

# BIOINFORMATICS COURSE

# Hands on Pathway Analysis

Bioinformatics Course UEB-VHIR  
June 2022

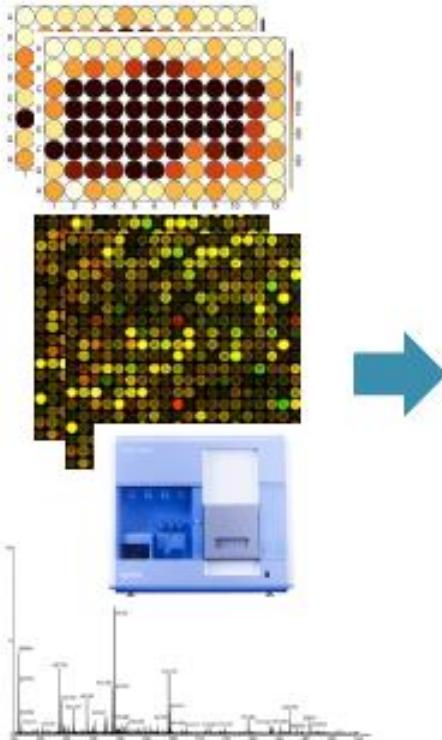
Mireia Ferrer<sup>1</sup>, Álex Sánchez<sup>1,2</sup>, Esther Camacho<sup>1</sup>, Angel Blanco<sup>1,2</sup>, Berta Miró<sup>1</sup>

<sup>1</sup> Unitat d'Estadística i Bioinformàtica (UEB) VHIR

<sup>2</sup> Departament de Genètica Microbiologia i Estadística, UB

# Hands on Pathway Analysis

What's the biology behind a list of genes ?



From individual genes

From groups of genes

# Hands on Pathway Analysis

## TABLE OF CONTENTS

1. Resources for functional analysis of gene lists
2. Navigating through different resources
  - a) Browsing the GO
  - b) Pathway visualization in Reactome
3. Case of study: RNA-seq experiment
  - a) Preparation of the gene list
  - b) Functional annotation of individual genes
  - c) Overrepresentation analysis of top DEG (g:Profiler / Reactome analysis tool)
  - d) Visualization and downstream analyses
4. A workflow in R (clusterProfiler)

# **Resources for functional analysis**

# Resources for functional analysis

## Annotation datasets

- Functional annotations can be very diverse: molecular functions, pathways (genes that work together to carry out a biological process), interactions, gene regulation, involvement in disease...
- Collections of gene sets:

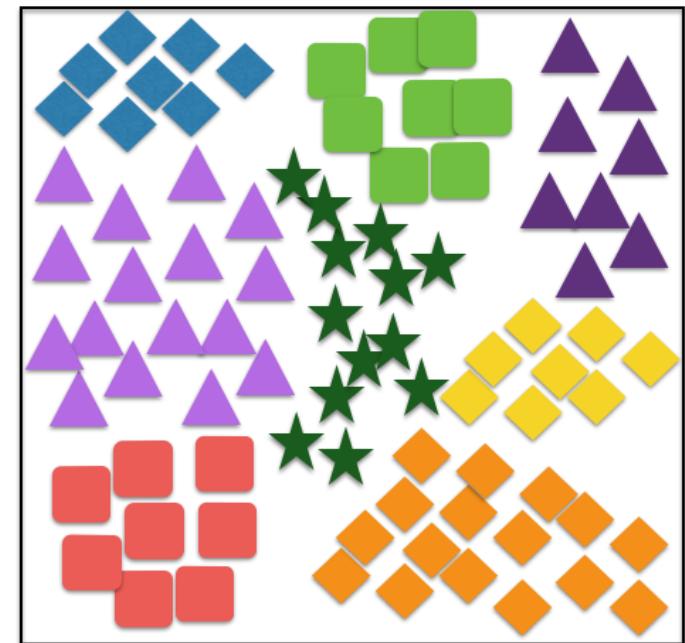
- Gene Ontology (GO)

- CC Cellular Components
- BP Biological Processes
- MF Molecular Function

- Pathways

- KEGG
- Reactome
- WikiPathways
- PantherDB

All known genes in a species  
(categorized into groups)



# Resources for functional analysis

## Annotation datasets

- Functional annotations can be very diverse: molecular functions, pathways (genes that work together to carry out a biological process), interactions, gene regulation, involvement in disease...

- Collections of gene sets:

- Gene Ontology (GO)

- CC Cellular Components
    - BP Biological Processes
    - MF Molecular Function

- Pathways (signaling, metabolic...)

- KEGG
    - Reactome
    - WikiPathways
    - PantherDB
    - BioCyc/HumanCyc

- Disease-related

- DisGeNet
    - Disease Ontology

- Cell/Tissue markers

- CellMarker

- Chromosomal location

- Regulatory motifs

- TRANSFAC (Transcr Fact)
    - miRTarBase (miRNA-target)

# Resources for functional analysis

## Annotation datasets

- MSigDB

**MSigDB**  
Molecular Signatures Database

### Annotation datasets

**Overview**

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- ▶ **Search** for gene sets by keyword.
- ▶ **Browse** gene sets by name or collection.
- ▶ **Examine** a gene set and its annotations. See, for example, the **HALLMARK\_APOPTOSIS** gene set page.
- ▶ **Download** gene sets.
- ▶ **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**

**License Terms**

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

Please [register](#) to download the GSEA software and the MSigDB gene sets, and to use our web tools. 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.

**Current Version**

MSigDB database v7.2 updated September 2020. [Release notes](#).

**Citing the MSigDB**

To cite your use of the Molecular Signatures Database (MSigDB), a joint project of UC San Diego and Broad Institute, please reference Subramanian, Tamayo, et al. (2005). PNAS.

### Molecular Signatures Database v7.2

#### Collections

The MSigDB gene sets are divided into 9 major 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** for each human chromosome and cytogenetic band.
- 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 microarray 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** defined directly from microarray gene expression data from immunologic studies.
- C8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

# Resources for functional analysis

## Web tools and software

- g:Profiler - <http://biit.cs.ut.ee/gprofiler/index.cgi>
- Reactome - <https://reactome.org/>
- DAVID - <http://david.abcc.ncifcrf.gov/tools.jsp>
- GeneMANIA - <http://www.genemania.org/>
- WebGestalt - <http://webgestalt.org/>
- AmiGO - <http://amigo.geneontology.org/amigo>
- QuickGO - <https://www.ebi.ac.uk/QuickGO/>
- ReviGO - <http://revigo.irb.hr/>
- GSEA - <http://software.broadinstitute.org/gsea/index.jsp>
- ClueGO - <http://apps.cytoscape.org/apps/cluego>
- PANTHER - <http://pantherdb.org/>
- GOrilla – <http://cbl-gorilla.cs.technion.ac.il/>

# Resources for functional analysis

## R packages

- clusterProfiler - <http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>
- GAGE/Pathview - <http://www.bioconductor.org/packages/release/bioc/html/gage.html>
- SPIA - <https://www.bioconductor.org/packages/release/bioc/html/SPIA.html>
- WGCNA - <https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/>

The screenshot shows the Bioconductor website interface. At the top, there's a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. Below the navigation bar, the URL [http://bioconductor.org/packages/release/BiocViews.html#\\_Pathways](http://bioconductor.org/packages/release/BiocViews.html#_Pathways) is displayed. The main content area is titled "All Packages". On the left, there's a sidebar with a tree view of package categories: ResearchField (902), StatisticalMethod (727), Technology (1251), WorkflowStep (1081) which includes Alignment (80), Annotation (124), BatchEffect (52), ExperimentalDesign (22), GenomeBrowsers (2), MultipleComparison (163), Normalization (128), Pathways (179) which includes BioCarta (7), GO (72), KEGG (51), Reactome (19). The main content area displays a table titled "Packages found under GO:" with the following data:

Package	Maintainer	Title	Rank
<a href="#">annotate</a>	Bioconductor Package Maintainer	Annotation for microarrays	17
<a href="#">GOSemSim</a>	Guangchuang Yu	GO-terms Semantic Similarity Measures	36
<a href="#">clusterProfiler</a>	Guangchuang Yu	statistical analysis and visualization of functional profiles for genes and gene clusters	38
<a href="#">enrichplot</a>	Guangchuang Yu	Visualization of Functional Enrichment Result	43
<a href="#">GSEABase</a>	Bioconductor Package Maintainer	Gene set enrichment data structures and methods	56
<a href="#">interactiveDisplayBase</a>	Shawn Balcombe	Base package for enabling powerful shiny web displays of Bioconductor objects	59

# Resources for functional analysis

## Considerations when choosing a tool

- Type of analysis performed
  - Annotation
  - Enrichment analysis (Overrepresentation vs GSEA)
  - Network-based analyses
  - Coexpression, ...
- Statistical test performed, algorithms used, multiple testing adjustment
- Source of information
  - Available genesets, databases
  - Updated? Curated?
  - Organisms supported
- Input data accepted
  - Genes, miRNA, SNPs, metabolomics
  - Flexibility in managing different types of Ids
- Output, visualization, postprocessing

# Hands on Pathway Analysis

## TABLE OF CONTENTS

1. Resources for functional analysis of gene lists
2. Navigating through different resources
  - a) Browsing the GO
  - b) Pathway visualization in Reactome
3. Case of study: RNA-seq experiment
  - a) Preparation of the gene list
  - b) Functional annotation of individual genes
  - c) Overrepresentation analysis of top DEG (g:Profiler / Reactome analysis tool)
  - d) Visualization and downstream analyses
4. A workflow in R (clusterProfiler)

# Browsing the Gene Ontology

# Browsing the Gene Ontology

## 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



The network of biological classes describing the current best representation of the "universe" of biology. The molecular functions, cellular locations, and processes gene products may carry out.

- [GO Ontology Overview](#)
- [Browse in AmiGO](#)
- [Download](#)



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

- [GO Annotations Overview](#)
- [Browse in AmiGO](#)
- [Download](#)



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

- [GO-CAM Overview](#)
- [Browse GO-CAMs](#)
- [Download](#)



Tools to curate, browse, search, visualize and download both the ontology and annotations. Bioinformatic Guides (Notebooks) and simple API access to integrate GO into your research.

- [GO Tools Overview](#)
- [GO APIs Guide](#)
- [GO Github](#)

Current release 2020-12-08: 44,117 GO terms | 7,963,579 annotations  
1,562,091 gene products | 4743 species (see statistics)

### GO Enrichment Analysis ?

Powered by PANTHER

Your gene IDs here...

biological process

Homo sapiens

Examples

Launch

Note: can use UniProt ID/AC, Gene Name, Gene Symbol, MGI ID

# Browsing the Gene Ontology

## Queries that can be done with the GO

- AmiGO / QuickGO
- Hierarchy
- Search for a term like apoptosis
- Search for the annotations of a given gene
- Do enrichment analysis of a gene list

# Browsing the Gene Ontology

- Functional annotation of individual genes with QuickGO

The screenshot shows the QuickGO search interface. At the top, there is a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and EMBL-EBI. Below the navigation bar is a search bar with the text 'Quick GO' and a 'Search' button. Underneath the search bar are links for Help, Contact, API, and Basket. The main content area displays the results for the gene product 'LACTB'. On the left, there is a sidebar with filters for UniProtKB (13 results), TrEMBL (3), Swiss-Prot (3), Type (MiRNA, Protein), Organism (Homo Sapiens selected), and Proteome status (Reference Proteomes). The main table lists the results for LACTB, showing columns for Database, ID, Name, Type, Taxon, and Annotations. The first result is Serine beta-lactamase-like protein LACTB, mitochondrial, with a UniProtKB ID P83111, categorized as a PROTEIN for Homo sapiens, with 14 annotations.

## Results for gene product "LACTB"

UniProtKB

13 results

TrEMBL (3)

Swiss-Prot (3)

Type

MiRNA (7)

Protein (6)

Organism

Homo Sapiens

Proteome status

Reference Proteomes  
(Gene Centric, Canonical) (3)

Database	ID	Name	Type	Taxon	Annotations
UniProtKB	P83111	Serine beta-lactamase-like protein LACTB, mitochondrial	PROTEIN	Homo sapiens	<a href="#">14 annotations</a>
UniProtKB	H0YNN5	Serine beta-lactamase-like protein LACTB, mitochondrial	PROTEIN	Homo sapiens	
UniProtKB	Q53H82	Endoribonuclease LACTB2	PROTEIN	Homo sapiens	<a href="#">16 annotations</a>
UniProtKB	A8MY62	Putative beta-lactamase-like 1	PROTEIN	Homo sapiens	
UniProtKB	A0A024R811	Lactamase, beta 2, isoform CRA_a	PROTEIN	Homo sapiens	
UniProtKB	H0Y608	Putative beta-lactamase-like 1	PROTEIN	Homo sapiens	

# Browsing the Gene Ontology

- Functional annotation of individual genes with QuickGO

Quick GO

Search

Help Contact API Basket

## GO annotations

Taxon ▾ ✓ Gene Products ▾ GO terms ▾ References ▾ Aspect ▾ Evidence ▾ Extension ▾ More ▾ Clear all

Annotations Statistics

Customise Export 5 annotations

- Biological Process?
- Molecular function?
- Cellular localization?

Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With / From	Taxon	Assigned	
UniProtKB:A0A2K6LNQ6	LACTB	involved_in	GO:0019216	regulation of lipid metabolic process	ECO:0000265	GO_REF:0000107	UniProtKB:P83111 more...	61621 Rhinopithecus bieti	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	enables	GO:0008233	peptidase activity	ECO:0000265	GO_REF:0000107	UniProtKB:P83111 more...	61621 Rhinopithecus bieti	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	involved_in	GO:0006508	proteolysis	ECO:0000265	GO_REF:0000107	UniProtKB:P83111 more...	61621 Rhinopithecus bieti	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	part_of	GO:0005829	cytosol	ECO:0000265	GO_REF:0000107	UniProtKB:P83111 more...	61621 Rhinopithecus bieti	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	part_of	GO:0005739	mitochondrion	ECO:0000265	GO_REF:0000107	UniProtKB:P83111 more...	61621 Rhinopithecus bieti	Ensembl

# Hands on Pathway Analysis

## TABLE OF CONTENTS

1. Resources for functional analysis of gene lists
2. Navigating through different resources
  - a) Browsing the GO
  - b) Pathway visualization in Reactome
3. Case of study: RNA-seq experiment
  - a) Preparation of the gene list
  - b) Functional annotation of individual genes
  - c) Overrepresentation analysis of top DEG (g:Profiler / Reactome analysis tool)
  - d) Visualization and downstream analyses
4. A workflow in R (clusterProfiler)

# Browsing the Reactome Pathways database

# Browsing the Reactome Pathways

<https://reactome.org/>

The screenshot shows the Reactome website homepage. At the top, there's a navigation bar with links for About, Content, Docs, Tools, Community, and Download. Below the navigation is a search bar with placeholder text "e.g. O95631, NTN1, signaling by EGFR, glucose" and a "Go!" button. The main content area features four large blue icons with labels: "Pathway Browser" (with a cluster icon), "Analyze Data" (with a bar chart icon), "ReactomeFlViz" (with a network icon), and "Documentation" (with a document icon). Below these are descriptions: "Visualize and interact with Reactome biological pathways" for Pathway Browser, "Merges pathway identifier mapping, over-representation, and expression analysis" for Analyze Data, "Designed to find pathways and network patterns related to cancer and other types of diseases" for ReactomeFlViz, and "Information to browse the database and use its principal tools for data analysis" for Documentation. A black banner at the bottom has the text "USE REACTOME GRAPH DATABASE IN YOUR PROJECT" and a "LEARN MORE" button. On the left of this banner is a yellow-bordered box containing the "Why Reactome" section, which includes a brief description of the database and a note about its Chinese mirror site. To the right of the banner is a "Tweets" section from the ChEMBL Database Twitter account.

Tutorials: <https://reactome.org/userguide>  
[https://www.ebi.ac.uk/training/online/courses/reactome-tools-for-analysing-biological-pathways/#vf-tabs\\_section--overview](https://www.ebi.ac.uk/training/online/courses/reactome-tools-for-analysing-biological-pathways/#vf-tabs_section--overview)

# Browsing the Reactome Pathways

## Navigating through Reactome Pathway database

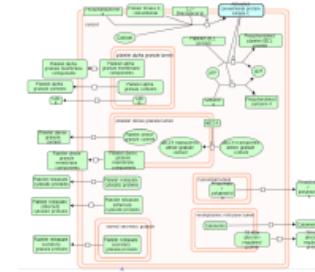
Reactome is...

Free, online, open-source curated database of pathways and reactions in human biology

Authored by expert biologists, maintained by Reactome editorial staff (curators)

Mapped to cellular compartment

Extensively cross-referenced



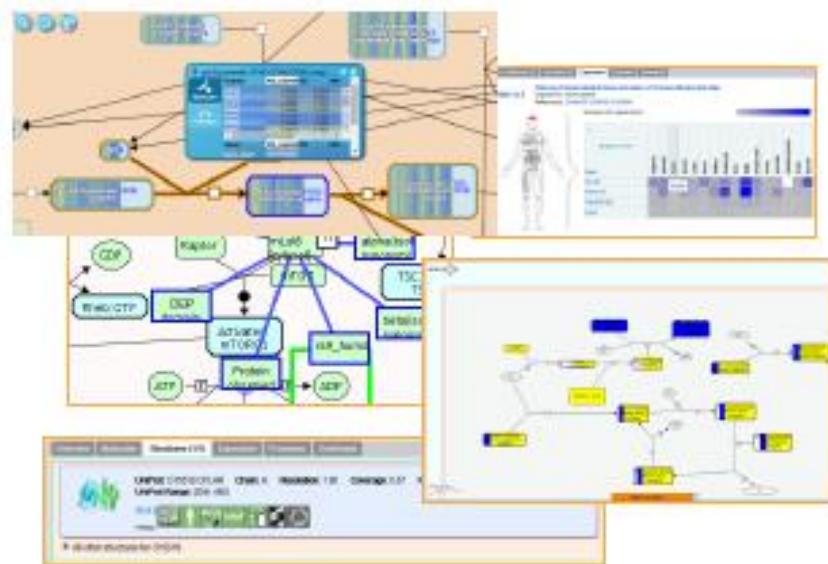
Entrez Gene



# Browsing the Reactome Pathways

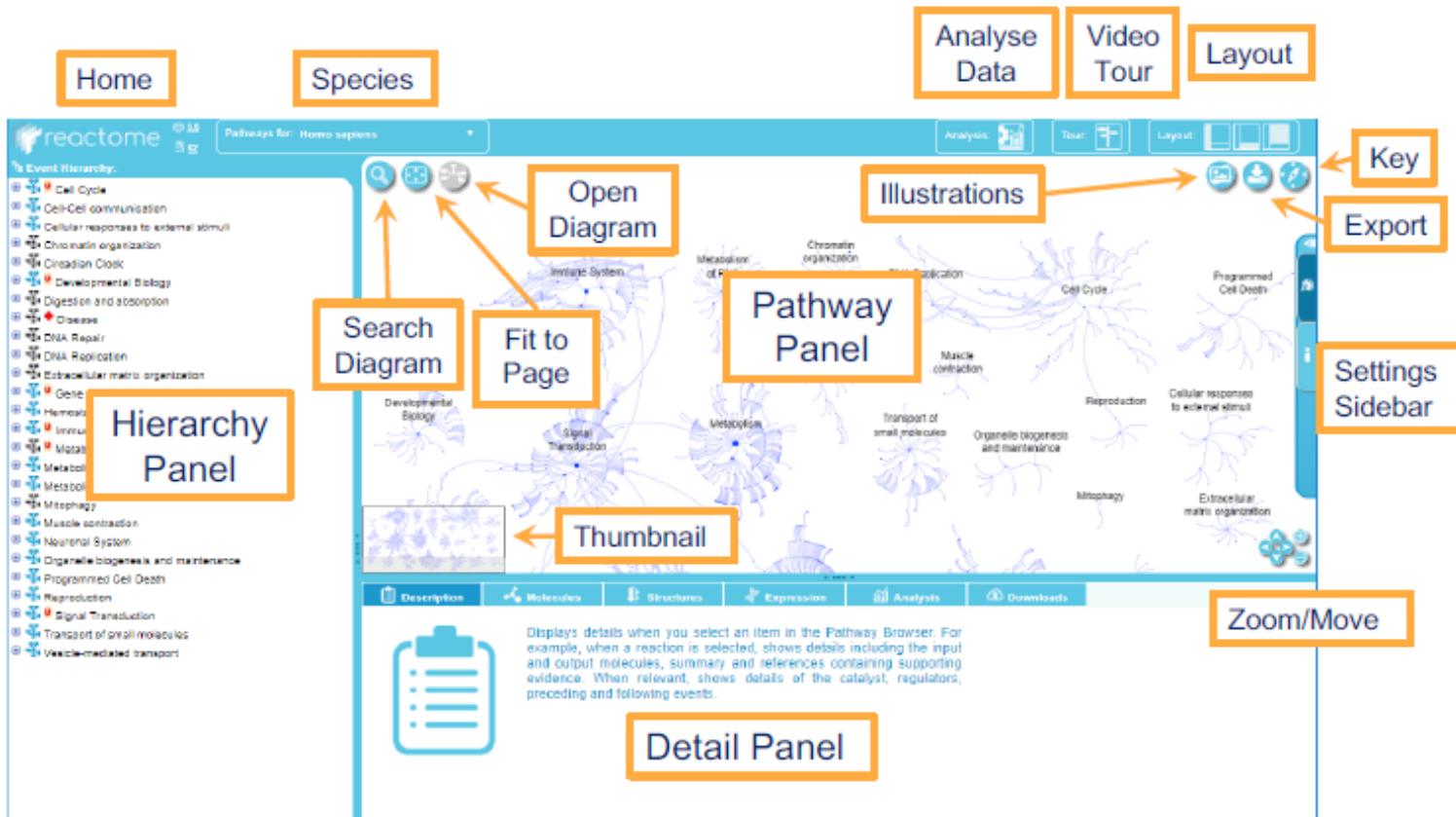
## Reactome Tools

- Interactive Pathway Browser
- Analysis
  - Over-representation
  - Pathway topology
  - Expression overlay
- Molecular Interaction overlay
- Species Comparison



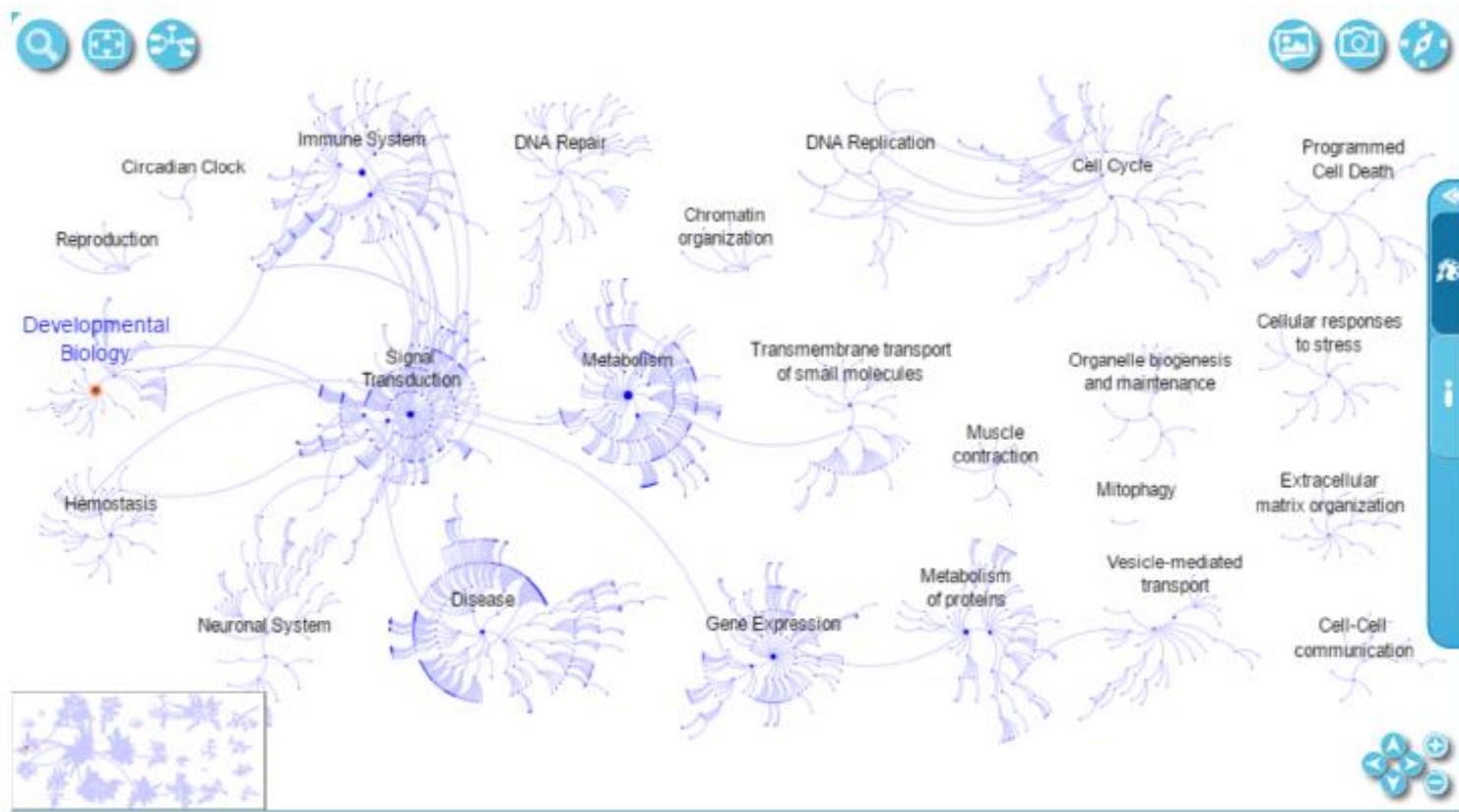
# Browsing the Reactome Pathways

## The Pathway Browser



# Browsing the Reactome Pathways

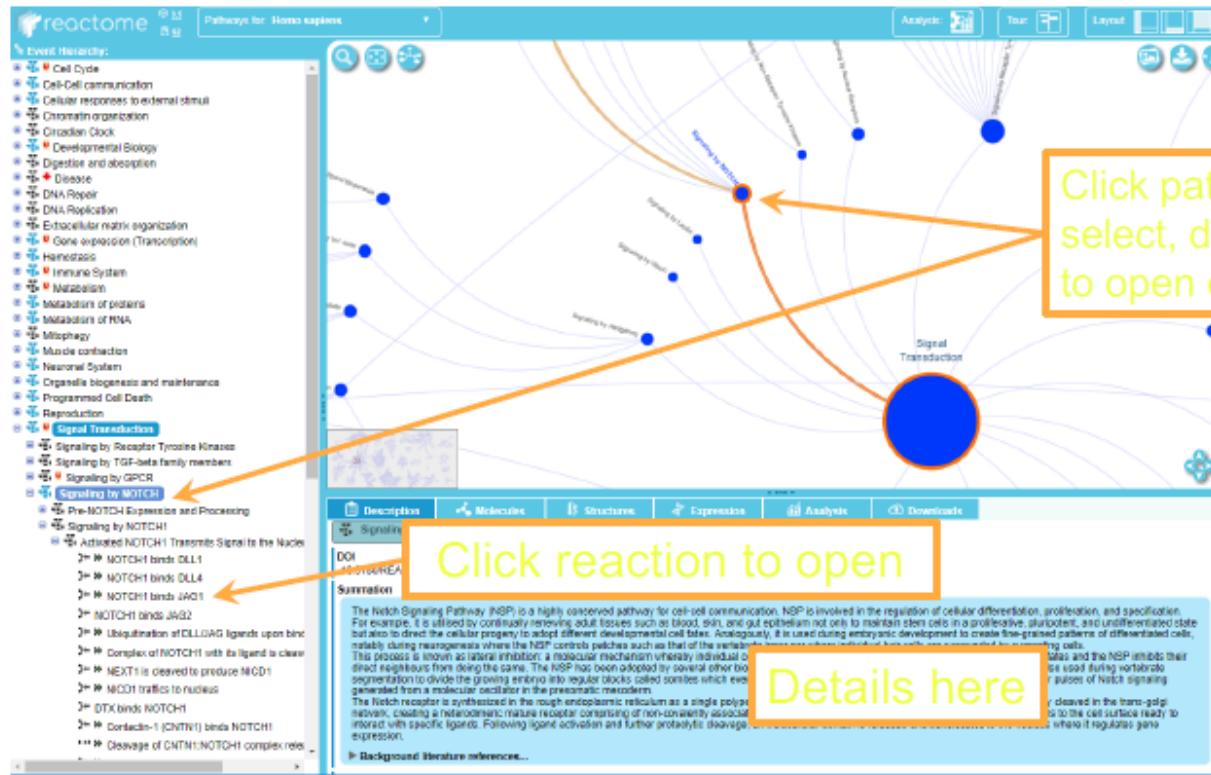
## Pathway Overview



# Browsing the Reactome Pathways

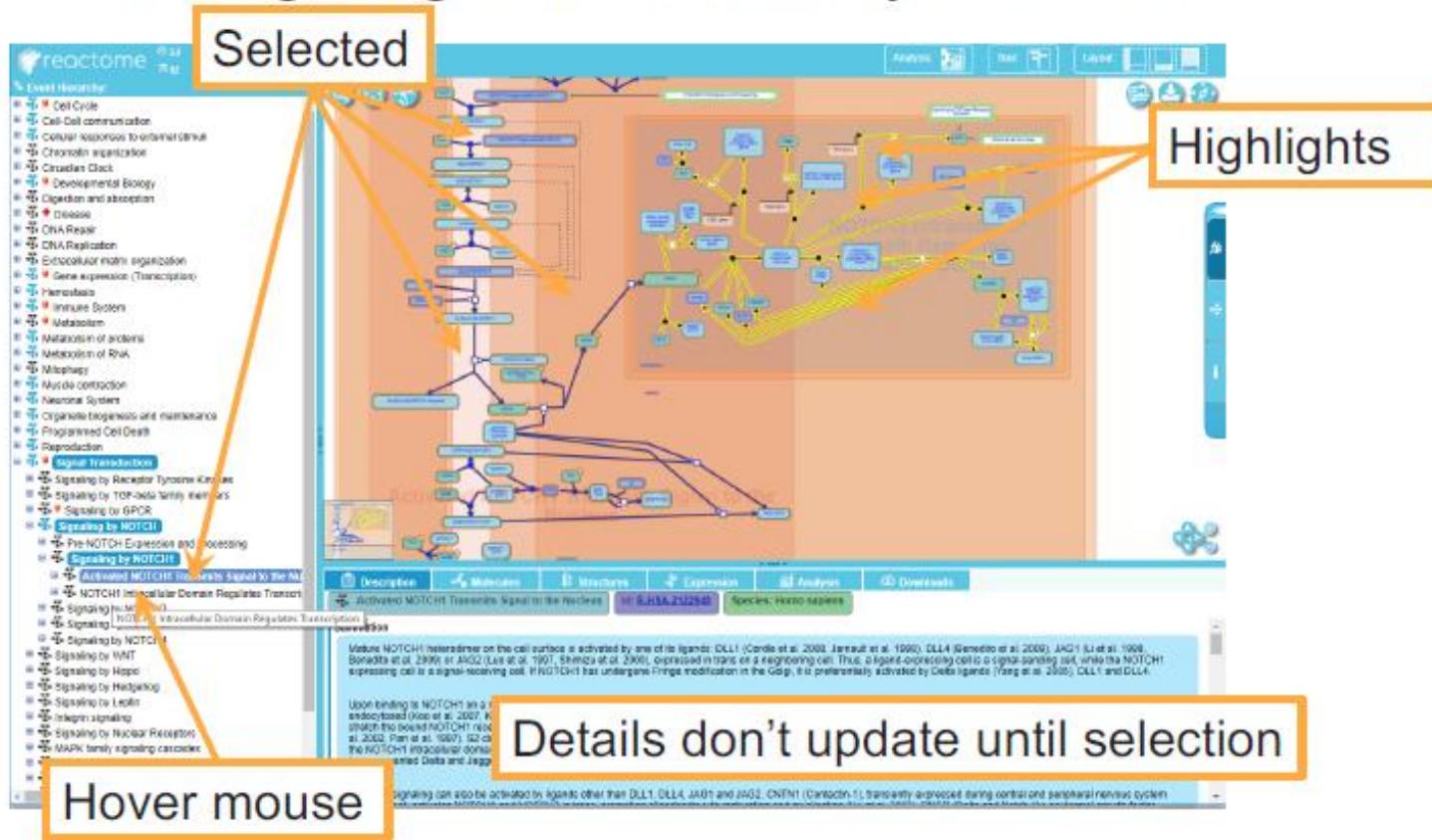
## Navigating in the Pathway Browser

Home button



# Browsing the Reactome Pathways

## Navigating in the Pathway Browser



The screenshot shows the Reactome Pathway Browser interface. A search bar at the top contains the text "NOTCH". On the left, a navigation tree is shown under the heading "Selected", which is highlighted with an orange box. The tree includes categories like Cell Cycle, Cell-Cell communication, and Signal Transduction. Under "Signal Transduction", the "Signaling by NOTCH" section is expanded, also highlighted with an orange box. The main panel displays a complex network of pathways, with the NOTCH signaling pathway highlighted in orange. A vertical sidebar on the right shows a scrollable list of terms. At the bottom, there are tabs for Description, Molecules, Structures, Expression, Analysis, and Downloads. A tooltip at the bottom provides a detailed description of the NOTCH signaling process.

**Selected**

**Highlights**

**Hover mouse**

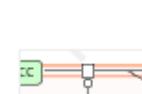
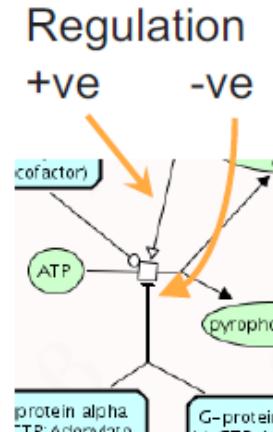
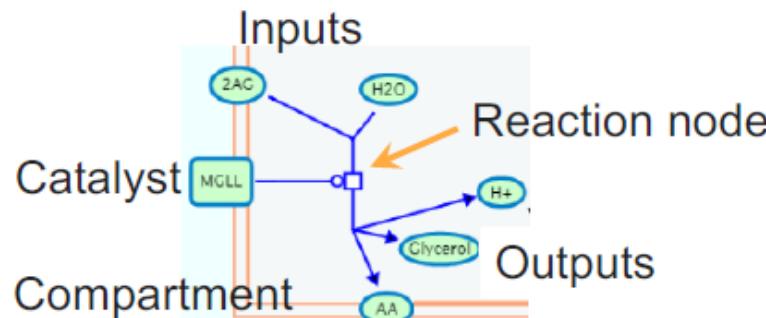
**Details don't update until selection**

Mature NOTCH1 heterodimer on the cell surface is activated by one of its ligands: DLL1 (Cordel et al. 2002; Jansson et al. 1996), DLL4 (Gremet et al. 2000), JAG1 (Li et al. 1998; Bonafida et al. 2000) or JAG2 (Lee et al. 1997; Shimizu et al. 2000), expressed in trans on a neighboring cell. Thus, a ligand-expressing cell is a signal-sending cell, while the NOTCH-expressing cell is a signal-receiving cell. If NOTCH1 has undergone Fringe modification in the Gag, it is preferentially activated by DLL1 ligand (Yang et al. 2005). DLL1 and DLL4, upon binding to NOTCH1 in a 2:1 endocytosed (Kao et al. 2007). It triggers the p70S6K/TAK1/NF- $\kappa$ B pathway (Yan et al. 1998; Li et al. 1999). The NOTCH1 receptor undergoes proteolytic cleavage by the γ-secretase complex (Wang et al. 2003; Kao et al. 2007). The released intracellular domain (NICD) translocates to the nucleus where it activates transcriptional programs.

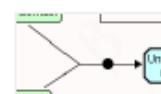
# Browsing the Reactome Pathways

## The Pathway Browser - Pathway Diagrams

Ovals are small molecules (or sets of)  
 Green boxes are proteins,  
 Blue are complexes,  
 Blue with double-boundary are sets



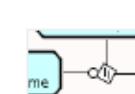
Transition



Binding



Dissociation



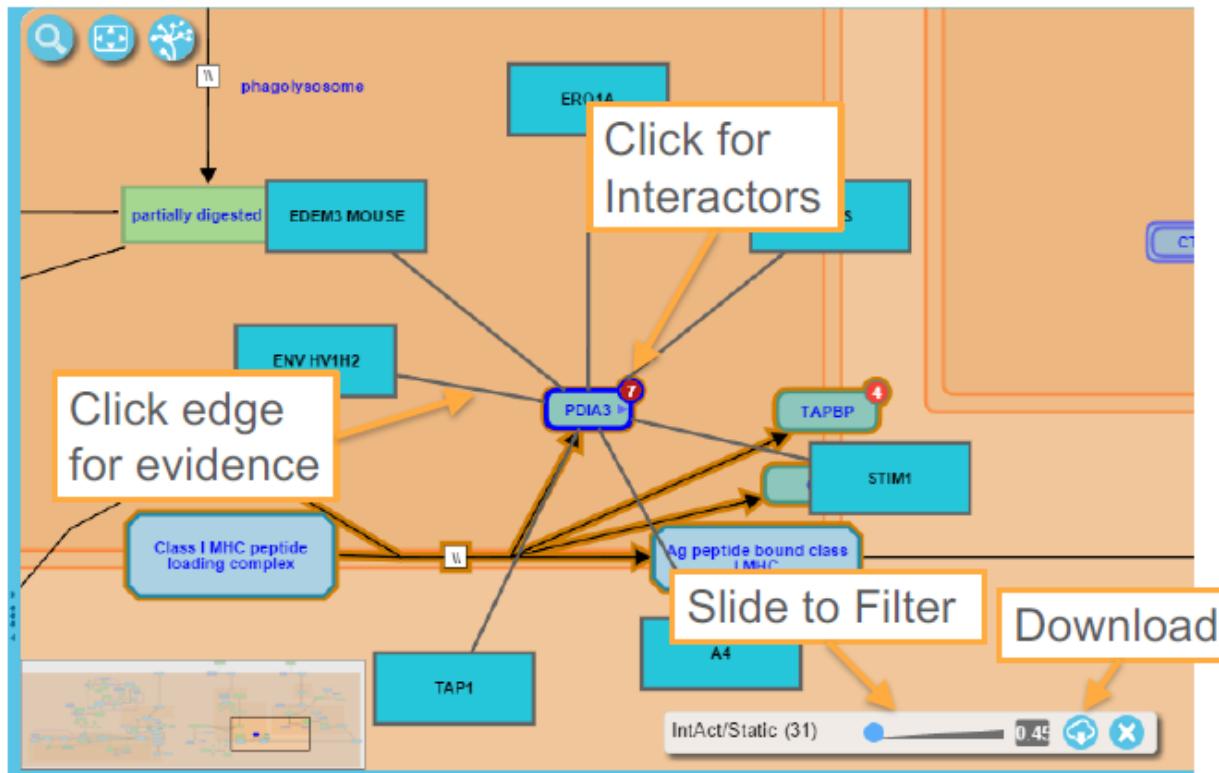
Omitted



Uncertain

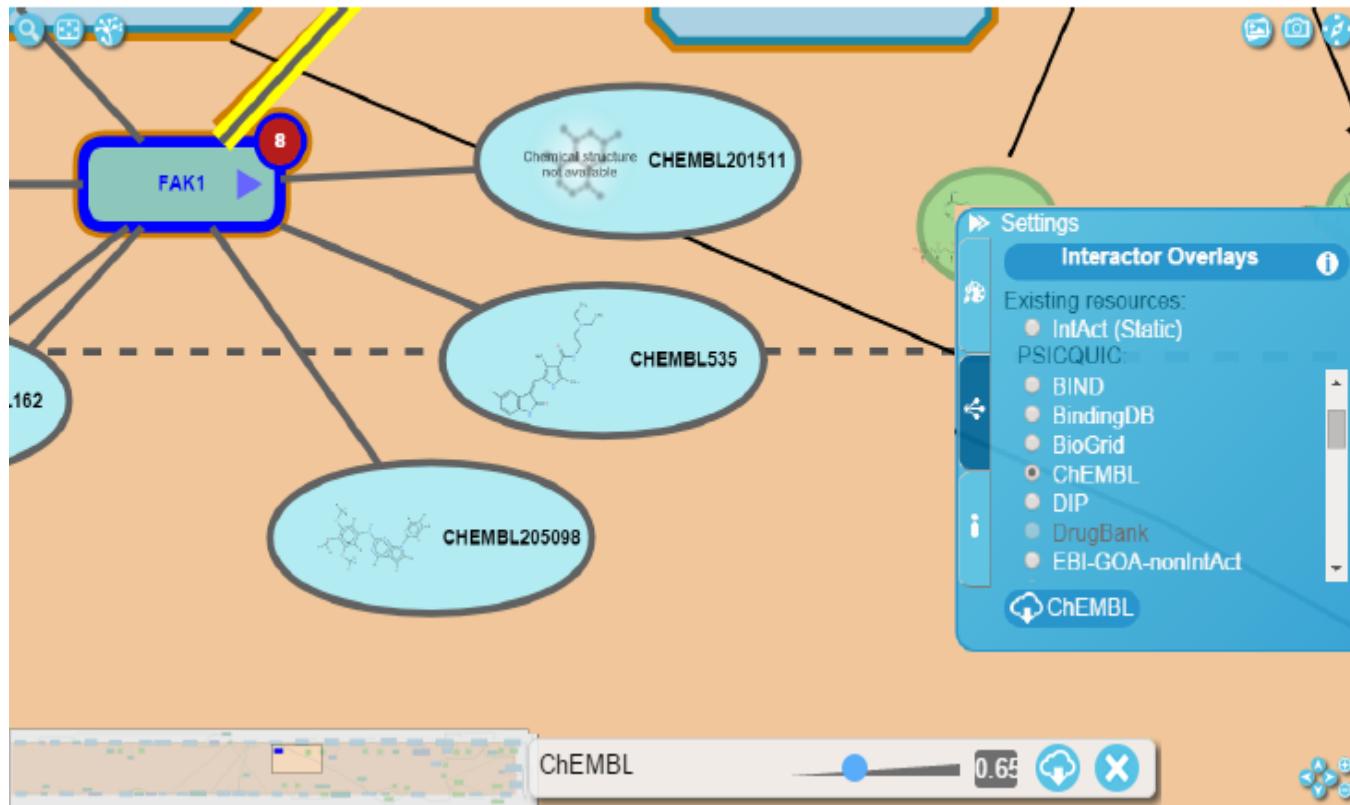
# Browsing the Reactome Pathways

## Interactors



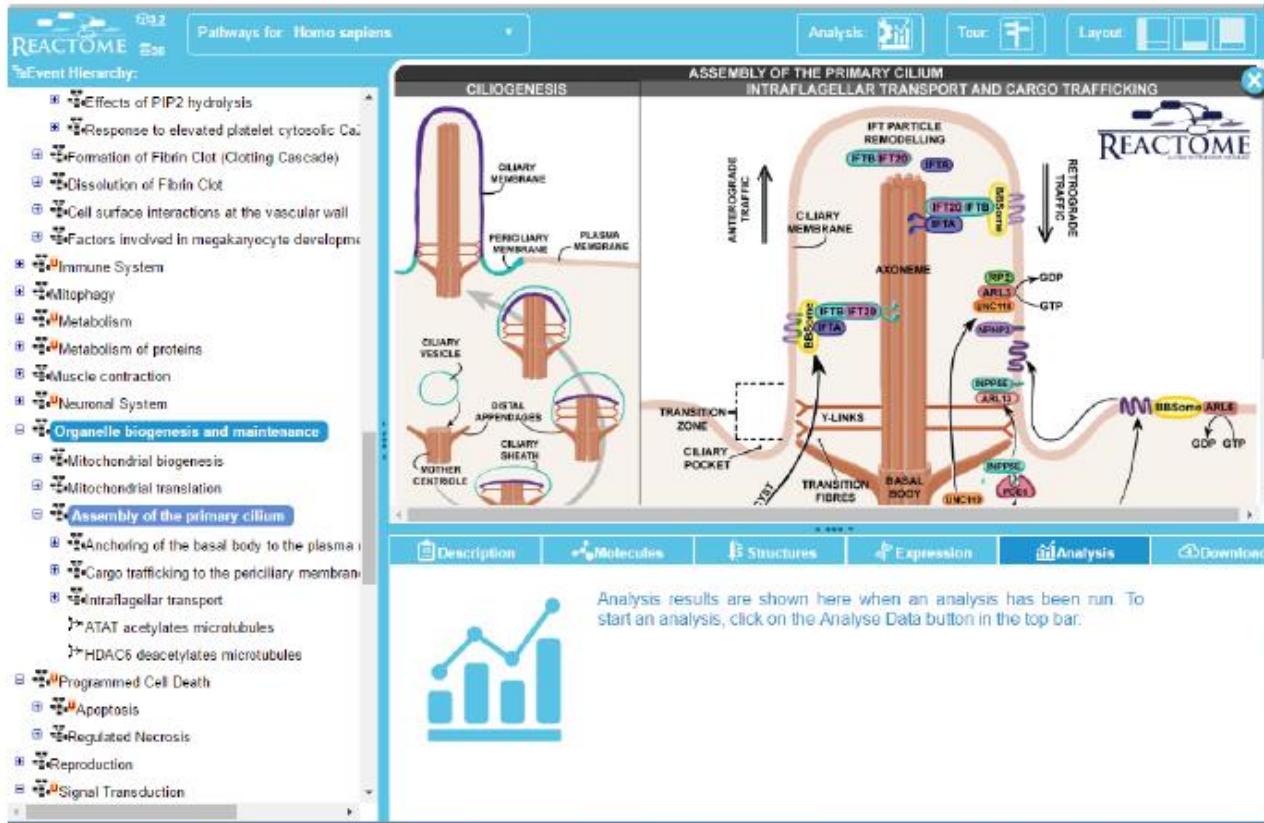
# Browsing the Reactome Pathways

## Molecular Interaction Overlay – Set source



# Browsing the Reactome Pathways

## Show Illustration

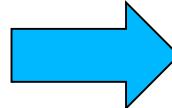


# Browsing the Reactome Pathways

## Searching the Reactome

### Simple text search

The simple text search tool is located top right of the Home page. Several other Reactome pages include this search tool. To search Reactome type a word, phrase or identifier in the search box:



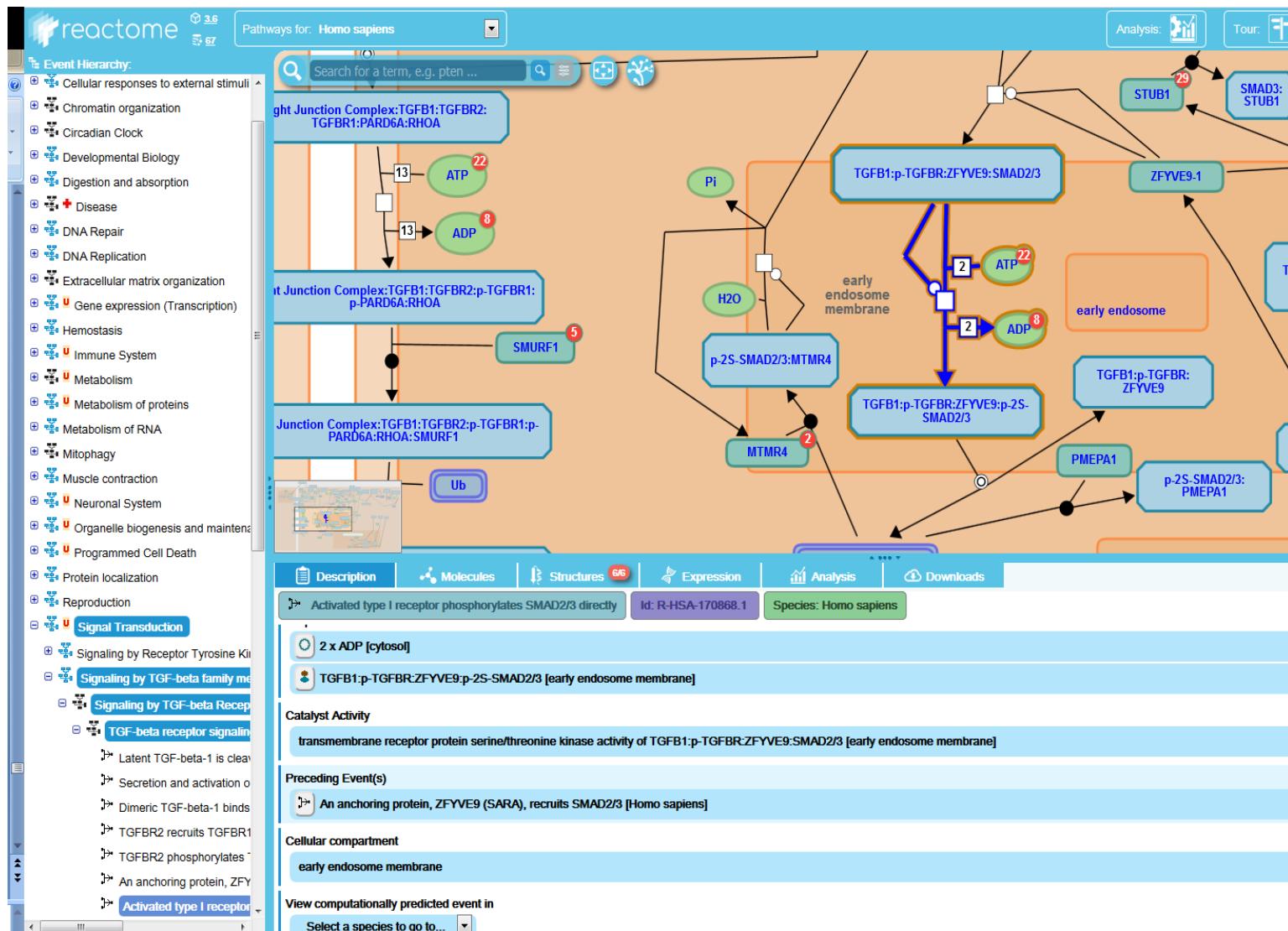
Different types of results (Pathway, Reactions, Prot...)

# Browsing the Reactome Pathways

## Searching the Reactome

1. Find the reaction Activated type I receptor phosphorylates SMAD2/3 directly. What pathway does it belong to?
2. In which cellular compartment does this reaction take place?
3. What is the GO molecular function associated with the catalyst?
4. What references verify this reaction?
5. Is this reaction predicted to occur in *Canis familiaris*? In *Saccharomyces cerevisiae*?
6. Is this event likely to occur in liver?
7. Are 3D structures available for TGF-beta1?

# Browsing the Reactome Pathways



# Hands on Pathway Analysis

## TABLE OF CONTENTS

1. Resources for functional analysis of gene lists
2. Navigating through different resources
  - a) Browsing the GO
  - b) Pathway visualization in Reactome
3. Case of study: RNA-seq experiment
  - a) Preparation of the gene list
  - b) Functional annotation of individual genes
  - c) Overrepresentation analysis of top DEG (g:Profiler / Reactome analysis tool)
  - d) Visualization and downstream analyses
4. A workflow in R (clusterProfiler)

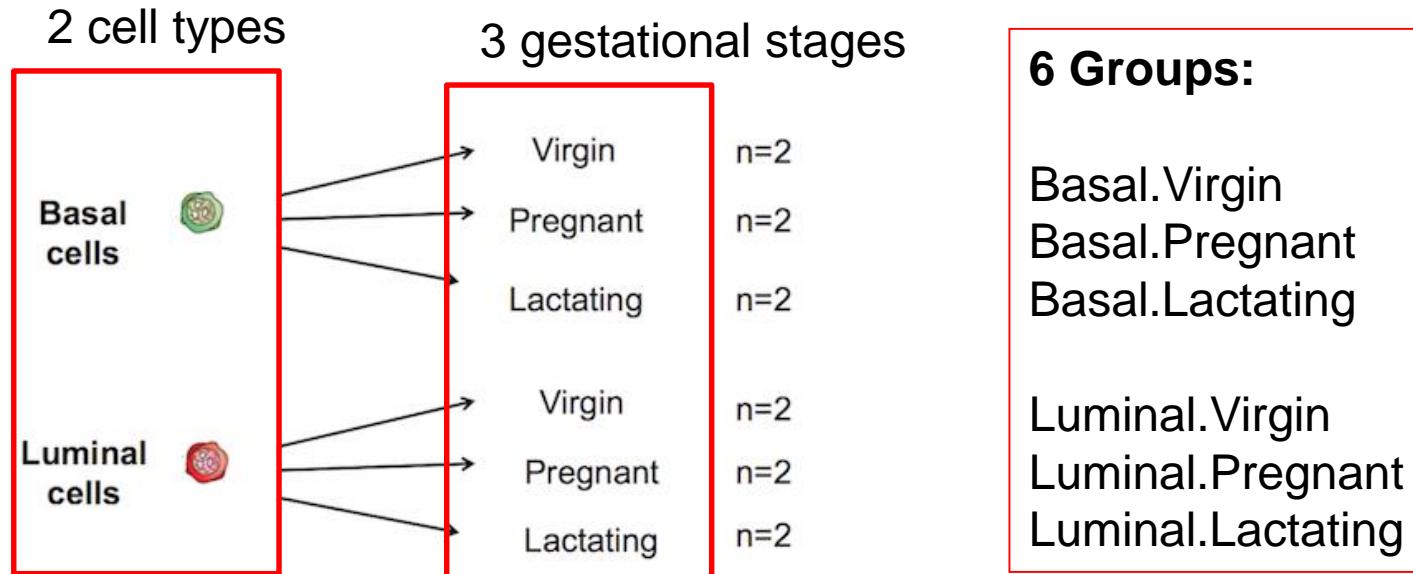


# Case of Study

# Case of study

## Study overview

Dataset: RNA-seq data of mouse mammary gland ([GSE60450](#))



We analyzed the differential gene expression in the following comparisons:

- Basal.Pregnant vs. Basal.Lactating
- Luminal.Pregnant vs. Luminal.Lactating

# Case of study

## Study overview

**Objective:** Our experiment in RNA-seq returned some genes differentially expressed in **basal.pregnant vs basal.lactate cells**. We want now to extract some biological meaning from these lists, such as **which biological processes or pathways the differentially expressed genes are implicated in**.

- What type(s) of test/study can you perform to get this information?
- What is the input data in such tests?
- Do you know some databases where you can find information on biological processes/pathways?

# Case of study

## Preparation of the gene list

- Prepare a Gene List with the “top” up-regulated genes and one with the “top” down-regulated genes according to the following statistical criteria:
  - **Up-regulated genes:** genes with an **adjusted p-value < 0.05 AND logFC > 2**
  - **Down-regulated genes:** genes with an **adjusted p-value < 0.05 AND logFC < -2**
- Under these criteria, how many genes are up-regulated? And how many down-regulated?
- What is the **Ensembl ID** of your top **most up/down-regulated gene?** And its **Gene Symbol / UniProt ID** (SwissProt)?

Tip: You can use g:Convert tool from g:Profiler for ID conversion

# Case of study

## Preparation of the gene list

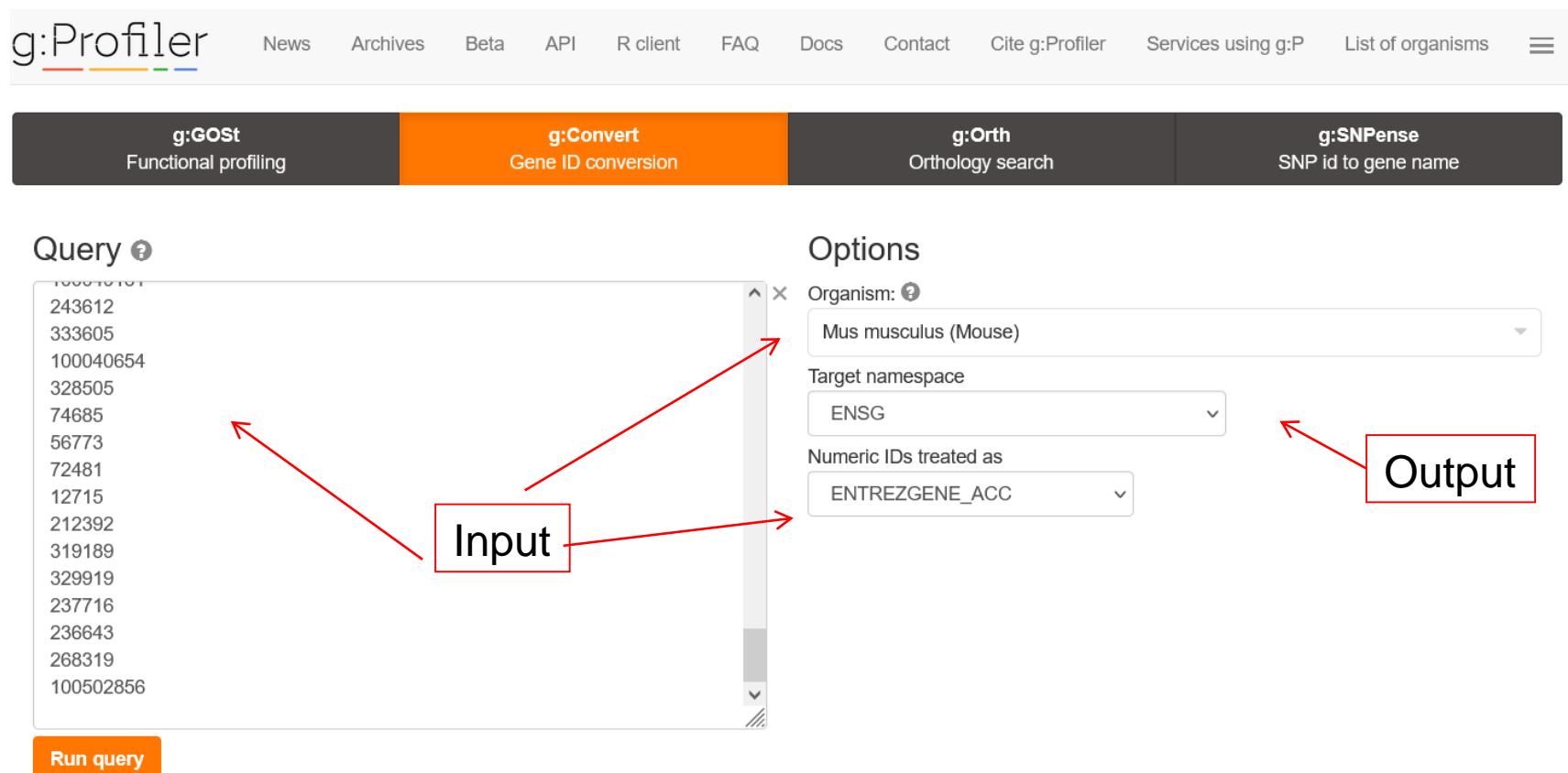
- Prepare a Gene List with the “top” up-regulated genes and one with the “top” down-regulated genes according to the following statistical criteria:
  - **Up-regulated genes:** genes with an **adjusted p-value < 0.05 AND logFC > 2**
  - **Down-regulated genes:** genes with an **adjusted p-value < 0.05 AND logFC < -2**
- Under these criteria, how many genes are up-regulated? And how many down-regulated? **294 up, 101 down**
- What is the **Ensembl ID** of your top **most up/down-regulated gene?** And its **Gene Symbol / UniProt ID** (SwissProt)?

Tip: You can use g:Convert tool from g:Profiler for ID conversion

# Case of study

## Preparation of the gene list

### ID mapping with g:Profiler – g:Convert



**Query** ⓘ

243612  
333605  
100040654  
328505  
74685  
56773  
72481  
12715  
212392  
319189  
329919  
237716  
236643  
268319  
100502856

**Run query**

**Input**

**Options**

Organism: ⓘ  
Mus musculus (Mouse)

Target namespace  
ENSG

Numeric IDs treated as  
ENTREZGENE\_ACC

**Output**

**g:Convert** enables to convert between various gene, protein, microarray probe and numerous other types of namespaces. We provide at least 40 types of IDs for more than 60 species. The 98 different namespaces supported for human include Ensembl, Refseq,

Illumina, Entrezgene and Uniprot identifiers. All namespaces are obtained through matching them via Ensembl gene identifiers as a reference.

# Case of study

## Preparation of the gene list

### ID mapping with g:Profiler

[Export to CSV](#) [Show query URL](#) [Show short link](#)

initial alias	converted alias	name	description	namespace
381290	ENSMUSG00000026463	Atp2b4	ATPase, Ca++ transporting, plasma membrane 4 [Source:MGI Symbol;Acc:MGI:88111]	ENTREZGENE_ACC
226101	ENSMUSG00000048612	Myof	myoferlin [Source:MGI Symbol;Acc:MGI:1919192]	ENTREZGENE_ACC
16012	ENSMUSG00000023046	Igfbp6	insulin-like growth factor binding protein 6 [Source:MGI Symbol;Acc:MGI:96441]	ENTREZGENE_ACC
16669	ENSMUSG00000020911	Krt19	keratin 19 [Source:MGI Symbol;Acc:MGI:96693]	ENTREZGENE_ACC
231991	ENSMUSG00000053007	Creb5	cAMP responsive element binding protein 5 [Source:MGI Symbol;Acc:MGI:2443973]	ENTREZGENE_ACC
211577	None	None	None	ENTREZGENE_ACC
270150	ENSMUSG00000070306	Ccdc153	coiled-coil domain containing 153 [Source:MGI Symbol;Acc:MGI:2448587]	ENTREZGENE_ACC
21953	ENSMUSG00000031097	Tnni2	troponin I, skeletal, fast 2 [Source:MGI Symbol;Acc:MGI:105070]	ENTREZGENE_ACC
12992	ENSMUSG00000061388	Csn1s2b	casein alpha s2-like B [Source:MGI Symbol;Acc:MGI:105312]	ENTREZGENE_ACC
170761	ENSMUSG00000032105	Pdzd3	PDZ domain containing 3 [Source:MGI Symbol;Acc:MGI:2429554]	ENTREZGENE_ACC
67111	ENSMUSG00000029413	Naaa	N-acylethanolamine acid amidase [Source:MGI Symbol;Acc:MGI:1914361]	ENTREZGENE_ACC
232016	ENSMUSG00000037973	Itprid1	ITPR interacting domain containing 1 [Source:MGI Symbol;Acc:MGI:2685304]	ENTREZGENE_ACC
76123	ENSMUSG00000027883	Gpsm2	G-protein signalling modulator 2 (AGS3-like, C. elegans) [Source:MGI Symbol;Acc:MGI:1...	ENTREZGENE_ACC
18019	ENSMUSG00000027544	Nfatc2	nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 2 [Source:MGI Sy...	ENTREZGENE_ACC
67971	ENSMUSG00000014846	Tppp3	tubulin polymerization-promoting protein family member 3 [Source:MGI Symbol;Acc:M...]	ENTREZGENE_ACC
74134	ENSMUSG00000040703	Cyp2s1	cytochrome P450, family 2, subfamily s, polypeptide 1 [Source:MGI Symbol;Acc:MGI:192...	ENTREZGENE_ACC
12833	ENSMUSG0000001119	Col6a1	collagen, type VI, alpha 1 [Source:MGI Symbol;Acc:MGI:88459]	ENTREZGENE_ACC
66234	ENSMUSG00000031604	Msmo1	methylsterol monooxygenase 1 [Source:MGI Symbol;Acc:MGI:1913484]	ENTREZGENE_ACC
12834	ENSMUSG00000020241	Col6a2	collagen, type VI, alpha 2 [Source:MGI Symbol;Acc:MGI:88460]	ENTREZGENE_ACC
21827	ENSMUSG00000028047	Thbs3	thrombospondin 3 [Source:MGI Symbol;Acc:MGI:98739]	ENTREZGENE_ACC
235505	ENSMUSG00000046186	Cd109	CD109 antigen [Source:MGI Symbol;Acc:MGI:2445221]	ENTREZGENE_ACC
59031	ENSMUSG00000036599	Chst12	carbohydrate sulfotransferase 12 [Source:MGI Symbol;Acc:MGI:1929064]	ENTREZGENE_ACC
57266	ENSMUSG00000021508	Cxcl14	chemokine (C-X-C motif) ligand 14 [Source:MGI Symbol;Acc:MGI:1888514]	ENTREZGENE_ACC
242122	ENSMUSG00000051076	Vtcn1	V-set domain containing T cell activation inhibitor 1 [Source:MGI Symbol;Acc:MGI:30396...	ENTREZGENE_ACC
320311	ENSMUSG00000047496	Rnf152	ring finger protein 152 [Source:MGI Symbol;Acc:MGI:2443787]	ENTREZGENE_ACC
107769	ENSMUSG00000038623	Tm6sf1	transmembrane 6 superfamily member 1 [Source:MGI Symbol;Acc:MGI:1933209]	ENTREZGENE_ACC
329738	ENSMUSG00000049565	Aknad1	AKNA domain containing 1 [Source:MGI Symbol;Acc:MGI:3584453]	ENTREZGENE_ACC
12722	ENSMUSG00000040212	Fmp2	epithelial membrane protein 2 [Source:MGI Symbol;Acc:MGI:1009720]	ENTREZGENE_ACC

# Case of study

## Functional annotation of individual genes

### What is known of gene X?

- Get information for one of the genes in some database (Pubmed, UniProt, NCBI-Gene...)
- Look for GO terms (CC, MF, BP) associated to your favorite gene:

- Tip: Search by gene/protein symbol in QuickGO:

<https://www.ebi.ac.uk/QuickGO/>

# Case of study

## Functional annotation of individual genes

*Note that doing this for every one of your genes may take you some long time...*

PubMed ▼ **GNAQ**  
Create RSS Create alert Advanced

Format: Summary ▼ Sort by: Most Recent ▼ Per page: 20 ▼ Send to ▼

See 271 articles about **GNAQ** gene function  
See also: [GNAQ G protein subunit alpha q](#) in the Gene database  
**gnaq** in [Homo sapiens](#) [Mus musculus](#) [Rattus norvegicus](#) All 160 Gene records  
See also: [38 tests for GNAQ in the Genetic Testing Registry](#)

Search results  
Items: 1 to 20 of 370

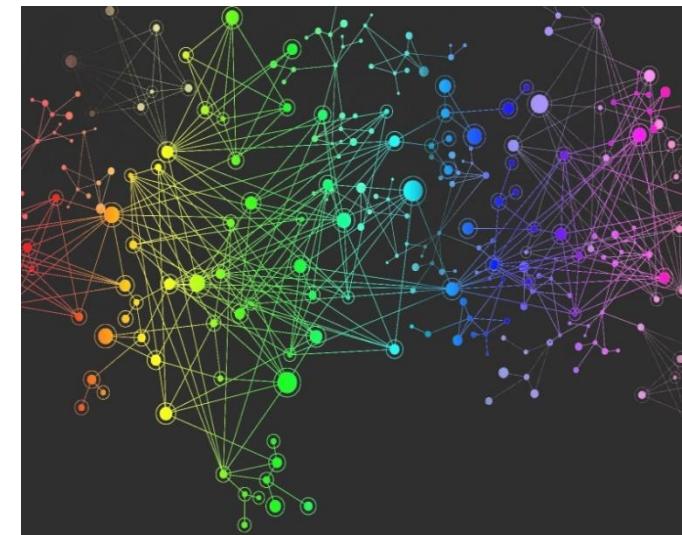
<< First < Prev Page 1 of 19 Next > Last >>



# Case of study

## Functional annotation of individual genes

- Genes usually don't act on their own
- Extracting information at the “group” level may be more meaningful and facilitate interpretation of the processes involved in the condition of study



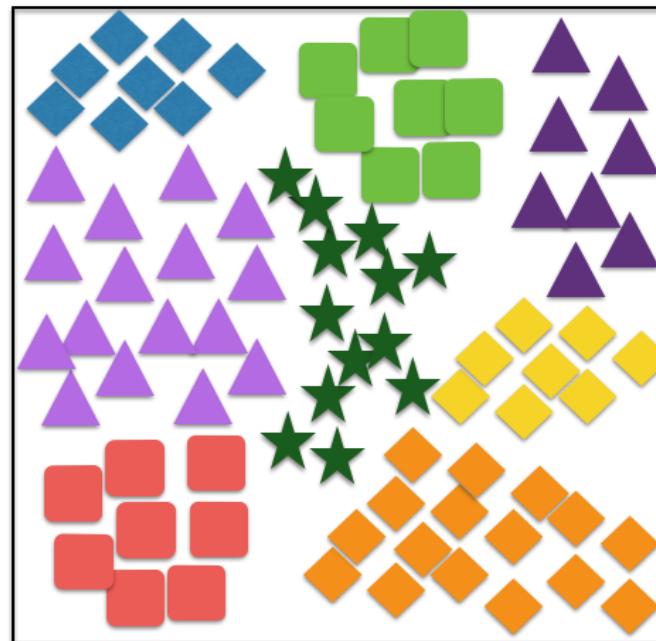
# Overrepresentation analysis

# Case of study

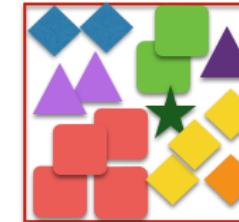
## Overrepresentation analysis

- Perform a statistical test (eg. Fisher exact test) to determine whether any gene sets (GO terms/ pathways) are “surprisingly” enriched in your gene list

All known genes in a species  
(categorized into groups)



Gene list



# Case of study

## Overrepresentation analysis

### What do we need?

Our list of genes



Selected genes from topTable

The universe from which comes our list of genes



Remaining genes in microarray/genome

Gene sets of annotations



From annotation database  
(eg. *.gmt* file /*.gaf* file for GO)

<http://geneontology.org/docs/download-go-annotations/>

Structure of relationships between annotation terms



From annotation database  
(eg. *.obo* file for GO)

<http://geneontology.org/docs/download-ontology/>

# Overrepresentation analysis

## g:Profiler – g:GOST

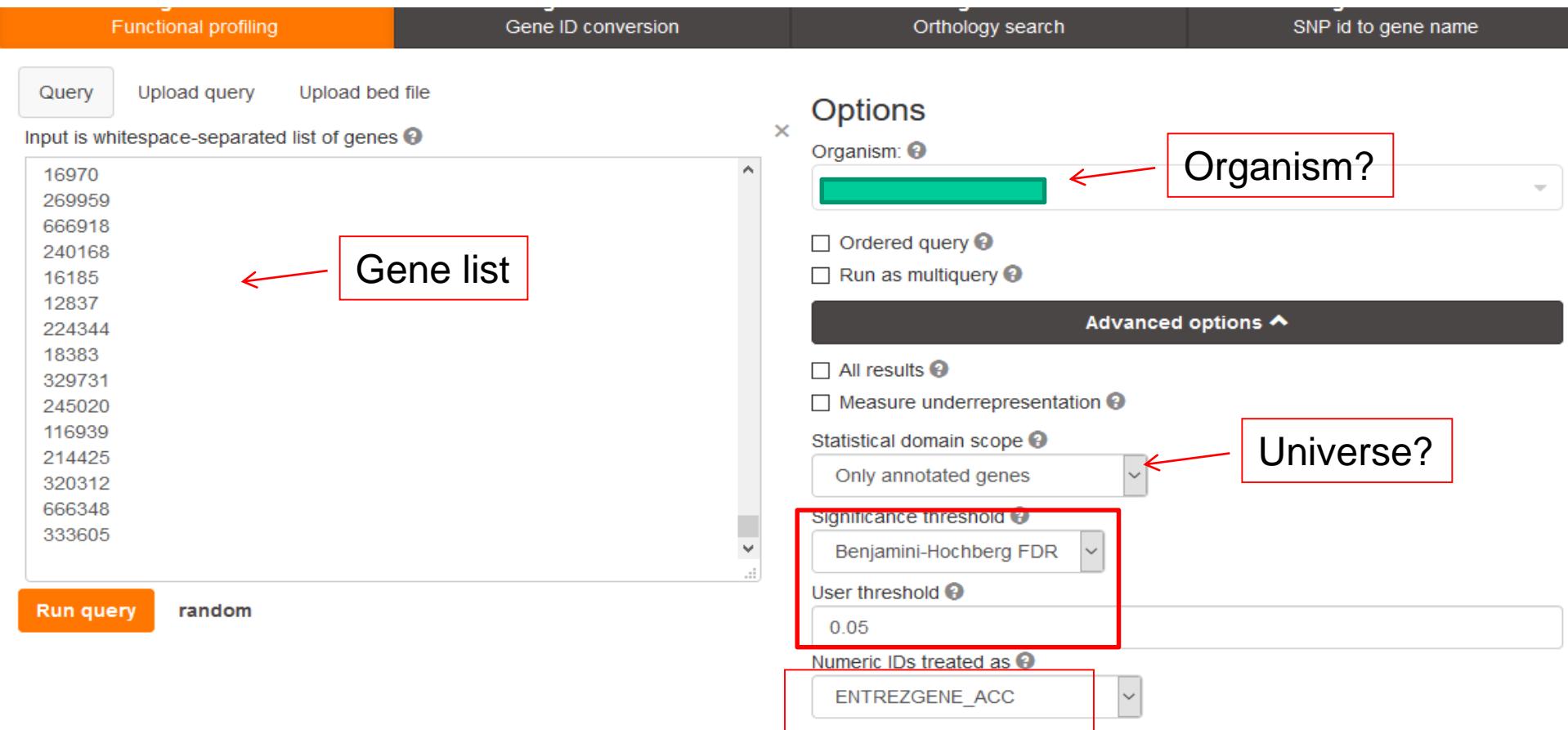
- ✓ Allows to perform enrichment analysis across different annotation databases
- ✓ Accepts different types of IDs as input
- ✓ P-value adjustments for multiple comparisons
- ✓ Source of evidence (experimental, computational, ...)
- ✓ Also for ranked lists

The screenshot shows the 'Data sources' section of the g:Profiler interface. At the top, there are buttons for 'select all', 'clear all', and 'Show data versions'. Below this, there are several sections of checked checkboxes:

- Gene Ontology**: GO molecular function, GO cellular component, GO biological process, No electronic GO annotations
- biological pathways**: KEGG, Reactome, WikiPathways
- regulatory motifs in DNA**: TRANSFAC, miRTarBase
- protein databases**: Human Protein Atlas, CORUM
- Human phenotype ontology**: HP

# Overrepresentation analysis

## g:Profiler – g:GOSt



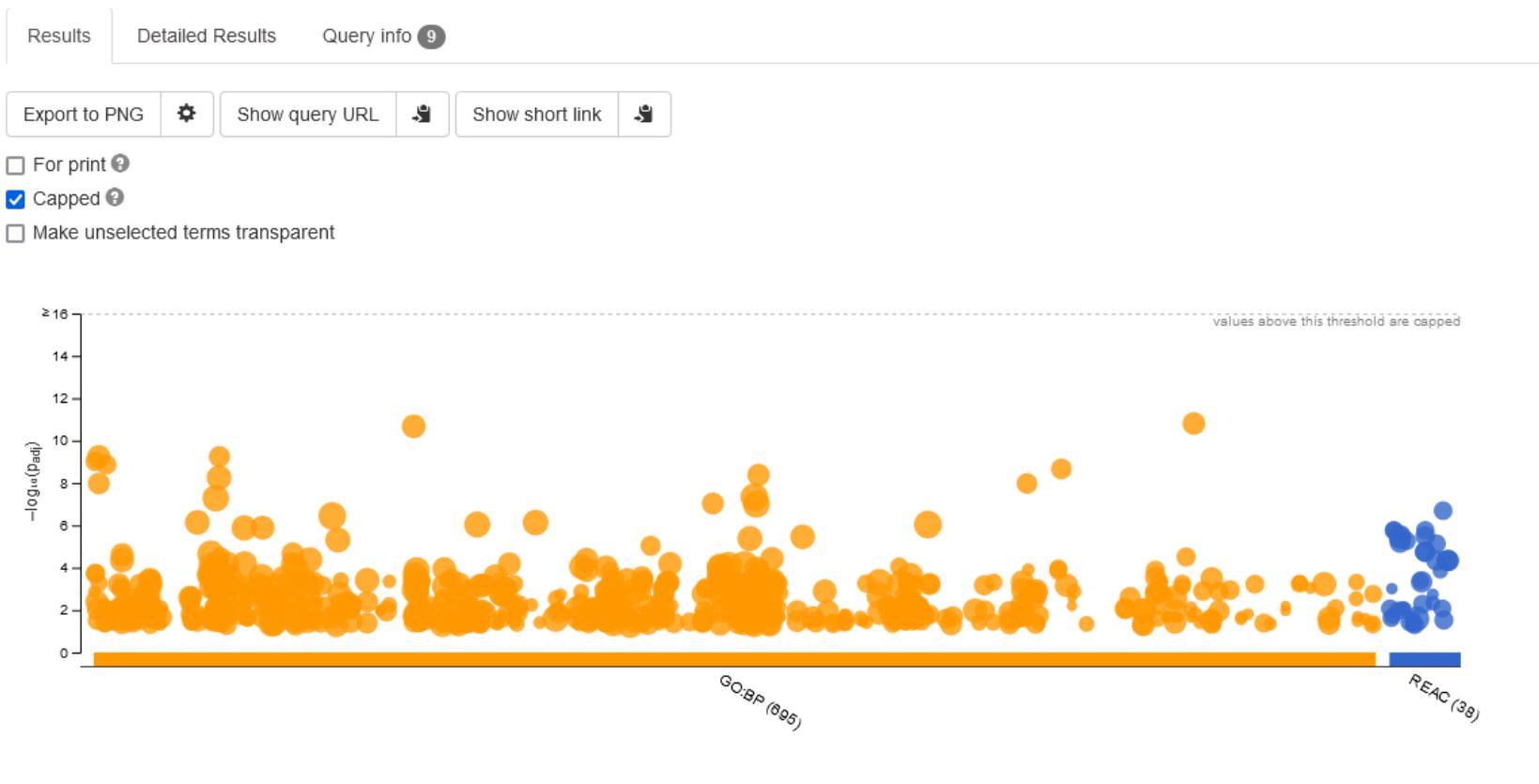
The screenshot shows the g:Profiler interface with the following components:

- Top Navigation:** Functional profiling, Gene ID conversion, Orthology search, SNP id to gene name.
- Left Panel:** A "Query" tab is selected. It includes "Upload query" and "Upload bed file" buttons. A text input field says "Input is whitespace-separated list of genes". Below it is a list of gene IDs:
  - 16970
  - 269959
  - 666918
  - 240168
  - 16185
  - 12837
  - 224344
  - 18383
  - 329731
  - 245020
  - 116939
  - 214425
  - 320312
  - 666348
  - 333605
- Bottom Left Buttons:** Run query (orange), random.
- Right Panel (Options):** An "Options" panel is open.
  - Organism?**: A dropdown menu labeled "Organism" with a red box around it and an arrow pointing to it from the left.
  - Universe?**: A dropdown menu labeled "Statistical domain scope" with a red box around it and an arrow pointing to it from the right.
  - Significance threshold**: A dropdown menu set to "Benjamini-Hochberg FDR" with a red box around it.
  - User threshold**: A text input field set to "0.05" with a red box around it.
  - Numeric IDs treated as**: A dropdown menu set to "ENTREZGENE\_ACC" with a red box around it.

# Overrepresentation analysis

## g:Profiler – g:GOSt

### g:Profiler graphical output



# Overrepresentation analysis

## g:Profiler – g:GOSt

# g:Profiler graphical output

### g:Profiler specific vocabulary:

Query (Q): genes in my gene list

Term (T): tested gene-set (pathway)

Q&T (common genes): number of genes that overlap between my gene list and the tested gene-sets

(Q&T) / T: overlap normalized by the gene-set size

p-value: FDR (corrected for multiple hypothesis testing)



### > Gene-sets (pathways/processes) that were found enriched

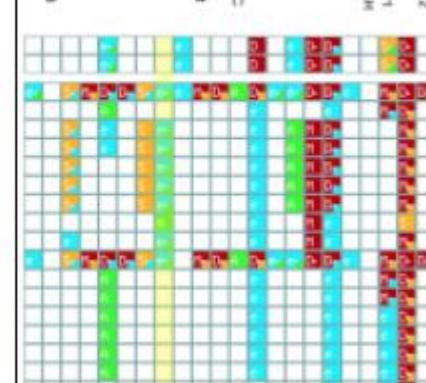
source	term name	term ID
BP	negative regulation of biological process	GO:0048519
BP	negative regulation of cellular process	GO:0048523
BP	single-organism process	GO:0044699
BP	death	GO:0015625
BP	developmental process	GO:0032592
BP	single-organism developmental process	GO:0044767
BP	multicellular organismal development	GO:0007275
BP	anatomical structure development	GO:0048896
BP	system development	GO:0048731
BP	tissue development	GO:0009888
BP	organ development	GO:0048513
BP	single-organism cellular process	GO:0044763
BP	cell death	GO:0006259
BP	programmed cell death	GO:0012951
BP	apoptotic process	GO:0006925
BP	regulation of cell death	GO:0051941
BP	regulation of programmed cell death	GO:0043067
BP	regulation of apoptotic process	GO:0042981

size of the overlap  
significance

size of the overlap

significance

### > Genes in input list



# Overrepresentation analysis

## GORilla

<http://cbl-gorilla.cs.technion.ac.il/>

- ✓ Direct visualization of enriched terms in hierarchy
- ✓ Needs to specify the universe
- ✓ Also for ranked lists
- ✗ Only for GO

# Overrepresentation analysis

## GORILLA

<http://cbl-gorilla.cs.technion.ac.il/>

### Step 1: Choose organism

Mus musculus

### Step 2: Choose running mode

- Single ranked list of genes     Two unranked lists of genes (target and background lists)

### Step 3: Paste a ranked list of gene/protein names

Names should be separated by an <ENTER>. The preferred format is gene symbol. Other supported formats are: gene and protein RefSeq, Uniprot, Unigene and Ensembl.

Target set:

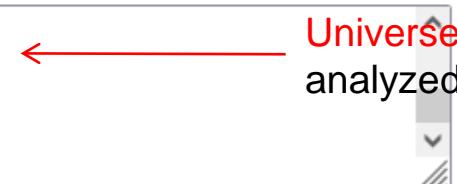
Myix2  
F630040K05Rik  
Cers3  
Gm10658  
Fer114  
Chac1  
Stra61



Or upload a file:  No se ha seleccionado ningún archivo.

Background set:

Gdf11  
Rcor1  
Smarcb1  
Zfp551  
Kbtbd6  
Psen2



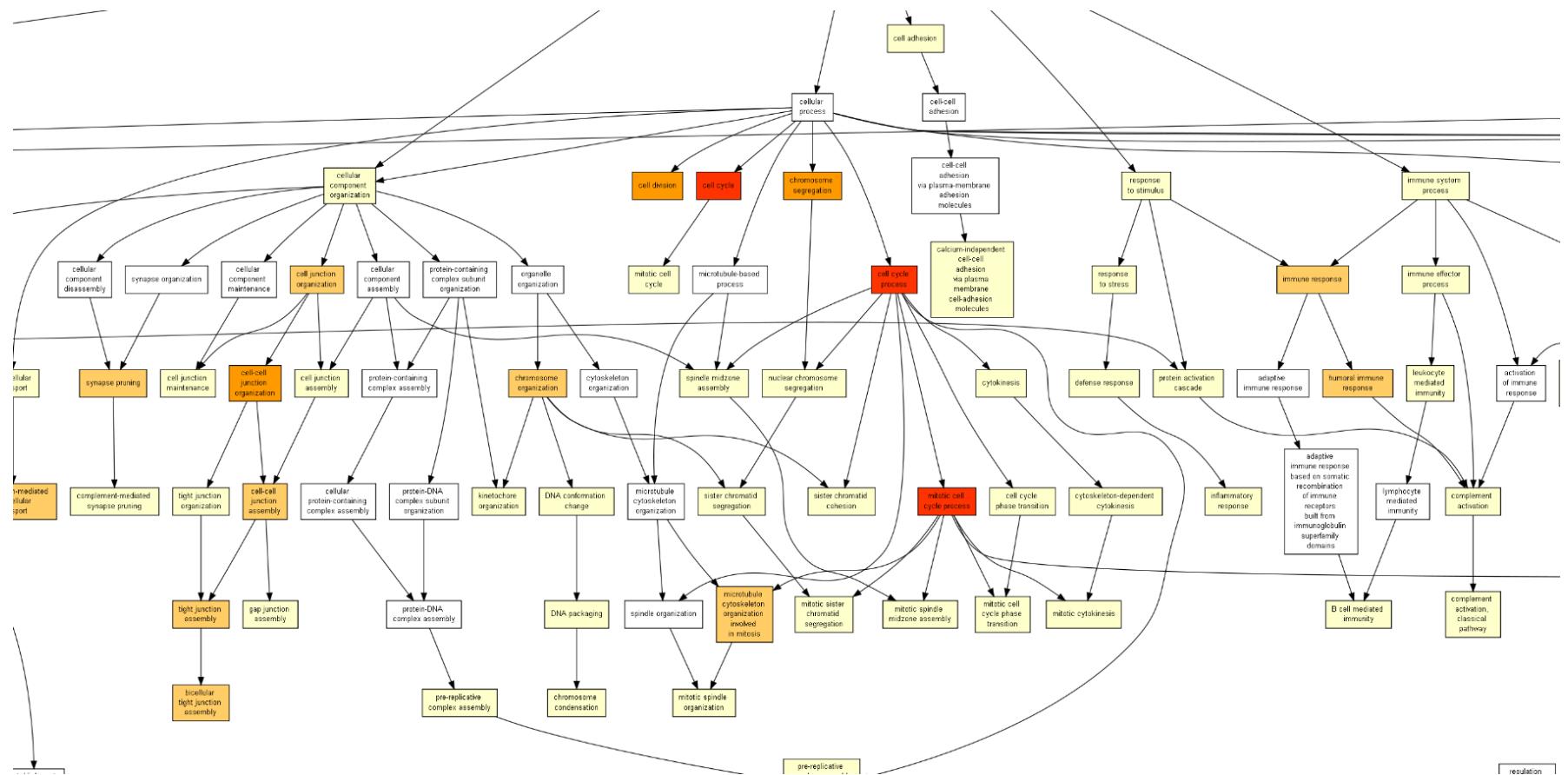
Or upload a file:  No se ha seleccionado ningún archivo.

### Step 4: Choose an ontology

- Process     Function     Component     All

# Overrepresentation analysis

## GORILLA



# Overrepresentation analysis

## Reactome Pathways analysis tool

### Analysis

The screenshot shows the Reactome homepage. At the top, there's a search bar with placeholder text "e.g. cdk1, mitogen, signaltransit, glucose". Below the search bar are four blue icons with text labels: "Pathway Browser" (with a subtext "Hausius and Interact with Reactome biological pathways"), "Analyze Data" (with a subtext "Integrate pathway analysis, mapping, overrepresentation and expression analysis" and circled in orange), "ReactomeFIViz" (with a subtext "Designed to find pathways and networks patient relevant to your samples"), and "Documentation" (with a subtext "Information to browse the database").

This screenshot shows a mobile or tablet view of the Reactome website. It features a large graph visualization on the right and a sidebar on the left with the title "USE REACTOME GRAPH DATABASE IN YOUR PROJECT". The sidebar contains a section titled "Why Reactome" with text about the database's mission and partners like EMBL-EBI, European Bioinformatics Institute, and CIBER-BBN.

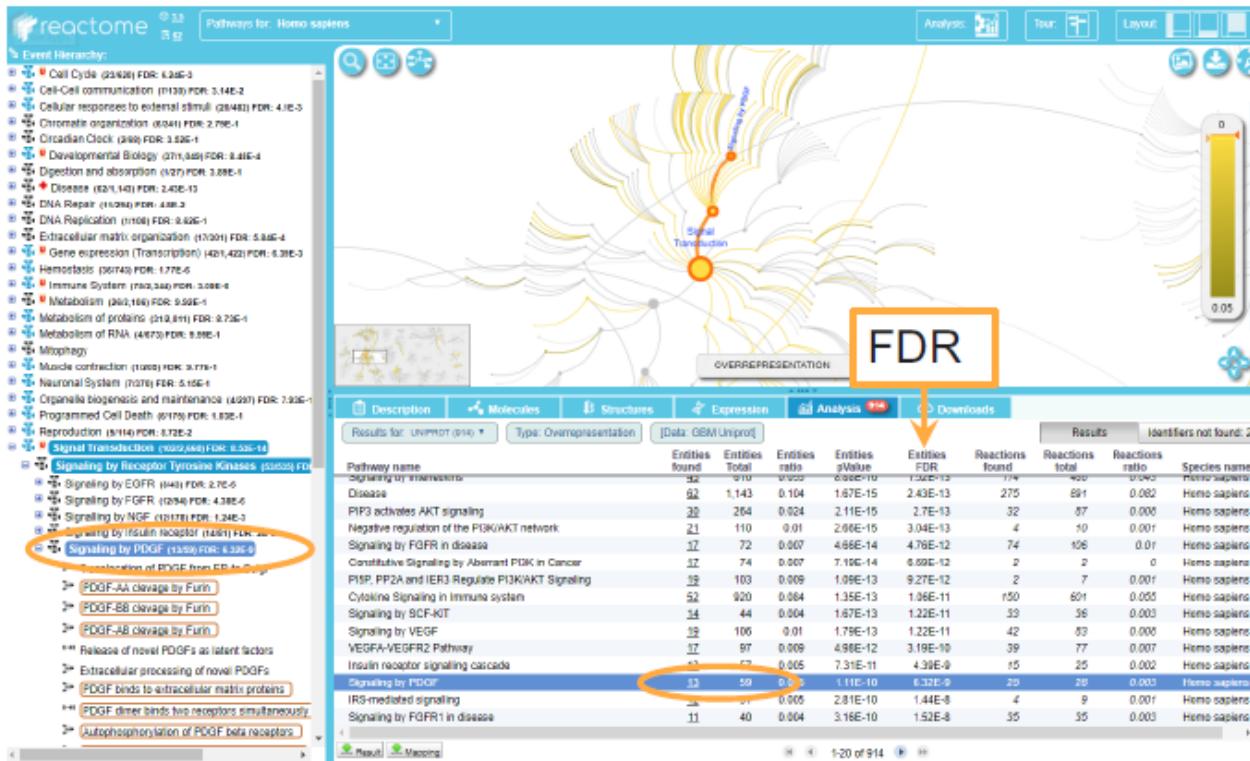
This screenshot shows the Reactome Pathway Browser. It displays a complex network of biological pathways as a graph. A specific node in the center is highlighted with a yellow box and labeled "Signal Transduction". To the left of the graph is a sidebar with a tree view of pathway categories. At the bottom of the screen, there are several tabs: "Description", "Molecules", "Reactions", "Enzymes", "Kinases", and "Downloads". A blue icon representing a clipboard is also present at the bottom.

Familiarize with  
Reactome using example  
data

# Overrepresentation analysis

## Reactome Pathways analysis tool

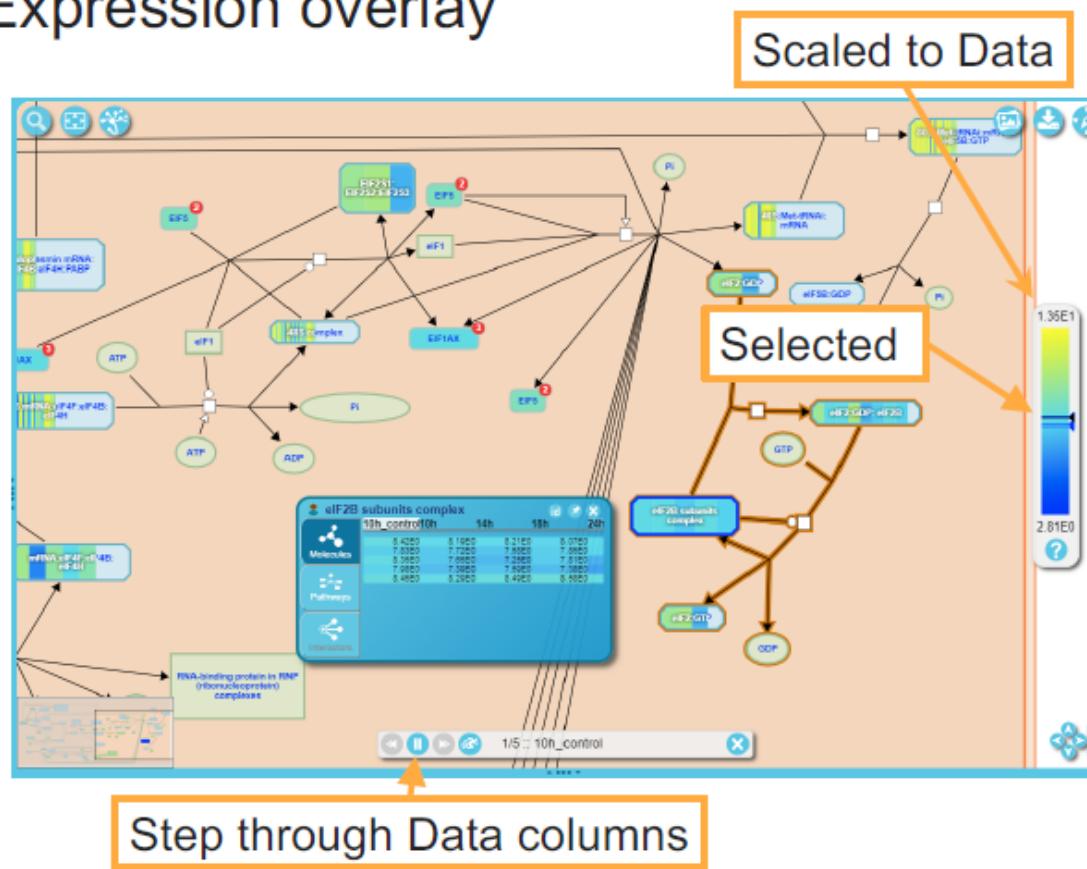
### Analysis Result – Over-representation



# Overrepresentation analysis

## Reactome Pathways analysis tool

### Expression overlay



# Overrepresentation analysis

## Reactome Pathways analysis tool

Try with your list of genes:

- Perform an overrepresentation analysis of the top up-regulated genes in B.PregVSLac comparison using Reactome Analysis web-based tool.
  - Examine the results and different outputs/tabs
  - What is the parent pathway of most of enriched terms?
  - Download a complete report of the results

# Overrepresentation analysis

## Reactome Pathways analysis tool

The screenshot shows the Reactome Pathways analysis tool interface. The top navigation bar includes a search bar for "Pathways for: Homo sapiens", citation and analysis buttons, a tour button, and layout options. A sidebar on the left displays the "Event Hierarchy" with a tree view of biological pathways, including "Autophagy", "Cell Cycle", and "Cell Cycle Checkpoints". The main area features a detailed pathway map with various biological processes like "DNA Repair", "Signal Transduction", and "Metabolism". A color scale on the right indicates overrepresentation, ranging from 0 (yellow) to 0.05 (blue). Below the map is a table titled "Overrepresentation analysis results for TOTAL Data submitted with no name". The table has columns for Pathway name, Entities found, Entities Total, Entities ratio, Entities pValue, Entities FDR, Reactions found, Reactions total, Reactions ratio, and Species. Key rows include "Cell Cycle, Mitotic" (40 entities, 0.042 ratio, 7.8E-9 pValue), "Cell Cycle" (44 entities, 0.051 ratio, 3.49E-8 pValue), and "Resolution of Sister Chromatid Cohesion" (17 entities, 0.009 ratio, 4.7E-8 pValue).

Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Species
Cell Cycle, Mitotic	40	596	0.042	7.8E-9	5.51E-6	177	350	0.026	Homo Sapiens
Cell Cycle	44	734	0.051	3.49E-8	1.1E-5	203	449	0.033	Homo Sapiens
Resolution of Sister Chromatid Cohesion	17	134	0.009	4.7E-8	1.1E-5	7	8	0.001	Homo Sapiens
Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	14	94	0.007	1.09E-7	1.54E-5	4	4	0	Homo Sapiens
Amplification of signal from the kinetochores	14	94	0.007	1.09E-7	1.54E-5	4	4	0	Homo Sapiens
Mitotic Spindle Checkpoint	15	111	0.008	1.32E-7	1.54E-5	6	7	0.001	Homo Sapiens
Cell Cycle Checkpoints	23	280	0.02	5.06E-7	5.06E-5	14	56	0.004	Homo Sapiens
Interleukin-10 signaling	12	86	0.006	1.72E-6	1.51E-4	3	15	0.001	Homo Sapiens
EML4 and NUDC in mitotic spindle formation	14	121	0.008	2.09E-6	1.63E-4	2	5	0	Homo Sapiens
RHO GTPases Activate Formins	15	149	0.01	4.82E-6	3.37E-4	4	27	0.002	Homo Sapiens

# Overrepresentation analysis

## Other tools

DAVID: <https://david.ncifcrf.gov/home.jsp>

The screenshot shows the DAVID Bioinformatics Database Analysis Wizard interface. At the top, there's a logo for DAVID Bioinformatics Database and another for LHRI. The menu bar includes Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, About DAVID, and About LHRI. Below the menu, there are tabs for Upload, List (which is selected), and Background. A sidebar on the left titled 'Gene List Manager' allows users to select species (e.g., Homo sapiens, Unknown) and manage gene lists (e.g., 'demolist2'). A main panel titled 'Analysis Wizard' displays the current gene list ('demolist2') and background ('Homo sapiens'). It also features a 'Tell us how you like the tool' section and a 'Contact us for questions' link. The main content area is divided into two sections: 'Step 1. Successfully submitted gene list' (with a blue checkmark) and 'Step 2. Analyze above gene list with one of DAVID tools'. Step 2 includes a large blue downward arrow, a 'Which DAVID tools to use?' link, and a list of available tools: Functional Annotation Tool (Functional Annotation Clustering, Functional Annotation Chart, Functional Annotation Table), Gene Functional Classification Tool, Gene ID Conversion Tool, and Gene Name Batch Viewer.

# Overrepresentation analysis

## Other tools

**DAVID:** <https://david.ncifcrf.gov/home.jsp>


**DAVID Bioinformatics Resources**  
 Laboratory of Human Retrovirology and  
 Immunoinformatics (LHRI)



**Functional Annotation Clustering**

Current Gene List: demolist2  
 Current Background: Homo sapiens  
 372 DAVID IDs

Options    Classification Stringency: Medium

Kappa Similarity	Similarity Term Overlap: 3	Similarity Threshold: 0.50	
Classification	Initial Group Membership: 3	Final Group Membership: 3	Multiple Linkage Threshold: 0.50
Enrichment Thresholds	EASE: 1.0		
Display	<input type="checkbox"/> Fold Change <input type="checkbox"/> Bonferroni <input checked="" type="checkbox"/> Benjamini <input type="checkbox"/> FDR <input type="checkbox"/> LT,PH,PT		

**91 Cluster(s)**

Annotation Cluster 1	Enrichment Score: 3.22	G	Count	P_Value	Benjamini
<input type="checkbox"/> GOTERM_BP_DIRECT	<a href="#">negative regulation of transcription from RNA polymerase II promoter</a>	RT	42	3.9E-7	1.0E-3
<input type="checkbox"/> GOTERM_BP_DIRECT	<a href="#">positive regulation of transcription from RNA polymerase II promoter</a>	RT	46	3.0E-6	2.7E-3
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">transcriptional repressor activity, RNA polymerase II transcription regulatory region</a>	RT	20	2.0E-5	3.3E-3
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">sequence-specific binding</a>	RT	26	6.2E-5	8.2E-3
<input type="checkbox"/> GOTERM_CC_DIRECT	<a href="#">chromatin</a>	RT	37	1.1E-4	7.0E-3
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">sequence-specific double-stranded DNA binding</a>	RT	25	1.7E-4	1.9E-2
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">DNA binding</a>	RT	44	2.1E-4	1.9E-2
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">transcriptional activator activity, RNA polymerase II transcription regulatory region</a>	RT	22	3.0E-4	1.9E-2
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">sequence-specific binding</a>	RT	17	5.7E-4	3.1E-2
<input type="checkbox"/> GOTERM_MF_DIRECT	<a href="#">RNA polymerase II core promoter proximal region sequence-specific DNA binding</a>	RT	39	1.5E-3	5.2E-2
<input type="checkbox"/> GOTERM_BP_DIRECT	<a href="#">regulation of transcription from RNA polymerase II promoter</a>	RT	49	1.8E-3	1.1E-1

 [Download File](#)

# Overrepresentation analysis

## Other tools

PantherDB: <http://www.pantherdb.org/>

 GENEONTOLOGY Unifying Biology

 PANTHER Classification System

LOGIN REGISTER CONTACT US

Home About PANTHER Data PANTHER Tools PANTHER Services Workspace Downloads Help/Tutorial

PANTHER17.0 Released.

Search

All

Quick links

Whole genome function views  
Genome statistics  
Data Version  
PANTHER API  
FAQ  
How to cite PANTHER  
Recent publication describing PANTHER

News

PANTHER17.0 Released.  
Click for additional info.

Newsletter subscription

Enter your Email:

Gene List Analysis      Browse      Sequence Search      cSNP Scoring      Keyword Search

Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page.

**Help Tips**  
Steps:  
► 1. Select list and list type to analyze  
► 2. Select Organism  
► 3. Select operation  
[Using enhancer data](#)

1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs: **Supported IDs**  separate IDs by a space or comma  
Upload  No s'ha seleccionat cap fitxer.  
IDs:   
Please [login](#) to be able to select lists from your workspace.

Select List  ID List  
Type:  Previously exported text search results  
 Workspace list  
 PANTHER Generic Mapping  
 ID's from Reference Proteome Genome  
Organism for id list   
 VCF File Flanking region   Search Enhancer Data

2. Select organism.

Homo sapiens  
Mus musculus  
Rattus norvegicus  
Gallus gallus  
Danio rerio

# Overrepresentation analysis

## Other tools



### WEB-based GEne SeT AnaLysis Toolkit

WebGestalt

Translating gene lists into biological insights...

<http://www.webgestalt.org/>

[ORA Sample Run](#) | [GSEA Sample Run](#) | [NTA Sample Run](#) | [Phosphosite Sample Run](#) (New in 2019!) | [External Examples](#) | [Manual \(PDF, Web\)](#) | [Citation](#) |  
[GOView](#) | [WebGestaltR](#) | [WebGestalt 2017](#)

**Basic parameters**

**Organism of Interest** ⓘ Mus musculus

**Method of Interest** ⓘ Over-Representation Analysis (ORA)

**Functional Database** ⓘ pathway

+ Reactome

**Gene List**

**Select Gene ID Type** ⓘ NCBI Entrez gene

**Upload Gene List** ⓘ Click to upload Reset

OR

100038566  
74562  
69065  
74152

**Reference Gene List**

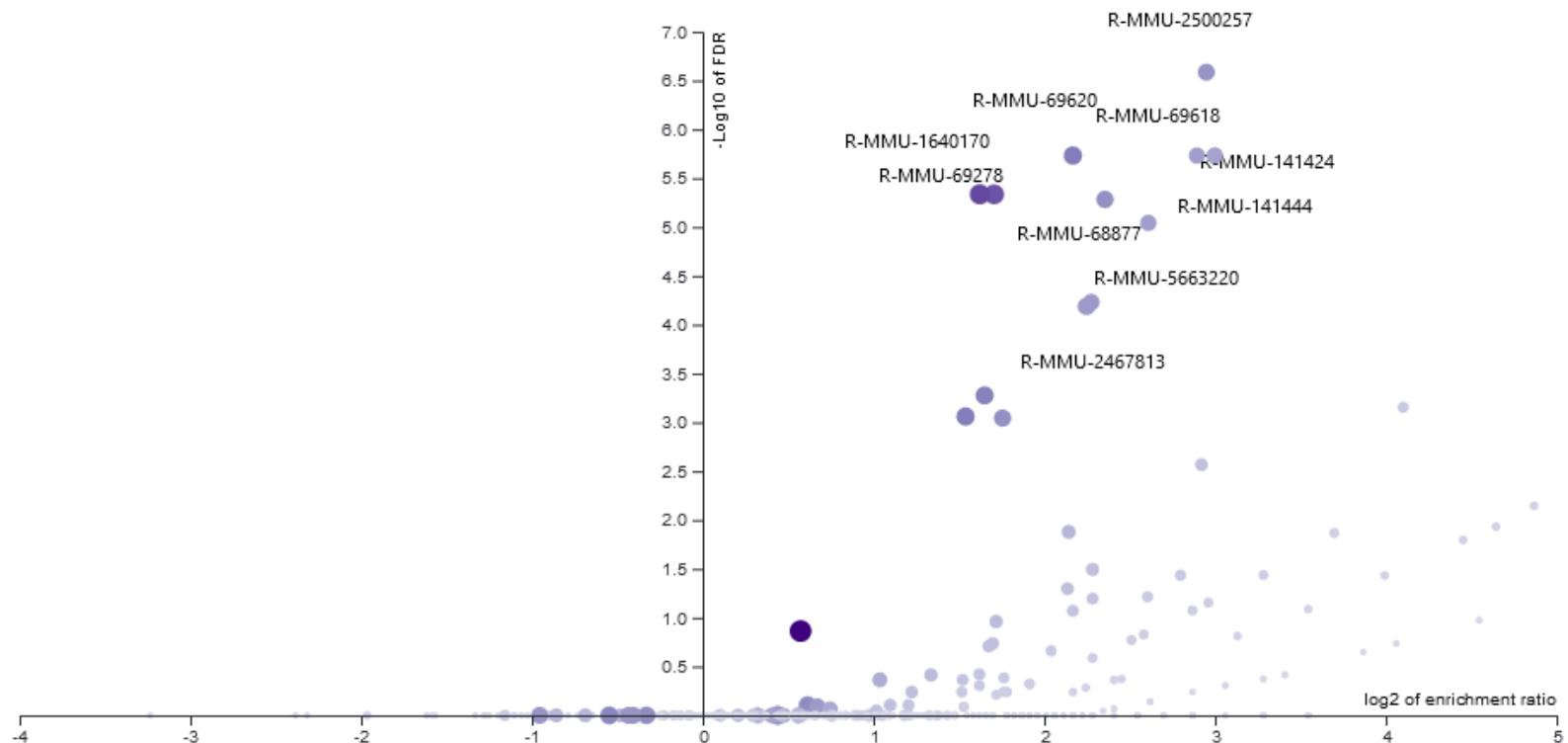
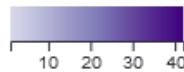
**Select Reference Set** ⓘ genome

**Upload User Reference Set** ⓘ Select the ID type of reference set

**File and Select ID type** ⓘ Click to upload Reset

# Overrepresentation analysis

## Other tools



Label: geneSet



PNG



Download

Draw Link

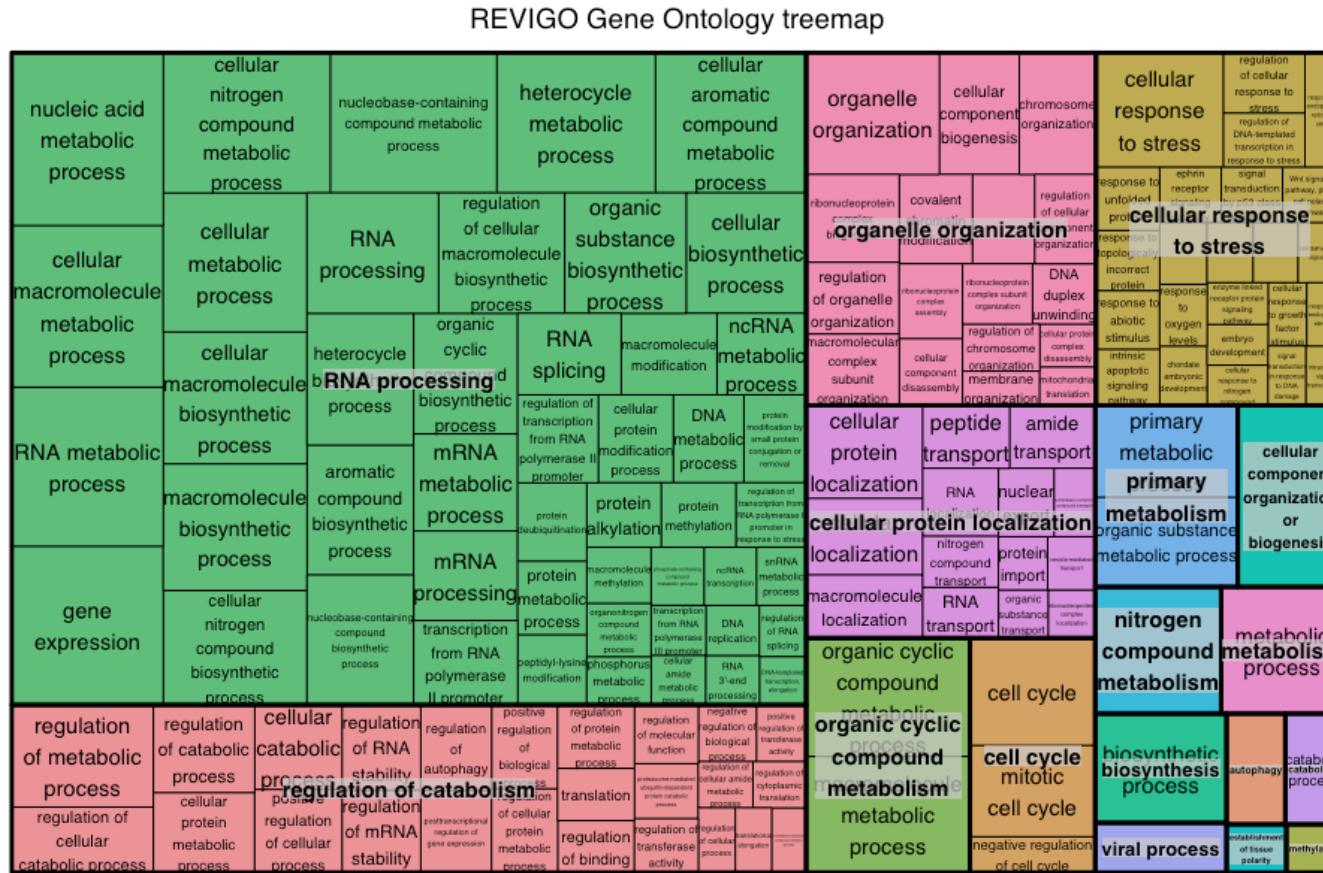


Pan & Zoom

# **Visualization and downstream analysis**

# Visualization and downstream analysis

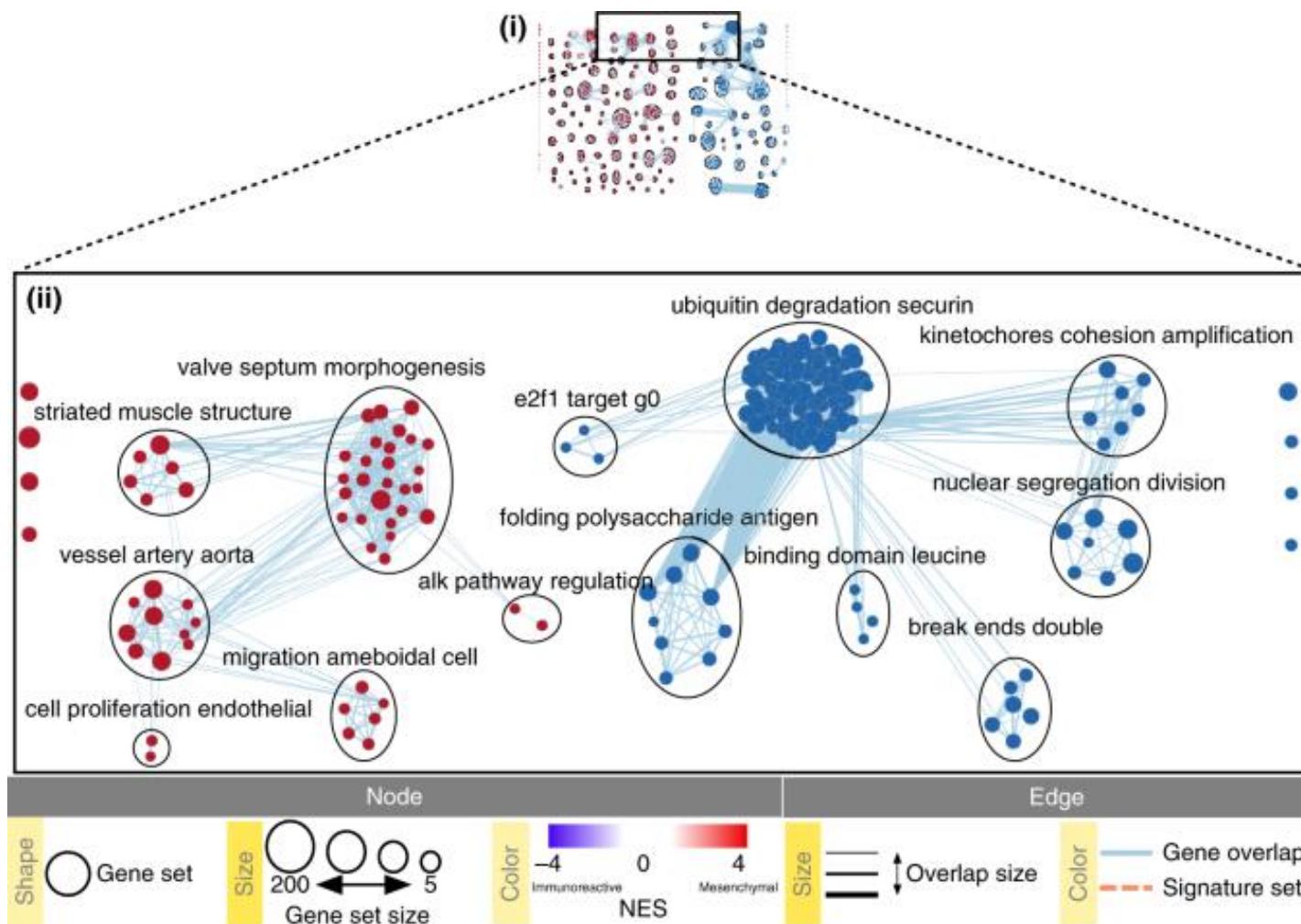
**REVIGO** is a web-based tool that can take our list of GO terms, collapse redundant terms by semantic similarity, and summarize them graphically.



# Visualization and downstream analysis

## EnrichmentMap Pipeline (Cytoscape)

- Nature Protocols: <https://www.nature.com/articles/s41596-018-0103-9>



# Hands on Pathway Analysis

## TABLE OF CONTENTS

1. Resources for functional analysis of gene lists
2. Navigating through different resources
  - a) Browsing the GO
  - b) Pathway visualization in Reactome
3. Case of study: RNA-seq experiment
  - a) Preparation of the gene list
  - b) Functional annotation of individual genes
  - c) Overrepresentation analysis of top DEG (g:Profiler / Reactome analysis tool)
  - d) Visualization and downstream analyses
4. A workflow in R (**clusterProfiler**)

# Hands On Pathway Analysis

## A workflow in R

Follow file: *Practicum\_EnrichmentAnalysis.Rmd / html*

# References

- [https://uebvhir.github.io/Pathway\\_Analysis-Guidelines/](https://uebvhir.github.io/Pathway_Analysis-Guidelines/)
- Galaxy tutorial:  
<https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-genes-to-pathways/tutorial.html>
- Training materials:  
[https://hbctraining.github.io/DGE\\_workshop\\_salmon/lessons/functional\\_analysis\\_2019.html](https://hbctraining.github.io/DGE_workshop_salmon/lessons/functional_analysis_2019.html)  
  
[https://www.ebi.ac.uk/training/online/courses/reactome-tools-for-analysing-biological-pathways/#vf-tabs\\_section--overview](https://www.ebi.ac.uk/training/online/courses/reactome-tools-for-analysing-biological-pathways/#vf-tabs_section--overview)