

Functional analysis of gene lists

A practical exploration of resources

Bioinformatics Course UEB-VHIR
November 2020

Ricardo Gonzalo¹, Mireia Ferrer¹, Àlex Sánchez^{1,2}
Berta Miró¹, Angel Blanco^{1,2}

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR

2 Departament de Genètica Microbiologia i Estadística, UB

Functional Analysis of gene lists

What's the biology behind a list of genes ?



Functional Analysis of gene lists

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 (airway data)
 - 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


Web tools

- 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/>



[Home](#)
[Install](#)
[Help](#)
[Developers](#)
[About](#)

[Home](#) » [BiocViews](#)

All Packages

http://bioconductor.org/packages/release/BiocViews.html#___Pathways

Bioconductor version 3.12 (Release)

Autocomplete [biocViews](#) search:

[ResearchField](#) (902)
[StatisticalMethod](#) (727)
[Technology](#) (1251)
[WorkflowStep](#) (1081)
 [Alignment](#) (80)
 [Annotation](#) (124)
 [BatchEffect](#) (52)
 [ExperimentalDesign](#) (22)
 [GenomeBrowsers](#) (2)
 [MultipleComparison](#) (163)
 [Normalization](#) (128)
 [Pathways](#) (179)
 [BioCarta](#) (7)
 [GO](#) (72)
 [KEGG](#) (51)
 [Reactome](#) (19)

Packages found under GO:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show entries Search table:

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

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, visualisation, post-processing

Resources for functional analysis

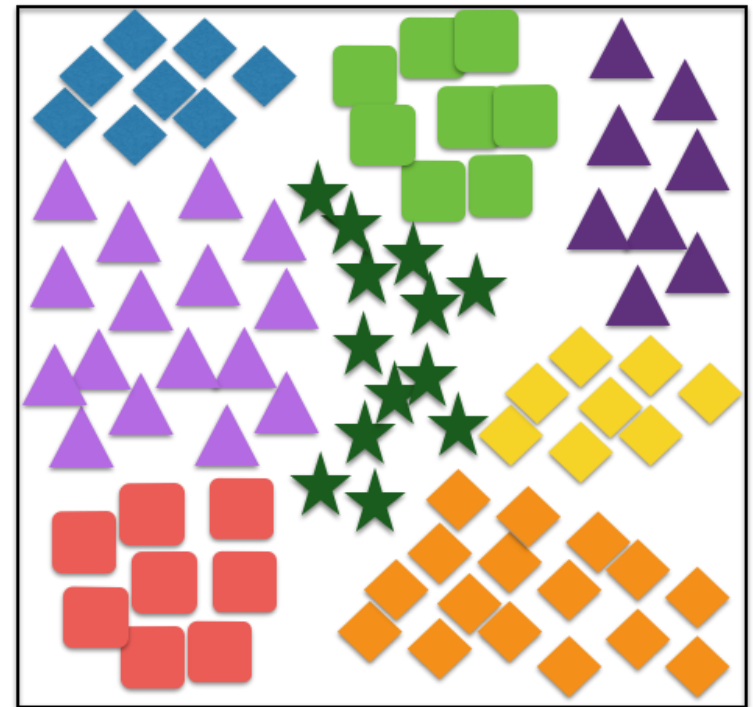
Resources of functional annotations

- 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

- The Gene Ontology
- Pathway databases
- Disease-related DB
- Cell/tissue markers
- Chromosome locations
- Transcription factor DB
- Micro-RNA / targets DB

All known genes in a species
(categorized into groups)



Resources of functional annotations

- [Gene Ontology](#) (GO)
 - Cellular Components / Biological Processes / Molecular Function
- Pathways Databases
 - [KEGG](#), [Reactome](#), [WikiPathways](#) [PantherDB](#)
- Disease-related
 - [DisGeNet](#), [Disease Ontology](#)
- Cell/Tissue markers databases
 - [CellMarker](#), [Cellfinder](#)
- Chromosomal location
- [Transcription Factors binding sites DBs](#), TRANSFAC
- MicroRNA-Target interactions DB

Resources for functional analysis



MSigDB
Molecular Signatures
Database

Molecular Signatures Database v7.2

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*

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.

Browsing the Gene Ontology

Functional Analysis of gene lists

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

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 >

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, 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

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 Reactome Pathways database

5. Introduction to Reactome and Pathway visualization

<https://reactome.org/>

The screenshot shows the Reactome website homepage. At the top is the Reactome logo and a navigation bar with links: About, Content, Docs, Tools, Community, and Download. Below this is a search bar with the text "Find Reactions, Proteins and Pathways" and a search input field containing "e.g. O95631, NTN1, signaling by EGFR, glucose". A "Go!" button is to the right of the input field. Below the search bar are four main sections, each with an icon and a title:

- Pathway Browser**: Visualize and interact with Reactome biological pathways. Icon: A tree diagram.
- Analyze Data**: Merges pathway identifier mapping, over-representation, and expression analysis. Icon: A bar chart.
- ReactomeFIViz**: Designed to find pathways and network patterns related to cancer and other types of diseases. Icon: A network diagram.
- Documentation**: Information to browse the database and use its principal tools for data analysis. Icon: A document with lines of text.

Below these sections is a dark banner with the text "USE REACTOME GRAPH DATABASE IN YOUR PROJECT" and a "LEARN MORE" button. At the bottom, there are two columns of content:

- Why Reactome**: A section with a magnifying glass icon. Text: "Reactome is a free, open-source, curated and peer-reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. If you use Reactome in Asia, we suggest using our Chinese mirror site at reactome.ncpsb.org."
- Tweets**: A section with a Twitter icon. Text: "reactome Retweeted ChEMBL Database @ChEMBL #ChEMBL is 10 years old this year :) For more about the celebrations see: chembl.blogspot.com/2019/02/chembl..."

Tutorials: <http://www.reactome.org/userguide/Usersguide.html>
<https://www.ebi.ac.uk/training/online/>

5. Introduction to Reactome and Pathway visualization

Reactome is...

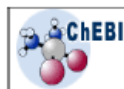
Free, online, open-source curated
database of pathways and
reactions in human biology

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

Mapped to cellular compartment

Extensively cross-referenced

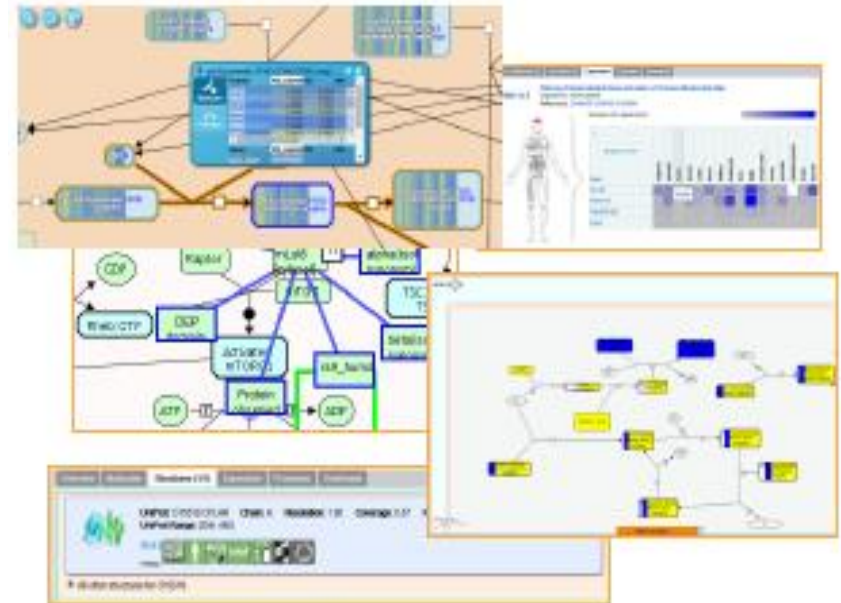
Entrez Gene



5. Introduction to Reactome and Pathway visualization

Reactome Tools

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



5. Introduction to Reactome and Pathway visualization

The Pathway Browser

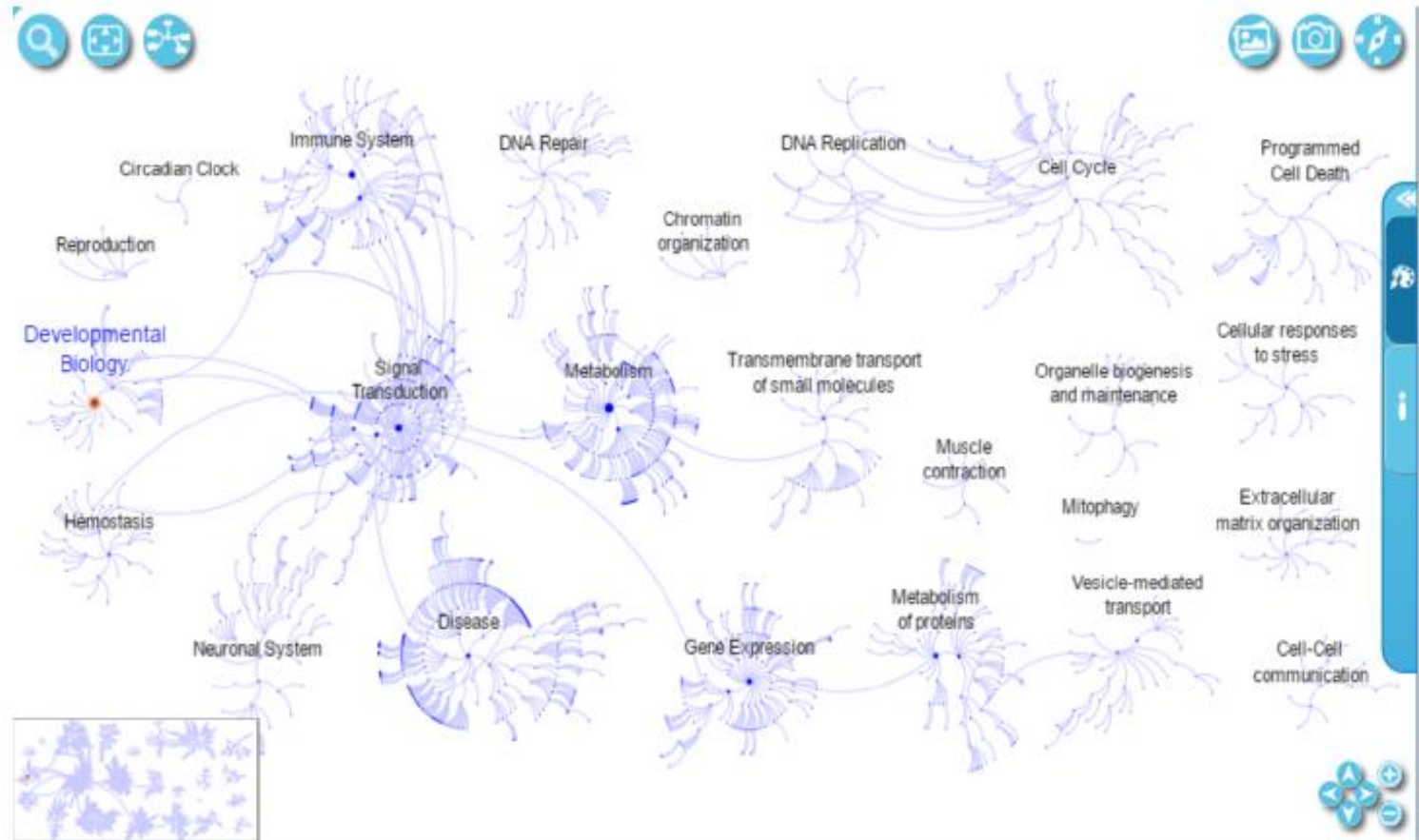
The image shows the Reactome Pathway Browser interface with several components labeled:

- Home**: A button in the top navigation bar.
- Species**: A button in the top navigation bar.
- Analyse Data**: A button in the top navigation bar.
- Video Tour**: A button in the top navigation bar.
- Layout**: A button in the top navigation bar.
- Key**: A button in the top right corner.
- Export**: A button in the top right corner.
- Settings Sidebar**: A sidebar on the right side of the interface.
- Zoom/Move**: A button in the bottom right corner.
- Open Diagram**: A button in the top left corner of the main panel.
- Search Diagram**: A button in the top left corner of the main panel.
- Fit to Page**: A button in the top left corner of the main panel.
- Pathway Panel**: The main area displaying the pathway diagram.
- Thumbnail**: A small preview of the pathway diagram in the bottom left corner.
- Detail Panel**: A panel at the bottom of the interface displaying details about the selected item.
- Hierarchy Panel**: A sidebar on the left side of the interface showing the event hierarchy.
- Illustrations**: A button in the top right corner of the main panel.

The interface also includes a top navigation bar with buttons for Home, Species, Analyse Data, Video Tour, and Layout. The main panel displays a complex pathway diagram with various biological processes labeled, such as Cell Cycle, Metabolism, and Signal Transduction. The bottom of the interface features a tabbed interface with tabs for Description, Molecules, Structures, Expression, Analysis, and Downloads. The Detail Panel at the bottom provides information about the selected item, including its description, input and output molecules, and supporting evidence.

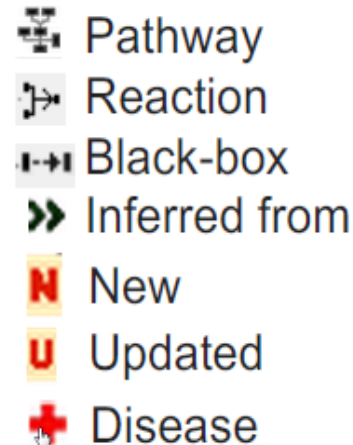
5. Introduction to Reactome and Pathway visualization

Pathway Overview



5. Introduction to Reactome and Pathway visualization

Hierarchy Panel



5. Introduction to Reactome and Pathway visualization

Navigating in the Pathway Browser

Home button

The screenshot displays the Reactome Pathway Browser interface. On the left, a sidebar lists various biological processes, with 'Signal Transduction' and 'Signaling by NOTCH' highlighted. The main area shows a network diagram of pathways. An orange box labeled 'Home button' points to the 'reactome' logo. Another orange box labeled 'Click pathway to select, double-click to open diagram' points to a pathway node in the network. A third orange box labeled 'Click reaction to open' points to a reaction in the 'Notch Signaling' pathway. A fourth orange box labeled 'Details here' points to the 'Description' tab in the pathway details panel.

reactome

Pathways for: Homo sapiens

Analysis: Tour Layout

Signal Transduction

Click pathway to select, double-click to open diagram

Click reaction to open

Details here

Signaling by NOTCH

Notch Signaling

DO1

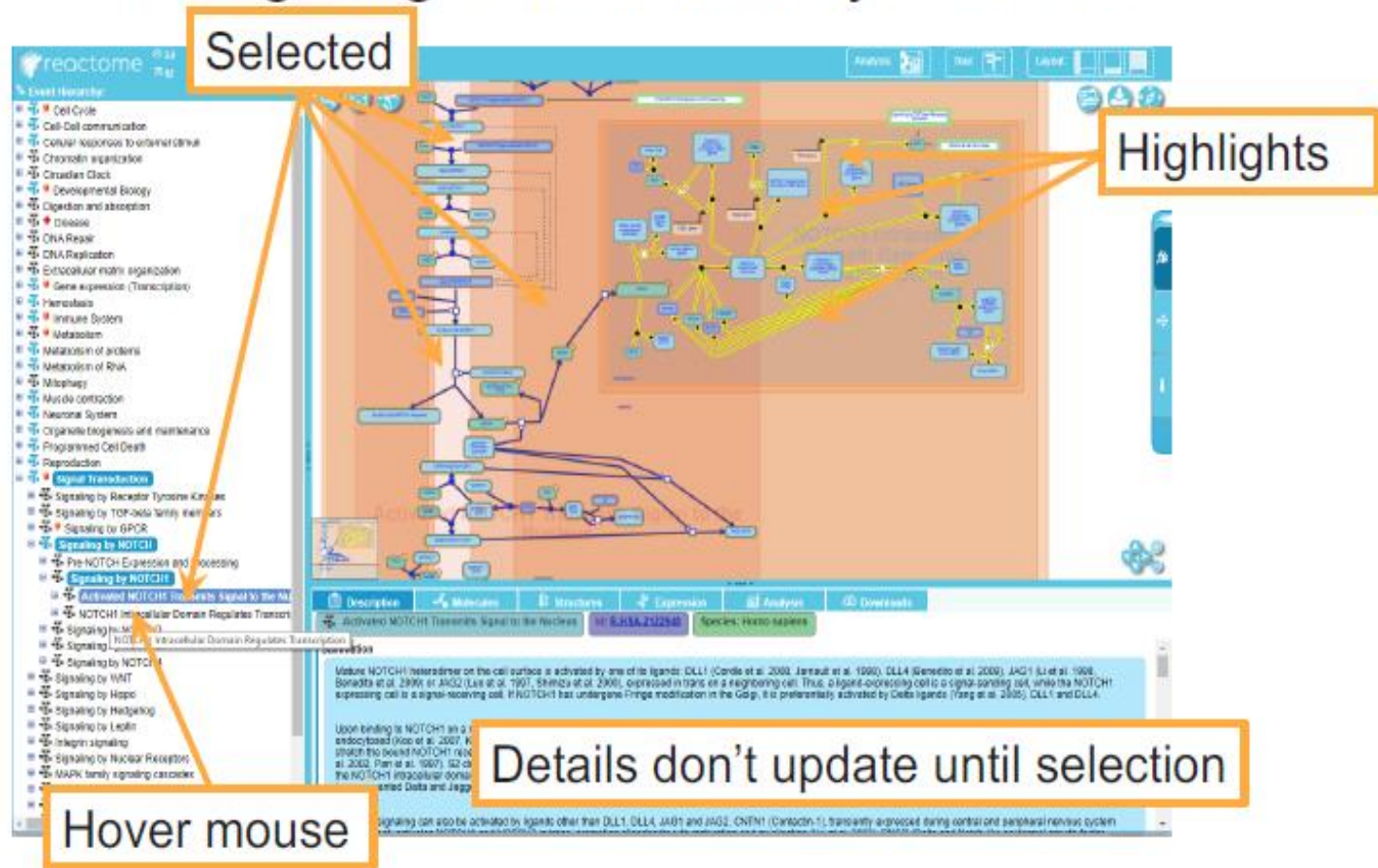
Summation

The Notch Signaling Pathway (NSP) is a highly conserved pathway for cell-cell communication. NSP is involved in the regulation of cellular differentiation, proliferation, and specification. For example, it is utilized by continually renewing adult tissues such as blood, skin, and gut epithelium not only to maintain stem cells in a proliferative, pluripotent, and undifferentiated state but also to direct the cellular progeny to adopt different developmental cell fates. Analogously, it is used during embryonic development to create fine-grained patterns of differentiated cells, notably during neurogenesis where the NSP controls switches such as that of the embryonic stem cell. This process is known as lateral inhibition, a molecular mechanism whereby individual cells direct neighbours from doing the same. The NSP has been adopted by several other biological systems to divide the growing embryo into regular blocks called somites which eventually generate the vertebral column. The NSP is also involved in the regulation of the cell cycle and the cell fate decisions during embryonic development. The NSP is involved in the regulation of the cell cycle and the cell fate decisions during embryonic development. The NSP is involved in the regulation of the cell cycle and the cell fate decisions during embryonic development.

Background literature references...

5. Introduction to Reactome and Pathway visualization

Navigating in the Pathway Browser



5. Introduction to Reactome and Pathway visualization

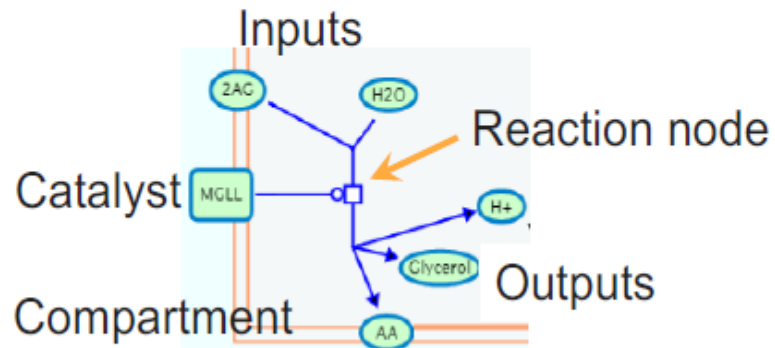
The Pathway Browser - Pathway Diagrams

Ovals are small molecules (or sets of)

Green boxes are proteins,

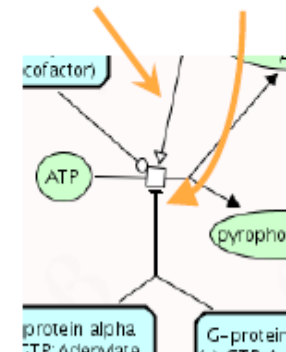
Blue are complexes,

Blue with double-boundary are sets



Regulation

+ve -ve



Transition



Binding



Dissociation



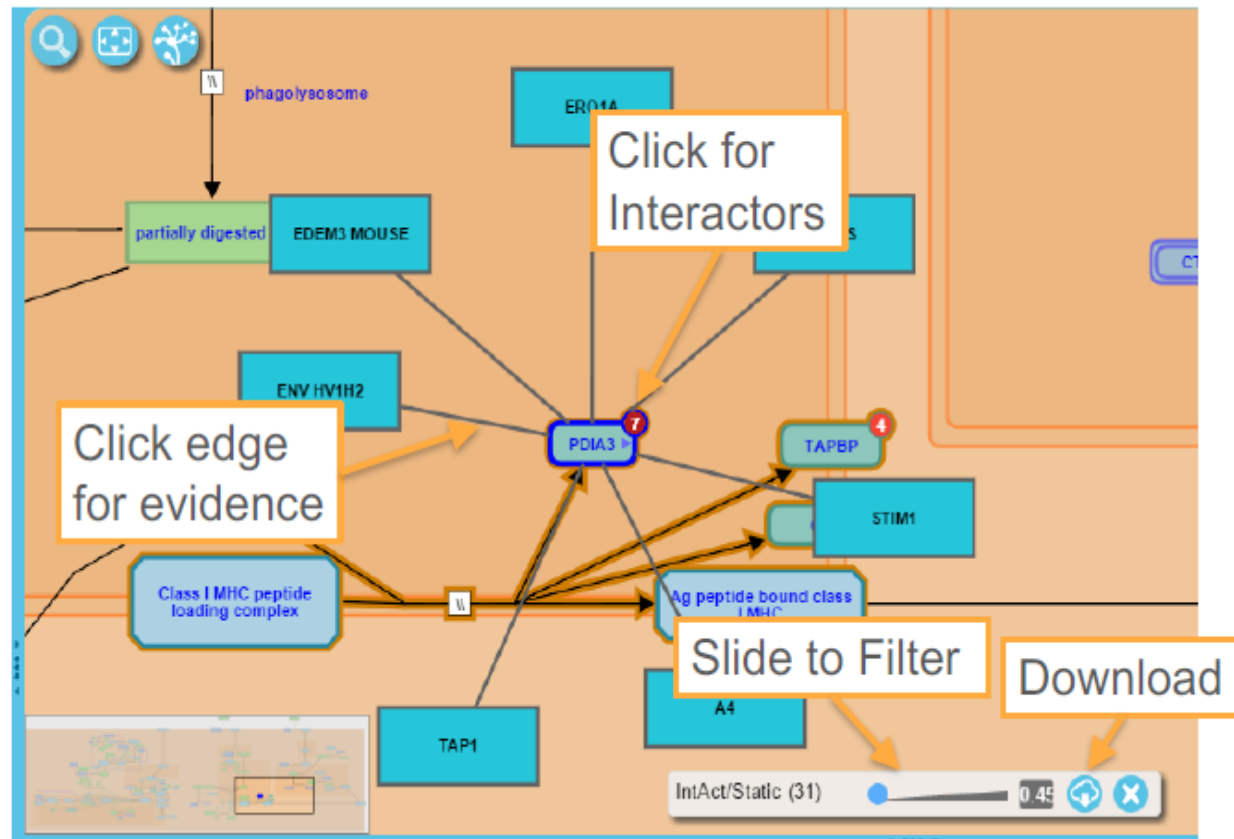
Omitted



Uncertain

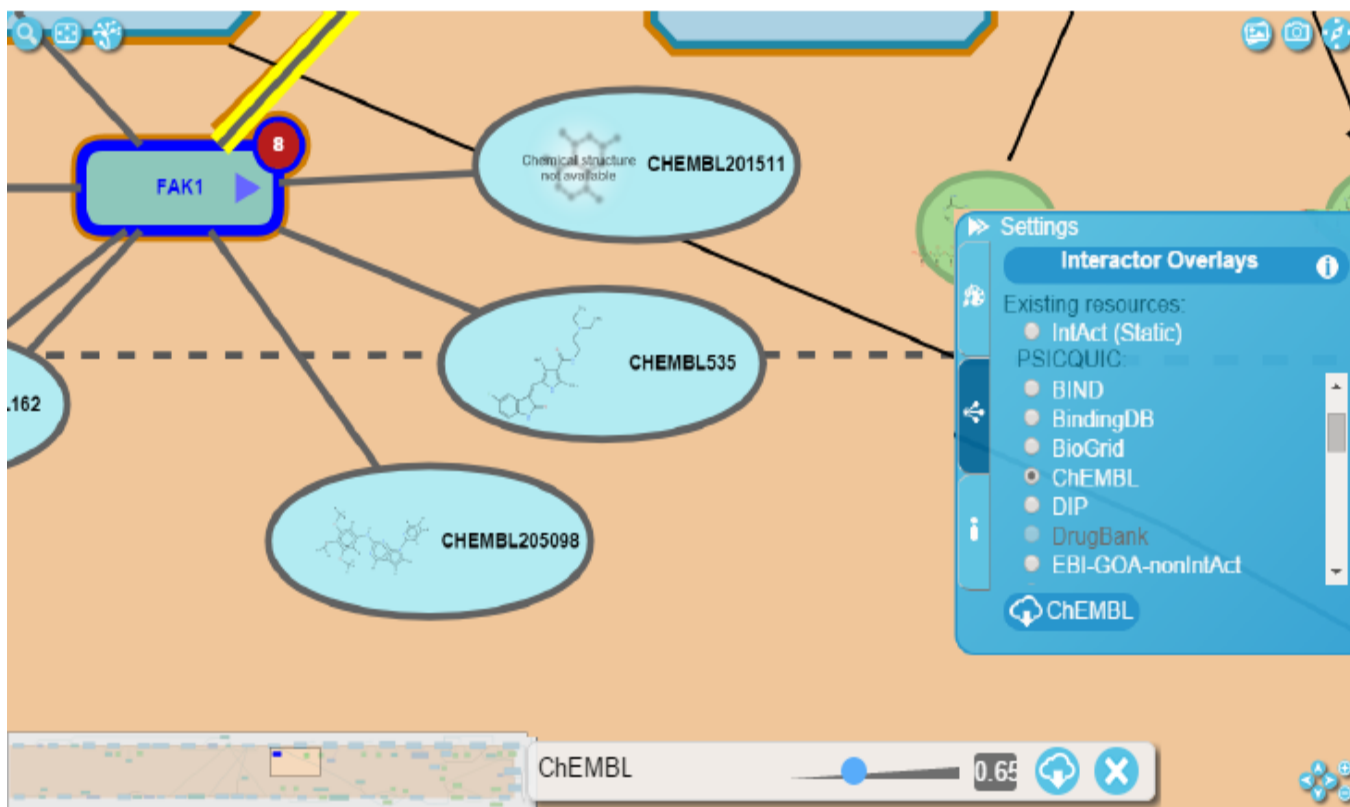
5. Introduction to Reactome and Pathway visualization

Interactors



5. Introduction to Reactome and Pathway visualization

Molecular Interaction Overlay – Set source



5. Introduction to Reactome and Pathway visualization

Show Illustration

REACTOME v9.2.2 Pathways for Homo sapiens Analysis Tour Layout

Event Hierarchy:

- Effects of PIP2 hydrolysis
- Response to elevated platelet cytosolic Ca²⁺
- Formation of Fibrin Clot (Clotting Cascade)
- Dissolution of Fibrin Clot
- Cell surface interactions at the vascular wall
- Factors involved in megakaryocyte development
- Immune System
- Mitophagy
- Metabolism
- Metabolism of proteins
- Muscle contraction
- Neuronal System
- Organelle biogenesis and maintenance**
 - Mitochondrial biogenesis
 - Mitochondrial translation
 - Assembly of the primary cilium**
 - Anchoring of the basal body to the plasma membrane
 - Cargo trafficking to the periciliary membrane
 - Intraflagellar transport
 - ATAT acetylates microtubules
 - HDAC6 deacetylates microtubules
 - Programmed Cell Death
 - Apoptosis
 - Regulated Necrosis
 - Reproduction
 - Signal Transduction

ASSEMBLY OF THE PRIMARY CILIUM
INTRAFLAGELLAR TRANSPORT AND CARGO TRAFFICKING

Diagram illustrating the assembly of the primary cilium, focusing on intraflagellar transport (IFT) and cargo trafficking. The diagram shows the cilium structure, including the basal body, transition zone, and axoneme. Key components and processes labeled include:

- CILIOGENESIS:** Ciliary vesicle, distal appendages, ciliary sheath, mother centriole.
- INTRAFLAGELLAR TRANSPORT:** Anterograde transport (IFT particles: IFTB, IFT20, IFT1), Retrograde transport (BBSome, ARLE).
- Basal Body and Transition Zone:** Y-links, transition fibres, basal body, transition pocket.
- Cargo Trafficking:** BBSome, ARLE, BBSome, ARLE, BBSome, ARLE.

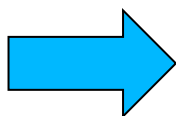
Analysis results are shown here when an analysis has been run. To start an analysis, click on the Analyse Data button in the top bar.

5. Introduction to Reactome and Pathway visualization

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...)

5. Introduction to Reactome and Pathway visualization

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?

Case Study

[PLoS One](#). 2014 Jun 13;9(6):e99625. doi: 10.1371/journal.pone.0099625. eCollection 2014.

RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells.

[Himes BE](#)¹, [Jiang X](#)², [Wagner P](#)², [Hu R](#)², [Wang Q](#)², [Klanderman B](#)³, [Whitaker RM](#)⁴, [Duan Q](#)⁴, [Lasky-Su J](#)⁴, [Nikolos C](#)⁵, [Jester W](#)⁵, [Johnson M](#)⁵, [Panettieri RA Jr](#)⁵, [Tantisira KG](#)⁴, [Weiss ST](#)⁶, [Lu Q](#)².

Introduction to Functional Analysis

The differential expression analysis performed on RNA-seq data (airway dataset) returned some genes differentially expressed in **DEX-Treated vs Untreated 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**.

```
res <- results(dds, contrast=c("dex","trt","untrt"))
res
```

log2 fold change (MLE): **dex trt vs untrt**

Wald test p-value: dex trt vs untrt

DataFrame with 31604 rows and 6 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSG00000000003.14	739.940717	-0.3611537	0.106869	-3.379419	0.000726392	0.00531137
ENSG000000000419.12	511.735722	0.2063147	0.128665	1.603509	0.108822318	0.29318870
ENSG000000000457.13	314.194855	0.0378308	0.158633	0.238479	0.811509461	0.92255697
ENSG000000000460.16	79.793622	-0.1152590	0.314991	-0.365912	0.714430444	0.87298038
ENSG000000000938.12	0.307267	-1.3691185	3.503764	-0.390757	0.695977205	NA
...
ENSG00000285979.1	38.353886	0.3423657	0.359511	0.952310	0.340940	0.600750
ENSG00000285987.1	1.562508	0.7064145	1.547295	0.456548	0.647996	NA
ENSG00000285990.1	0.642315	0.3647333	3.433276	0.106235	0.915396	NA
ENSG00000285991.1	11.276284	-0.1165515	0.748601	-0.155692	0.876275	0.952921
ENSG00000285994.1	3.651041	-0.0960094	1.068660	-0.089841	0.928414	NA

Exercise 2

- 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

Tip: You can use excel (with caution) or **R** to filter your TopTable

- Under these criteria, how many genes are up-regulated? And how many down-regulated?
- What is the **Entrez ID** of your top **most up/down-regulated gene**? And its **UniProt ID** (SwissProt)?

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

Functional annotation of individual genes

Id mapping

Searching for a gene in different databases

2. Preparation of the Gene list

ID mapping with g:Profiler

g:Profiler^β Contact Documentation Cite g:Profiler Archives

g:GOST Functional profiling g:Convert Gene ID conversion g:Orth Orthology search g:SNPense SNP id to gene name

Query ?

1552807_a_at
210982_s_at
203915_at
208894_at
202833_s_at
201137_s_at
205242_at
209924_at
223502_s_at
204670_x_at
219505_at
228532_at
211429_s_at
229560_at
32128_at
212671 s at

Run query

Options

Organism: ?
Homo sapiens (Human)

Target namespace
ENTREZGENE_ACC

Numeric IDs treated as
AFFY_HUGENE_2_0_ST_V1

Output

Input

2. Preparation of the Gene list

ID mapping with g:Profiler

Run query

Export to CSV

Show query URL

Copy URL to clipboard

initial alias	converted alias	name	namespace
1552807_A_AT	89790	SIGLEC10	AFFY_HG_U133_PLUS_2
210982_S_AT	3122	HLA-DRA	AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C, HGNC:4...
203915_AT	4283	CXCL9	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
208894_AT	3122	HLA-DRA	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
202833_S_AT	5265	SERPINA1	AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C, HGNC:4...
201137_S_AT	3115	HLA-DPB1	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
205242_AT	10563	CXCL13	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
209924_AT	6362	CCL18	AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C, HGNC:4...
223502_S_AT	10673	TNFSF13B	AFFY_HG_U133B, AFFY_HG_U133C, AFFY_HG_U133D, HGNC:4...
204670_X_AT	105369230	HLA-DRB1	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
204670_X_AT	3123	HLA-DRB1	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
219505_AT	51816	ADA2	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
228532_AT	128346	C1orf162	AFFY_HG_U133B, AFFY_HG_U133C, AFFY_HG_U133D, HGNC:4...
211429_S_AT	5265	SERPINA1	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
209969_S_AT	6772	STAT1	AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C, HGNC:4...
209619_AT	972	CD74	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
206881_S_AT	None	None	
207111_AT	2015	ADGRE1	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
212788_X_AT	2512	FTL	AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C, HGNC:4...
204118_AT	962	CD48	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
202902_S_AT	1520	CTSS	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C
229390_AT	441168	CALHM6	AFFY_HG_U133B, AFFY_HG_U133C, AFFY_HG_U133D, HGNC:4...
204588_S_AT	9056	SLC7A7	AFFY_HG_FOCUS, AFFY_HG_U133A, AFFY_HG_U133B, AFFY_HG_U133C

Export results as
gProfiler_IDconvert_GSE7586_sel.csv

Some ambiguities

Some missing

2. Preparation of the Gene list

ID mapping with g:Profiler

228055_at
220578_at
202464_s_at
215925_s_at
215633_x_at
204103_at
214084_x_at
226354_at
203729_at
200701_at
206991_s_at
205270_s_at
209901_x_at
221269_s_at
220330_s_at

Organism:
Homo sapiens (Human)

Target namespace
UNIPROTSSWISSPROT

Numeric IDs treated as
AFFY_HUGENE_2_0_ST_V1

Run query

Export to CSV Show query URL Copy URL to clipboard

initial alias	converted alias	name	description	namespace
1552807_A_AT	Q96LC7	SIGLEC10	sialic acid binding Ig like lectin 10 [Source:HGNC Symbol;Acc:HGNC:15620]	AFFY_HG_U133_PLUS_2
210982_S_AT	P01903	HLA-DRA	major histocompatibility complex, class II, DR alpha [Source:HGNC Symbol;Acc:HGNC:4...	AFFY_HG_U133A, AFFY_HG_U133,
203915_AT	Q07325	CXCL9	C-X-C motif chemokine ligand 9 [Source:HGNC Symbol;Acc:HGNC:7098]	AFFY_HG_FOCUS, AFFY_HG_U133
208894_AT	P01903	HLA-DRA	major histocompatibility complex, class II, DR alpha [Source:HGNC Symbol;Acc:HGNC:4...	AFFY_HG_FOCUS, AFFY_HG_U133
202833_S_AT	P01009	SERPINA1	serpin family A member 1 [Source:HGNC Symbol;Acc:HGNC:8941]	AFFY_HG_U133A, AFFY_HG_U133,
201137_S_AT	P04440	HLA-DPB1	major histocompatibility complex, class II, DP beta 1 [Source:HGNC Symbol;Acc:HGNC:4...	AFFY_HG_FOCUS, AFFY_HG_U133
205242_AT	Q43927	CXCL13	C-X-C motif chemokine ligand 13 [Source:HGNC Symbol;Acc:HGNC:10639]	AFFY_HG_FOCUS, AFFY_HG_U133
209924_AT	P55774	CCL18		AFFY_HG_U133A, AFFY_HG_U133,
223502_S_AT	Q9V275	TNFSF13B		AFFY_HG_U133B, AFFY_HG_U133,
204670_X_AT	P01911	HLA-DRB1		NC4... AFFY_HG_FOCUS, AFFY_HG_U133
204670_X_AT	P04229	HLA-DRB1		NC4... AFFY_HG_FOCUS, AFFY_HG_U133
204670_X_AT	Q29974	HLA-DRB1		NC4... AFFY_HG_FOCUS, AFFY_HG_U133
204670_X_AT	Q9GT93	HLA-DRB1		NC4... AFFY_HG_FOCUS, AFFY_HG_U133
219505_AT	Q9NZK5	ADA2	adenosine deaminase 2 [Source:HGNC Symbol;Acc:HGNC:1839]	AFFY_HG_FOCUS, AFFY_HG_U133
228532_AT	Q8NEQ5	C1orf162	chromosome 1 open reading frame 162 [Source:HGNC Symbol;Acc:HGNC:28344]	AFFY_HG_U133B, AFFY_HG_U133,
211429_S_AT	P01009	SERPINA1	serpin family A member 1 [Source:HGNC Symbol;Acc:HGNC:8941]	AFFY_HG_FOCUS, AFFY_HG_U133
229560_AT	Q9NR97	TLR8	toll like receptor 8 [Source:HGNC Symbol;Acc:HGNC:15632]	AFFY_HG_U133B, AFFY_HG_U133,
221138_AT	P55774	CCL18	C-X-C motif chemokine ligand 18 [Source:HGNC Symbol;Acc:HGNC:10639]	AFFY_HG_FOCUS, AFFY_HG_U133

Different protein isoforms (check in UniProt)

3. 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/>

3. Functional annotation of individual genes

- Functional annotation of individual genes with QuickGO



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 <i>LACTB</i> , mitochondrial	PROTEIN	Homo sapiens	14 annotations
UniProtKB	H0YNN5	Serine beta-lactamase-like protein <i>LACTB</i> , mitochondrial	PROTEIN	Homo sapiens	
UniProtKB	Q53H82	Endoribonuclease <i>LACTB2</i>	PROTEIN	Homo sapiens	16 annotations
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	

3. Functional annotation of individual genes

- Functional annotation of individual genes with QuickGO

[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

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  IEA	GO_REF:0000107	UniProtKB:P83111 more...	61621 <i>Rhinopithecus bieti</i>	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	enables	GO:0008233    peptidase activity	ECO:0000265  IEA	GO_REF:0000107	UniProtKB:P83111 more...	61621 <i>Rhinopithecus bieti</i>	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	involved_in	GO:0006508    proteolysis	ECO:0000265  IEA	GO_REF:0000107	UniProtKB:P83111 more...	61621 <i>Rhinopithecus bieti</i>	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	part_of	GO:0005829    cytosol	ECO:0000265  IEA	GO_REF:0000107	UniProtKB:P83111 more...	61621 <i>Rhinopithecus bieti</i>	Ensembl
UniProtKB:A0A2K6LNQ6	LACTB	part_of	GO:0005739    mitochondrion	ECO:0000265  IEA	GO_REF:0000107	UniProtKB:P83111 more...	61621 <i>Rhinopithecus bieti</i>	Ensembl

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

3. Functional annotation of individual genes

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

PubMed [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

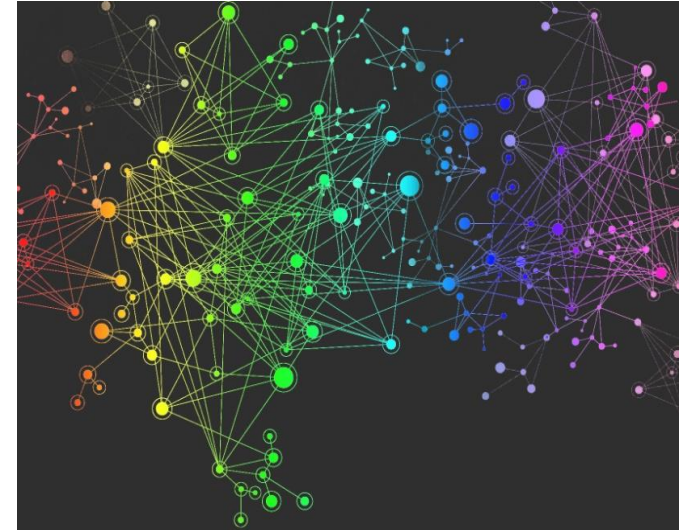
Items: 1 to 20 of 370

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



4. Enrichment analysis of selected genes

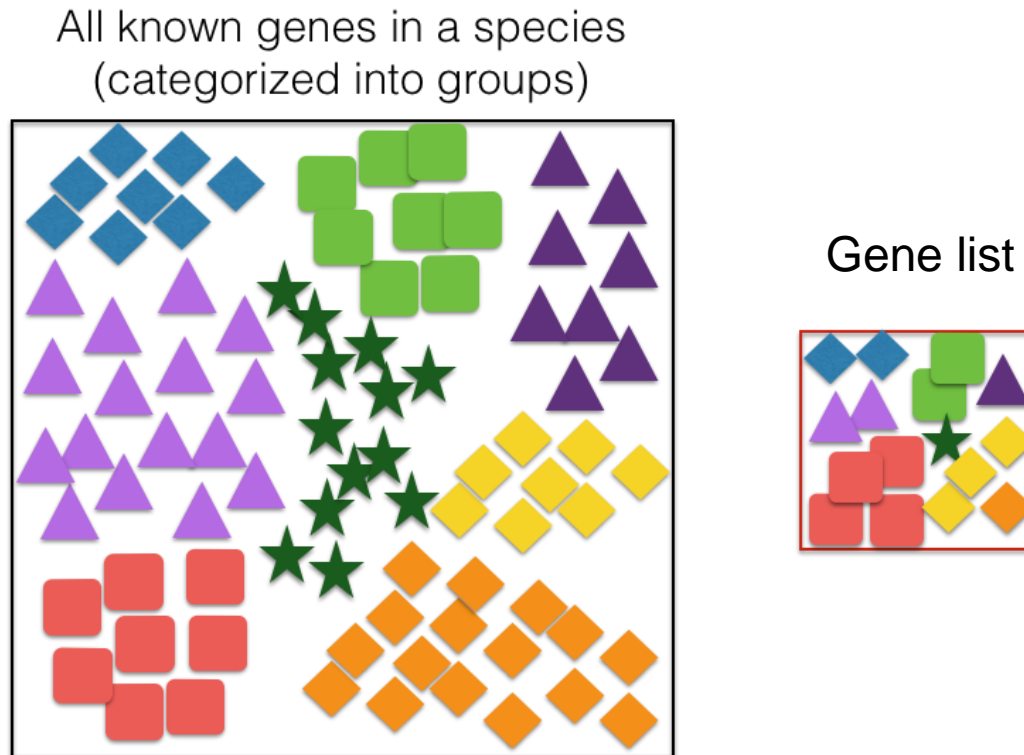
- Genes usually don't act on their own
- Extracting information at the “group” level may be more meaningful and facilitate interpretation of the processes involved in the condition of study



Overrepresentation analysis

G:profiler
Gorilla
Reactome

- you can determine the probability of having the observed proportion of genes associated with a specific category in your gene list based on the proportion of genes associated with the same category in the background set (gene categorizations for the appropriate organism).



4. Enrichment analysis of selected genes

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. *.gaf* file for GO)

❑ Structure of relationships between annotation terms

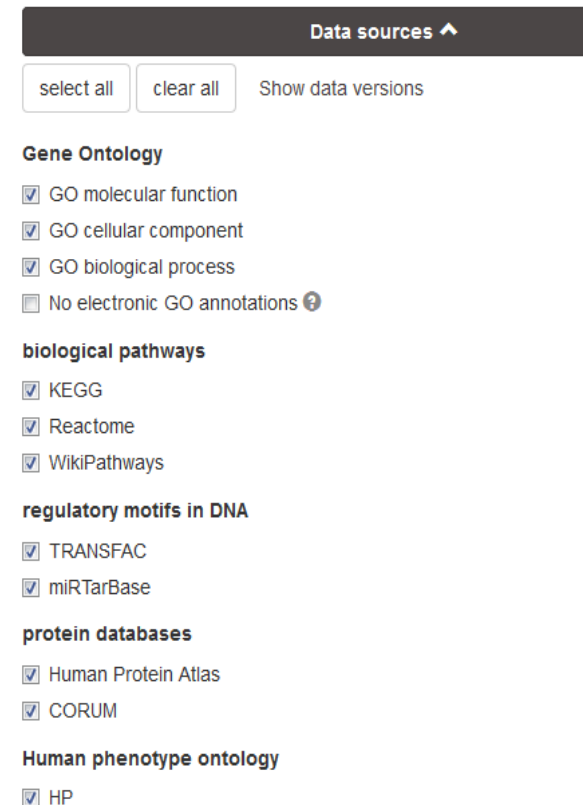
From annotation database
(eg. *.obo* file for GO)

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

4. Enrichment analysis of selected genes

g:Profiler

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



The screenshot shows the 'Data sources' panel in the g:Profiler web interface. At the top, there is a dark grey header with the text 'Data sources' and an upward-pointing arrow. Below the header, there are two buttons: 'select all' and 'clear all', followed by a link 'Show data versions'. The panel is organized into several categories, each with a bold heading and a list of items with 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

4. Enrichment analysis of selected genes

g:Profiler

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

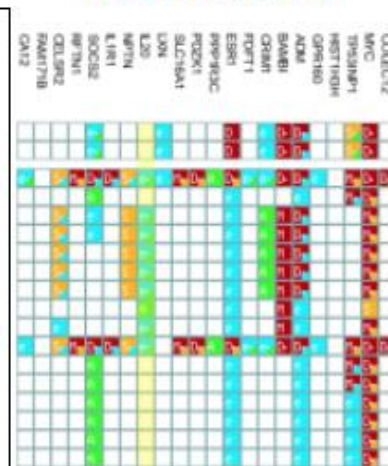
source	term name	term ID
BP	negative regulation of biological process	GO:0048519
BP	negative regulation of cellular process	GO:0048523
BP	single-organism process	GO:0044699
BP	death	GO:0016265
BP	developmental process	GO:0032502
BP	single-organism developmental process	GO:0044767
BP	Multicellular organismal development	GO:0007275
BP	anatomical structure development	GO:0048056
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:0006219
BP	programmed cell death	GO:0012901
BP	apoptotic process	GO:0006915
BP	regulation of cell death	GO:0010941
BP	regulation of programmed cell death	GO:0043067
BP	regulation of apoptotic process	GO:0042981

size of the overlap

significance

> Genes in input list

n. of term genes	n. of query genes	n. of common genes	p-value
4340	391	127	2.29e-03
3985	391	119	2.17e-03
13478	391	305	1.49e-02
1945	391	69	3.41e-03
9560	391	150	1.82e-02
5476	391	148	2.00e-02
4696	391	132	1.16e-02
4925	391	138	7.09e-03
4114	391	122	2.11e-03
1763	391	63	8.57e-03
2992	391	95	3.79e-03
12236	391	282	2.32e-02
1941	391	69	3.17e-03
1850	391	68	1.16e-03
1830	391	68	7.76e-04
1467	391	61	7.24e-05
1395	391	60	2.83e-05
1383	391	60	2.06e-05



4. Enrichment analysis of selected genes

g:Profiler

What GO or Reactome terms are enriched in your gene list?

- Tip: Use g:GOSt tool in g:Profiler for functional enrichment analysis

4. Enrichment analysis of selected genes

g:Profiler

g:GOST
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

Query

Upload bed file

1552807_a_at
210982_s_at
203915_at
208894_at
202833_s_at
201137_s_at
205242_at
209924_at
223502_s_at
204670_x_at
219505_at
228532_at
211429_s_at
229560_at
32128_at
212671_s_at

Run query

random

example

Options

Organism: Homo sapiens (Human)

☐ Ordered query ?
☐ Run as multiquery ?

Advanced options ^

☐ All results ?
☐ Measure underrepresentation ?
Statistical domain scope ?
Only annotated genes
Significance threshold ?
Benjamini-Hochberg FDR
User threshold ?
0.05
Numeric IDs treated as ?
AFFY_HUGENE_2_0_ST_V1

Data sources ^

select all

clear all

Show data versions

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

name.gmt

ENSG.gmt

Gene list

Universe

Statistical threshold

databases

4. Enrichment analysis of selected genes

g:Profiler

- When mapping ambiguous symbols, pay attention to
 - Pseudogenes, fusion genes, read-throughs
 - Apparently unrelated symbols, from e.g. matching aliases
 - Families of very similar proteins (histones)
 - Number of annotated GO terms (Ensembl alignment errors, etc)

200748_S_AT

- ENSG00000237264: ferritin heavy chain 1 pseudogene 11 [Source:HGNC Symbol;Acc:HGNC:3981] [Number of GO annotations: 0]
- ENSG00000226564: ferritin heavy chain 1 pseudogene 20 [Source:HGNC Symbol;Acc:HGNC:37639] [Number of GO annotations: 0]
- ENSG00000227376: ferritin heavy chain 1 pseudogene 16 [Source:HGNC Symbol;Acc:HGNC:3986] [Number of GO annotations: 0]
- ENSG00000167996: ferritin heavy chain 1 [Source:HGNC Symbol;Acc:HGNC:3976] [Number of GO annotations: 138]
- ENSG00000234975: ferritin heavy chain 1 pseudogene 2 [Source:HGNC Symbol;Acc:HGNC:3989] [Number of GO annotations: 0]

With the most terms

Clear

203882_AT

- ENSG00000213928: interferon regulatory factor 9 [Source:HGNC Symbol;Acc:HGNC:6131] [Number of GO annotations: 124]
- ENSG00000259529: novel transcript [Number of GO annotations: 38]

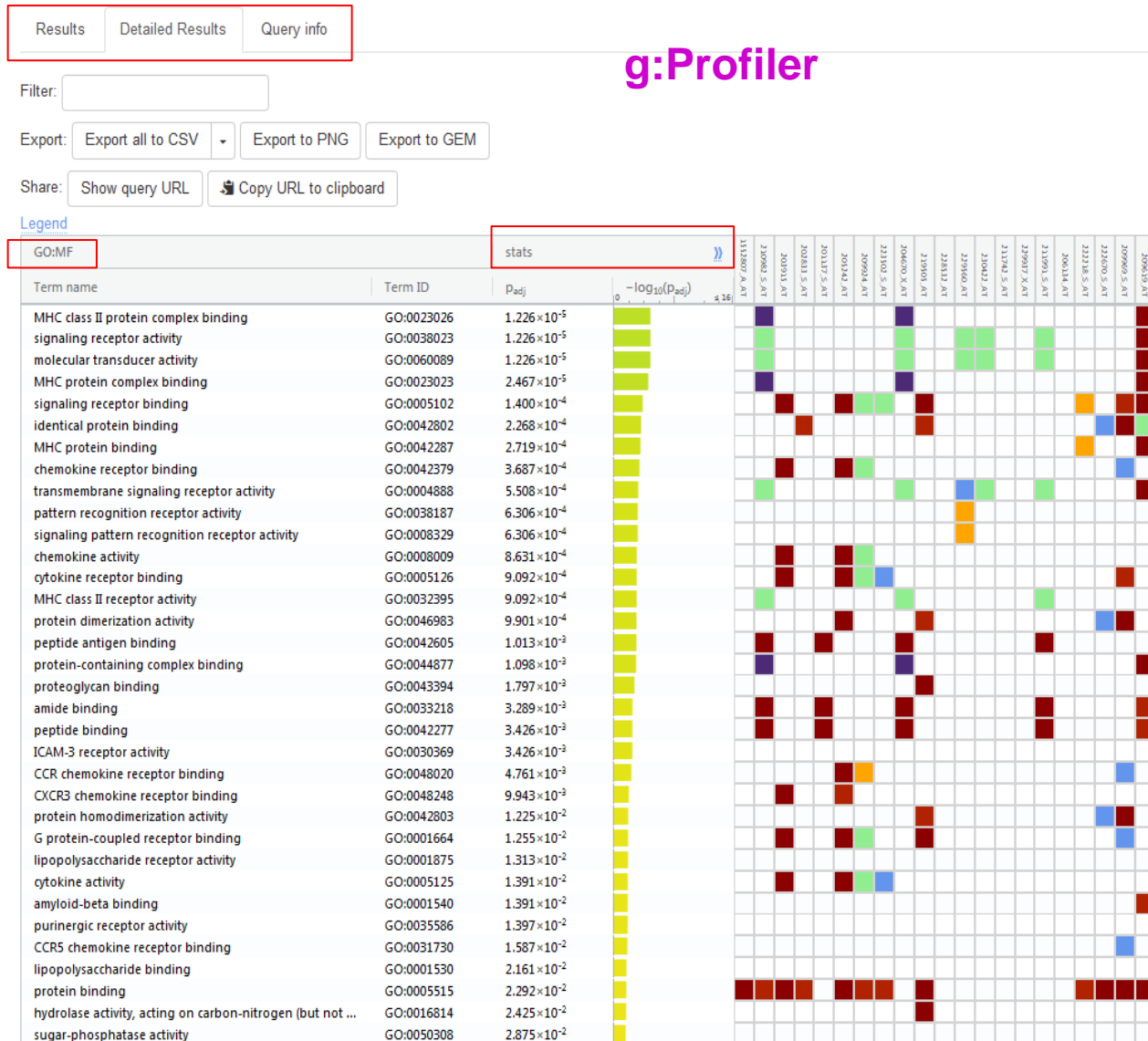
With the most terms

Clear

204006_S_AT

- ENSG00000162747: Fc fragment of IgG receptor IIb [Source:HGNC Symbol;Acc:HGNC:3620] [Number of GO annotations: 71]
- ENSG00000203747: Fc fragment of IgG receptor IIIa [Source:HGNC Symbol;Acc:HGNC:3619] [Number of GO annotations: 63]

4. Enrichment analysis of selected genes



4. Enrichment analysis of selected genes

GORilla

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

- ✓ Direct visualization of enriched terms in hierarchy
- ✓ Needs to specify the universe
- ✗ Only for GO
- ✗ Gene/Protein names as input

4. Enrichment analysis of selected genes

GOrilla

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

[Running example](#)

[Usage instructions](#)

[GOrilla News](#)

[Refer](#)

Step 1: [Choose organism](#)

Homo sapiens

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:

XXXX
MARCH1
HAVCR2
NCKAP1L
CD84
CCL18

← Gene list

Or upload a file: No s'ha seleccionat cap fitxer.

Background set:

LOC100506675
LINCO1191
DNMBP
PRKXP1
TMEM41B
BBIP1

← Universe (take it from file
topTab_GSE7586_annot.csv)

Or upload a file: No s'ha seleccionat cap fitxer.

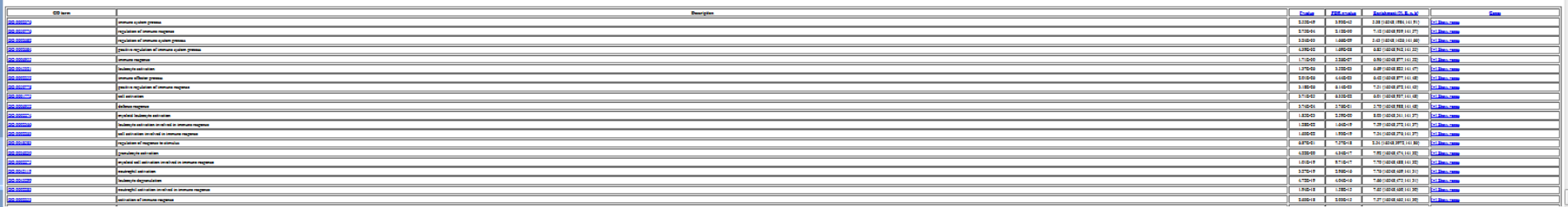
Step 4: [Choose an ontology](#)

☒ Process ☐ Function ☐ Component ☐ All

Advanced parameters

P-value threshold:

Analysis name: (optional)



6. Reactome Pathway Enrichment analysis of selected genes


Try with your dataset (list of genes): *selected_logFC_reactome.csv*

- What pathways are enriched below $FDR < 0.05$?
- Which has the major number of matching entities? What is its hierarchy?
- Explore “Diseases” section: are there similar pathways found in other infections?
- ...?

5. Introduction to Reactome and Pathway visualization

Analysis

Familiarize with Reactome using example data







USE REACTOME GRAPH DATABASE IN YOUR PROJECT

Why Reactome

Reactome is a free, open-source, curated and peer-reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genomics analysis, modeling systems biology and education. The current version of Reactome open release September 22, 2017.

If you use Reactome in a lab, we suggest citing our Chinese mirror site at: reactome.org/cn.

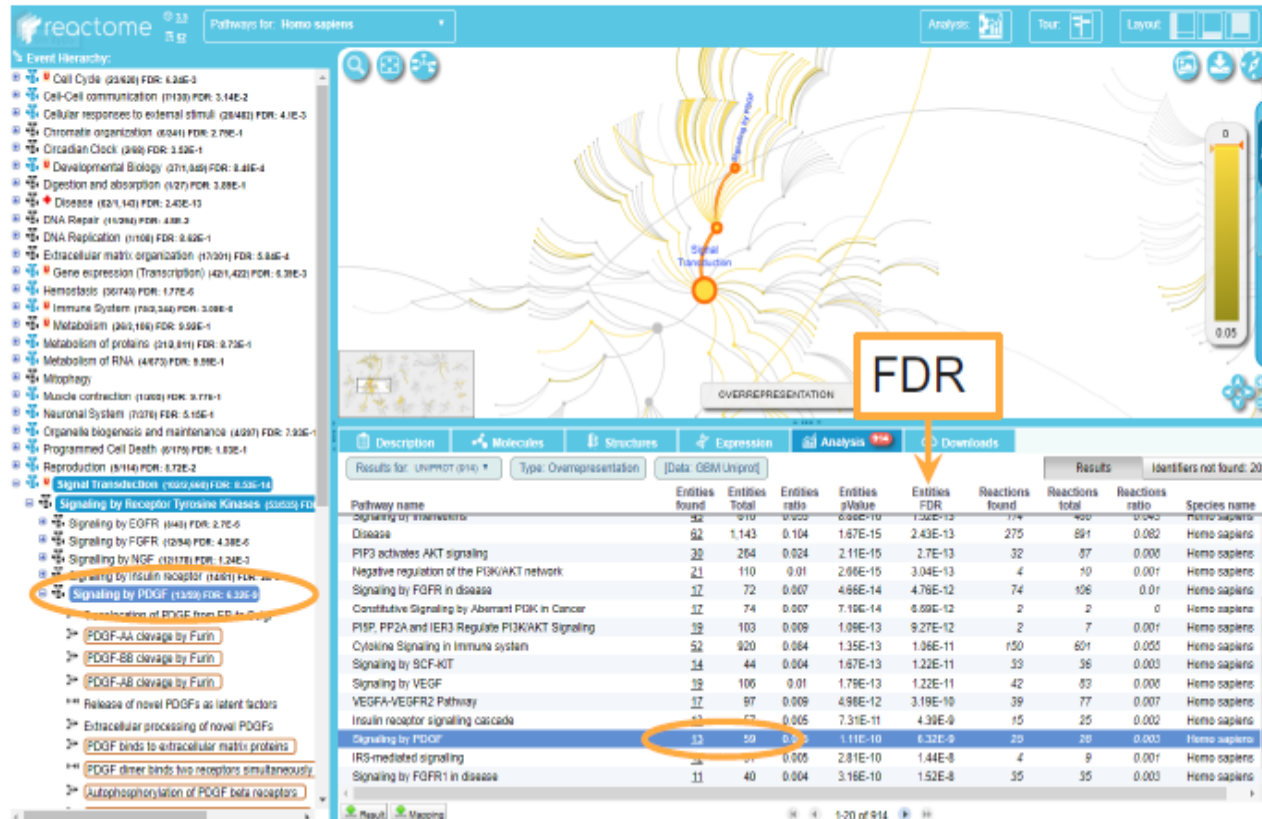
EMBL-EBI    

The development of Reactome is supported financially from the European Institute of Innovation and Research (EIR) and the European Molecular Biology Laboratory.

The screenshot shows the Reactome database interface. On the left is a sidebar with a list of biological processes, including 'Signal Transduction', 'Cell Cycle', 'Metabolism of PABA', and 'Muscle Contraction'. The main area displays a complex pathway diagram with various nodes and connecting lines. A red circle highlights the 'Analyze' button in the top right corner. Below the diagram, there is a section titled 'Description' which provides details about the selected pathway, including its name, location, and a brief description of its function.

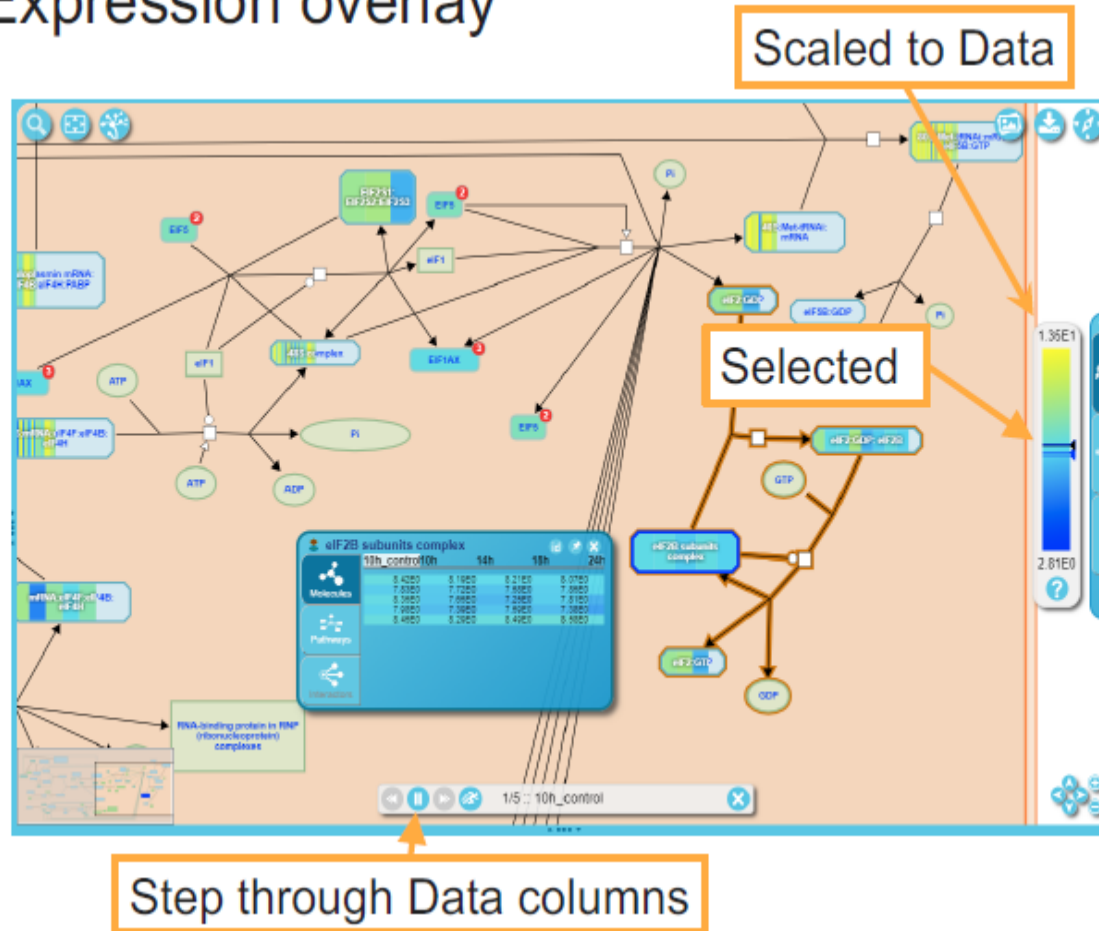
5. Introduction to Reactome and Pathway visualization

Analysis Result – Over-representation



5. Introduction to Reactome and Pathway visualization

Expression overlay



Visualization and downstream analysis

REVIGO
EnrichmentMaps
Pathview

[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 of pathways of interest

EnrichmentMaps

A workflow in R

clusterProfileR

Workflow gsea with clusterprofiler:

<https://learn.gencore.bio.nyu.edu/rna-seq-analysis/gene-set-enrichment-analysis/>

https://hbctraining.github.io/DGE_workshop_salmon/lessons/functional_analysis_2019.html

https://hbctraining.github.io/DGE_workshop/lessons/functional_analysis_other_methods.html

<http://biit.cs.ut.ee/gprofiler/page/r>