# Análisis transcriptómicos de la expresión génica

## Máster Universitario en Bioinformática

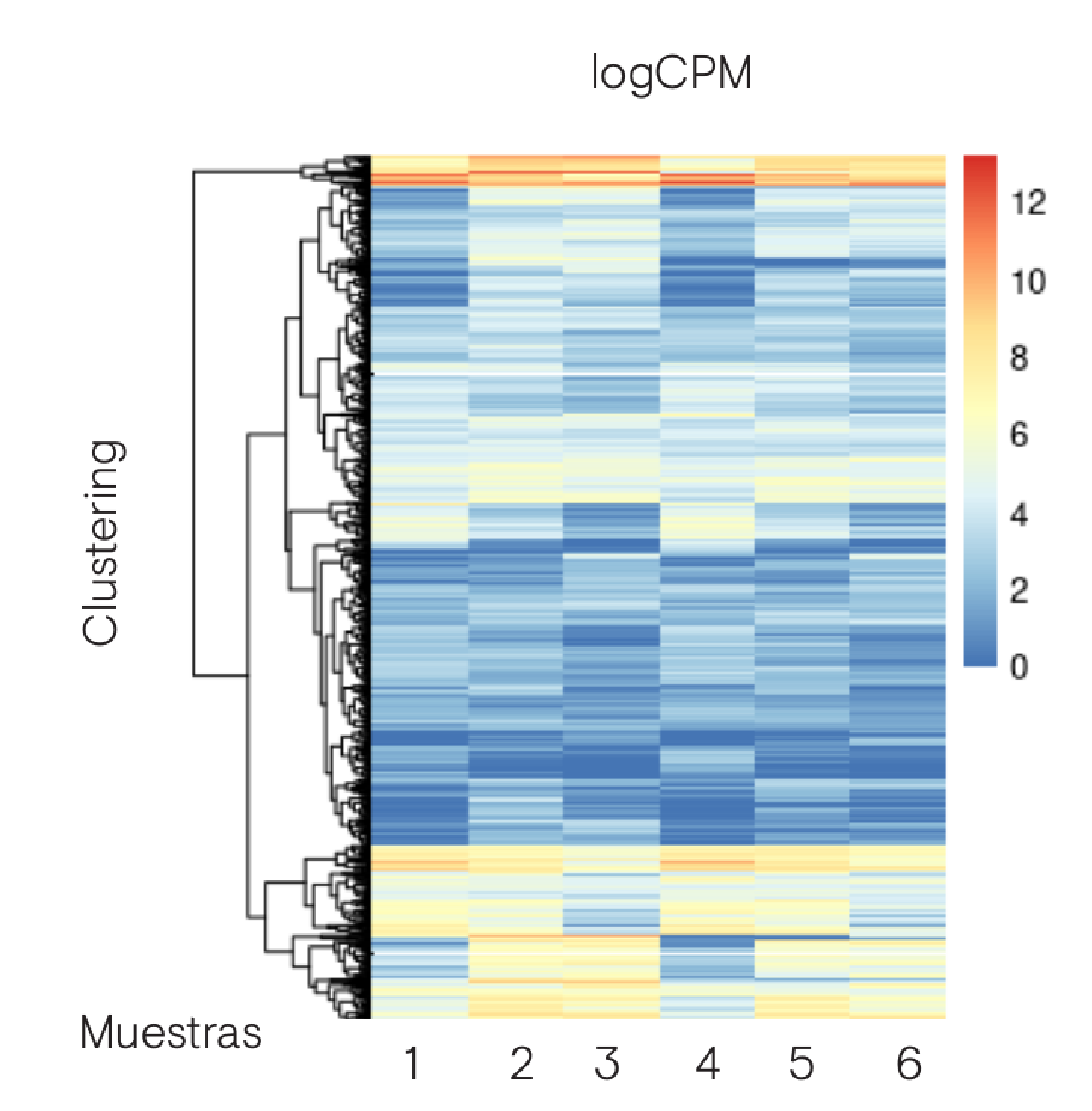
## Sesión 12

Universidad Internacional de Valencia

Dra. Paula Soler Vila
paula.solerv@professor.universidadviu.com

De:
Planeta Formación y Universidades

# Bloque V: Exploración y visualización de resultados



# Objetivos

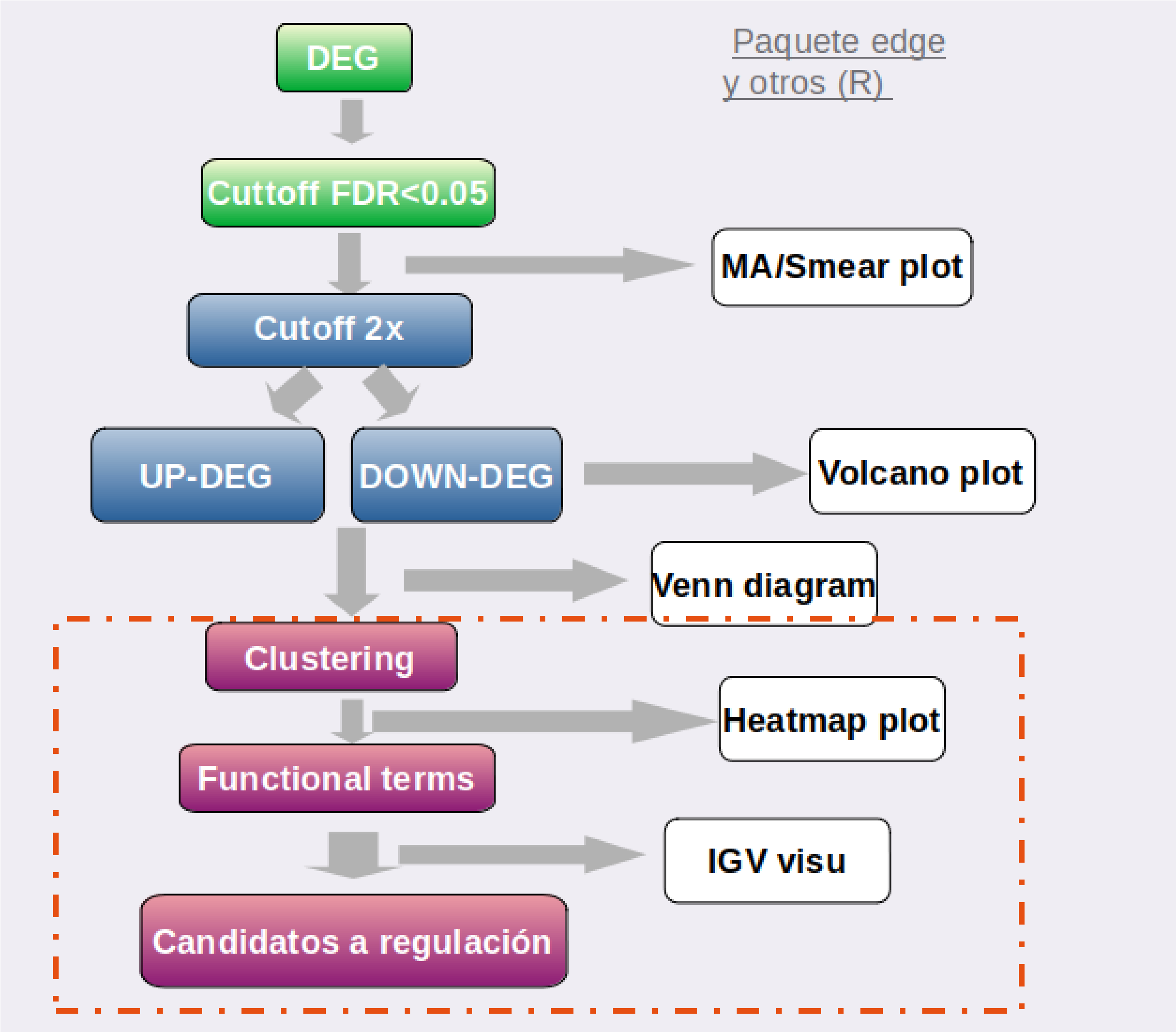
1. Conocer las posibilidades del “**agrupamiento jerárquico**” en la exploración de datos de RNA-seq y en la visualización de patrones de expresión de genes DE.

2. Generar gráficos de **heatmap** a partir de recuentos normalizados ajustando los datos a diferentes **escalas** y métodos de cálculo de **distancias**.

3. Conocer y analizar la **ontología génica** y su aplicación en datos de expresión.

4. Calcular el **enriquecimiento funcional** de listas de genes DE.

# Flujo de trabajo para la exploración y visualización de resultados



Paquete edge
y otros (R)

DEG

Cuttoff FDR<0.05

MA/Smear plot

Cutoff 2x

UP-DEG

DOWN-DEG

Volcano plot

Venn diagram

Clustering

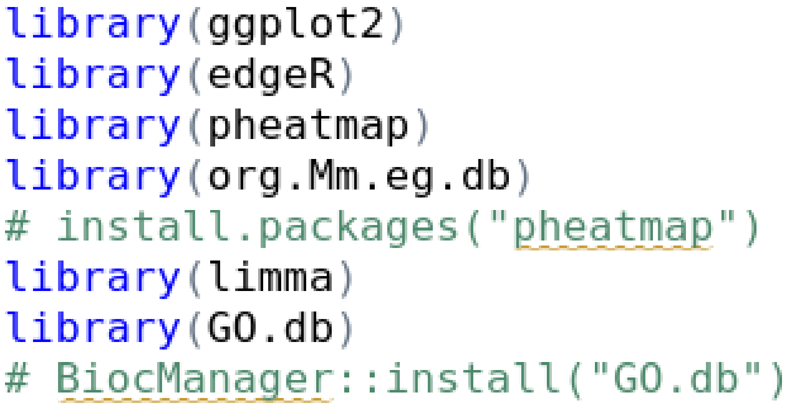
Heatmap plot

Functional terms

IGV visu

Candidatos a regulación

LIBRARY

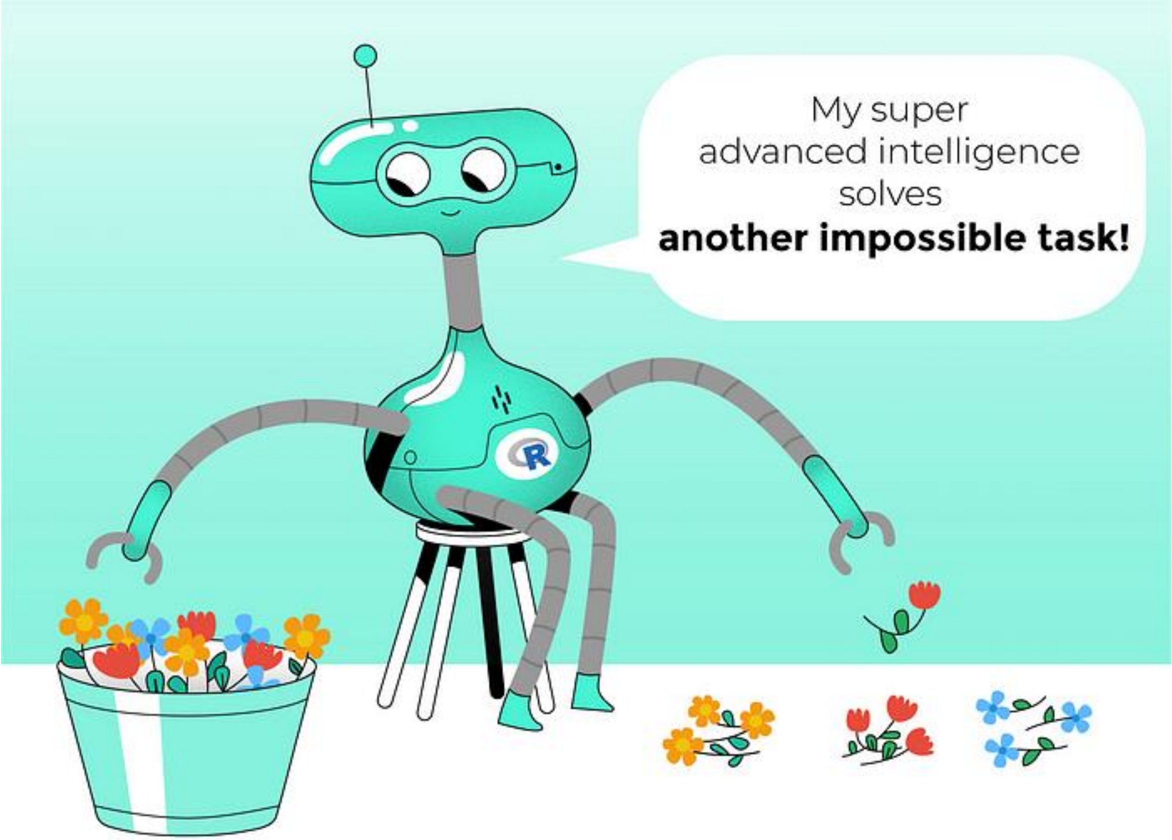


```
library(ggplot2)
library(edgeR)
library(pheatmap)
library(org.Mm.eg.db)
# install.packages("pheatmap")
library(limma)
library(GO.db)
# BiocManager::install("GO.db")
```

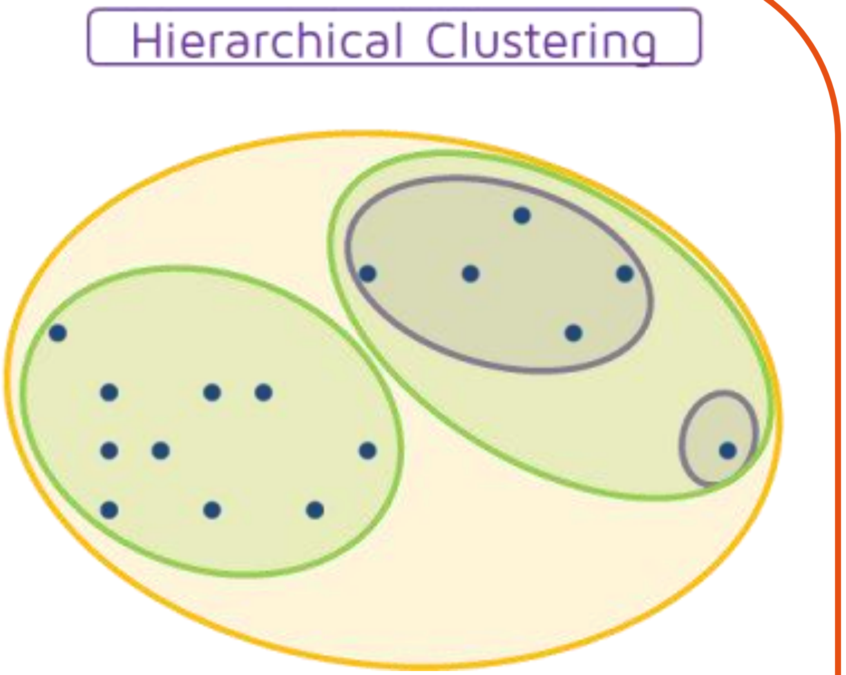
# ¿A qué hace referencia el término *clustering*?

Técnicas *no supervisadas* cuya finalidad es encontrar patrones o grupos (*clusters*) dentro de un conjunto de observaciones que comparten patrones comunes

Las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las observaciones de otros grupos.

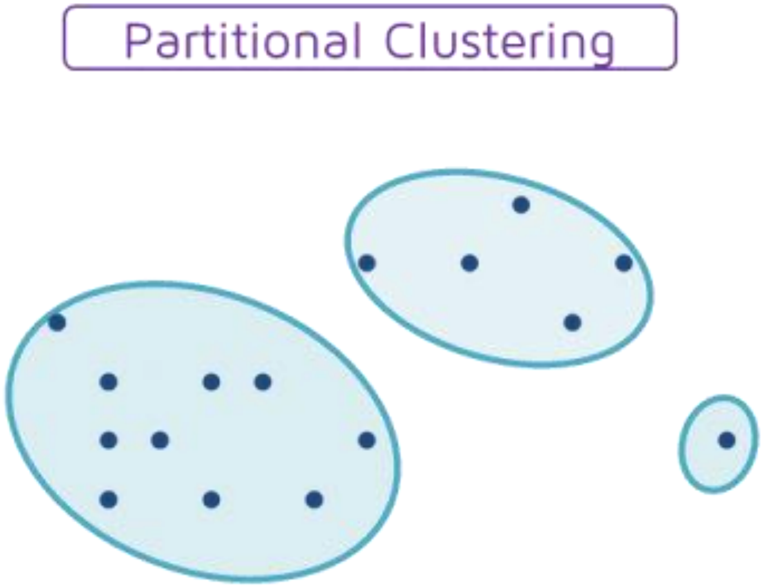

My super advanced intelligence solves **another impossible task!**
R

# ¿A qué hace referencia el término *clustering*?



***Agrupamiento jerárquico (Hierarchical Clustering)***

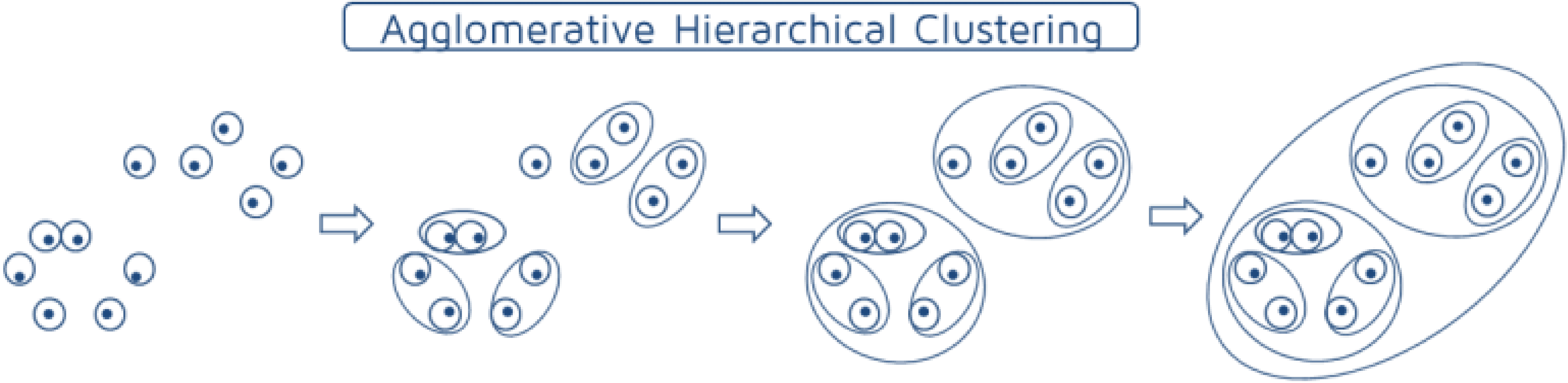
- Cada observación pertenece a múltiples grupos
- Hay solapamiento



***Agrupamiento particional (Partitional Clustering)***

- Cada observación pertenece únicamente a un grupo
- No hay solapamiento
- **K-means clustering**

# ¿A qué hace referencia el término *clustering*?



Se van agrupando atendiendo a su cercanía

# Agrupación jerárquica aglomerativa

1. Inicialización

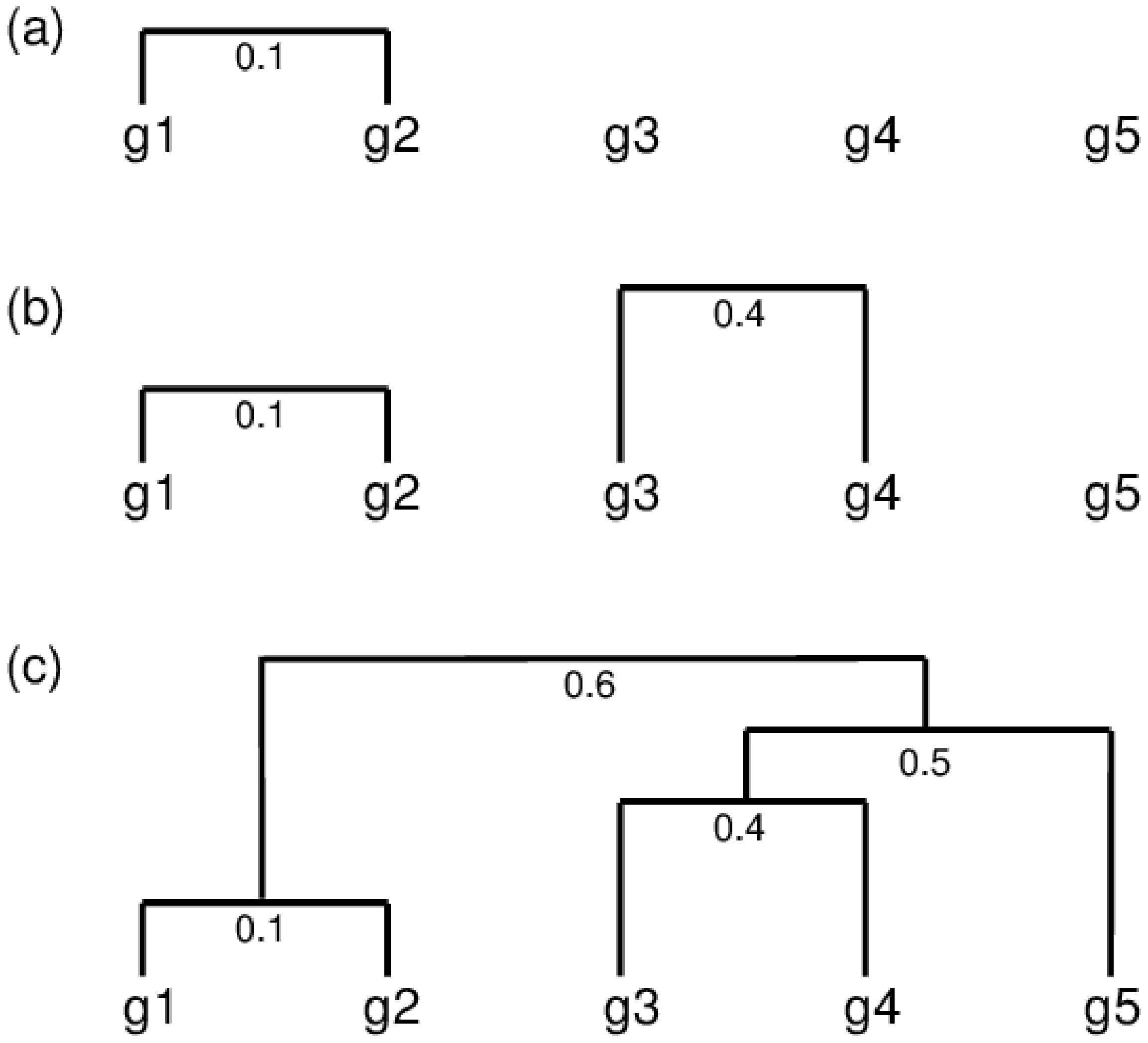
2. Cálculo de distancias

Medidas de distancias

3. Fusión de clusters

Métodos de aglomeramiento (*linkage*)

4. **Dendrograma** final



# Cálculo de las distancias entre observaciones

**Definir y cuantificar la similitud/distancia entre las observaciones**

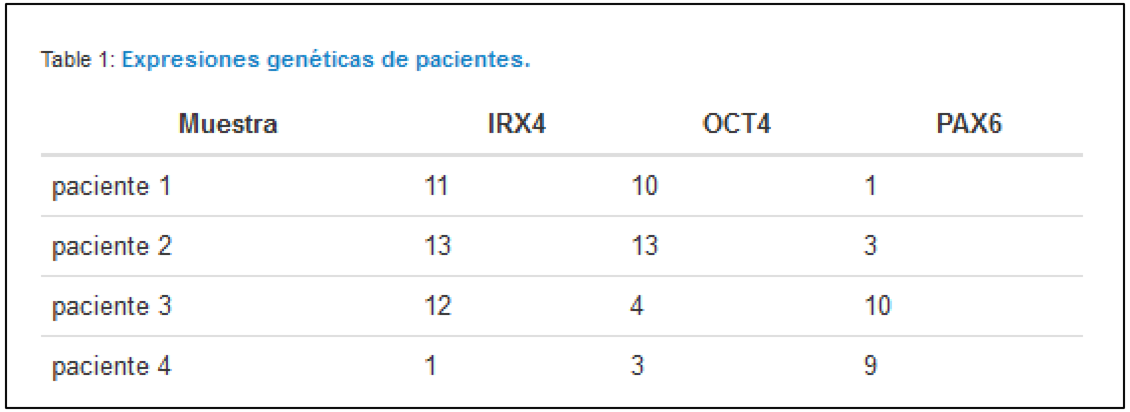


Table 1: Expresiones genéticas de pacientes.

| Muestra | IRX4 | OCT4 | PAX6 |
|---|---|---|---|
| paciente 1 | 11 | 10 | 1 |
| paciente 2 | 13 | 13 | 3 |
| paciente 3 | 12 | 4 | 10 |
| paciente 4 | 1 | 3 | 9 |

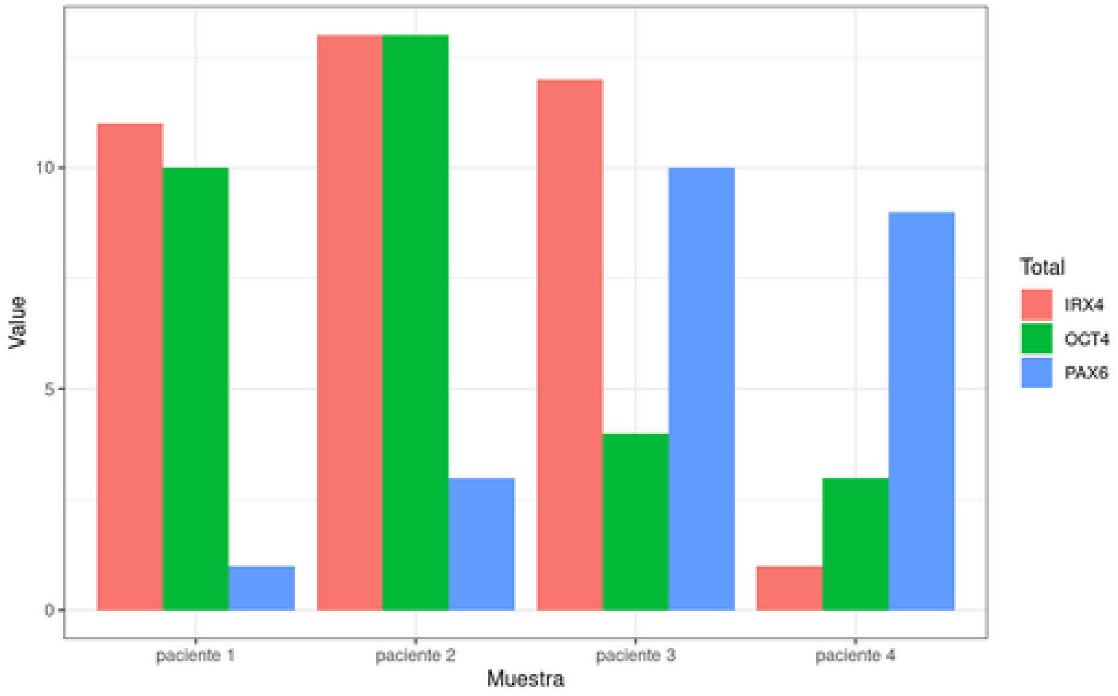
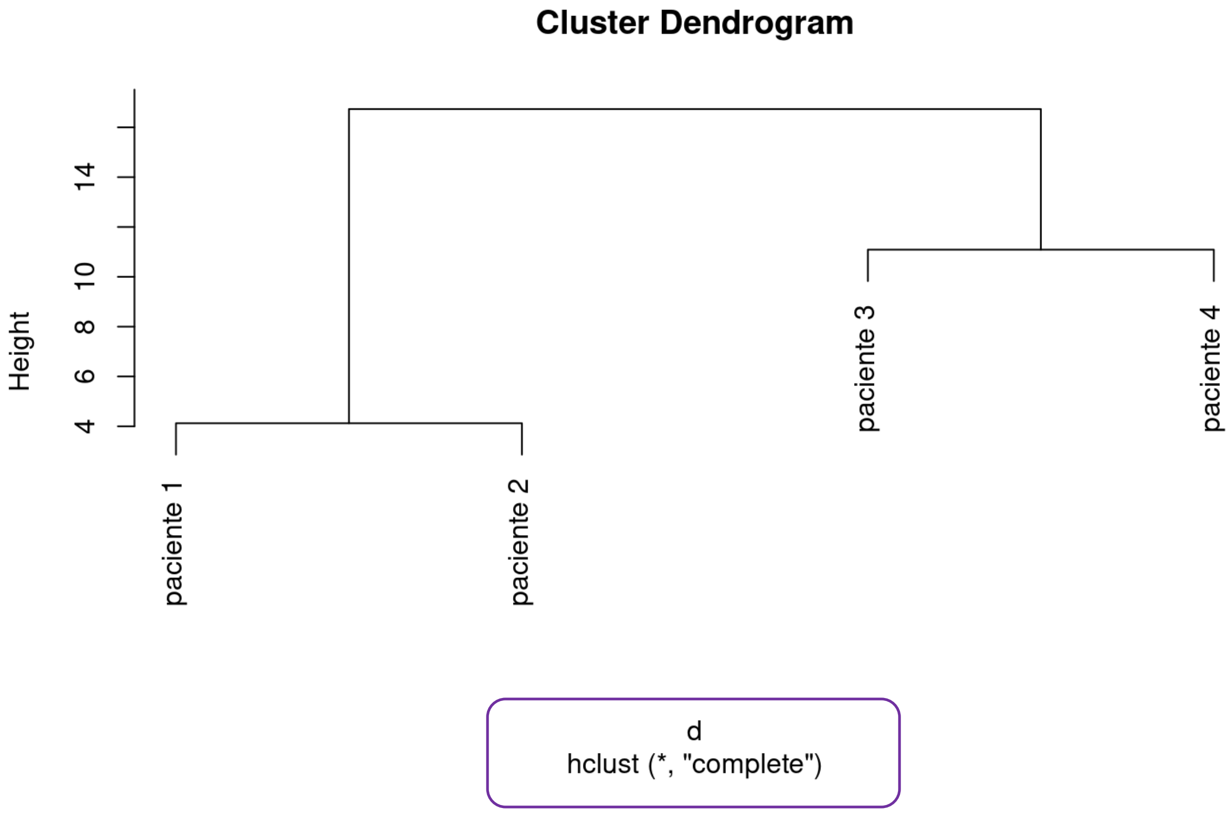


Figure 1: Valores de expresión génica para diferentes pacientes
Paciente 1 y paciente 2 tienen valores de expresión genética similares.

# Cálculo de las **distancias** entre observaciones

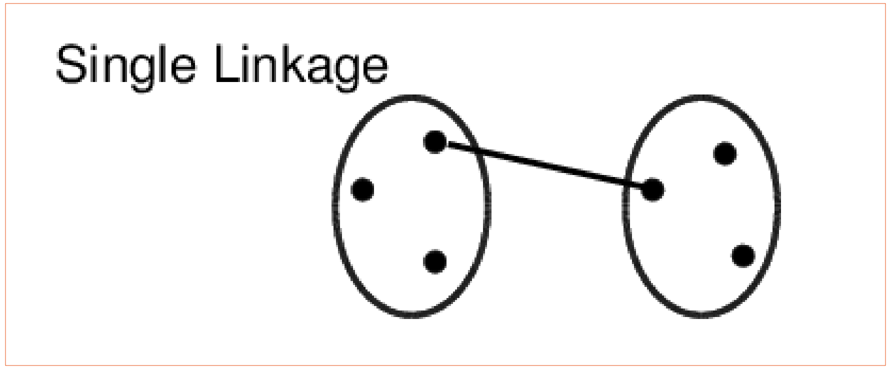
**Definir y cuantificar la similitud/distancia entre las observaciones**

- **Distancia euclídea**
- Distancia de Manhattan
- Índice de Jaccard
- Correlación de Pearson
- Correlación de Spearman …

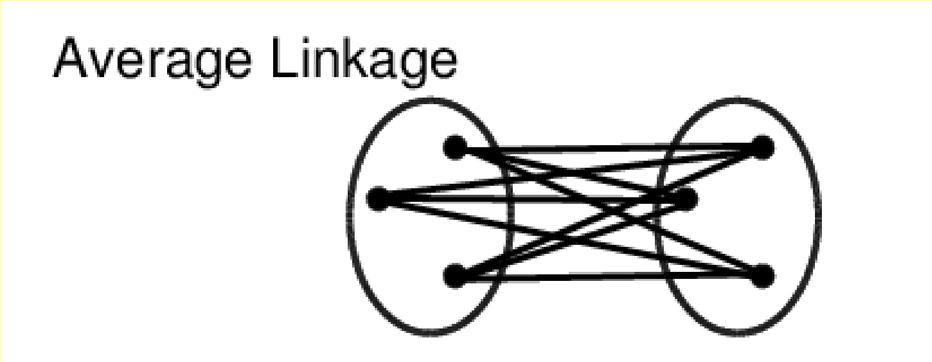


# **Linkage**: métodos de aglomeración

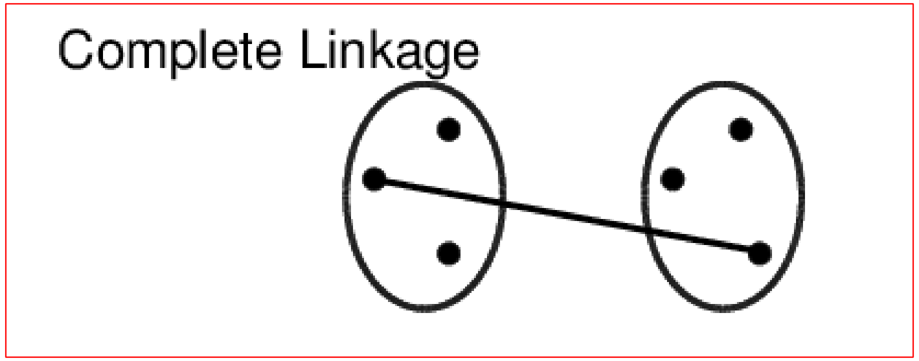
Estrategias que pueden ser empleadas a la hora de unir los *clusters* en las diversas etapas o niveles de un procedimiento jerárquico



**Single linkage: la distancia de similitud viene dada por la mínima distancia entre sus componentes**



**No hay ningún método bueno o malo, no hay ningún método que sea el más óptimo, por tanto lo mejor es hacerlo con varios y comparar la robustez**



**Complete linkage: viene determinado por la similitud de sus elementos más dispares, aquellos genes que se encuentran a una mayor distancia**

# Agrupación jerárquica aglomerativa

1. Inicialización

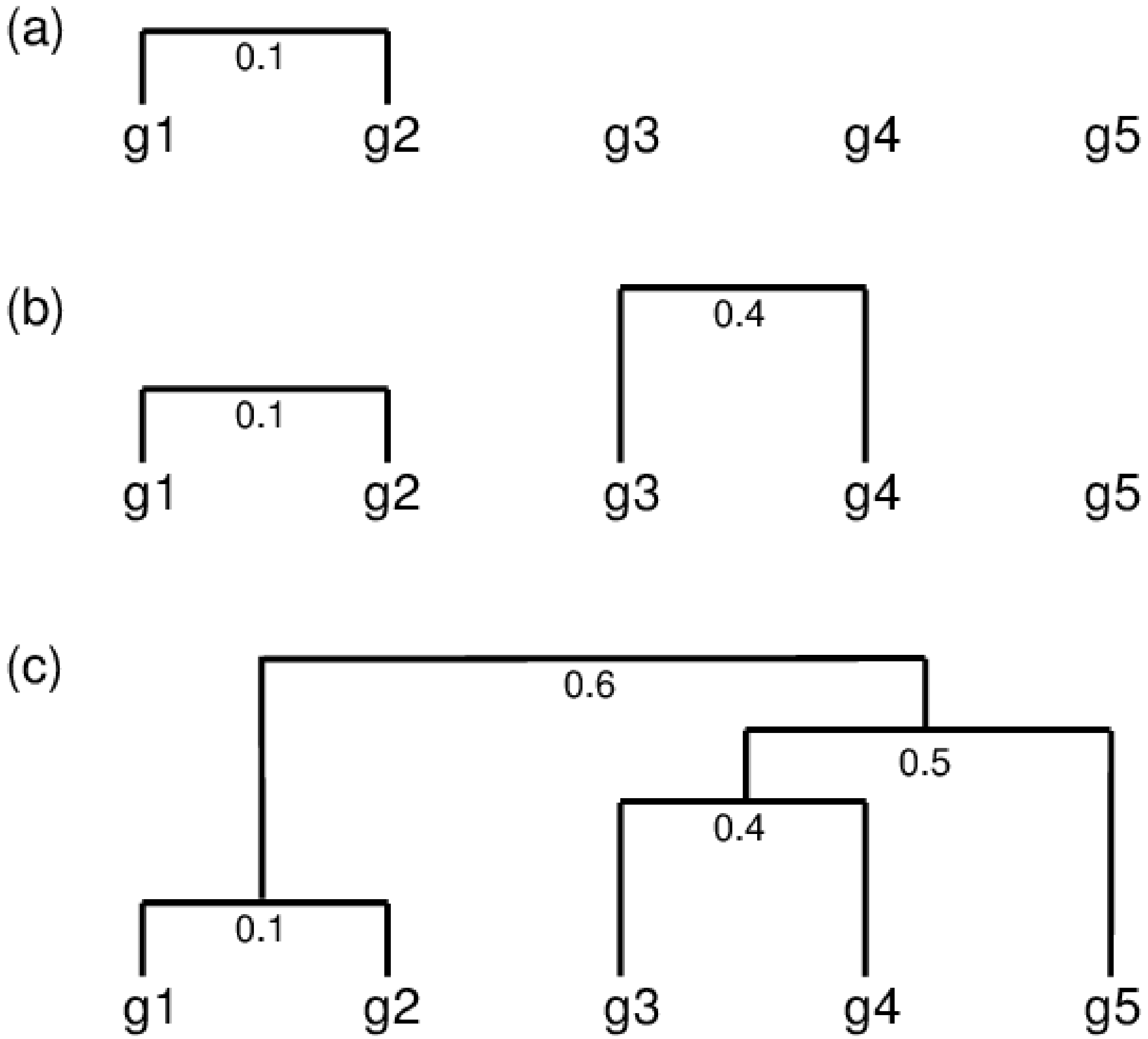
2. Cálculo de distancias

Medidas de distancias

3. Fusión de clusters

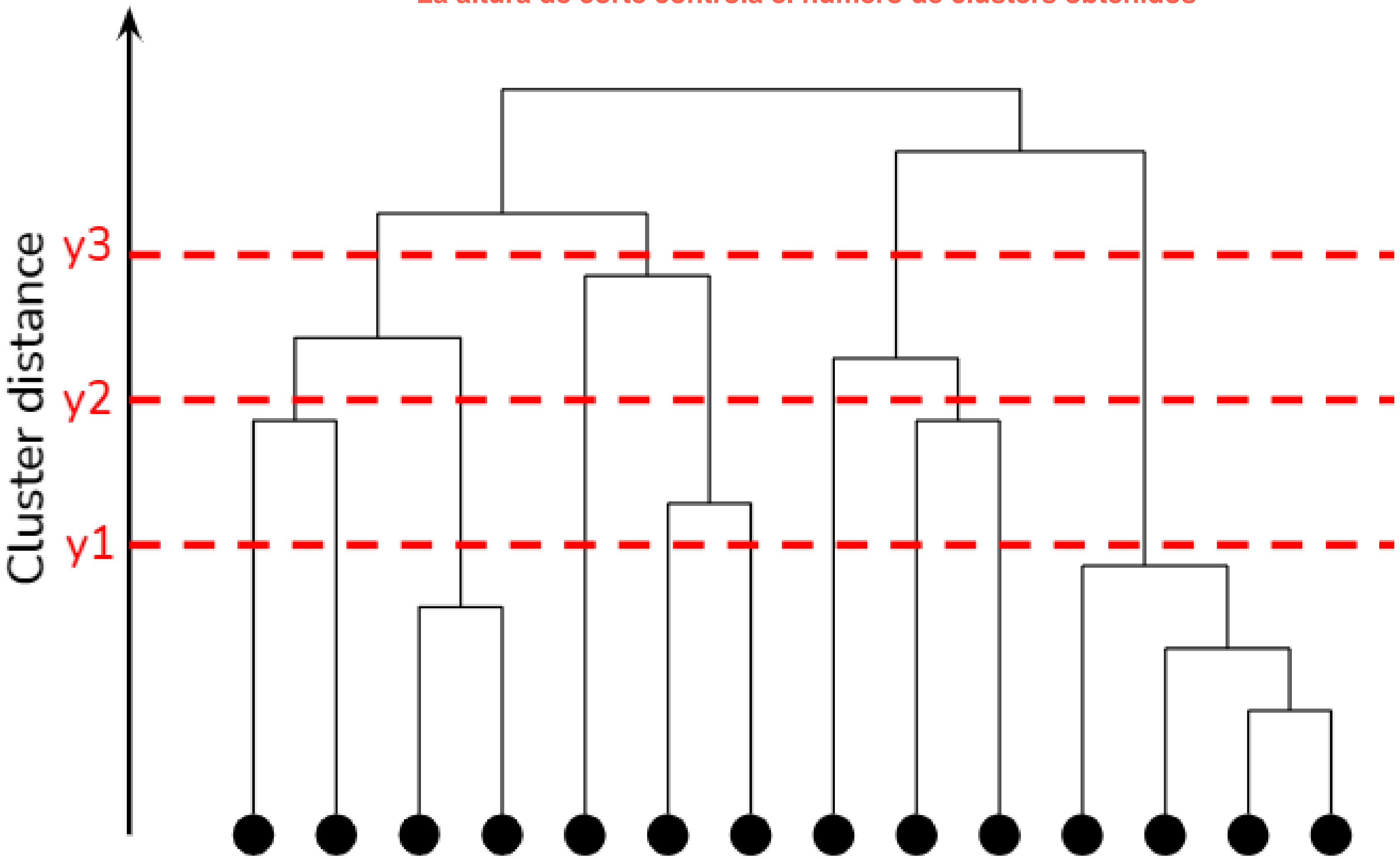
Métodos de aglomeramiento (*linkage*)

4. **Dendrograma** final

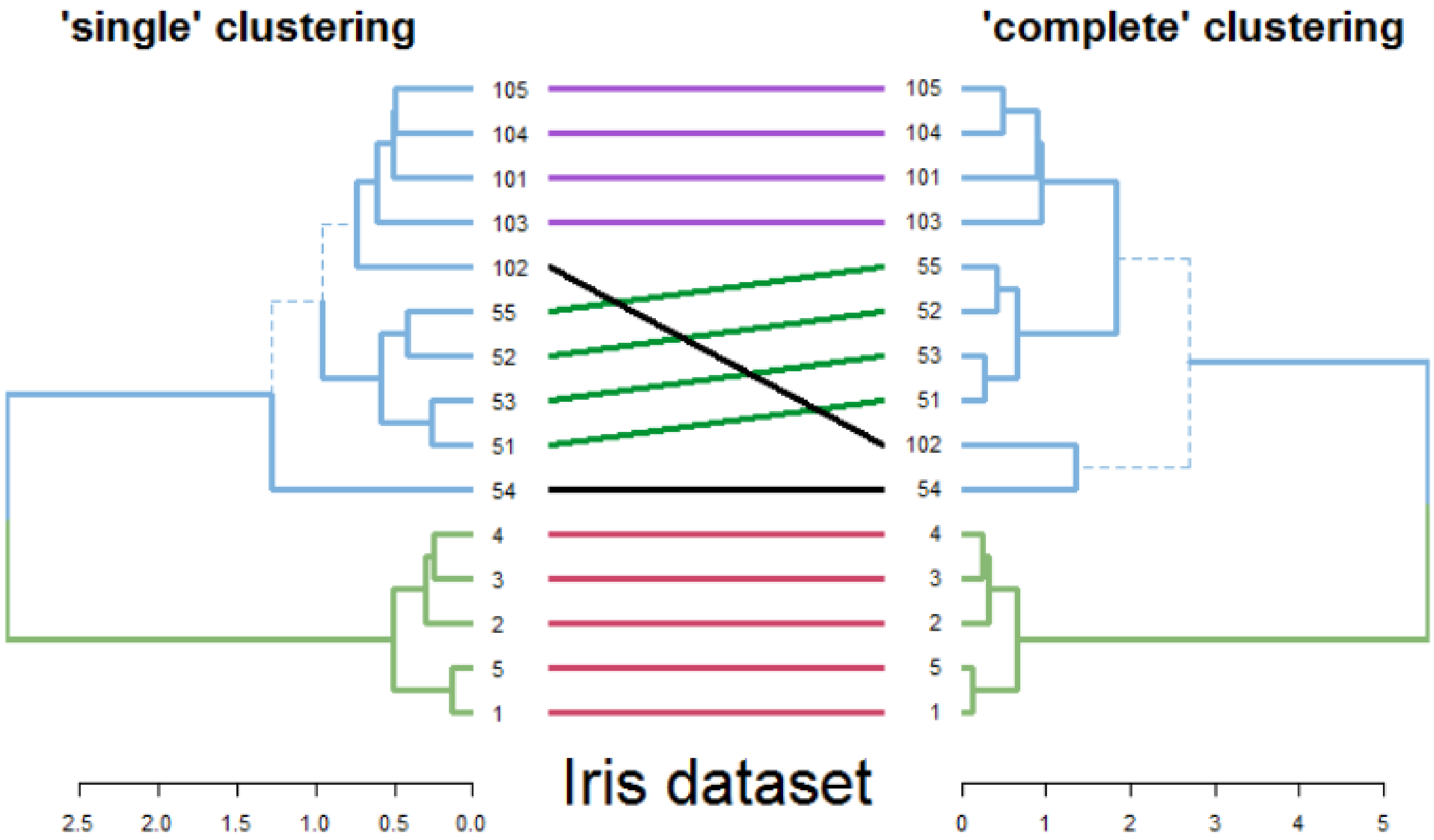


# Dendrograma

La altura de corte controla el número de clusters obtenidos

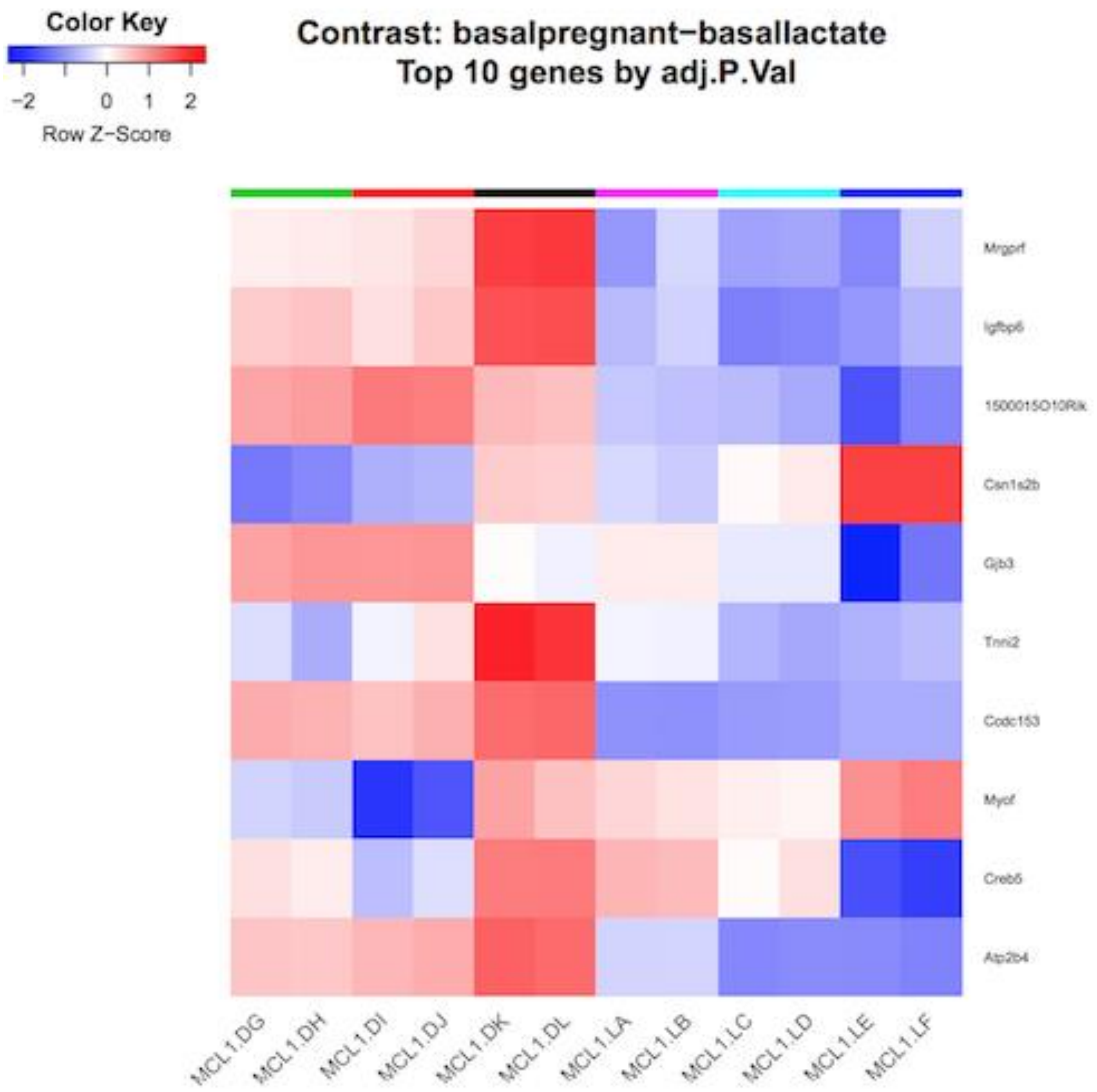


# Comparación de dendrogramas



***Función: tanglegram***

# Heatmap: Visualización de la expresión de DEG

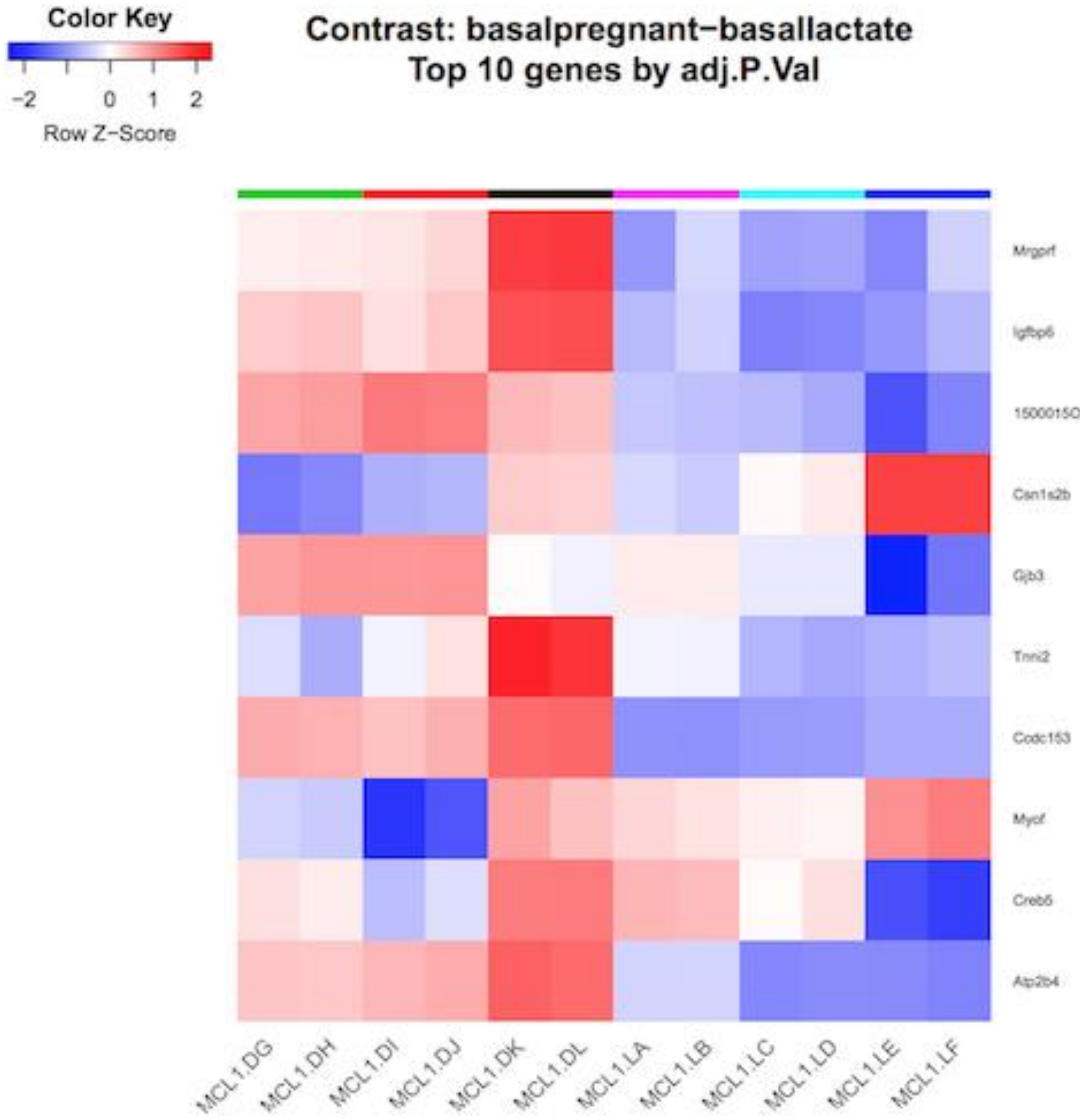


1 Extracción de los **conteos normalizados** para estos genes

**Colores frios son genes que estarán menos expresados, mientras que los colores cálidos son genes que están mas expresados**

2 **Escalado**
Calcular el **Z-score** de los conteos normalizados

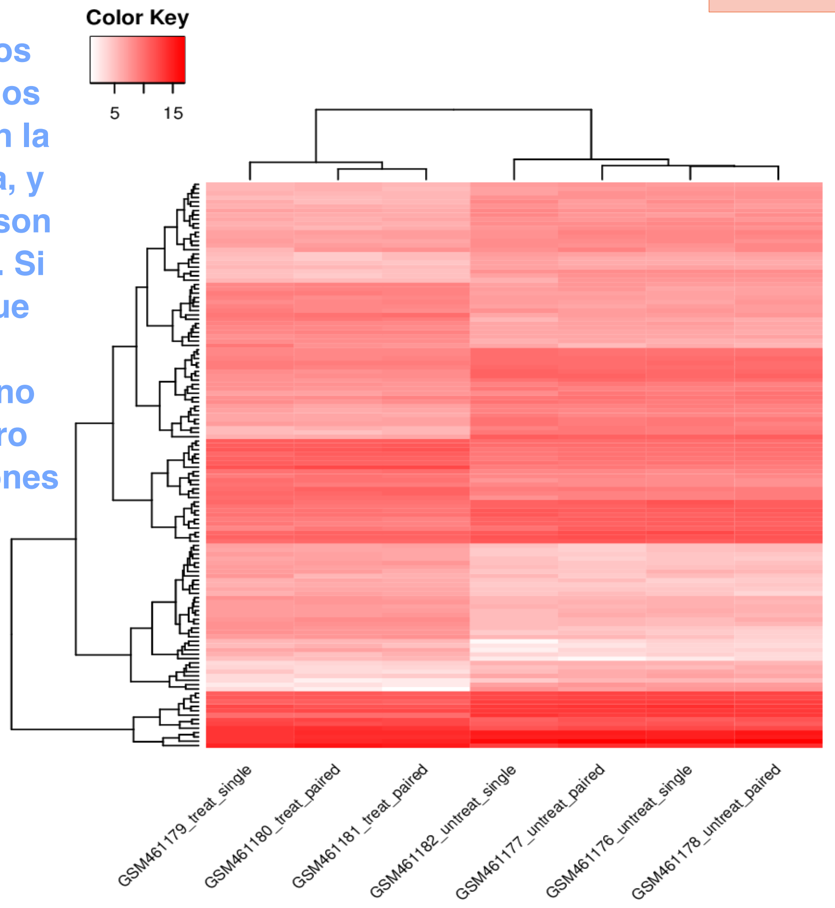
# Heatmap: Visualización de la expresión de DEG



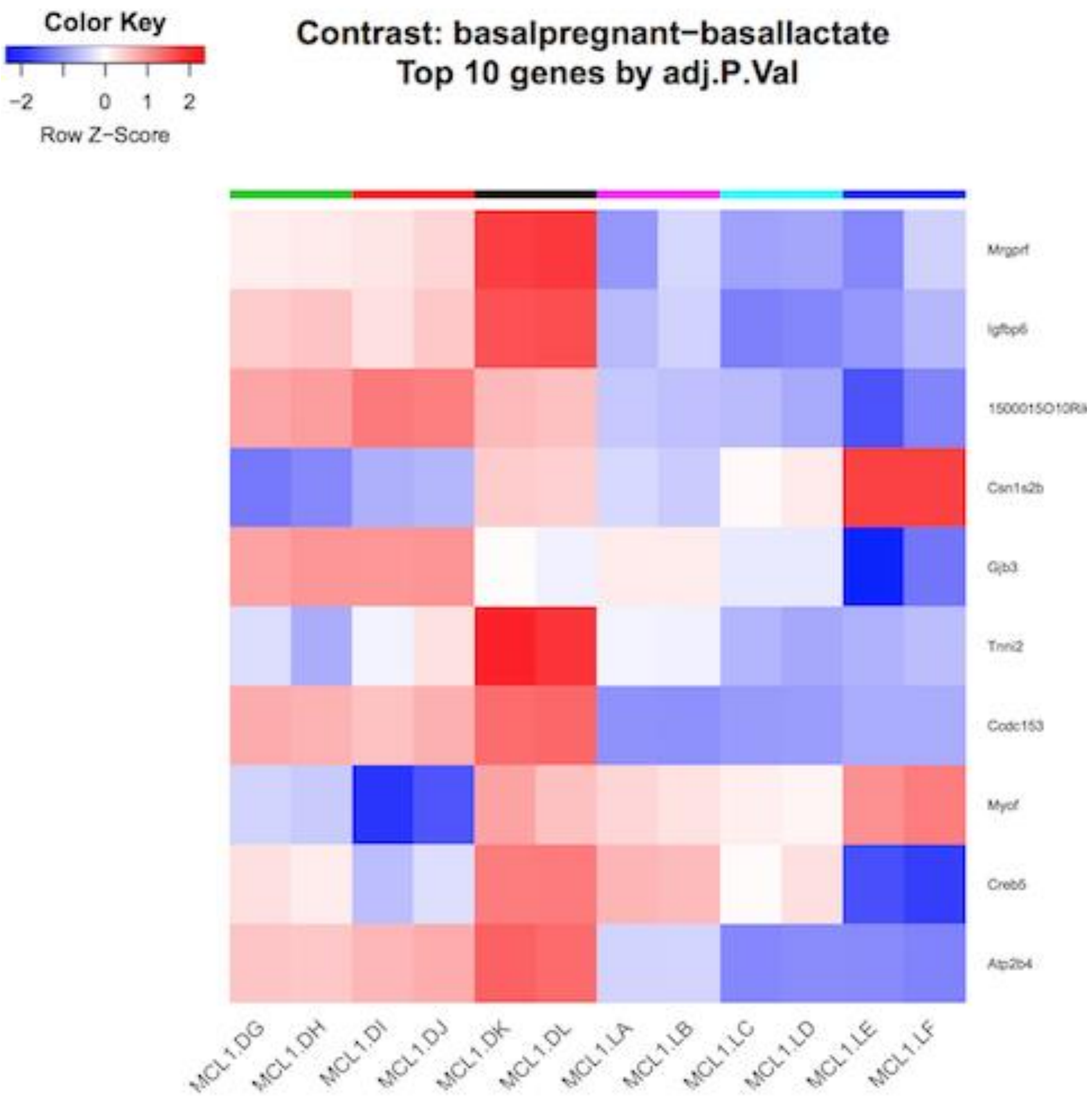
1 Extracción de los **conteos normalizados** para estos genes

**logCPM**

**Cuando escalamos los datos todos los genes tendrán la misma escala, y por tanto, ya son comparables. Si no, genes que están muy expresados no dejan ver otro tipo de relaciones**



# Heatmap: Visualización de la expresión de DEG



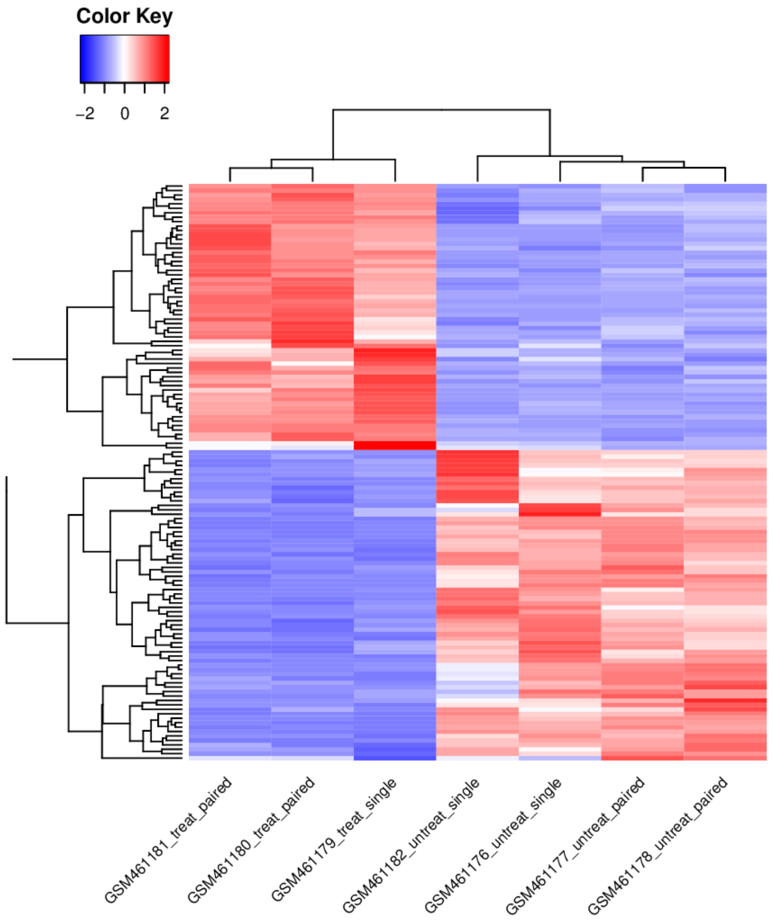
2

**Escalado**

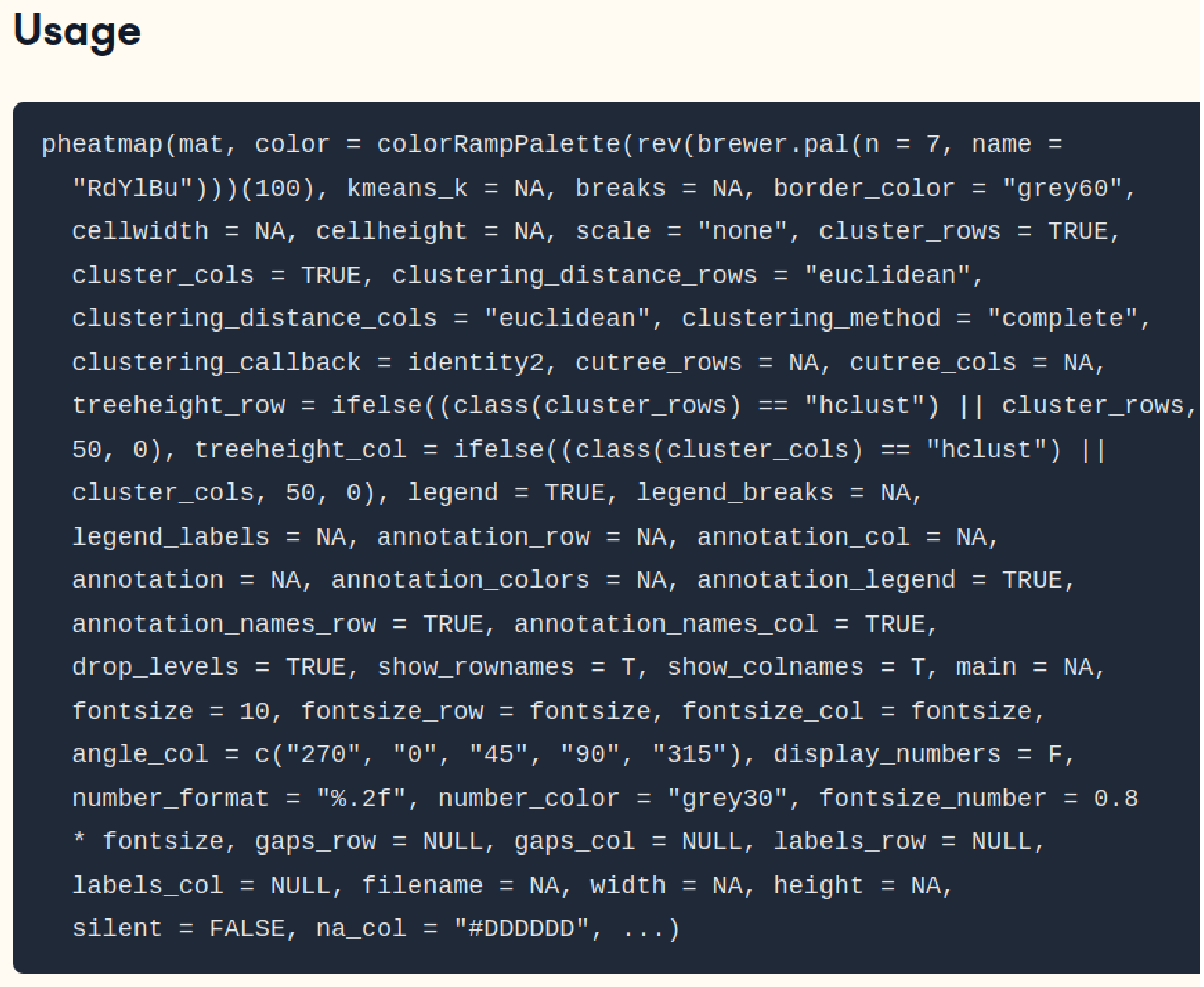
Calcular el **Z-score** de los conteos normalizados

***Scale function***

**El concepto de escalado es transformar el logCPM en Zscore, que va a calcular la media y la desviación estándar de un gen a lo largo de todas las muestras**



# Heatmap: Visualización de la expresión de DEG

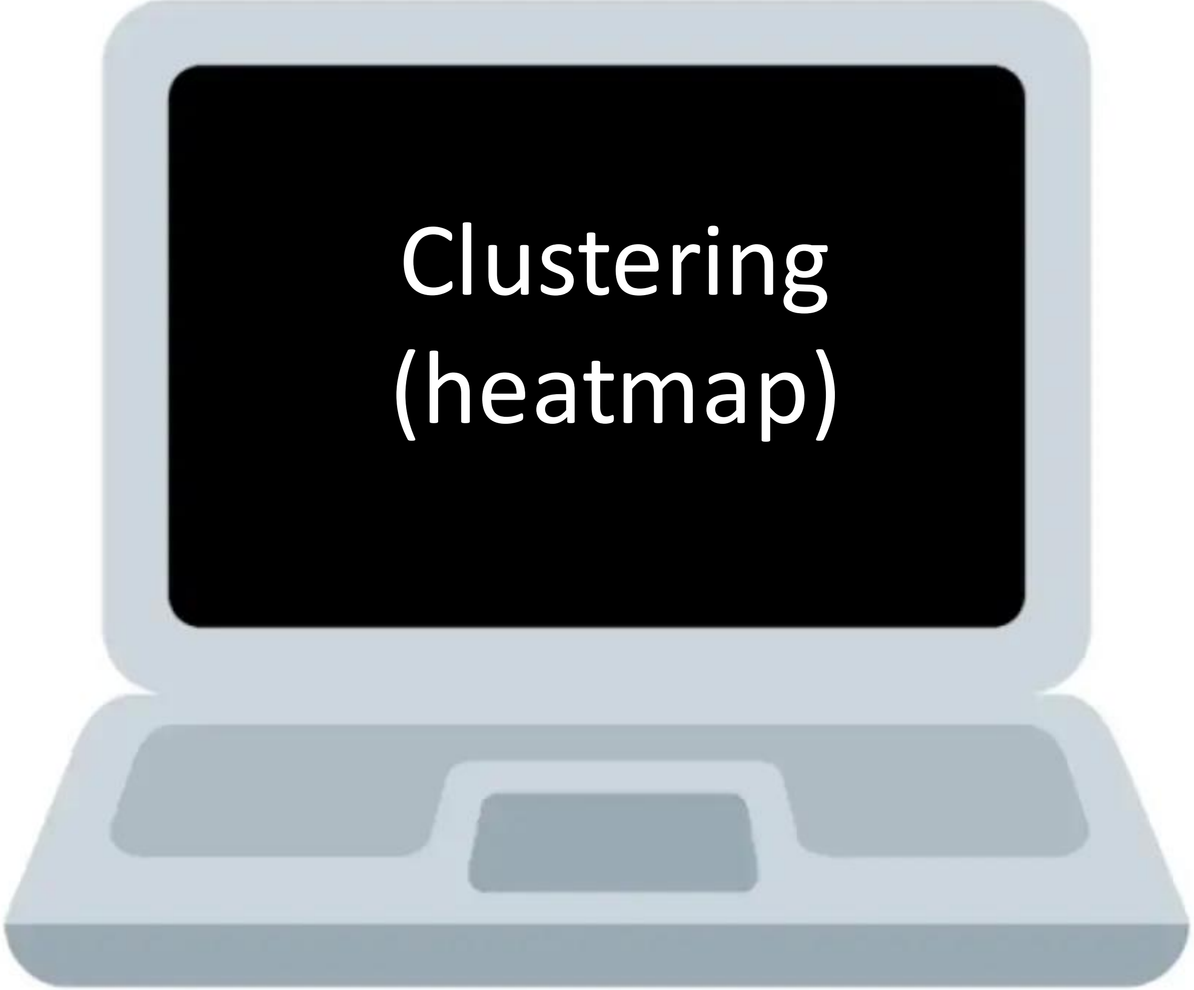


## Usage

```
pheatmap(mat, color = colorRampPalette(rev(brewer.pal(n = 7, name =
  "RdYlBu")))(100), kmeans_k = NA, breaks = NA, border_color = "grey60",
  cellwidth = NA, cellheight = NA, scale = "none", cluster_rows = TRUE,
  cluster_cols = TRUE, clustering_distance_rows = "euclidean",
  clustering_distance_cols = "euclidean", clustering_method = "complete",
  clustering_callback = identity2, cutree_rows = NA, cutree_cols = NA,
  treeheight_row = ifelse((class(cluster_rows) == "hclust") || cluster_rows,
  50, 0), treeheight_col = ifelse((class(cluster_cols) == "hclust") ||
  cluster_cols, 50, 0), legend = TRUE, legend_breaks = NA,
  legend_labels = NA, annotation_row = NA, annotation_col = NA,
  annotation = NA, annotation_colors = NA, annotation_legend = TRUE,
  annotation_names_row = TRUE, annotation_names_col = TRUE,
  drop_levels = TRUE, show_rownames = T, show_colnames = T, main = NA,
  fontsize = 10, fontsize_row = fontsize, fontsize_col = fontsize,
  angle_col = c("270", "0", "45", "90", "315"), display_numbers = F,
  number_format = "%.2f", number_color = "grey30", fontsize_number = 0.8
  * fontsize, gaps_row = NULL, gaps_col = NULL, labels_row = NULL,
  labels_col = NULL, filename = NA, width = NA, height = NA,
  silent = FALSE, na_col = "#DDDDDD", ...)
```

# PRACTIQUEMOS

amazon WorkSpaces



Clustering (heatmap)

# Heatmap: Visualización de la expresión de DEG

```
# 1. Transform the data
logcpm <- cpm(y, log = TRUE)
rownames(logcpm) <- y$genes$SYMBOL
colnames(logcpm) <- paste(y$samples$group, 1:2, sep="-")
head(logcpm)

# 2. Selection of the top genes
DEG <- res_corrected$table[res_corrected$table$FDR <= 0.01 &
                   abs(res_corrected$table$logFC) >= 3.5 ,]

DEG_selection <- logcpm[na.omit(DEG$SYMBOL),]

# 3. Creation of pheatmap.
pheatmap(DEG_selection, scale = "row",
       cluster_rows = T,
       cluster_cols = T,
       clustering_distance_rows = "euclidean",
       clustering_distance_cols = "euclidean",
       clustering_method = "ward.D2",
       cutree_cols = 1, cutree_rows = 2,
       display_numbers = T, fontsize_number = 6, fontsize_row = 7, border_color = NA)
```

# Ontología Génica

# Ontología génica

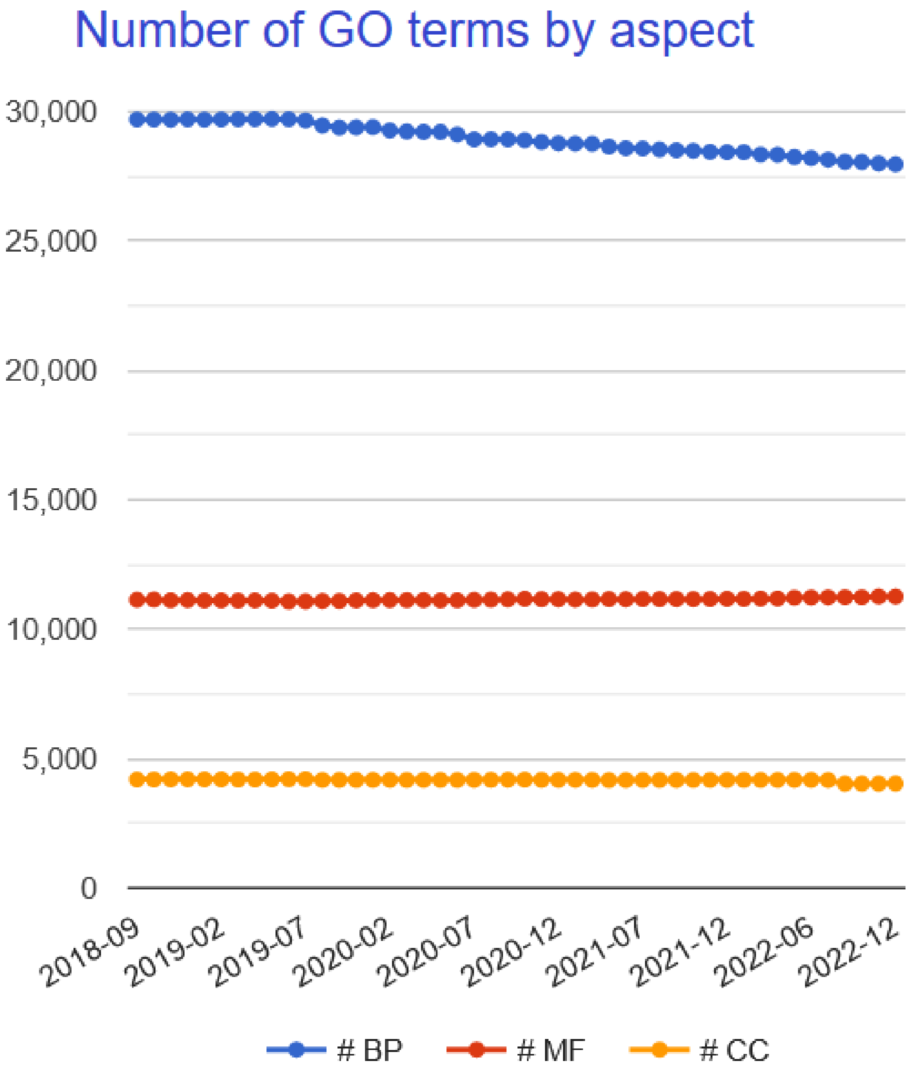
La **ontología génica** (**GO**, por sus siglas en inglés) es el conjunto de términos o de vocabulario estructurado creado para **describir y categorizar los genes y sus productos**.

Actualmente, la mayoría de las principales bases de datos de plantas, animales y microorganismos forman parte del proyecto, y **un total de 45.003 términos** —aplicables a una amplia variedad de organismos biológicos— aparecen registrados en su base de datos.

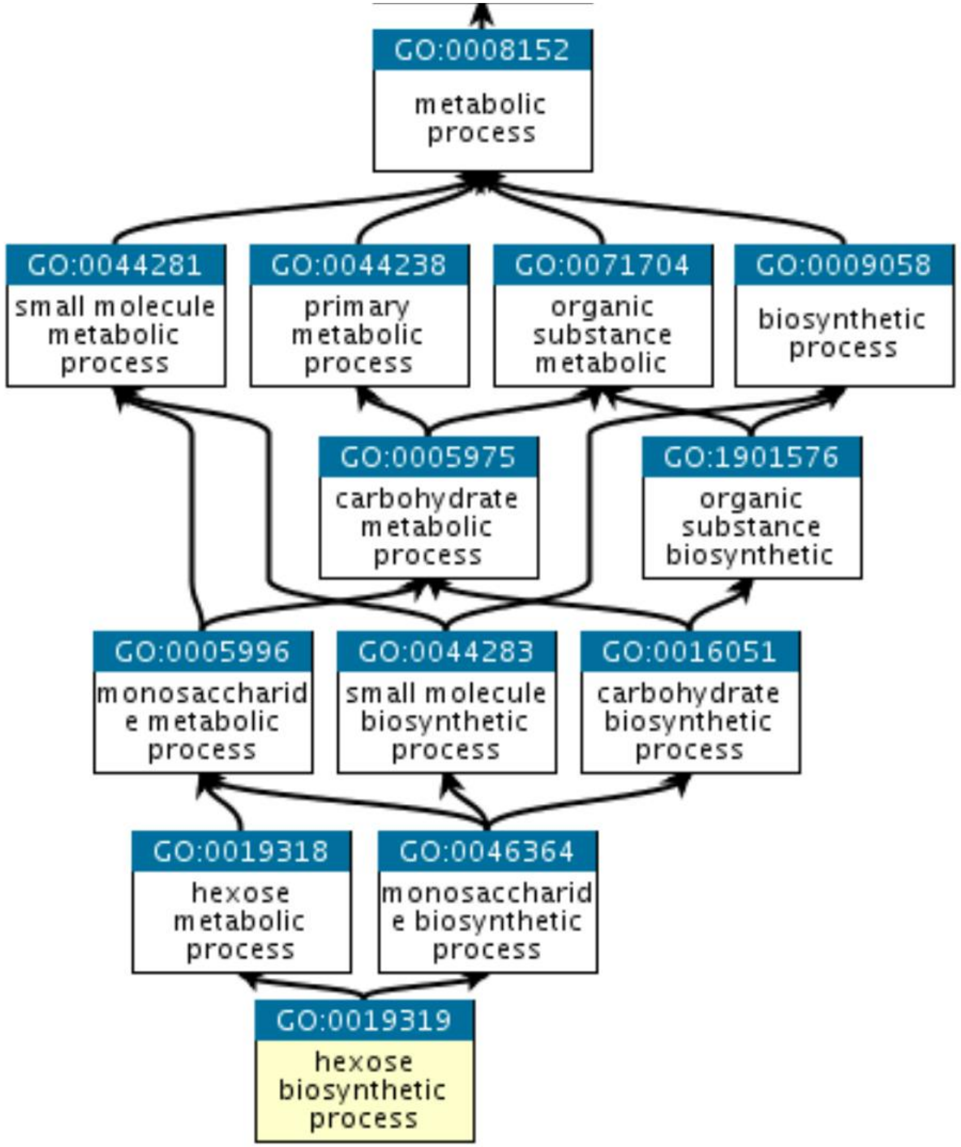
# Ontología génica se divide en:

## Ontología

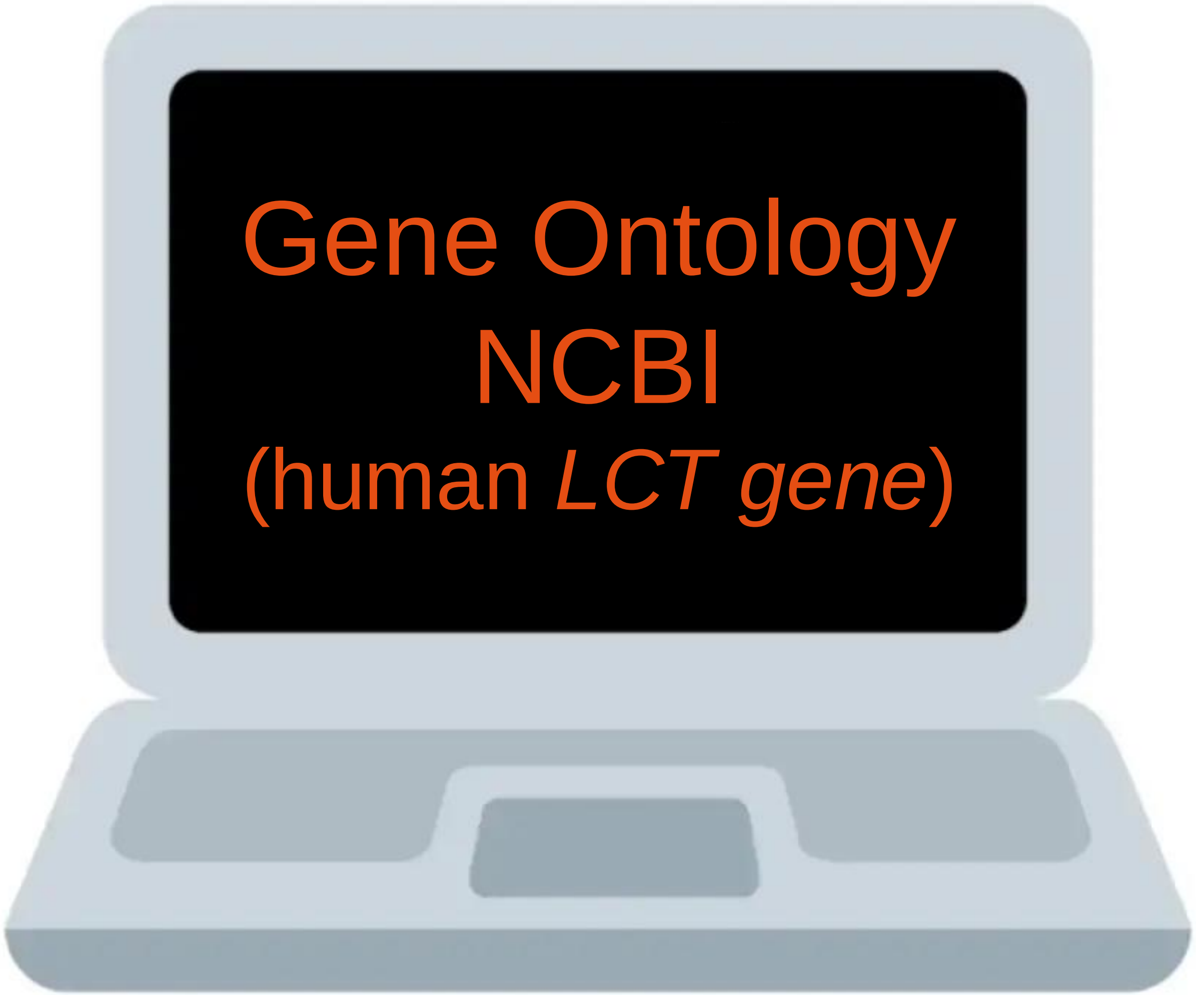
- La función molecular de los productos génicos.
- Su participación en los procesos biológicos.
- Su localización celular.



# Ontología génica


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

# PRACTIQUEMOS

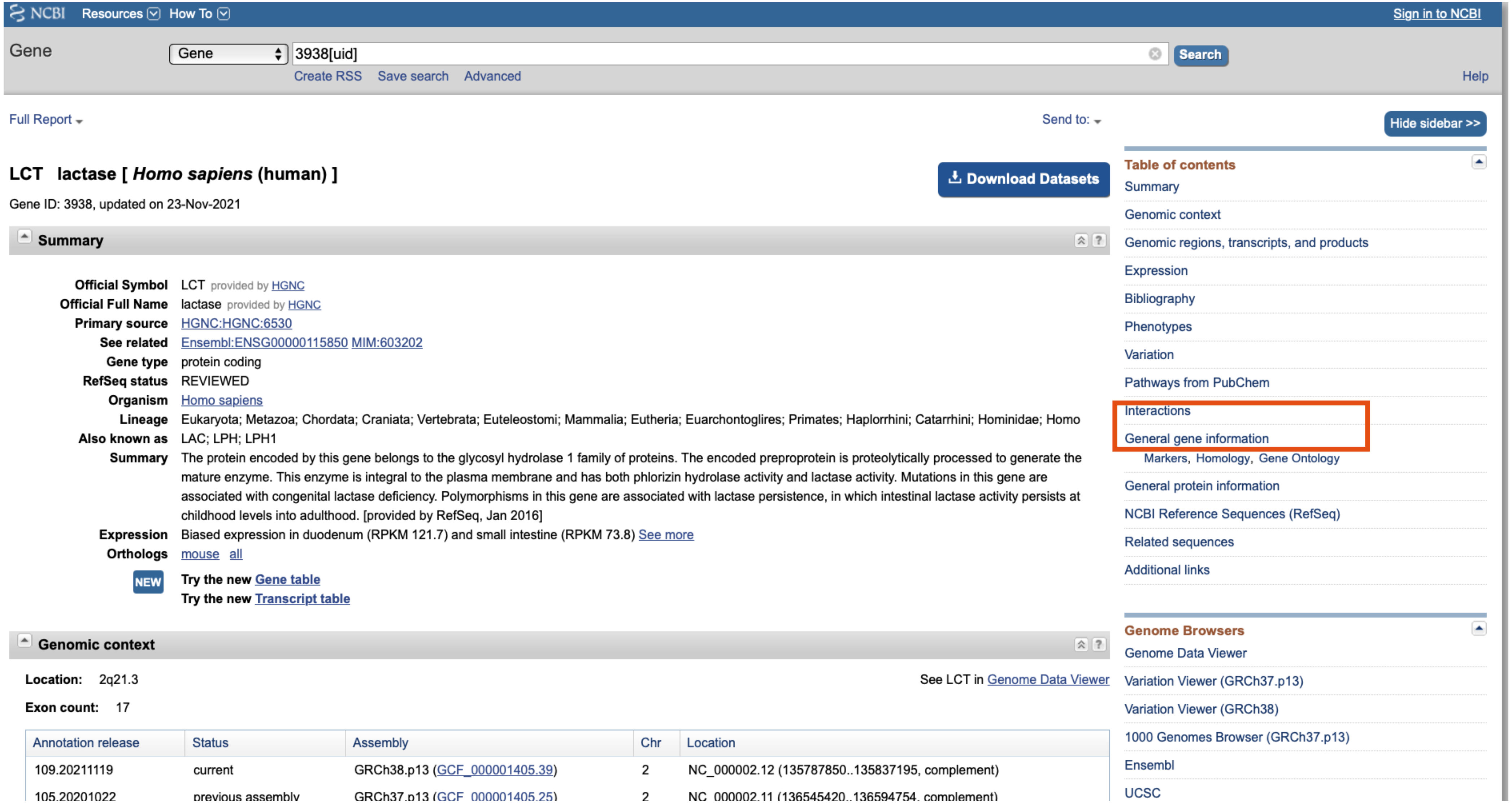


Gene Ontology

NCBI

(human *LCT gene*)

# Ontología génica-> Ejemplo (gen de la lactasa)



NCBI Resources How To Sign in to NCBI

Gene | Gene | 3938[uid] | Search

Create RSS Save search Advanced Help

Full Report ▾ Send to: ▾ Hide sidebar >>

**LCT lactase [ *Homo sapiens* (human) ]**

Download Datasets

Gene ID: 3938, updated on 23-Nov-2021

## Summary

|  |  |
|---|---|
| **Official Symbol** | LCT provided by HGNC |
| **Official Full Name** | lactase provided by HGNC |
| **Primary source** | HGNC:HGNC:6530 |
| **See related** | Ensembl:ENSG00000115850 MIM:603202 |
| **Gene type** | protein coding |
| **RefSeq status** | REVIEWED |
| **Organism** | Homo sapiens |
| **Lineage** | Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo |
| **Also known as** | LAC; LPH; LPH1 |
| **Summary** | The protein encoded by this gene belongs to the glycosyl hydrolase 1 family of proteins. The encoded preproprotein is proteolytically processed to generate the mature enzyme. This enzyme is integral to the plasma membrane and has both phlorizin hydrolase activity and lactase activity. Mutations in this gene are associated with congenital lactase deficiency. Polymorphisms in this gene are associated with lactase persistence, in which intestinal lactase activity persists at childhood levels into adulthood. [provided by RefSeq, Jan 2016] |
| **Expression** | Biased expression in duodenum (RPKM 121.7) and small intestine (RPKM 73.8) See more |
| **Orthologs** | mouse all |

NEW **Try the new Gene table**
**Try the new Transcript table**

## Genomic context

**Location:** 2q21.3 See LCT in Genome Data Viewer

**Exon count:** 17

| Annotation release | Status | Assembly | Chr | Location |
|---|---|---|---|---|
| 109.20211119 | current | GRCh38.p13 (GCF_000001405.39) | 2 | NC_000002.12 (135787850..135837195, complement) |
| 105.20201022 | previous assembly | GRCh37.p13 (GCF_000001405.25) | 2 | NC_000002.11 (136545420..136594754, complement) |

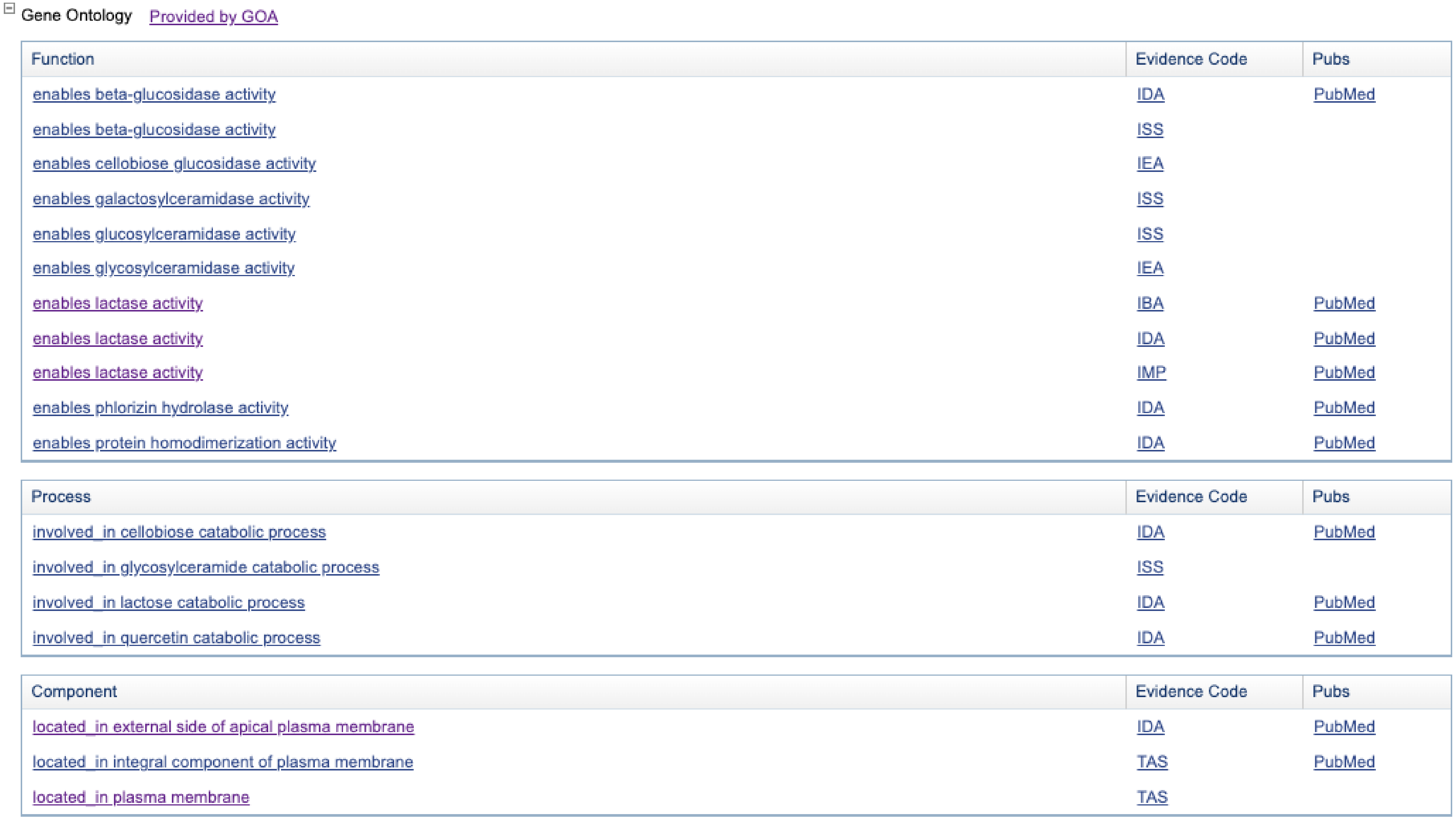
**Table of contents**

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Expression
- Bibliography
- Phenotypes
- Variation
- Pathways from PubChem
- Interactions
- General gene information
  - Markers, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

**Genome Browsers**

- Genome Data Viewer
- Variation Viewer (GRCh37.p13)
- Variation Viewer (GRCh38)
- 1000 Genomes Browser (GRCh37.p13)
- Ensembl
- UCSC

# Ontología génica-> ejemplo



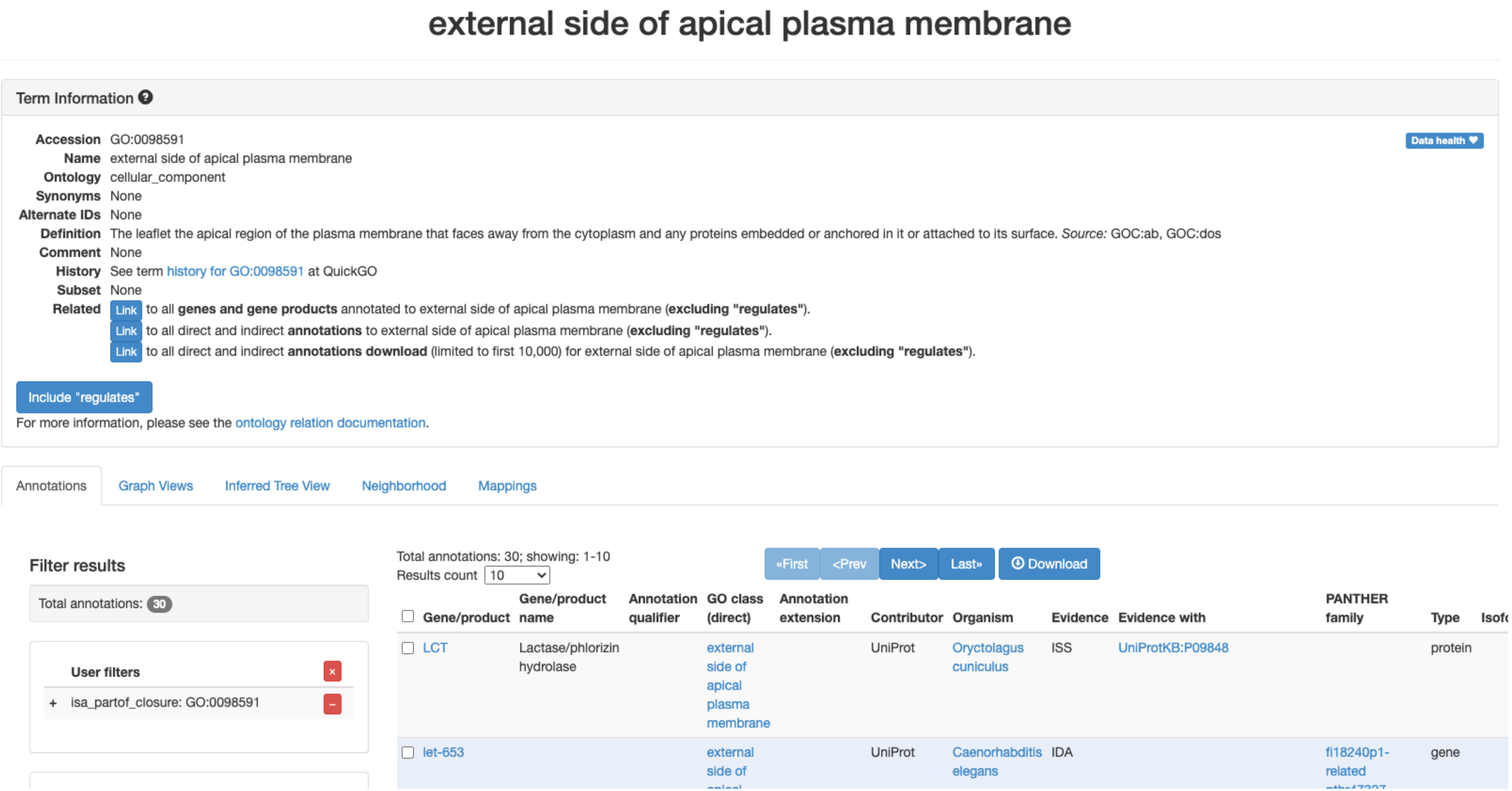
⊟ Gene Ontology Provided by GOA

| Function | Evidence Code | Pubs |
|---|---|---|
| enables beta-glucosidase activity | IDA | PubMed |
| enables beta-glucosidase activity | ISS | |
| enables cellobiose glucosidase activity | IEA | |
| enables galactosylceramidase activity | ISS | |
| enables glucosylceramidase activity | ISS | |
| enables glycosylceramidase activity | IEA | |
| enables lactase activity | IBA | PubMed |
| enables lactase activity | IDA | PubMed |
| enables lactase activity | IMP | PubMed |
| enables phlorizin hydrolase activity | IDA | PubMed |
| enables protein homodimerization activity | IDA | PubMed |

| Process | Evidence Code | Pubs |
|---|---|---|
| involved_in cellobiose catabolic process | IDA | PubMed |
| involved_in glycosylceramide catabolic process | ISS | |
| involved_in lactose catabolic process | IDA | PubMed |
| involved_in quercetin catabolic process | IDA | PubMed |

| Component | Evidence Code | Pubs |
|---|---|---|
| located_in external side of apical plasma membrane | IDA | PubMed |
| located_in integral component of plasma membrane | TAS | PubMed |
| located_in plasma membrane | TAS | |

# Ontología génica-> ejemplo



## external side of apical plasma membrane

### Term Information

Data health

**Accession** GO:0098591

**Name** external side of apical plasma membrane

**Ontology** cellular_component

**Synonyms** None

**Alternate IDs** None

**Definition** The leaflet the apical region of the plasma membrane that faces away from the cytoplasm and any proteins embedded or anchored in it or attached to its surface. *Source:* GOC:ab, GOC:dos

**Comment** None

**History** See term history for GO:0098591 at QuickGO

**Subset** None

**Related** Link to all **genes and gene products** annotated to external side of apical plasma membrane (**excluding "regulates"**).
Link to all direct and indirect **annotations** to external side of apical plasma membrane (**excluding "regulates"**).
Link to all direct and indirect **annotations download** (limited to first 10,000) for external side of apical plasma membrane (**excluding "regulates"**).

Include "regulates"

For more information, please see the ontology relation documentation.

**Filter results**

Total annotations: 30

**User filters**

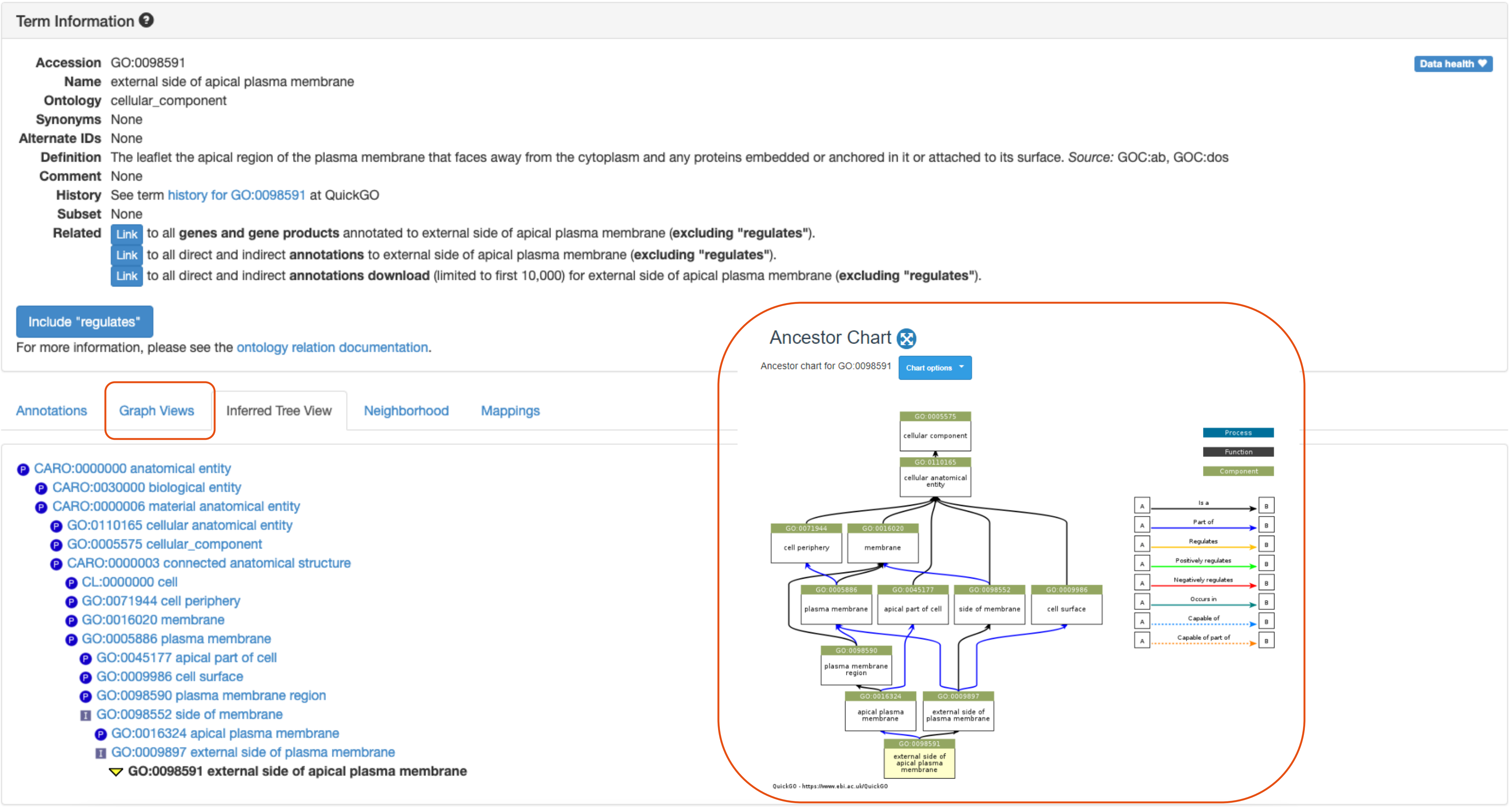
+ isa_partof_closure: GO:0098591

Total annotations: 30; showing: 1-10

Results count 10

| □ | Gene/product | Gene/product name | Annotation qualifier | GO class (direct) | Annotation extension | Contributor | Organism | Evidence | Evidence with | PANTHER family | Type | Isof |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| □ | LCT | Lactase/phlorizin hydrolase | | external side of apical plasma membrane | | UniProt | Oryctolagus cuniculus | ISS | UniProtKB:P09848 | | protein | |
| □ | let-653 | | | external side of | | UniProt | Caenorhabditis elegans | IDA | | fi18240p1-related | gene | |

# Ontología génica-> ejemplo



Term Information

Data health

| | |
|---|---|
| **Accession** | GO:0098591 |
| **Name** | external side of apical plasma membrane |
| **Ontology** | cellular_component |
| **Synonyms** | None |
| **Alternate IDs** | None |
| **Definition** | The leaflet the apical region of the plasma membrane that faces away from the cytoplasm and any proteins embedded or anchored in it or attached to its surface. *Source:* GOC:ab, GOC:dos |
| **Comment** | None |
| **History** | See term history for GO:0098591 at QuickGO |
| **Subset** | None |
| **Related** | Link to all **genes and gene products** annotated to external side of apical plasma membrane (**excluding "regulates"**).<br>Link to all direct and indirect **annotations** to external side of apical plasma membrane (**excluding "regulates"**).<br>Link to all direct and indirect **annotations download** (limited to first 10,000) for external side of apical plasma membrane (**excluding "regulates"**). |

Include "regulates"

For more information, please see the ontology relation documentation.

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings

- P CARO:0000000 anatomical entity
  - P CARO:0030000 biological entity
  - P CARO:0000006 material anatomical entity
    - P GO:0110165 cellular anatomical entity
    - P GO:0005575 cellular_component
    - P CARO:0000003 connected anatomical structure
      - P CL:0000000 cell
      - P GO:0071944 cell periphery
      - P GO:0016020 membrane
      - P GO:0005886 plasma membrane
        - P GO:0045177 apical part of cell
        - P GO:0009986 cell surface
        - P GO:0098590 plasma membrane region
        - I GO:0098552 side of membrane
          - P GO:0016324 apical plasma membrane
          - I GO:0009897 external side of plasma membrane
            - ▽ **GO:0098591 external side of apical plasma membrane**

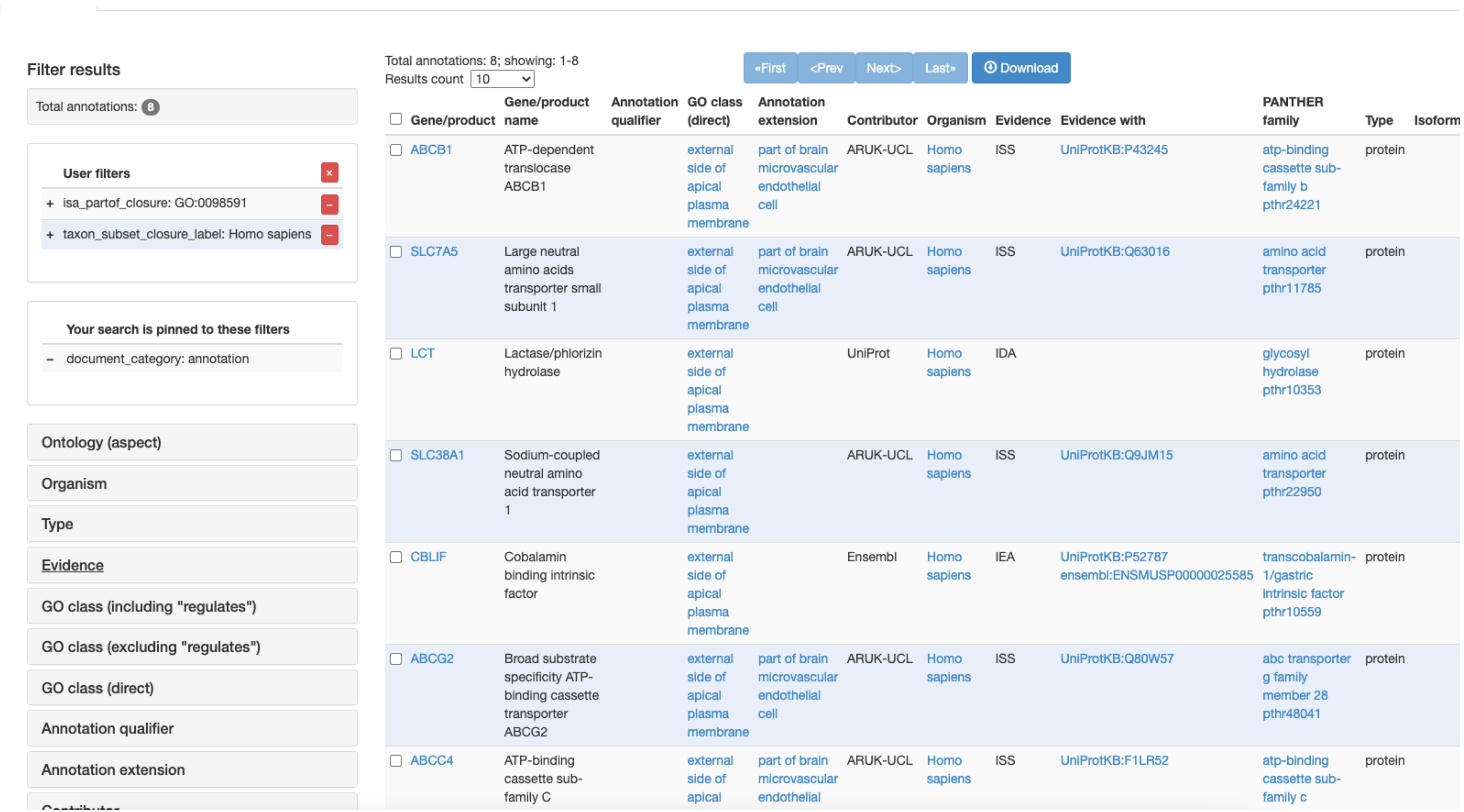
Ancestor Chart

Ancestor chart for GO:0098591

Chart options

QuickGO - https://www.ebi.ac.uk/QuickGO

# Ontología génica-> ejemplo



**Filter results**

Total annotations: 8

**User filters** ×

+ isa_partof_closure: GO:0098591 –

+ taxon_subset_closure_label: Homo sapiens –

**Your search is pinned to these filters**

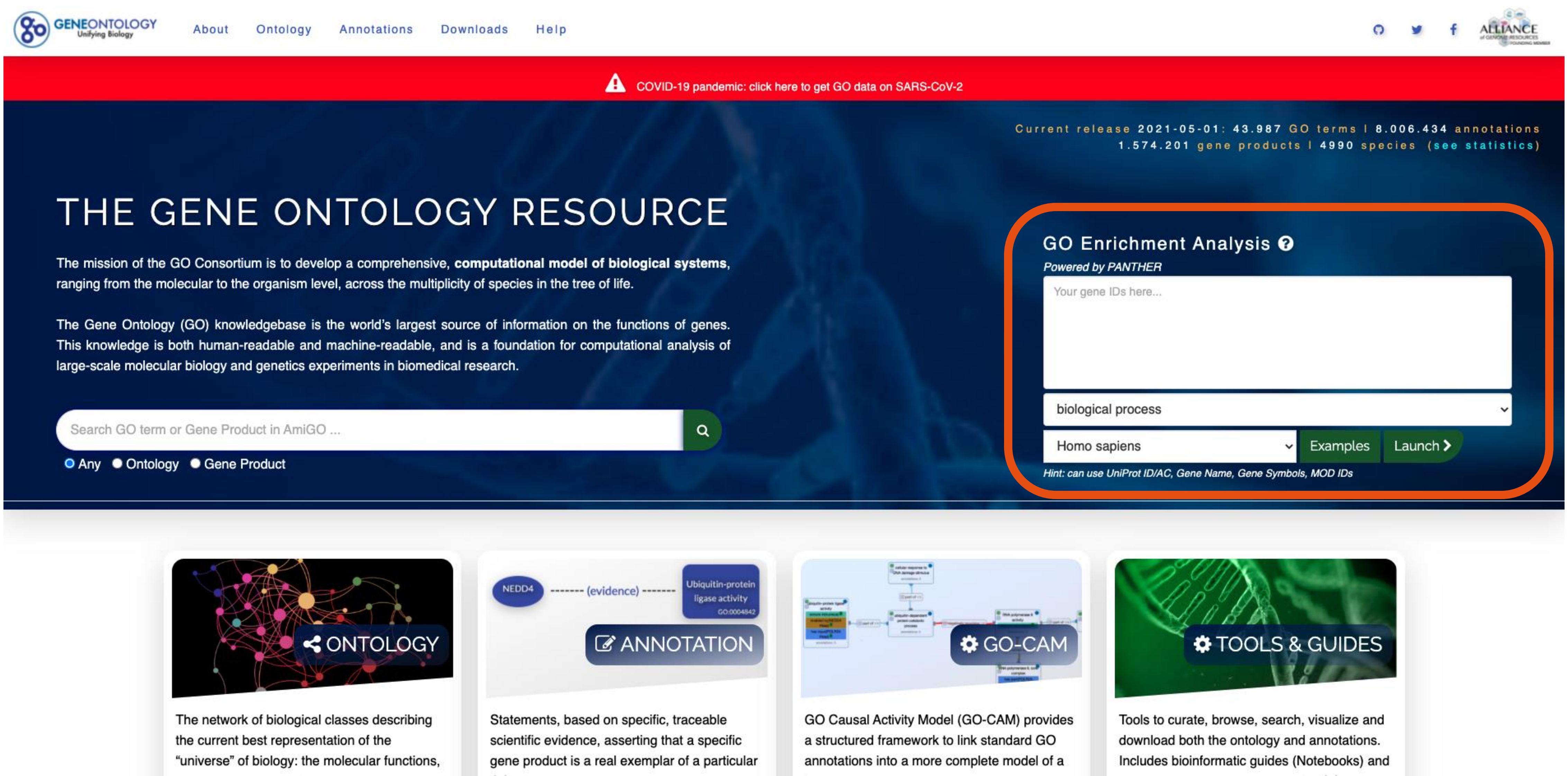
– document_category: annotation

Total annotations: 8; showing: 1-8

Results count 10

| □ Gene/product | Gene/product name | Annotation qualifier | GO class (direct) | Annotation extension | Contributor | Organism | Evidence | Evidence with | PANTHER family | Type | Isoform |
|---|---|---|---|---|---|---|---|---|---|---|---|
| □ ABCB1 | ATP-dependent translocase ABCB1 | | external side of apical plasma membrane | part of brain microvascular endothelial cell | ARUK-UCL | Homo sapiens | ISS | UniProtKB:P43245 | atp-binding cassette sub-family b pthr24221 | protein | |
| □ SLC7A5 | Large neutral amino acids transporter small subunit 1 | | external side of apical plasma membrane | part of brain microvascular endothelial cell | ARUK-UCL | Homo sapiens | ISS | UniProtKB:Q63016 | amino acid transporter pthr11785 | protein | |
| □ LCT | Lactase/phlorizin hydrolase | | external side of apical plasma membrane | | UniProt | Homo sapiens | IDA | | glycosyl hydrolase pthr10353 | protein | |
| □ SLC38A1 | Sodium-coupled neutral amino acid transporter 1 | | external side of apical plasma membrane | | ARUK-UCL | Homo sapiens | ISS | UniProtKB:Q9JM15 | amino acid transporter pthr22950 | protein | |
| □ CBLIF | Cobalamin binding intrinsic factor | | external side of apical plasma membrane | | Ensembl | Homo sapiens | IEA | UniProtKB:P52787<br>ensembl:ENSMUSP00000025585 | transcobalamin-1/gastric intrinsic factor pthr10559 | protein | |
| □ ABCG2 | Broad substrate specificity ATP-binding cassette transporter ABCG2 | | external side of apical plasma membrane | part of brain microvascular endothelial cell | ARUK-UCL | Homo sapiens | ISS | UniProtKB:Q80W57 | abc transporter g family member 28 pthr48041 | protein | |
| □ ABCC4 | ATP-binding cassette sub-family C | | external side of apical | part of brain microvascular endothelial | ARUK-UCL | Homo sapiens | ISS | UniProtKB:F1LR52 | atp-binding cassette sub-family c | protein | |

# Ontología génica *anotación/análisis*



COVID-19 pandemic: click here to get GO data on SARS-CoV-2

Current release 2021-05-01: 43.987 GO terms | 8.006.434 annotations
1.574.201 gene products | 4990 species (see statistics)

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Search GO term or Gene Product in AmiGO ...

● Any ○ Ontology ○ Gene Product

GO Enrichment Analysis

*Powered by PANTHER*

Your gene IDs here...

biological process

Homo sapiens

Examples

Launch

*Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs*

The network of biological classes describing the current best representation of the "universe" of biology: the molecular functions,

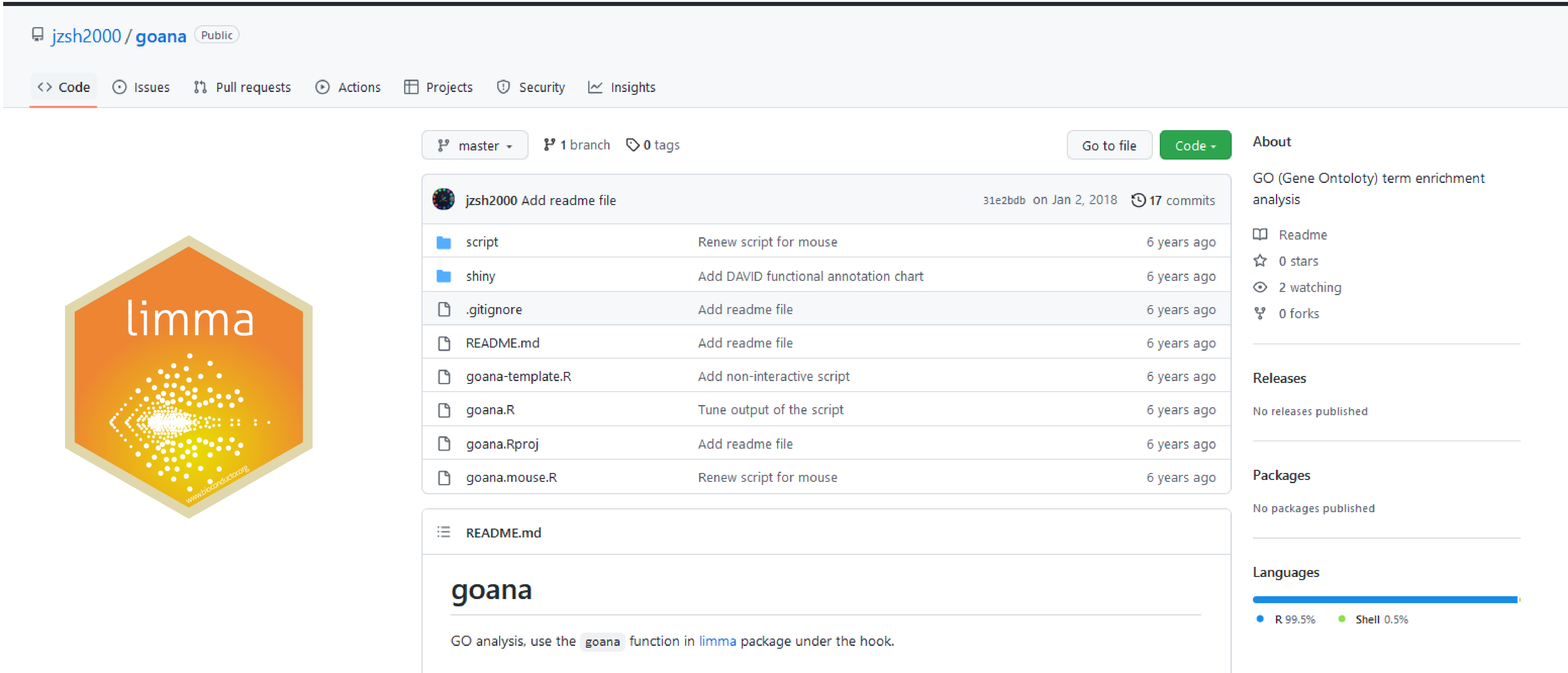
Statements, based on specific, traceable scientific evidence, asserting that a specific gene product is a real exemplar of a particular

GO Causal Activity Model (GO-CAM) provides a structured framework to link standard GO annotations into a more complete model of a

Tools to curate, browse, search, visualize and download both the ontology and annotations. Includes bioinformatic guides (Notebooks) and

http://geneontology.org/

# Análisis de enriquecimiento de Ontología Genética (GO)



jzsh2000 Add readme file 31e2bdb on Jan 2, 2018 17 commits

| | | |
|---|---|---|
| script | Renew script for mouse | 6 years ago |
| shiny | Add DAVID functional annotation chart | 6 years ago |
| .gitignore | Add readme file | 6 years ago |
| README.md | Add readme file | 6 years ago |
| goana-template.R | Add non-interactive script | 6 years ago |
| goana.R | Tune output of the script | 6 years ago |
| goana.Rproj | Add readme file | 6 years ago |
| goana.mouse.R | Renew script for mouse | 6 years ago |

README.md

## goana

GO analysis, use the `goana` function in limma package under the hook.

About

GO (Gene Ontoloty) term enrichment analysis

Readme
0 stars
2 watching
0 forks

Releases

No releases published

Packages

No packages published

Languages

R 99.5% Shell 0.5%

*https://github.com/jzsh2000/goana*

10/09/2024

# PRACTIQUEMOS

amazon WorkSpaces

viu

Universidad Internacional de Valencia

universidadviu.com

De:

Planeta Formación y Universidades