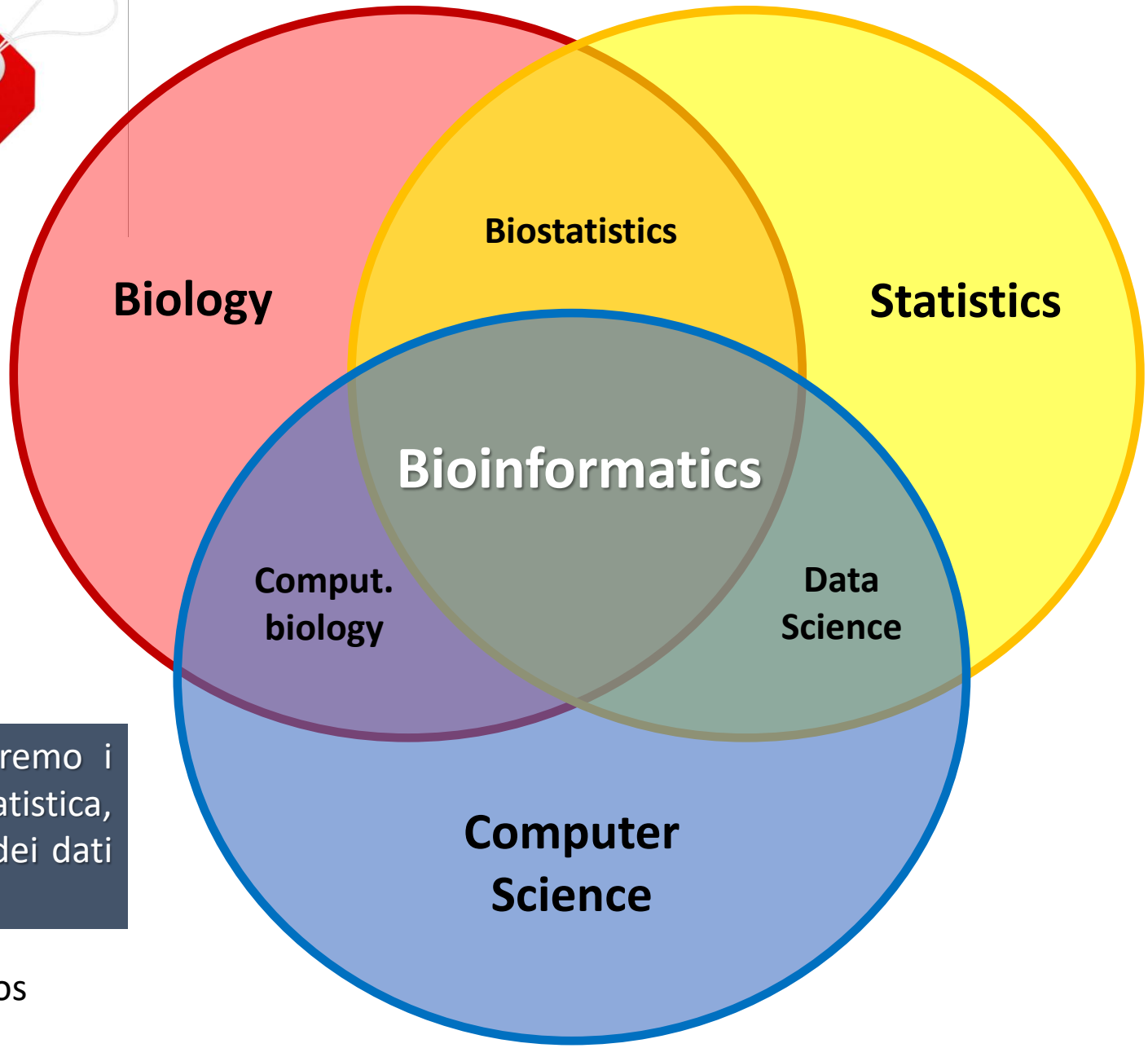
Analisi del trascrittoma con R

PhD. Estevao Barcelos
Università degli Studi di Perugia - UNIPG



2024 Training on Microarray/RNA-seq data analysis using R Studio/Geo2R NCBI pipeline



Obiettivo: In questa lezione approfondiremo i principi della trascrittomica e della statistica, concentrandoci in particolare sull'analisi dei dati di microarray utilizzando R.

https://github.com/estevaobarcelos

Innanzitutto...

- **INTRODUCTION TO R AND RSTUDIO**
  - **R** è un linguaggio di programmazione specifico per la statistica e la grafica computazionali. È uno dei linguaggi più ampiamente utilizzati da statistici, analisti di dati e ricercatori per gestire, manipolare, analizzare e visualizzare.
  - **Rstudio** è un ambiente di sviluppo integrato per R che consente agli utenti di interagire più facilmente con R integrando diversi aspetti della scripting.

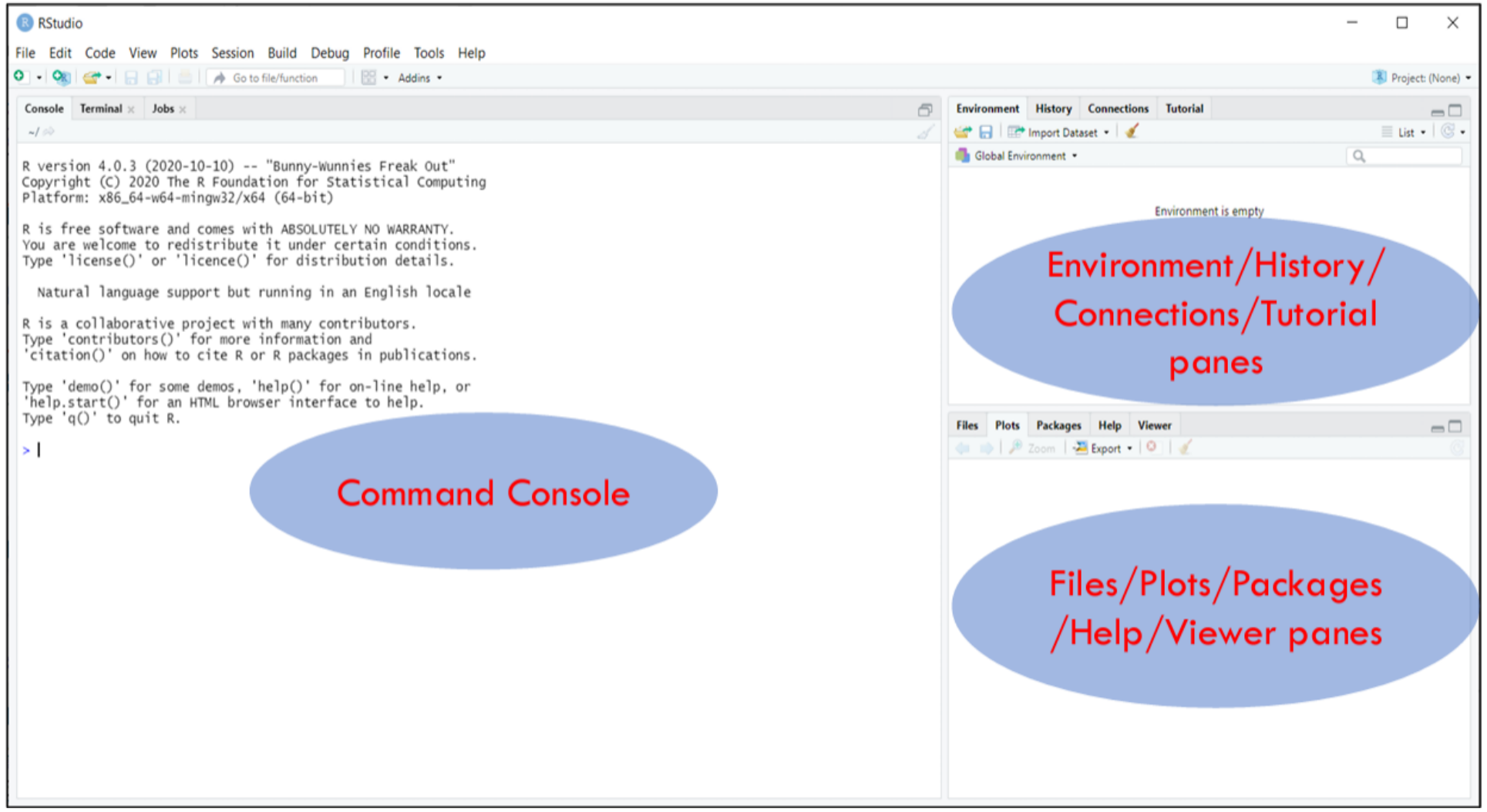
In order to use RStudio, R needs to be installed first.

R and Rstudio Installation

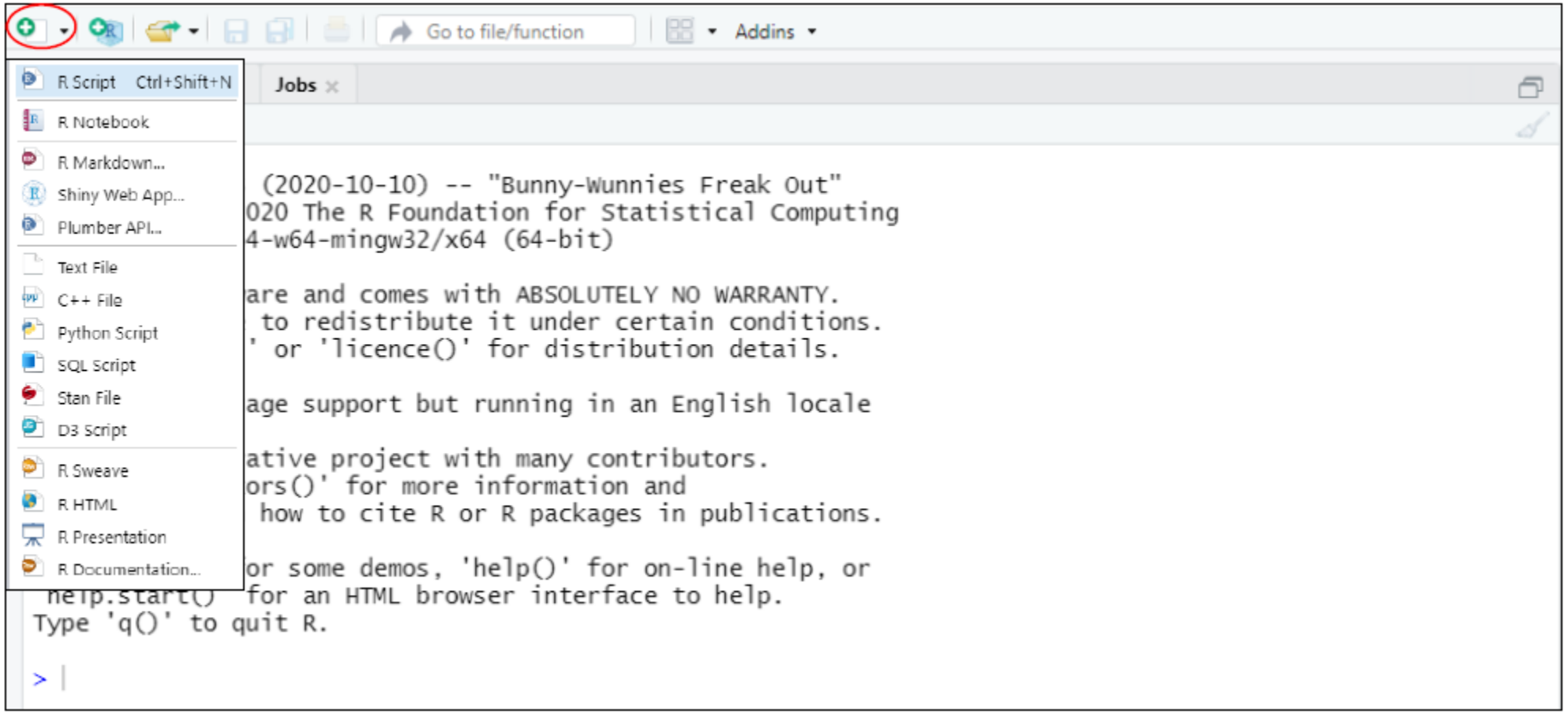
- **Step 1:**
  - Vai al sito Comprehensive R Archive Network (CRAN) e clicca su “Download R for Windows” (or “Download R for MacOS”). **https://cran.r-project.org**
- **Step 2:**
  - Clicca su collegamento sottodirectory “base”.
- **Step 3:**
  - Clicca su “Download R-4.4.2 for Windows”. Il collegamento consente di scaricare un'estensione di installazione (.exe file).
- **Step 4:**
  - Eseguire il file .exe e seguire la procedura guidata di installazione accettando le impostazioni predefinite.

  Una volta installato R, puoi procedere all'installazione di RStudio.
- **Step 5:**
  - Vai al sito Web di download di RStudio e fai clic su tasto “Download RStudio for Windows” (or the link for the MacOS version).
- **Step 6:**
  - Eseguire il file .exe e seguire le istruzioni di installazione.

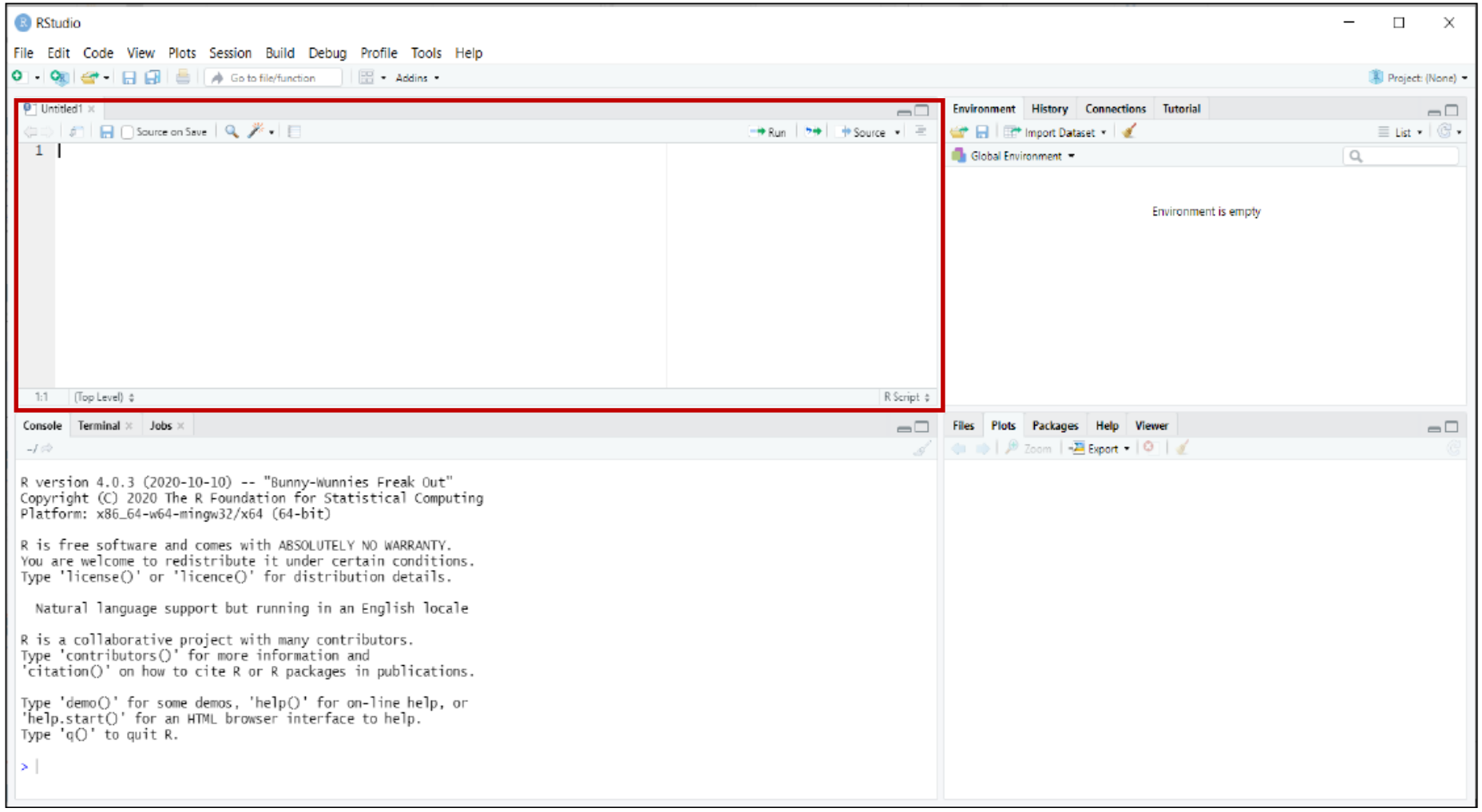
Rstudio Interface



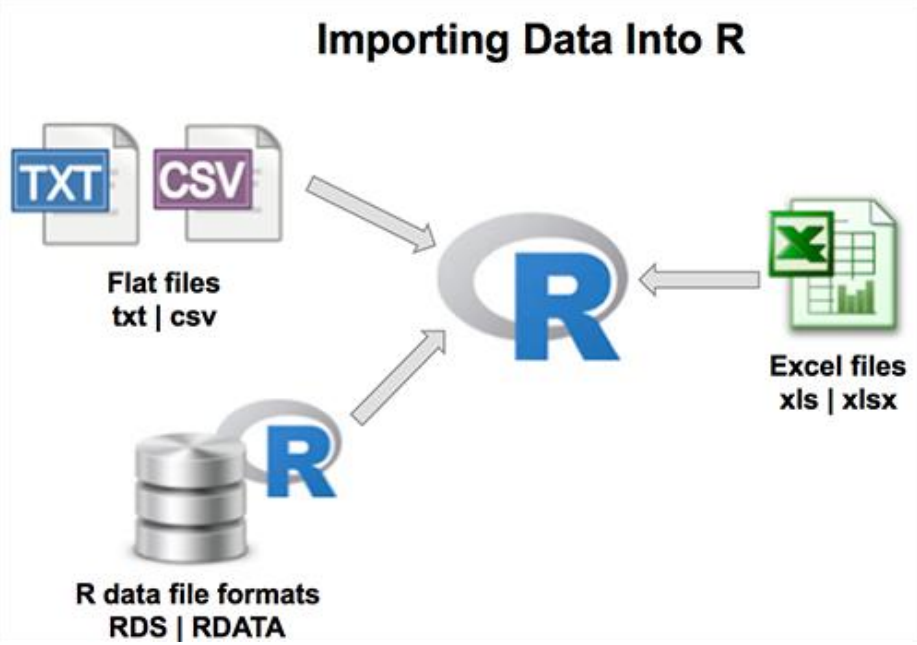
Rstudio: command console

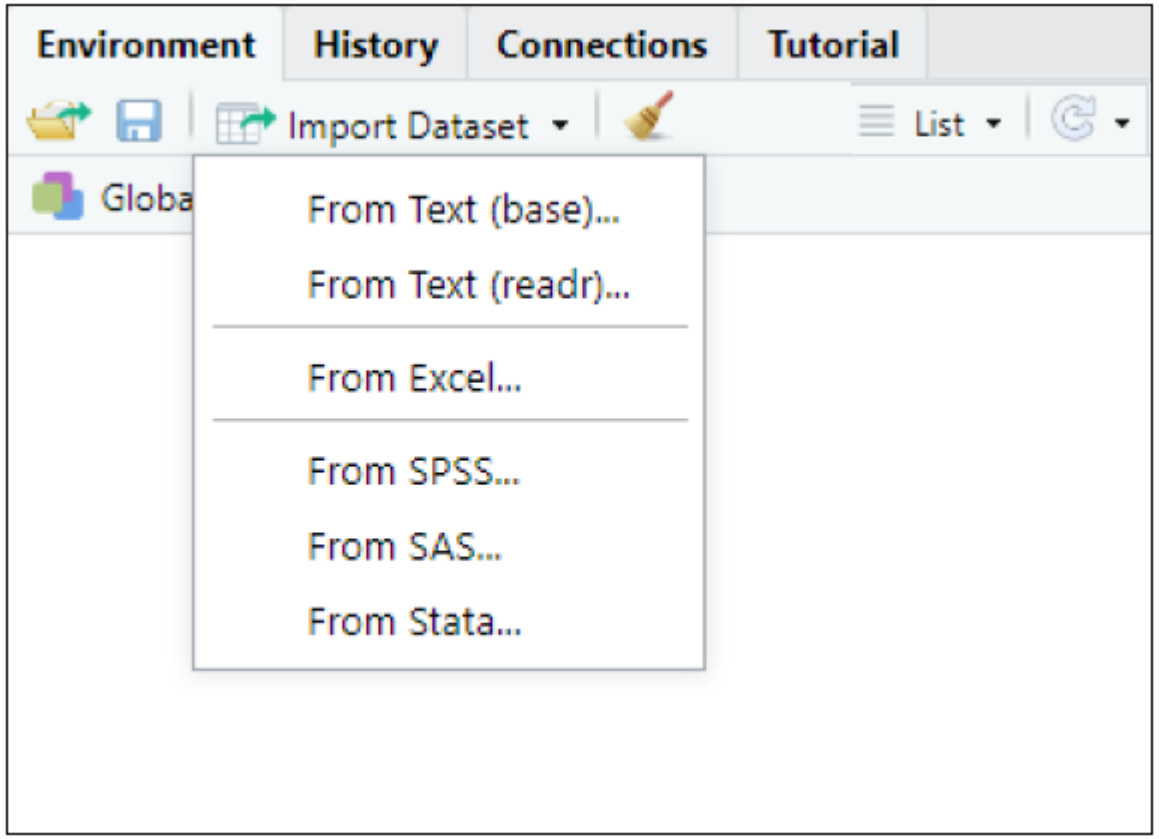


Rstudio interface with editor window open

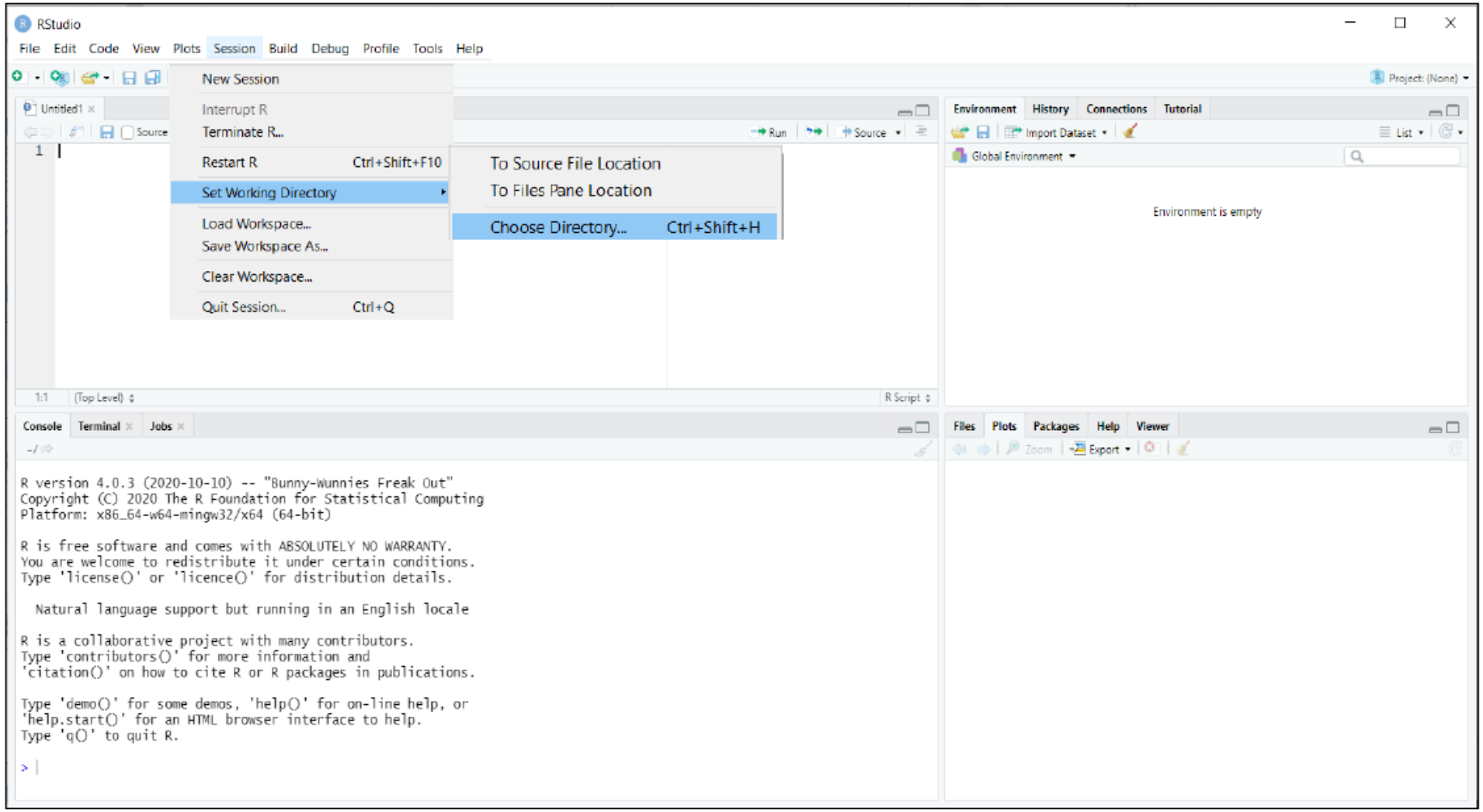


Import dataset tab -scheda importazione dataset





Setting the Working Directory



Get and Set the Working Directory

Getwd function

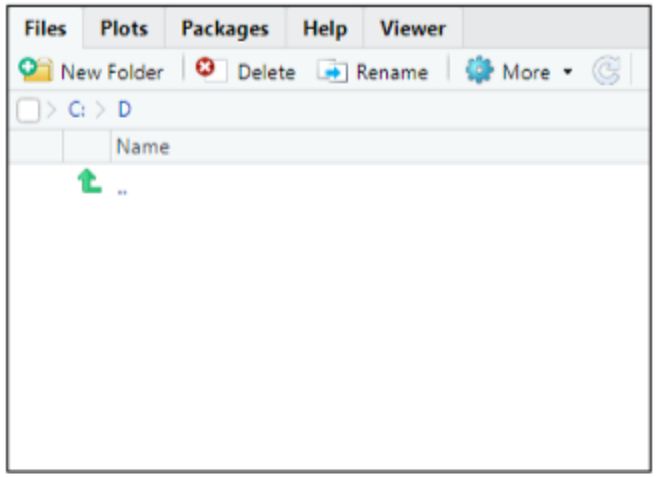
```r
# Find the path of your working directory
getwd()
```

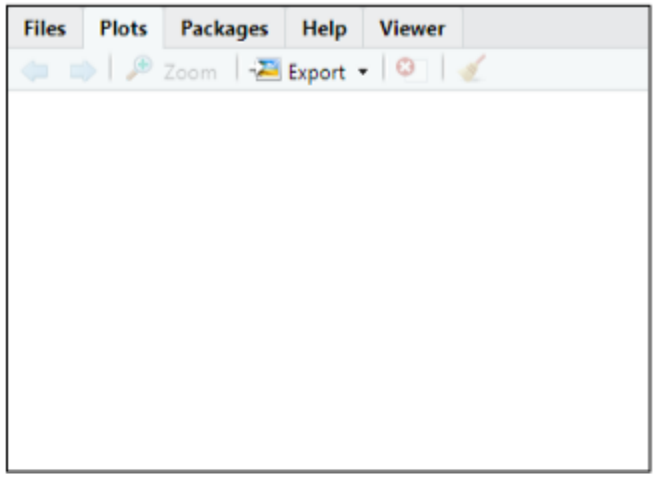
Setwd function

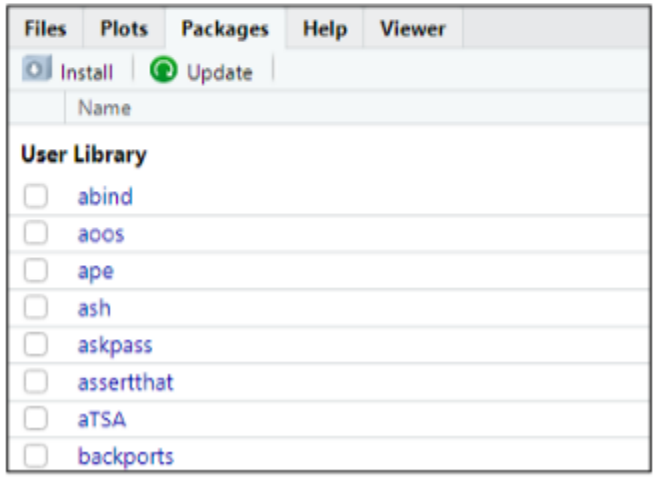
```r
# Set the path of your working directory
setwd("My\\Path")
setwd("My/Path") # Equivalent
```

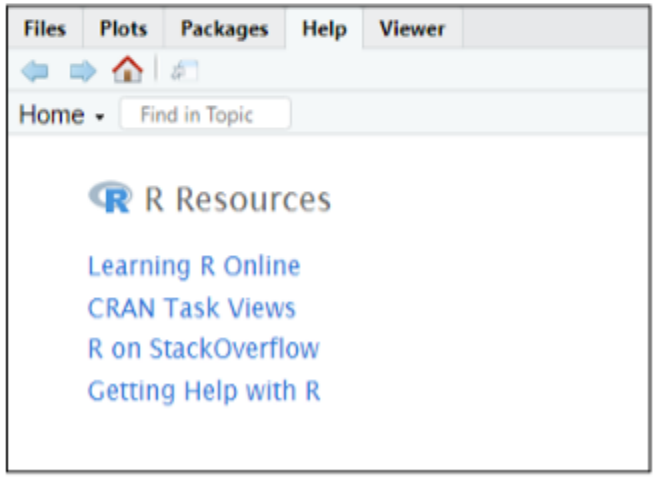
```r
dir()
list.files() # Equivalent
```

Rstudio: Files/Plots/Packages/Help/Viewer panels





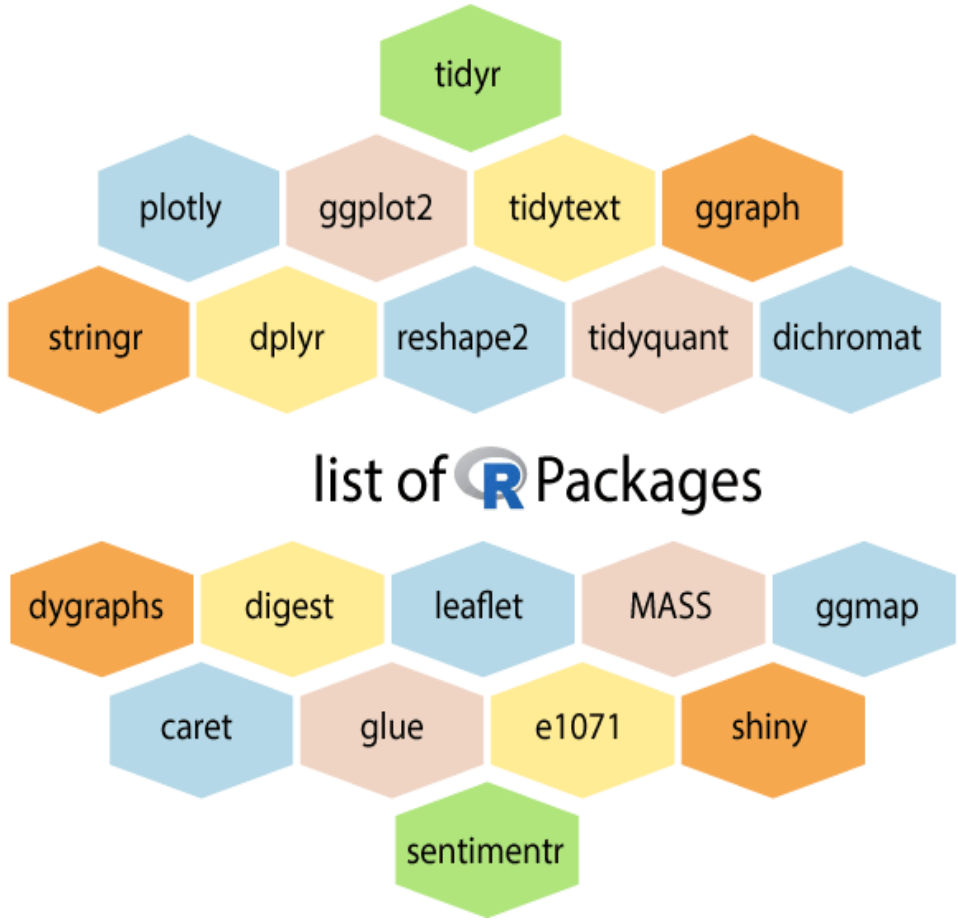




Packages

- **Packages**: estensioni che contengono codice, dati e documentazione in un formato standardizzato che può essere installato e utilizzato dagli utenti di R per risolvere specifici problemi analitici.

- La versione base di R include già molti pacchetti utili che consentono di eseguire attività elementari come calcoli semplici, esplorazione dei dati e caricamento di file di dati di testo.

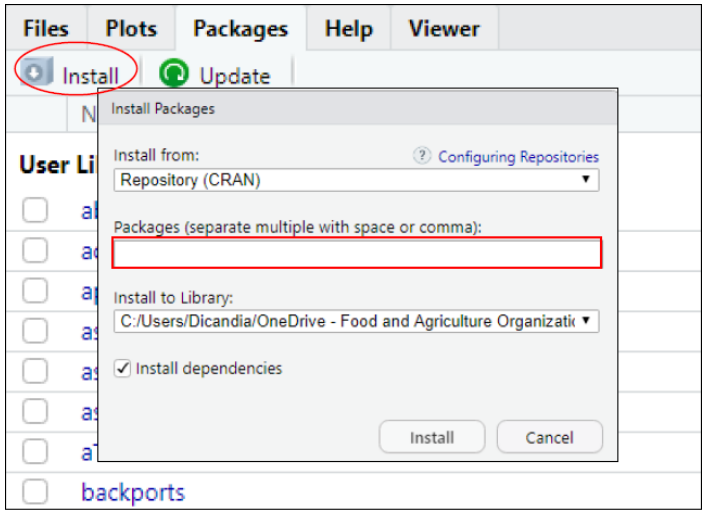


Packages (2)

- A) Per **installare un pacchetto** da CRAN (il repository ufficiale per i pacchetti R forniti dagli utenti) e quindi caricarlo, utilizzare i seguenti comandi:

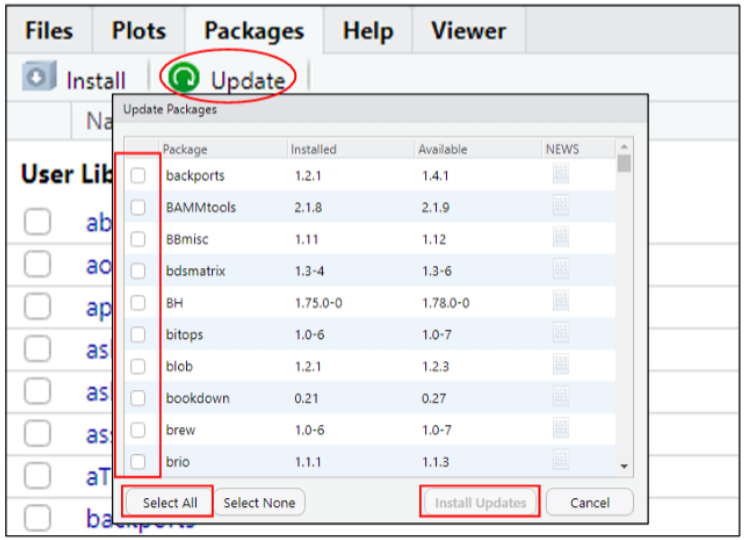
```
install.packages("name of the package")
library("name of the package")
```

- B) Un altro modo per installare i pacchetti è utilizzare la scheda Installa pacchetti:



Packages (3)

- Talvolta i pacchetti R vengono aggiornati per migliorarne o modificarne la funzionalità. Si consiglia di aggiornare occasionalmente i pacchetti installati sul computer.
- Puoi aggiornare i pacchetti R installati in RStudio cliccando sul pulsante Aggiorna nella barra degli strumenti nel pannello Pacchetti.



BiocManager

- Il **BiocManager** è un pacchetto R progettato per facilitare l'installazione, l'aggiornamento e la gestione dei pacchetti **Bioconductor**, una piattaforma open-source dedicata all'analisi di dati biologici come genomica e trascrittomica.
- Offre strumenti per garantire compatibilità tra i pacchetti e la versione di R in uso, semplificando il lavoro dei ricercatori nel campo della bioinformatica.

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```

List of packages used in the training

https://www.ncbi.nlm.nih.gov/geo/

- **GEOquery**: è uno strumento R progettato per interagire con il database Gene Expression Omnibus (GEO) e consente di scaricare, importare e gestire i dati di esperimenti di espressione genica.

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GEOquery")
```

getGEO() **getGSEMatrix()** **Meta()** **GSMList()**

getGEOSuppFiles() **GEOquery::getGPL()** **GPLList()**

List of packages used in the training

- **limma**: (Linear Models for Microarray Data) è un pacchetto per l'analisi dei dati di espressione genica derivanti da tecnologie microarray o RNA-seq.

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("limma")
```

normalizeBetweenArrays()

lmFit()

eBayes()

topTable()

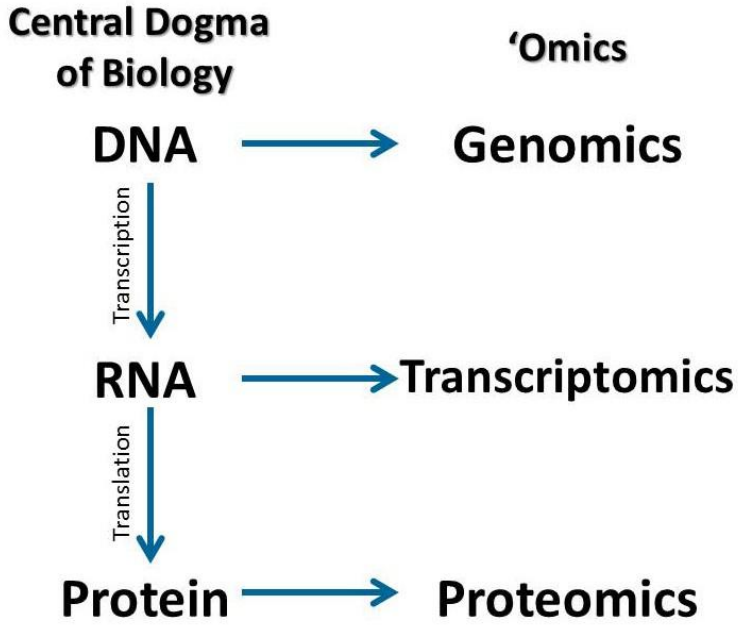
plotMA()

vennDiagram()

L’analisi del trascrittoma

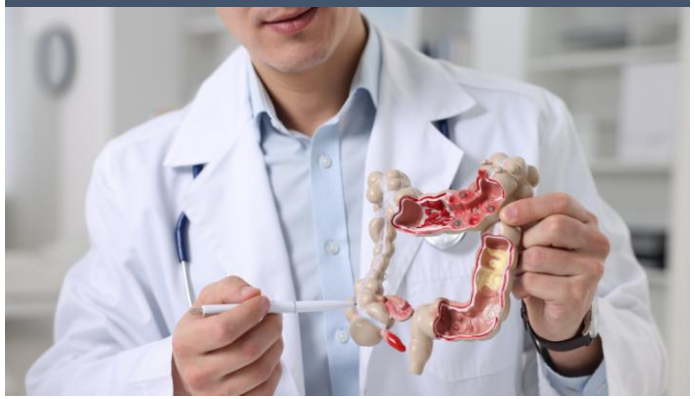
"Mentre la genomica analizza le sequenze di DNA, la trascrittomica si concentra sull'attività dei geni, rivelando quali vengono 'accesi' e 'spenti' e in quale quantità."

- La conoscenza del genoma di un organismo multicellulare è molto importante per comprendere il suo potenziale funzionale…
- Per comprendere come funzionano le diverse tipologie di cellule, organi o tessuti, o come avviene il differenziamento cellulare, è necessario conoscere quale porzione di informazione genetica viene utilizzata, ovvero quali geni vengono espressi.
- Assumiamo che il livello di mRNA trascritti da un dato gene sia un buon indicatore del livello di espressione delle proteine corrispodenti.

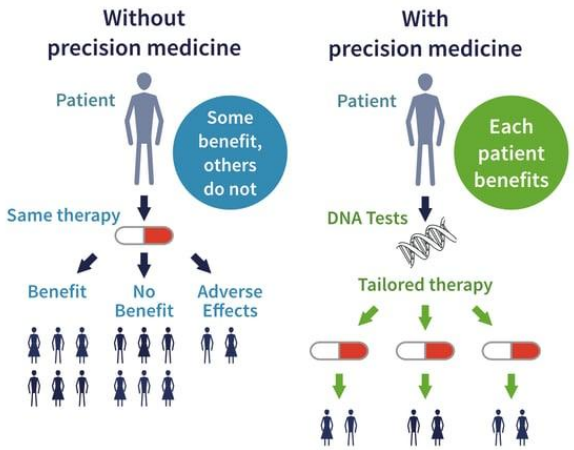


Perché la trascrittomica è importante?

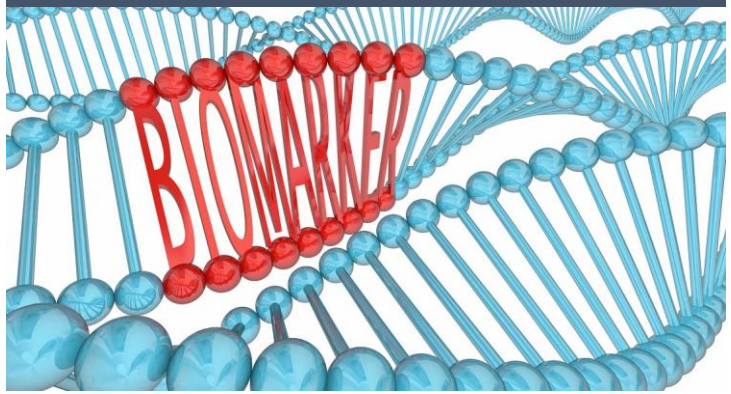
Ricerca sulle malattie



Sviluppo di terapie personalizzate

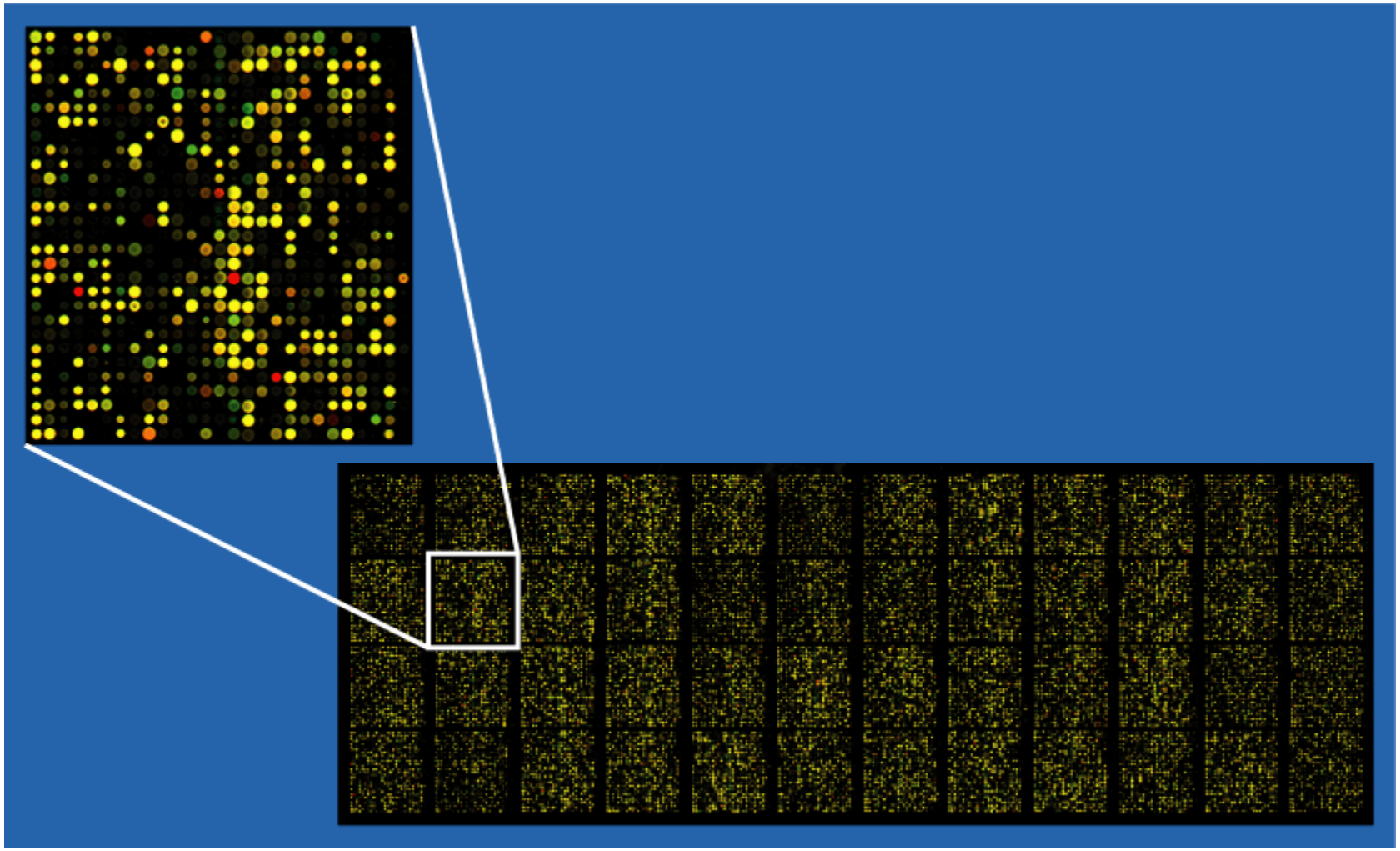


Scoperta di biomarcatori



Comprendere l'impatto ambientale sull'espressione genica.

Microarray

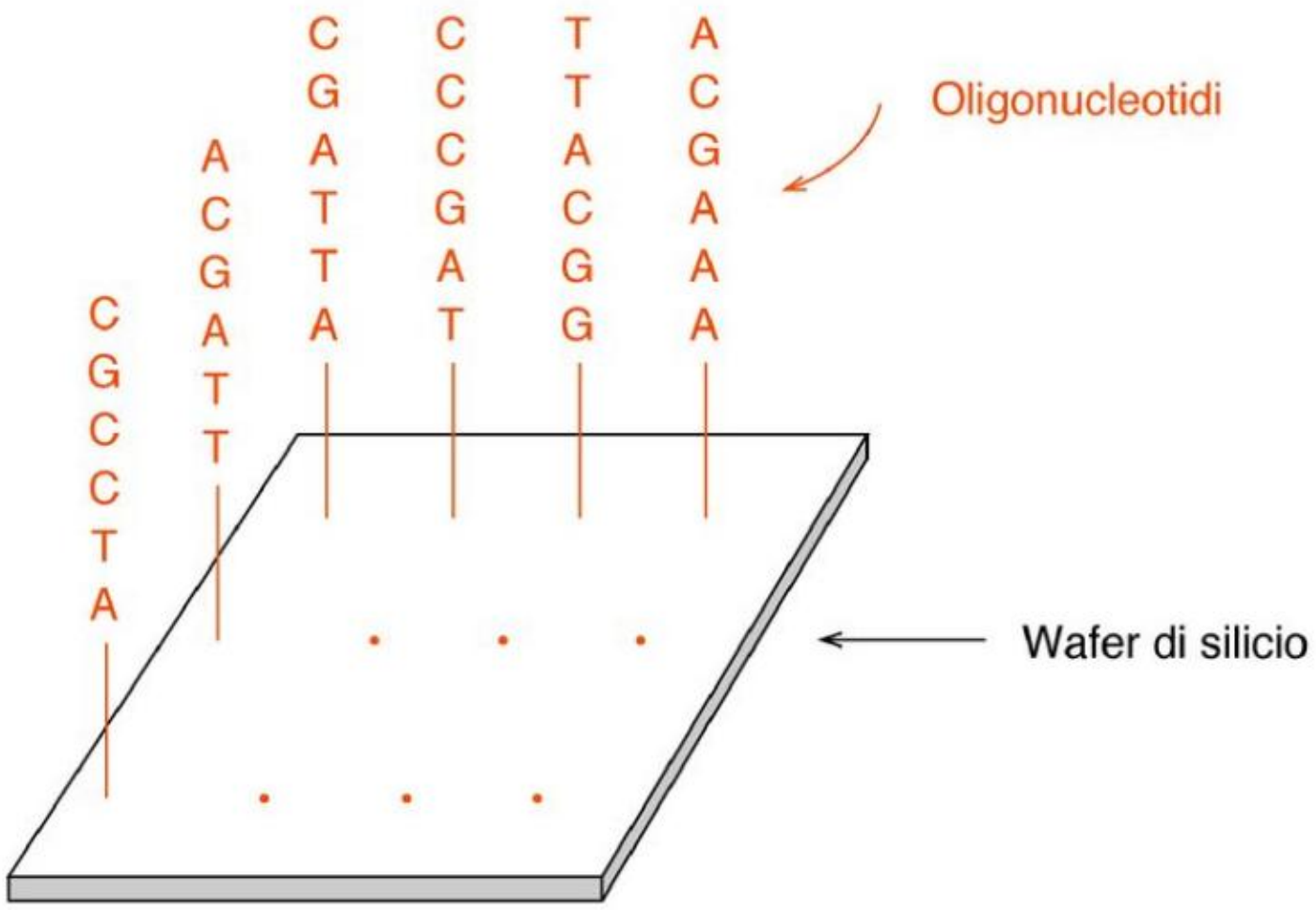


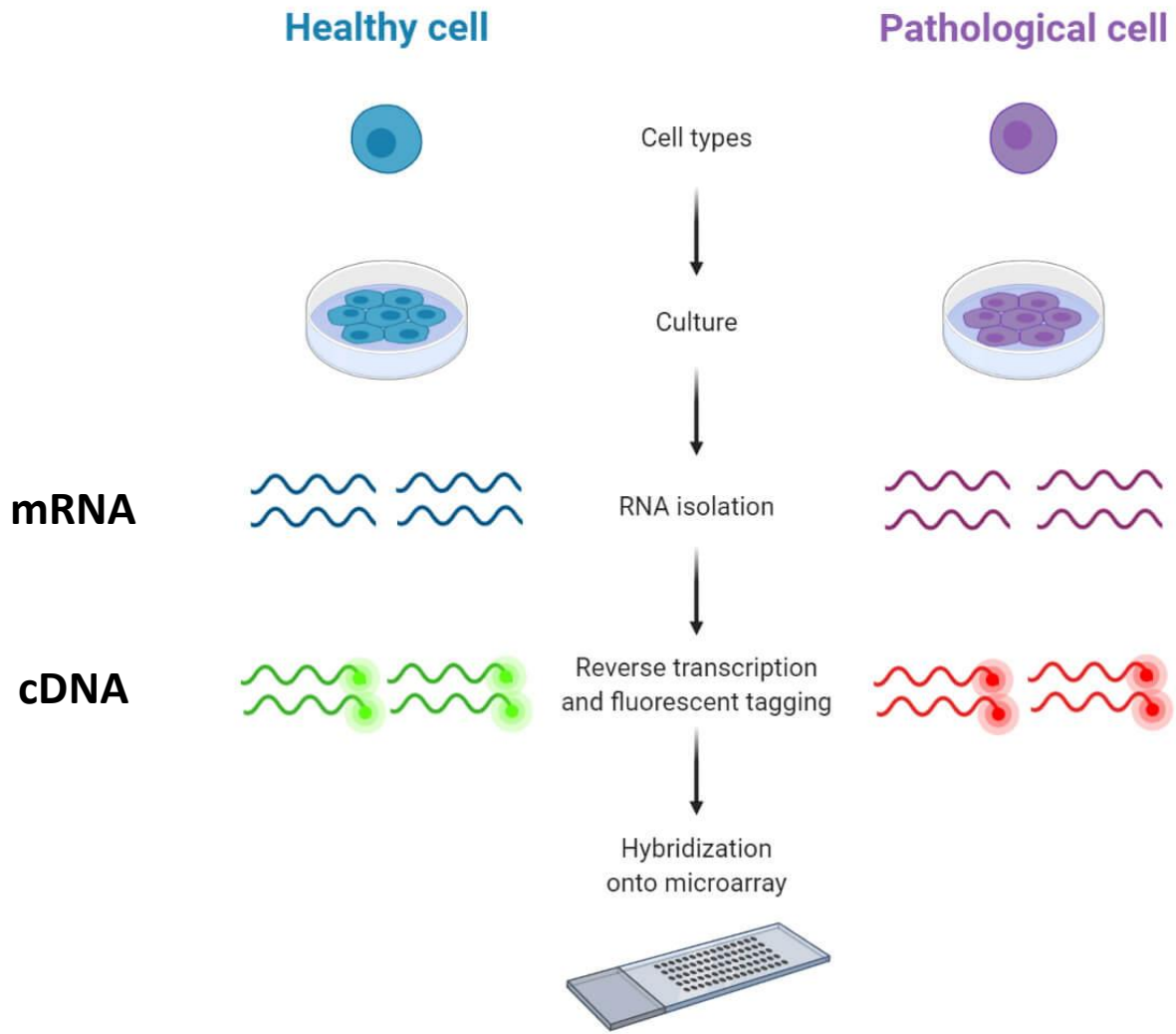
Analisi di una libreria di cDNA attraverso microarray

- ***Microarray***: superficie (una piastra/chip) che contiene migliaia di sonde di DNA immobilizzate in posizioni ben definite.
  - Ogni **sonda** corrisponde a un **gene specifico**.

- ***Campione***: estrazione di RNA cellulare, conversione a cDNA con marcatura con fluorescenza di tutti i cDNA (diversa fluorescenza pr i diversi campioni di Ibridazione ed esame microarray.

- ***Chip a DNA***
  - Microarray ad alta densità con oligo sintetizzati in superficie (1 milione/cm2),
  - Uno o più oligo corrispondono a un DNA espresso (gene).





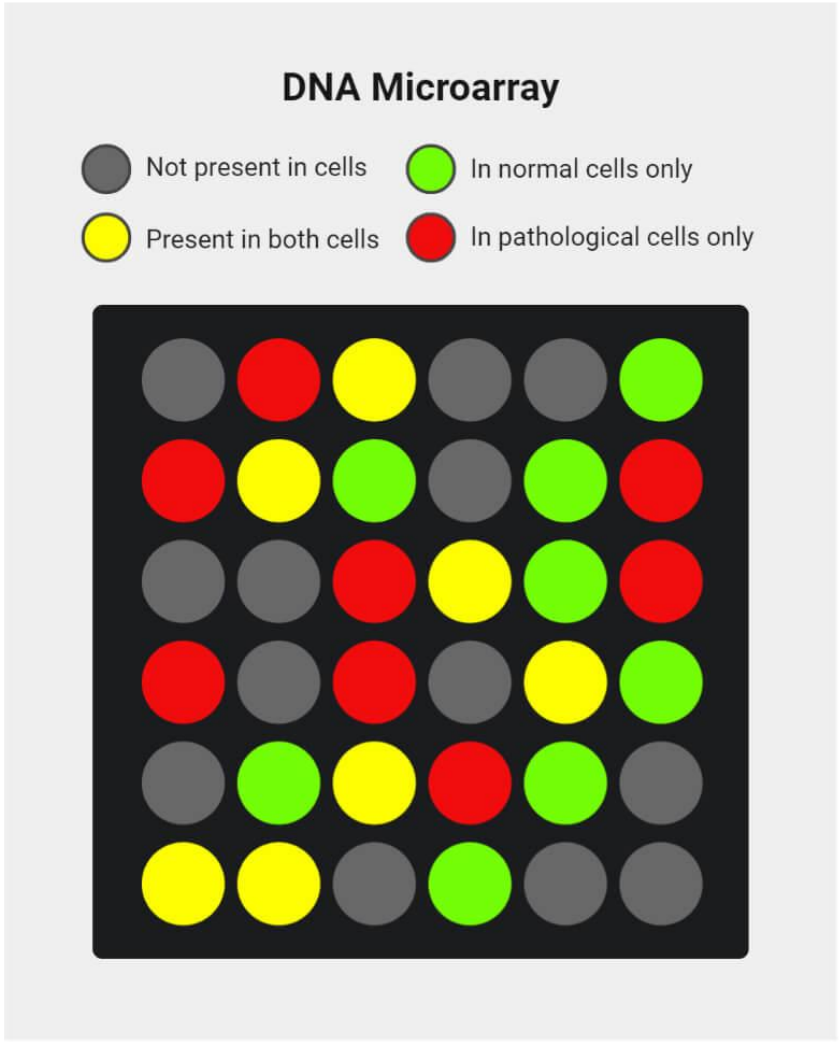
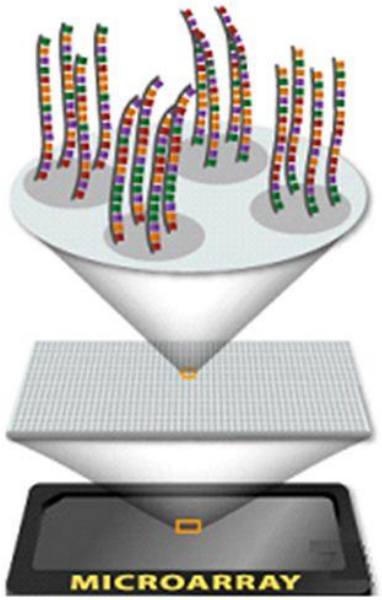
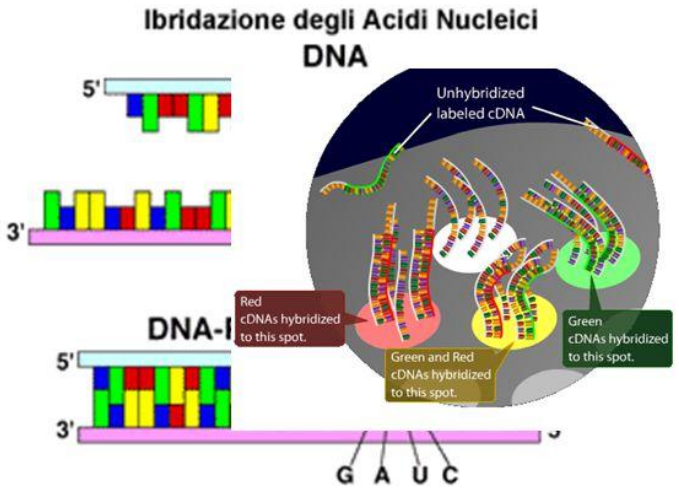
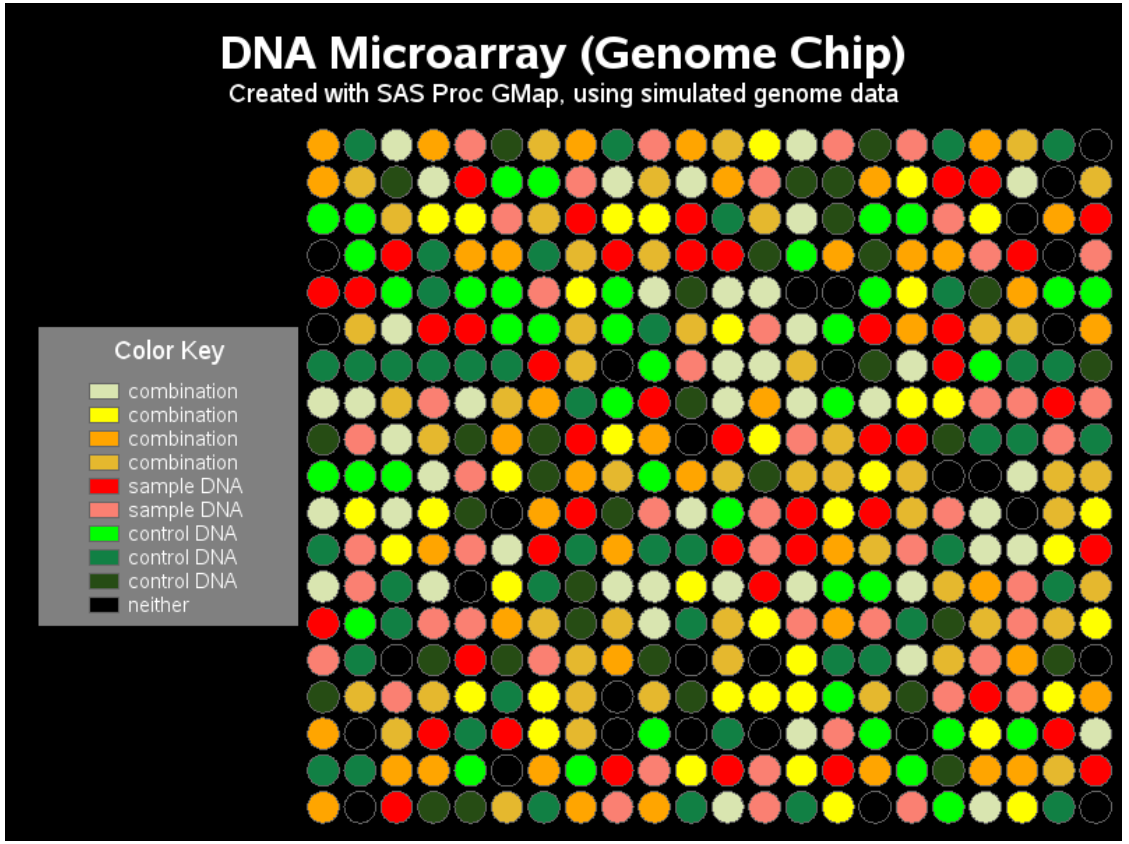


Image By Sagar Aryal, created using biorender.com

Microarray per l'analisi dell'espressione genica

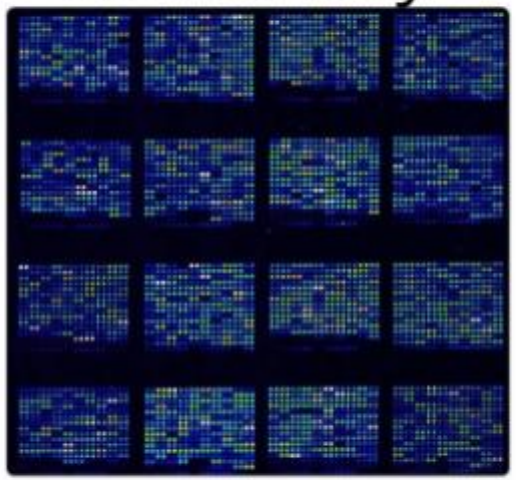






Surescan Dx

Microarray

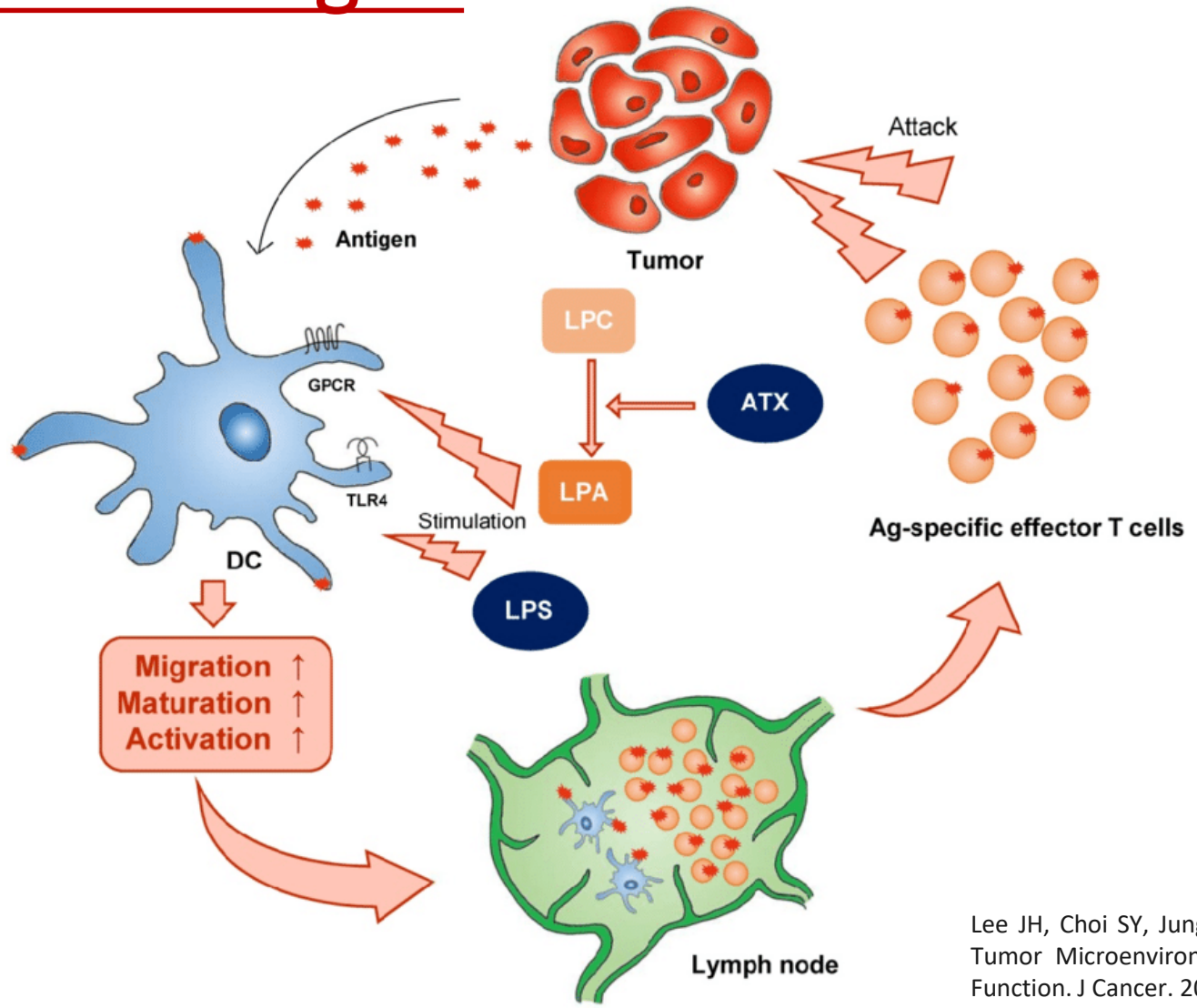


Gene chip 3000 7g

https://www.youtube.com/watch?v=wZN070rl7VA&t=139s

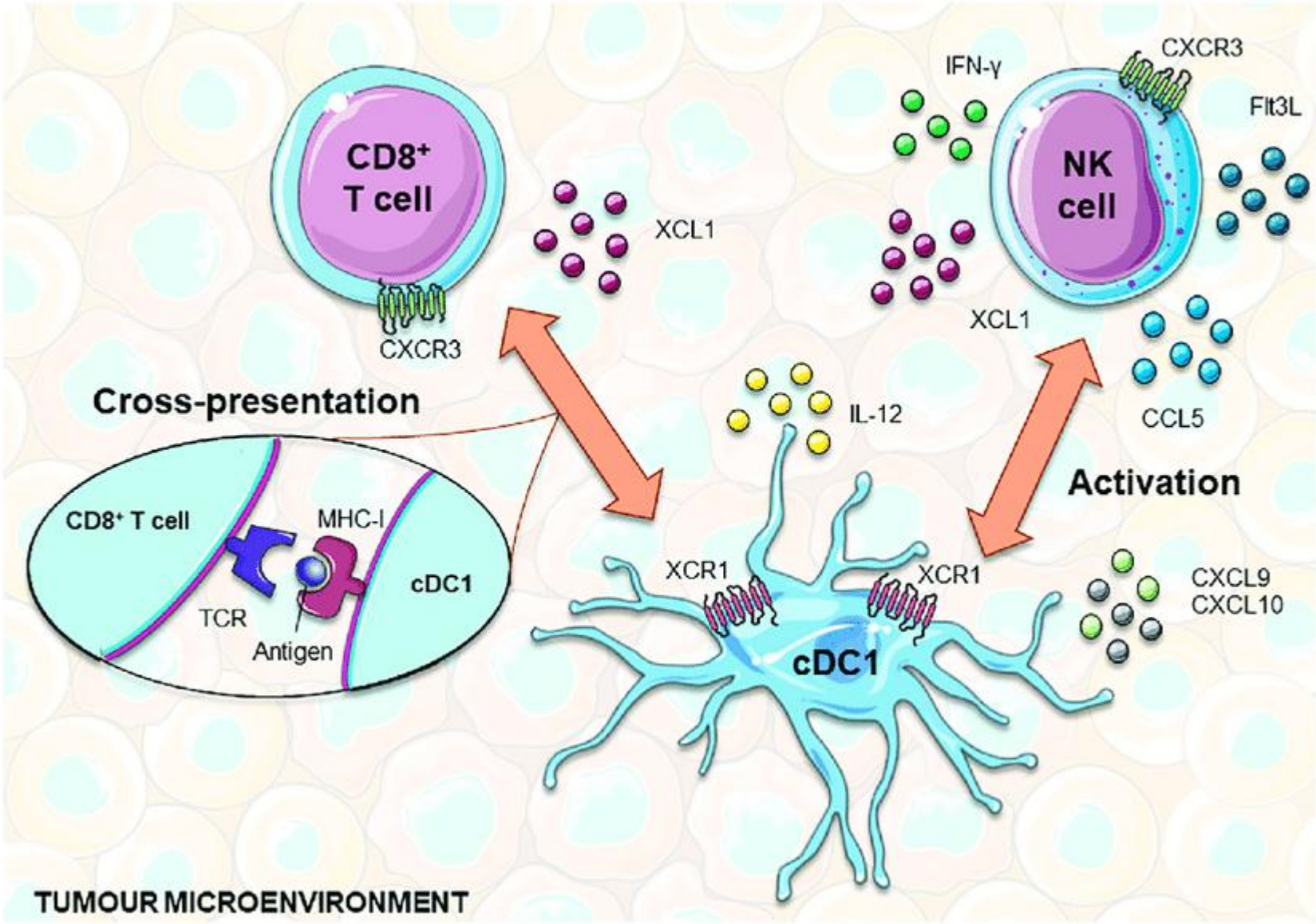
Questione Biologica

Classical stimulators induce DC activation. Classical stimulators (e.g. LPA and LPS) influence DC function.

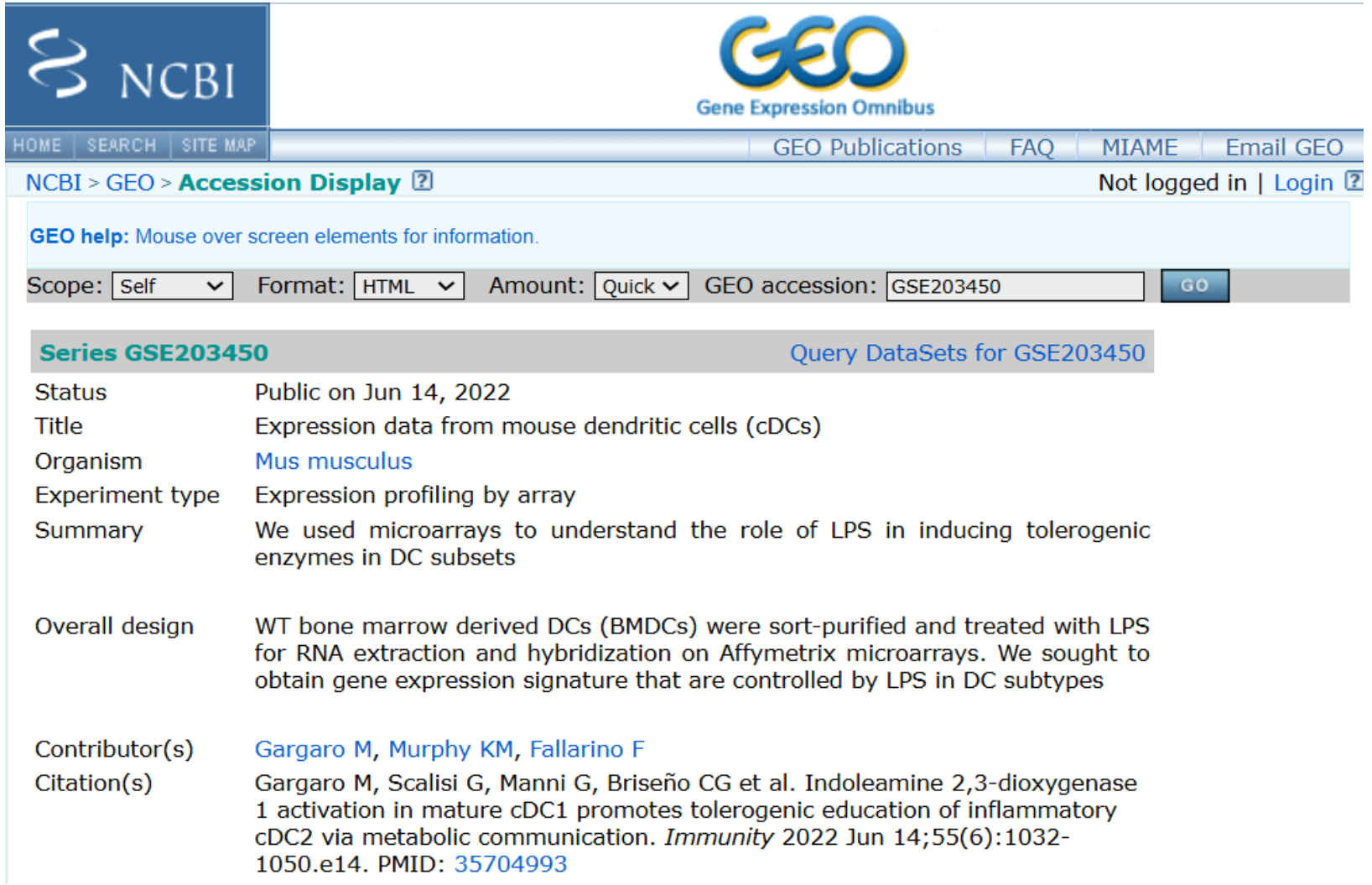


LPS: Agonista TLR4

Lee JH, Choi SY, Jung NC, Song JY, Seo HG, Lee HS, Lim DS. The Effect of the Tumor Microenvironment and Tumor-Derived Metabolites on Dendritic Cell Function. J Cancer. 2020 Jan 1;11(4):769-775.

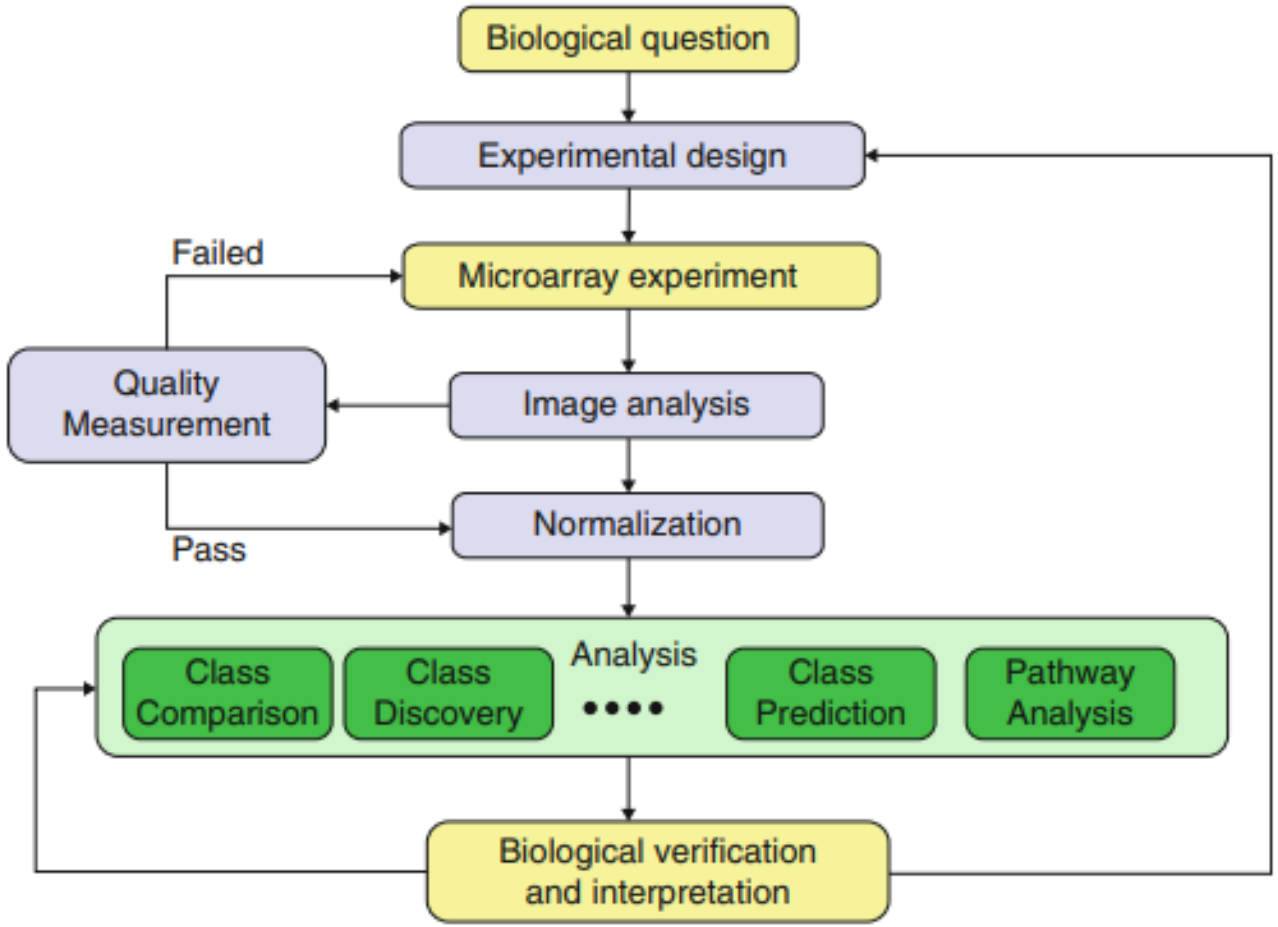

CD8⁺
T cell
CXCR3
XCL1
IFN-γ
CXCR3
Flt3L
NK
cell
XCL1
CCL5
Cross-presentation
IL-12
Activation
CD8⁺ T cell
MHC-I
TCR
Antigen
cDC1
XCR1
XCR1
cDC1
CXCL9
CXCL10
TUMOUR MICROENVIRONMENT

GEO – Gene Expression Omnibus



GEO accession: GSE203450

Pipeline Microarray Analysis



Part 1. Load required libraries

```
#-------------------------------------------------------------------------#
# Pre-processing Microarray Data - QC, Normalization, DEGs and Visualization
#-------------------------------------------------------------------------#
# Analysis of datasets from: GSE203450 (Microarray Affymetric Data from DCs)


# Set Working Directory.-------------------------------------------------
setwd("") # Set the Working Directory on your computer.

# Part 1. Load required libraries.=========================================
  print("loading libraries...")

# Install Bioconductor Manager (needed for some packages).---------------------
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

# Install the required libraries.----------------------------------------------
BiocManager::install("GEOquery")
library(GEOquery)

# BiocManager::install("limma")
library(limma)
```

Altri modi per aprire le biblioteche

- **require("Nome del pacchetto")**

```
> if (require(GEOquery)) {
+     print("Pacchetto caricato con successo!")
+ }
```

- Per i pacchetti non disponibili su CRAN, caricare da altre fonti come ***GitHub***. Utilizza devtools o remotes per installare e caricare:

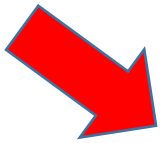
```
> if (!require(devtools)) install.packages("devtools")
  devtools::install_github("username/repository")
  library(package_name)

> devtools::install_github("tidyverse/ggplot2")
  library(ggplot2)
```

Part 2. Import and explore data from GEO

```
# Part 2. Import and explore data from GEO.=====================================
  print("Importing data...")

  id <- "GSE203450"
  gse <- getGEO(id, GSEMatrix =TRUE, AnnotGPL=TRUE)

  # Get some information about the file.----------------------------------------
  list(gse)
  names(pData(gse[[1]]))      # print the sample info.
  length(gse)                 # check how many platforms was used.

  # Select the dataset.--------------------------------------------------------
  gse <- gse[[1]]

  # if more than one dataset is present, you can analyse the other dataset by
  # changing the number inside the [[...]]
      # e.g. gse <- gse[[2]]

  # We can also use that kind of line.
  # if (length(gse) > 1) idx <- grep("GPL6246", attr(gse, "names")) else idx <- 1


   # Explore the expression dataframe.
   pData(gse)                    # print the sample info.
   fData(gse)                    # print the gene annotation.
   exprs(gse)                    # print the expression data.
```

```
>   list(gse)
[[1]]
ExpressionSet (storageMode: lockedEnvironment)
assayData: 28853 features, 18 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM6172191 GSM6172192 ... GSM6172208 (18 total)
  varLabels: title geo_accession ... tissue:ch1 (38 total)
  varMetadata: labelDescription
featureData
  featureNames: 10344614 10344616 ... 10608630 (28853 total)
  fvarLabels: ID Gene title ... GO:Component ID (21 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 35704993
Annotation: GPL6246

>   names(pData(gse[[1]]))      # print the sample info.
 [1] "title"                 "geo_accession"           "status"                  "submission_date"
 [5] "last_update_date"      "type"                    "channel_count"           "source_name_ch1"
 [9] "organism_ch1"          "characteristics_ch1"     "characteristics_ch1.1"   "characteristics_ch1.2"
[13] "treatment_protocol_ch1" "growth_protocol_ch1"     "molecule_ch1"            "extract_protocol_ch1"
[17] "label_ch1"             "label_protocol_ch1"      "taxid_ch1"               "hyb_protocol"
[21] "scan_protocol"         "description"             "data_processing"         "platform_id"
[25] "contact_name"          "contact_phone"           "contact_department"      "contact_institute"
[29] "contact_address"       "contact_city"            "contact_state"           "contact_zip/postal_code"
[33] "contact_country"       "supplementary_file"      "data_row_count"          "genotype:ch1"
[37] "strain:ch1"            "tissue:ch1"

>   length(gse)               # check how many platforms was used.
[1] 1
```

```
>     pData(gse)                    # print the sample info.
                                 title geo_accession                status submission_date last_update_date type
GSM6172191          WT pDC_untreated_rep1    GSM6172191 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172192          WT pDC_untreated_rep2    GSM6172192 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172193          WT pDC_untreated_rep3    GSM6172193 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172194   WT pDC_activated (LPS)_rep1    GSM6172194 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172195   WT pDC_activated (LPS)_rep2    GSM6172195 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172196   WT pDC_activated (LPS)_rep3    GSM6172196 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172197         WT cDC1_untreated_rep1    GSM6172197 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172198         WT cDC1_untreated_rep2    GSM6172198 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172199         WT cDC1_untreated_rep3    GSM6172199 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172200 WT cDC1_activated (LPS)_rep1    GSM6172200 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172201 WT cDC1_activated (LPS)_rep2    GSM6172201 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172202 WT cDC1_activated (LPS)_rep3    GSM6172202 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172203         WT cDC2_untreated_rep1    GSM6172203 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172204         WT cDC2_untreated_rep2    GSM6172204 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172205         WT cDC2_untreated_rep3    GSM6172205 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172206 WT cDC2_activated (LPS)_rep1    GSM6172206 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172207 WT cDC2_activated (LPS)_rep2    GSM6172207 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
GSM6172208 WT cDC2_activated (LPS)_rep3    GSM6172208 Public on Jun 14 2022     May 20 2022      Jun 14 2022  RNA
           channel_count                   source_name_ch1 organism_ch1              characteristics_ch1
GSM6172191             1  pDC isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172192             1  pDC isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172193             1  pDC isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172194             1  pDC isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172195             1  pDC isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172196             1  pDC isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172197             1 cDC1 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172198             1 cDC1 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172199             1 cDC1 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172200             1 cDC1 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172201             1 cDC1 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172202             1 cDC1 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172203             1 cDC2 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
GSM6172204             1 cDC2 isolated from WT Flt3-BMDCs Mus musculus tissue: Bone marrow derived DCs
```



```
>     fData(gse)                    # print the gene annotation.
                ID                                                                                                  Gene title
10344614 10344614                                                                                          predicted gene 2889
10344616 10344616
10344618 10344618
10344620 10344620                                                                                          predicted gene 10568
10344622 10344622
10344624 10344624                                                                                                lysophospholipase 1
10344633 10344633                                                                                transcription elongation factor A (SII) 1
10344637 10344637                                                                      ATPase, H+ transporting, lysosomal V1 subunit H
10344653 10344653                                                                                        opioid receptor, kappa 1
10344658 10344658                                                                                  RB1-inducible coiled-coil 1
10344674 10344674                                                                    family with sequence similarity 150, member A
10344679 10344679                                                                          suppression of tumorigenicity 18
10344705 10344705
10344707 10344707                                                  protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1
10344713 10344713                                                                 S-adenosylhomocysteine hydrolase///predicted gene 4737
10344715 10344715                                                                                      predicted gene, 30414
10344717 10344717
10344719 10344719                                                                  ring finger protein 7 pseudogene///ring finger protein 7
10344721 10344721
10344723 10344723                                                                                  ribosome biogenesis regulator 1
10344725 10344725                                                                        alcohol dehydrogenase, iron containing, 1
10344741 10344741 heterogeneous nuclear ribonucleoprotein A3 pseudogene///heterogeneous nuclear ribonucleoprotein A3
10344743 10344743                                                                                  RIKEN cDNA 3110035E14 gene
10344750 10344750                                                                          serum/glucocorticoid regulated kinase 3
10344772 10344772                                                                  minichromosome maintenance domain containing 2
10344789 10344789                                                                centrosome and spindle pole associated protein 1
10344797 10344797                                                                centrosome and spindle pole associated protein 1
10344799 10344799                                                                centrosome and spindle pole associated protein 1
10344801 10344801                                                                centrosome and spindle pole associated protein 1
10344803 10344803                                                                centrosome and spindle pole associated protein 1
```

```
>     exprs(gse)                    # print the expression data.
          GSM6172191 GSM6172192 GSM6172193 GSM6172194 GSM6172195 GSM6172196 GSM6172197 GSM6172198 GSM6172199 GSM6172200
10344614    109.847     88.484    107.828    102.969    103.437    114.278    112.667    109.742    116.739    129.456
10344616      6.427      6.455      7.069      6.717      6.587      7.585      6.362      6.687      6.393      6.200
10344618      9.538      8.063      8.625      8.244      9.297      9.263      8.470      8.333      8.131      9.918
10344620     27.262     26.002     25.437     21.966     23.633     25.847     26.793     29.451     25.089     30.966
10344622    169.643    146.881    143.827    134.991    177.279    170.648    133.768    153.171    181.502    145.824
10344624    266.608    239.226    250.661    218.589    228.583    271.410    237.238    247.015    290.231    222.051
10344633    751.896    539.310    677.006    654.438    635.193    677.426    647.876    727.015    588.344    693.568
10344637    285.090    367.982    303.612    358.949    354.451    362.860    233.057    262.295    268.416    240.799
10344653     18.618     15.632     18.166     16.192     14.598     14.217     21.143     16.423     16.219     12.631
10344658    236.535    285.525    254.945    242.434    274.885    221.629    267.975    265.944    276.128    304.479
10344674     14.193     13.903     14.432     14.665     12.702     14.385     14.049     13.202     15.903     15.740
10344679     64.290     74.270     75.216     87.059     85.975     79.868     60.216     65.746     75.252     97.496
10344705    111.926     99.328    101.250     96.119    113.184     96.298    128.605    125.420    113.499    136.842
10344707    263.257    299.612    310.842    234.002    239.783    224.276    274.102    261.410    224.838    250.679
10344713    264.830    264.658    241.719    204.200    232.481    304.528    191.195    208.750    240.040    219.350
10344715     29.725     27.262     24.757     26.610     27.515     26.435     25.673     24.806     25.196     26.940
10344717     51.700     61.483     93.079     62.350     63.386     42.714     63.403     84.022     73.543     51.950
10344719     90.737     88.424     96.520     85.889     79.839    101.214     88.499     89.431     97.980     92.323
10344721      6.153      6.544      6.397      6.319      6.784      6.637      6.329      7.031      6.142      6.603
10344723    210.302    346.573    249.945    225.517    274.970    248.559    213.344    227.629    279.853    205.144
10344725     66.629     67.864     70.005     88.766     84.158     81.080     63.289     62.486     73.613     62.853
10344741    607.019    892.090    849.388    836.131    862.923    719.113    693.475    792.967    686.353    761.475
```

Part 3. Group membership for all samples

```
# Part 3. Group membership for all samples.=====================================
print("Grouping the samples...")

# Select one comparison between groups.
cdc1_lps <- "XXXXXX111000XXXXXX"           # (cDC1 LPS vs cDC1 Ctrl)

# Split Samples.-------------------------------------------------------------
sml <- strsplit(cdc1_lps, split ="")[[1]]
```

0 è la condizione che voglio studiare e 1 è il mio controllo

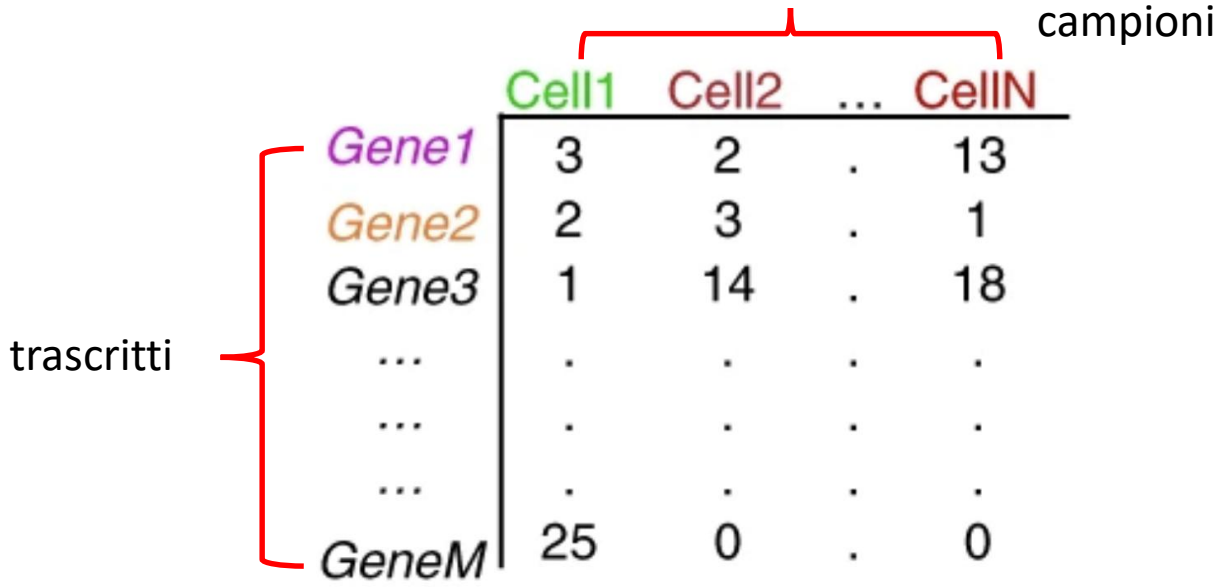
Part 4. Filter out excluded samples and create the Expression matrix

```
# Part 4. Filter out excluded samples and Create the expression matrix.=========
print("Filtering out excluded samples...")
sel <- which(sml != "X")   # excluded samples marked as "X".
sml <- sml[sel]
gse <- gse[ ,sel]

head(gse)

# Create the expression matrix.----------------------------------------------
ex <- exprs(gse)
ex[which(ex <= 0)] <- NaN

head(ex)
hist(ex)
```

sostituisce valori uguali o inferiori a 0 con "NaN" (Not a Number), perché valori negativi o pari a zero non hanno alcun significato biologico in termini di espressione genica e possono falsare l'analisi.

Matrice di Espressione

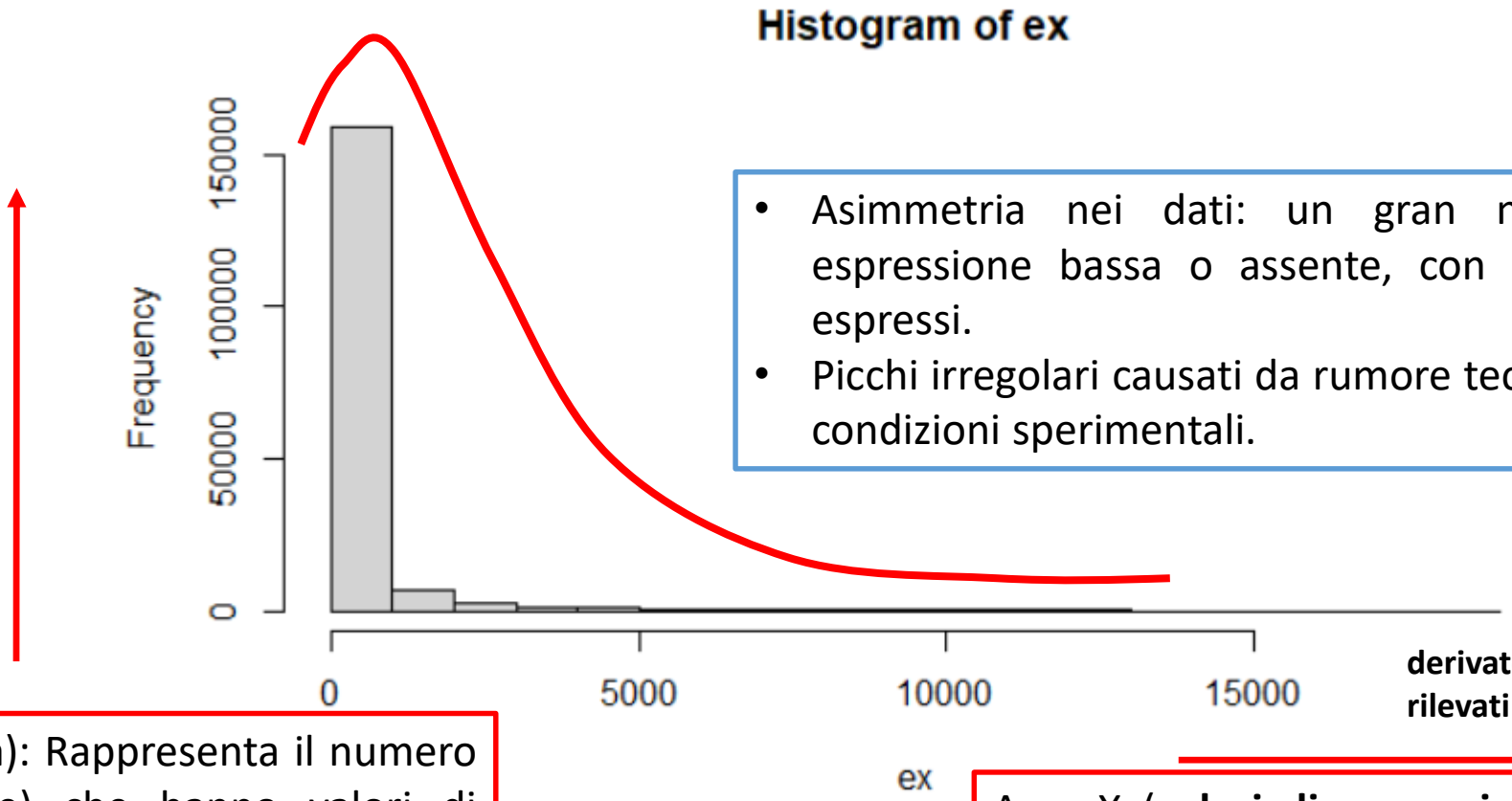
- L'expression matrix (matrice di espressione) è una rappresentazione fondamentale dei dati di espressione genica.
- Si tratta di una matrice in cui le **righe** rappresentano **geni o trascritti** e le **colonne** rappresentano **campioni o condizioni** sperimentali.



Questa matrice contiene i valori di espressione per ciascun gene o sonda sul microarray.

Istogramma - Espressione

L'istogramma rappresenta il numero di geni (o sonde) i cui valori di espressione rientrano in un certo intervallo.



- Asimmetria nei dati: un gran numero di geni con espressione bassa o assente, con meno geni altamente espressi.
- Picchi irregolari causati da rumore tecnico o differenze tra le condizioni sperimentali.

derivato dall'intensità dei segnali rilevati sul microarray

Asse Y (**frequenza**): Rappresenta il numero di geni (o sonde) che hanno valori di espressione all'interno di ciascun intervallo (o intervallo).

Asse X (**valori di espressione**): rappresenta i livelli di espressione genica (valori numerici) per le sonde o i geni nel set di dati.

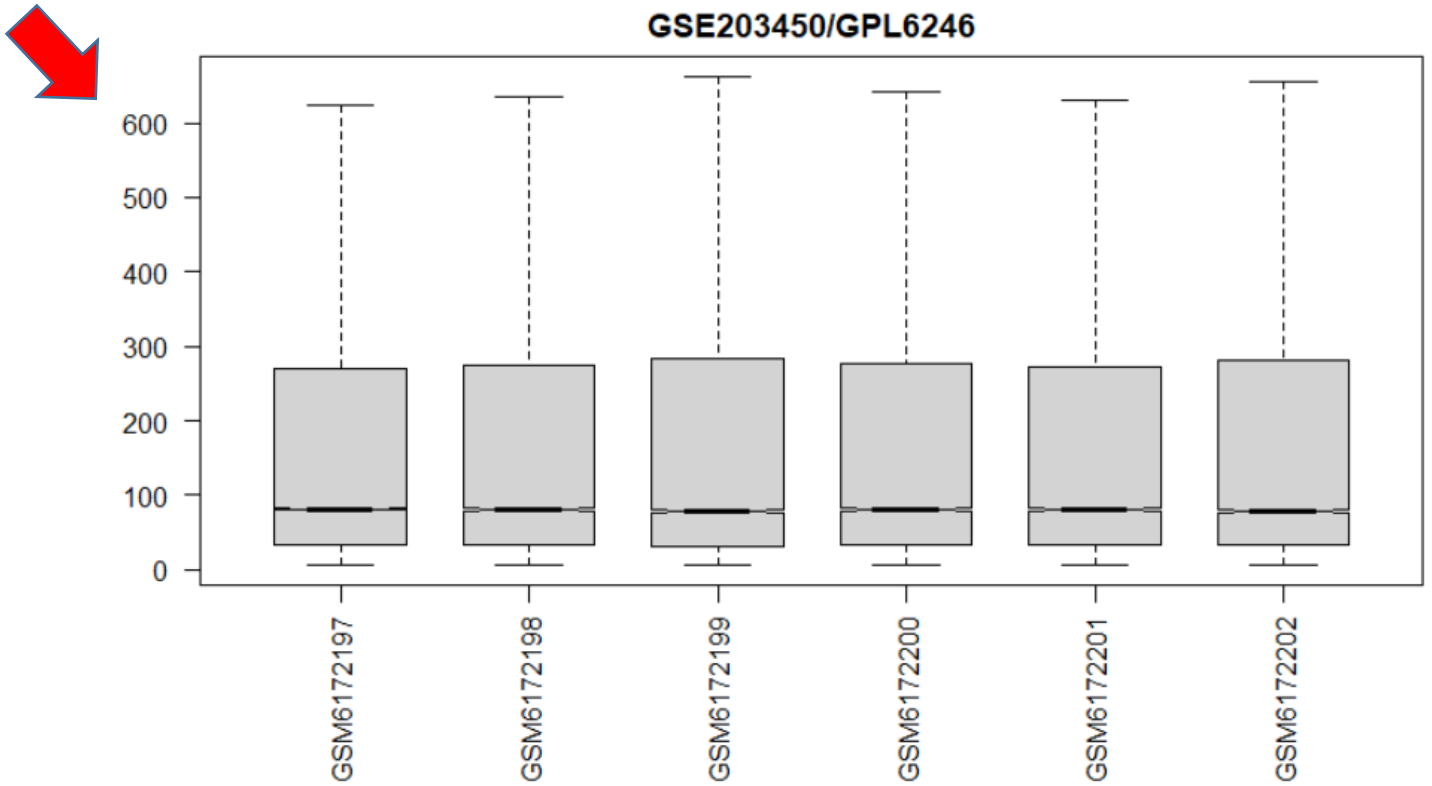
Part 5. Log2 transformation and Normalization of the data

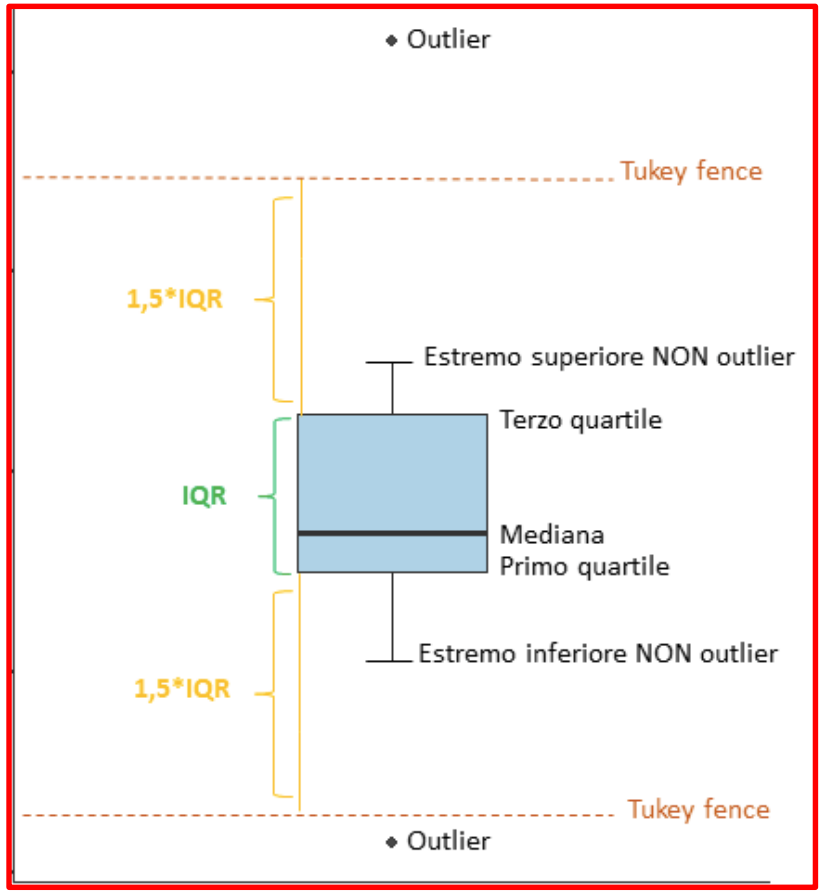
```
# Part 5. Log2 transformation and Normalization of the data.====================
print("Transforming and Normalizing data...")

# Box-and-whisker plot (all samples before normalization).-------------------
par(mar=c(7,4,2,1))
title <- paste ("GSE203450", "/", annotation(gse), sep ="")

# pdf("Microarray_data_before_transformation.pdf")
boxplot(ex, boxwex=0.7, notch=T, main=title, outline=FALSE, las=2)
# dev.off()
```

boxwex = Regola la larghezza relativa dei riquadri nel boxplot.
notch = rappresenta l'intervallo di confidenza attorno alla mediana.
main = Imposta il titolo del grafico.
outline = «outliers». comune nei microarray, ma non sempre rilevante.
las = Imposta l'orientamento delle etichette dell'asse X.

BoxPlot – before transformation

Il boxplot fornisce informazioni sulla mediana, sui quartili e sui valori anomali dei valori di espressione genica per ciascun campione.



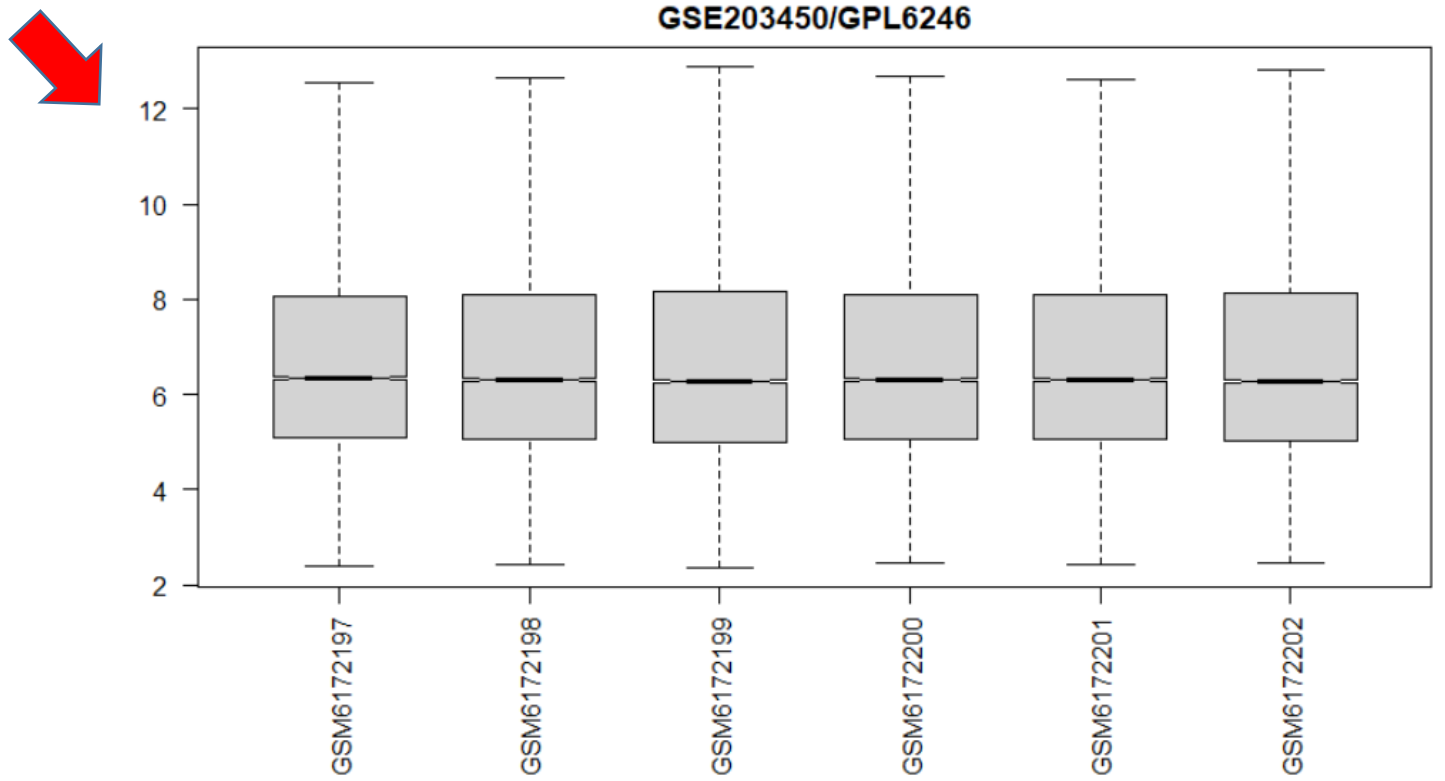


È particolarmente utile per confrontare le distribuzioni tra diversi campioni e verificare eventuali errori sperimentali o la necessità di normalizzazione.

Part 5. Log2 transformation and Normalization of the data

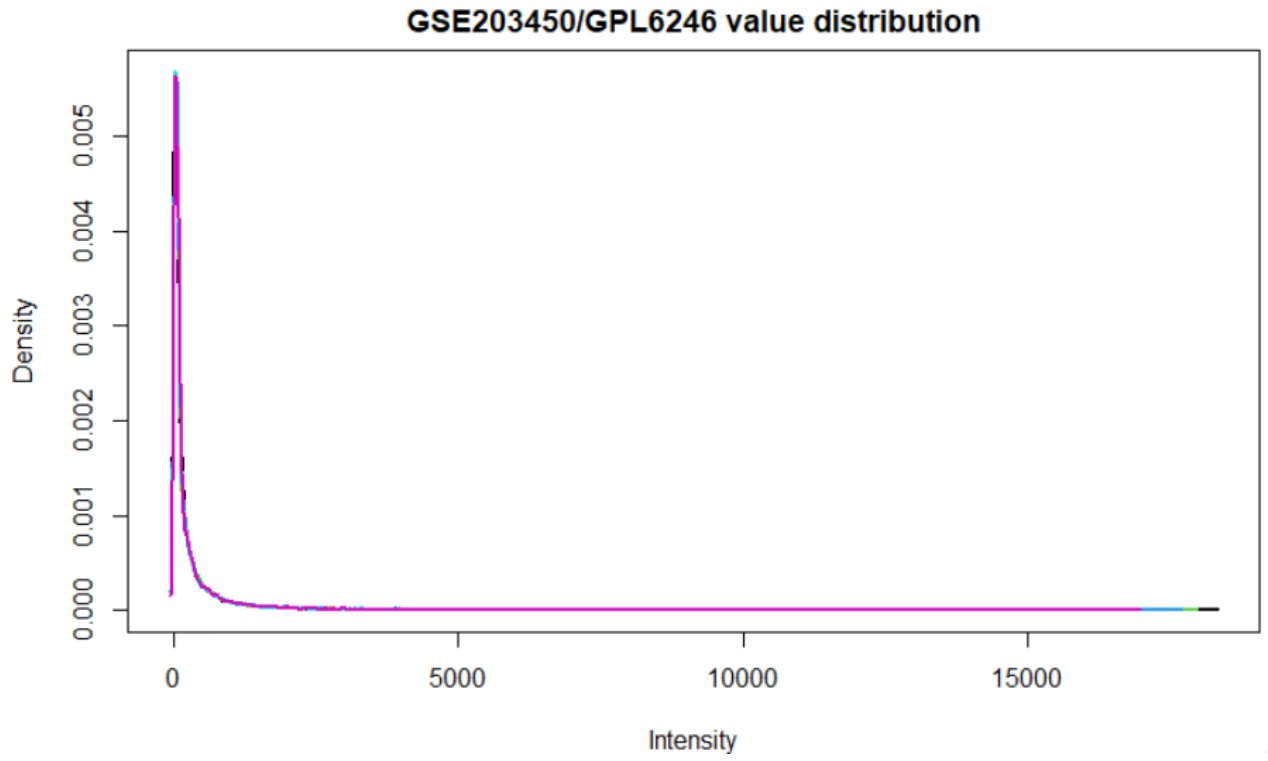
```
# Log2 transformation.-------------------------------------------------------
exprs(gse) <- log2(ex)        # Log2 transform.

# Box-and-whisker plot (after log2 transformation).--------------------------
par(mar=c(7,4,2,1))
title <- paste ("GSE203450", "/", annotation(gse), sep ="")

# pdf("Microarray_data_after_transformation.pdf")
boxplot(exprs(gse), boxwex=0.7, notch=T, main=title, outline=FALSE, las=2)
# dev.off()

# Expression value distribution plot.----------------------------------------
par(mar=c(4,4,2,1))
title <- paste ("GSE203450", "/", annotation(gse), " value distribution",
                sep ="")

# pdf("Densities_after_normalization.pdf")
plotDensities(ex, main=title, legend=F)
# dev.off()
```

BoxPlot – after transformation



Trasformando i nostri dati non solo normalizziamo le osservazioni, ma anche i residui. La normalizzazione rende i modelli di training meno sensibili alla scala delle caratteristiche, quindi possiamo risolvere meglio i coefficienti.

Densità – Distribuzione

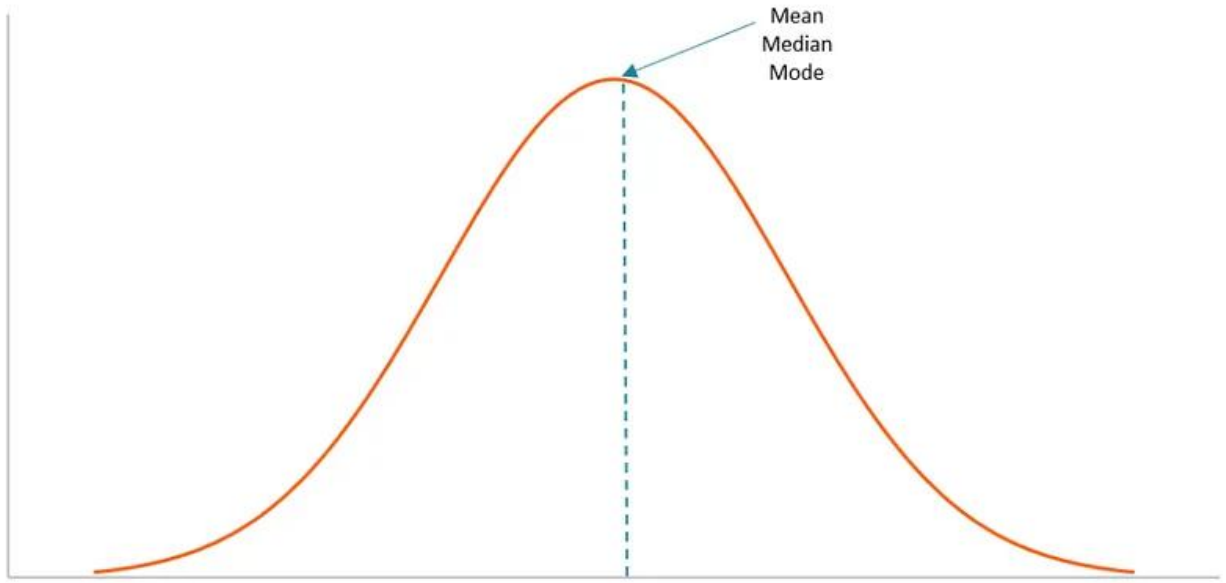


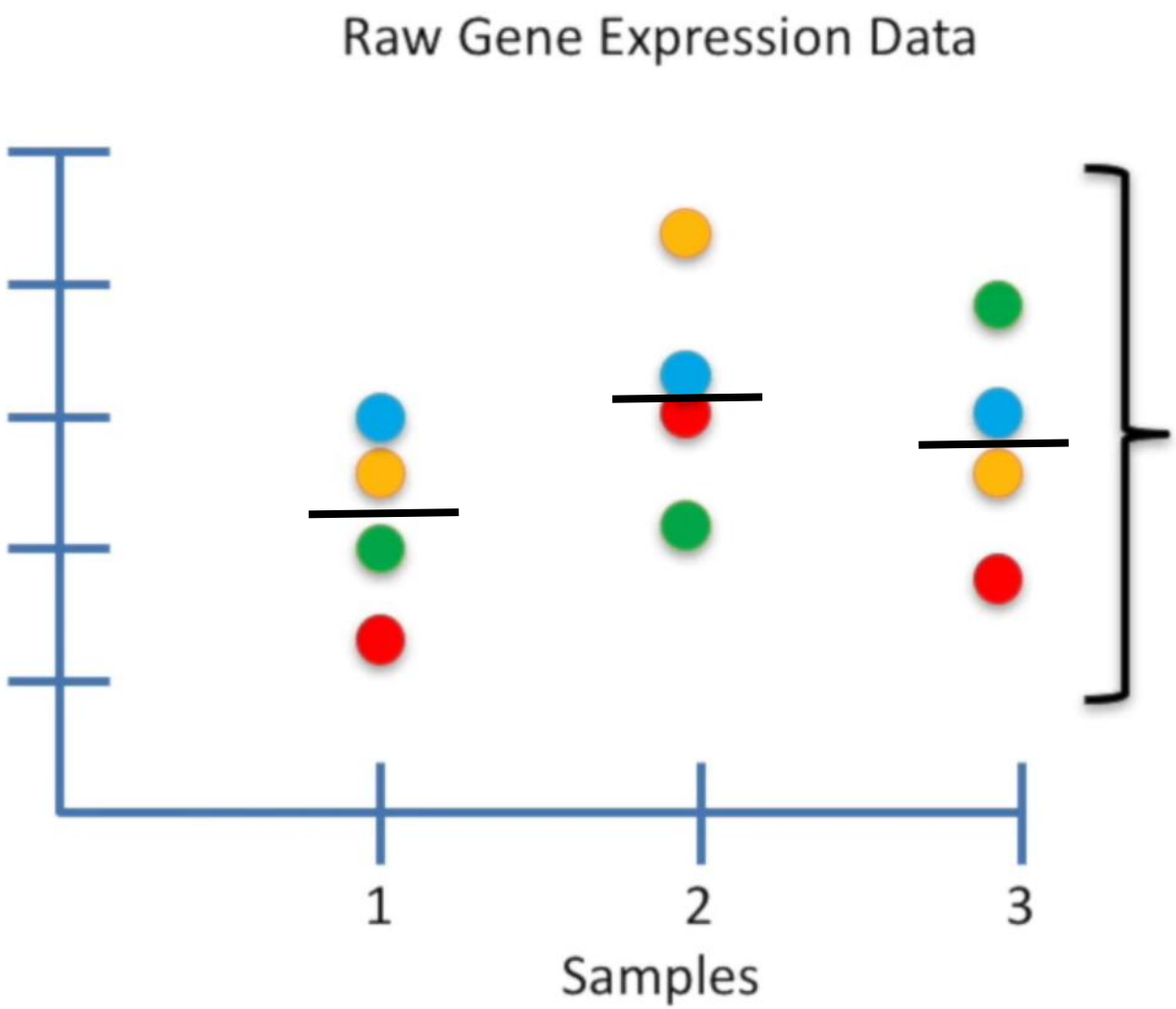
Il diagramma della densità è una visualizzazione che rappresenta la distribuzione di probabilità stimata dei valori di espressione genica per ciascun campione. In altre parole:

- Visualizza la frequenza relativa dei valori di espressione lungo una scala continua.
- Ogni riga del grafico corrisponde ad un campione (colonna della matrice ex).
- Il grafico della densità è un'alternativa fluida all'istogramma.

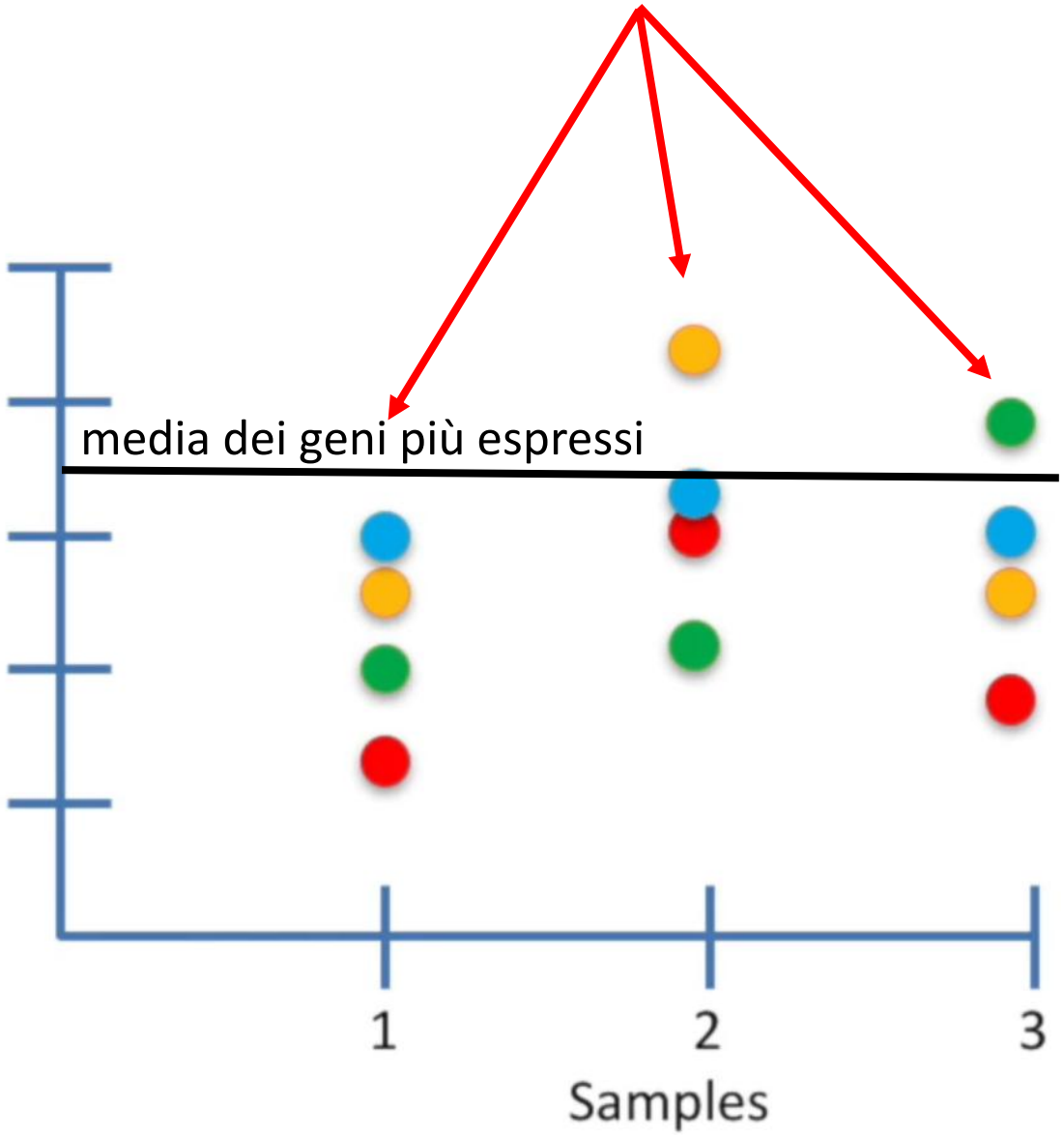
Normalizzazione - Quantile

- L'obiettivo generale della Normalizzazione dei nostri dati è creare una distribuzione più normale (*__Gaussiana__*), ovvero una curva a campana. **In generale, le distribuzioni normali tendono a produrre risultati migliori** in un modello perché ci sono osservazioni pressoché uguali sopra e sotto la media e la media e la mediana sono le stesse.
- I modelli vengono eseguiti presupponendo che i dati siano distribuiti normalmente.

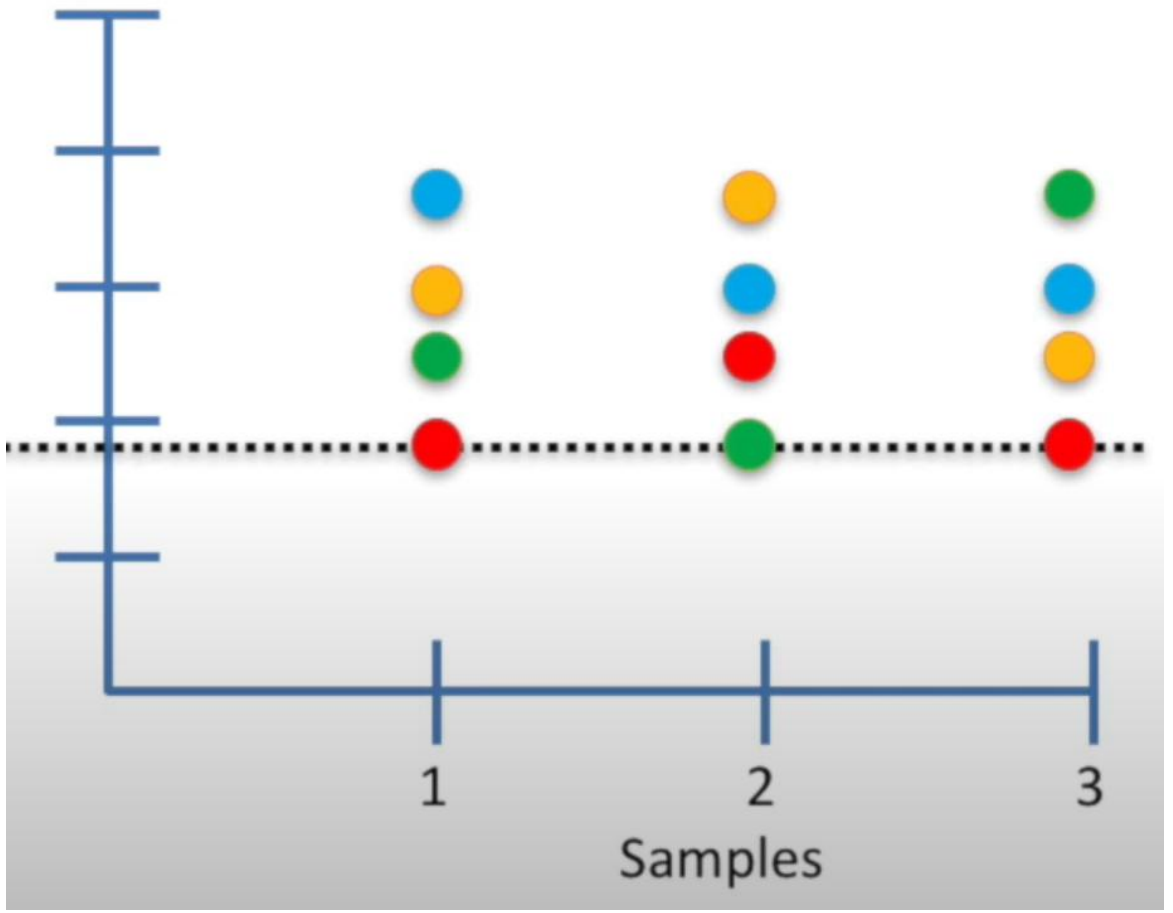




- Ogni campione ha un valore medio diverso, suggerendo che dobbiamo **compensare le diverse intensità complessive della luce.**
- **Normalizzazione quantile (Quantile Normalization)** può correggere questo artefatto tecnico.



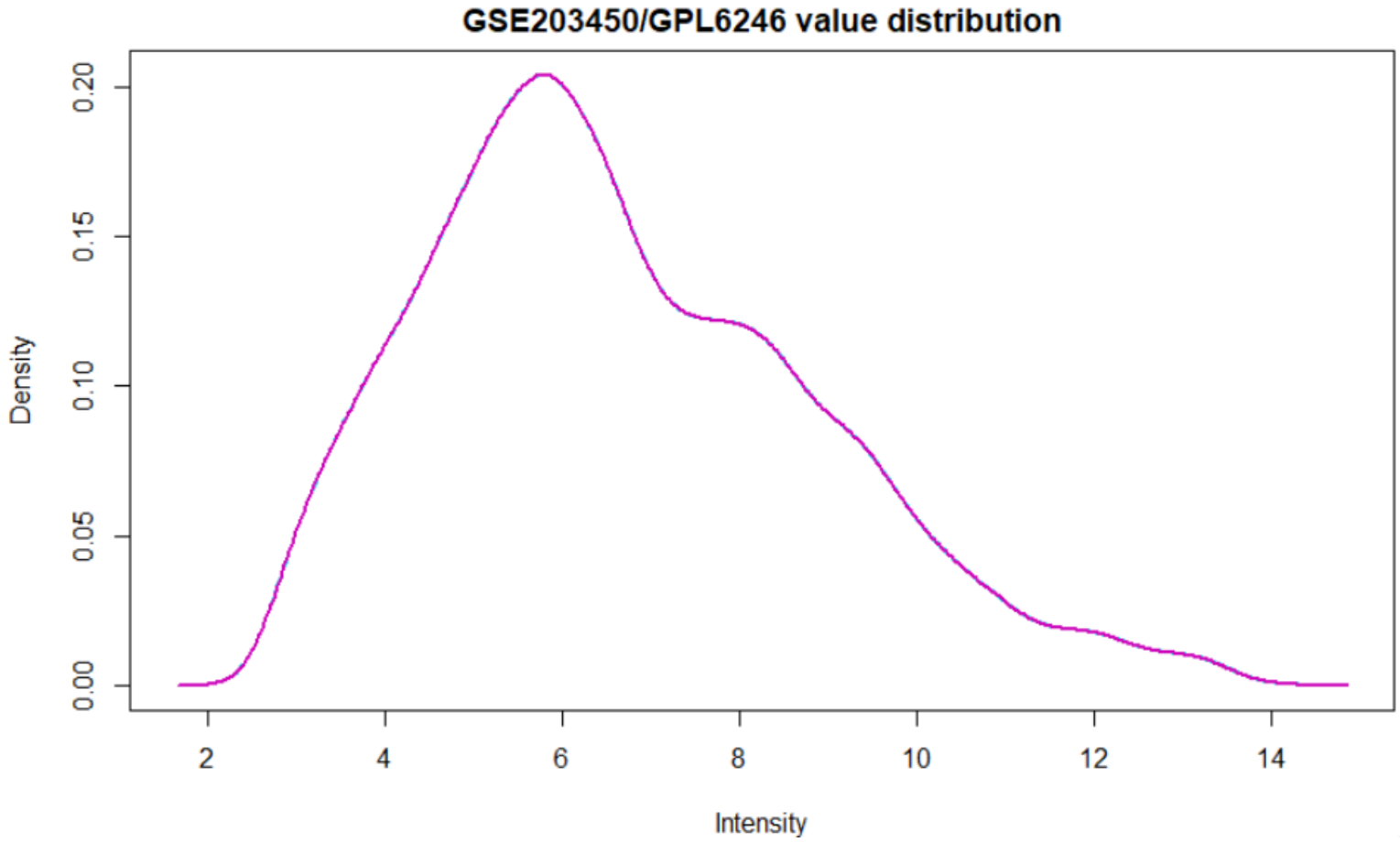
Quantile Normalized Data –
Original gene orders preserved



Part 6. Normalization log-ratios with limma

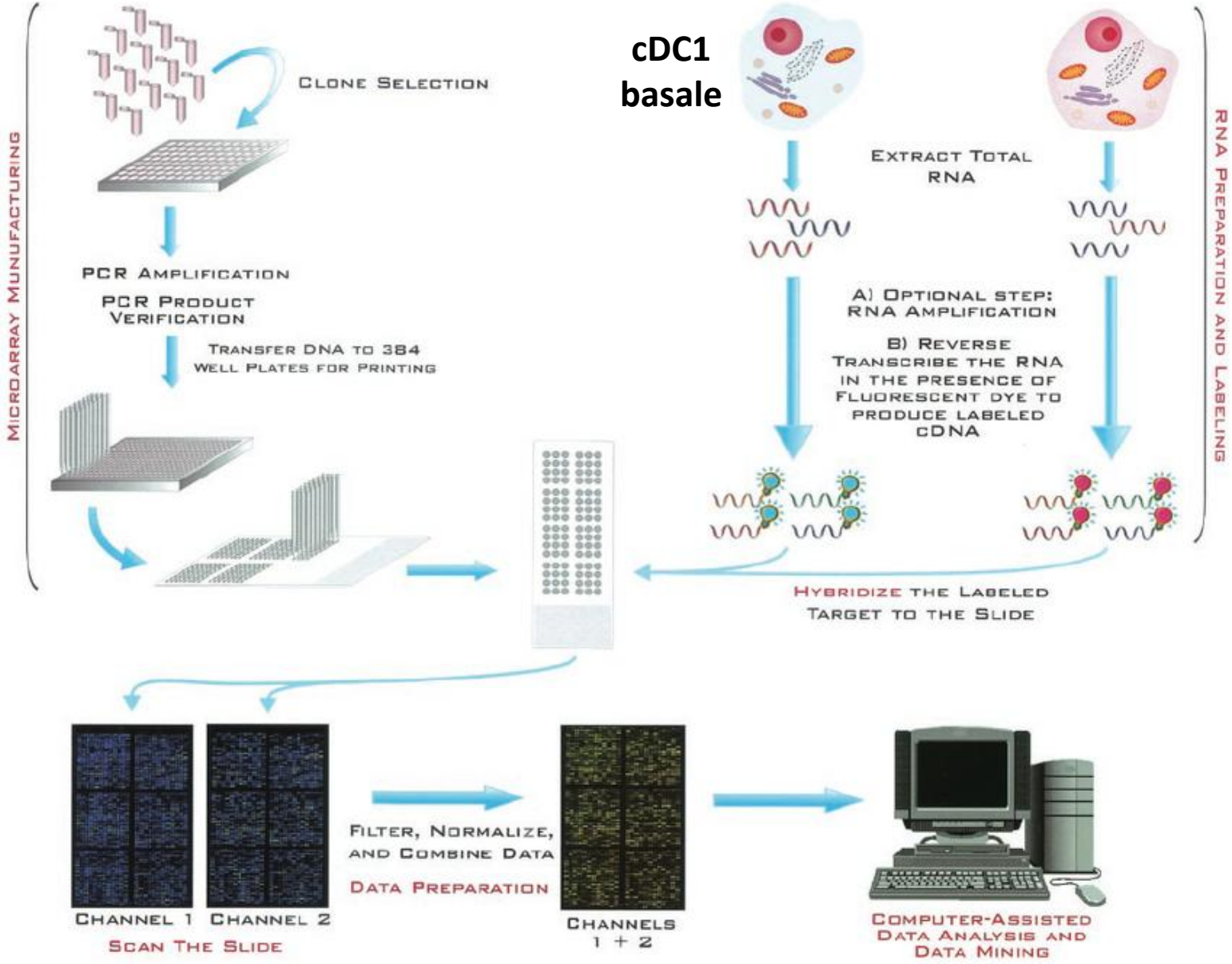
```
# Part 6. Normalization log-ratios with Limma.=================================
print("Normalizing data with Limma...")
exprs(gse) <- normalizeBetweenArrays(exprs(gse))        # normalize data
gse

plotDensities(gse, main=title, legend=F)
```

Densità – Distribuzione



Next lesson...

- PCA / UMAP
- Differential Expressed Genes (DEGs)
- Volcano Plot
- Enrichment Analysis
  - Gene Ontology
  - GSEA



cDC1 + LPS

E. coli LPS (0.250 µg/ml) for 24 h

[Application of Microarrays to the Analysis of Gene Expression in Cancer]

Riepilogo – GEO2R Pipeline

GEO Datasets

Gene Expression Omnibus
https://www.ncbi.nlm.nih.gov/geo/

- **getGEO()**

I dati GEO hanno quattro tipi di entità tra cui **GEO Platform** (**GPL**), **GEO Sample** (**GSM**), **GEO Series** (**GSE**) e curated **GEO DataSet** (**GDS**).

Probe ID
Gene/transcript

Sample

```
>     exprs(gse)                          # print
         GSM6172191 GSM6172192 GSM6172193 GS
10344614    109.847     88.484    107.828
10344616      6.427      6.455      7.069
10344618      9.538      8.063      8.625
10344620     27.262     26.002     25.437
10344622    169.643    146.881    143.827
10344624    266.608    239.226    250.661
10344633    751.896    539.310    677.006
10344637    285.090    367.982    303.612
10344653     18.618     15.632     18.166
10344658    236.535    285.525    254.945
```

raw expression

Riepilogo – GEO2R Pipeline

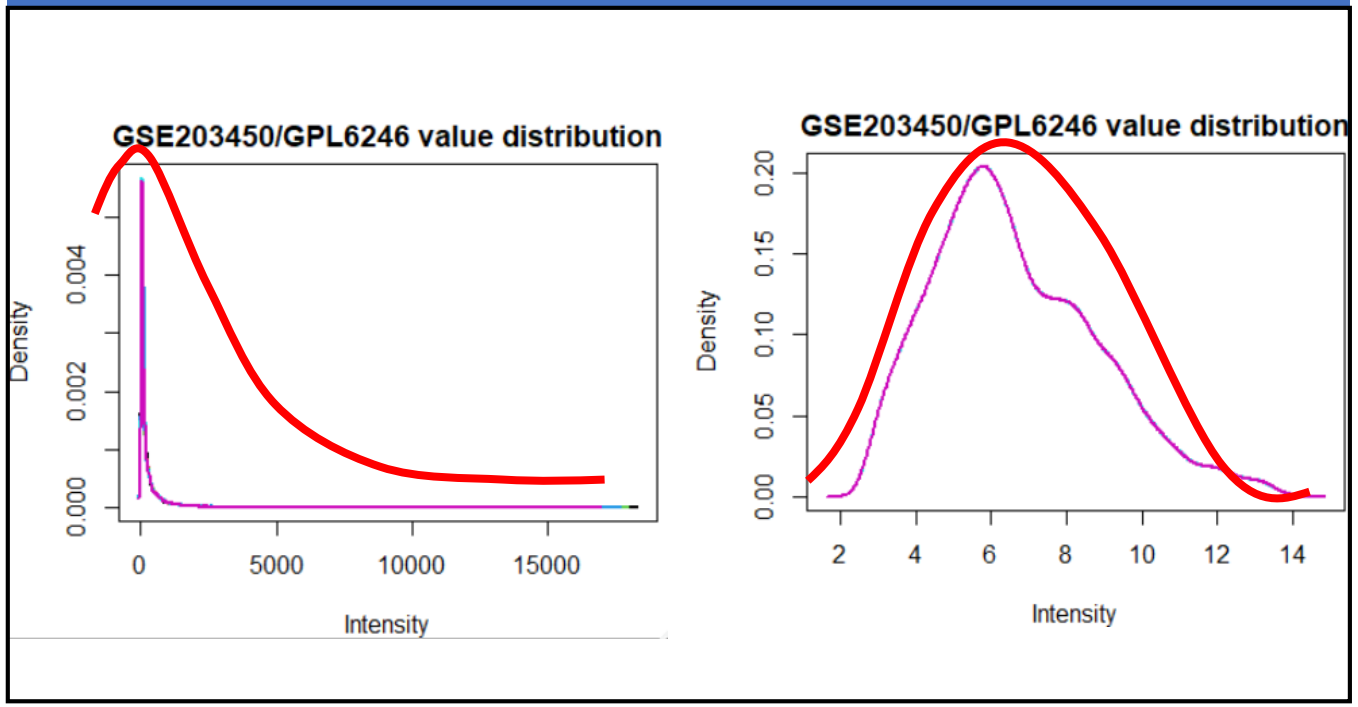
Matrice di Espressione

Campioni/samples

Trascritti/features

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Log transf/Normalizzazione Quantile



Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr;4(2):249-64.

Part 7. Assign samples to groups and set up design matrix

```
# Part 7. Assign samples to groups and set up design matrix.====================

# Define groups to compare.-----------------------------------------------------
groups <- make.names(c("cDC1_LPS","cDC1_ctrl"))

gs <- factor(sml)
levels(gs) <- groups
gse$group <- gs
design <- model.matrix(~group + 0, gse)
colnames(design) <- levels(gs)

gse <- gse[complete.cases(exprs(gse)), ] # skip missing values
```

Converte sml in un fattore categorico

Reimposta i nomi delle categorie nel fattore (gs)

Memorizza le informazioni sul gruppo sperimentale direttamente nell'oggetto gse

Questa parte del codice è una preparazione per l'analisi statistica differenziale con il pacchetto Limma. **Imposta i gruppi sperimentali**, associa i campioni a questi gruppi e crea una matrice di progettazione, essenziale per la modellazione lineare utilizzata nell'analisi.

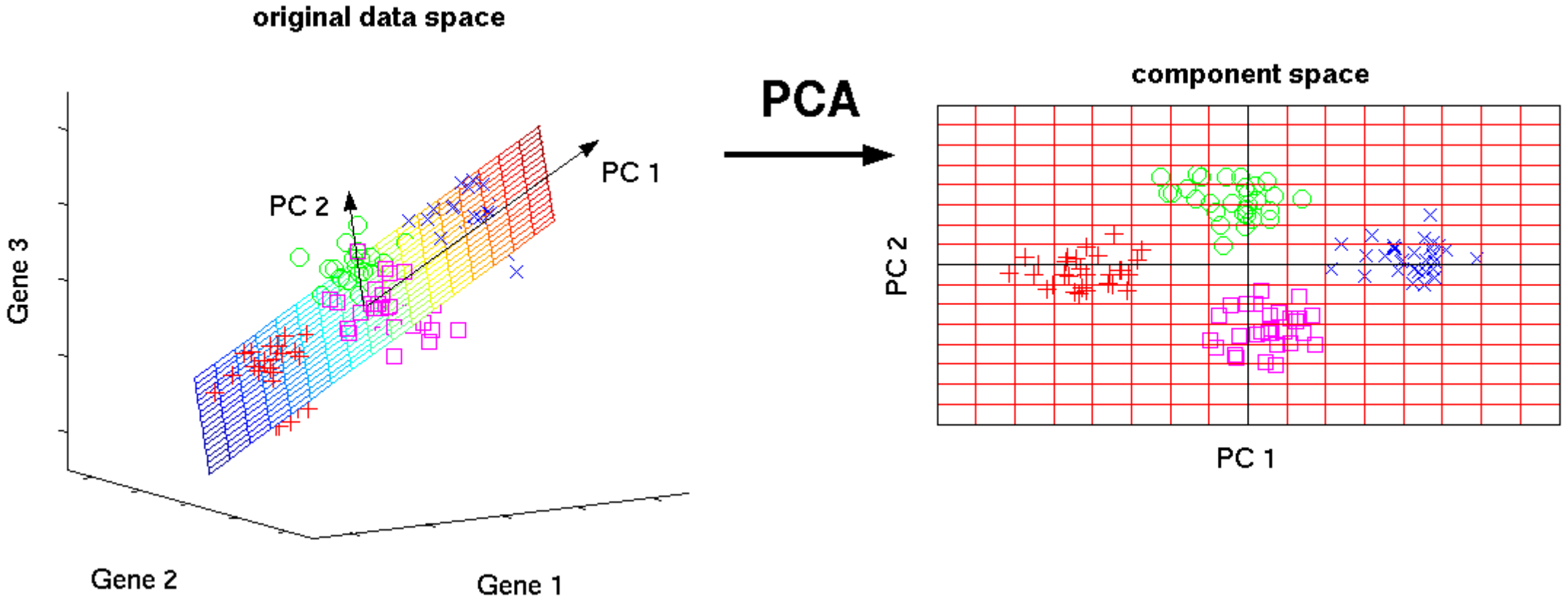
Modello Matrice Design

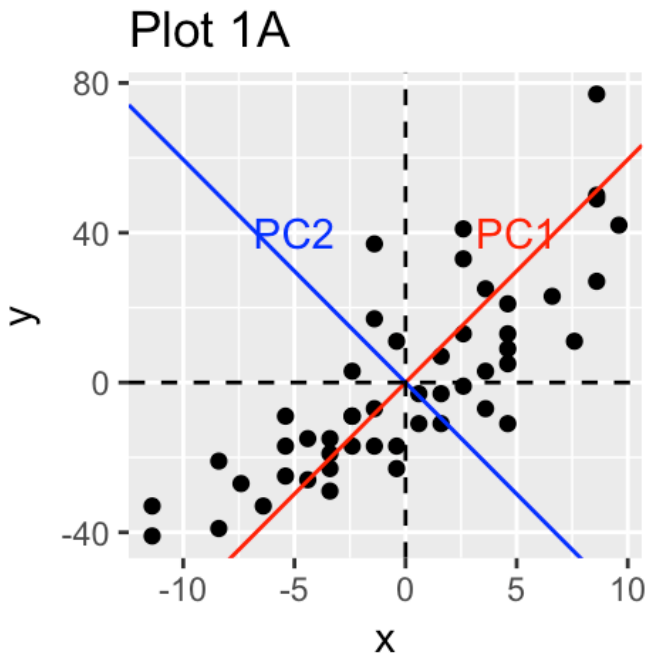
cDC1_LPS	cDC1_ctrl
1	0
1	0
1	0
0	1
0	1
0	1

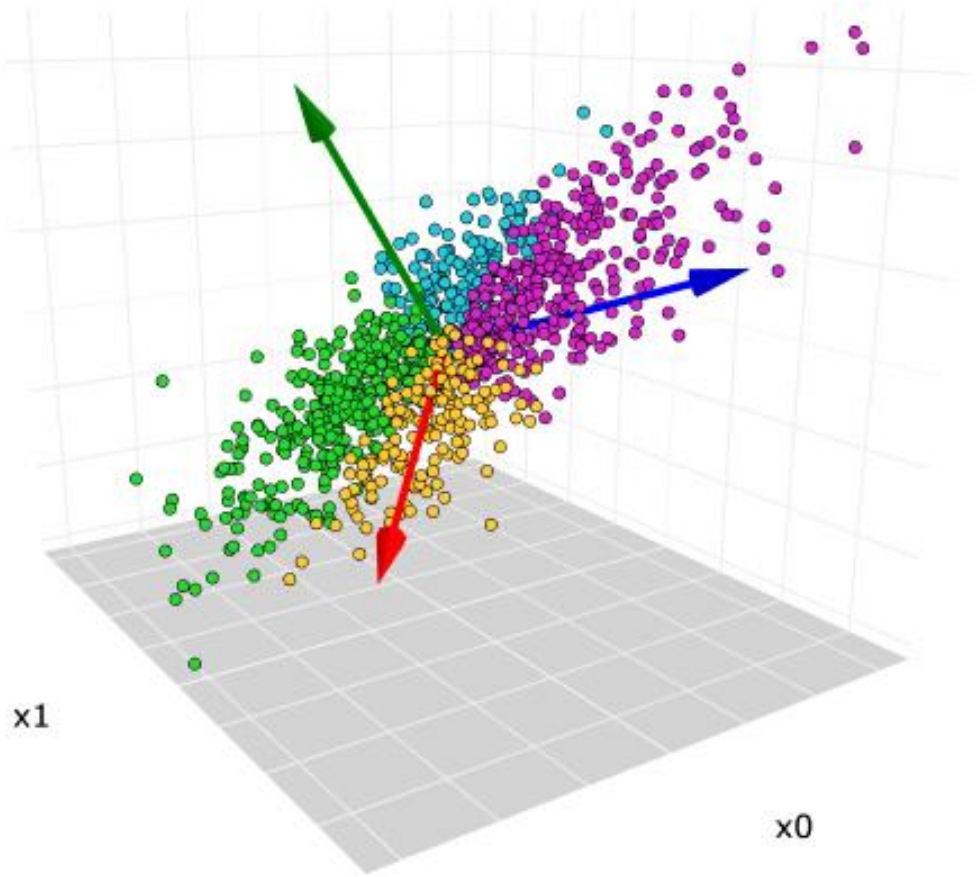
Part 7. Assign samples to groups and set up design matrix

```
# UMAP plot (dimensionality reduction).----------------------------------------
# library("maptools")  # point labels without overlaps
library("umap")
library("car")

ex <- na.omit(ex) # eliminate rows with NAs
ex <- ex[!duplicated(ex), ]  # remove duplicates

ump <- umap(t(ex), n_neighbors = 3, random_state = 123)
par(mar=c(3,3,2,6), xpd=TRUE)
plot(ump$layout, main="UMAP plot, nbrs=3", xlab="", ylab="", col=gs, pch=20, cex=1.5)
legend("topright", inset=c(-0.15,0), legend=levels(gs), pch=20,
       col=1:nlevels(gs), title="Group", pt.cex=1.5)
pointLabel(ump$layout, labels = rownames(ump$layout), method="SANN", cex=0.6)
```

PCA (Principal Component Analysis) è una tecnica di riduzione dimensionale che trasforma dati complessi in un insieme di variabili principali (componenti) per catturare quanta più variazione possibile e facilitare la visualizzazione e l'analisi.

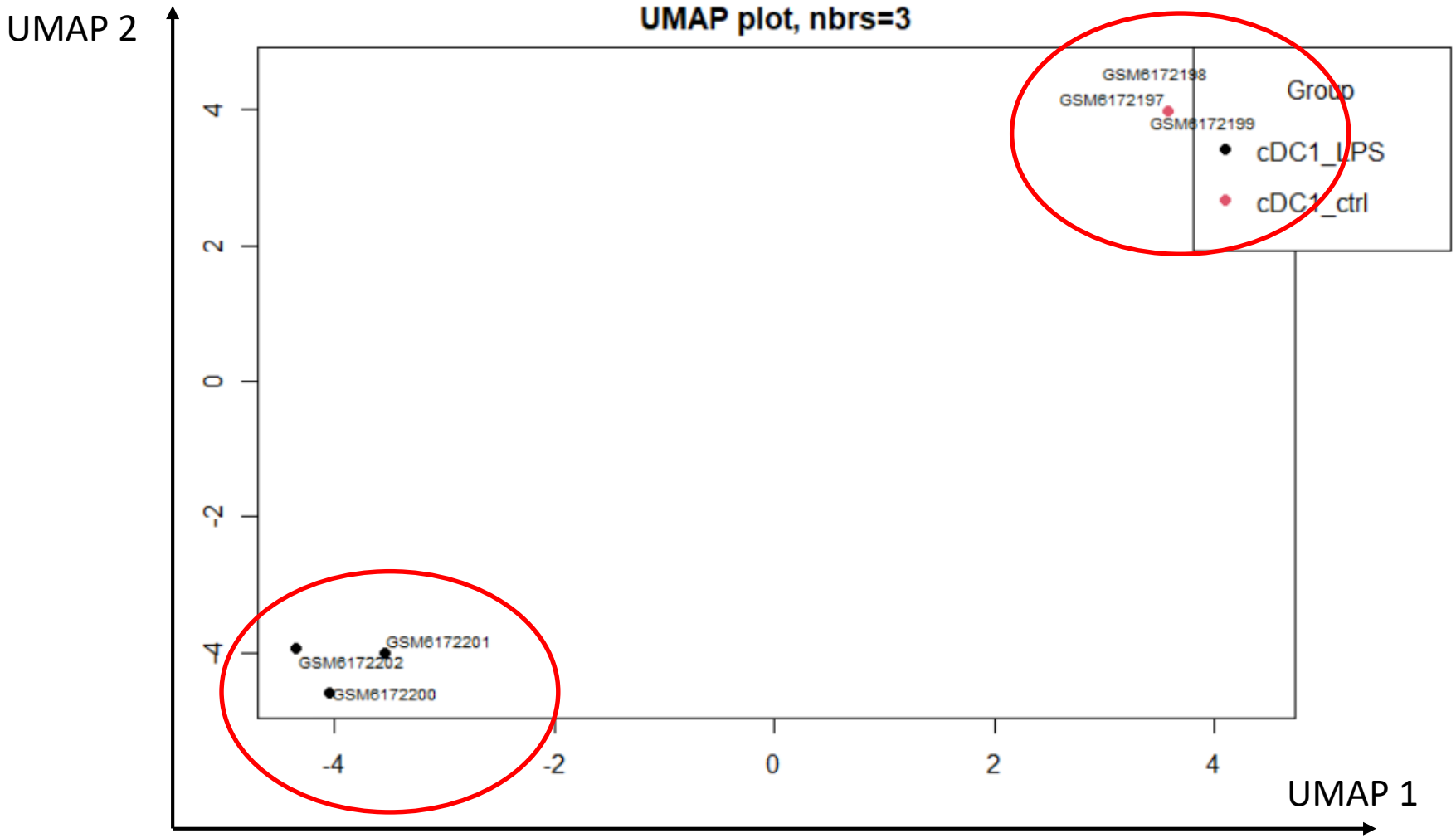






UMAP (**Uniform Manifold Approximation and Projection**) è una tecnica di riduzione della dimensionalità che preserva sia la struttura locale che globale dei dati in spazi ad alta dimensionalità. È ampiamente utilizzato per la visualizzazione 2D o 3D, soprattutto nell'analisi di dati complessi come la genomica e l'apprendimento automatico, evidenziando cluster e modelli nei dati.

UMAP plot - riduzione della dimensionalità



Part 8. Differential Expression Genes

Precision Weight Calculation and Plotting the Mean-Variance Trend

Un oggetto contenente i dati sull'espressione genica.

La matrice di design per l'esperimento

```
# Part 8. Differential expression with Limma-Voom.==============================
print("Calculating the differential gene expression...")

# Calculate precision weights and show plot of mean-variance trend.-------------
v <- vooma(gse, design, plot=T)

v$genes <- fData(gse) # attach gene annotations
```

Fit the Linear Model

```
# Fit linear model.-----------------------------------------------------------
fit  <- lmFit(v)
```

Questa funzione adatta un modello lineare ai dati (dopo la modellazione della varianza con vooma)

Part 8. Differential Expression Genes

Set Up Contrasts of Interest

Crea un contrasto, un confronto tra i due gruppi nell'esperimento. Il contrasto è scritto in un modo che sottrae il secondo gruppo dal primo (ad esempio, "gruppo1 - gruppo2").

```
# Set up contrasts of interest and recalculate model coefficients.--------------
cts <- c(paste(groups[1],"-",groups[2],sep=""))
cont.matrix <- makeContrasts(contrasts=cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
```

Questa funzione crea una matrice di contrasto dalle definizioni di contrasto (cts) basate sulla matrice di progettazione (design).

Questa funzione applica la matrice di contrasto (cont.matrix) al modello lineare precedentemente adattato (fit).

Compute Statistics and Table of Top Significant Genes

Questa funzione applica la moderazione empirica di Bayes agli errori standard del modello lineare. Questo passaggio stabilizza le stime della varianza e migliora la potenza statistica dei test di ipotesi.

```
# Compute statistics and table of top significant genes.-----------------------
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
```

Bayesian statistics

Questa funzione estrae i geni maggiormente espressi in modo differenziale in base al modello adattato.

Inferenza Statistica

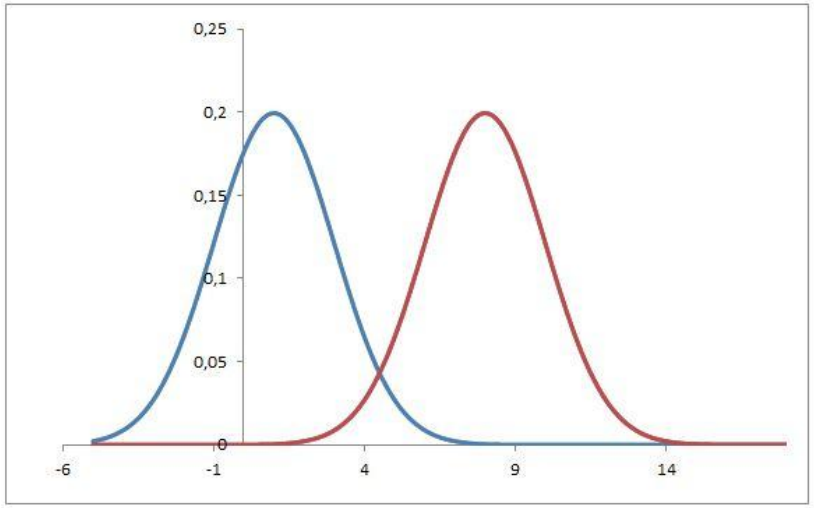
- Spesso capita di voler confermare i valori di una variabile in due popolazioni. In questo caso, assumendo che la variabile sia numerica continua, la domanda è: **la variabile ha una media significativamente diversa nelle due popolazioni?**
- **Verifica d'ipotesi**
  - L'ipotesi Nulla (H0)
  - L'ipotesi Alternativa (H1) -> La negazione dell'ipotesi Nulla

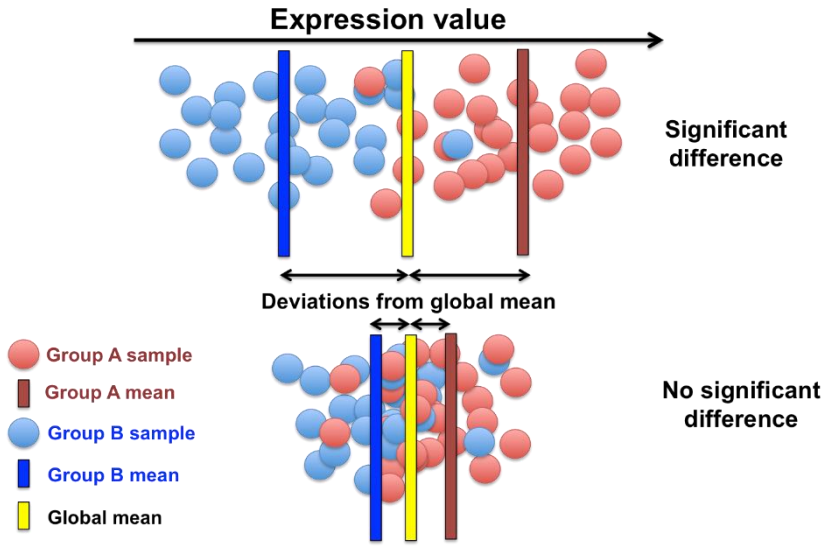
$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 \leq \mu_2$$

Livello di significatività (*p-value*)

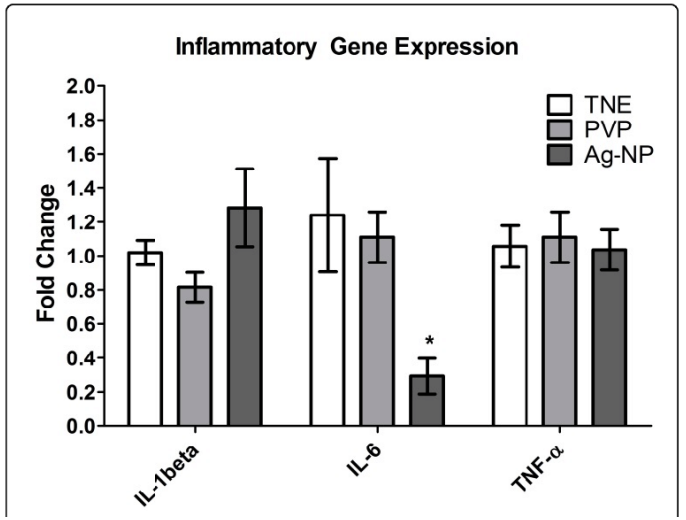
- Per avere maggiori informazioni sull'effettiva probabilità di osservare un certo valore del test nella distribuzione sotto H0, viene solitamente riportato il *p-value*.
- Il *p-value*, quindi, è una probabilità con valori che vanno da 0 a 1: valori picolli del *p-value* portano a rifiutare H0.





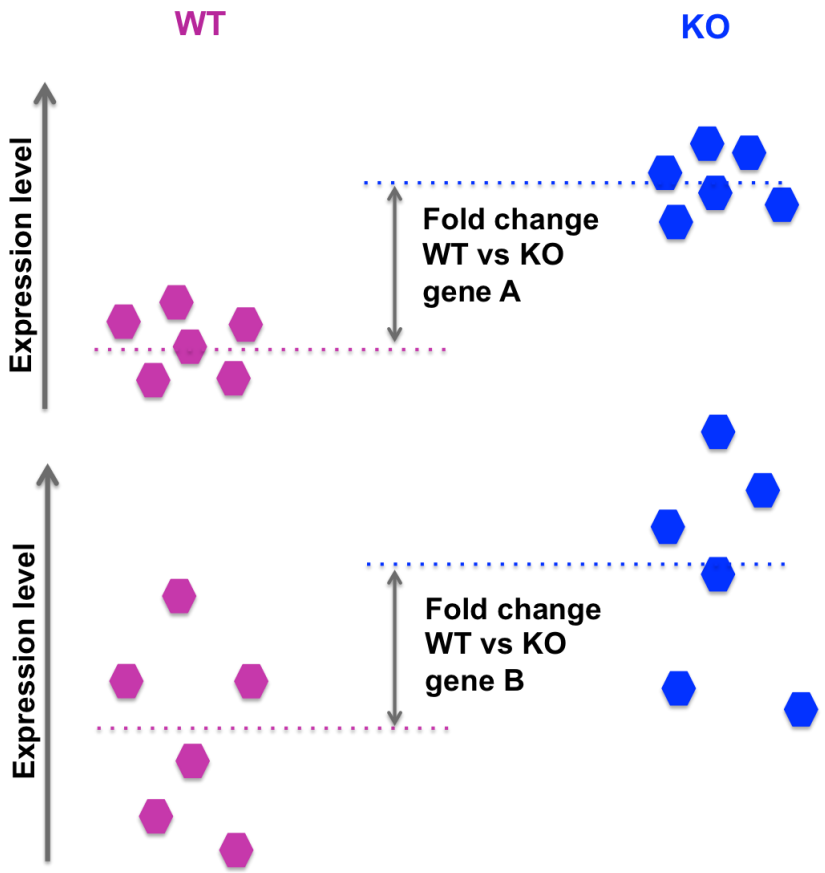
Fold Change

- Il fold change è un modo semplice per descrivere quanto un valore ***cambia*** rispetto a un riferimento.
  - Se il valore raddoppia rispetto al riferimento, il fold change è 2.
  - Se il valore si dimezza, il fold change è 0,5.
  - Un fold change di 1 significa che non c'è stato alcun cambiamento.



Log-Fold Change (*logFC*)

- Il fold change è spesso trasformato in scala logaritmica, come il log2 fold change:
  - Log2(Fold Change) = 1 → il gene ha **raddoppiato** l'espressione.
  - Log2(Fold Change) = -1 → l'espressione si è **dimezzata**.
  - Log2(Fold Change) = 0 → **nessun cambiamento** nell'espressione.



Part 9. Summarize test results

Adjusted P-Values and Histogram, Summarize Test Results

p-value = significatività statistica
Fold-change = entità del cambiamento

```
# Part 9. Visualize adj p-values, Venn diagram and QQ plot.=====================
print("Visualizing adj p-values and venn diagram...")

# Build histogram of P-values for all genes. Normal test.----------------------
# assumption is that most genes are not differentially expressed.
tT2 <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
hist(tT2$adj.P.Val, col = "grey", border = "white", xlab = "P-adj",
     ylab = "Number of genes", main = "P-adj value distribution")

# Summarize test results as "up", "down" or "not expressed".--------------------
dT <- decideTests(fit2, adjust.method="fdr", p.value=0.05, lfc=0.5)
```

Classifica i geni in categorie in base alle soglie statistiche

LogFC threshold

L'output, dT, è una matrice in cui ogni elemento indica se un gene è:

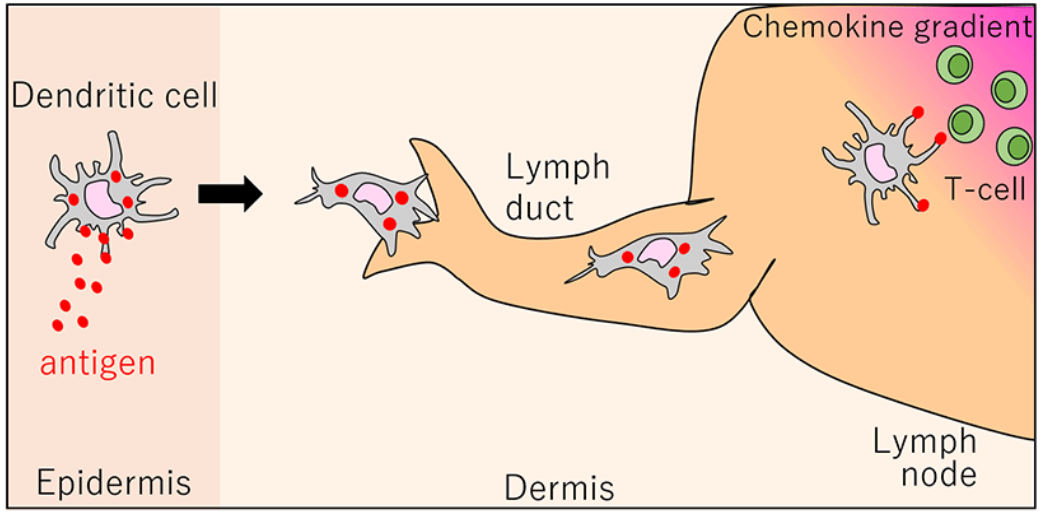
- 1: Upregulated.
- -1: Downregulated.
- 0: Not significantly differentially expressed.

Part 10. DEGs List

Creating the DEGs List and Annotating Differential Expression Status

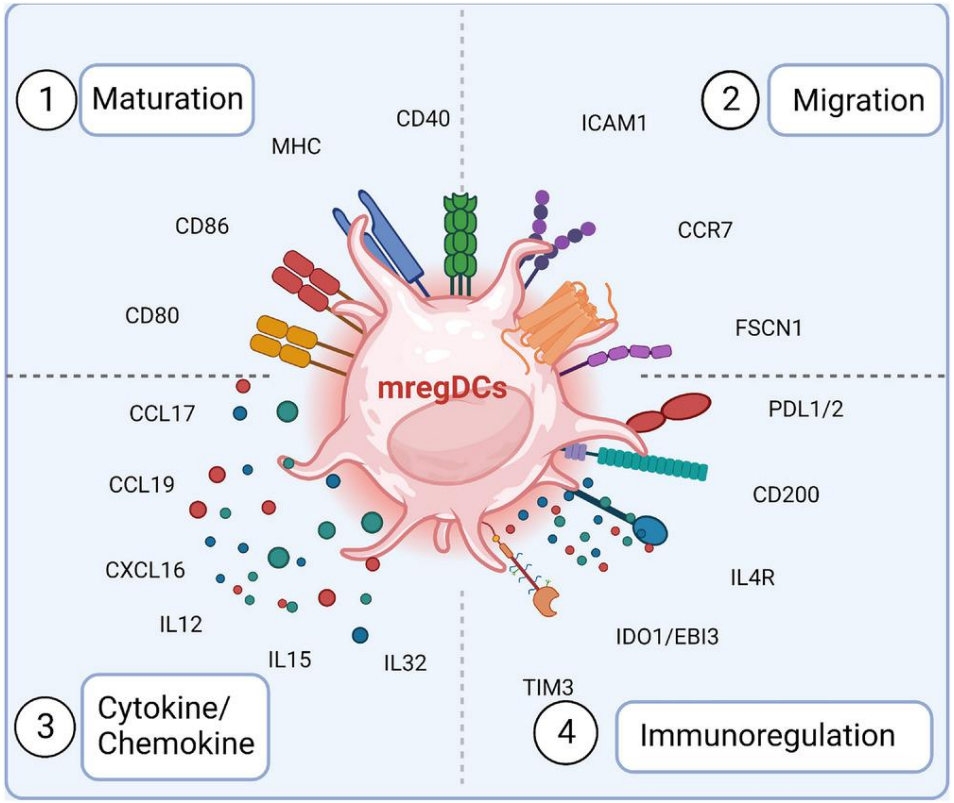
```
# Part 10. DEGs list.=============================================================
print("Creating a DEGs list...")
library(tidyverse)
DEG <- tT2[c(3,1,22,26)] %>%
  setNames(c("gene","id","logFC","padj"))

# DEGs selection (Define groups and cut-off points of logFC and padj).----------
DEG$Enriched <- "NS"
DEG$Enriched[DEG$logFC > 0.5 & DEG$padj < 0.05] <- "cDC1_LPS"
DEG$Enriched[DEG$logFC < -0.5 & DEG$padj < 0.05] <- "cDC1_ctrl"
#
```

- Later, around 4 h after LPS activation, DCs show recovery of migratory ability and start to progressively lose their antigen uptake function until the mature stage in which they show poor antigen uptake and migratory activity (Granucci et al., 1999).



https://bsw3.naist.jp/eng/bsedge/0011.html

Granucci F, Ferrero E, Foti M, Aggujaro D, Vettoretto K, Ricciardi-Castagnoli P. Early events in dendritic cell maturation induced by LPS. Microbes Infect. 1999 Nov;1(13):1079-84.

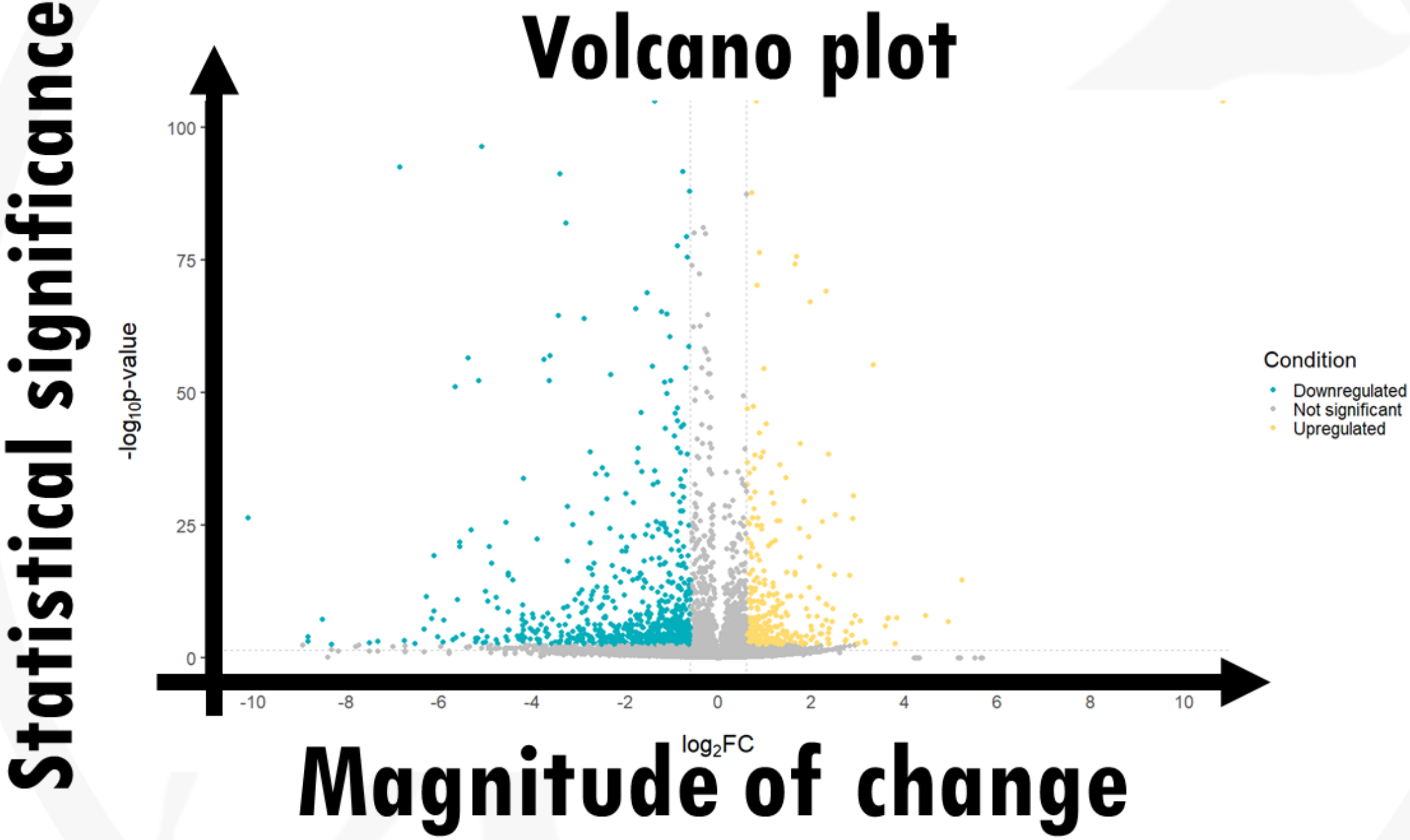


Clinical & Translational Med, Volume: 13, Issue: 2, First published: 17 February 2023, DOI: (10.1002/ctm2.1199)

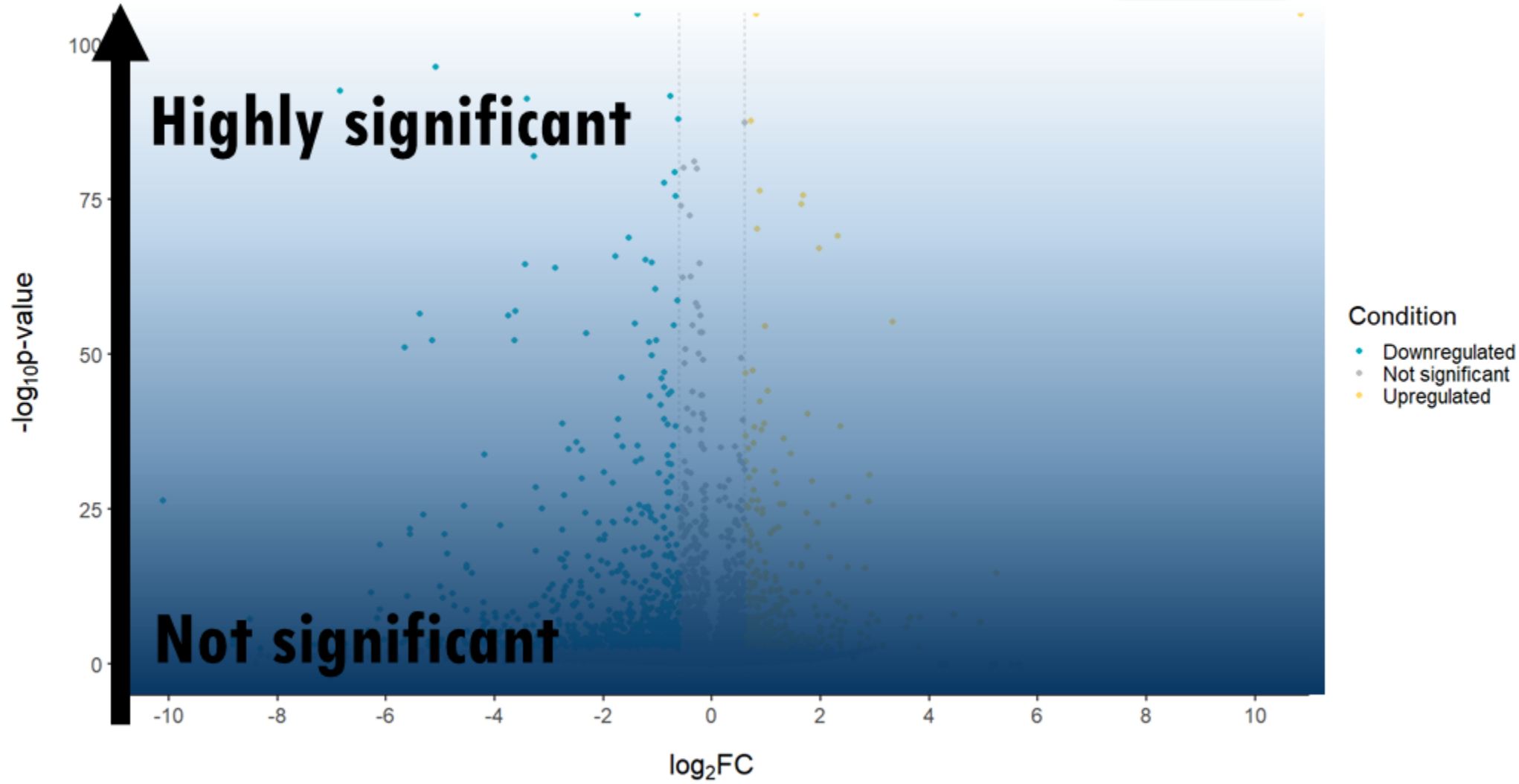
Part 11. Volcano plot Representation (logFC =|0.5|)

```
# a) Select genes to highlight.================================================

genes <- DEG %>%
  filter(gene %in% c("Il6","Il1b","Tnf","Cd86","Cd80","Clec9a","Xcr1"))

# Visualization Volcano Plot.-----------------------------------------------
ggplot(data = DEG,
       aes(x = logFC,
           y = -log10(padj))) +
  geom_point(aes(colour = Enriched),
             alpha = 0.5,
             shape = 16,
             size = 1) +
  geom_point(data = genes,
             shape = 21,
             size = 2,
             fill = "black",
             colour = "black") +
  theme_classic() +
  geom_hline(yintercept = -log10(0.05),
             linetype = "dashed") +
  geom_vline(xintercept = c(log2(0.7071), log2(1.4142)),
             linetype = "dashed") +
  geom_label_repel(data = genes, # Add labels last to appear as the top layer
                   max.overlaps = Inf,
                   aes(label = gene),fontface = 'italic',
                   force = 1, nudge_y = 0.5) +
  scale_colour_manual(values = cols) +
  # scale_x_continuous(breaks = c(seq(-4, 5, 2)),
  #                    limits = c(-4, 5),)  +
  ggtitle("BMDCs/cDC1-ctrl versus cDC1-LPS enriched genes")
```

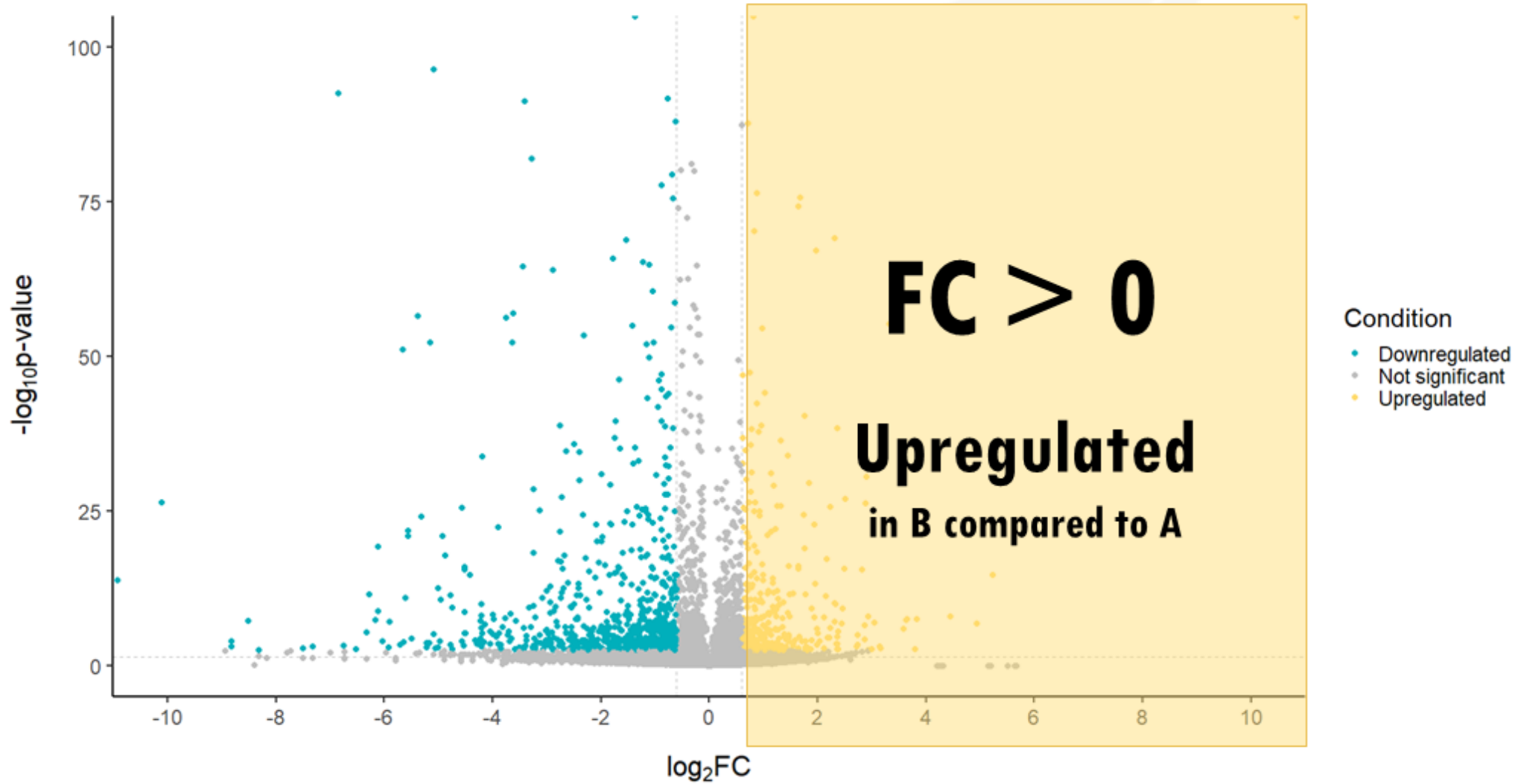
Volcano plot



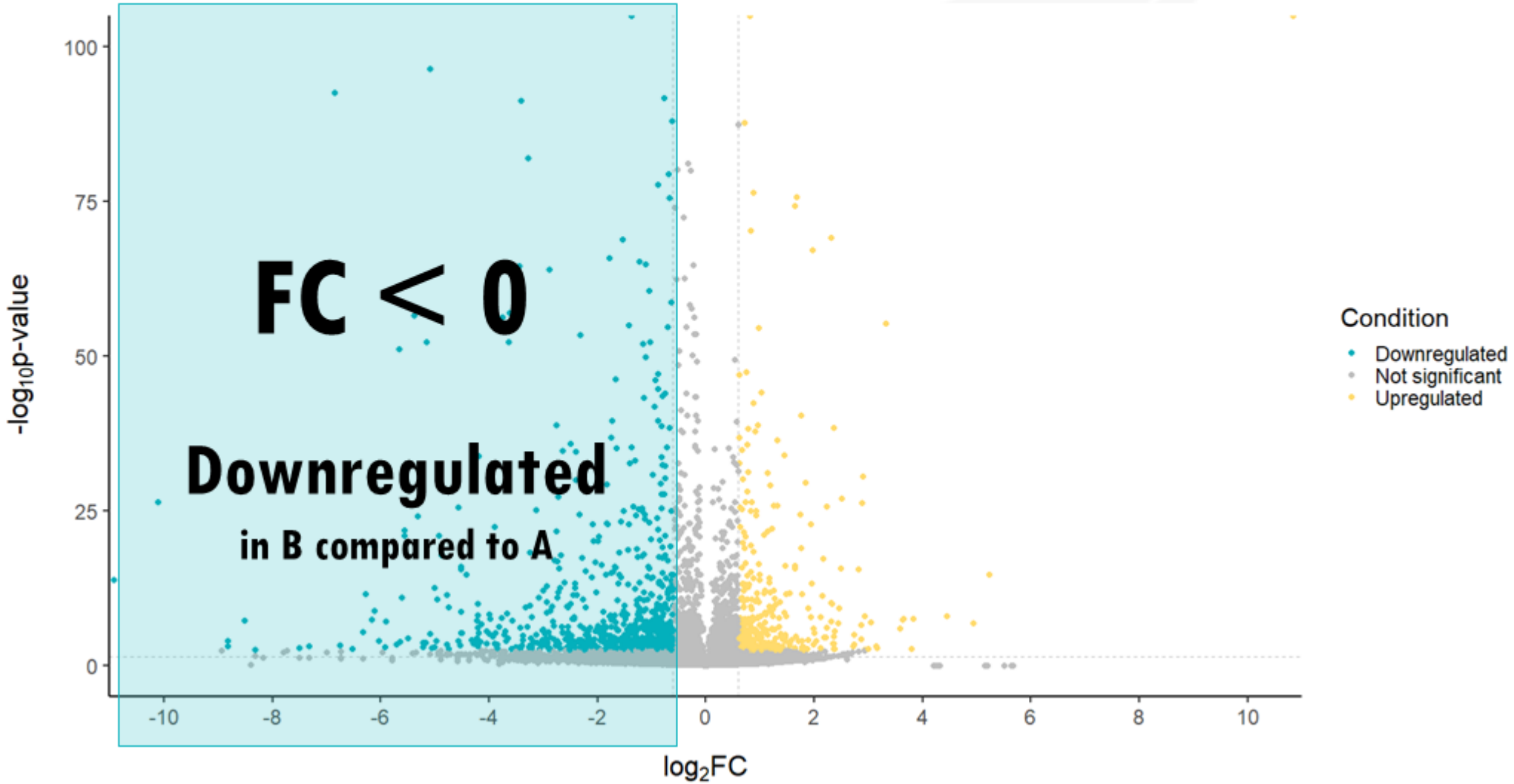
Statistical significance: p-value



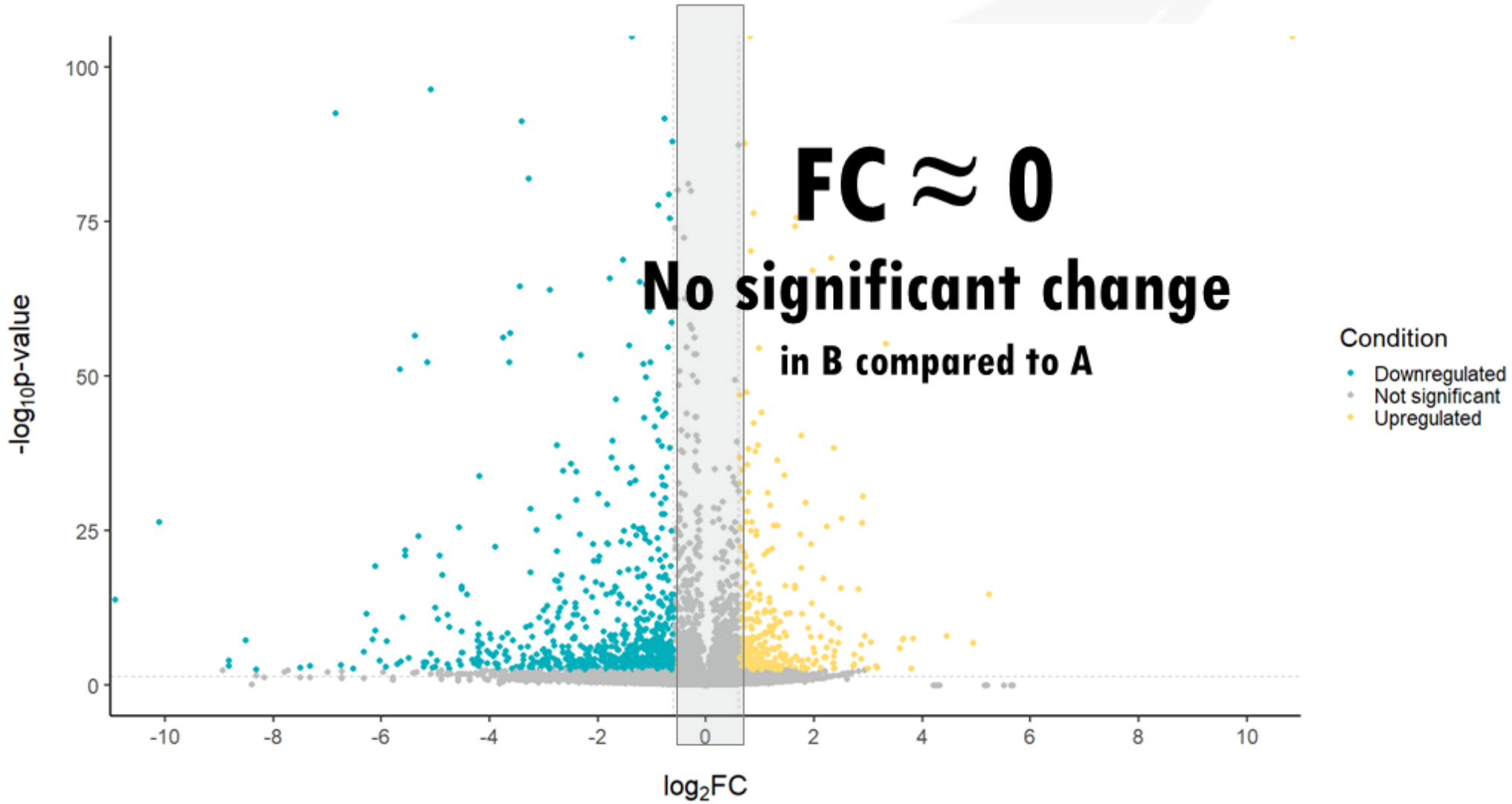
$$FC = \frac{\text{expression of gene X in B}}{\text{expression of gene X in A}}$$



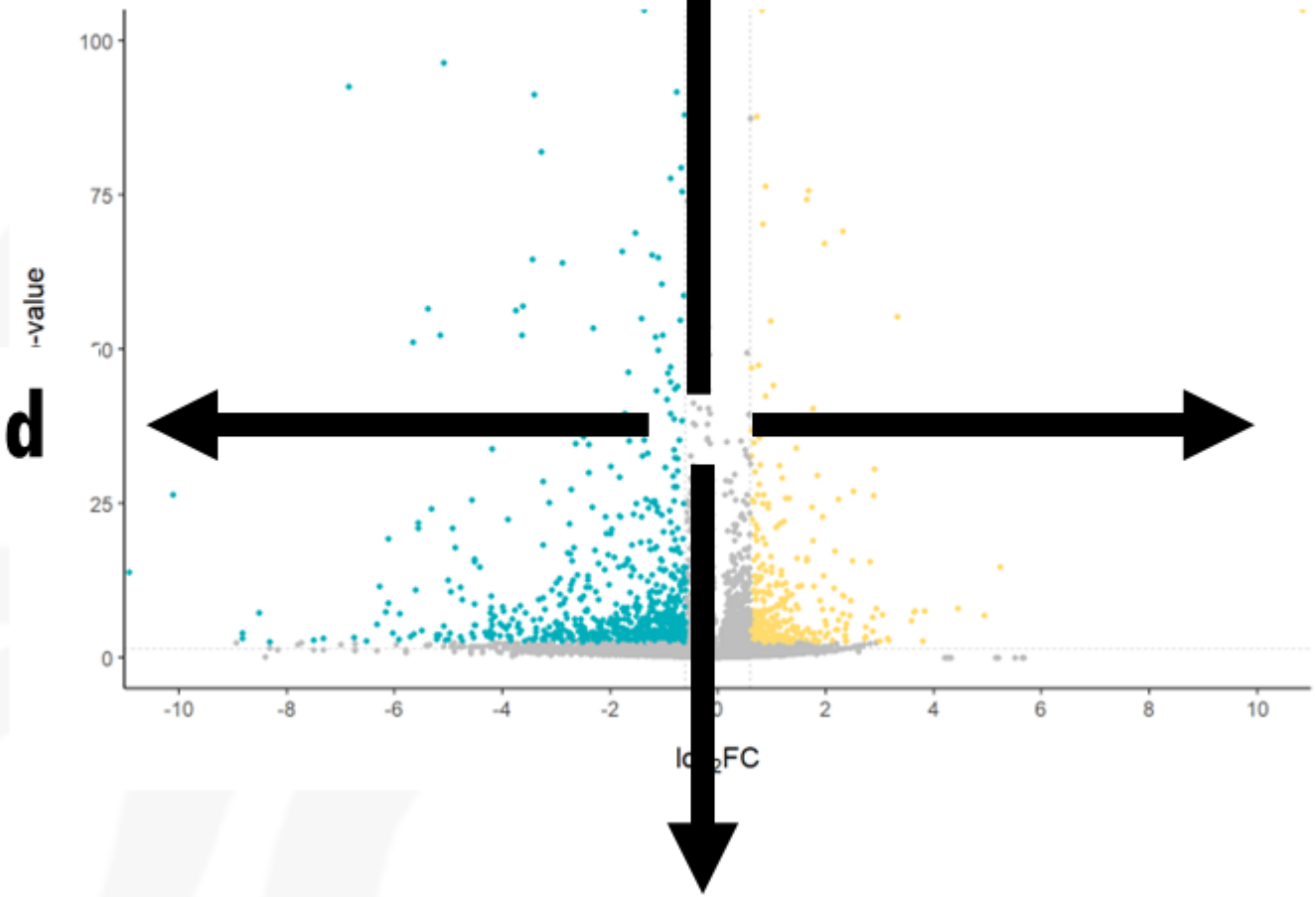
$$FC = \frac{\text{expression of gene X in B}}{\text{expression of gene X in A}}$$



$$FC = \frac{\text{expression of gene X in } \mathbf{B}}{\text{expression of gene X in } \mathbf{A}}$$



Highly significant



Downregulated

Upregulated

Not significant

Enrichment Analysis

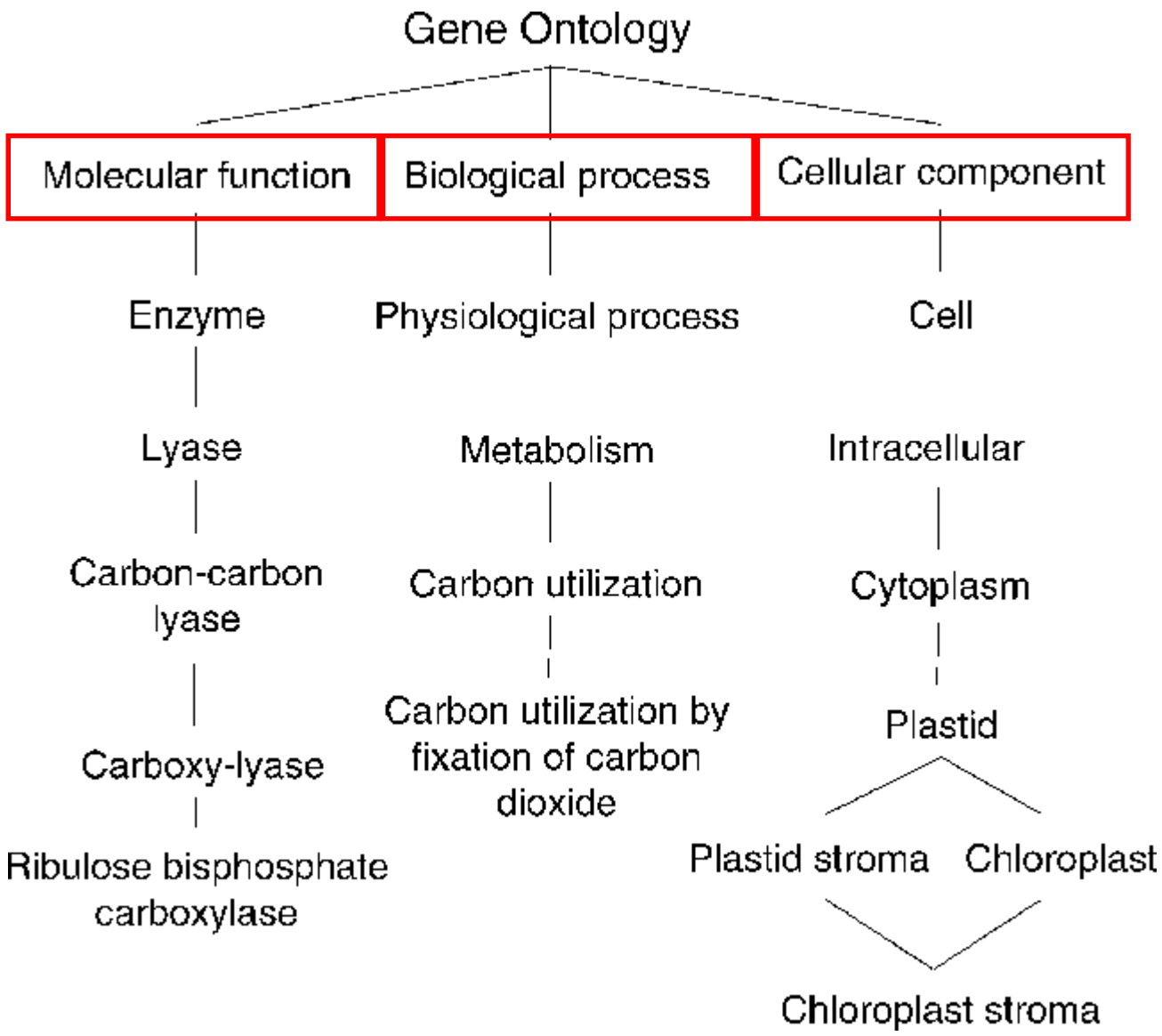
- L'analisi di arricchimento è un metodo che identifica se insiemi predefiniti di geni o tratti sono significativamente sovrarappresentati in un elenco di geni, aiutando a interpretare processi o funzioni biologici associati a condizioni specifiche.

Differenza essenziale:
Gene Ontology (GO): si concentra sull'associazione funzionale di uno specifico sottoinsieme di geni (come quelli espressi in modo differenziale) con funzioni o processi.
GSEA: esamina tutti i geni e identifica set di geni predefiniti arricchiti in un contesto più ampio, anche senza utilizzare un limite rigoroso.

Part 12. GO (Gene Ontology) Enrichment Analysis

```
# Part 12. GO (Gene Ontology) enrichment Analysis.==============================

# BiocManager::install("DOSE")
library(DOSE)

# BiocManager::install("clusterProfiler")
library(clusterProfiler)

# BiocManager::install("org.Mm.eg.db")
library(org.Mm.eg.db)
```

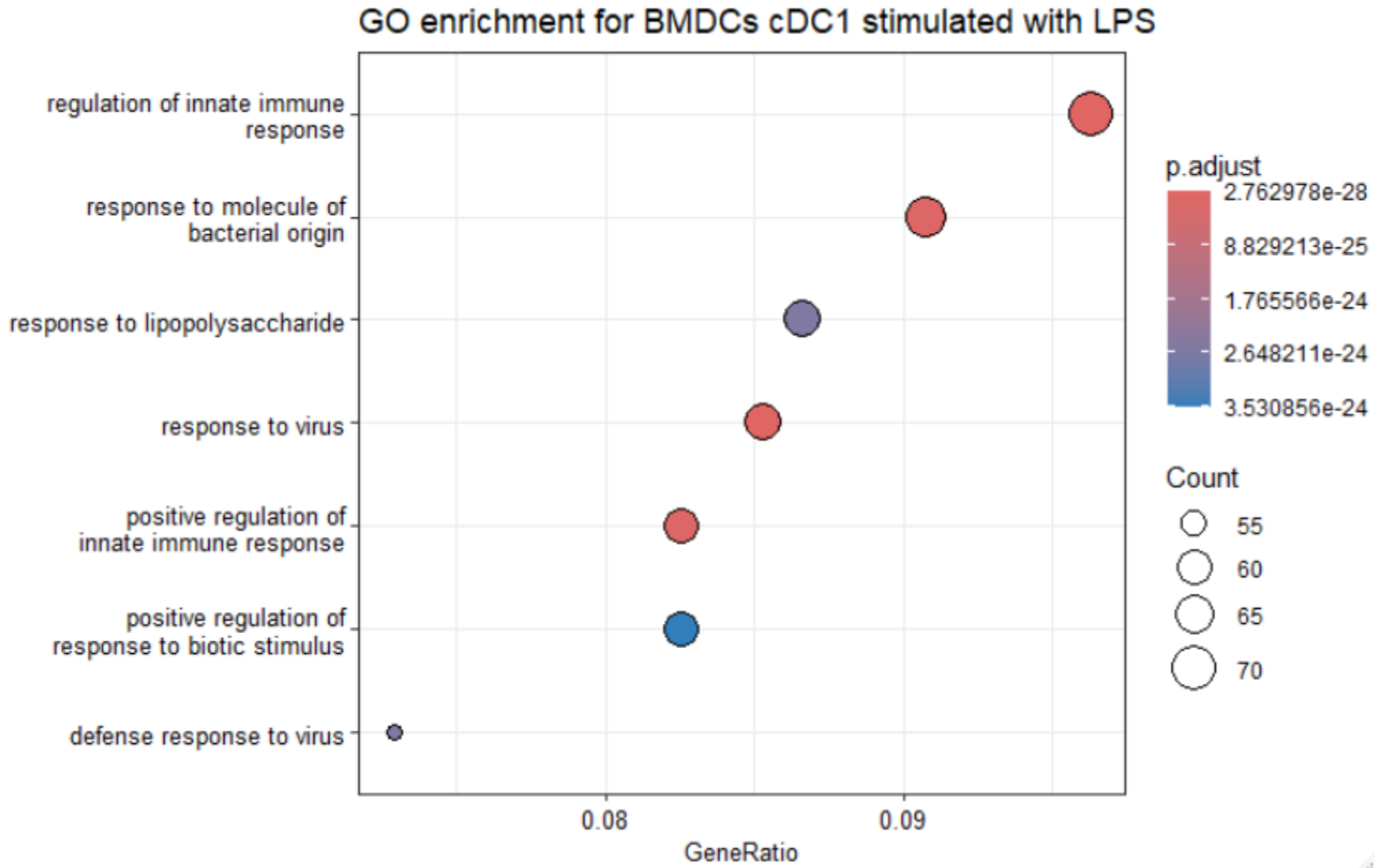
```
# Select only group 2 for analysis.-------------------------------------------
DEG_LPS <- filter(DEG, Enriched == "cDC1_LPS")
DEG_LPS$entrez = mapIds(org.Mm.eg.db,
                        keys=as.character(DEG_LPS$gene),
                        column = "ENTREZID",
                        keytype = "SYMBOL",
                        multiVals = "first")
DEG_LPS <- DEG_LPS$entrez              # Create a list with genes.
go_cDC1_LPS <- enrichGO(gene = DEG_LPS,
                        OrgDb = org.Mm.eg.db,
                        pvalueCutoff = 0.05,
                        qvalueCutoff = 0.05,
                        ont="all",                  #BP, CC, MF or all
                        readable = T)
```



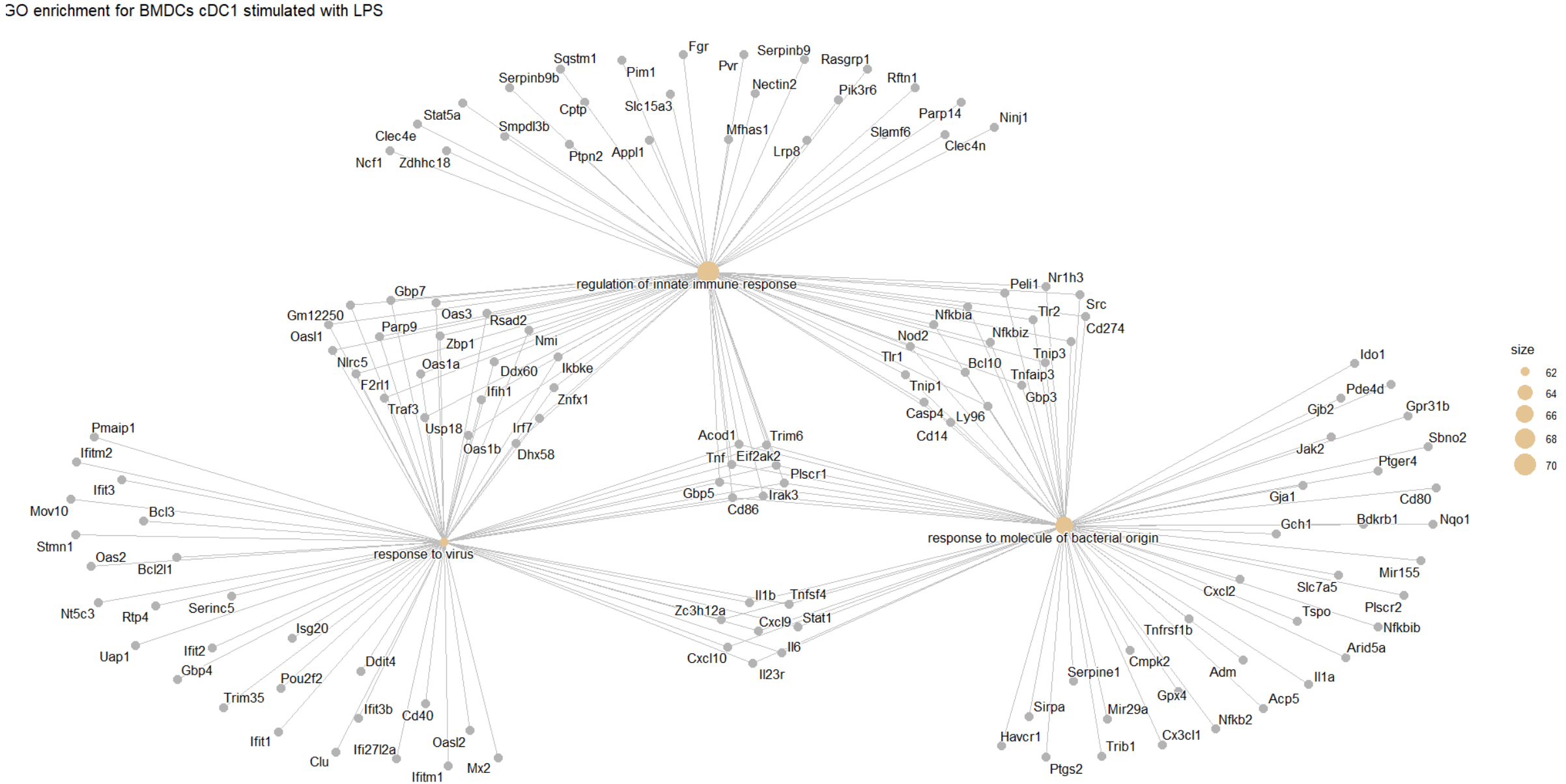
Gerarchico

Gene Ontology: Un sistema di classificazione che raggruppa i geni in base alle loro funzioni biologiche, alla posizione cellulare o al ruolo molecolare.

- **Obiettivo**: Identificare quali processi biologici, componenti cellulari o funzioni molecolari sono associati a un insieme di geni.
- **Input**: un elenco di geni precedentemente selezionati (ad esempio: geni espressi in modo differenziale).
- **Risultato**: termini GO (funzioni o processi) arricchiti, cioè più rappresentati in questo elenco di geni di quanto ci si aspetterebbe per caso.


GO enrichment for BMDCs cDC1 stimulated with LPS
regulation of innate immune response
response to molecule of bacterial origin
response to lipopolysaccharide
response to virus
positive regulation of innate immune response
positive regulation of response to biotic stimulus
defense response to virus
0.08
0.09
GeneRatio
p.adjust
2.762978e-28
8.829213e-25
1.765566e-24
2.648211e-24
3.530856e-24
Count
55
60
65
70

Category Network Plot - Cnetplot



Part 13. Gene Set Enrichment Analysis

```
# Part 13. GSEA.=================================================================
print("Gene Set Enrichment Analysis...")

DEG$entrez = mapIds(org.Mm.eg.db, keys=as.character(DEG$gene), column = "ENTREZID",
                    keytype = "SYMBOL", multiVals = "first")
genelist <- DEG$entrez                          # Create a list with genes.

rnk_gsea <- dplyr::bind_cols(DEG$gene,
                             as.numeric(-log10(DEG$padj) *
                                          sign(DEG$logFC)))
# rnk_gsea <- rnk_gsea[order(rnk_gsea[,2], decreasing = TRUE),]
rnk_gsea <- rnk_gsea[!is.na(rnk_gsea$...2),]
rnk_gsea_clean <- rnk_gsea[rnk_gsea$...2 !=-Inf,]
rnk_gsea_clean <- rnk_gsea_clean[rnk_gsea_clean$...2 !=Inf,]


write.table(rnk_gsea_clean, file = "BMDC_DC1_ctrl_vs_LPS.rnk",
            quote = FALSE,
            row.names = FALSE,
            col.names = FALSE,
            sep = "\t")             # load this file into the GSEA program.

# End of script.
################################################################################
```

```
Sept3   -3.83629341624994
Parvg   -3.83629341624994
Il2ra   3.83629341624994
Clec4e  3.68019522299916
Ccr2    -3.68019522299916
G530011O06Rik   3.68019522299916
Tnip3   3.68019522299916
Fas     3.68019522299916
Dpep3   -3.68019522299916
Tm6sf1  -3.62576295323193
F630028O10Rik   -3.62576295323193
Cacna1e -3.62576295323193
Nrp2    3.62576295323193
Cd80    3.62576295323193
Cd70    3.62576295323193
Gbp7    3.62103317804584
Srgap3  -3.62103317804584
Oasl2   3.62103317804584
Lpar6   -3.62103317804584
Dusp14  3.62103317804584
Sigmar1 -3.62103317804584
Tmigd3///Adora3 -3.62103317804584
Srgap3  -3.62103317804584
Dpep2   -3.62103317804584
Pydc4   3.62103317804584
Tlr3    -3.62103317804584
Ifit1   3.62103317804584
Tbxas1  -3.5240045795282
Arsb    -3.52711110016329
Cd80    3.51728416281519
Susd2   3.38296157379116
Nmrk1   3.38296157379116
Ikzf4   3.38218734696675
Itgb7   -3.3501543787326
Hepacam2        -3.31851934032522
Edil3   3.31851934032522
Tnfsf9  3.31611944197327
Gsn     -3.31611944197327
Rtp4    3.31611944197327
```

UC San Diego

BROAD INSTITUTE

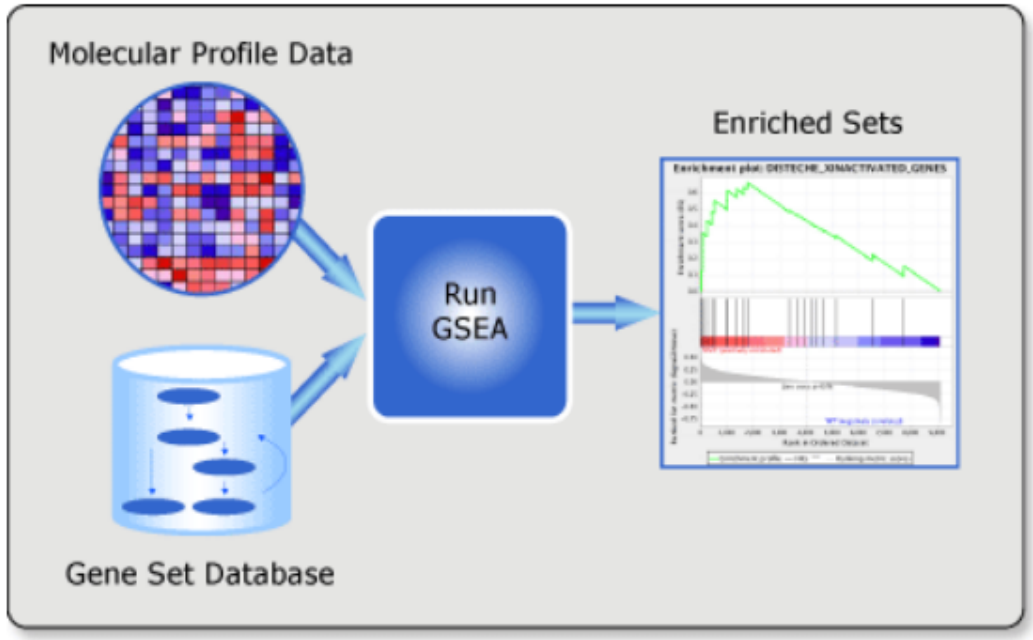
Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

- **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- **View documentation** describing GSEA and MSigDB.
- View guidelines for **using RNA-seq datasets with GSEA**.
- Use the **GenePattern** platform to run analyses, including classical GSEA and a variation designed for single-sample analysis (**ssGSEA**).

What's New

9-Aug-2024: MSigDB 2024.1 provides collection updates for GO, Reactome, WikiPathways, and more along with numerous new set additions for Human and Mouse Databases. Additionally, gene data has been updated to Ensembl 112. See the release notes for details.



License Terms

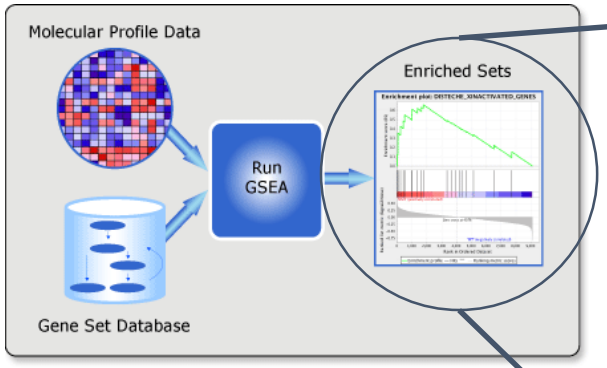
GSEA and MSigDB are available for use under these license terms.

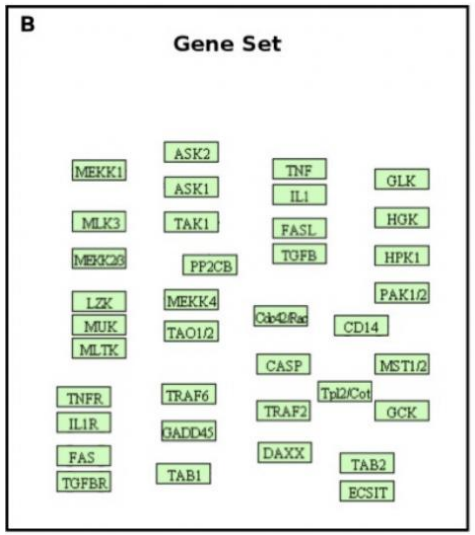
Please register to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

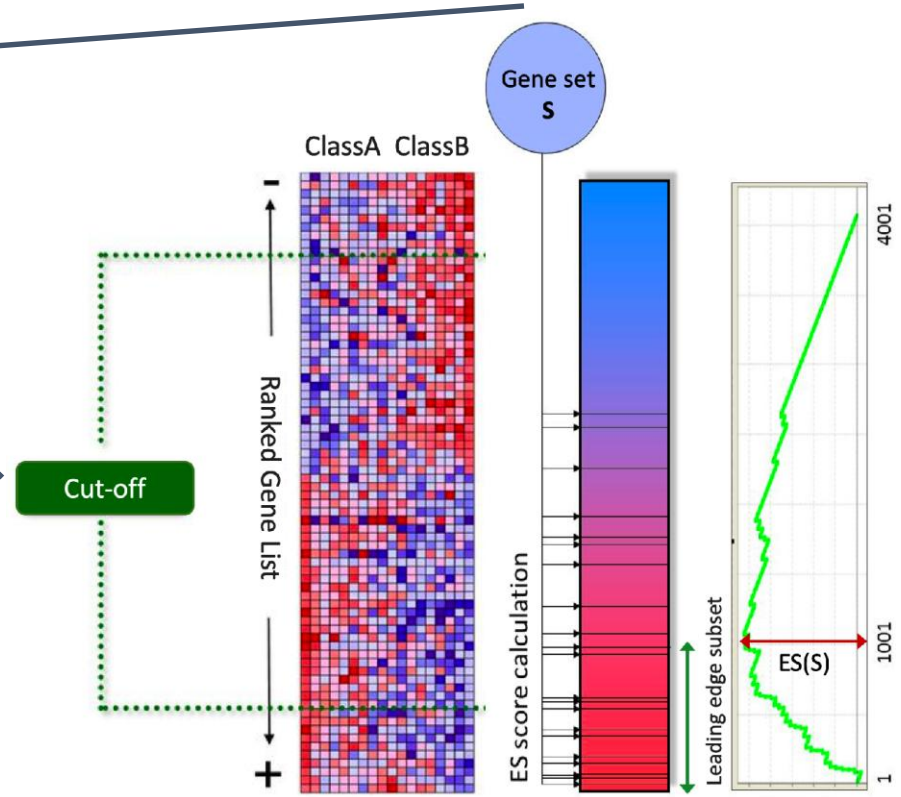
Citing GSEA

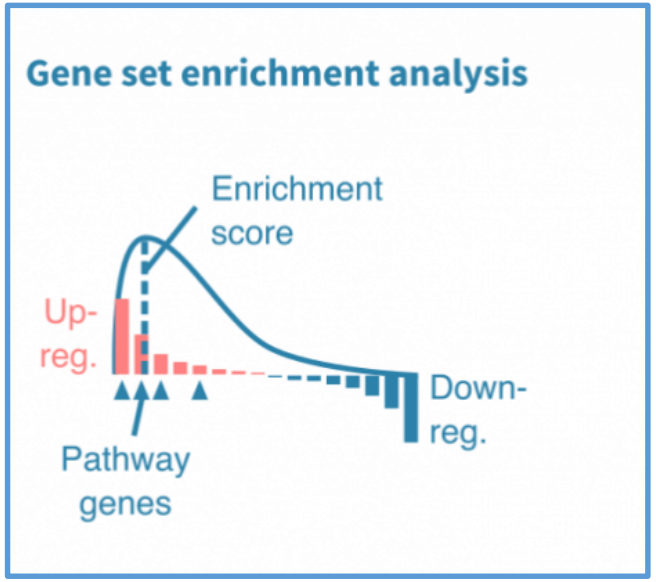
Non gerarchico

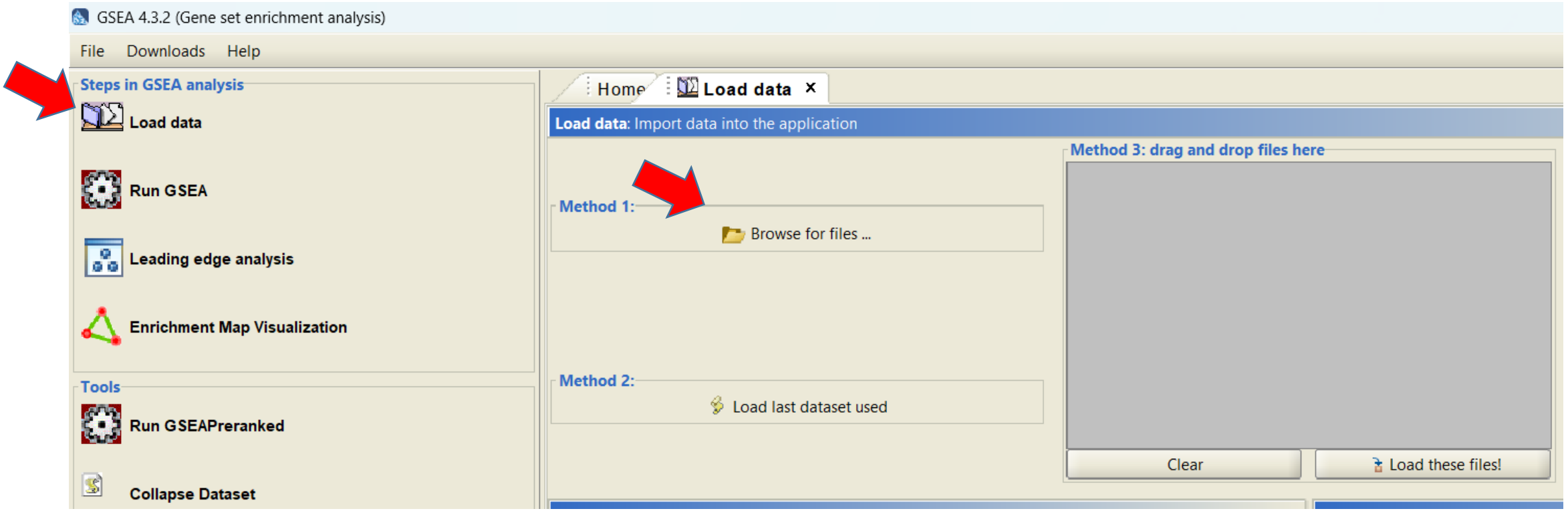
Gene Set Enrichment Analysis (GSEA)

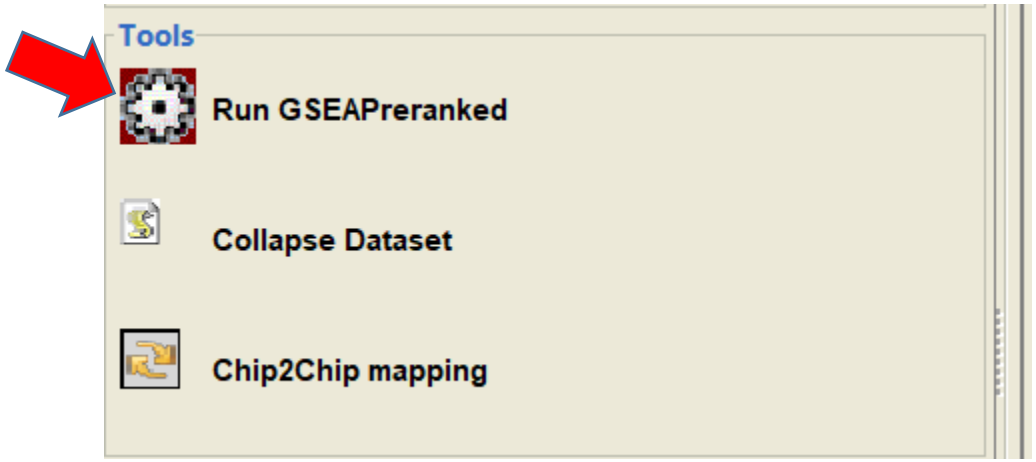


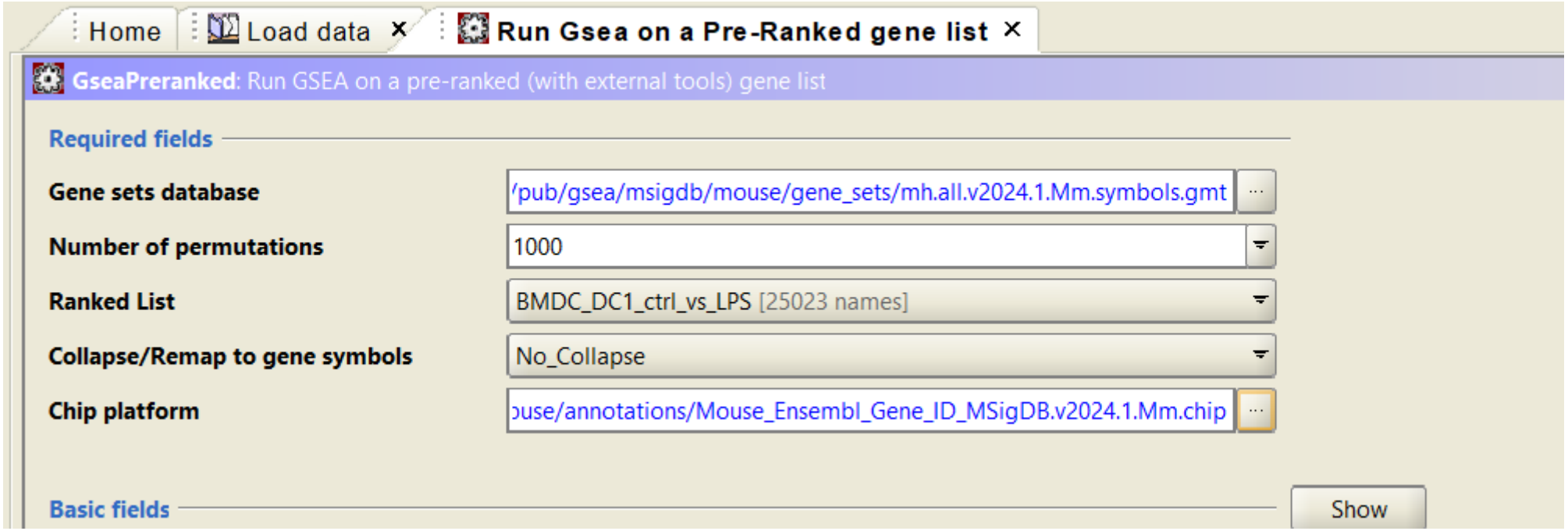







GSEA 4.3.2 (Gene set enrichment analysis)
File
Downloads
Help
Steps in GSEA analysis
Load data
Run GSEA
Leading edge analysis
Enrichment Map Visualization
Tools
Run GSEAPreranked
Collapse Dataset
Home
Load data
Load data: Import data into the application
Method 1:
Browse for files ...
Method 2:
Load last dataset used
Method 3: drag and drop files here
Clear
Load these files!


Tools
Run GSEAPreranked
Collapse Dataset
Chip2Chip mapping



Collapse: Quando sono presenti più sonde o trascritti mappati sullo stesso gene, GSEA applica un criterio per ridurre questi dati (ad esempio selezionando il valore di espressione più alto o utilizzando una media) e associa il valore finale al gene corrispondente.
No collapse: Ogni ID di sonda o trascrizione viene trattato individualmente. Ciò può portare a più voci per lo stesso gene se ad esso sono associati più sonde o trascrizioni.

Premere RUN

GSEA Report for Dataset BMDC_DC1_ctrl_vs_LPS

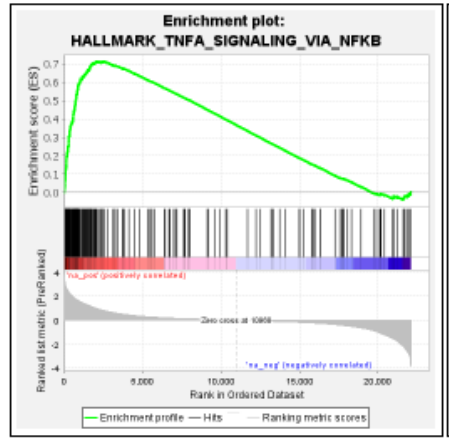
Enrichment in phenotype: na

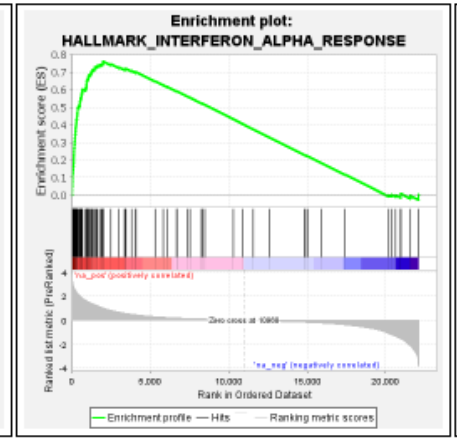
- 40 / 50 gene sets are upregulated in phenotype **na_pos**
- 31 gene sets are significant at FDR < 25%
- 20 gene sets are significantly enriched at nominal pvalue < 1%
- 23 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

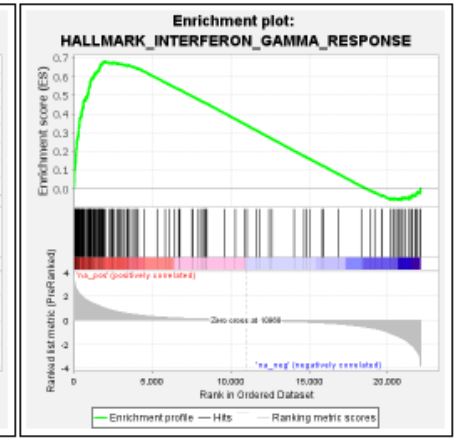
Enrichment in phenotype: na

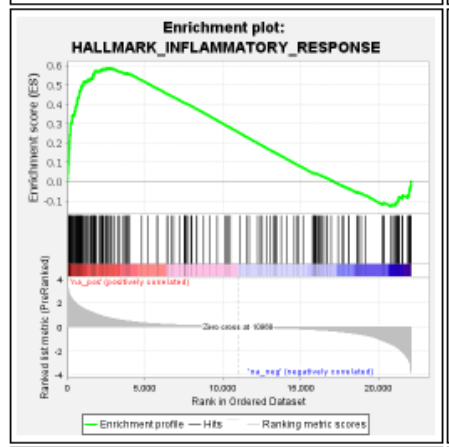
- 10 / 50 gene sets are upregulated in phenotype **na_neg**
- 1 gene sets are significantly enriched at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 2 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

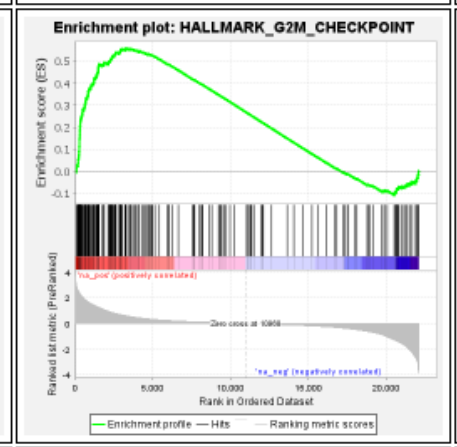
Table: Snapshot of enrichment results

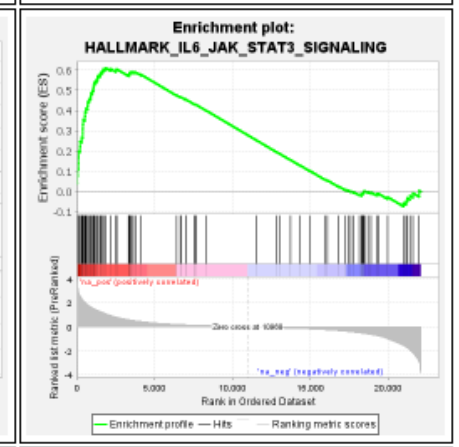


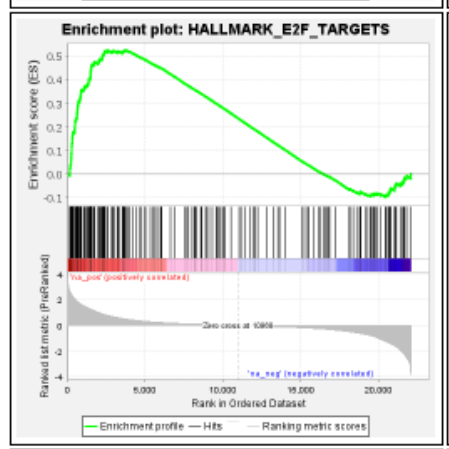


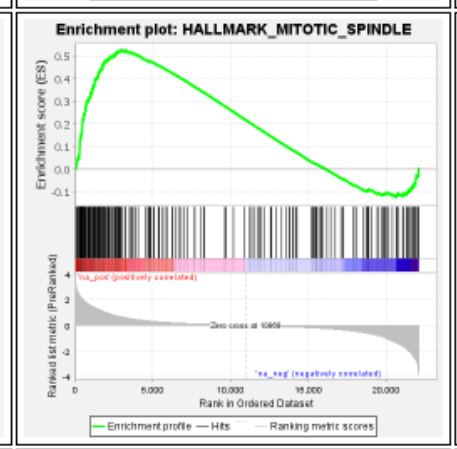


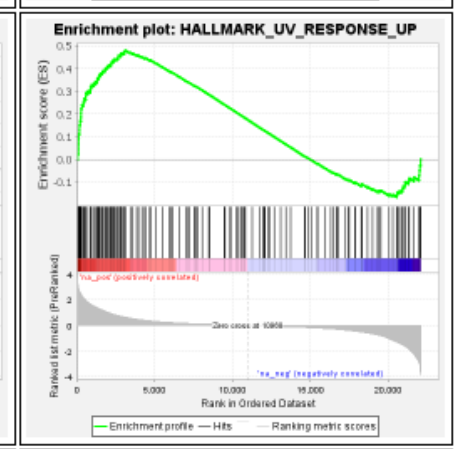












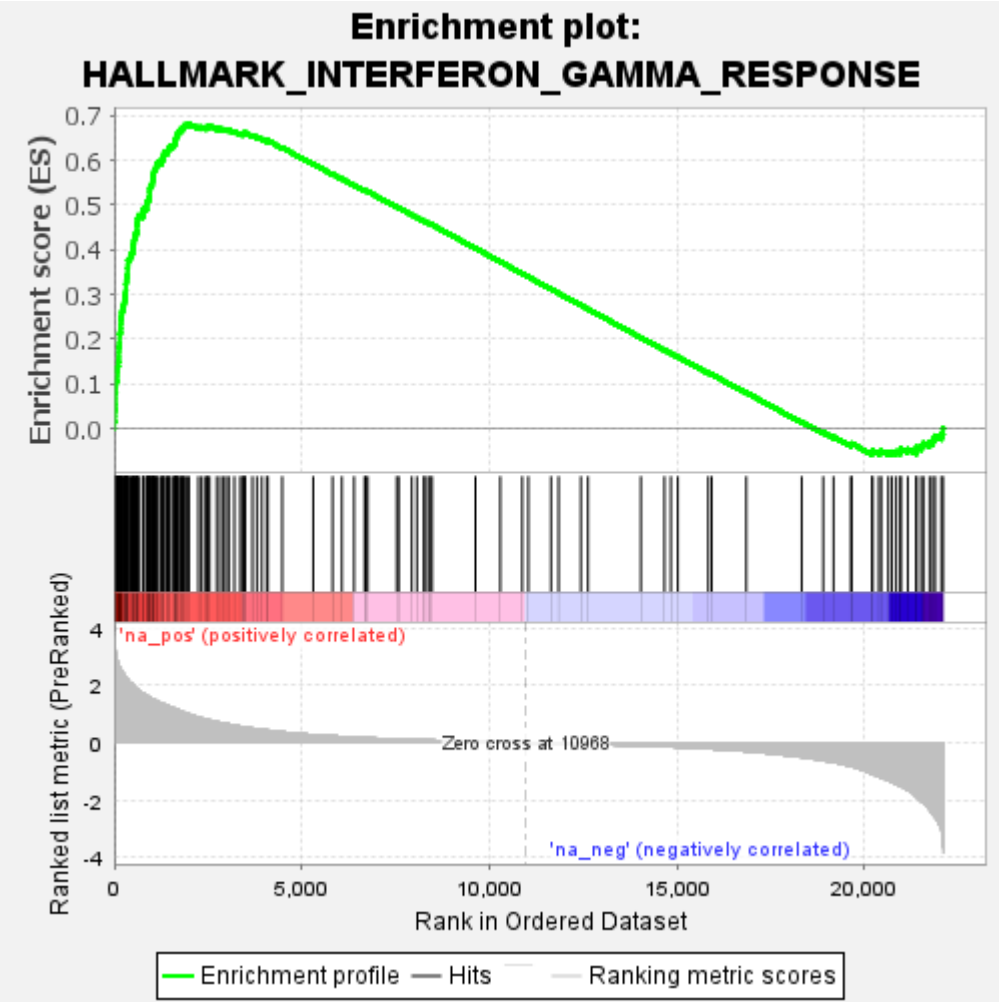


Table: GSEA Results Summary

Dataset	BMDC_DC1_ctrl_vs_LPS
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	HALLMARK_INTERFERON_GAMMA_RESPONSE
Enrichment Score (ES)	0.68222123
Normalized Enrichment Score (NES)	2.372906
Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

MSigDB Home

Human Collections
- About
- Browse
- Search
- Investigate
- Gene Families

Mouse Collections
- About
- Browse
- Search
- Investigate

Help

UC San Diego

BROAD INSTITUTE

MSigDB
Molecular Signatures Database

Molecular Signatures Database

Overview

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS human gene set page.
- **Browse** gene sets by name or collection.
- **Search** for gene sets by keyword.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEx**
- **Download** gene sets.

License Terms

GSEA and MSigDB are available for use under these license terms.

Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented expression data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

MSigDB Home

Human Collections
- About
- Browse
- Search
- Investigate
- Gene Families

Mouse Collections
- About
- Browse
- Search
- Investigate

Help

Browse Mouse Gene Sets

UC San Diego BROAD INSTITUTE

Gene set name: Search *(Enter full or partial name)*

By first letter: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

By collection: [about the MSigDB Mouse collections]

- **MH** (orthology-mapped hallmark gene sets, 50 gene sets)
- **M1** (positional gene sets, 341 gene sets)
  - by chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y MT
- **M2** (curated gene sets, 2710 gene sets)
  - **CGP** (chemical and genetic pertubations, 980 gene sets)
  - **CP** (canonical pathways, 1730 gene sets)
    - **CP:BIOCARTA** (BioCarta gene sets, 252 gene sets)
    - **CP:REACTOME** (Reactome gene sets, 1289 gene sets)
    - **CP:WIKIPATHWAYS** (WikiPathways gene sets, 189 gene sets)
- **M3** (regulatory gene sets, 2047 gene sets)
  - **GTRD** (GTRD transcription factor targets, 279 gene sets)
  - **MIRDB** (miRDB microRNA targets, 1768 gene sets)
- **M5** (ontology gene sets, 10678 gene sets)

- **GO:BP** (GO biological process, 7713 gene sets)
- **GO:CC** (GO cellular component, 1028 gene sets)
- **GO:MF** (GO molecular function, 1845 gene sets)

- **MPT** (Mouse Phenotype Ontology MP Tumor, 92 gene sets)

- **M8** (cell type signature gene sets, 233 gene sets)

Click on a gene set name to view its gene set page.

Back to Top

GOBP_ACUTE_INFLAMMATORY_RESPONSE_TO_ANTIGENIC_STIMULUS
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_ENDOGENOUS_ANTIGEN
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_EXOGENOUS_ANTIGEN
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_EXOGENOUS_PEPTIDE_ANTIGEN
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_EXOGENOUS_PEPTIDE_ANTIGEN_VIA_MHC_CLASS_I
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_EXOGENOUS_PEPTIDE_ANTIGEN_VIA_MHC_CLASS_II
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN_VIA_MHC_CLASS_I
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_OR_POLYSACCHARIDE_ANTIGEN_VIA_MHC_CLASS_II
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_VIA_MHC_CLASS_IB
GOBP_ANTIGEN_RECEPTOR_MEDIATED_SIGN
GOBP_NEGATIVE_REGULATION_OF_INFLAMMATORY_RESPONSE_TO_ANTIGENIC_STIMULUS
GOBP_PEPTIDE_ANTIGEN_ASSEMBLY_WITH_MHC_CLASS_I_PROTEIN_COMPLEX
GOBP_PEPTIDE_ANTIGEN_ASSEMBLY_WITH_MHC_CLASS_II_PROTEIN_COMPLEX
GOBP_PEPTIDE_ANTIGEN_ASSEMBLY_WITH_MHC_PROTEIN_COMPLEX
GOBP_POSITIVE_REGULATION_OF_ACUTE_INFLAMMATORY_RESPONSE_TO_ANTIGENIC_STIMULUS
GOBP_POSITIVE_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_POSITIVE_REGULATION_OF_ANTIGEN_RECEPTOR_MEDIATED_SIGNALING_PATHWAY
GOBP_POSITIVE_REGULATION_OF_DENDRITIC_CELL_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_POSITIVE_REGULATION_OF_INFLAMMATORY_RESPONSE_TO_ANTIGENIC_STIMULUS
GOBP_REGULATION_OF_ACUTE_INFLAMMATORY_RESPONSE_TO_ANTIGENIC_STIMULUS
GOBP_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_T_CELL_ACTIVATION_VIA_T_CELL_RECEPTOR_CONTACT_WITH_ANTIGEN_BOUND_TO_MHC_MOLECULE_ON_ANTIGEN_PRESENTING_CELL
GOBP_T_CELL_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_TOLERANCE_INDUCTION_TO_SELF_ANTIGEN
GOLDRATH_ANTIGEN_RESPONSE
GOMF_ANTIGEN_BINDING
GOMF_PEPTIDE_ANTIGEN_BINDING
GOMF_PROTEIN_ANTIGEN_BINDING
REACTOME_ANTIGEN_ACTIVATES_B_CELL_RECEPTOR_BCR_LEADING_TO_GENERATION_OF_SECOND_MESSENGERS
REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC
REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION
REACTOME_ANTIGEN_PROCESSING_UBIQUITINATION_PROTEASOME_DEGRADATION
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING_PRESENTATION
REACTOME_CROSS_PRESENTATION_OF_PARTICULATE_EXOGENOUS_ANTIGENS_PHAGOSOMES

Mouse Gene Set: GOBP_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION

For the Human gene set with the same name, see GOBP_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION

Standard name	GOBP_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION
Systematic name	MM4309
Brief description	Any process that modulates the frequency, rate, or extent of antigen processing and presentation. [GOC:add]
Full description or abstract	
Collection	M5: Ontology GO: Gene Ontology GO:BP: GO Biological Process
Source publication	
Exact source	GO:0002577
Related gene sets	
External links	http://amigo.geneontology.org/amigo/term/GO:0002577
Filtered by similarity ?	
Source species	Mus musculus
Contributed by	Gene Ontology (Gene Ontology Consortium)
Source platform or identifier namespace	Mouse_NCBI_Gene_ID
Dataset references	
Download gene set	format: grp | gmt | xml | json | TSV metadata
Compute overlaps ?	(show collections to investigate for overlap with this gene set)
Compendia expression profiles ?	NG-CHM interactive heatmaps (*Please note that clustering takes a few seconds*) Mouse Transcriptomic BodyMap compendium Legacy heatmaps (PNG) Mouse Transcriptomic BodyMap compendium

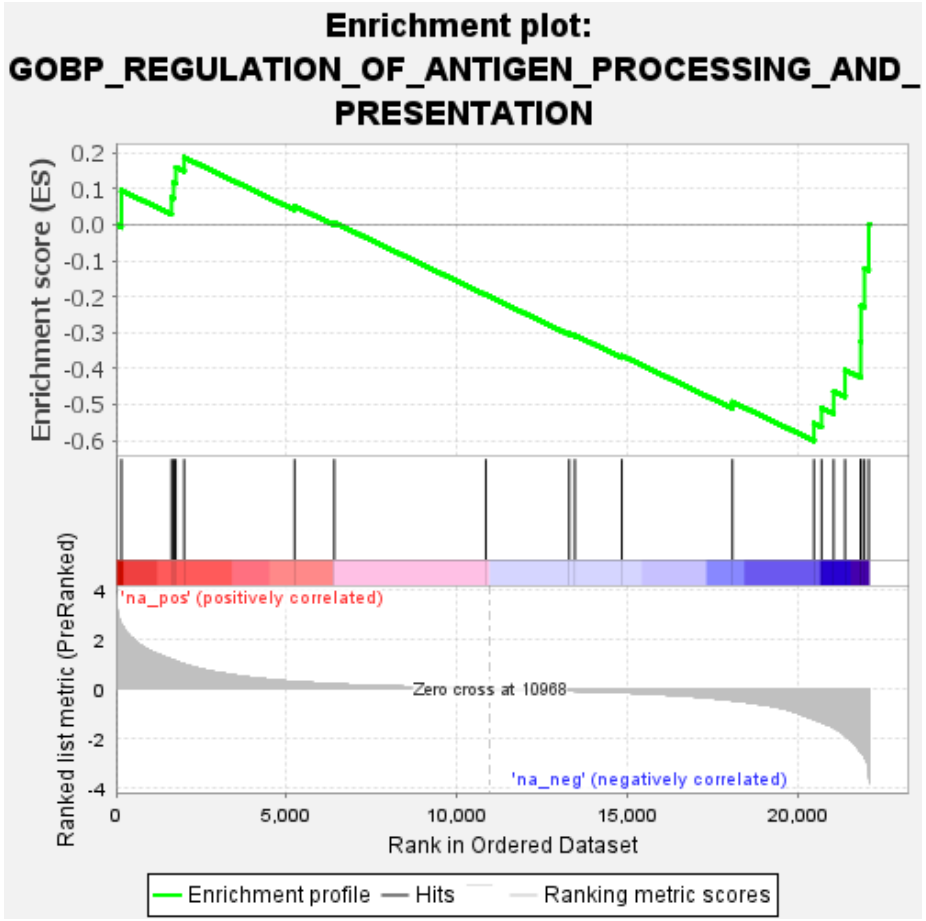
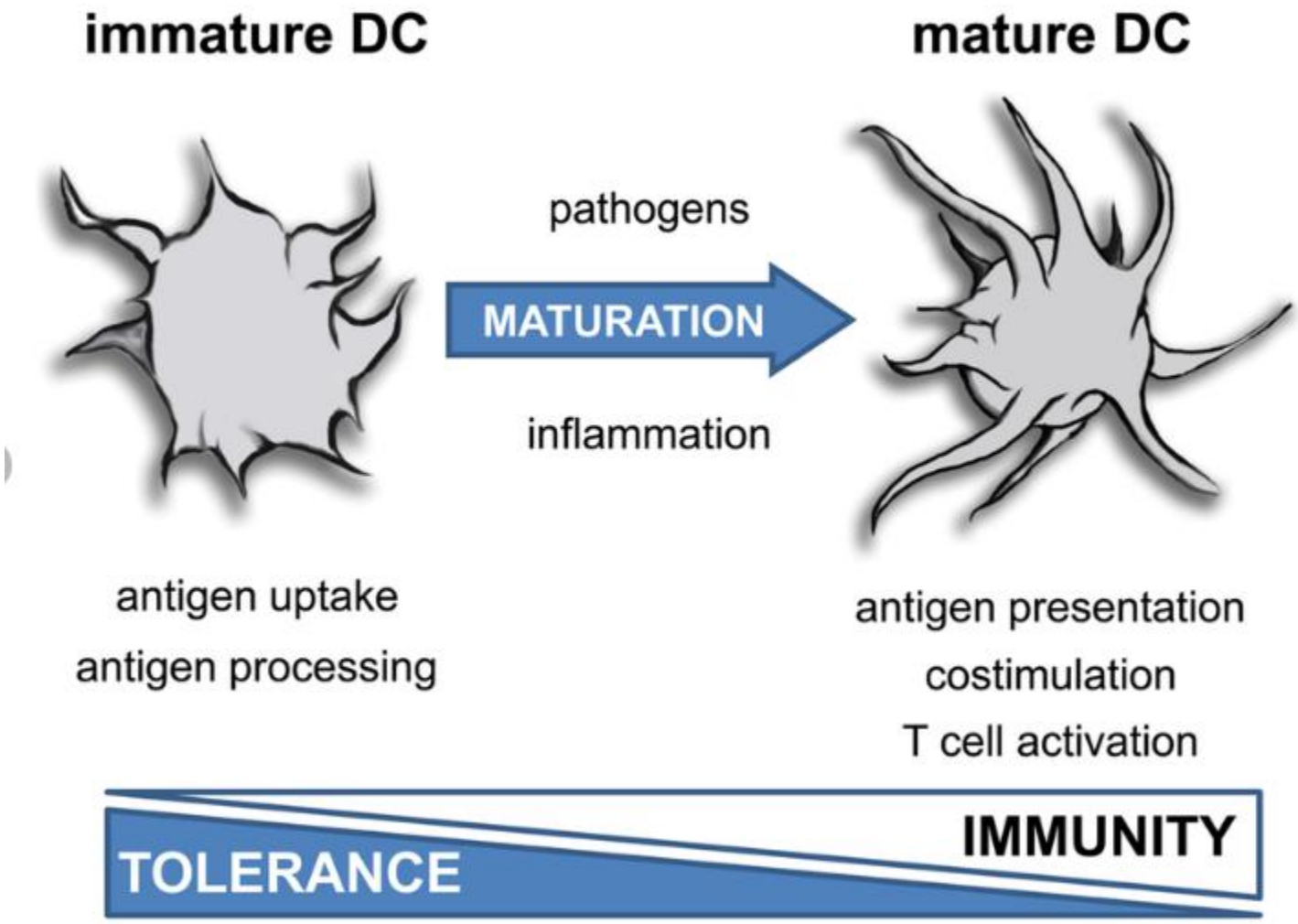


Table: GSEA Results Summary

Dataset	BMDC_DC1_ctrl_vs_LPS
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_neg
GeneSet	GOBP_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION
Enrichment Score (ES)	-0.60123724
Normalized Enrichment Score (NES)	-1.4314585
Nominal p-value	0.037950665
FDR q-value	0.037950665
FWER p-Value	0.02

Conclusion


immature DC
mature DC
pathogens
MATURATION
inflammation
antigen uptake
antigen processing
antigen presentation
costimulation
T cell activation
TOLERANCE
IMMUNITY

Troubleshooting - packages

1) Verifica la versione R:
version()

2) Errore nel compilare il codice fonte:
*Windows: Installa **Rtools**.*
*Mac: Installa **Xcode Command Line Tools**.*

3) Verificare che ci siano dipendenze ausenti:
*install.packages(c("**curl**", "**jsonlite**"))*

4) Errori nella connessione con il CRAN:
chooseCRANmirror() # choose another repository
install.packages("httr2")

5) Errore di autorizzazioni: #admin
install.packages("httr2",
lib = "percorso/della/cartella/con/permesso")

6) Installazione manuale: #RTolls
Download .tar.gz from CRAN
install.packages("percorso/del/httr2.tar.gz",
repository = NULL, tipo = "fonte")