

# Pathway Analysis

Research Informatics Core

November 02, 2023

## Contents

<b>1 Morning</b>	<b>2</b>
1.1 Gene ID conversion using DAVID . . . . .	2
1.2 Gene ID conversion using UniProt . . . . .	8
1.3 Gene ID conversion using BioMart . . . . .	13
1.4 Using QuickGO . . . . .	19
1.5 Using MSigDB . . . . .	25
1.6 Browse KEGG . . . . .	35
1.7 Pathway Enrichment in DAVID . . . . .	45
1.8 <i>INSTRUCTOR DEMONSTRATION:</i> Pathway Enrichment in Ingenuity Pathway Analysis (IPA) . . . . .	51
1.9 Pathway Enrichment in R . . . . .	55
1.10 Pathway Enrichment in GSEA . . . . .	57
<b>2 Afternoon</b>	<b>62</b>
2.1 Barplot visualization . . . . .	62
2.2 Pathway comparison plots . . . . .	65
2.3 Heatmap of expression patterns in pathway . . . . .	71
2.4 Pathway analysis in variant calling data . . . . .	73
2.5 Using STRING . . . . .	78
2.6 Using STITCH . . . . .	90
2.7 Using miRNet . . . . .	100
2.8 <i>INSTRUCTOR DEMONSTRATION:</i> Network analysis using IPA . . . . .	105

# 1 Morning

## 1.1 Gene ID conversion using DAVID

### 1.1.1 Prepare input data

Download the Ensembl Gene list data we'll be using:

[https://wd.cri.uic.edu/pathway/GeneID\\_Conversion\\_list.txt](https://wd.cri.uic.edu/pathway/GeneID_Conversion_list.txt)

### 1.1.2 Navigate to DAVID

Navigate to DAVID <https://david.ncifcrf.gov>, and choose “Start Analysis” from the top menu.

The screenshot shows the DAVID Bioinformatics homepage. The top navigation bar includes links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Terms of Service, About DAVID, and About LHR. The main content area has two sections: 'Overview' and 'Hot Links'. The 'Hot Links' section contains several links: 'Post Doctoral Fellow position available in LHR!' (with a note about updates), 'DAVID Forum' (a forum for users to ask questions), 'FAQ' (Frequently Asked Questions), 'LHR Publications' (publications from the Laboratory of Human Retrovirology and Immunoinformatics), and 'DAVID Publications' (publications about DAVID). The 'Overview' section provides a brief introduction to DAVID's comprehensive set of functional annotation tools for understanding biological meaning behind large lists of genes.

### 1.1.3 Upload to DAVID

- Step 1: Click “Choose File” and browse to the gene list text file that downloaded
- Step 2: Choose “ENSEMBL\_GENE\_ID” from the Select Identifier drop-down
- Step 3: Select the “Gene List” radio button
- Step 4: Submit List to DAVID.

**Analysis Wizard**

Upload List Background

**Upload Gene List**

Demolist 1 Demolist 2  
Upload Help

**Step 1: Enter Gene List**

A: Paste a list

Box A

Clear

Or

B: Choose From a File

Choose File **GenelD\_Conversion\_list.txt**

Multi-List File ?

**Step 2: Select Identifier**

ENSEMBL\_GENE\_ID

1007\_s\_at  
1053\_at  
117\_at  
121\_at  
1255\_g\_at  
1294\_at  
1316\_at  
1320\_at  
1405\_i\_at  
1431\_at  
1438\_at  
1487\_at  
1494\_f\_at  
1598\_g\_at

**Step 3: List Type**

Gene List  
 Background

**Step 4: Submit List**

Tell us how you like the tool  
Contact us for questions

Step 1. Submit your gene list through left panel.

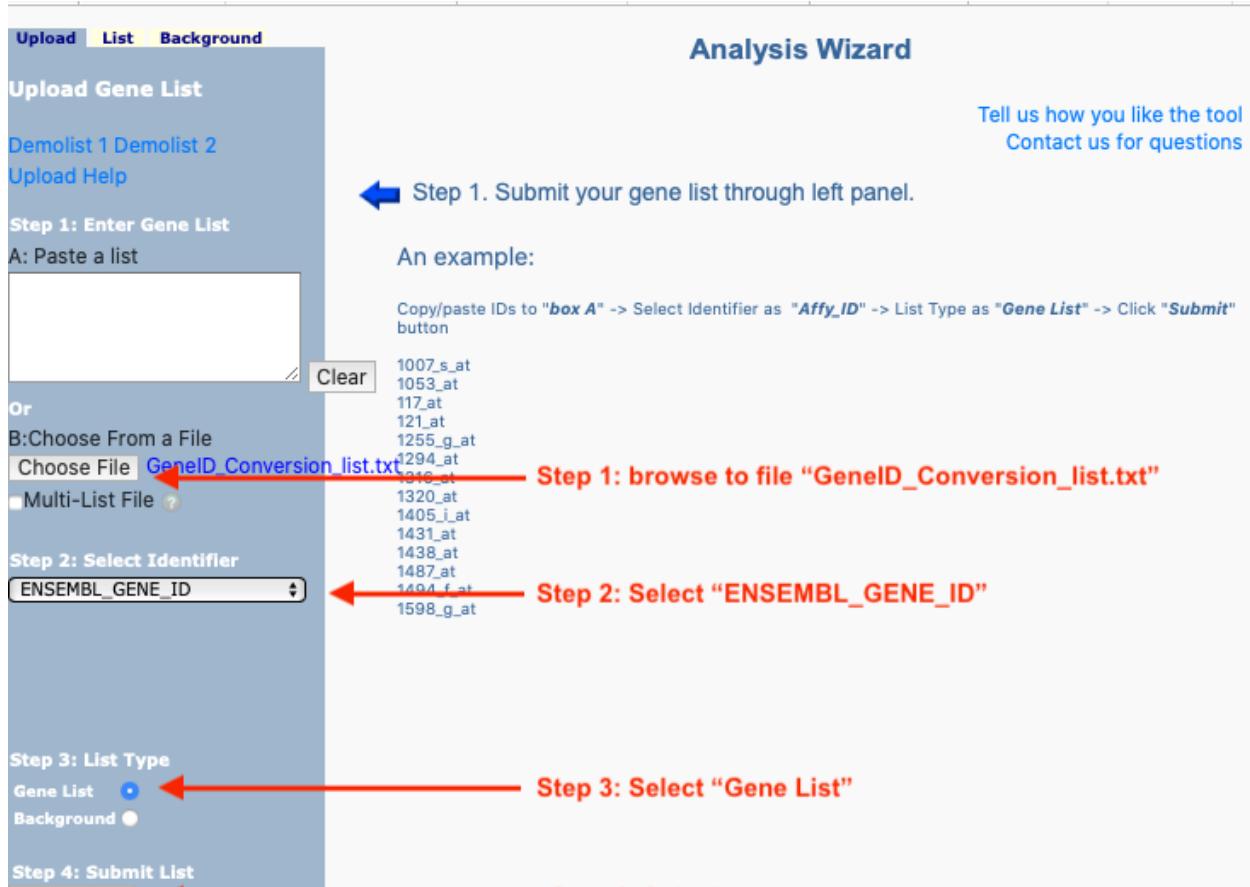
An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy\_ID" -> List Type as "Gene List" -> Click "Submit" button

Step 1: browse to file “GenelD\_Conversion\_list.txt”

Step 2: Select “ENSEMBL\_GENE\_ID”

Step 3: Select “Gene List”



#### 1.1.4 Submit to conversion tool

- Step 5: Choose “GeneID Conversion Tool”

**Analysis Wizard**

**Gene List Manager**

Select to limit annotations by one or more species [Help](#)

- Use All Species -  
Mus musculus(42)  
Unknown(4)

Select Species

List Manager Help  
GenelD\_Conversion\_list

Select List to:  
Use Rename  
Remove Combine  
Show Gene List

View Unmapped Ids

**Step 1. Successfully submitted gene list**  
Current Gene List: GenelD\_Conversion\_list  
Current Background: Mus musculus

**Step 2. Analyze above gene list with one of DAVID tools**

↓

Which DAVID tools to use?

Functional Annotation Tool

- Functional Annotation Clustering
- Functional Annotation Chart
- Functional Annotation Table

Gene Functional Classification Tool

Gene ID Conversion Tool **(circled)**

Gene Name Batch Viewer

Tell us how you like the tool  
Contact us for questions

The screenshot shows the Analysis Wizard interface. On the left, the Gene List Manager panel displays a list of species: 'Mus musculus(42)' and 'Unknown(4)'. Below this is a 'Select Species' button. Under 'List Manager', there's a link to 'GenelD\_Conversion\_list'. At the bottom of this panel are buttons for 'Use', 'Rename', 'Remove', 'Combine', and 'Show Gene List'. A 'View Unmapped Ids' link is also present. On the right, the main 'Analysis Wizard' section is titled 'Step 1. Successfully submitted gene list' with details about the current gene list and background. It then transitions to 'Step 2. Analyze above gene list with one of DAVID tools'. A downward arrow points to a list of DAVID tools: 'Functional Annotation Tool' (with sub-options for clustering, chart, and table), 'Gene Functional Classification Tool', 'Gene ID Conversion Tool' (which is circled in red), and 'Gene Name Batch Viewer'. At the top right of the main area, there are links to 'Tell us how you like the tool' and 'Contact us for questions'.

- **Step 5:** Choose “UNIPROT\_ID” from the drop down menu
- **Step 6:** Enter “Mus musculus” for the species
- **Step 7:** Click on “Submit to Conversion Tool” to submit the gene list for conversion\*

**Upload** **List** **Background**

### Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -  
Mus musculus(42)  
Unknown(4)

[Select Species](#)

[List Manager](#) [Help](#)  
[GeneID\\_Conversion\\_list](#)

**Select List to:**  
[Use](#) [Rename](#)  
[Remove](#) [Combine](#)  
[Show Gene List](#)

[View Unmapped Ids](#)

### Gene ID Conversion Tool

[Help and Tool Manual](#)

**Option 1:**  
Convert the gene list being selected in left panel to [UNIPROT\\_ID](#)

For species:

[Submit to Conversion Tool](#)

**Option 2:** [Go Back to Submission Form](#)

### 1.1.5 Results from David

- Review “Gene Accession Conversion Statistics”
  - Our `GeneID_Conversion_list.txt` had 46 genes in the list and 42 are converted
- Right-click on the “Download File” link on right hand corner above the results to save as a file. Save it as `david_id_conversion_results.tsv`.
  - Open in Excel and review

**Gene Accession Conversion Tool**

Output: for many genes there is one-to-many relationship

Download the results  Help

Conversion Summary			Submit Converted List to DAVID as a Gene List		Submit Converted List to DAVID as a Background	
ID Count	In DAVID DB	Conversion	From	To	Species	David Gene Name
42	Yes	Successful	ENSMUSG00000016918	A0A0G2JDD4_MOUSE	Mus musculus	sulfatase 1(Sulf1)
0	Yes	None	ENSMUSG00000016918	SULF1_MOUSE	Mus musculus	sulfatase 1(Sulf1)
0	No	None	ENSMUSG00000016918	A0A0B7WR86_MOUSE	Mus musculus	sulfatase 1(Sulf1)
0	Ambiguous	Pending	ENSMUSG00000016918	A0A0R4J2D2_MOUSE	Mus musculus	sulfatase 1(Sulf1)
<b>Total Unique User IDs: 42</b>			ENSMUSG00000016918	A0A0B7WS44_MOUSE	Mus musculus	sulfatase 1(Sulf1)
<b>Summary of Ambiguous Gene IDs</b>			ENSMUSG00000026121	A0A0A6YX48_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)
<b>All Possible Sources For Ambiguous IDs</b>			ENSMUSG00000026121	SEM4C_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)
<b>Ambiguous ID</b>			ENSMUSG00000026121	Q69ZB9_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)
			ENSMUSG00000026121	A0A0A6YXK9_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)
			ENSMUSG00000031790	Q6GX96_MOUSE	Mus musculus	matrix metallopeptidase 15(Mmp15)
			ENSMUSG00000031790	MMP15_MOUSE	Mus musculus	matrix metallopeptidase 15(Mmp15)
			ENSMUSG00000041859	Q3UI57_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)
			ENSMUSG00000041859	MCM3_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)
			ENSMUSG00000041859	Q3ULD6_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)
			ENSMUSG00000041859	Q3UZH2_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)

- Observations:
  - Open `david_id_conversion.tsv` in Excel
  - 329 UniProt IDs records are returned for 42 ENSEMBL ID
  - Many genes have one to many relationship

A	B	C	D	E	F	G	H	I	J	K	L	M
From	To	Species	Gene Name									
2 ENSMUSG00000016918	A0A0G2JDD4_MOUSE	Mus musculus	sulfatase 1(Sulf1)									
3 ENSMUSG00000016918	SULF1_MOUSE	Mus musculus	sulfatase 1(Sulf1)									
4 ENSMUSG00000016918	A0A087WR86_MOUSE	Mus musculus	sulfatase 1(Sulf1)									
5 ENSMUSG00000016918	A0A0R4J2D2_MOUSE	Mus musculus	sulfatase 1(Sulf1)									
6 ENSMUSG00000016918	A0A087WS44_MOUSE	Mus musculus	sulfatase 1(Sulf1)									
7 ENSMUSG00000026121	A0A0A6YX48_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)									
8 ENSMUSG00000026121	SEM4C_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)									
9 ENSMUSG00000026121	Q692B9_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)									
10 ENSMUSG00000026121	A0A0A6YXK9_MOUSE	Mus musculus	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4C(Sema4c)									
11 ENSMUSG00000031790	Q6GX96_MOUSE	Mus musculus	matrix metallopeptidase 15(Mmp15)									
12 ENSMUSG00000031790	MMP15_MOUSE	Mus musculus	matrix metallopeptidase 15(Mmp15)									
13 ENSMUSG00000041859	Q3UI57_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)									
14 ENSMUSG00000041859	MCM3_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)									
15 ENSMUSG00000041859	Q3ULD6_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)									
16 ENSMUSG00000041859	Q3UZH2_MOUSE	Mus musculus	minichromosome maintenance complex component 3(Mcm3)									
17 ENSMUSG00000042807	HECW2_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hebw2)									
18 ENSMUSG00000042807	A3KP87_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hebw2)									
19 ENSMUSG00000042807	Q8BQD5_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hebw2)									
20 ENSMUSG00000042807	Q8CB85_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hebw2)									
21 ENSMUSG00000026042	C05A2_MOUSE	Mus musculus	collagen, type V, alpha 2(Col5a2)									
22 ENSMUSG00000026042	Q8BNA3_MOUSE	Mus musculus	collagen, type V, alpha 2(Col5a2)									
23 ENSMUSG00000027330	Q3TZX9_MOUSE	Mus musculus	cell division cycle 25B(Cdc25b)									
24 ENSMUSG00000027330	Q9DBN8_MOUSE	Mus musculus	cell division cycle 25B(Cdc25b)									
25 ENSMUSG00000027330	MPIP2_MOUSE	Mus musculus	cell division cycle 25B(Cdc25b)									
26 ENSMUSG00000027330	Q3U535_MOUSE	Mus musculus	cell division cycle 25B(Cdc25b)									

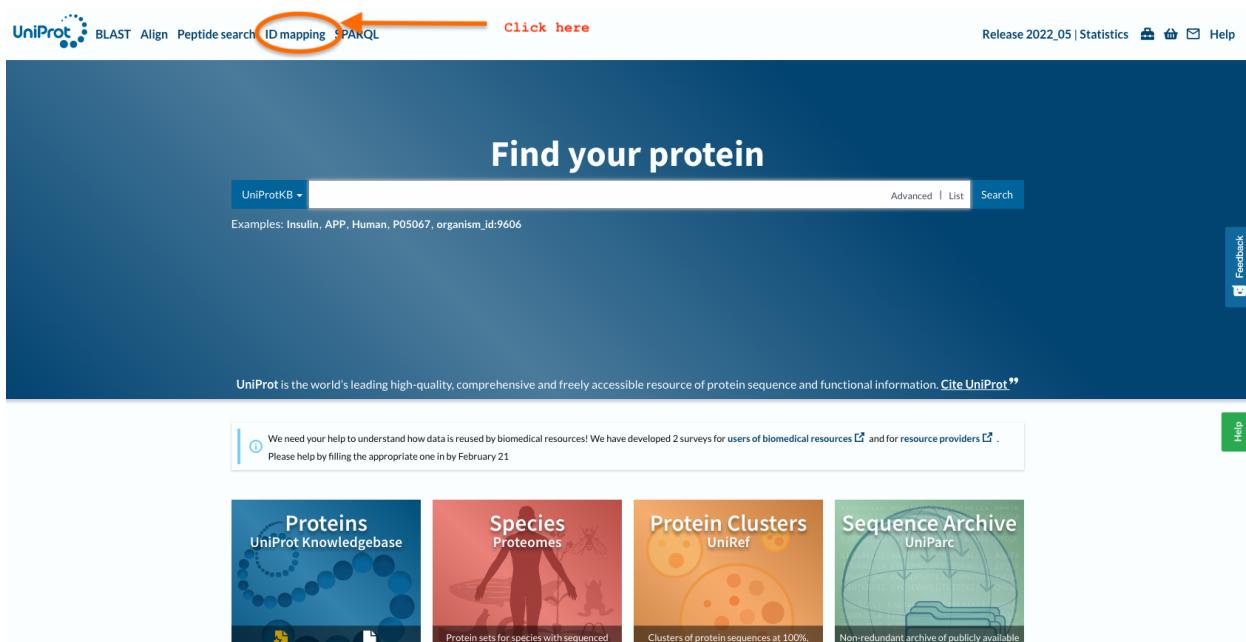
## 1.2 Gene ID conversion using UniProt

We're going to use the same Ensemble gene list as in the DAVID exercise:

[https://wd.cri.uic.edu/pathway/GeneID\\_Conversion\\_list.txt](https://wd.cri.uic.edu/pathway/GeneID_Conversion_list.txt)

### 1.2.1 Navigate to UniProt

Navigate to UniProt <https://www.uniprot.org>, and choose “ID mapping” from top menu bar



The screenshot shows the UniProt homepage with a dark blue header. In the top left, there's a logo with the word "UniProt". To its right are links for "BLAST", "Align", "Peptide search", "ID mapping" (which is circled in orange and has an arrow pointing to it), and "SPARQL". On the far right of the header, it says "Release 2022\_05 | Statistics" followed by icons for user profile, help, and feedback. Below the header, the main title "Find your protein" is centered. Underneath it is a search bar with dropdown menus for "UniProtKB" and "UniProtSprot", and buttons for "Advanced", "List", and "Search". A note below the search bar says "Examples: Insulin, APP, Human, P05067, organism\_id:9606". At the bottom of the page, there's a banner with text about data reuse surveys and a "Feedback" button. Below the banner are four colored boxes: blue for "Proteins UniProt Knowledgebase", red for "Species Proteomes", orange for "Protein Clusters UniRef", and green for "Sequence Archive UniParc". Each box contains a small icon and a descriptive subtitle.

### 1.2.2 Upload to UniProt

- **Step 1** Click “1. load from a text file”
  - Click “Choose File” and browse to the gene list text file that was downloaded
- **Step 2** Under “2. Select options”
  - Choose “ENSEMBL” on the Select Identifier drop-down “From”
  - Choose “UniProtKB” on the Select Identifier drop-down “To”
  - In “Name Your ID Mapping Job” put *ensemble\_to\_uniprot*
  - Click “Map 46 IDS”
  - To view mapping results click on *Completed* by your job name *ensemble\_to\_uniprot* under *Tool results*

The screenshot shows the UniProt ID mapping interface. At the top, there's a navigation bar with links for BLAST, Align, Peptide search, ID mapping, SPARQL, and UniProtKB. Below the navigation bar, the main title is "Retrieve/ID mapping". A text input field is labeled "Enter one or more IDs (100,000 max). You may also [load from a text file](#). Separate IDs by whitespace (space, tab, newline) or commas." Below the input field, there are two dropdown menus: "From database" set to "Ensembl" and "To database" set to "UniProtKB". A red circle highlights the "load from a text file" link. Further down, there's a field labeled "Name your ID Mapping job" containing "my job title", which is also highlighted with a red box. On the right side of the page, there are "Feedback" and "Help" buttons.

### 1.2.3 Results from UniProt

UniProt Knowledgebase has two databases

- **Swiss-Prot:** Proteins that are manually annotated and reviewed
- **TrEMBL:** Proteins that are automatically annotated and not reviewed

Observations:

- 41 out 46 Ensemble identifiers are successfully mapped to 103 UniProtKB IDs
  - For many Ensembl gene IDs we have multiple UniProt IDs
  - Out of 103 records, 39 records have Swiss-Prot IDs and 64 have TrEMBL IDs
- Click pencil icon by ‘Customize columns’ - to review other available annotations.

**ID mapping 103 results found for Ensembl → UniProtKB**

From	Entry	Entry Name	Protein Names	Gene Names	Organism	Length
ENSMUSG00000025903	P97823	LYPA1_MOUSE	Acyl-protein thioesterase 1[...]	Lypa1,Apt1,Pla1a	Mus musculus (Mouse)	230 AA
ENSMUSG00000025903	D3YUG4	D3YUG4_MOUSE	palmitoyl-protein hydrolase[...]	Lypa1	Mus musculus (Mouse)	142 AA
ENSMUSG00000025903	D3Z111	D3Z111_MOUSE	palmitoyl-protein hydrolase[...]	Lypa1	Mus musculus (Mouse)	196 AA
ENSMUSG00000025903	D3Z269	D3Z269_MOUSE	palmitoyl-protein hydrolase[...]	Lypa1	Mus musculus (Mouse)	142 AA
ENSMUSG00000025903	J3QP56	J3QP56_MOUSE	palmitoyl-protein hydrolase[...]	Lypa1	Mus musculus (Mouse)	224 AA
ENSMUSG00000025903	J3QQ63	J3QQ63_MOUSE	palmitoyl-protein hydrolase[...]	Lypa1	Mus musculus (Mouse)	92 AA
ENSMUSG00000025912	AOA087WPK9	AOA087WPK9_MOUSE	Myb-related protein A	Mybl1	Mus musculus (Mouse)	267 AA
ENSMUSG00000025912	AOA0R4J132	AOA0R4J132_MOUSE	Myb-related protein A	Mybl1	Mus musculus (Mouse)	751 AA
ENSMUSG00000025912	E9QLX9	E9QLX9_MOUSE	Myb-related protein A	Mybl1	Mus musculus (Mouse)	691 AA
ENSMUSG00000016918	Q8K007	SLF1_MOUSE	Extracellular sulfatase Sulf-1[...]	Sulf1,Kiaa1077	Mus musculus (Mouse)	870 AA
ENSMUSG00000016918	AOA087WR86	AOA087WR86_MOUSE	Extracellular sulfatase Sulf-1	Sulf1	Mus musculus (Mouse)	372 AA
ENSMUSG00000016918	AOA087WS44	AOA087WS44_MOUSE	Extracellular sulfatase Sulf-1	Sulf1	Mus musculus (Mouse)	37 AA
ENSMUSG00000016918	AOAG2JD4	AOAG2JD4_MOUSE	Extracellular sulfatase Sulf-1	Sulf1	Mus musculus (Mouse)	7 AA
ENSMUSG00000016918	AOA0R4J2D2	AOA0R4J2D2_MOUSE	Extracellular sulfatase Sulf-1	Sulf1	Mus musculus (Mouse)	862 AA

- Review other available annotations
  - Select “Gene ontology IDs” under “Gene Ontology (GO)”
    - Hit close button once done

**Customize columns**

Reviewed x Entry Name x Protein names x Gene Names x Organism x Length x Gene Ontology IDs x

Data 7 External links

Search for available columns

Names & Taxonomy	Gene Names	Organism	Length	Gene Ontology IDs
cyl-protein thioesterase 1 [..]	Lypia1, Apt1, Pla1a	Mus musculus (Mouse)	230 AA	GO:0005737 GO:0005829 GO:0005783 GO:0005739 GO:0031965 9 more IDs
palmitoyl-protein hydrolase [..]	Lypia1	Mus musculus (Mouse)	142 AA	GO:0016787
palmitoyl-protein hydrolase [..]	Lypia1	Mus musculus (Mouse)	196 AA	GO:0016787
palmitoyl-protein hydrolase [..]	Lypia1	Mus musculus (Mouse)	142 AA	GO:0016787

- Review results with other added annotations
- Download the results. Various formats are available e.g fasta, gff, tab, excel

Download

Download selected (0)

Download all (103)

**FASTA (canonical)**

**✓ FASTA (canonical & isoform)**

TSV  
Excel  
JSON  
XML  
RDF/XML  
Text  
GFF  
List

Advanced | List Search   

### l → UniProtKB

Customize columns  

Protein Names	Gene Names	Organism	Length	Gene Ontology IDs	Gene Ontology
cyl-protein dioesterase 1 <sup>[...]</sup>	Lypla1, Apt1, Pla1a	Mus musculus (Mouse)	230 AA	GO:0005737 GO:0005829 GO:0005783 GO:0005739 GO:0031965 9 more IDs	cytoplasm <sup>[?]</sup> cytosol <sup>[?]</sup> endoplasmic reticulum <sup>[?]</sup> mitochondrion <sup>[?]</sup> nuclear membrane <sup>[?]</sup> More terms
palmitoyl-protein lydrolase <sup>[...]</sup>	Lypla1	Mus musculus (Mouse)	142 AA	GO:0016787	hydrolase activity <sup>[?]</sup>
palmitoyl-protein lydrolase <sup>[...]</sup>	Lypla1	Mus musculus (Mouse)	196 AA	GO:0016787	hydrolase activity <sup>[?]</sup>
palmitoyl-protein	Lypla1	Mus	142 AA	GO:0016787	hydrolase activity <sup>[?]</sup>

## 1.3 Gene ID conversion using BioMart

We're going to use the same Ensembl gene list as in the DAVID exercise:

[https://wd.cri.uic.edu/pathway/GeneID\\_Conversion\\_list.txt](https://wd.cri.uic.edu/pathway/GeneID_Conversion_list.txt)

### 1.3.1 Navigate to BioMart

Navigate to Ensembl BioMart <https://useast.ensembl.org/info/data/biomart/index.html>, and choose BioMart from top menu bar.

The screenshot shows the Ensembl BioMart interface. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. The BioMart link is highlighted with a red arrow. Below the navigation bar, there's a search bar labeled "Search all species..." and a "Login/Register" button. On the left, there are two sidebar sections: "In this section" and "On this page". The "In this section" sidebar lists links for Biomart BioC R package, Biomart Perl API, Biomart RESTful access, Combining species datasets, and How to use BioMart. The "On this page" sidebar lists BioMart tutorials and FAQs, and BioMart R package, RESTful access. In the main content area, the URL is "Help & Documentation / Accessing Ensembl Data / BioMart". The title is "Extracting data with BioMart". A sub-section header "Tables of Ensembl data can be downloaded via the highly customisable BioMart data mining tool." is followed by a paragraph about the tool's ease of use. Below this, there's a table with columns for Ensembl Gene ID, Ensembl Transcript ID, HOGC ID, and HOGC symbol. The table contains 15 rows of data, corresponding to the 15 entries in the provided gene list. At the bottom of the table, there's a note: "Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results".

Ensembl Gene ID	Ensembl Transcript ID	HOGC ID	HOGC symbol
ENSG00000140908	ENST000000471846	HOGC-14668	MMEL1
ENSG00000140908	ENST00000076112	HOGC-14668	MMEL1
ENSG00000140908	ENST00000076112	HOGC-14668	MMEL1
ENSG00000140908	ENST000000504800	HOGC-14668	MMEL1
ENSG00000140908	ENST000000504800	HOGC-14668	MMEL1
ENSG00000140908	ENST000000504800	HOGC-14668	MMEL1
ENSG00000140908	ENST000000504800	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000491141	HOGC-14668	MMEL1
ENSG00000140908	ENST000000511099	HOGC-14668	MMEL1
ENSG00000140908	ENST000000511099	HOGC-14668	MMEL1
ENSG00000140908	ENST000000228750	ENST00000432429	

### 1.3.2 Choose Database(s)

- **Step 1** Choose “Ensembl Genes 110” from “CHOOSE DATABASE”
- **Step 2** Choose “Mouse genes (GRCm39)” from “CHOOSE DATASET”

Note: This will automatically load left panel with Filters and Attributes.

The screenshot shows the Ensembl BioMart search interface. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there are buttons for Login/Register and a search bar labeled "Search all species..." with a magnifying glass icon. Below the navigation bar, there are three tabs: "New", "Count", and "Results". The "Results" tab is selected. In the main content area, there are two dropdown menus under the heading "Dataset". The first dropdown is set to "Ensembl Genes 110" and the second is set to "Mouse genes (GRCm39)". Above these dropdowns, there are three buttons: "URL", "XML", and "Perl". Below the dropdowns, there are three sections: "Filters" (with "[None selected]"), "Attributes" (listing Gene stable ID, Gene stable ID version, Transcript stable ID, and Transcript stable ID version), and another "Dataset" section with "[None Selected]".

### 1.3.3 Upload to BioMart

- **Step 1** Click “Filters”. This will open “Filter” panel in the middle.
- **Step 2** Click “+” sign beside “GENE” on right hand side. We can filter query results based on any one of the seven options, e.g. Region, Gene.

The screenshot shows the Ensembl BioMart interface. At the top, there is a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, More, and a search bar. Below the navigation bar, there are buttons for New, Count, and Results, and links for URL, XML, Perl, and Help. On the left, there is a sidebar with sections for Dataset (Mouse genes (GRCm39)), Filters (None selected), and Attributes (Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version). The main area has a heading "Please restrict your query using criteria below" followed by a note "(If filter values are truncated in any lists, hover over the list item to see the full text)". A large panel on the right contains a list of filter categories with checkboxes: REGION, GENE, PHENOTYPE, GENE ONTOLOGY, MULTI SPECIES COMPARISONS, PROTEIN DOMAINS AND FAMILIES, and VARIANT.

- **Step 3**

- Clicking on “Gene” will expand and will show option to upload the file.
- Click “Choose File” and browse to the gene list text file that we downloaded **GeneID\_Conversion\_list.txt**. Note we are using Ensembl Gene IDs as our input so we do not need to change anything in select identifier dropdown.

The screenshot shows the Ensembl BioMart search interface. On the left, there's a sidebar with 'Dataset' set to 'Mouse genes (GRCm39)' and 'Filters' expanded, showing options like 'Gene stable ID(s)' and 'Attributes'. The main area has 'REGION:' and 'GENE:' sections. Under 'GENE:', the 'Input external references ID list [Max 500 advised]' checkbox is checked, and a file input field contains 'GeneID\_Conversion\_list.txt'. Below it, another file input field for 'AFFY MG U74A probe ID(s)' also contains 'GeneID\_Conversion\_list.txt'. Both fields have 'Only' selected in dropdown menus.

**Select Attributes:** Click “Attributes” in the left panel. This will open “Attributes” panel in the middle.

- **Step 1:** “Features” is selected by default

Please select columns to be included in the output and hit 'Results' when ready  
Missing non coding genes in your mart query output, please check the following [FAQ](#)

Features       Variant (Germline)  
 Structures       Sequences  
 Homologues (Max select 6 orthologues)

GENE:  
 EXTERNAL:  
 PROTEIN DOMAINS AND FAMILIES:

- **Step 2:** Click “+” beside “External”

- Select “Go term accession” under “GO”
- Select “UniProtKB Gene Name ID” under External References (max 3)
- Select “UniProtKB/Swiss-Prot ID” under External References (max 3)
- Select “UniProtKB/TrEMBL ID” under External References (max 3)

GO term accession       GO term name       GO term definition  
 GOSlim GOA Accession(s)       GOSlim GOA Description  
 BioGRID Interaction data, The General Repository for Interaction Datasets ID  
 CCDS ID  
 ChEMBL ID  
 EntrezGene transcript name ID  
 European Nucleotide Archive ID  
 Expression Atlas ID  
 GeneDB ID  
 HGNC ID  
 HGVS nomenclature  
 INSDC protein ID  
 MEROPS - the Peptidase Database ID  
 MGI description  
 MGI symbol  
 MGI ID  
 MGI transcript name ID  
 miRBase accession  
 Reactome gene ID  
 Reactome transcript ID  
 RefSeq mRNA ID  
 RefSeq mRNA predicted ID  
 RefSeq ncRNA ID  
 RefSeq ncRNA predicted ID  
 RefSeq peptide ID  
 RefSeq peptide predicted ID  
 RPRD1B  
 RFAM transcript name ID  
 RNAcentral ID  
 Transcript name ID  
 UCSC Stable ID  
 UniParc ID  
 UniProtKB Gene Name symbol  
 UniProtKB Gene Name ID

### 1.3.4 Results from BioMart

- Results
  - Make sure all attributes of interest are selected.
  - Click “Results” in top left panel menu.

### 1.3.5 Saving the results

- Top 10 results will be displayed if your query generates the results.
- To Save the results to file: “Export all results to”
  - Choose “compressed file (.gz)” from drop down
  - Choose “TSV” from the dropdown
  - Choose “Unique results only”
  - Click go
  - Look for `mart_export.txt.gz` file in the Downloads folder.

Note for larger query one can choose to get Email notification.

The screenshot shows the Ensembl BioMart interface. At the top, there are links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there is a search bar for "Search all species..." and a "Login/Register" link. Below the header, there are buttons for "New", "Count", and "Results". The main area has tabs for "Dataset" (selected), "Filters", and "Attributes". Under "Dataset", it says "Mouse genes (GRCm39)". Under "Filters", there is a field for "Gene stable ID(s) [e.g., ENSMUSG000000000001]: [ID-list specified]". Under "Attributes", there is a list of options: Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version, GO term accession, UniProtKB Gene Name ID, UniProtKB/Swiss-Prot ID, and UniProtKB/TrEMBL ID. The main content area shows a table of results with the following columns: Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version, GO term accession, UniProtKB Gene Name ID, UniProtKB/Swiss-Prot ID, and UniProtKB/TrEMBL ID. There are 10 rows of data, each corresponding to the gene ENSMUSG000000000386. The table includes export options at the top: "Export all results to" (File, TSV, Unique results only), "Email notification to" (input field), and "View" (rows as HTML, Unique results only).

Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version	GO term accession	UniProtKB Gene Name ID	UniProtKB/Swiss-Prot ID	UniProtKB/TrEMBL ID
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000247756	ENSMUST00000247756.1	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000247757	ENSMUST00000247757.1	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000155233	ENSMUST00000155233.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			
ENSMUSG000000000386	ENSMUSG000000000386.19	ENSMUST00000113768	ENSMUST00000113768.8	Q3UD61			

## 1.4 Using QuickGO

### 1.4.1 Navigate to QuickGO

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

### 1.4.2 Search using GO ID

- Type “GO:0006641” in the search box.
- Click on “GO:0006641” from the returned results

The screenshot shows the QuickGO search interface. At the top, the search bar contains "GO:0006641". Below the search bar, the title "Gene Ontology and GO Annotations" is displayed. On the left, there are navigation links: Help, Contact, API, and Basket. The main content area is titled "Search" and shows the results for the entered GO ID. The first section, "Terms", lists several GO terms with their counts: GO:0006641 (34,278 annotations), GO:0000811 (9,673 annotations), GO:0070966 (7,341 annotations), GO:0001591 (454 annotations), and GO:0042735 (47 annotations). A link "Show all 8 results" is also present. The second section, "Gene products", lists gene products with their annotation counts: A0A1L7NQ74 (1 annotation), USLRM3 (1 annotation), USLRM5 (1 annotation), BSKUM7 (6 annotations), and BSKUL2 (6 annotations). A link "Show all 283 results" is available. On the right side, there is a sidebar titled "Area of interest" which includes a "GO slim annotations" section. At the bottom right of the main content area, there is a "Screenshot" link.

- View GO ID Results
  - Explore results by clicking on left menu for tabs, e.g. “Synonyms”, “Ancestor Chart”, or “Child terms”

**Quick GO**

Help | Contact | API | Basket | Search

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

**GO:0006641** P JSON

## triglyceride metabolic process

**Biological Process**

**Definition (GO:0006641 GONUTS page)**

The chemical reactions and pathways involving triglyceride, any triester of glycerol. The three fatty acid residues may all be the same or differ in any permutation. Triglycerides are important components of plant oils, animal fats and animal plasma lipoproteins.

33,926 annotations

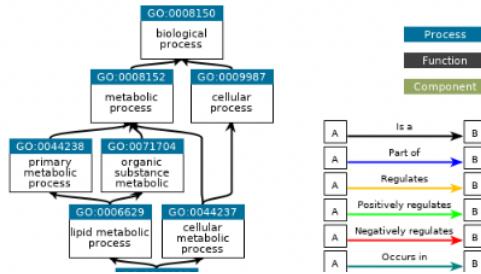
### Synonyms

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.

Synonym	Type
triacylglycerol metabolism	exact
triglyceride metabolism	exact
triacylglycerol metabolic process	exact

### Ancestor Chart

Ancestor chart for GO:0006641 | Chart options ▾



- Filter data
  - Click on “34,278 annotations” as shown in the previous image
  - Explore different tabs like “Taxon”, “Go terms”
  - Click “Taxon” to filter the results
    - \* Click “Homo sapiens”
    - \* Click “Apply”

**GO annotations**

**Click on Taxon to filter the results**

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

83333 Escherichia coli (strain K12)  
2759 Eukaryota  
**9606 Homo sapiens** **Click "Homo sapien"**  
3398 Magnoliophyta (flowering plants)  
40674 Mammalia  
10090 Mus musculus

...or enter taxon identifiers:  
9606,100090,10116,...

Visit UniProt Taxonomy to find identifiers for other taxonomic groups

Add **Click "Apply"** Reset Apply

Include descendants Exact match

			Evidence	Reference	With / From	Taxon	Assigned By	Annotation Extension
			ECO:0000256 IEA	GO_REF:0000002	InterPro:IPR027251	1445577 Colletotrichum fioriniae PJ7	InterPro	
			ECO:0000322 IEA	GO_REF:0000041	UniPathway:UPA00282	927661 Cryptosporangium arvum DSM 44712	UniProt	
			ECO:0000256 IEA	GO_REF:0000104	UniRule:UR000121046	568076 Metarhizium robertsii	UniProt	
		triglyceride biosynthetic process	ECO:0000256 IEA	GO_REF:0000104	UniRule:UR000121046	1432141 Rhizophagus irregularis (strain DAOM 197198w)	UniProt	
UniProtKB:A0A016SXY9	Acey_s0158.g3246	involved_in	GO:0019432 (P) ECO:0000256 triglyceride biosynthetic process IEA	GO_REF:0000002	InterPro:IPR027251	53326 Ancylostoma ceylanicum	InterPro	
UniProtKB:A0A016SYQ5	Acey_s0158.g3246	involved_in	GO:0019432 (P) ECO:0000256 triglyceride biosynthetic process IEA	GO_REF:0000002	InterPro:IPR027251	53326 Ancylostoma ceylanicum	InterPro	
UniProtKB:A0A016SYS7	Acey_s0158.g3246	involved_in	GO:0019432 (P) ECO:0000256 triglyceride biosynthetic process IEA	GO_REF:0000002	InterPro:IPR027251	53326 Ancylostoma ceylanicum	InterPro	

- Click “Export” to download the data

The screenshot shows the QuickGO annotations interface. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, About us, and EMBL-EBI logo. Below the navigation bar, there is a search bar and a menu bar with options: Taxon, Gene Products, GO terms, References, Aspect, Evidence, Extension, More, and Clear all. The main content area is titled "GO annotations". It features a table with the following columns: Gene Product, Evidence, Reference, With / From, Taxon, Assigned By, and Annotation Extension. The table contains several rows of data, each corresponding to a UniProtKB entry and its associated GO terms and evidence codes.

Gene Product	Evidence	Reference	With / From	Taxon	Assigned By	Annotation Extension				
UniProtKB:A0A024R	lycende biosynthetic process	ECO:0000265 IEA	GO_REF:0000107	9606 Homo sapiens	Ensembl					
UniProtKB:A0A04R	lycende metabolic process	ECO:0000265 IEA	GO_REF:0000107	9606 Homo sapiens	Ensembl					
UniProtKB:A0A1B1R	lycende catabolic process	ECO:0000265 IEA	GO_REF:0000107	9606 Homo sapiens	Ensembl					
UniProtKB:A0A1B1RVA9	LPL involved_in	GO:0019433 P IAA IFA	triglyceride catabolic process	ECO:0000265 IEA	GO_REF:0000104	UniRule:UR000141825	9606 Homo sapiens	UniProt		
UniProtKB:A0A384P5Q0	A0A384P5Q0 involved_in	GO:0006641 P IAA IFA	triglyceride metabolic process	ECO:0000265 IEA	GO_REF:0000107	UniProtKB:P24270 more...	9606 Homo sapiens	Ensembl		
UniProtKB:A0K2M5	A0K2M5 involved_in	GO:0019433 P IAA IFA	triglyceride catabolic process	ECO:0000265 IEA	GO_REF:0000104	UniRule:UR000141825	9606 Homo sapiens	UniProt		
UniProtKB:A0K8W7	A0K8W7 involved_in	GO:0019433 P IAA IFA	triglyceride catabolic process	ECO:0007322 IEA	GO_REF:0000041	UniPathwayUPA00256	9606 Homo sapiens	UniProt		
UniProtKB:B3KRV7	B3KRV7 involved_in	GO:0019433 P IAA IFA	triglyceride catabolic process	ECO:0000265 IEA	GO_REF:0000104	UniRule:UR000141825	9606 Homo sapiens	UniProt		
UniProtKB:K7ERI9	APOC1 involved_in	GO:0006641 P IAA IFA	triglyceride met...	Screenshot	ECO:0000265 IEA	GO_REF:0000107	UniProtKB:P34928 more...	9606 Homo sapiens	Ensembl	

- Click “Statistics” to get Summary

The screenshot shows a software interface with a header containing tabs: Summary, GO ID, Aspect, Evidence, Reference, Taxon, and Assigned By. The 'Summary' tab is selected. Below the tabs, there is a message: "Your current result set contains 173 annotations to 85 distinct gene products." A note below states: "Note: Figures displayed on this, and the other pages in this dialog relate to the annotation set that you have selected; if you perform any filtering or ID mapping operations, the statistics displayed will alter accordingly."

### 1.4.3 Search using keyword

- Navigate to QuickGO <https://www.ebi.ac.uk/QuickGO/>
- Type “p53” in the search box
- Explore the results

The screenshot shows the QuickGO search interface. At the top, there is a navigation bar with links for Help, Contact, API, and Basket. The main search bar contains the query "p53". Below the search bar, the results are displayed under two main sections: "Terms" and "Gene products".

**Terms**

- GO:0002039 (F) p53 binding (13,465 annotations)
- GO:0072331 (P) signal transduction by p53 class mediator (6,650 annotations)
- GO:1901796 (P) regulation of signal transduction by p53 class mediator (5,998 annotations)
- GO:0072332 (P) intrinsic apoptotic signaling pathway by p53 class mediator (3,242 annotations)
- GO:0043516 (P) regulation of DNA damage response, signal transduction by p53 class mediator (3,285 annotations)

Show all 33 results

**Gene products**

- A0A075BAE6 SAM domain-containing protein (12 annotations)
- A0A0B4K7P1 p53, isoform E (61 annotations)
- A0A5B9DV18 p53 DNA-binding domain-containing protein (9 annotations)
- Q35873 Cellular tumor antigen p53 (14 annotations)
- Q533U3 Cellular tumor antigen p53 (23 annotations)

Show all 11,809 results Screenshot

area of interest

0 annotations

EMBL-|

## 1.5 Using MSigDB

### 1.5.1 Navigate to MSigDB

Navigate to MSigDB <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>

**Molecular Signatures Database**

**Human Collections**

- H hallmark gene sets
- C1 positional gene sets
- C2 curated gene sets
- C3 regulatory target gene sets
- C4 computational gene sets
- C5 ontology gene sets
- C6 oncogenic signature gene sets
- C7 immunologic signature gene sets
- C8 cell type signature gene sets

**Overview**

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

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

**License Terms**

GSEA and MSigDB are available for use under [these license terms](#).

### 1.5.2 Explore different collections

For this exercise we will highlight several different collections.

- Hallmark and KEGG pathway collections given as examples below
- Also note:
  - C2: Reactome pathways
  - C3: miRNA and TF target genes
  - C5: Gene Ontology databases

The screenshot shows the MSigDB Human Gene Sets browser. At the top, there's a navigation bar with links for 'MSigDB Home', 'Human Collections' (selected), 'About', 'Browse', 'Search', 'Investigate', and 'Gene Families'. Below this is a secondary navigation bar for 'Mouse Collections' with links for 'About', 'Browse', 'Search', and 'Investigate'. The main content area has a header 'Browse Human Gene Sets' with the 'UC San Diego' and 'BROAD INSTITUTE' logos. It features a search bar with placeholder '(Enter full or partial name)' and a dropdown menu for 'By first letter' (A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z). There's also a section for 'By collection' which lists various gene set categories: H (Hallmark gene sets, 50 gene sets), C1 (positional gene sets, 300 gene sets), C2 (curated gene sets, 6495 gene sets), C3 (regulatory gene sets, 3713 gene sets), MIR (microRNA targets, 2598 gene sets), C4 (computational gene sets, 858 gene sets), C5 (miRNA and TF target genes, 1115 gene sets), and C6 (Gene Ontology databases, 491 gene sets). A 'Screenshot' button is located at the bottom right of the list.

- **Hallmark Collection:** Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. Currently it has total of 50 gene sets.

- Click “Browse Gene Sets” on the left panel
- Click “H” to expand the Hallmark collection
  - \* Scroll to the bottom of the page. It will list of all Gene sets for Hallmark Collection.
  - \* Click on any one of the gene set collection to view the details and to download the gene set.

**HSigDB Home**

**Human Collections**

- » About
- » Search
- » Investigate
- » Gene Families

**House Collections**

- » About
- » Browse
- » Search
- » Investigate

**Help**

**Browse Human Gene Sets**

UCSan Diego 

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

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

**By collection:** [about the MSigDB Human collections]

- » **H** (hallmark gene sets, 50 gene sets)
- » **C1** (positional gene sets, 300 gene sets)
- » by chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT
- » **C2** (curation gene sets, 6495 gene sets)
  - » **CGP** (chemical and genetic perturbations, 3405 gene sets)
  - » **CP** (canonical pathways, 309 gene sets)
    - » **CP:BIOCARTA** (BioCarta gene sets, 292 gene sets)
    - » **CP:KEGG** (KEGG gene sets, 186 gene sets)
    - » **CP:PID** (PID gene sets, 196 gene sets)
    - » **CP:REACTIONE** (Reactome gene sets, 1654 gene sets)
    - » **CP:WIKIPATHWAYS** (WikiPathways gene sets, 733 gene sets)
  - » **C3** (regulatory gene sets, 3713 gene sets)
    - » **MIR** (microRNA targets, 2598 gene sets)
      - » **MIR:MIR\_LEGACY** (legacy microRNA targets, 221 gene sets)
      - » **MIR:MIRDB** (mirDB microRNA targets, 2377 gene sets)
    - » **TFT** (all transcription factor targets, 1115 gene sets)
      - » **TFT:TFT** (CTTD transcription factor targets, 505 gene sets)
      - » **TFT:TFT\_LEGACY** (legacy transcription factor targets, 610 gene sets)
  - » **C4** (computational gene sets, 858 gene sets)
    - » **CGN** (cancer gene neighborhoods, 427 gene sets)
    - » **CM** (cancer modules, 431 gene sets)
  - » **C5** (ontology gene sets, 15937 gene sets)
    - » **GO** (Gene Ontology, 10532 gene sets)
      - » **GO:BP** (GO biological process, 7751 gene sets)
      - » **GO:CC** (GO cellular component, 1009 gene sets)
      - » **GO:MF** (GO molecular function, 1772 gene sets)
    - » **HPO** (Human Phenotype Ontology, 5405 gene sets)
  - » **C6** (oncogenic gene sets, 189 gene sets)
  - » **C7** (immunologic gene sets, 5219 gene sets)
    - » **IMMUNESIGDB** (ImmuneSigDB gene sets, 4872 gene sets)
    - » **VAX** (vaccine response gene sets, 347 gene sets)
  - » **C8** (cell type signature gene sets, 830 gene sets)

Click on a gene set name to view its gene set page. [\[Back to Top\]](#)

HALLMARK_ADOPOGENESIS	HALLMARK_GDN_NEUROPATHOLOGY	HALLMARK_NOTCH_SIGNALING
HALLMARK_ALLODRAFT_REJECTION	HALLMARK_GLUCOSESS	HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_ANDROGEN_RESPONSE	HALLMARK_HEDGEHOG_SIGNALING	HALLMARK_P33_PATHWAY
HALLMARK_ANGIOGENESIS	HALLMARK_HEME_METABOLISM	HALLMARK_PANCREAS_BETA_CELLS
HALLMARK_APICAL_JUNCTION	HALLMARK_HIF_SIGNALING	HALLMARK_PENK_SIGNALING
HALLMARK_APICAL_SURFACE	HALLMARK_I2_STATS_SIGNALING	HALLMARK_PENK_AKT_MTOR_SIGNALING
HALLMARK_APOTOPSIS	HALLMARK_JAK_STAT3_SIGNALING	HALLMARK_PROTEIN_SECRETION
HALLMARK_ARACHIDONATE_METABOLISM	HALLMARK_LIPID_AMINOACID_RESPONSE	HALLMARK.REACTIV_OXYGEN_SPECIES_PA
HALLMARK_CHOLESTEROL_HOMEOSTASIS	HALLMARK_INTERFERON_ALPHA_RESPONSE	HALLMARK_SPERMATOGENESIS
HALLMARK_COAGULATION	HALLMARK_INTERFERON_GAMMA_RESPONSE	HALLMARK_TGF_BETA_SIGNALING
HALLMARK_COMPLEMENT	HALLMARK_KRAS_SIGNALING_DN	

- Review the details for the given gene set collection.

**GSEA**  
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

UC San Diego  
BROAD INSTITUTE

## Gene Set: HALLMARKADIPOGENESIS

<b>Standard name</b>	HALLMARKADIPOGENESIS
<b>Systematic name</b>	M5905
<b>Brief description</b>	Genes up-regulated during adipocyte differentiation (adipogenesis).
<b>Full description or abstract</b>	
<b>Collection</b>	H: hallmark gene sets
<b>Source publication</b>	Pubmed 26771021 Authors: Liberzon A,Birger C,Thorvaldsdóttir H,Ghandi M,Mesirov JP,Tamayo P,
<b>Exact source</b>	
<b>Related gene sets</b>	(show 49 additional gene sets from the source publication) (show 47 gene sets from the same authors) (show 36 founder gene sets for this hallmark gene set)
<b>External links</b>	
<b>Filtered by similarity</b>	
<b>Organism</b>	Homo sapiens
<b>Contributed by</b>	Arthur Liberzon (MSigDB Team)
<b>Source platform</b>	HUMAN_GENE_SYMBOL
<b>Dataset references</b>	(show 4 hallmark refinement datasets) (show 3 hallmark validation datasets)
<b>Download gene set</b>	format: grp   text   gmt   gmx   xml
<b>Compute overlaps</b>	(show collections to investigate for overlap with this gene set)
<b>Compendia expression profiles</b>	GTEX compendium Human tissue compendium (Novartis) Global Cancer Map (Broad Institute) NCI-60 cell lines (National Cancer Institute)
<b>Advanced query</b>	Further investigate these 200 genes
<b>Gene families</b>	Categorize these 200 genes by gene family
<b>Show members</b>	(show 200 member mapped to 200 genes)
<b>Version history</b>	5.0: First introduced

See MSigDB license terms here. Please note that certain gene sets have special access terms.

- **C2 collection:** Curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts. It has various sub-collections for e.g. CP:KEGG represents canonical pathway's in KEGG's database.

- Click “Browse Gene Sets” on the left panel
- Click “C2” to expand the C2 collection
  - \* Click “CP:KEGG” to expand the collection
  - \* Scroll to the bottom of the page. It will list of all Gene sets for CP:KEGG collection.
  - \* Click on any one of the gene set collection to view the details

- Review the details for the given gene set collection.

**UC San Diego**

**BROAD INSTITUTE**

**GSEA**  
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

**CP:KEGG collection**

## Gene Set: KEGG\_ABC\_TRANSPORTERS

<b>Standard name</b>	KEGG_ABC_TRANSPORTERS
<b>Systematic name</b>	M11911
<b>Brief description</b>	ABC transporters
<b>Full description or abstract</b>	The ATP-binding cassette (ABC) transporters form one of the largest known protein families, and are widespread in bacteria, archaea, and eukaryotes. They couple ATP hydrolysis to active transport of a wide variety of substrates such as ions, sugars, lipids, sterols, peptides, proteins, and drugs. The structure of a prokaryotic ABC transporter usually consists of three components; typically two integral membrane proteins each having six transmembrane segments, two peripheral proteins that bind and hydrolyze ATP, and a periplasmic (or lipoprotein) substrate-binding protein. Many of the genes for the three components form operons as in fact observed in many bacterial and archaeal genomes. On the other hand, in a typical eukaryotic ABC transporter, the membrane spanning protein and the ATP-binding protein are fused, forming a multi-domain protein with the membrane-spanning domain (MSD) and the nucleotide-binding domain (NBD).
<b>Collection</b>	C2: curated gene sets CP: Canonical pathways CP:KEGG: KEGG gene sets
<b>Source publication</b>	
<b>Exact source</b>	hsa02010
<b>Related gene sets</b>	
<b>External links</b>	<a href="http://www.genome.jp/kegg/pathway/hsa/hsa02010.html">http://www.genome.jp/kegg/pathway/hsa/hsa02010.html</a>
<b>Organism</b>	Homo sapiens
<b>Contributed by</b>	KEGG (Kyoto Encyclopedia of Genes and Genomes)
<b>Source platform</b>	EntrezGeneIDs
<b>Dataset references</b>	<a href="#">Data download options</a>
<b>Download gene set</b>	format: grp   text   gmt   gmix   xml (show collections to Investigate for overlap with this gene set)
<b>Compute overlaps ?</b>	
<b>Compendia expression profiles ?</b>	GTEX compendium Human tissue compendium (Novartis) Global Cancer Map (Broad Institute) NCI-60 cell lines (National Cancer Institute)
<b>Advanced query</b>	Further investigate these 44 genes
<b>Gene families ?</b>	Categorize these 44 genes by gene family
<b>Show members</b>	(show 44 members mapped to 44 genes)
<b>Version history</b>	

### 1.5.3 Search gene sets in the collection browsing

- Click “Browse” in left menu bar. Search term “lung” in Gene set name and see what collections come back. Scroll all the way at bottom to review the collections results.
- Click “BLANCO\_MELO\_COVID19\_SARS\_COV\_2\_POS\_PATIENT\_LUNG\_TISSUE\_DN” from the returned results for review

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

[Back to Top](#)

BLANCO_MELO_COVID19_SARS_COV_2_POS_	GSE36078_UNTREATED_VS_ADS_T425A_HEX	STEARMAN_LUNG_CANCER_EARLY_VS_LATE_UP
PATIENT_LUNG_TISSUE_DN	ON_INF_MOUSE_LUNG_DC_DN	SWEET_LUNG_CANCER_KRAS_DN
BLANCO_MELO_COVID19_SARS_COV_2_POS_	GSE36078_UNTREATED_VS_ADS_T425A_HEX	SWEET_LUNG_CANCER_KRAS_UP
PATIENT_LUNG_TISSUE_UP	ON_INF_MOUSE_LUNG_DC_UP	TOMIDA_LUNG_CANCER_POOR_SURVIVAL
CHEN_LUNG_CANCER_SURVIVAL	GSE36078_WT_VS_IL1R_KO_LUNG_DC_AFTE	TRAVAGLINI_LUNG_ADVENTITIAL_FIBROBL
DESCARTES_FETAL_LUNG_BRONCHIOLAR_AN	R_ADS_INF_DN	AST_CELL
D_ALVEOLAR_EPITHELIAL_CELLS	GSE36078_WT_VS_IL1R_KO_LUNG_DC_AFTE	TRAVAGLINI_LUNG_AIRWAY_SMOOTH_MUSCL
DESCARTES_FETAL_LUNG_CILIATED_EPITH	R_ADS_INF_UP	E_CELL
ELIAL_CELLS	GSE36078_WT_VS_IL1R_KO_LUNG_DC_AFTE	TRAVAGLINI_LUNG_ALVEOLAR_EPITHELIAL
DESCARTES_FETAL_LUNG_CSH1_CSH2_POSI	R_ADS_T425A_HEXON_INF_DN	TYPE_1_CELL
TIVE_CELLS	GSE36078_WT_VS_IL1R_KO_LUNG_DC_AFTE	TRAVAGLINI_LUNG_ALVEOLAR_EPITHELIAL
DESCARTES_FETAL_LUNG_LYMPHATIC_ENDO	R_ADS_T425A_HEXON_INF_UP	TYPE_2_CELL
THELIAL_CELLS	GSE36078_WT_VS_IL1R_KO_LUNG_DC_DN	TRAVAGLINI_LUNG_ALVEOLAR_FIBROBLAST
DESCARTES_FETAL_LUNG LYMPHOID_CELLS	GSE36078_WT_VS_IL1R_KO_LUNG_DC_UP	_CELL
DESCARTES_FETAL_LUNG_MEGAKARYOCYTES	GSE36392_EOSINOPHIL_VS_MAC_IL25_TRE	TRAVAGLINI_LUNG_ARTERY_CELL
DESCARTES_FETAL_LUNG_MESOTHELIAL_CE	ATED_LUNG_DN	TRAVAGLINI_LUNG_B_CELL
LLS	GSE36392_EOSINOPHIL_VS_MAC_IL25_TRE	TRAVAGLINI_LUNG_BASAL_CELL
DESCARTES_FETAL_LUNG_MYELOID_CELLS	ATED_LUNG_UP	TRAVAGLINI_LUNG_BASOPHIL_MAST_1_CEL
DESCARTES_FETAL_LUNG_NEUROENDOCRINE	GSE36392_EOSINOPHIL_VS_NEUTROPHIL_I	L
_CELLS	L25_TREATED_LUNG_DN	TRAVAGLINI_LUNG_BASOPHIL_MAST_2_CEL
DESCARTES_FETAL_LUNG_SQUAMOUS_EPITH	GSE36392_EOSINOPHIL_VS_NEUTROPHIL_I	L
ELIAL_CELLS	L25_TREATED_LUNG_UP	TRAVAGLINI_LUNG_BRONCHIAL_VESSEL_1
DESCARTES_FETAL_LUNG_STROMAL_CELLS	GSE36392_MAC_VS_NEUTROPHIL_IL25_TRE	CELL
DESCARTES_FETAL_LUNG_VASCULAR_ENDOT	ATED_LUNG_DN	TRAVAGLINI_LUNG_BRONCHIAL_VESSEL_2
HELIALL_CELLS	GSE36392_MAC_VS_NEUTROPHIL_IL25_TRE	CELL
DESCARTES_FETAL_LUNG_VISCERAL_NEURO	ATED_LUNG_UP	TRAVAGLINI_LUNG_CAPILLARY_AEROCYTE
NS	GSE36392_TYPE_2_MYELOID_VS_EOSINOPH	CELL
DING_LUNG_CANCER_BY_MUTATION_RATE	IL_IL25_TREATED_LUNG_DN	TRAVAGLINI_LUNG_CAPILLARY_CELL
DING_LUNG_CANCER_EXPRESSION_BY_COPY	GSE36392_TYPE_2_MYELOID_VS_EOSINOPH	TRAVAGLINI_LUNG_CAPILLARY_INTERMEDI
_NUMBER	IL_IL25_TREATED_LUNG_UP	ATE_1_CELL
DING_LUNG_CANCER_MUTATED_FREQUENTLY	GSE36392_TYPE_2_MYELOID_VS_MAC_IL25	TRAVAGLINI_LUNG_CAPILLARY_INTERMEDI
DING_LUNG_CANCER_MUTATED_RECURRENTL	_TREATED_LUNG_DN	ATE_2_CELL
Y	GSE36392_TYPE_2_MYELOID_VS_MAC_IL25	TRAVAGLINI_LUNG_CD4_MEMORY_EFFECTOR
DING_LUNG_CANCER_MUTATED_SIGNIFICAN	_TREATED_LUNG_UP	_T_CELL
TLY	GSE36392_TYPE_2_MYELOID_VS_NEUTROPH	TRAVAGLINI_LUNG_CD4_NAIVE_T_CELL
FALVELLA_SMOKERS_WITH_LUNG_CANCER	IL_IL25_TREATED_LUNG_DN	TRAVAGLINI_LUNG_CD8_MEMORY_EFFECTOR
GOBP_BUD_ELONGATION_INVOLVED_IN_LUN	GSE36392_TYPE_2_MYELOID_VS_NEUTROPH	T CELL
G_BRANCHING	IL IL25 TREATED LUNG UP	

- Review results for “BLANCO\_MELO\_COVID19\_SARS\_COV\_2\_POS\_PATIENT\_LUNG\_TISSUE\_DN”
- Download gene set of interest

## Human Gene Set:

### BLANCO\_MELO\_COVID19\_SARS\_COV\_2\_POS\_PATIENT\_LUNG\_TISSUE\_DN

<b>Standard name</b>	BLANCO_MELO_COVID19_SARS_COV_2_POS_PATIENT_LUNG_TISSUE_DN
<b>Systematic name</b>	M34029
<b>Brief description</b>	Genes strongly down-regulated ( $\log_2(\text{FC}) < -3.58$ , $\text{padj} < 0.05$ ) in post mortem lung tissue from COVID-19 patients vs uninfected biopsy.
<b>Full description or abstract</b>	Transcriptional profiling of post-mortem lung samples from COVID-19-positive patients (N=2) compared with biopsied healthy lung tissue from uninfected individuals.
<b>Collection</b>	C2: Curated CGP: Chemical and Genetic Perturbations
<b>Source publication</b>	<a href="#">Pubmed 32416070</a> Authors: Blanco-Melo D,Nilsson-Payant BE,Liu WC,Uhl S,Hoagland D,Moller R,Jordan TX,Oishi K,Panis M,Sachs D,Wang TT,Schwartz RE,Lim JK,Albrecht RA,tenOever BR
<b>Exact source</b>	<a href="#">Table S4. Differential Gene Expression Analysis of COVID-19 Patients, Related to Figure 4. <math>\log_2(\text{FC}) &lt; -3.58</math>, <math>\text{padj} &lt; 0.05</math></a>
<b>Related gene sets</b>	(show 29 additional gene sets from the source publication) (show 22 gene sets from the same authors)
<b>External links</b>	
<b>Filtered by similarity</b>	
<b>Organism</b>	Homo sapiens
<b>Contributed by</b>	Anthony Castanza (MSigDB Team)
<b>Source platform</b>	HUMAN_GENE_SYMBOL
<b>Dataset references</b>	(show 1 datasets)
<b>Download gene set</b>	format: grp   gmt   xml   json   TSV metadata
<b>Compute overlaps</b>	(show collections to investigate for overlap with this gene set)
<b>Compendia expression profiles</b>	GTEX compendium Human tissue compendium (Novartis) Global Cancer Map (Broad Institute) NCI-60 cell lines (National Cancer Institute)
<b>Advanced query</b>	Further investigate these 175 genes
<b>Gene families</b>	Categorize these 175 genes by gene family
<b>Show members</b>	(show 175 members mapped to 175 genes)
<b>Version history</b>	

See [MSigDB license terms](#) here. Please note that certain gene sets have special access terms.

#### 1.5.4 Search gene sets in all of MSigDB by keyword

- Click “Search” in the left menu
  - Requires to register on the website to search.
- Type “Apoptosis” in *Keywords* bar
- Leave all other values as default

The screenshot shows the GSEA Gene Set Enrichment Analysis website. At the top, there is a navigation bar with links for GSEA Home, Downloads, Molecular Signatures Database, Documentation, Contact, and Team. On the far left, there is a sidebar with two main sections: Human Collections (About, Browse, Search, Investigate, Gene Families) and Mouse Collections (About, Browse, **Search**, Investigate, Help). The main content area is titled "Search Mouse Gene Sets". It includes a search bar with the placeholder "To search by full or partial gene set name, or to browse an alphabetical list, see the [Browse Gene Sets page](#).", a "Keywords" input field containing "Apoptosis" (with a note: "(supports boolean operators AND and OR, and wildcard searches with \*)"), a "search" button, and two filter panels. The "Search Filters" panel has two tabs: "collection" (listing M1: positional gene sets, M2: curated gene sets, M3: regulatory gene sets, M4: GTRD transcription factor targets, M5: ontology gene sets) and "source organism contributor" (listing all sources, Homo sapiens, Mus musculus). A note at the bottom says "control-click to select multiple lines". Logos for UC San Diego and the Broad Institute are visible in the top right corner.

## Review results

[MSigDB Home](#)

**Human Collections**

- ▶ About
- ▶ Browse
- ▶ Search
- ▶ Investigate
- ▶ Gene Families

**Mouse Collections**

- ▶ About
- ▶ Browse
- ▶ **Search**
- ▶ Investigate

Help

---

**Search Mouse Gene Sets**

UC San Diego 

To search by full or partial gene set name, or to browse an alphabetical list, see the [Browse Gene Sets page](#).

Search by keyword, collection, source organism, or contributor: [?](#)

**Keywords:**   
(supports boolean operators AND and OR, and wildcard searches with \*)

**Search Filters:**

collection	source organism	contributor
all collections	Homo sapiens	all contributors
M1: orthology-mapped hallmark gene sets	Mus musculus	Belgian Nuclear Research Centre
M2: curated gene sets		BioCarta
M3: positional gene sets		Broad Institute
-CGP: chemical and genetic perturbations		Dana-Farber Cancer Institute
-CP: canonical pathways		Gene Ontology Consortium
M4: regulatory gene sets		Michigan State University
-GTRD: GTRD transcription factor targets		MSigDB Team
-MIRDB: miRDB microRNA targets		Pierre and Marie Curie University
M5: ontology gene sets		Reacome

control-click to select multiple lines

found 115 gene sets

click on rows to select gene sets, click a gene set name to view the gene set page

select all 115   **0** gene sets selected  

<< < **1** 2 3 4 5 6 7 8 9 10 11 12 > >> 10

name	# genes	description	collections	source organism	contributor
BIOCARTA_ACH_PATHWAY	14	Role of nicotinic acetylcholine receptors in the regulation of <a href="#">apoptosis</a>	M2 CP	Mus musculus	BioCarta
BIOCARTA_AKT_PATHWAY	21	AKT Signaling Pathway	M2 CP	Mus musculus	BioCarta
BIOCARTA_ASBCEL_PATHWAY	10	Antigen Dependent B Cell Activation	M2 CP	Mus musculus	BioCarta
BIOCARTA_ATM_PATHWAY	10	ATM Signaling Pathway	M2 CP	Mus musculus	BioCarta

## 1.6 Browse KEGG

1. Open a web browser to <https://www.genome.jp/kegg>
2. Click on **KEGG PATHWAY** in the *Data-oriented entry points*.
3. In search (keywords) box enter “VEGF” and click **Go**.

4. On *Search Results* page, click **map04370** to view the pathway entry for the *VEGF signaling pathway*.

### 1.6.1 Pathway entry view

The KEGG Pathway entry page has the following details for the pathway **map04370**.



### PATHWAY: map04370

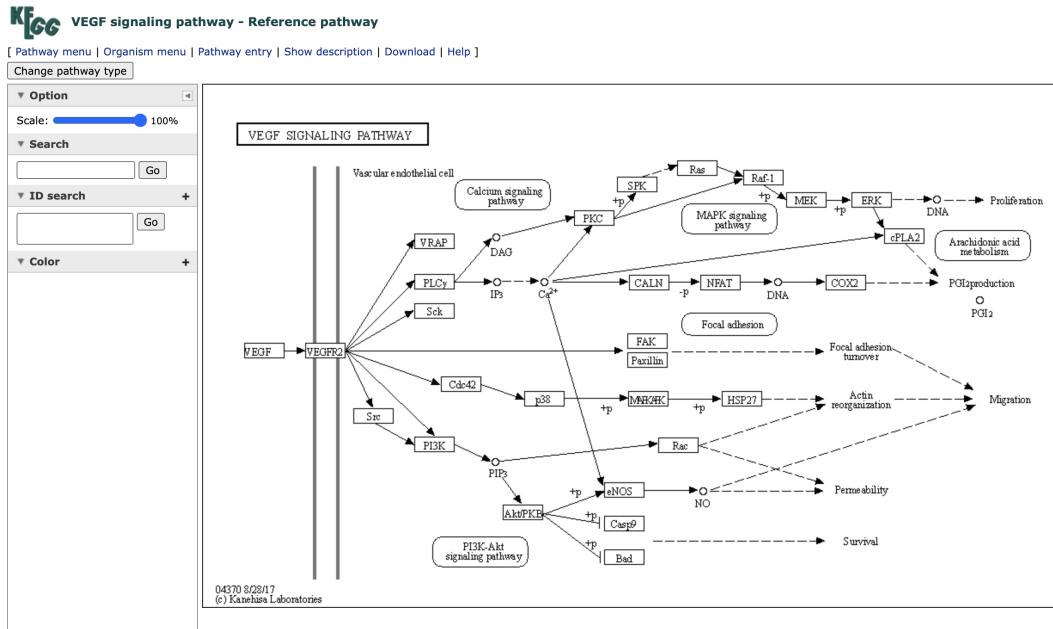
[Help](#)

<b>Entry</b>	map04370	Pathway
<b>Name</b>	VEGF signaling pathway	
<b>Description</b>	There is now much evidence that VEGFR-2 is the major mediator of VEGF-driven responses in endothelial cells and it is considered to be a crucial signal transducer in both physiologic and pathologic angiogenesis. The binding of VEGF to VEGFR-2 leads to a cascade of different signaling pathways, resulting in the up-regulation of genes involved in mediating the proliferation and migration of endothelial cells and promoting their survival and vascular permeability. For example, the binding of VEGF to VEGFR-2 leads to dimerization of the receptor, followed by intracellular activation of the PLCgamma;PKC-Raf kinase-MEK-mitogen-activated protein kinase (MAPK) pathway and subsequent initiation of DNA synthesis and cell growth, whereas activation of the phosphatidylinositol 3'-kinase (PI3K)-Akt pathway leads to increased endothelial-cell survival. Activation of PI3K, FAK, and p38 MAPK is implicated in cell migration signaling.	
<b>Class</b>	Environmental Information Processing; Signal transduction <a href="#">BRITE hierarchy</a>	
<b>Pathway map</b>	<a href="#">map04370 VEGF signaling pathway</a>	
	<a href="#">Ortholog table</a>	
<b>Other DBs</b>	GO: 0048010	
<b>Reference</b>	PMID: <a href="#">13678960</a>	
<b>Authors</b>	Cross MJ, Dixelius J, Matsumoto T, Claesson-Welsh L.	
<b>Title</b>	VEGF-receptor signal transduction.	
<b>Journal</b>	Trends Biochem Sci 28:488-94 (2003) DOI: <a href="#">10.1016/S0968-0004(03)00193-2</a>	

- **Entry** - The Entry ID and type (Pathway).
- **Name** - Name of the pathway,
- **Description** - Short description/definition of the Entry.
- **Class** - BRITE categories for the pathway
- **Pathway map** - Map of the pathway
- **Other DBs** - Links to other databases for this entry.
  - *GO* - Link to associated gene ontology term.
- **References** - Important references.

## 1.6.2 Pathway View

- Click on the picture of the pathway in the **Pathway view** section.



The KEGG Pathway view, gives a graphical representation of the pathway. In the pathway map, you can observe the following.

- Proteins/Enzymes are depicted as rectangles with their EC number.
- Compounds are depicted as small circles with their name adjacent.
- Reactions are denoted as arrows from substrate to product with associated enzyme components above/below the line.
- Upstream or downstream pathways are depicted as rectangles with rounded corners.

In this view, you can click on any of the protein/enzymes, compounds or associated pathways to view the details for that entry.

Above the pathway map are the following links

- Pathway menu** - Opens the pathway menu (list of all pathways) for this pathway.
- Organism menu** - Opens the Organism menu (list of all organisms) for this pathway.
- Pathway entry** - Shows the entry (details) page for this pathway.
- Show description** - Show the description for the pathway.
- Download** - Download the pathway information in png format

Side Panel

- Color** - Opens dialog box that allows a user to provide custom colors for entries on the pathway.

Note the **Change pathway type** button. This opens the organism menu (list of organisms/genomes) listing those organisms that have at least some portion of this pathway.

### 1.6.3 Orthology Details Page

1. Click on the **VEGF** protein rectangle in the left side of the pathway map. This will open the *Orthology Details* page.

<b>Entry</b>	K05448	KO
<b>Symbol</b>	VEGFA	
<b>Name</b>	vascular endothelial growth factor A	
<b>Pathway</b>	map01521 EGFR tyrosine kinase inhibitor resistance map04010 MAPK signaling pathway map04014 Ras signaling pathway map04015 Rap1 signaling pathway map04020 Calcium signaling pathway map04066 HIF-1 signaling pathway map04151 PI3K-Akt signaling pathway map04370 VEGF signaling pathway map04510 Focal adhesion map04926 Relaxin signaling pathway map04933 AGE-RAGE signaling pathway in diabetic complications map05163 Human cytomegalovirus infection map05165 Human papillomavirus infection map05167 Kaposi sarcoma-associated herpesvirus infection map05200 Pathways in cancer map05205 Proteoglycans in cancer map05206 MicroRNAs in cancer map05207 Chemical carcinogenesis - receptor activation map05208 Chemical carcinogenesis - reactive oxygen species map05211 Renal cell carcinoma map05212 Pancreatic cancer map05219 Bladder cancer map05323 Rheumatoid arthritis map05418 Fluid shear stress and atherosclerosis	
<b>Disease</b>	H01456 Diabetic nephropathy H01457 Diabetic retinopathy H01459 Diabetic neuropathy H01529 Avascular necrosis of femoral head H01709 Glucocorticoid-induced osteonecrosis H02559 Microvascular complications of diabetes	
<b>Brite</b>	KEGG Orthology (KO) [BR: <a href="#">ko00001</a> ] 09130 Environmental Information Processing 09132 Signal transduction 04010 MAPK signaling pathway K05448 VEGFA; vascular endothelial growth factor A 04014 Ras signaling pathway K05448 VEGFA; vascular endothelial growth factor A 04015 Rap1 signaling pathway K05448 VEGFA; vascular endothelial growth factor A 04370 VEGF signaling pathway K05448 VEGFA; vascular endothelial growth factor A 04066 HIF-1 signaling pathway K05448 VEGFA; vascular endothelial growth factor A 04020 Calcium signaling pathway	

- Entry - The Entry ID and type (KO for KEGG Orthology).
- Name - Name of the entry.

- **Pathway** - List of KEGG Pathway(s) of which this entry is a member.
- **Disease** - List of KEGG Disease of which this entry is a member.
- **BRITE** - Functional hierarchies for this entry.
- **Other DBs** - Links to other databases for this entry.
  - *GO* - Link to associated gene ontology term.
- **References** - Important references.

#### 1.6.4 Organism specific pathway view

1. Go back to browser window with the Pathway view page for **map04370**
2. Click on the **Change pathway type** button on top left.

### ▼ Reference

```
map Reference pathway
ko Reference pathway (KO only)
```

### ▼ Organism specific

#### ▼ Animals

##### ▼ Mammals

hsa	Homo sapiens (human)	28/28
ptr	Pan troglodytes (chimpanzee)	28/28
pps	Pan paniscus (bonobo)	28/28
ggo	Gorilla gorilla gorilla (western lowland gorilla)	28/28
pon	Pongo abelii (Sumatran orangutan)	28/28
nle	Nomascus leucogenys (northern white-cheeked gibbon)	27/28
hmh	Hylobates moloch (silvery gibbon)	28/28
mcc	Macaca mulatta (rhesus monkey)	28/28
mcf	Macaca fascicularis (crab-eating macaque)	28/28
mthb	Macaca thibetana thibetana (Pere David's macaque)	28/28
mni	Macaca nemestrina (pig-tailed macaque)	28/28
csab	Chlorocebus sabaeus (green monkey)	28/28
caty	Cercocebus atys (sooty mangabey)	28/28
panu	Papio anubis (olive baboon)	28/28
tge	Theropithecus gelada (gelada)	28/28
mlou	Mandrillus leucophaeus (drill)	27/28

This list displays the organism specific versions of the pathway that are available. The number to the right of the organism name denotes the fraction of genes present in that organism.

3. Click on **hsa** to open the Human specific version of the pathway. In this view of the pathway the molecules that are present in the selected organism are highlighted in green.



## VEGF signaling pathway - Homo sapiens (human)

[ Pathway menu | Organism menu | Pathway entry | Show description | Download | Help ]

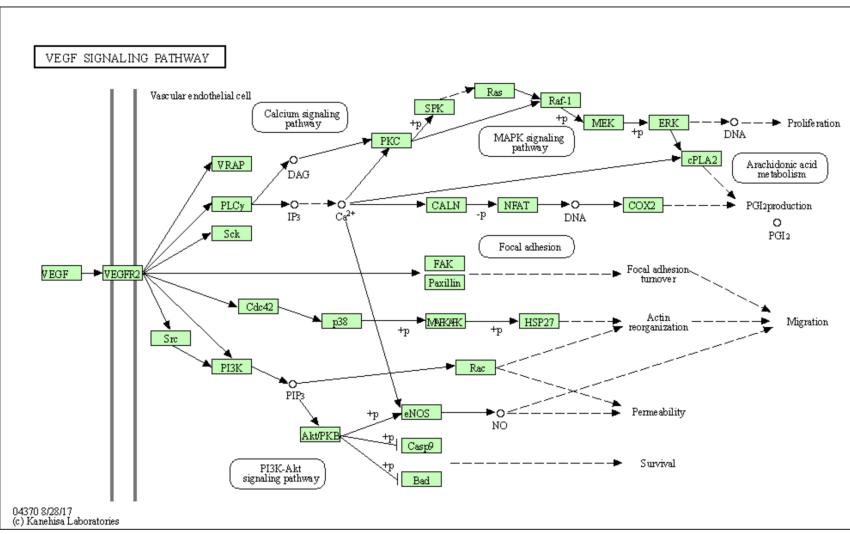
Change pathway type

▼ Option  
Scale:  100%

▼ Search  Go

▼ ID search  Go

▼ Color +



## 1.6.5 Gene Details Page

- Click on the **VEGF** protein rectangle in the left side of the organism specific pathway map. The KEGG Gene Details page has the following details for the Gene **hsa:7422**.

**Kegg** Homo sapiens (human): 7422 [Help](#)

Entry	7422	CDS	T01001
Symbol	VEGFA, MVCD1, VEGF, VPF		
Name	(RefSeq) vascular endothelial growth factor A		
KO	K05448 vascular endothelial growth factor A		
Organism	hsa Homo sapiens (human)		
Pathway	<a href="#">hsa01521 EGFR tyrosine kinase inhibitor resistance</a> <a href="#">hsa04010 MAPK signaling pathway</a> <a href="#">hsa04014 Ras signaling pathway</a> <a href="#">hsa04015 Rapi signaling pathway</a> <a href="#">hsa04016 Calcium signaling pathway</a> <a href="#">hsa04016 PI3K-Akt signaling pathway</a> <a href="#">hsa04151 PI3K-Akt signaling pathway</a> <a href="#">hsa04170 VEGF signaling pathway</a> <a href="#">hsa04510 Focal adhesion</a> <a href="#">hsa04926 Relxin signaling pathway</a> <a href="#">hsa04933 AGE-RAGE signaling pathway in diabetic complications</a> <a href="#">hsa05163 Human cytomegalovirus infection</a> <a href="#">hsa05164 Human papillomavirus infection</a> <a href="#">hsa05167 Kaposi's sarcoma-associated herpesvirus infection</a> <a href="#">hsa05200 Pathways in cancer</a> <a href="#">hsa05205 Proteoglycans in cancer</a> <a href="#">hsa05206 MicroRNAs in cancer</a> <a href="#">hsa05207 Chemical carcinogenesis - receptor activation</a> <a href="#">hsa05208 Chemical carcinogenesis - reactive oxygen species</a> <a href="#">hsa05210 Renal cell carcinoma</a> <a href="#">hsa05212 Bladder cancer</a> <a href="#">hsa05219 Bladder cancer</a> <a href="#">hsa05323 Rheumatoid arthritis</a> <a href="#">hsa05418 Fluid shear stress and atherosclerosis</a>		
Network	<a href="#">nt06114 PI3K signaling (viruses)</a> <a href="#">nt06126 Chemokine signaling (viruses)</a> <a href="#">nt06130 Cytokine-cytokine receptor interaction (viruses)</a> <a href="#">nt06131 Human T-lymphotoxin-associated herpesvirus (KSHV)</a> <a href="#">nt06214 PI3K signaling</a> <a href="#">nt06219 JAK-STAT signaling</a> <a href="#">nt06224 CXCR signaling</a> <a href="#">nt06226 GPCR signaling</a> <a href="#">nt06227 Nucleotide receptor signaling</a> <a href="#">nt06262 Pancreatic cancer</a> <a href="#">nt06263 Renal cell carcinoma</a> <a href="#">nt06306 Cushing syndrome</a>		
Element	<a href="#">N00079 HIF-1 signaling pathway</a> <a href="#">N00080 Loss of function to HIF-1 signaling pathway</a> <a href="#">N00081 Mutation-inactivation of HIF-1 to HIF-1 signaling pathway</a> <a href="#">N00095 ERBB2 overexpression to Egr-Jak-STAT signaling pathway</a>		

- Entry** - The Entry ID, type (CDS for coding sequence), and ID of source genome.
- Symbol** - Name of the entry
- Name KO** - Short description/definition of the Entry. Include link to associated KEGG Orthology (KO) entry.
- Organism** - Source genome/organism.
- Pathway** - List of *Organism specific* KEGG Pathway(s) of which this entry is a member.
- Network** - List of *perturbed molecular* networks
- Disease** - List of *Disease*
- Drug Target** - List of *Drug Target*
- BRITE** - Functional hierarchies for this entry.
- SSDB** - Links to sequence similarity database.
  - Ortholog* - Orthologous genes found in other organisms.
  - Paralog* - Sequence similar genes in the same genome.
  - Gene cluster* - Details of the local gene cluster.
  - GFIT* - Sequence similarity tables.
- Motif** - List of Pfam protein motifs.
- Other DBs** - Links to other databases for this entry.
  - NCBI-ProteinID* - Link to this entry in NCBI protein database.
  - Uniprot* - Link to this entry in Uniprot database.
- Structure** - Link to PDB database.
- LinkDB** - Link to all databases.
- Position** - Location of the gene in the genome.
- AA seq** - Translated sequence of the gene.
- NT seq** - Nucleotide sequence of the gene.

## 1.7 Pathway Enrichment in DAVID

### 1.7.1 Prepare input data

Download the RNA-seq data we'll be using:

[https://wd.cri.uic.edu/pathway/RNAseq\\_example.xlsx](https://wd.cri.uic.edu/pathway/RNAseq_example.xlsx)

Next, we want to generate a list of differentially expressed genes:

- Open the file in Excel
- Go to the *example\_diff* tab
- In the drop-down, over column F (*Infection/Control : QValue*), filter for values less than 0.05

The screenshot shows an Excel spreadsheet titled "RNAseq\_example.xlsx" with the "example\_diff" tab selected. The data consists of 281 rows of gene expression values. The columns are labeled "# Gene ID", "Gene name", and four "Infection/Control" metrics: "logC", "logPM", "pValue", and "Control". A filter dialog is open on the right side of the screen, titled "Infection/Control : QValue". It has two main sections: "Sort" and "Filter". Under "Sort", "A" (Ascending) is selected for the first column and "Z" (Descending) is selected for the second column. Under "Filter", "Less Than" is selected for the condition, and "0.05" is entered into the value field. The "And" radio button is selected. At the bottom of the filter dialog is a "Clear Filter" button.

# Gene ID	Gene name	Infection/Control : logC	Infection/Control : logPM	Infection/Control : pValue	Infection/Control : Control	
49	ENSMUSG00000067879	-1.20	8.01	4.18E-04	6.19E-03	
52	ENSMUSG00000025915	0.88	4.07	4.54E-04	6.65E-03	
59	ENSMUSG00000056763	-0.78	5.88	5.01E-04	7.21E-03	
102	ENSMUSG00000100398	2.36	-0.46	4.60E-03	0.05	
121	ENSMUSG00000041859	Mcm3	1.85	3.31	1.54E-12	6.20E-11
126	ENSMUSG00000041779	Tram2	1.46	1.78	3.06E-03	0.03
137	ENSMUSG00000041670	Rims1	-0.69	7.56	1.96E-04	3.20E-03
157	ENSMUSG00000101249	Gm29216	-8.28	3.91	5.45E-04	7.72E-03
198	ENSMUSG00000026126	Ptpn18	1.02	4.22	2.18E-04	3.52E-03
220	ENSMUSG00000047180	Neurl3	4.17	2.15	9.48E-20	5.88E-18
221	ENSMUSG00000037447	Arid5a	2.42	4.54	1.13E-20	7.32E-19
230	ENSMUSG00000026121	Sema4c	0.78	4.91	1.06E-03	0.01
238	ENSMUSG00000026117	Zap70	1.03	1.42	4.77E-03	0.05
250	ENSMUSG00000060771	Tsga10	-0.96	3.74	2.95E-04	4.58E-03
252	ENSMUSG00000026088	Mitd1	1.90	4.93	3.79E-09	1.17E-07
270	ENSMUSG00000048234	Rnf149	1.51	2.93	7.78E-08	2.11E-06
277	ENSMUSG00000026073	Il1r2	5.10	2.47	9.30E-30	1.01E-27
278	ENSMUSG00000026072	Il1r1	1.30	2.58	2.50E-04	3.97E-03
279	ENSMUSG00000070942	Il1rl2	2.12	5.94E-03	1.80E-03	0.02
281	ENSMUSG00000026070	Il18r1	3.64	-0.55	1.62E-04	2.69E-03

Save these genes to a text file:

- Select all of the genes in column A (excluding the title line)
  - We're using Ensembl IDs for the most precise gene identification
- Open a new Excel file and paste them
- Save As tab-delimited text to **RNAseq\_example.txt**
  - You can download this file also from:

[https://wd.cri.uic.edu/pathway/RNAseq\\_example.txt](https://wd.cri.uic.edu/pathway/RNAseq_example.txt)

### 1.7.2 Upload to DAVID

Navigate to DAVID <https://david.ncifcrf.gov>, and follow the same steps as before to upload the new data set. First choose “Start Analysis” from the top menu.

On the next page:

- Click “Browse...” and browse to the gene list text file that downloaded
- Choose “ENSEMBL\_GENE\_ID” from the Select Identifier drop-down
- Select the “Gene List” radio button
- Submit List to DAVID.

After uploading, click the “Functional Annotation Tool” link.

*Note that DAVID has auto-detected the species from the ENSEMBL IDs. If we had uploaded gene symbols, we would need to select the correct species here.*

The screenshot shows the DAVID Bioinformatics Database Analysis Wizard interface. The top navigation bar includes links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, About DAVID, and About LHRI. The main content area is titled "Analysis Wizard" and "DAVID Bioinformatics Resources, NIAID/NIH". On the left, a sidebar titled "Gene List Manager" shows a list of species: "Use All Species - Mus musculus(2534)", "Unknown(105)", and a "Select Species" button. Below this is a "List Manager" section with "RNASeq\_example" selected. Under "Select List to:", there are buttons for "Use", "Rename", "Remove", "Combine", and "Show Gene List". A blue arrow points down to the "Step 2. Analyze above gene list with one of DAVID tools" section. This section lists three expandable options: "Functional Annotation Tool" (with sub-options: Functional Annotation Clustering, Functional Annotation Chart, Functional Annotation Table), "Gene Functional Classification Tool", "Gene ID Conversion Tool", and "Gene Name Batch Viewer". A "Tell us how you like the tool" and "Contact us for questions" link is located in the top right corner.

We can choose which pathways to run enrichment against. DAVID has these organized into several categories. For today, we will enrich against Gene Ontology Biological Processes, and KEGG pathways (i.e., two databases of biological pathways).

- Uncheck the “Check Defaults” button
- Open the **Gene\_Ontology** section and select the GOTERM\_BP\_FAT option
  - The different numbered options (GOTERM\_BP\_1, GOTERM\_BP\_2, etc.) allow you to select different levels of the GO hierarchy
  - ALL is all terms
  - DIRECT omits parent terms
  - FAT is almost all terms, but the very broad ones are filtered out. **We recommend this option**, as it gives more specificity with a larger set of terms, but removes the very broad ones that would be uninformative.
  - Click on the *CHART* buttons next to each to get a quick view of how many pathways are included in the analysis, and what their enrichment statistics are.
- Open the **Pathways** section and select KEGG\_PATHWAYS

Other databases here may be useful for other contexts, so it can be valuable to explore them. For example, for the other GO categories:

- GO MF has types of molecular functions, such as cytokine binding or kinase activity, which can be useful if you’re trying to characterize classes of molecules.
- GO CC has localization of molecules in the cell. This can be useful if, for example, you were looking for cell surface markers in a set of differentially expressed genes from a single-cell experiment.

There are also pathways from Reactome available.

### 1.7.3 Results from DAVID

Select “Functional Annotation Chart” for the pathway enrichment statistics per term. Right-click on the *Download File* link to save as a file.

- NOTE: by default DAVID only downloads terms with at least 2 molecules and a nominal p-value < 0.1. If you’re running a pathway comparison, as we will in the afternoon, you may want to relax these thresholds for the purposes of the comparison. You can do this in the *Rerun Using Options* drop-down on this page.

The screenshot shows the DAVID Bioinformatics Functional Annotation Chart interface. At the top, it displays the U.S. Department of Health & Human Services and National Institutes of Health logos. Below the header, the DAVID Bioinformatics logo is shown. The main title is "Functional Annotation Chart". Underneath, it shows the "Current Gene List: RNAseq\_example", "Current Background: Mus musculus", and "2533 DAVID IDs". There are buttons for "Options", "Rerun Using Options", "Create Sublist", and "3282 chart records". On the right, there is a "Help and Manual" link and a "Download File" button with a file icon. Below these are navigation icons for page 1, 2, 3, 4, and so on. The central part of the screen is a table with the following columns: Sublist, Category, Term, RT, Genes, Count, %, P-Value, and Benjamini. The table lists various biological processes, each with a checkbox in the Sublist column and a blue bar chart representing the number of genes. The rows are color-coded in yellow or orange, likely indicating significant pathways.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_FAT	defense response	RT	614	24.2	3.0E-131	3.1E-127	
	GOTERM_BP_FAT	regulation of immune system process	RT	565	22.3	2.3E-121	1.2E-117	
	GOTERM_BP_FAT	immune response	RT	605	23.9	1.8E-111	6.4E-108	
	GOTERM_BP_FAT	positive regulation of response to stimulus	RT	672	26.5	5.4E-98	1.2E-94	
	GOTERM_BP_FAT	cytokine production	RT	338	13.3	5.9E-98	1.2E-94	
	GOTERM_BP_FAT	regulation of defense response	RT	329	13.0	9.4E-98	1.6E-94	
	GOTERM_BP_FAT	positive regulation of immune system process	RT	426	16.8	1.4E-95	2.1E-92	
	GOTERM_BP_FAT	regulation of response to stress	RT	491	19.4	3.0E-92	3.9E-89	
	GOTERM_BP_FAT	regulation of immune response	RT	382	15.1	9.9E-91	1.2E-87	
	GOTERM_BP_FAT	regulation of cytokine production	RT	323	12.8	4.2E-90	4.4E-87	
	GOTERM_BP_FAT	innate immune response	RT	383	15.1	2.7E-87	2.6E-84	
	GOTERM_BP_FAT	positive regulation of defense response	RT	230	9.1	1.0E-84	9.0E-82	
	GOTERM_BP_FAT	immune effector process	RT	366	14.4	2.5E-83	2.1E-80	
	GOTERM_BP_FAT	response to external stimulus	RT	705	27.8	1.3E-82	9.8E-80	
	GOTERM_BP_FAT	cell activation	RT	420	16.6	6.4E-82	4.5E-79	
	GOTERM_BP_FAT	leukocyte activation	RT	388	15.3	9.6E-78	6.3E-75	
	GOTERM_BP_FAT	inflammatory response	RT	282	11.1	3.1E-75	1.9E-72	
	GOTERM_BP_FAT	positive regulation of multicellular organismal process	RT	523	20.6	1.1E-72	6.5E-70	
	GOTERM_BP_FAT	regulation of innate immune response	RT	201	7.9	4.2E-72	2.3E-69	
	GOTERM_BP_FAT	intracellular signal transduction	RT	611	24.1	1.1E-71	5.8E-69	
	GOTERM_BP_FAT	positive regulation of cytokine production	RT	228	9.0	6.8E-71	3.4E-68	

Open in Excel to view

- Please note following picture is from last year. Results would little bit different when you re-run it.
- Use the FDR column for statistical significance.
- This table also gives you the different counts that were used in Fisher's Exact test.
  - Note that the *Pop Total* column is the total number of genes annotated in the database. Similarly, the *List Total* is the subset of your DEGs that are annotated to the database. Both of these numbers will change a bit with different databases (e.g., GO BP vs KEGG).
- Looking through the terms, we can see a large number related to immune response, as would be expected from a viral infection.

	A	C	D	E	F	G	H	I	K	L	M		
1	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichm	Bonferroni	Benjamini	FDR
2	GOTERM_BF GO:0002376		223	8.80378997	9.67E-80	ENSMUSG0C	2237	503	19672	3.89869633	6.36E-76	6.36E-76	6.01E-76
3	GOTERM_M_BF GO:0006954		157	6.19818397	4.83E-51	ENSMUSG0C	2237	379	1972	3.64286414	3.17E-47	1.59E-47	1.59E-47
4	GOTE_M_BF GO:0051607		100	3.94788788	7.16E-35	ENSMUSG0C	2237	229	1972	3.84013993	4.71E-31	1.57E-31	1.8E-31
5	GOTE_M_BF GO:0045087		189	7.46150809	1.60E-32	ENSMUSG0C	2237	683	1972	2.43345675	1.05E-28	2.63E-29	2.8E-29
6	GOTE_M_BF GO:0035458		45	1.77654955	4.32E-29	ENSMUSG0C	2237	60	1972	6.59544032	2.84E-25	5.68E-26	5.6E-26
7	KEGG_PATH mmu05169:t		98	3.86893012	2.05E-28	ENSMUSG0C	1176	231	841	3.22546898	6.65E-26	6.67E-26	4.3E-26
8	GOTE_M_BF GO:0032760		62	2.44769049	3.29E-27	ENSMUSG0C	2237	118	1972	4.62053446	2.17E-23	3.61E-24	3.1E-24
9	KEGG_PATH mmu04060:t		104	4.1058034	5.59E-23	ENSMUSG0C	1176	292	841	2.70787904	1.82E-20	9.11E-21	6.4E-21
10	KEGG_PATH mmu05140:L		44	1.73707067	2.95E-21	ENSMUSG0C	1176	70	841	4.77896016	9.60E-19	3.21E-19	2.7E-19
11	31		40	1.57915515	5.02E-21	ENSMUSG0C	2237	65	1972	5.41164334	3.31E-17	4.72E-18	4.6E-18
12	Database	71	35	1.38176076	1.22E-20	ENSMUSG0C	2237	2	1972	6.03504343	8.00E-17	1.00E-17	9.5E-18
13		46	52	2.0529017	1.72E-20	ENSMUSG0C	2237	2	1972	4.19526479	1.13E-16	1.25E-17	1.9E-17
14	KEGG_PATH mmu04064:t		54	2.13185946	1.90E-20	ENSMUSG0C	2237	1	841	3.91005831	6.16E-18	1.55E-18	1.4E-18
15	GOTERM_BF GO:0032755:		52	2.0529017	4.67E-20	ENSMUSG0C	2237	2	1972	4.11967444	3.07E-16	3.07E-17	2.90E-17
16	GOTERM_BF G Intersection	16636	6.74E-20	ENSMUSG0C	2237	39	1972	6.76455418					
17	GOTERM_BF G	Intersection	15985	7.25E-20	ENSMUSG0C	2237	473	1972	3.23297474				
18	GOTERM_BF GO:0009615		46	1.81602842	1.89E-19	ENSI		91	1972	4.44527846			
19	KEGG_PATH mmu04668:t		55	2.17133833	2.07E-19	ENSI	1176	113	841	3.70052224			
20	GOTERM_BF GO:0043123		58	2.28977497	2.58E-18	ENSI	2237	145	1972	3.51756817			
21	KEGG_PATH mmu04380:t		57	2.25029609	7.26E-18	ENSMUSG0C	1176	128	841	3.38566247			
22	GOTERM_BF GO:0006915		138	5.44808527	2.50E-17	ENSMUSG0C	2237	584	1972	2.07801544	1.65E-13	1.10E-14	1.04E-14
23	KEGG_PATH mmu05162:t		60	2.36873273	8.81E-17	ENSMUSG0C	1176	146	841	3.12447582	3.61E-14	4.10E-15	3.03E-15
24	GOTERM_BF GO:0071222		86	3.39518358	9.52E-17	ENSMUSG0C	2237	204	10672	2.77737128	7.30E-13	3.91E-14	3.70E-14
25	GOTERM_BF GO:0002250		80	3.1583103	1.07E-16	ENSMUSG0C	2237	2	95678	7.30E-13	4.13E-14	3.90E-14	
26	GOTERM_BF GO:0006935		53	2.09238058	1.47E-16	ENSMUSG0C	2237	1	19241	7.30E-13	5.36E-14	5.06E-14	
27	GOTERM_BF GO:0050729		43	1.69759179	1.92E-16	ENSMUSG0C	2237	93	19672	4.06600622	1.46E-12	6.64E-14	6.28E-14
28	KEGG_PATH mmu04061:t		46	1.81602842	4.36E-16	ENSMUSG0C	1176	95	8941	3.68139993	1.44E-13	1.78E-14	1.31E-14
29	KEGG_PATH mmu05152:t		67	2.64508488	5.43E-16	ENSMUSG0C	1176	180	8941	2.82996504	1.80E-13	1.97E-14	1.45E-14
30	GOTERM_BF GO:0032729		41	1.61863403	5.73E-16	ENSMUSG0C	2237	87	19672	4.14426135	3.65E-12	1.89E-13	1.78E-13
31	GOTERM_BF GO:0019221		57	2.25029609	6.98E-16	ENSMUSG0C	2237	156	19672	3.21316323	4.38E-12	2.19E-13	2.07E-13
32	GOTERM_BF GO:0030593		39	1.53967627	7.60E-16	ENSMUSG0C	2237	80	19672	4.28703621	5.11E-12	2.27E-13	2.15E-13
33	KEGG_PATH mmu05164:t		65	2.56612712	9.57E-16	ENSMUSG0C	1176	173	8941	2.8565776	3.25E-13	3.12E-14	2.31E-14

A second result from DAVID is a functional annotation clustering, where it attempts to group terms that are similar into “clusters” of terms.

Select “Functional Annotation Clustering” to open this result. There are options to adjust the way that the clustering is performed. Again right-click on the *Download File* link to save as a file david\_cluster.txt.

This can provide a useful way to parse a large list of similar terms. For example:

- Annotation Cluster 1 relates to chemotaxis, signaling pathway
- Annotation Cluster 2 has several viral infection terms
- Annotation Cluster 3 relates to cell differentiation, but also some diseases

**National Institutes of Health  
DAVID Bioinformatics**

## Functional Annotation Clustering

Help and Manual

Current Gene List: RNAseq\_example  
 Current Background: Mus musculus  
 2533 DAVID IDs

Options Classification Stringency Medium ▾  
[Rerun using options](#) [Create Sublist](#)

**710 Cluster(s)**

Annotation Cluster 1		Enrichment Score: 74.48	G	Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	defense response	RT	614	3.0E-131	3.1E-127
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of immune system process	RT	565	2.3E-121	1.2E-117
<input type="checkbox"/>	GOTERM_BP_FAT	immune response	RT	605	1.8E-111	6.4E-108
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of defense response	RT	329	9.4E-98	1.6E-94
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of immune response	RT	382	9.9E-91	1.2E-87
<input type="checkbox"/>	GOTERM_BP_FAT	innate immune response	RT	383	2.7E-87	2.6E-84
<input type="checkbox"/>	GOTERM_BP_FAT	immune effector process	RT	366	2.5E-83	2.1E-80
<input type="checkbox"/>	GOTERM_BP_FAT	response to external stimulus	RT	705	1.3E-82	9.8E-80
<input type="checkbox"/>	GOTERM_BP_FAT	response to biotic stimulus	RT	432	1.0E-67	4.5E-65
<input type="checkbox"/>	GOTERM_BP_FAT	response to external biotic stimulus	RT	409	8.4E-62	3.0E-59
<input type="checkbox"/>	GOTERM_BP_FAT	response to other organism	RT	408	1.2E-61	4.2E-59
<input type="checkbox"/>	GOTERM_BP_FAT	defense response to other organism	RT	255	1.1E-44	1.5E-42
<input type="checkbox"/>	GOTERM_BP_FAT	response to bacterium	RT	295	5.9E-39	5.5E-37
<input type="checkbox"/>	GOTERM_BP_FAT	response to molecule of bacterial origin	RT	153	1.5E-31	9.7E-30
<input type="checkbox"/>	GOTERM_BP_FAT	defense response to bacterium	RT	139	6.0E-16	1.8E-14
Annotation Cluster 2		Enrichment Score: 72.97	G	Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	cytokine production	RT	338	5.9E-98	1.2E-94
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of cytokine production	RT	323	4.2E-90	4.4E-87
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of multicellular organismal process	RT	523	1.1E-72	6.5E-70
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of cytokine production	RT	228	6.8E-71	3.4E-68

[Screenshot](#)

## 1.8 INSTRUCTOR DEMONSTRATION: Pathway Enrichment in Ingenuity Pathway Analysis (IPA)

We're going to use the same RNA-seq data set as in the DAVID exercise:

[https://wd.cri.uic.edu/pathway/RNAseq\\_example.xlsx](https://wd.cri.uic.edu/pathway/RNAseq_example.xlsx)

### 1.8.1 Upload data to IPA

We will save the entire “example\_diff” tab as a tab-delimited text file to upload to IPA

- We can filter for q-value and/or fold-change within IPA, which is more convenient if we want to try different thresholds later.
- Uploading the fold-changes also allows IPA to compute its z-scores and make inferences about up- or down-regulation of different pathways.
- Before uploading to IPA, you may want to make a new project first. In our case, we already have a test project created for the workshop.

To upload to IPA, go to *File > Upload Dataset*, and browse to the text file made above. IPA will automatically parse the file in the next dialog:

- Use *INFER OBSERVATIONS*: automatically identify the columns. Usually IPA does this correctly, but it's good to double-check:
  - ID column identified as Ensembl
  - Ignore second column: don't need the gene symbol if we have the Ensembl ID
  - Expr Log ratio is correct, observation name is “Infection/Control : LogFC”
  - Ignore fourth column: don't need the logCPM
  - Set fifth column to ignored: don't need p-value, since we'll use FDR
  - FDR is correct, observation name is also “Infection/Control : LogFC”
- Check *Dataset Summary* tab: how many IDs are recognized by IPA. “SHOW DETAILS” button gives more information about unmapped IDs.
- Click *SAVE*
  - IPA starts uploading the file
  - Ignore warning about missing metadata
  - Choose which project to save it to

Dataset Upload - RNAseq\_diff.txt

1. Select File Format: Flexible Format

2. Contains Column Header:  Yes  No

3. Select Identifier Type: Please assign at least one column below as “ID”, and assign the identifier type(s). Assign additional columns as ID to improve mapping coverage if desired.

4. Array platform used for experiments: Not specified/applicable

5. Use the dropdown menus to specify the column names that contain identifiers and observations. For observations, select the appropriate measurement value type.

Can double-check how many molecules were recognized

Have IPA guess the column types

ID/Observation Name	ID	Ignore	Infection/Co...	Ignore	Ignore	Infection/Co...
Measurement/Annotation	Ensembl					
1	# Gene ID	Gene name	Infection/Control ...	Infection/Control ...	Infection/Control ...	Infection/Control ...
2	ENSMUSC00000...	Xkr4	0.18	2.54	0.81	1.00
3	ENSMUSC00000...	Gm7341	-3.03	-2.72	0.78	1.00
4	ENSMUSC00000...	Gm19938	-0.44	5.39	0.04	0.24
5	ENSMUSC00000...	Rp1	-1.88	-2.16	0.32	0.95
6	ENSMUSC00000...	Sox17	-0.01	2.51	1.00	1.00
7	ENSMUSC00000...	Gm37587	-1.42	-2.44	0.78	1.00

Keep ENSEMBL ID, log fold-change, and Q-Value (FDR) columns

After loading, IPA opens a annotated dataset view, which lets you know how IPA is interpreting all of the genes. We can close this window.

### 1.8.2 Start core analysis

Pathway enrichment analysis is part of the “Core Analysis”. Other modules include enrichment against upstream regulators, disease biomarkers, toxicity functions, and others. IPA runs everything at once.

- Browse to the project/dataset (“Workshop Project/Dataset Files”
  - Other folders are generated automatically, used for different analyses
- Right-click on the dataset file and choose New > Core Analysis
  - Leave the defaults in the next screen: expression analysis, and basing z-scores on expression log ratio.
- Next customize the analysis: **set thresholds for FDR and/or fold-change.**
  - Set Expr False Discovery Rate (q-value) threshold to **0.05**.
  - *We won't set a threshold for the expression log ratio, but if, for example, you wanted to limit to a 2-fold change, you would choose -1 for Down and 1 for Up.*
  - **RECALCULATE** to update gene list.
  - *You can also customize how the databases are used in the other sections, but we'll leave them all as defaults for now.*
- Click **RUN ANALYSIS**
  - Confirm which project to save to. Results automatically saved in the *Analyses* folder.
  - Job is submitted to IPA's cloud server. Runs in background, and sends email notification when complete. Jobs typically take a couple minutes to run. You can close IPA while it runs.

For today, we've run this data set previously, so we'll skip to those results.

The screenshot shows the IPA software interface with the following details:

- General Settings:** Population of genes to consider for p-value calculations: Reference Set - Ingenuity Knowledge Base (Genes Only).
- Relationships to consider:** Affects networks and upstream regulator analysis. Options: Direct and Indirect Relationships (selected) and Direct Relationships.
- Optional Analyses:** My Project, My Pathways, My Lists.
- Analysis Filter Summary:** Contains options for filtering analysis-ready molecules based on various biological processes and chemical entities.
- Set Cutoffs:**
  - Dataset Column: Infection/Control : logFC
  - Measurement Value Type: Expr Log Ratio
  - Range: -11.2952 to 8.9306
  - Cutoff: 0.05
  - Buttons: Recalculate, Set FDR < 0.05 threshold
  - Text: 2554 analysis-ready molecules across observations (709 Down and 1845 Up)
- Preview Dataset RNAseq\_diff:**
  - Analysis-Ready (2554): Mapped IDs (2560), Unmapped IDs (23), All IDs (2593), Metadata.
  - Table view showing columns: ID, Symbol, Entrez Gene Name, Location, Type(s), Drug(s).
  - Data rows (partial):
 

ID	Symbol	Entrez Gene Name	Location	Type(s)	Drug(s)
ENSMUS000000043644	0610009L18Rik	RIKEN cDNA 0610009L18 gene	Other	other	
ENSMUS000000092203	1110038812Rik	RIKEN cDNA 1110038812 gene	Other	other	
ENSMUS000000111867	1190001M18Rik	RIKEN cDNA 1190001M18 gene	Other	other	
ENSMUS00000047150	17000047C19Rik	RIKEN cDNA 17000047C19 gene	Other	other	
ENSMUS000000100147	1700047M11Rik	RIKEN cDNA 1700047M11 gene	Other	other	
ENSMUS000000055007	1700109K09Rik	RIKEN cDNA 1700109K09 gene	Other	other	
  - Buttons: Run Analysis, Cancel.

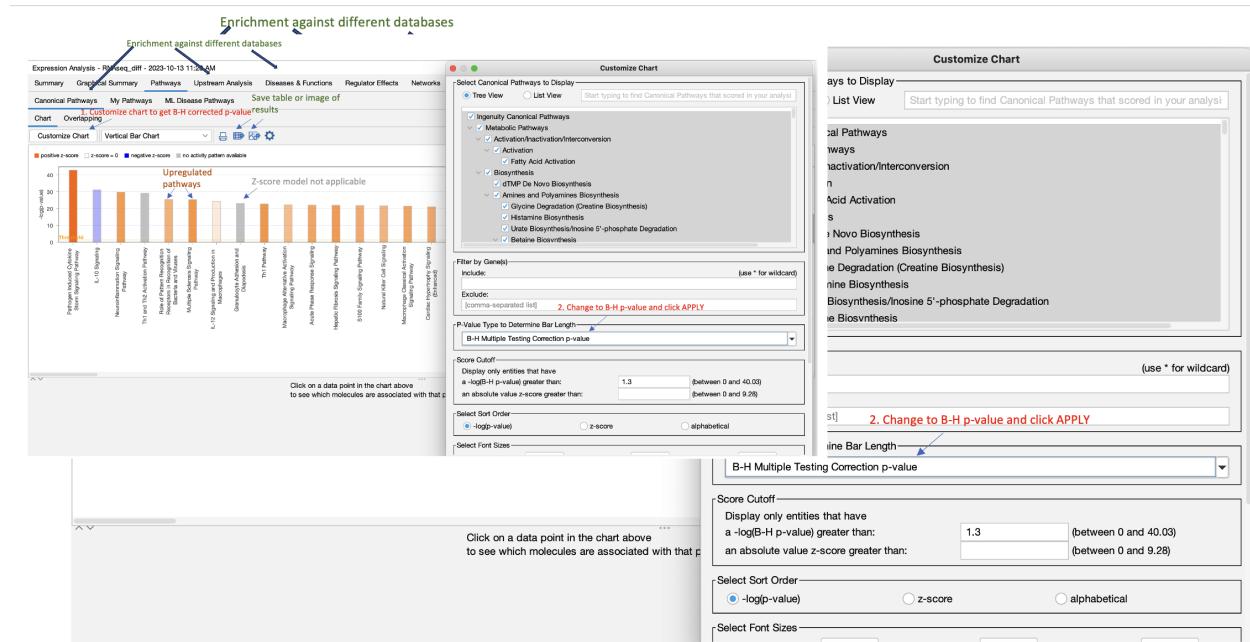
### 1.8.3 View IPA results

Browse to the **Analyses** folder and double-click a result to open the core analysis.

- Different tabs for different databases. In particular:
  - Pathways -> Canonical Pathways are biological/functional pathways, conceptually similar to GO BP or KEGG pathways.
  - Upstream Analysis compare target molecules of regulators to the input. Useful for inferring which regulators may be responsible for the changes in the data.
  - Diseases and Functions are biomarkers for different disease or broad functional categories.

Start with the *Canonical Pathways* tab.

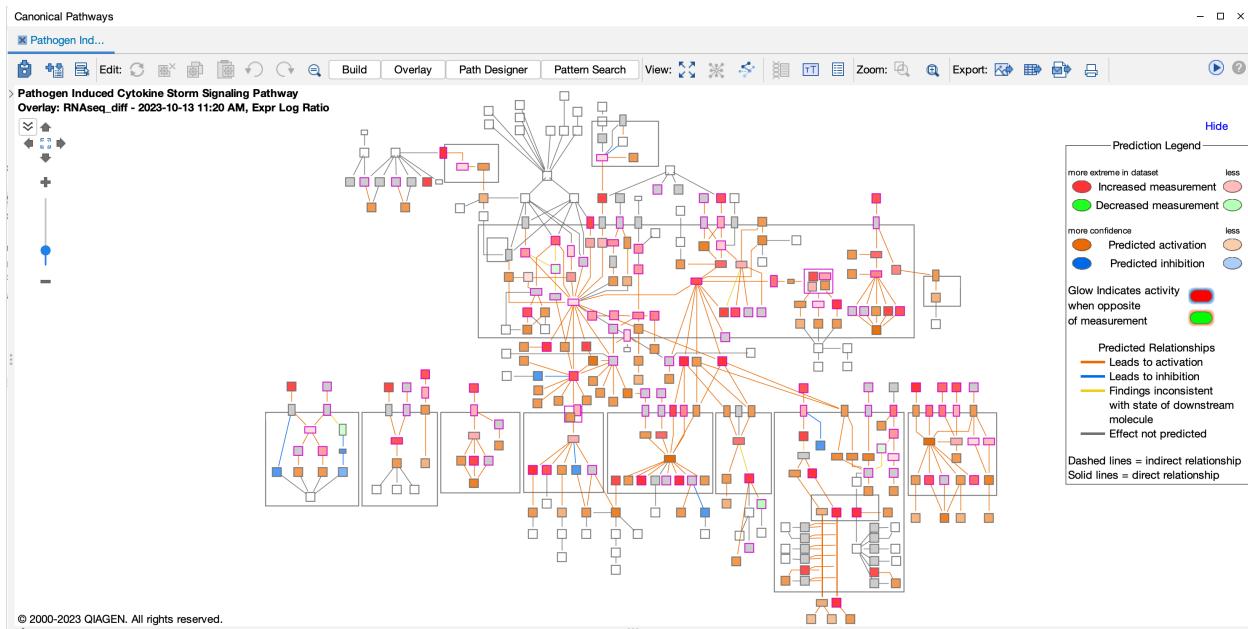
- The bar plot shows significance in  $-\log_{10}$  scale (i.e., 2 = p-value of 0.01). Bars are colored by z-score; gray bars indicate that the z-scoring model could not be computed for that pathway.
- IMPORTANT:** IPA reports nominal p-value by default. **This is wrong:** it guarantees many false positives. IPA *does* compute the B-H corrected p-value (FDR), just need to turn it on:
  - Click the **CUSTOMIZE CHART** button.
  - At the bottom, change **Select Scoring Method** to *B-H Multiple Testing Correction p-value*, then **APPLY**.
  - Need to make similar adjustments in other tabs.
- Positive z-scores indicate up-regulated pathways. Since our comparison was infection/control, this means up-regulated in infection.
- You can also export these results in tabular text or Excel formats, or the barplot in a PDF.



#### 1.8.4 Look at a specific IPA pathway map

If you want more detail into the structure of a particular pathway, double-click it (twice). This will open an interactive image.

- There are a number of ways to manipulate these images, which we will detail more in the afternoon.



## 1.9 Pathway Enrichment in R

### 1.9.1 Data input (overview)

We will be using the same RNA-seq data set as before:

[https://wd.cri.uic.edu/pathway/RNAseq\\_example.xlsx](https://wd.cri.uic.edu/pathway/RNAseq_example.xlsx)

- Use `RNAseq_example.txt` file from the DAVID exercise
- Compare to a set of anti-viral genes, which were determined from a literature review
- Also need the total number of genes in the genome, from the “example\_norm” tab (save as tab-delimited text to `RNAseq_norm.txt`)

*For consistency, we will obtain these data sets from RIC github repository. However, you can run the same exercise by saving tab-delimited text files from Excel and loading into R.*

### 1.9.2 Enrichment test

Open R Studio to perform this analysis, navigating to the folder where you saved your `RNAseq_example.txt` and `RNAseq_norm.txt` files.

```
# read in gene list from our server
degs <- read.table("https://wd.cri.uic.edu/pathway/RNAseq_example.txt")
# remake as a vector
degs <- degs[,1]
# read in the antiviral gene list from RIC github repository
antiviral <- read.table("https://wd.cri.uic.edu/pathway/antiviral_list.txt")
# remake as a vector
antiviral <- antiviral[,1]
head(antiviral)

## [1] "ENSMUSG00000026104" "ENSMUSG00000042349" "ENSMUSG00000026896"
## [4] "ENSMUSG00000027639" "ENSMUSG00000028037" "ENSMUSG00000040296"

# read in norm table from our server
norm <- read.table("https://wd.cri.uic.edu/pathway/RNAseq_norm.txt",
  header=T, row.names=1, sep="\t")
all_genes <- rownames(norm)
# check how big each list is
length(degs)

## [1] 2633
length(antiviral)

## [1] 44
length(all_genes)

## [1] 25931

# obtain true/false vectors for all genes based on intersection
# with degs or antiviral
antiviral_list <- all_genes %in% antiviral
degs_list <- all_genes %in% degs
# we can use table to confirm the number of genes in each
table(antiviral_list)

## antiviral_list
## FALSE  TRUE
## 25887    44
```

```



```

Antiviral genes are very strongly enriched: odds ratio is ~70x, p-value is very small.

### 1.9.3 More efficient processing (bonus exercise)

Note: you may want to write a function to allow you do this calculation more rapidly. *We will skip this for today, but here's an example:*

```

pathway_enrich <- function( degs, pathway, all_genes ){
  pathway_list <- all_genes %in% pathway
  degs_list <- all_genes %in% degs
  fet.table <- table(data.frame(pathway_list, degs_list))
  fet <- fisher.test(fet.table)
  # return the odds ratio and p-value as a vector
  result <- c(fet$p.value, fet$estimate)
  names(result)[1] = "p.value"
  return(result)
}
pathway_enrich(degs, antiviral, all_genes)

##      p.value    odds ratio
## 9.138488e-34 7.010235e+01

```

## 1.10 Pathway Enrichment in GSEA

### 1.10.1 Prepare input .txt and .cls files

We will be using the same RNA-seq data set as before:

[https://wd.cri.uic.edu/pathway/RNAseq\\_example.xlsx](https://wd.cri.uic.edu/pathway/RNAseq_example.xlsx)

Input .txt file prepared from the normalized expression table (“example\_norm” tab). *Again, for consistency we will obtain this file from RIC github repository. However, you can run the same exercise by saving tab-delimited text files from Excel and loading into R.*

Open R Studio to prepare the input files. Use ENSEMBL IDs to go with GSEA’s built-in CHIP file for converting to gene symbols. Preparation steps:

- Remove genes with average CPM < 1
- Log-scale the expression

1. Read from our server; the `colClasses` parameter tells R to ignore the columns named Gene.name

```
norm <- read.table("https://wd.cri.uic.edu/pathway/RNAseq_norm.txt",
  header=T, row.names=1, sep="\t", colClasses=c(Gene.name=NULL))
```

2. Subset to CPM > 1

```
norm <- norm[rowMeans(norm) > 1,]
```

3. log-scale with a pseudocount

```
norm <- log2(norm + 0.1)
```

4. Duplicate two extra columns of the gene IDs and change the names of the new columns

```
norm <- cbind(rownames(norm), rownames(norm), norm)
colnames(norm)[1:2] = c("NAME", "DESCRIPTION")
```

5. Write the file

```
write.table(norm, "RNAseq_for_gsea.txt", col.names=T, row.names=F, quote=F, sep="\t")
```

6. Make the .cls file. Use the `append=T` parameter in `write.table` to keep adding to the same file, and transpose vectors so they’re written along the rows.

```
groups <- c(rep("Control",4),rep("Infection",4))
write.table(t(c(8, 2, 1)), "RNAseq_for_gsea.cls", col.names=F, row.names=F)
write.table("# Control Infection", "RNAseq_for_gsea.cls", col.names=F, row.names=F,
  append=T, quote=F)
write.table(t(groups), "RNAseq_for_gsea.cls", col.names=F, row.names=F, append=T, quote=F)
```

NOTE: you can also download these completed files if you have trouble with the above steps:

[https://wd.cri.uic.edu/pathway/RNAseq\\_for\\_gsea.txt](https://wd.cri.uic.edu/pathway/RNAseq_for_gsea.txt)

[https://wd.cri.uic.edu/pathway/RNAseq\\_for\\_gsea.cls](https://wd.cri.uic.edu/pathway/RNAseq_for_gsea.cls)

### 1.10.2 Go to GSEA

Go to the *GenePattern* cloud website: <https://cloud.genepattern.org/gp/pages/login.jsf>

After logging in, enter “GSEA” under the Modules search box, and pick the first option (GSEA).

- NOTE: if you’re using your own ranking, you should select the *GSEAPreranked* module.

The screenshot shows the GenePattern web interface. At the top, there is a navigation bar with links for 'Modules & Pipelines', 'Suites', 'Job Results', 'Resources', and 'Help'. Below the navigation bar, a search bar contains the text 'GSEA'. Underneath the search bar, a message says 'No Jobs Processing' and a 'Browse Modules' button is available. On the left side, there are sections for 'Favorite Modules' and 'Recent Modules', both of which list 'GSEA'. The main area is titled 'Search: GSEA' and displays a list of modules. One module, 'GSEA', is highlighted with a red circle. Other listed modules include 'CollapseDataset', 'ConstellationMap', 'GSEA LeadingEdgeViewer', 'GSEA Preranked', 'SsGSEA', 'TCGA.SampleSelection', and 'Txiimport.DESeq2'. Each module entry includes a brief description and a link to its documentation.

### 1.10.3 Set up and run GSEA

- **expression dataset:** Select our .txt file. The help info says to use a .gct or .res file, but .txt works also.
- **gene sets database:** Let's use the GO database. Select *m5.go\_bp.v2023.1.Mm.symbols.gmt [Gene Ontology]*.
  - Here's where you could upload a .gmt file if you wanted to test a custom gene set.
- **number of permutations:** Leave at 1000 for now, though you may want to consider 10,000 for more accurate p-values.
- **phenotype labels:** Select our .cls file.
- **permutation type:** Set to *gene\_set*, since we have <7 replicates per group.
- **chip platform file:** Browse to *Mouse\_ENSEMBL\_Gene\_ID\_MSigDB.v2023.1.Mm.chip*, to match the ENSEMBL IDs to gene symbols.
- **output file:** Set to *gsea\_out.zip*.
- Then *Run* the job at the bottom.

**Basic parameters**  
These parameters are essential for the analysis.

<b>expression dataset*</b>	<input type="button" value="Hide Files... (Selected 1 files)"/> <input type="button" value="Click 'upload file' and browse to RNaseq_for_gsea.txt"/>	<input type="checkbox"/> Batch
<b>gene sets database*</b>	<input type="button" value="Select a file or Upload your own file"/> <input type="button" value="-- select an option --, -- select an option --, m5.go_bp.v2023.1.Mm.symbols.gmt"/>	<input type="checkbox"/> Batch
Gene sets database from GSEA website. Upload a gene set if your gene set is not listed as a choice from MSigDB.		
<b>number of permutations*</b>	1000	<input type="checkbox"/> Batch
<b>phenotype labels*</b>	<input type="button" value="Hide Files... (Selected 1 files)"/> <input type="button" value="Click 'upload file' and browse to RNaseq_for_gsea.cls"/>	<input type="checkbox"/> Batch
Cls file - .cls, must be binary		
<b>target profile</b>	<input type="text"/>	<input type="checkbox"/> Batch
Name of the target phenotype profile. Only applicable if class file defines continuous labels (one or more phenotype profiles). Leave blank if class file defines a discrete phenotype.		
<b>permutation type*</b>	gene_set	<input type="checkbox"/> Batch
Type of permutations to perform		
<b>collapse dataset*</b>	Collapse	<input type="checkbox"/> Batch
Select whether to collapse each probe set in the expression dataset into a single vector for the gene, which gets identified by its gene symbol. It is also possible to remap symbols from one namespace to another without collapsing (an error will occur if multiple source genes map to a single destination gene). No_Collapse will use the dataset as-is.		
<b>chip platform file</b>	<input type="button" value="Select a file or Upload your own file"/> <input type="button" value="Select Mouse_ENSEMBL_Gene_ID_MSigDB.v2023.1.Mm.chip"/>	<input type="checkbox"/> Batch
DNA Chip (array) annotation file from GSEA website. Upload your own chip file if the one corresponding to your DNA Microarray platform is not listed in the drop-down menu. A chip file is only required if collapse dataset is set to true.		
<b>output file name</b>	<input type="text" value="gsea_out.zip"/>	<input type="checkbox"/> Batch
Name of the output ZIP file.		

You can close your browser while the job runs. When you log back in, you can check the status or see results.

#### 1.10.4 Download results

After results are finished, browse to the *Jobs* tab and click on the run name. Choose *Download Job* on the right.

- Note that you can also view the code to run GSEA in Java, MATLAB, R, or Python if you want to learn how.
- **For the impatient:** You can download the results from our server as well...

[https://wd.cri.uic.edu/pathway/gsea\\_out.zip](https://wd.cri.uic.edu/pathway/gsea_out.zip)

The screenshot shows the GenePattern interface with the following details:

- Header:** GenePattern, Modules & Pipelines, Suites, Job Results, Resources, Help.
- Job Status Bar:** Shows 'No Jobs Processing' and a Refresh button.
- Job Details:** Job ID: 494779, Name: GSEA (494779).
  - Job Status:** View the job status page for this job.
  - Download Job:** Download a copy of this job, including all result files. (This is the target of the red arrow.)
  - Reload Job:** Reload this job using the same input parameters.
  - Delete:** Delete the job.
- View Code:** Options to View Java Code, View MATLAB Code, View R Code, and View Python Code.
- File Download:** A file named 787646.tsv (1.3 MB) is listed with a download link.
- Timestamp:** Last modified: Thu Feb 23 16:54:26 UTC 2023.
- Log:** Log entries starting with 'Feb 23 16:54:26 UTC 2023'.
- Footer:** Footer text including 'ver on 3/9/23 11:15 PM'.

### 1.10.5 View results

Unpack the zip folder, and find the main result file, **index.html**, double-click to open in a browser.

- There are two sets of results, one for gene sets up-regulated in Control, and another for gene sets up-regulated in Infection.
- Click on the Detailed enrichment results in html or Excel to see details (note: Excel is really just tab-delimited text). Html is more interactive.

#### GSEA Report for Dataset RNAseq\_for\_gsea

##### Enrichment in phenotype: Control (4 samples)

- 568 / 3711 gene sets are upregulated in phenotype **Control**
- 57 gene sets are significant at FDR < 25%
- 62 gene sets are significantly enriched at nominal pvalue < 1%
- 109 gene sets are significantly enriched at nominal pvalue < 5%
- **Snapshot** of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in TSV](#) format (tab delimited text)
- [Guide to interpret results](#)

##### Enrichment in phenotype: Infection (4 samples)

- 3143 / 3711 gene sets are upregulated in phenotype **Infection**
- 1374 gene sets are significantly enriched at FDR < 25%
- 907 gene sets are significantly enriched at nominal pvalue < 1%
- 1174 gene sets are significantly enriched at nominal pvalue < 5%
- **Snapshot** of enrichment results
- Detailed [enrichment results in html](#) format See up-regulated pathways in infection
- Detailed [enrichment results in TSV](#) format (tab delimited text)
- [Guide to interpret results](#)

Within the detailed view:

- See the FDR q-value for statistical significance
- The pathway name link will take you to MSigDB
- The *Details* link will take you to the leading edge figure and statistics, gene list, plus a heatmap

Table: Gene sets enriched in phenotype Infection (4 samples) [ <a href="#">plain text format</a> ]									
	GS follow link to MSigDB	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	GOBP_ADAPTIVE_IMMUNE_RESPONSE <span style="color:red">Link to gene set on MSigDB</span>	289	-0.83	-2.19	0.000	0.000	0.000	1241	tags=55%, list=9%, signal=59%
2	GOBP_POSITIVE_REGULATION_OF_RESPONSE_TO_BIOTIC_STIMULUS	127	-0.85	-2.18	0.000	0.000	0.000	1185	tags=50%, list=9%, signal=54%
3	GOBP_INNATE_IMMUNE_RESPONSE	496	-0.83	-2.18	0.000	0.000	0.000	1297	tags=48%, list=10%, signal=49%
4	GOBP_DEFENSE_RESPONSE_TO_BACTERIUM	128	-0.85	-2.17	0.000	0.000	0.000	1388	tags=45%, list=10%, signal=72%
5	GOBP_RESPONSE_TO_INTERFERON_GAMMA	101	-0.85	-2.16	0.000	0.000	0.000	1041	tags=55%, list=8%, signal=60%
6	GOBP_IMMUNE_EFFECTOR_PROCESS	404	-0.81	-2.15	0.000	0.000	0.000	1305	tags=47%, list=10%, signal=51%

## 2 Afternoon

### 2.1 Barplot visualization

We're using chart results from an analysis on DAVID.

Open R Studio to perform this analysis:

```
# read in the results from DAVID
# the quote="" helps us to parse pathway names that have ' or " in them
david <- read.table("https://wd.cri.uic.edu/pathway/david_chart.txt",
  header=T, sep="\t", quote="")
# check how the names are interpreted in R
colnames(david)

## [1] "Category"          "Term"            "Count"           "X."
## [5] "PValue"             "Genes"            "List.Total"      "Pop.Hits"
## [9] "Pop.Total"          "Fold.Enrichment" "Bonferroni"      "Benjamini"
## [13] "FDR"

# sort by significance
david <- david[order(david$FDR),]
# for now, we'll focus on the KEGG pathways with FDR < 1%
david.kegg <- david[david$Category=="KEGG_PATHWAY" & david$FDR < 0.01,]
# log-scale the FDR and the enrichment ratio
david.kegg$logFDR <- -log10(david.kegg$FDR)
# the KEGG pathway names have IDs in them, fix it so that we just plot the name
head(david.kegg$Term)

## [1] "mmu04060:Cytokine-cytokine receptor interaction"
## [2] "mmu05140:Leishmaniasis"
## [3] "mmu04668:TNF signaling pathway"
## [4] "mmu05168:Herpes simplex infection"
## [5] "mmu04380:Osteoclast differentiation"
## [6] "mmu05162:Measles"

david.kegg$Term <- gsub("mmu[0-9]*:", "", david.kegg$Term)
head(david.kegg$Term)

## [1] "Cytokine-cytokine receptor interaction"
## [2] "Leishmaniasis"
## [3] "TNF signaling pathway"
## [4] "Herpes simplex infection"
## [5] "Osteoclast differentiation"
## [6] "Measles"

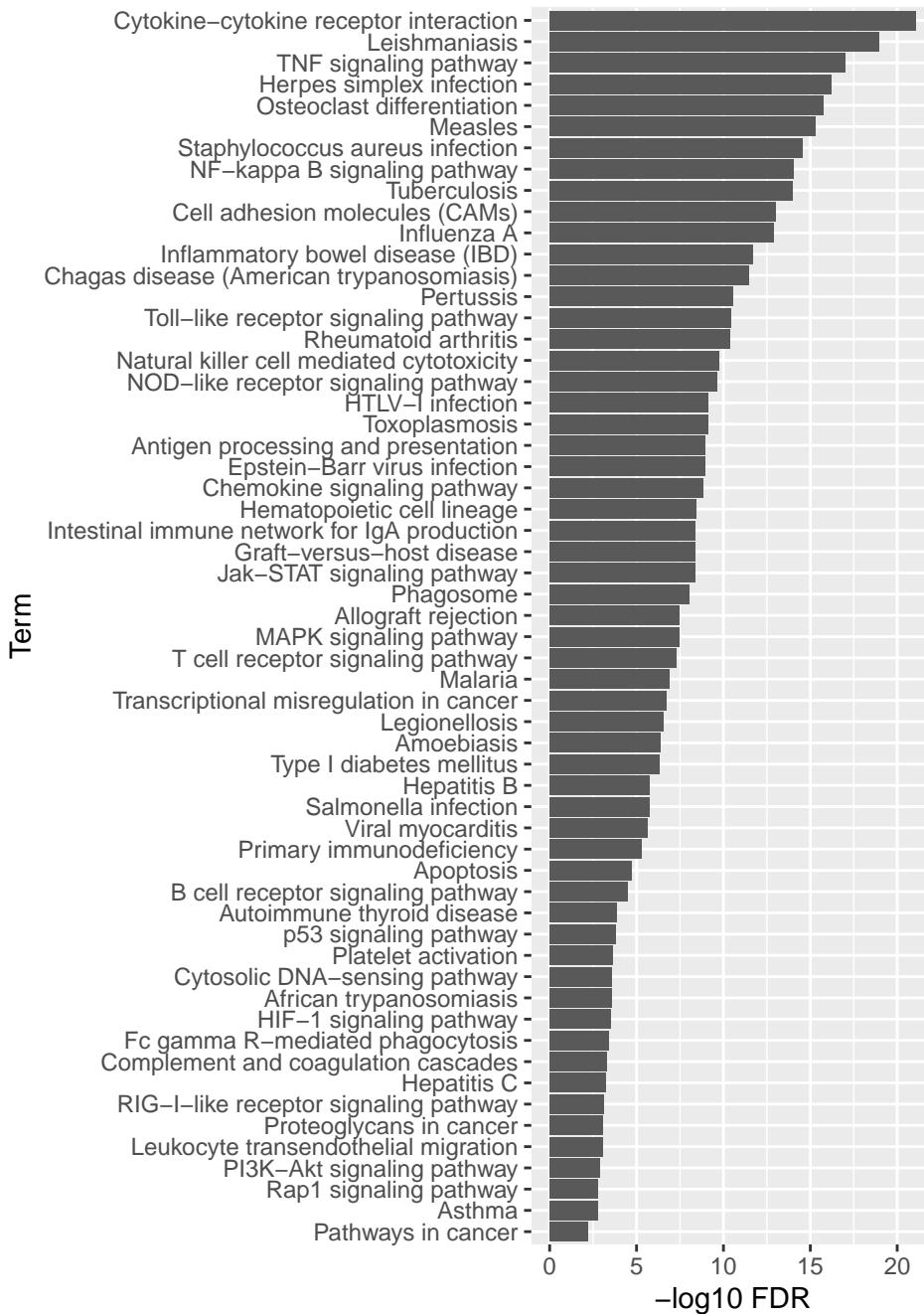
# load the ggplot2 library
library(ggplot2)
```

*NOTES FOR PLOT:*

- Use `coord_flip()` to plot the bars horizontally
- Use `scale_x_discrete()` to prevent `ggplot2` from reordering the columns
- Use `rev()` to reverse the order so that the most significant is on top

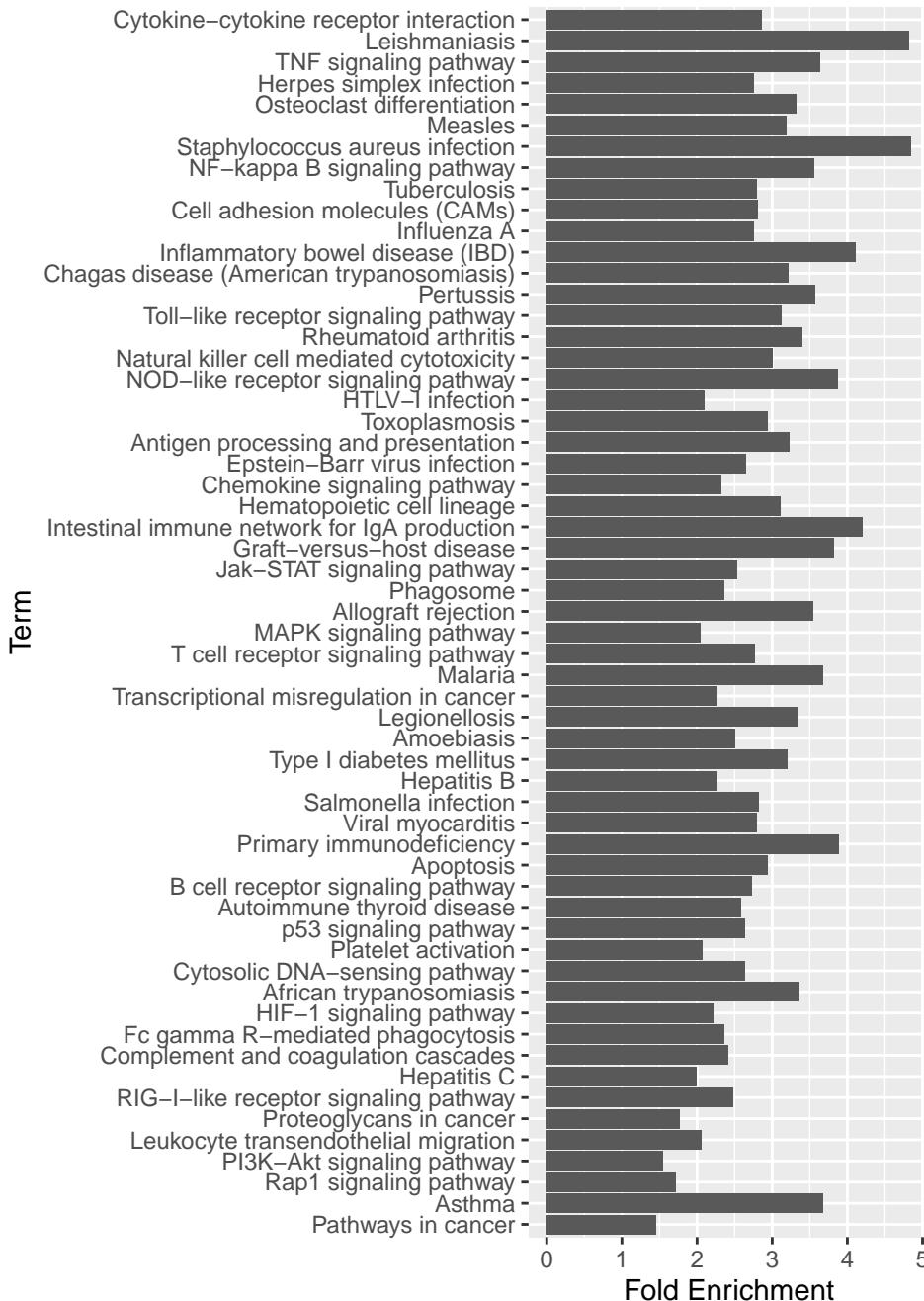
### 2.1.1 Plot $-\log_{10}$ FDR

```
ggplot(data=david.kegg, aes(x=Term, y=logFDR)) + geom_col() + coord_flip() +
  scale_x_discrete(limits=rev(david.kegg$Term)) + labs(y="-log10 FDR")
```



## 2.1.2 Plot the fold enrichment

```
ggplot(data=david.kegg, aes(x=Term, y=Fold.Enrichment)) + geom_col() + coord_flip() +
  scale_x_discrete(limits=rev(david.kegg$Term)) + labs(y="Fold Enrichment")
```



## 2.2 Pathway comparison plots

Compare pathway statistics for KEGG pathways in 3 different pathway enrichment results. These were generated on DAVID from three different gene lists generated from an RNA-seq clustering analysis.

For these files, we generated the Functional Annotation Chart in DAVID using a minimum of 1 molecule and a nominal p-value < 1, to include all terms in the comparison. We'll filter for overall significance below.

### 2.2.1 Prepare combined data set

Download the three result files, and combine together into a single data frame. This takes a bit of futzing:

- Fix column names
  - Merge data sets together and fix missing values
  - Subset to just significant terms
  - Log-scale FDR corrected p-values
  - Remove KEGG IDs from pathway names
1. Download each file

```
kegg1 <- read.table("https://wd.cri.uic.edu/pathway/cluster1_KEGG.txt",
  header=T, sep="\t", quote="")
kegg2 <- read.table("https://wd.cri.uic.edu/pathway/cluster2_KEGG.txt",
  header=T, sep="\t", quote="")
kegg3 <- read.table("https://wd.cri.uic.edu/pathway/cluster3_KEGG.txt",
  header=T, sep="\t", quote="")
```

2. We'll make the comparison on the basis of the FDR, so get just this value

```
kegg1_subset <- kegg1[,c("Term", "FDR")]
kegg2_subset <- kegg2[,c("Term", "FDR")]
kegg3_subset <- kegg3[,c("Term", "FDR")]
```

3. Use the cluster name for the column names

```
colnames(kegg1_subset)[2] <- "Cluster1"
colnames(kegg2_subset)[2] <- "Cluster2"
colnames(kegg3_subset)[2] <- "Cluster3"
```

4. Now we want to combine these results.

- We could use `merge()` to combine 2 data frames, but we have 3 to combine.
- The `Reduce` function repeatedly applies a given function to a list of inputs. So, this will apply `merge` across all three results

```
kegg_merged <- Reduce(function(x,y) merge(x=x, y=y, by="Term", all=T),
  list(kegg1_subset, kegg2_subset, kegg3_subset))
```

5. Missing values are not significant, so set them to FDR = 1

```
head(kegg_merged)
```

```
##                                     Term Cluster1 Cluster2
## 1 mmu00010:Glycolysis / Gluconeogenesis 1.0000000 0.9465988
## 2 mmu00020:Citrate cycle (TCA cycle) 0.9999999      NA
## 3 mmu00030:Pentose phosphate pathway 0.9999998 1.0000000
## 4 mmu00040:Pentose and glucuronate interconversions 1.0000000 1.0000000
## 5 mmu00051:Fructose and mannose metabolism 1.0000000 0.9999998
## 6 mmu00052:Galactose metabolism 1.0000000 1.0000000
##   Cluster3
## 1 1.0000000
```

```

## 2 0.9995364
## 3      NA
## 4 1.0000000
## 5 1.0000000
## 6 1.0000000

kegg_merged[is.na(kegg_merged)] <- 1

```

6. Remake this as a data frame with terms as the row names

```

kegg_df <- data.frame(kegg_merged[,c(2:ncol(kegg_merged))])
rownames(kegg_df) = kegg_merged[,1]

```

7. Subset to the set of terms with FDR < 0.05 for at least one cluster

```

kegg_subset <- kegg_df[apply(kegg_df, 1, min) < 0.05 ,]

```

8. Remove the term IDs in the names

```

rownames(kegg_subset) <- gsub("mmu[0-9]*:", "", rownames(kegg_subset))

```

9. This file is now nicely organized and scaled

```

head(kegg_subset)

```

```

##                               Cluster1     Cluster2     Cluster3
## Oxidative phosphorylation 1.349328e-03 3.585680e-08 0.999782250
## Pyrimidine metabolism    9.422153e-01 3.476666e-02 0.009506058
## Ribosome                  6.922510e-10 2.964080e-16 1.000000000
## DNA replication           7.539350e-06 2.945990e-04 0.963056903
## Spliceosome                3.786304e-01 1.554310e-14 0.999992409
## Base excision repair       1.000000e+00 9.999999e-01 0.002142762

```

## 2.2.2 Plot as side-by-side barplots

We need to do a bit more reformatting to work with *ggplot2*, as it requires data in the long format to plot in side-by-side barplots.

1. Sort on most significant (lowest value). Could also use average significance: just replace “max” with “mean” in the apply.

```

kegg_subset2 <- kegg_subset[order(apply(kegg_subset, 1, min)),]

```

2. Convert to long format using the *tidyverse* package
  - a. First need to add back the row names as a column.
  - b. Set the column name for the row names.
  - c. Then use *pivot\_longer* to convert the data frame into long format.

```

library(tidyverse)
kegg_subset2 <- data.frame(Term=rownames(kegg_subset2), kegg_subset2)
head(kegg_subset2)

```

```

##                               Term     Cluster1
## Ribosome                   Ribosome 6.922510e-10
## Spliceosome                Spliceosome 3.786304e-01
## Parkinson's disease        Parkinson's disease 2.707503e-02
## Huntington's disease      Huntington's disease 6.105167e-02
## Systemic lupus erythematosus Systemic lupus erythematosus 9.526860e-10
## Alzheimer's disease        Alzheimer's disease 1.193811e-01
##                               Cluster2     Cluster3

```

```

## Ribosome          2.96408e-16 1.0000000000
## Spliceosome      1.55431e-14 0.999992409
## Parkinson's disease 1.23347e-10 0.981089407
## Huntington's disease 6.58956e-10 0.851762867
## Systemic lupus erythematosus 1.00000e+00 0.000195325
## Alzheimer's disease 1.23151e-09 0.826030417

kegg_long <- pivot_longer(kegg_subset2, !Term, names_to = "cluster")
head(kegg_long)

```

```

## # A tibble: 6 x 3
##   Term     cluster   value
##   <chr>    <chr>     <dbl>
## 1 Ribosome Cluster1 6.92e-10
## 2 Ribosome Cluster2 2.96e-16
## 3 Ribosome Cluster3 1    e+ 0
## 4 Spliceosome Cluster1 3.79e- 1
## 5 Spliceosome Cluster2 1.55e-14
## 6 Spliceosome Cluster3 1.00e+ 0

```

### 3. Create barplot using `ggplot2`

- `fill=variable` will color based on the different clusters in the variable column
- `position=dodge` will put the bars next to each other
- `scale_x_discrete`, To order the items using the `kegg_subset2` list.
- `scale_y_continuous`, This will use our custom made reverse log transformation for the scale.

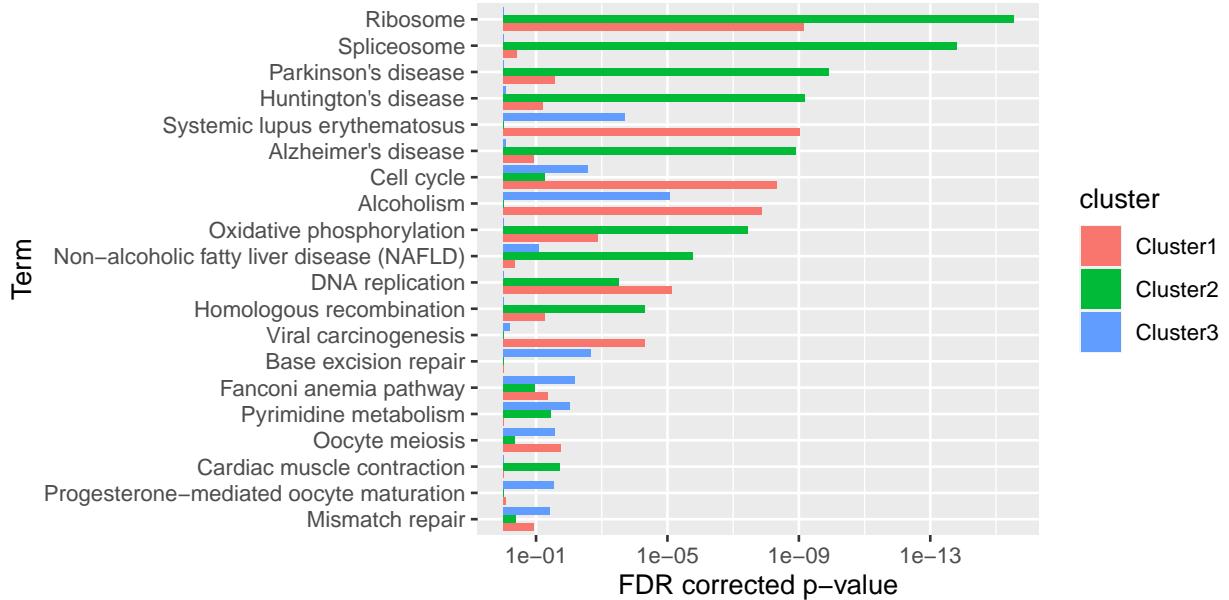
```

library(ggplot2)
library(scales)

# We only really need to do this once, but this will allow us
# to quickly create a reverse log scale in ggplot2
reverselog_trans <- function(base=exp(1)) {
  trans <- function(x) { -log(x, base) }
  inv <- function(x) { base ^ (-x) }
  trans_new(paste0("reverselog-", format(base)), trans, inv,
            log_breaks(base=base),
            domain = c(1e-100, Inf))
}

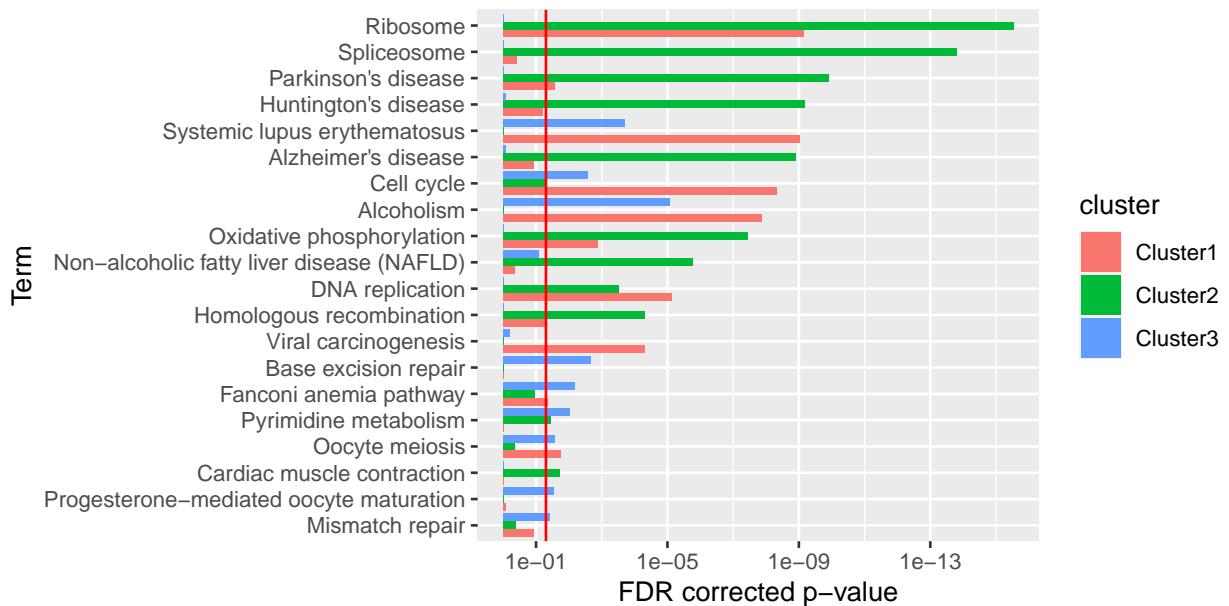
ggplot(data=kegg_long, aes(x=Term, y=value, fill=cluster)) +
  geom_col(position='dodge') + coord_flip() +
  scale_x_discrete(limits=rev(kegg_subset2$Term)) +
  scale_y_continuous("FDR corrected p-value", trans=reverselog_trans(10))

```



4. (OPTIONAL) Use `geom_hline` to create a red line at an FDR value of 0.05.

```
ggplot(data=kegg_long, aes(x=Term, y=value, fill=cluster)) +
  geom_col(position='dodge') + coord_flip() +
  scale_x_discrete(limits=rev(kegg_subset2$Term)) +
  scale_y_continuous("FDR corrected p-value", trans=reverselog_trans(10)) +
  geom_hline(yintercept = 0.05, color="red")
```



### 2.2.3 Plot as a heatmap

Also plot the prepared table as a heatmap.

1. Make a color scale going from white (non significant) to red (most significant) with the *circlize* package set the middle of the color range to be 0.05 significance

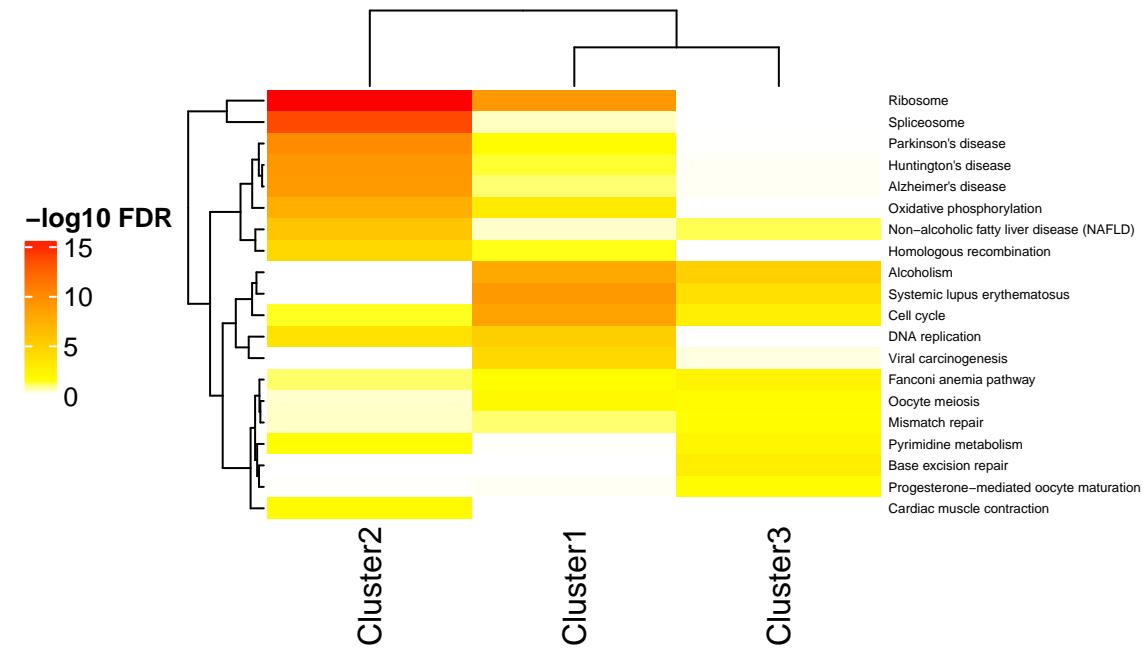
```
library(circlize)
```

```
kegg_subset_revlog <- -log10(kegg_subset)
col_fun <- colorRamp2(c(0, -log10(0.05), max(unlist(kegg_subset_revlog))),  
c("white", "yellow", "red"))
```

2. Plot the heatmap, with the legend (colorbar) on the left to avoid running into the pathway names also make the row names smaller

```
library(ComplexHeatmap)
```

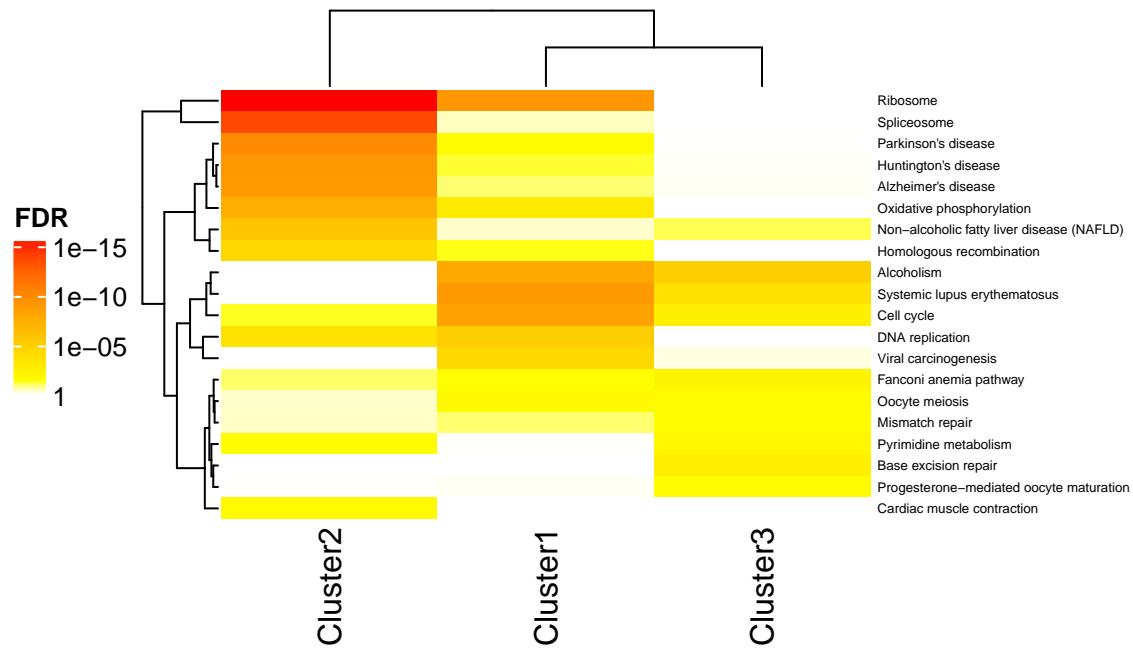
```
heatmap <- Heatmap(as.matrix(kegg_subset_revlog), name = "-log10 FDR", col = col_fun,  
row_names_gp = gpar(fontsize = 5))  
draw(heatmap, heatmap_legend_side = "left")
```



3. (OPTIONAL) Replot with the breaks as the actual FDR corrected p value on the legend.

```
# We saw the breaks were at 0, 5, 10, and 15  
breaks <- c(0, 5, 10, 15)
```

```
heatmap <- Heatmap(as.matrix(kegg_subset_revlog),  
col = col_fun,  
heatmap_legend_param = list(  
at = breaks,  
labels = 10 ^ (-1 * breaks),  
title = "FDR"),  
row_names_gp = gpar(fontsize = 5))  
draw(heatmap, heatmap_legend_side = "left")
```



## 2.3 Heatmap of expression patterns in pathway

### 2.3.1 Read in data sets

*NOTE: some of these are repeats of steps we ran before.* Pick the top KEGG pathway from our DAVID results and make a heatmap of the expression levels for genes in that pathway.

Use the results from DAVID from before, and the corresponding RNA-seq data set.

- Use the chart results from DAVID (**david\_chart.txt**).
- Use the normalized expression table (**RNAseq\_norm.txt**).

*For consistency we will download these from RIC github repository. However, you can run the same exercise with your own DAVID results as well.*

```
# read in the results from DAVID and RNA-seq normalized data
david <- read.table("https://wd.cri.uic.edu/pathway/david_chart.txt",
  header=T, sep="\t", quote="")
# sort by significance
david <- david[order(david$FDR),]
david.kegg <- david[david$Category=="KEGG_PATHWAY" & david$FDR < 0.01,]
# read in normalized expression
norm <- read.table("https://wd.cri.uic.edu/pathway/RNAseq_norm.txt",
  header=T, row.names=1, sep="\t", colClasses=c(Gene.name="NULL"))
```

### 2.3.2 Make heatmap for top KEGG pathway

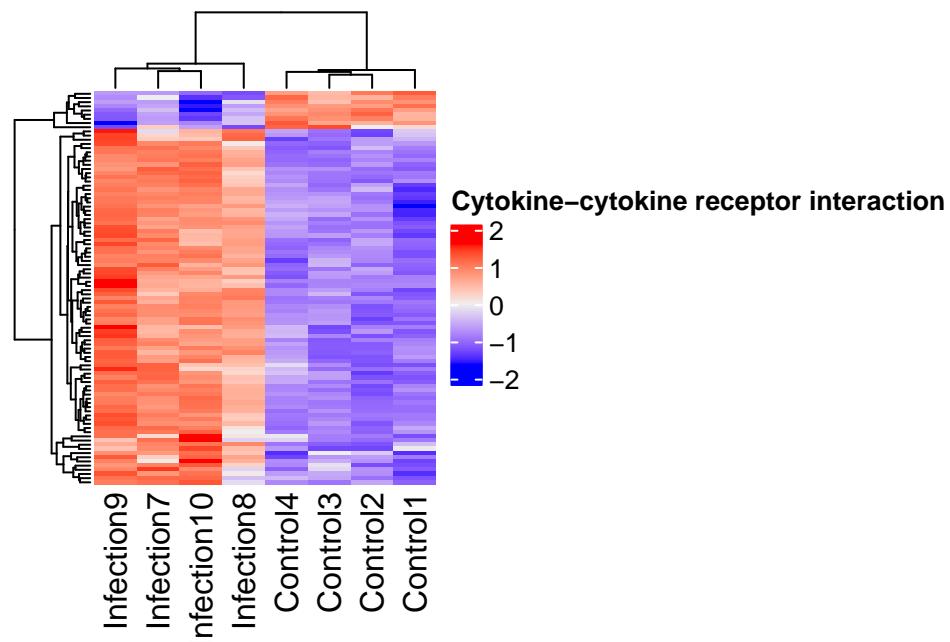
Subset normalized expression to the genes in the top KEGG pathway:

```
# name of the top pathway
top.kegg.name <- david.kegg[1,"Term"]
# remove the KEGG ID
top.kegg.name <- gsub("mmu[0-9]*:", "", top.kegg.name)
# get the gene list for the top pathway
# first remove the white space
top.kegg.genes <- gsub(" ", "", david.kegg[1,"Genes"])
# use string split to split the list into a vector by commas
# use unlist to turn it back into a vector
top.kegg.genes <- unlist(strsplit(top.kegg.genes, ","))
length(top.kegg.genes)
```

```
## [1] 94
# subset the norm table to these genes
top.kegg.norm <- norm[top.kegg.genes,]
dim(top.kegg.norm)
```

```
## [1] 94  8
# log-scale and z-score
top.kegg.norm <- log2(top.kegg.norm + 0.1)
top.kegg.norm <- t(scale(t(top.kegg.norm)))
```

```
# plot in a heatmap
library(ComplexHeatmap)
Heatmap(top.kegg.norm, name = top.kegg.name, show_row_names = FALSE)
```



## 2.4 Pathway analysis in variant calling data

This is a result from variant calling in a whole-exome variant calling data set (one sample, human). We'll do a pathway analysis of mutated genes, using two different filtering strategies:

- **Damaging:** Genes that contain non-synonymous variants with a predicted damaging effect (SIFT score).
- **Mutation burden:** Genes that contain >3 non-synonymous variants.

NOTES:

*SIFT (Sorting Interleaved From Tolerant) is a model for predicting the effects of amino acid substitutions on protein function, based on sequence homology and known physical properties of amino acids. The SIFT model scores amino acid substitutions on a range from 0 to 1, where 0 is biggest effect, 1 is least effect. Damaging effects are predicted for scores < 0.05.*

*The filtering strategies outlined below are only a couple examples of how to filter variants. Other options include combining the strategies (damaging + mutation burden), filtering also based on the genotype call (heterozygous vs homozygous), using a different threshold for mutation burden, including other prediction models of damage to function (polyPhen, phyloP, etc.), and looking at known information about the variants in databases such as dbSNP, 1000 Genomes/ExAC/gnomAD, COSMIC (for cancer mutations), ClinVar, etc.*

### 2.4.1 Filter variants in R

First, read in the variant table in R studio. It is available as a tab-delimited text file.

```
vars <- read.table("https://wd.cri.uic.edu/pathway/example_variants.txt",
  header=T, sep="\t")
# look at vars in the R studio variable browser, or with head
head(vars)

##   CHROM     POS REF ALT Genotype Func Gene      ExonicFunc
## 1 chr1    69270 A  G    1/1 exonic OR4F5 synonymous_SNV
## 2 chr1    69511 A  G    1/1 exonic OR4F5 nonsynonymous_SNV
## 3 chr1   930248 G  A    0/1 exonic SAMD11 nonsynonymous_SNV
## 4 chr1   952421 A  G    1/1 exonic NOC2L synonymous_SNV
## 5 chr1   953259 T  C    1/1 exonic NOC2L synonymous_SNV
## 6 chr1   953279 T  C    1/1 exonic NOC2L nonsynonymous_SNV
##
##                                     AACchange SIFT_score SIFT_pred
## 1 OR4F5:ENST00000335137.3:exon1:c.A180G:p.S60S          .
## 2 OR4F5:ENST00000335137.3:exon1:c.A421G:p.T141A        0.652      T
## 3 SAMD11:ENST00000342066.7:exon3:c.G166A:p.G56S        0.038      D
## 4 NOC2L:ENST00000327044.6:exon10:c.T1182C:p.T394T         .
## 5 NOC2L:ENST00000327044.6:exon9:c.A918G:p.E306E          .
## 6 NOC2L:ENST00000327044.6:exon9:c.A898G:p.I300V        1.0       T

# filtering strategy 1: damaging effects
# variants with SIFT prediction D for damaging, just the Gene column
filter1 <- as.character(vars[vars$SIFT_pred=="D","Gene"])
# some gene annotations have a comma-separated list of genes
# strsplit will split them, and unlist will give it all as a vector
filter1 <- unlist(strsplit(filter1,","))
filter1 <- unique(filter1)
head(filter1)

## [1] "SAMD11"  "PERM1"   "ATAD3B"  "SLC35E2" "CEP104"  "CA6"
length(filter1)
```

```

## [1] 806
# filtering strategy 2: mutation burden
# remove variants with synonymous or unknown change to translated sequence
filter2_all <- as.character(vars[vars$ExonicFunc!="synonymous_SNV" &
  vars$ExonicFunc!="unknown","Gene"] )
# split by commas again
filter2_all <- unlist(strsplit(filter2_all,","))
# generate a count per gene using the table function
filter2_table <- table(filter2_all)
head(filter2_table)

## filter2_all
##      A1BG     A2ML1    A4GALT    A4GNT      AAC5  AADACL2
##      1         2         1         1         1         1

# get the gene names from the table with counts bigger than 3
filter2 <- names(filter2_table)[filter2_table > 3]
head(filter2)

## [1] "ABCA13" "ACAN"    "ADGRF2"  "ADGRV1"  "AHNAK"   "AHNAK2"
length(filter2)

## [1] 245
# check the number of genes in common between the lists
length(intersect(filter1, filter2))

## [1] 122
# write both lists to tables
write.table(filter1,"filter_damaging.txt",col.names=F,row.names=F,quote=F)
write.table(filter2,"filter_mutation_load.txt",col.names=F,row.names=F,quote=F)

```

NOTE: these gene lists are available from our server as well:

- [https://wd.cri.uic.edu/pathway/filter\\_damaging.txt](https://wd.cri.uic.edu/pathway/filter_damaging.txt)
- [https://wd.cri.uic.edu/pathway/filter\\_mutation\\_load.txt](https://wd.cri.uic.edu/pathway/filter_mutation_load.txt)

## 2.4.2 Run pathway enrichment in DAVID for both lists

NOTE: the DAVID portion of this exercise may be left as homework, depending on workshop timing. The visualization parts of the exercise can be completed using DAVID results we have already generated.

Go to DAVID <https://david.ncifcrf.gov>, and follow similar steps as before:

- Set the identifier as **OFFICIAL\_GENE\_SYMBOL**
- Set as Gene List, and submit
  - It may take a little longer with uploading gene lists, as they will match to multiple species. DAVID will warn us about this, and we should select **Homo Sapiens** from the list.
- **NOTE:** Open a second tab in your browser to upload a second gene list at the same time. Once uploaded, both will appear in the *List Manager* on DAVID.
- After the upload, we'll just look at the GOTERM\_BP\_FAT database. Click the "Chart" button next to this database to open the enrichment table.

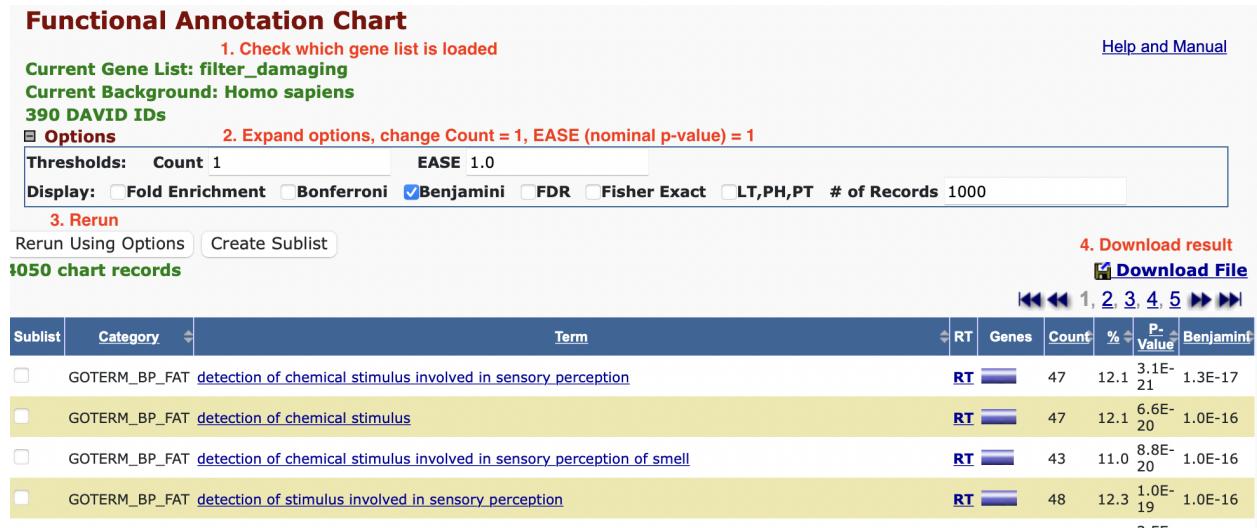
The screenshot shows the DAVID Annotation Summary Results interface. On the left, the Gene List Manager panel displays a list of species: Homo sapiens (390), Ursus americanus (362), and Pan troglodytes (317). A red arrow points to the 'Select Species' dropdown with the text 'Select "Homo sapiens" and click "Select Species"'.

The main panel shows the Annotation Summary Results for the current gene list 'filter\_damaging' against the background 'Homo sapiens'. It lists 390 DAVID IDs. The GOTERM\_BP\_FAT database is selected, indicated by a checked checkbox. A red arrow points to the 'Chart' button next to GOTERM\_BP\_FAT with the text 'Click "Chart" button next to GOTERM\_BP\_FAT'.

Annotation	Percentage	Count	Action
GOTERM_BP_1	86.2%	336	Chart
GOTERM_BP_2	86.2%	336	Chart
GOTERM_BP_3	85.9%	335	Chart
GOTERM_BP_4	85.4%	333	Chart
GOTERM_BP_5	83.6%	326	Chart
GOTERM_BP_ALL	86.2%	336	Chart
<b>GOTERM_BP_DIRECT</b>	<b>86.2%</b>	<b>336</b>	<b>Chart</b>
<b>GOTERM_BP_FAT</b>	<b>85.9%</b>	<b>335</b>	<b>Chart</b>
GOTERM_CC_1	92.3%	360	Chart
GOTERM_CC_2	90.8%	354	Chart
GOTERM_CC_3	90.8%	354	Chart
GOTERM_CC_4	89.5%	349	Chart
GOTERM_CC_5	73.8%	288	Chart

In the chart view, note which dataset we're looking at. Also modify the filters:

- Change the filtering options to set count=1, EASE=1
  - This will maximize the number of terms in the list, so we have less missing data in the intersection later
  - We will filter for significance later in R
- Right-click to download the chart report
- Repeat with the other gene list
  - Confirm that the other list is active when you open the chart view for it
- Save the files as:
  - **GOBP\_damaging.txt** and **GOBP\_mutation\_load.txt**



### 2.4.3 Comparison visualization in R

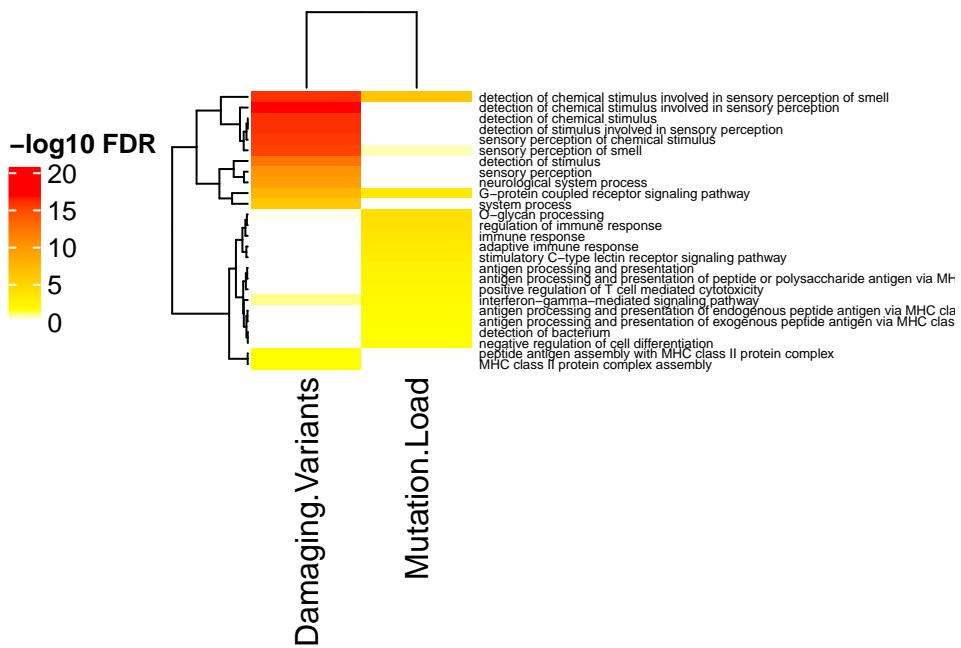
Prepare a heatmap comparison of the two results in R Studio. *For consistency, we will use the results stored on our server, but you can read your own results into R also.*

```
# read in tables; remember quote="" helps with parsing
# pathway names with quotes in their name
gobp.damaging <- read.table(
  "https://wd.cri.uic.edu/pathway/GOBP_damaging.txt",
  header=T, sep="\t", quote="")
gobp.mutation_load <- read.table(
  "https://wd.cri.uic.edu/pathway/GOBP_mutation_load.txt",
  header=T, sep="\t", quote="")
# process as we did before:
# subset the columns
sub.damaging <- gobp.damaging[,c("Term", "FDR")]
sub.mutation_load <- gobp.mutation_load[,c("Term", "FDR")]
# name the FDR column based on the gene list
colnames(sub.damaging)[2] <- "Damaging.Variants"
colnames(sub.mutation_load)[2] <- "Mutation.Load"
# merge and replace missing values with FDR = 1
gobp.merged <- merge(x=sub.damaging, y=sub.mutation_load, by="Term", all=T)
gobp.merged[is.na(gobp.merged)] <- 1
# prepare as a data frame with term IDs in the rownames
gobp.df <- data.frame(gobp.merged[,c(2:ncol(gobp.merged))])
rownames(gobp.df) <- gobp.merged[,1]
```

```

# subset to the significant terms and log-scale FDRs
gobp.subset <- gobp.df[apply(gobp.df, 1, min) < 0.05, ]
gobp.subset <- -log10(gobp.subset)
# remove the GO IDs from the term descriptions
rownames(gobp.subset) <- gsub("GO:[0-9]*~","",rownames(gobp.subset))
# plot in a heatmap, setting a color scale first
library(circlize)
col_fun <- colorRamp2(c(0, -log10(0.05), max(unlist(gobp.subset))), 
  c("white","yellow","red"))
library(ComplexHeatmap)
heatmap <- Heatmap(as.matrix(gobp.subset), name="-log10 FDR", col=col_fun,
  row_names_gp = gpar(fontsize = 5))
draw(heatmap, heatmap_legend_side = "left")

```



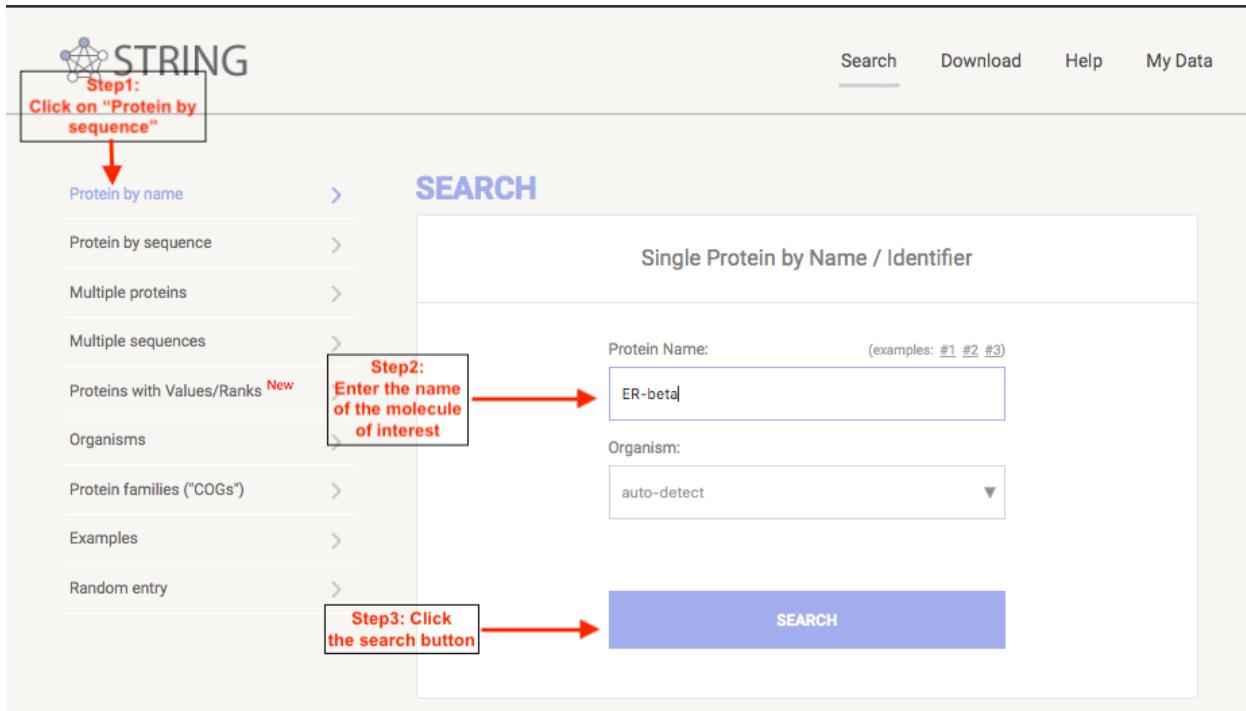
## 2.5 Using STRING

### 2.5.1 Navigate to STRING

Navigate to STRING <https://string-db.org/> Click on Search Button on Right hand side.

### 2.5.2 Search by molecule of interest

- Step 1: Click search button top right side of the menu.
- Step 2: Select “Protein by name” in the left side menu
- Step 3: Type “ER-beta” in the “Protein Name” search box
- Step 4: Click “SEARCH” button



- **Step 1:** Select the species of interest. For this exercise we will keep default selected choice “Homo sapiens”
- **Step 2:** Click “continue” button

 **STRING**

**Step1: Select species of interest.  
For this exercise keep default "Homo sapiens"**

Search Download Help My Data

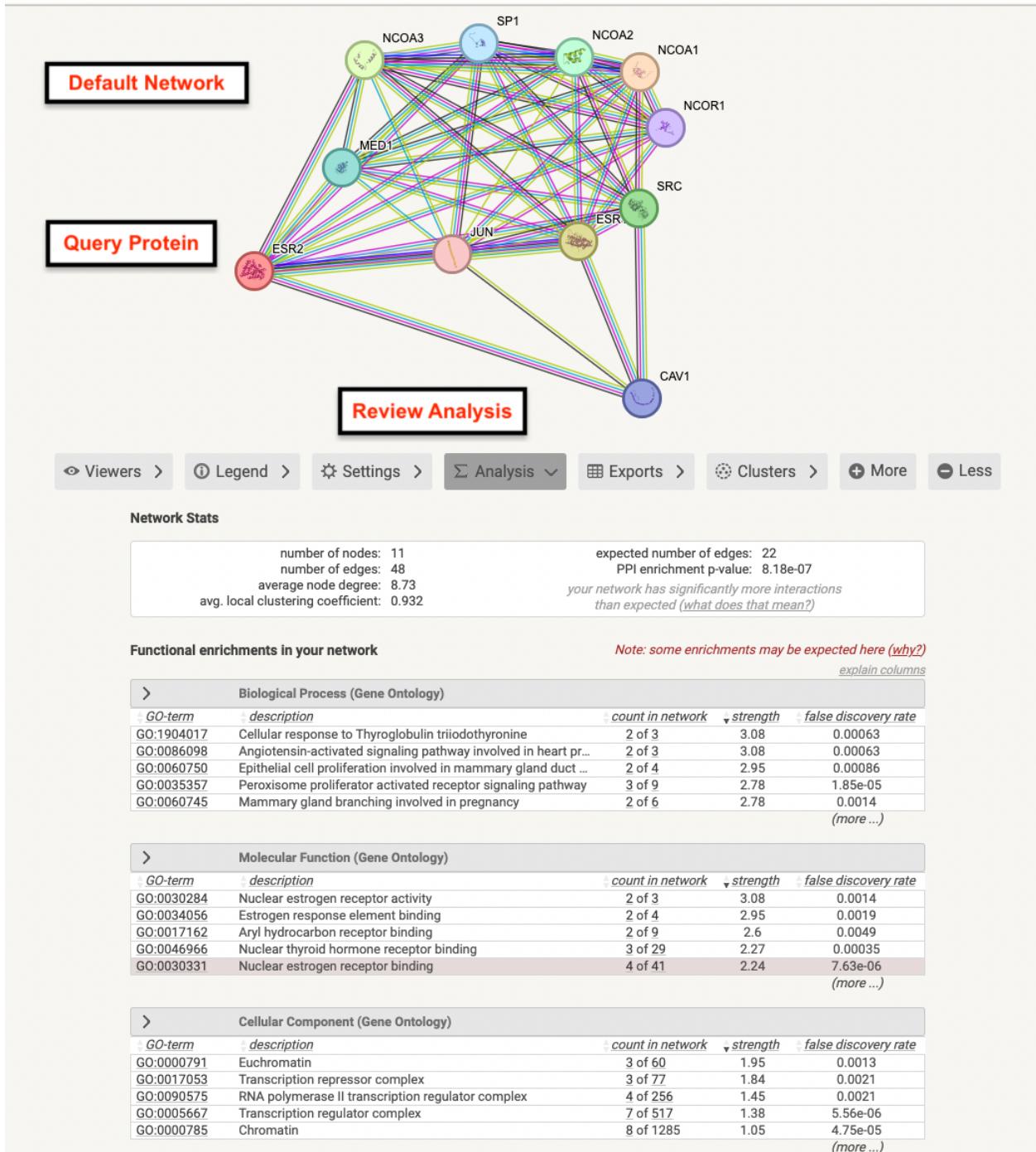
**Step2: Click "CONTINUE"**

There are several matches for 'ER-beta'. Please select one from the list below and press Continue to proceed.

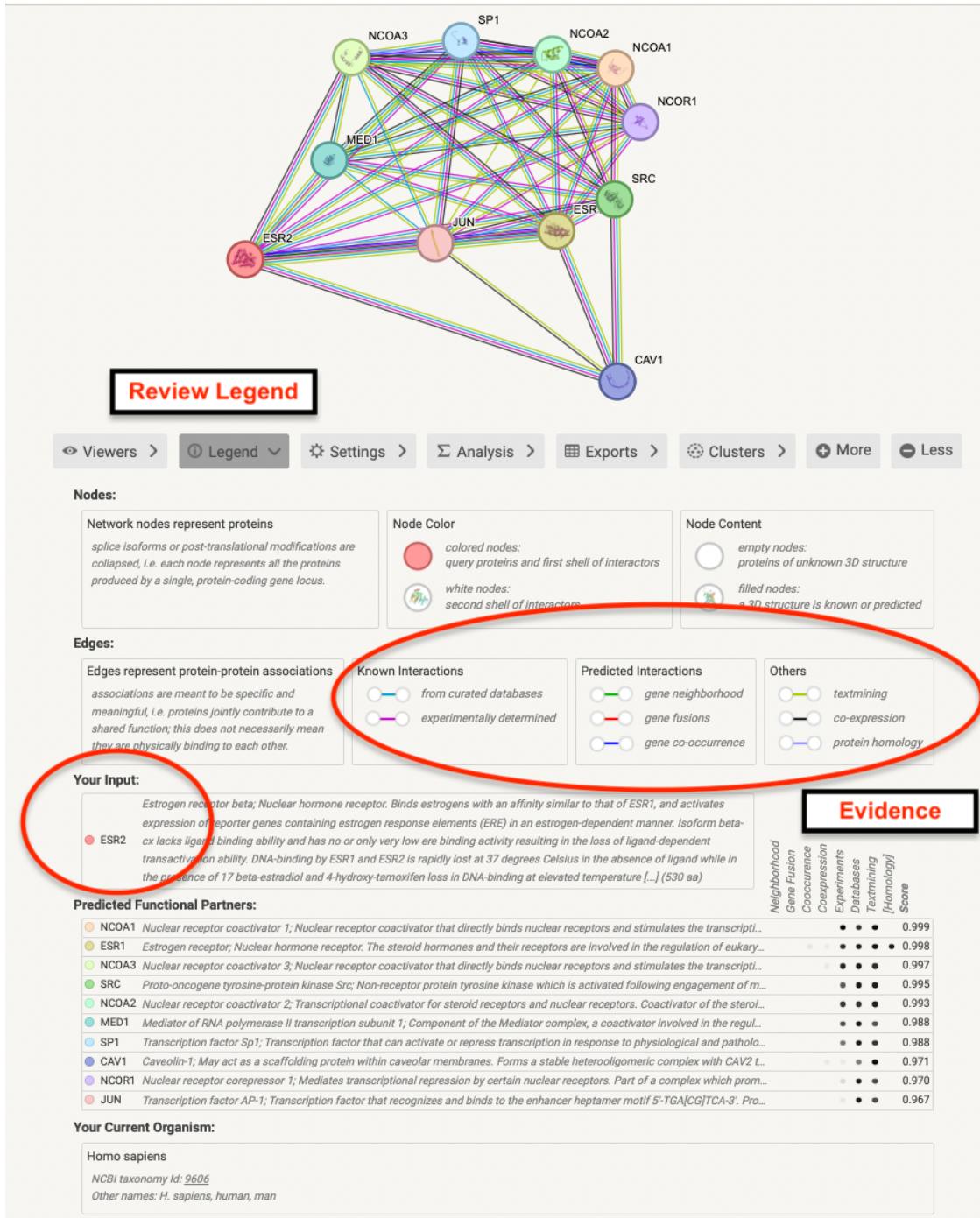
<- BACK CONTINUE ->

organism	protein
<input checked="" type="checkbox"/> <b>Homo sapiens</b>	ESR2 - Estrogen receptor beta; Nuclear hormone receptor. Binds estrogens with an affinity similar to that of ESR1, and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. Isoform beta-cx lacks ligand binding ability and has no or only very low ere binding activity resulting in the loss of ligand-dependent transactivation capability. DNA-binding by ESR1 and ESR2 is rapidly lost at 37 degrees Celsius in the absence of ligand while in the presence of 17 beta-estradiol and 4-hydroxy-tamoxifen loss in DNA-binding at elevated temperature [...] [a.k.a. <i>ESTRB</i> , <i>NR3A2</i> , <i>1QKM</i> , <i>ER-beta</i> , <i>ER-BETA</i> , ...]
<input type="checkbox"/> <i>Callithrix jacchus</i>	ESR2 - Estrogen receptor beta; Nuclear hormone receptor. Binds estrogens with an affinity similar to that of ESR1 (ER-alpha), and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. May play a role in ovarian follicular growth and maturation [a.k.a. <i>NR3A2</i> , <i>ENSCJAP00000036245</i> , <i>ACFV01048515</i> , <i>ER-beta</i> ]
<input type="checkbox"/> <i>Oreochromis niloticus</i>	<i>nr3a2</i> - Estrogen receptor beta; Binds estrogens with an affinity similar to that of ER- alpha, and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner; Belongs to the nuclear hormone receptor family. NR3 subfamily [a.k.a. <i>LOC100534515</i> , <i>esr2</i> , <i>XP_005475012.1</i> , <i>ER-beta</i> ]
<input type="checkbox"/> <i>Rattus norvegicus</i>	<i>Esr2</i> - Estrogen receptor beta; Binds estrogens with an affinity similar to that of ER- alpha, and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. Isoform 3 and isoform 4 are unable to bind DNA and activate transcription due to the truncation of the DNA binding domain. Isoform 2 shows loss of ligand binding affinity and suppresses ER- alpha and ER-beta1 mediated transcriptional activation and may act as a dominant negative regulator of estrogen action; Belongs to the nuclear hormone receptor family. NR3 subfamily [a.k.a. <i>Nr3a2</i> , <i>Erbeta</i> , <i>1HJ1</i> , <i>ER-beta</i> ]
<input type="checkbox"/> <i>Sus scrofa</i>	ESR2 - Estrogen receptor beta; Nuclear hormone receptor. Binds estrogens with an affinity similar to that of ESR1 (ER-alpha), and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. May play a role in ovarian follicular growth and maturation [a.k.a. <i>NR3A2</i> , <i>NP_001001533.1</i> , <i>ESR2-001</i> , <i>ER-beta</i> ]

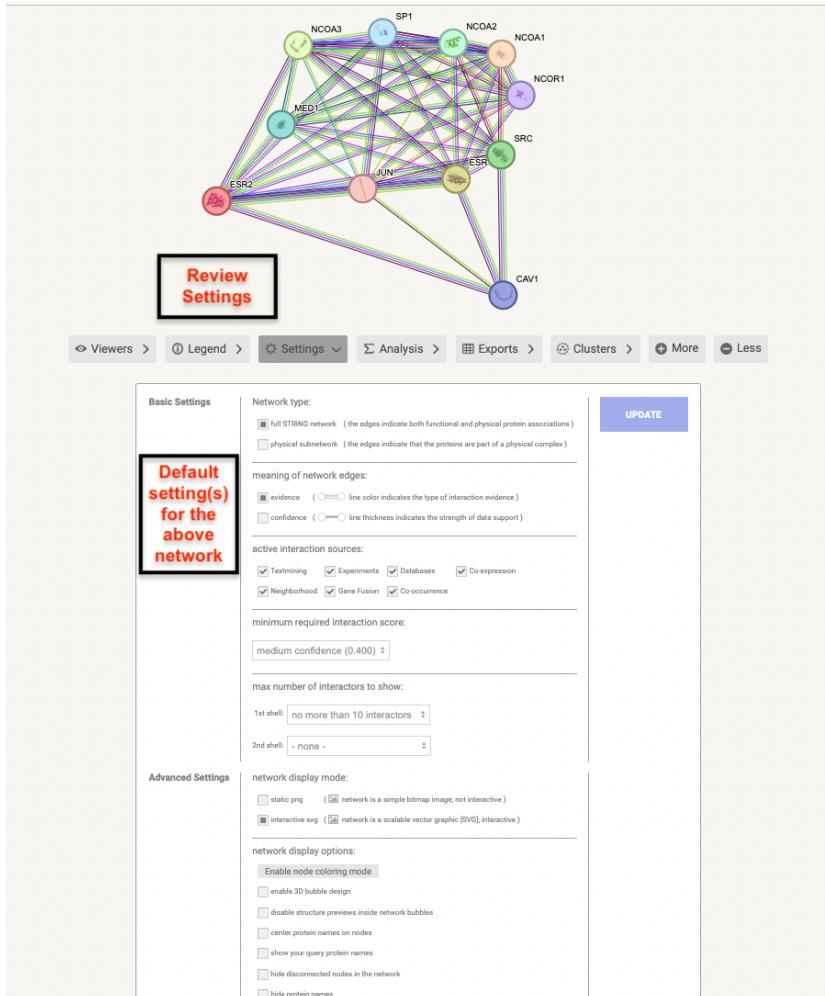
- Review network and analysis details
  - Review Network Stats and Functional enrichment for the network.
  - Click on node/edge to get more details for a given interaction.



- Click on “Legend”
  - Review edges colors for interaction type, e.g. pink = Experimental Evidence
  - Click on node/edge to get more details for a given interaction.

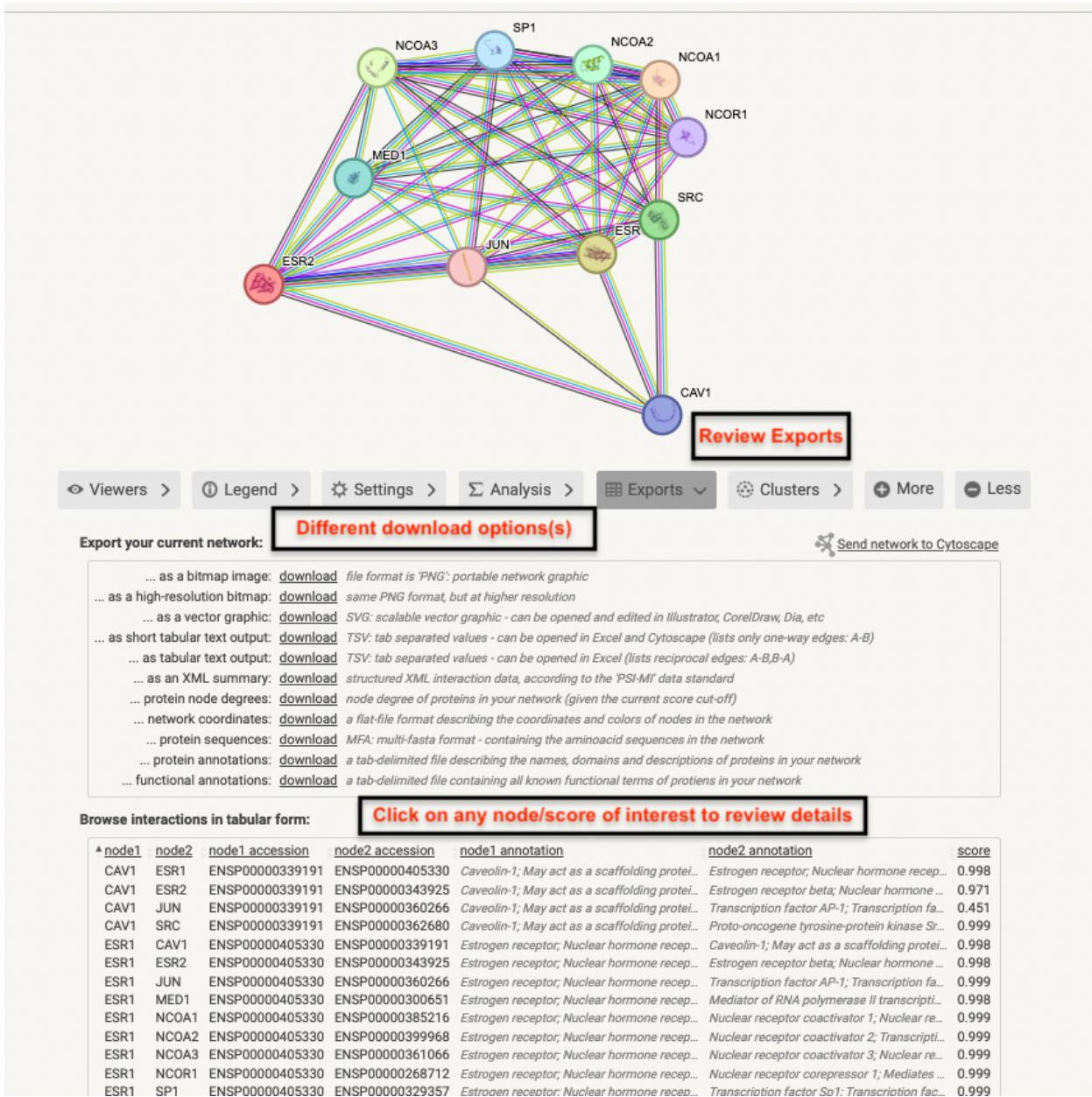


- Changing the view settings:
  - For default network settings as shown in the above/below network. Please choose following options if your network looks different than the figure.
    - \* Step 1: Under “Active interaction sources:” Select all evidence
    - \* Step 2: Under “minimum required interaction score:” select “Medium confidence (0.400)”
    - \* Step 3: Under “max number of interactors to show:” select “1st shell - no more than 10 interactions”
    - \* Step 4: Click “UPDATE” to update the network
  - Explore different settings add/remove extra nodes/edges (optional exercise)
    - \* Step 1: Under “Active interaction sources:” Select just “Experiments” as evidence
    - \* Step 2: Under “minimum required interaction score” select “Low confidence (0.150)”
    - \* Step 3: Under “max number of interactors to show:” 1st shell - no more than 10 interactions
      - Note: The 1st shell interactors are the proteins directly associated with input protein(s)/query. 2nd shell of interactors are the proteins associated with the proteins from the 1st shell or with input protein(s)/query.
    - \* Step 4: Click “UPDATE” to update the network



- **Export results:** Click on Export tab

- Check all download options
- Check the interaction table.



### 2.5.3 Search by multiple proteins

Download antiviral gene list.

[https://wd.cri.uic.edu/pathway/antiviral\\_list.txt](https://wd.cri.uic.edu/pathway/antiviral_list.txt)

- Step 1: Select “Multiple Protein” in the left side menu
- Step 2: Upload the “antiviral\_list.txt” by clicking “Browse”
- Step 3: Organism: auto-detect
- Step 4: Advanced Settings: Default
- Step 5: Click “SEARCH” button

NOTE: If STRING asks you to select an organism, type in “Mus musculus”.

The screenshot shows the STRING search interface. On the left, there's a sidebar with various search options: Protein by name, Protein by sequence, Multiple proteins (which is selected and highlighted in blue), Multiple sequences, Proteins with Values/Ranks, Organisms, Protein families ("COGs"), Examples, and Random entry. The main area is titled "SEARCH" and contains a form for "Multiple Proteins by Names / Identifiers". It has fields for "List Of Names:" (with a note "(one per line; examples: #1 #2 #3)" and a large text input area), "Organism:" (set to "auto-detect"), and "Advanced Settings" (with options for Network Type, Required score, and FDR stringency). A red circle highlights the "Browse ..." button next to the file input field, which contains the text "antiviral\_list.txt". At the bottom is a large blue "SEARCH" button.

- **Step 1:** Select proteins of interest. For this exercise we will keep all of the search results.
- **Step 2:** Click “CONTINUE” button.

The following proteins in *Mus musculus* appear to match your input.  
Please review the list, then click ‘Continue’ to proceed.

Search Download Help My Data

[← BACK](#) [↓ MAPPING](#) [CONTINUE →](#)

44 query items showing page 1 of 3 • first • previous • next • last

1) 'ENSMUSG00000026104':

Stat1 - Signal transducer and transcription activator that mediates cellular responses to interferons (IFNs), cytokine KITLG/SCF and other cytokines and other growth factors. Following type I IFN (IFN-alpha and IFN-beta) binding to cell surface receptors, signaling via protein kinases leads to activation of Jak kinases (TYK2 and JAK1) and to tyrosine phosphorylation of STAT1 and STAT2. The phosphorylated STATs dimerize and associate with ISGF3G/IRF-9 to form a complex termed ISGF3 transcription factor, that enters the nucleus. ISGF3 binds to the IFN stimulated response element (ISRE) to activate [...] [a.k.a. *Stat1*-015, OTTMUST00000121565, OTTMUST00000121563, [ENSMUSG00000026104](#)]

2) 'ENSMUSG00000042349':

Ikbke - Inhibitor of nuclear factor kappa-B kinase subunit epsilon; Serine/threonine kinase that plays an essential role in regulating inflammatory responses to viral infection, through the activation of the type I IFN, NF-kappa-B and STAT signaling. Also involved in TNFA and inflammatory cytokines, like Interleukin-1, signaling. Following activation of viral RNA sensors, such as RIG-I-like receptors, associates with DDX3X and phosphorylates interferon regulatory factors (IRFs), IRF3 and IRF7, as well as DDX3X. This activity allows subsequent homodimerization and nuclear translocation of the [...] [a.k.a. *Ikki*, *Ikke*, OTTMUST00000161764, [ENSMUSG00000042349](#)]

3) 'ENSMUSG00000026896':

Ifih1 - Interferon-induced helicase C domain-containing protein 1; Innate immune receptor which acts as a cytoplasmic sensor of viral nucleic acids and plays a major role in sensing viral infection and in the activation of a cascade of antiviral responses including the induction of type I interferons and proinflammatory cytokines. Its ligands include mRNA lacking 2'-O- methylation at their 5' cap and long-dsRNA (>1 kb in length). Upon ligand binding it associates with mitochondria antiviral signaling protein (MAVS/IPS1) which activates the IKK-related kinases: TBK1 and IKBKE which phosphorylate [...] [a.k.a. UPI00000E79DD, OTTMUST00000030862, uc008jvm.2, [ENSMUSG00000026896](#)]

4) 'ENSMUSG00000027639':

Samhd1 - Deoxynucleoside triphosphate triphosphohydrolase SAMHD1; Host restriction nuclease involved in defense response to virus. Has dNTPase activity and reduces cellular dNTP levels to levels too low for retroviral reverse transcription to occur. Blocks early-stage virus replication in dendritic and other myeloid cells. Likewise, suppresses LINE-1 retrotransposon activity. May play a role in mediating proinflammatory responses to TNF-alpha signaling. Has ribonuclease activity, acting on single-stranded RNA; Belongs to the SAMHD1 family [a.k.a. *Mg11*, *Mm.468781*, *AAH67198.1*, [ENSMUSG00000027639](#)]

5) 'ENSMUSG00000028037':

Ifi44 - Interferon-induced protein 44; This protein aggregates to form microtubular structures [a.k.a. *Mtpap44*, *BC026901*, *Q8BV66*, [ENSMUSG00000028037](#)]

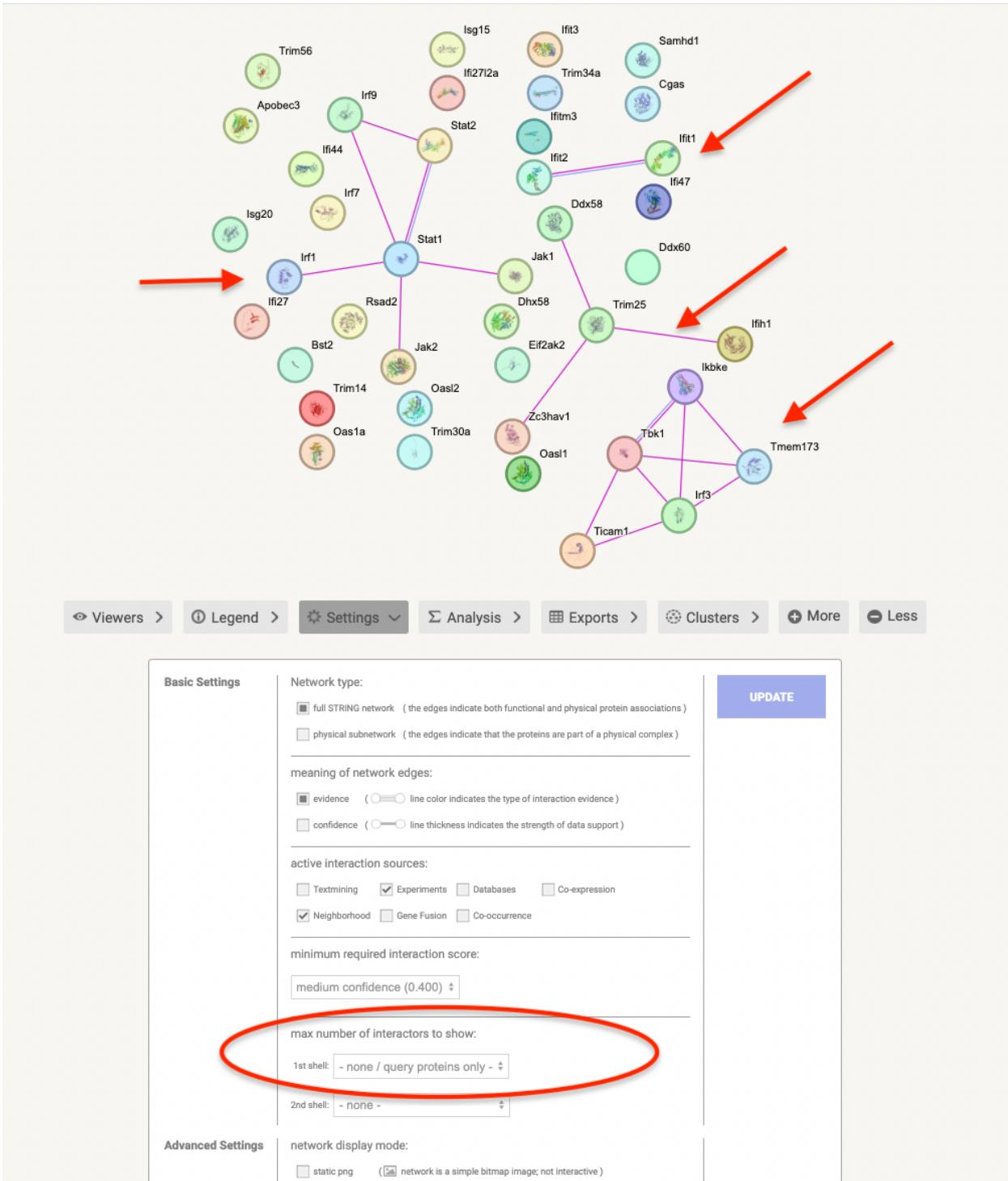
6) 'ENSMUSG00000040296':

Ddx58 - Probable ATP-dependent RNA helicase DDX58; Innate immune receptor which acts as a cytoplasmic sensor of viral nucleic acids and plays a major role in sensing viral infection and in the activation of a cascade of antiviral responses including the induction of type I interferons and proinflammatory cytokines. Its ligands include: 5'- triphosphorylated ssRNA and dsRNA and short dsRNA (<1 kb in length). In addition to the 5'-triphosphate moiety, blunt-end base pairing at the 5'-end of the RNA is very essential. Overhangs at the non-triphosphorylated end of the dsRNA RNA have no major impact [...] [a.k.a. *Ddx58-005*, *DDX58\_MOUSE*, *Ddx58-009*, [ENSMUSG00000040296](#)]

7) 'ENSMUSG00000039853':

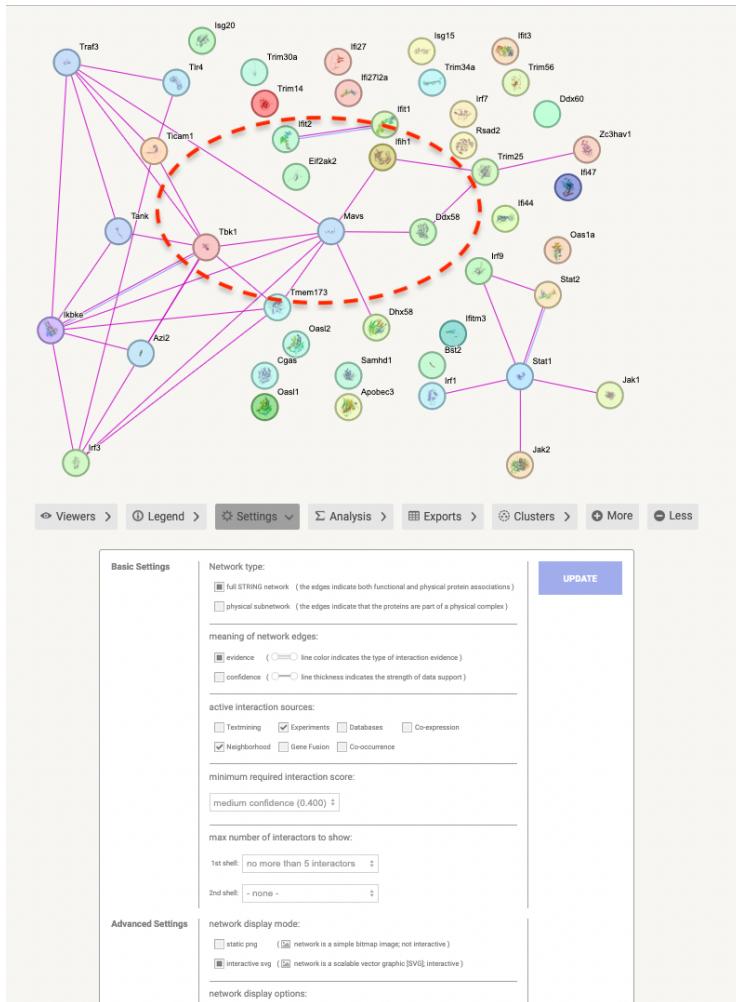
Trim14 - Tripartite motif-containing protein 14; Plays a role in the innate immune defense against viruses. Facilitates the type I IFN response by interacting with MAVS at the outer mitochondria membrane and thereby recruiting NF- kappa-B essential modulator IKBKG/NEMO to the MAVS signalosome, leading to the activation of both the IFN regulatory factor 3/IRF3 and NF-kappa-B pathways. Positively regulates the cGAS-induced type I interferon signaling pathway by stabilizing cGAS/MB21D1 and inhibiting its autophagic degradation. Inhibits the transcriptional activity of SPI1 in a dose-dependent manner [...] [a.k.a. *Kiaa0129*, *Pub*, *Q9D3G8*, [ENSMUSG00000039853](#)]

- Adjustments to view settings (**The network is shown using these settings**):
  - Please choose following options if your network looks different than the figure.
    - \* Step 1: Under “Active interaction sources:” Select “Experiments” and “Neighborhood””
    - \* Step 2: Under “minimum required interaction score:” select “Medium confidence (0.400)”
    - \* Step 3: Under “max number of interactors to show:” select “- none/query proteins only -”
      - Note: The 1st shell interactors are the proteins directly associated with input protein(s)/query. 2nd shell of interactors are the proteins associated with the proteins from the 1st shell or with input protein(s)/query.
    - \* Step 4: Click “UPDATE” to update the network
  - Observations.
    - \* There are small sub-connected networks and other are unconnected nodes (proteins)
    - \* Click on node/edge to get more details for a given interaction
    - \* Move around node(s) to get clear view



- Additional information: Click on “Analysis” tab
  - Click on node/edge in the network to get more details for a given interaction
  - Observations: Network stats
    - \* There are 41 nodes and 18 edges.

- Click on “Settings” again to update the network
  - Update the network by changing following settings
    - \* Step 1: Under “Active interaction sources:” Select “Experiments” and “Neighborhood”
    - \* Step 2: Under “minimum required interaction score:” select “Medium confidence (0.400)”
    - \* Step 3: Under “max number of interactors to show:” select “no more than 5 interactions”
      - Note: The 1st shell interactors are the proteins directly associated with input protein(s)/query. 2nd shell of interactors are the proteins associated with the proteins from the 1st shell or with input protein(s)/query.
    - \* Step 4: Click “UPDATE” to update the network
    - \* Move around nodes to see interactions. Move around single nodes
  - Observations
    - \* Tbk1: we see more nodes added around Tbk1



- Stats: click on “Analysis” tab
  - Observations: Network stats
    - \* There are 47 nodes and 40 edges. Earlier we had 42 nodes and 26 edges.

## 2.6 Using STITCH

### 2.6.1 Navigate to STITCH

Navigate to STITCH <http://stitch.embl.de/>

### 2.6.2 Search by molecule name

- Estradiol (endogenous small molecule)
  - Step 1: Select “Item by name” in the left side menu
  - Step 2: Type “Estradiol” in the “Item Name” search box
  - Step 2A: Leave default “Organism: auto-detect”
  - Step 3: Click “SEARCH” button

The screenshot shows the STITCH search interface. At the top, there is a navigation bar with "Version: 5.0", "LOGIN | REGISTER", and links for "Search", "Download", "Help", and "My Data". On the left, a sidebar lists search options: "Multiple names", "Chemical structure(s)", "Protein sequence(s)", "Examples", and "Random entry". A red box labeled "Step1: Click on 'Item by name'" highlights the "Item by name" link. The main search area is titled "SEARCH" and contains a sub-section for "Single Item by Name / Identifier" under "Small Molecule". It has fields for "Item Name:" (with "Estradiol" typed in) and "Organism:" (set to "auto-detect"). A red box labeled "Step2: Enter the name of the protein/small molecule/drug" highlights the "Item Name:" field. A red arrow points from this field to a red box labeled "Step3: Click 'SEARCH'". A final red arrow points from the "SEARCH" button to the right.

- For this exercise we will keep default selected choice “estradiol” in the *chemical* list.
- Click **CONTINUE** button

Version: 5.0      LOGIN | REGISTER

# STITCH

**Step1: Keep default selection**

Search   Download   Help   My Data

**Step2: Click “CONTINUE”**

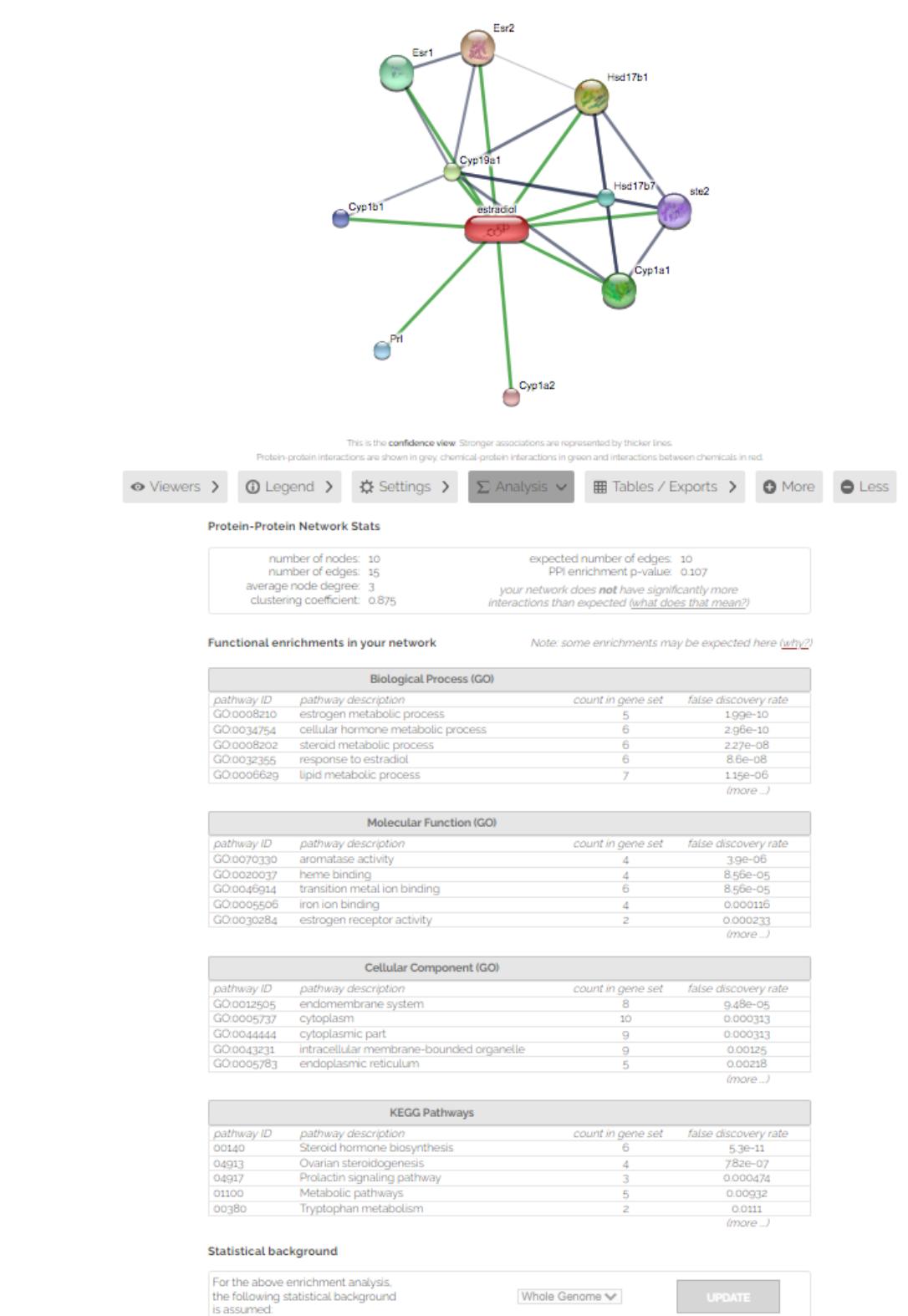
There are several matches for Estradiol! Please select one from the list below and press Continue to proceed.

**chemical**  estradiol

organism	protein
<input type="checkbox"/> Homo sapiens	SULT1E1 - sulfotransferase family 1E, estrogen-preferring, member 1: Sulfotransferase that utilizes 3'-phospho-5'-adenylyl sulfate (PAPS) as sulfonate donor to catalyze the sulfate conjugation of estradiol and estrone. May play a role in the regulation of estrogen receptor activity by metabolizing free estradiol. Maximally sulfates beta-estradiol and estrone at concentrations of 20 nM. Also sulfates dehydroepiandrosterone, pregnenolone, ethynodiol, equilenin, diethylstilbestrol and 1-naphthol, at significantly higher concentrations; however, cortisol, testosterone and dopamine are not sulfated
<input type="checkbox"/> Homo sapiens	CUEDC2 - CUE domain containing 2; Down-regulates ESR1 protein levels through the ubiquitination-proteasome pathway, regardless of the presence of 17 beta-estradiol. Also involved in 17 beta-estradiol-induced ESR1 degradation. Controls PGR protein levels through a similar mechanism
<input type="checkbox"/> Homo sapiens	ATAD2 - ATPase family, AAA domain containing 2; May be a transcriptional coactivator of the nuclear receptor ESR1 required to induce the expression of a subset of estradiol target genes, such as CCND1, MYC and E2F1. May play a role in the recruitment or occupancy of CREBPP at some ESR1 target gene promoters. May be required for histone hyperacetylation. Involved in the estrogen-induced cell proliferation and cell cycle progression of breast cancer cells
<input type="checkbox"/> Homo sapiens	SLCO1B3 - solute carrier organic anion transporter family, member 1B3; Mediates the Na <sup>(+)</sup> -independent uptake of organic anions such as 17-beta-glucuronosyl estradiol, taurocholate, triiodothyronine (T <sub>3</sub> ), leukotriene C4, dehydroepiandrosterone sulfate (DHEAS), methotrexate and sulfobromophthalein (BSP). Involved in the clearance of bile acids and organic anions from the liver
<input type="checkbox"/> Homo sapiens	SLC17A3 - solute carrier family 17 (sodium phosphate), member 3; Isoform 2: voltage-driven, multispecific, organic anion transporter able to transport para-aminohippurate (PAH), estrone sulfate, estradiol-17-beta-glucuronide, bumetanide, and ochratoxin A. Isoform 2 functions as urate efflux transporter on the apical side of renal proximal tubule and is likely to act as an exit path for organic anionic drugs as well as urate in vivo. May be involved in actively transporting phosphate into cells via Na <sup>(+)</sup> cotransport
<input type="checkbox"/> Homo sapiens	HSD17B2 - hydroxysteroid (17-beta) dehydrogenase 2; Capable of catalyzing the interconversion of testosterone and androstenedione, as well as estradiol and estrone. Also has 20-alpha-HSD activity. Uses NADH while EDH17B3 uses NADPH
<input type="checkbox"/> Homo sapiens	HSD17B12 - hydroxysteroid (17-beta) dehydrogenase 12; Catalyzes the transformation of estrone (E1) into estradiol (E2), suggesting a central role in estrogen formation. Its strong expression in ovary and mammary gland suggest that it may constitute the major enzyme responsible for the conversion of E1 to E2 in women. Also has 3-ketoacyl-CoA reductase activity, reducing both long chain 3-ketoacyl-CoAs and long chain fatty acyl-CoAs, suggesting a role in long fatty acid elongation
<input type="checkbox"/> Homo sapiens	SLCO1B1 - solute carrier organic anion transporter family, member 1B1; Mediates the Na <sup>(+)</sup> -independent uptake of organic anions such as pravastatin, taurocholate, methotrexate, dehydroepiandrosterone sulfate, 17-beta-glucuronosyl estradiol, estrone sulfate, prostaglandin E2, thromboxane B <sub>2</sub> , leukotriene C <sub>3</sub> , leukotriene E <sub>4</sub> , thyroxine and triiodothyronine. Involved in the clearance of bile acids and organic anions from the liver
<input type="checkbox"/> Homo sapiens	HSD17B6 - hydroxysteroid (17-beta) dehydrogenase 6 homolog (mouse); NAD-dependent oxidoreductase with broad substrate specificity that shows both oxidative and reductive activity ( <i>in vitro</i> ). Has 17-beta-hydroxysteroid dehydrogenase activity towards various steroids ( <i>in vitro</i> ). Converts 5-alpha-androstan-3-

**<- BACK**   **CONTINUE ->**

- Click on “Analysis”
  - Click on node/edge in the network to get more details for a given interaction.
  - Observations: Protein-Protein Network stats, Functional enrichment, etc.



### 2.6.3 Search by protein name

- ER-beta (protein)
  - Select “Item by name” in the left side menu
  - Type “ER-beta” in the “Item Name” search box (leave default “Organism: auto-detect”)
  - Click “SEARCH” button

The screenshot shows the STITCH search interface. At the top, there is a navigation bar with "Version: 5.0", "LOGIN | REGISTER", "Search", "Download", "Help", and "My Data". On the left, a sidebar lists search options: "Multiple names", "Chemical structure(s)", "Protein sequence(s)", "Examples", and "Random entry". A red box labeled "Step1: Click on 'item by name'" has an arrow pointing to the "Item by name" link. In the center, a main search form is titled "SEARCH". It has two sections: "Single Item by Name / Identifier" and "Multiple Items by Name / Identifier". The "Single Item" section contains fields for "Item Name" (set to "Protein" with examples "#1 #2 #3") and "Organism" (set to "auto-detect"). A red box labeled "Step2: Enter the name of the protein/small molecule/drug" has an arrow pointing to the "Item Name" input field, which contains "ER-beta". Another red box labeled "Step3: Click 'SEARCH'" has an arrow pointing to the blue "SEARCH" button.

© STITCH CONSORTIUM 2016



EMBL - European Molecular Biology Laboratory



SIB - Swiss Institute of Bioinformatics



CPR - NNF Center for Protein Research

ABOUT

Content

References

Contributors

INFO

Scores

Use scenarios

FAQs

ACCESS

Versions

APIs

Licensing

CREDITS

Funding

Datasources

Partners

Software

- Keep default species as “Homo sapiens”.
- Click “CONTINUE” button

Version: 5.0 [LOGIN](#) | [REGISTER](#)

# STITCH

**Step1: Select species of interest.  
For this exercise keep default "Homo sapiens"**

Search Download Help My Data **Step2: Click "CONTINUE"**

There are several matches for 'ER-beta'. Please select one from the list below and press Continue to proceed.

**<- BACK** **CONTINUE ->**

organism	protein
<input checked="" type="checkbox"/> Homo sapiens	ESR2 - estrogen receptor 2 (ER beta)
<input type="checkbox"/> Callithrix jacchus	ESR2 - estrogen receptor beta ; Nuclear hormone receptor. Binds estrogens with an affinity similar to that of ESR1 (ER-alpha), and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. May play a role in ovarian follicular growth and maturation
<input type="checkbox"/> Oreochromis niloticus	esr2 - Estrogen receptor beta : Binds estrogens with an affinity similar to that of ER- alpha, and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner
<input type="checkbox"/> Rattus norvegicus	Esr2 - estrogen receptor beta ; Binds estrogens with an affinity similar to that of ER- alpha, and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. Isoform 3 and isoform 4 are unable to bind DNA and activate transcription due to the truncation of the DNA binding domain. Isoform 2 shows loss of ligand binding affinity and suppresses ER- alpha and ER-beta1 mediated transcriptional activation and may act as a dominant negative regulator of estrogen action
<input type="checkbox"/> Sus scrofa	Esr2 - estrogen receptor beta ; Nuclear hormone receptor. Binds estrogens with an affinity similar to that of ESR1 (ER-alpha), and activates expression of reporter genes containing estrogen response elements (ERE) in an estrogen-dependent manner. May play a role in ovarian follicular growth and maturation
<input type="checkbox"/> Xenopus laevis	esr2 - estrogen receptor 2 (ER beta)

© STITCH CONSORTIUM 2016



EMBL - European Molecular Biology Laboratory



SIB - Swiss Institute of Bioinformatics



CPR - NNF Center for Protein Research

## ABOUT

Content

References

Contributors

## INFO

Scores

Use scenarios

FAQs

## ACCESS

Versions

APIs

Licensing

## CREDITS

Funding

Datasources

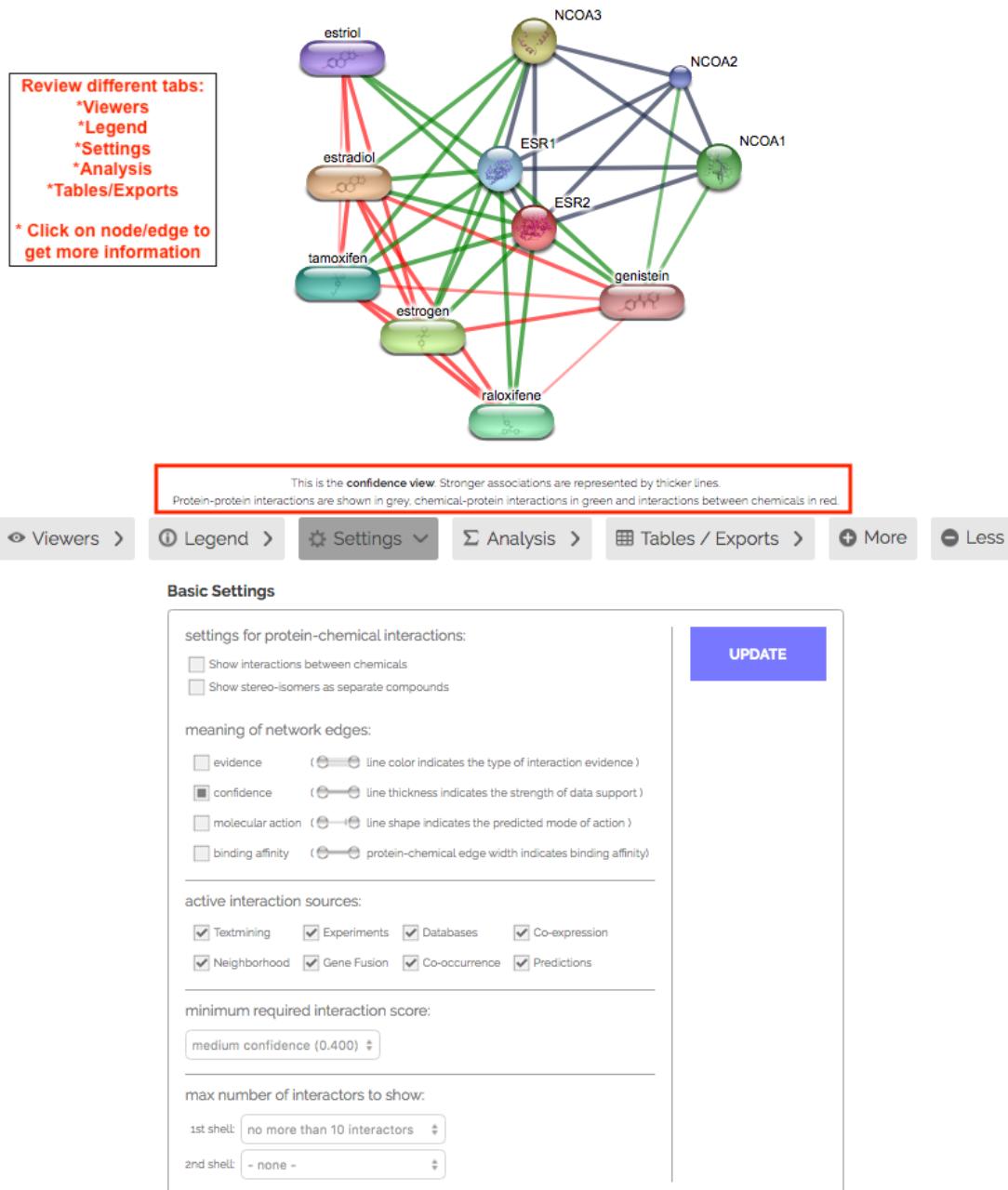
Partners

Software

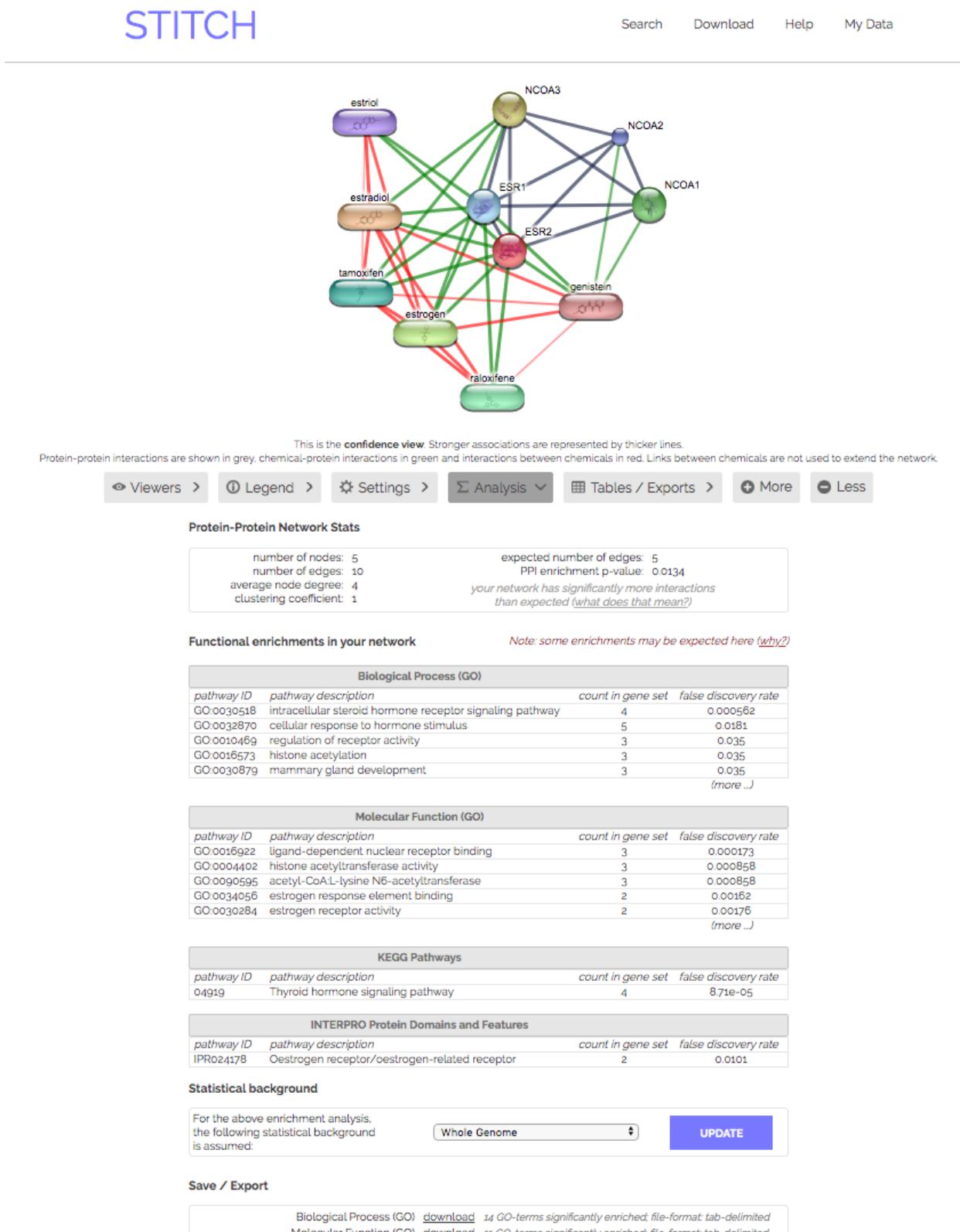
- Click on “Settings”
  - Review different settings for protein-chemical interaction

# STITCH

Search Download Help My Data



- Click on “Analysis”
  - Click on node/edge in the network to get more details for a given interaction.
  - Observations: Protein-Protein Network stats, Functional enrichment, etc.



## 2.6.4 Search by drug name

- Tamoxifen (drug)
  - Select “Item by name” in the left side menu
  - Type “Tamoxifen” in the “Item Name” search box (leave default “Organism: auto-detect”)
  - Click “SEARCH” button

Version: 5.0 [LOGIN](#) | [REGISTER](#)

**STITCH**

**SEARCH**

**Step1:** Click on "item-by-name"

**Step2:** Enter the name of the protein/small molecule/drug

**Step3:** Click "SEARCH"

Item by name

Multiple names

Chemical structure(s)

Protein sequence(s)

Examples

Random entry

Single Item by Name / Identifier

Item Name: **Drug** (examples: #1 #2 #3)

Tamoxifen

Organism:

auto-detect

SEARCH

The screenshot shows the STITCH search interface. At the top, there's a navigation bar with 'Version: 5.0' and links for 'LOGIN' and 'REGISTER'. Below the header, the 'STITCH' logo is displayed. On the left, a sidebar lists search options: 'Multiple names', 'Chemical structure(s)', 'Protein sequence(s)', 'Examples', and 'Random entry'. The main area is titled 'SEARCH' and contains a form for 'Single Item by Name / Identifier'. It has fields for 'Item Name:' (with placeholder 'Drug' and examples '#1 #2 #3') and 'Organism:' (set to 'auto-detect'). A red box highlights the 'Item Name:' field with the text 'Tamoxifen' inside, which is also circled in red. A red arrow points from the 'Item Name:' field to the 'SEARCH' button. Another red box highlights the 'SEARCH' button itself. Red annotations with arrows point to each of these three steps: Step 1 points to the 'Item by name' link in the sidebar; Step 2 points to the 'Item Name:' field; and Step 3 points to the 'SEARCH' button.

© STITCH CONSORTIUM 2016



EMBL - European Molecular Biology Laboratory



SIB - Swiss Institute of Bioinformatics



CPR - NNF Center for Protein Research

ABOUT

Content

References

Contributors

INFO

Scores

Use scenarios

FAQs

ACCESS

Versions

APIs

Licensing

CREDITS

Funding

Datasources

Partners

Software

- Leave default species to “Homo sapiens”
- Click “CONTINUE” button

Version: 5.0 [LOGIN](#) | [REGISTER](#)

# STITCH

**Step1: Keep default selection**

There are several matches for 'Tamoxifen'. Please select one from the list below and press Continue to proceed.

chemical  tamoxifen

organism	protein
<input type="checkbox"/> Homo sapiens	FOXL2 - forkhead box L2; Transcriptional regulator. Critical factor essential for ovary differentiation and maintenance, and repression of the genetic program for somatic testis determination. Prevents trans-differentiation of ovary to testis through transcriptional repression of the Sertoli cell-promoting gene SOX9 (By similarity). Has apoptotic activity in ovarian cells. Suppresses ESR1-mediated transcription of PTGS2/COX2 stimulated by tamoxifen (By similarity). Is a regulator of CYP19 expression (By similarity). Participates in SMAD3-dependent transcription of FST via the intronic SMAD-binding element (By similarity).
<input type="checkbox"/> Mus musculus	Foxl2 - forkhead box L2; Transcriptional regulator. Critical factor essential for ovary differentiation and maintenance, and repression of the genetic program for somatic testis determination. Prevents trans-differentiation of ovary to testis through transcriptional repression of the Sertoli cell-promoting gene SOX9 (By similarity). Has apoptotic activity in ovarian cells. Suppresses ESR1-mediated transcription of PTGS2/COX2 stimulated by tamoxifen (By similarity). Activates SIRT1 transcription under cellular stress conditions. Activates transcription of OSR2. Is a regulator of CYP19 expression. Is a transcriptional repressor (By similarity).
<input type="checkbox"/> Bos taurus	FOXL2 - Forkhead box protein L2 ; Transcriptional regulator. Critical factor essential for ovary differentiation and maintenance, and repression of the genetic program for somatic testis determination (By similarity). Prevents trans-differentiation of ovary to testis through transcriptional repression of the Sertoli cell-promoting gene SOX9 (By similarity). Has apoptotic activity in ovarian cells (By similarity). Suppresses ESR1-mediated transcription of PTGS2/COX2 stimulated by tamoxifen (By similarity). Activates SIRT1 transcription under cellular stress conditions (By similarity). Activates ...
<input type="checkbox"/> Sus scrofa	FOXL2 - Forkhead box protein L2 ; Transcriptional regulator. Critical factor essential for ovary differentiation and maintenance, and repression of the genetic program for somatic testis determination (By similarity). Prevents trans-differentiation of ovary to testis through transcriptional repression of the Sertoli cell-promoting gene SOX9 (By similarity). Has apoptotic activity in ovarian cells (By similarity). Suppresses ESR1-mediated transcription of PTGS2/COX2 stimulated by tamoxifen (By similarity). Activates SIRT1 transcription under cellular stress conditions (By similarity). Activates ...

**Step2: Click “CONTINUE”**

[CONTINUE →](#)

© STITCH CONSORTIUM 2016

 EMBL - European Molecular Biology Laboratory

 SIB - Swiss Institute of Bioinformatics

 CPR - NNF Center for Protein Research

## ABOUT

Content

References

Contributors

## INFO

Scores

Use scenarios

FAQs

## ACCESS

Versions

APIs

Licensing

## CREDITS

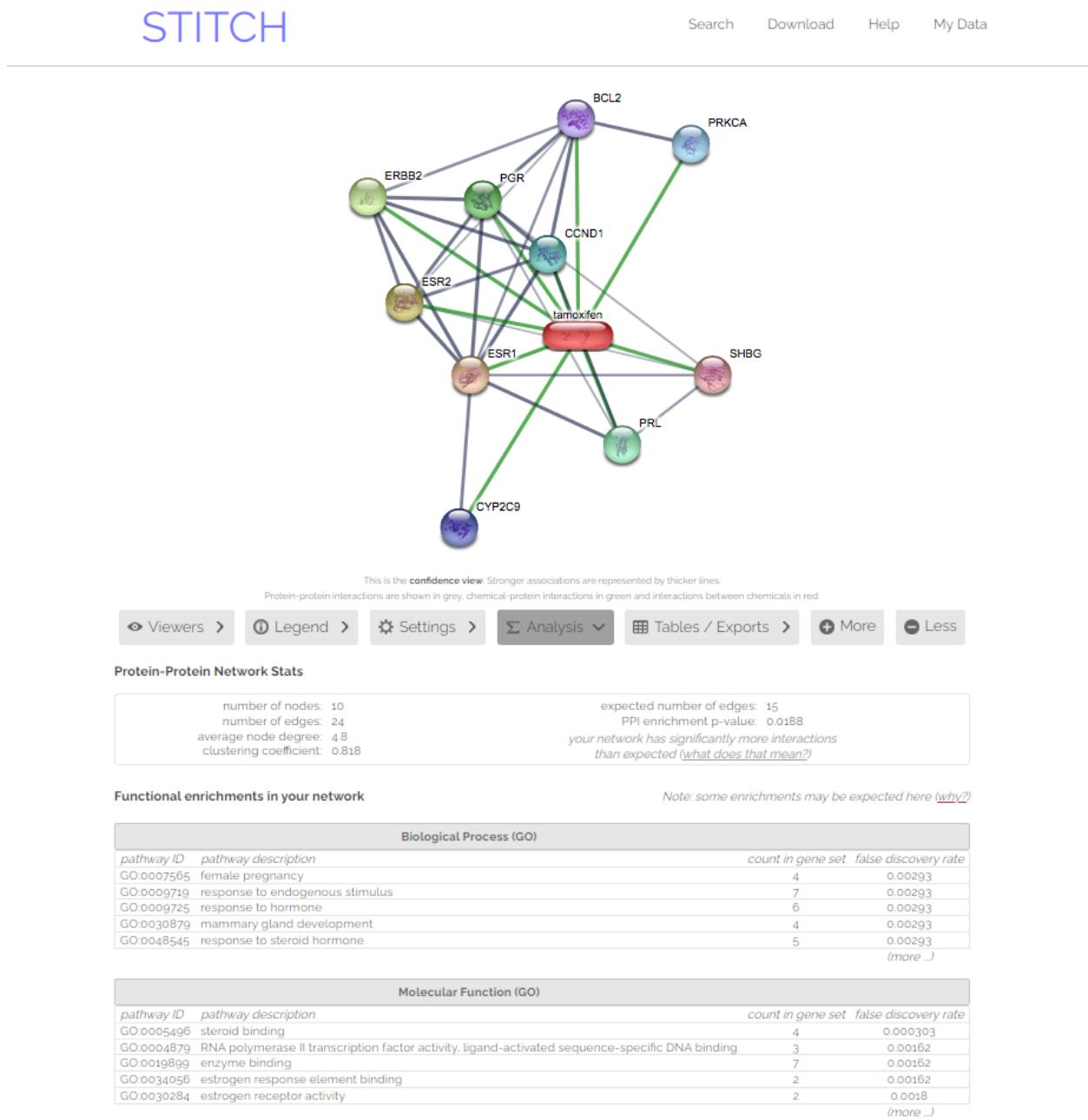
Funding

Datasources

Partners

Software

- Click on “Analysis”
  - Click on node/edge in the network to get more details for a given interaction.
  - Observations: Protein-Protein Network stats, Functional enrichment etc



## 2.7 Using miRNet

### 2.7.1 Prepare input data

Download the miRBase IDs

<https://wd.cri.uic.edu/pathway/miR-ids.txt>

### 2.7.2 Navigate to miRNet

<https://www.mirnet.ca/miRNet/home.xhtml>

Click on the “miRNAs” button on the home screen to upload our miR gene list.

The screenshot shows the miRNet home page. At the top, there is a navigation bar with links: Home, FAQs, Tutorials, Resources, Gallery, APIs, miRNetR, Updates, About, and Contact. Below the navigation bar, there is a large input selection grid. The grid has a header row labeled "Input Type" and a column labeled "Click on a module below to start". The grid rows correspond to different input types: "Mixed input types", "Upload a data table", "Select from a database", and "Upload a query list". The "Upload a query list" row contains seven buttons: "miRNAs" (which is highlighted with a red circle), "SNPs", "Genes", "ncRNAs", "Transcription factors", and "Xeno-miRs".

Input Type	Click on a module below to start					
Mixed input types				Multiple query types		
Upload a data table			Expression table	RT-qPCR data		
Select from a database		Diseases	Small compounds	Epigenetic modifiers	Xeno-miR explorer	
Upload a query list	miRNAs	SNPs	Genes	ncRNAs	Transcription factors	Xeno-miRs

### 2.7.3 Perform network analysis for 10 microRNAs

- **Step 1:** Select “H. sapiens (human)” from “Organism” dropdown
- **Step 2:** Select “miRBase ID” from “ID type” dropdown
- **Step 3:** Select “Exosomes” from “Tissue (humans only)” dropdown
- **Step 4:** Select “Genes miRTarbase v8.0” and “LncRNAs” from “Targets” dropdown
- **Step 5:** Copy paste following 10 miRBase IDs into miRNA list from `mir-ids.txt`
- **Step 6:** Click **Submit**.
- **Step 7:** Then click **Proceed**

The screenshot shows the miRNet web application interface. At the top, there's a navigation bar with links for Home, FAQs, Tutorials, Resources, Gallery, APIs, miRNetR, Updates, About, and Contact. Below the navigation is a breadcrumb trail: Home > Upload. The main area has a title "Enter a list of miRNAs below:" and a red note "Match settings as noted:". On the left, there are dropdown menus for Organism (set to H. sapiens (human)), ID type (set to miRBase ID), and Tissue (set to Exosomes [1250]). Under Targets, a dropdown menu is open, showing "Selections" and a list of options: Genes (miRTarBase v8), Genes (tarBase v8.0), Genes (miRecords), IncRNAs, circRNAs, Pseudogenes, and sncRNAs. The "IncRNAs" option is checked. Below this is a text input field containing "hsa-mir-194-5p" and "hsa-mir-24-3p". A red box highlights the text "Copy-paste miR IDs here". At the bottom, there are two buttons: "Submit" and "Click Submit First" (in red). A red note "Then Click Proceed" is positioned above the "Proceed" button. To the right, there's a "Do You Know?" section with information about miRNA ID formats and target selection.

**Match settings as noted:**

Organism: H. sapiens (human)

ID type: miRBase ID

Tissue (human only): Exosomes [1250]

Targets: Selections

Include PPI (gene only)

Include tf2gene

miRNA list (one entry per line):  
hsa-mir-194-5p  
hsa-mir-24-3p

**Do You Know?**

Please pay attention to the miRNA ID format. For miRBase ID, the format is "hsa-mir-1" (letters should be low-case) and for accession number is "MIMAT000001" (letters should be capital).

You can select one or multiple "Targets" to be included in the network, provided they have direct interactions with input miRNA list based on our knowledge

In addition, miRNet can automatically recognize different versions of miRBase IDs (v15-v22), as well as link pre-miRNAs to their mature forms.

**Click Submit First**

**Then Click Proceed**

## 2.7.4 Results overview

- **Step 1:** Right click “Browse” for mir2gene to open results in new tab to review the table.
- **Step 2:** Right click “Browse” for mir2lnc to open results in new tab to review the table.
- **Step 3:** Click “Minimum Network” to get smaller network.
  - Observe number of edges in the Networks table (before clicking on minimum network)
  - There are 6196 edges in the network based on our query
  - Observe number of edges in the network table after “minimum network”
  - Number of edges reduced to 90
- After filtering, click “Proceed”

**Interaction Tables**

The pair-wise interaction tables together with the supporting information are listed below. You can click the corresponding link to download (Download) or browse (View) the interaction tables, or click the Proceed button at the bottom to directly explore the results in a network context.

↳ mir2gene	[miRNA:10, gene: 4024]	↳ Browse	Step1: Right Click on "Browse" and open mir2gene results in new tab
↳ mir2lnc	[miRNA:10, lncRNA: 295]	↳ Browse	Step2: Right Click on "Browse" and open mir2lnc results in new tab

**Networks**

In some cases, multiple isolated networks will be generated, with a big 'continent' containing most of queries, and several small 'islands' containing one or a few queries. These networks will be available for visual analysis in the next step.

Networks	Queries	Nodes	Edges	Topology
↳ mirnet1	miRNA: 10;	lncRNA: 295; Gene: 4024; miRNA: 10;	6196	View

Step 3: To reduce the number of edges: Click "Minimum Network" -- This will update "Networks Table"

Our query results in a very big network with 6196 edges

Step4: Click "Proceed" to see the network

Network Tools: ?

- Degree Filter
- Betweenness Filter
- Shortest Path Filter
- Manual Batch Filter
- Minimum Network**
- Steiner Forest Network

Update Network

Reset Network

Previous

Downloads

Proceed

## 2.7.5 Review Interaction Tables results: microRNA - Gene interactions

- microRNA - Gene interactions (from “Browse” button)
  - Review Advanced filter for different options to filter based on Target, Method, Literature, etc.
  - Filtered results can be used further downstream to do network analysis

**2.0**

miRNet -- a miRNA-centric network visual analytics platform

Home FAQs Tutorials Resources Gallery APIs Updates About Support

Upload Network Builder Downloads Interaction Table

For cattle (*B. taurus*), chicken (*G. gallus*), pig (*S. scrofa*) and helminth (*S. mansoni*), the result will be mainly composed of interaction data predicted using miRanda. You can use the “Advanced Filter” to exclude the results based on the miRanda scores to keep more confident predictions.

**Step 1: Click on “Advanced Filter” to filter results**

**Advanced Filter** **Reset**

miRNA	Link	Tissue	Target:Gene	Link	Method	Literature	Action
hsa-mir-16-5p	miRBase	Exosomes	ABC B7	Entrez	CLASH	23622248	Delete
hsa-mir-16-5p	miRBase	Exosomes	ABCF1	Entrez	Proteomics	18668040	Delete
hsa-mir-16-5p	miRBase	Exosomes	ABL2	Entrez	PAR-CLIP//Sequencing	20371350, 23592263, 24398324, 23446348, 2	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACOX1	Entrez	HITS-CLIP	22473208	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACP2	Entrez	Proteomics//pSILAC	18668040	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACTB	Entrez	CLASH	23622248	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACTG1	Entrez	CLASH//Proteomics	18668040, 23622248	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACTN4	Entrez	CLASH	23622248	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACTN1	Entrez	CLASH	23622248	Delete
hsa-mir-16-5p	miRBase	Exosomes	ACVR2A	Entrez	Luciferase reporter assay//PAR-CLIP//W	20216554, 20371350	Delete
hsa-mir-16-5p	miRBase	Exosomes	ADK	Entrez	CLASH	23622248	Delete
hsa-mir-16-5p	miRBase	Exosomes	ADORA2A	Entrez	Luciferase reporter assay//qRT-PCR//W	27476546	Delete
hsa-mir-16-5p	miRBase	Exosomes	ADORA3	Entrez	HITS-CLIP	19536157	Delete
hsa-mir-16-5p	miRBase	Exosomes	ADRA2B	Entrez	HITS-CLIP	23824327	Delete
hsa-mir-16-5p	miRBase	Exosomes	ADSS	Entrez	Proteomics//pSILAC	18668040	Delete

(1 of 67) 1 2 3 4 5 6 7 8 9 10 > >> 15 +

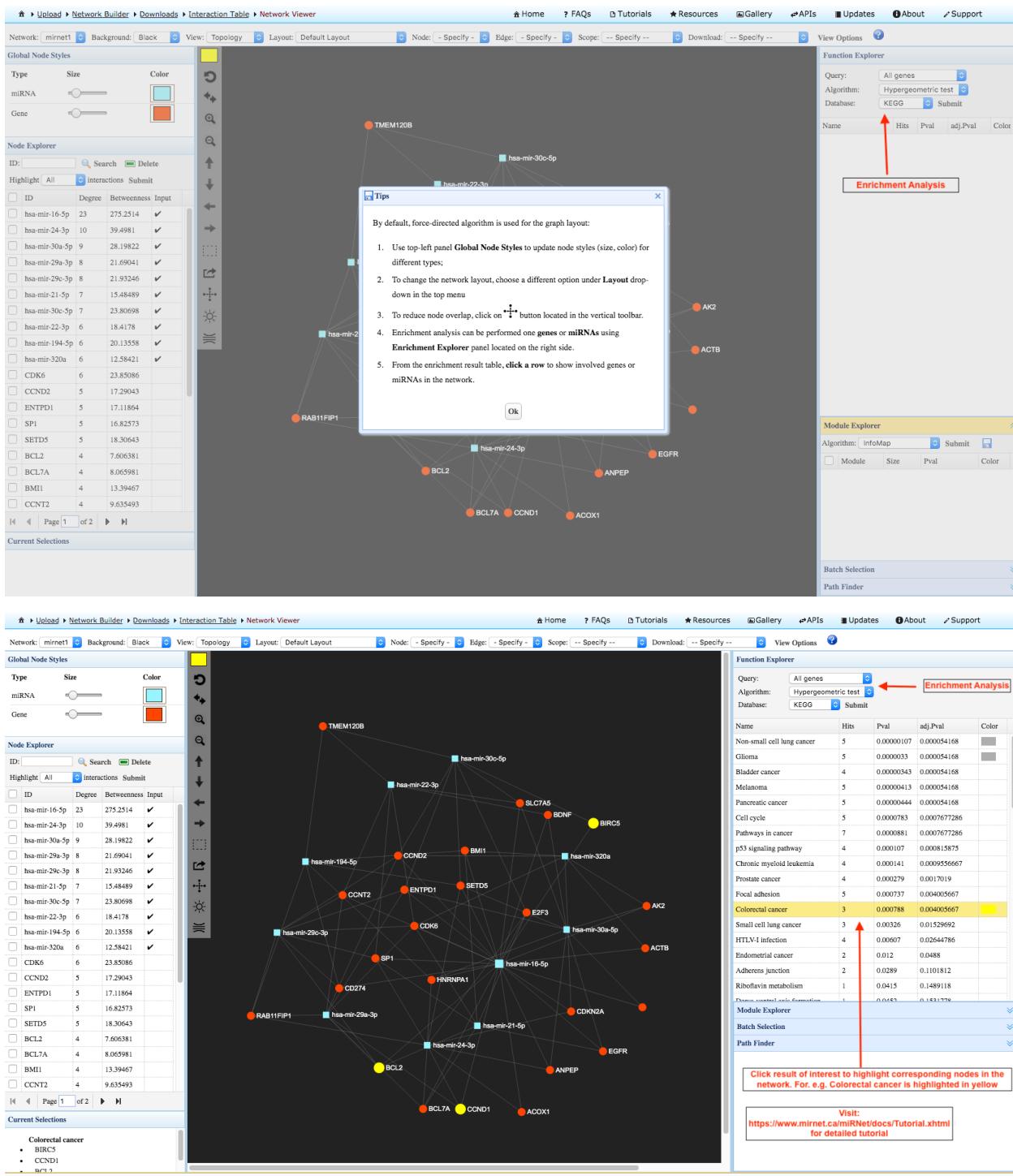
**Step 2: Click on “Proceed” to build the network based on filtered results. WE WILL NOT BE PERFORMING THIS ANALYSIS TODAY**

**Downloads** **Proceed**

Xia Lab @ McGill (last updated 2020-04-03)

## 2.7.6 Review network

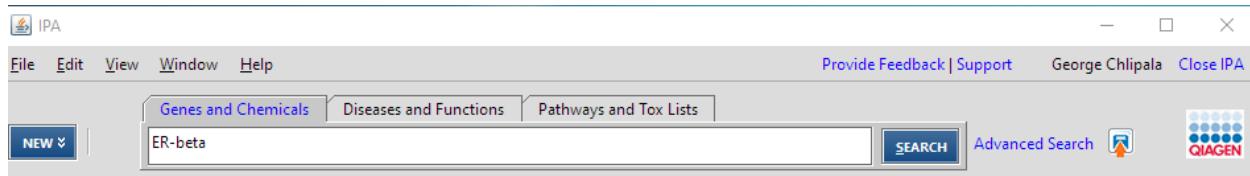
- Review “minimum network”
- Enrichment analysis options



## 2.8 INSTRUCTOR DEMONSTRATION: Network analysis using IPA

### 2.8.1 Build (Grow) a network from a single molecule

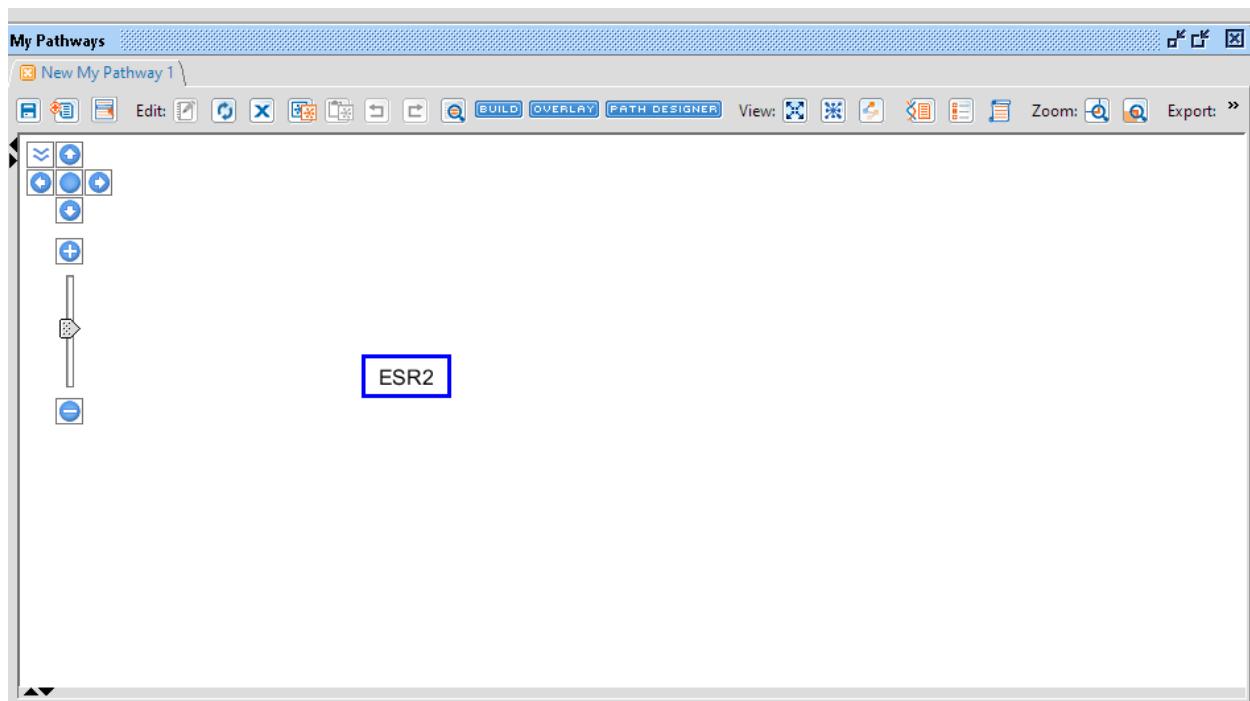
1. Enter “ER-beta” in search dialog and click **Search**.



2. Select “ESR2” and click **Add To My Pathway** and select *New My Pathway*.

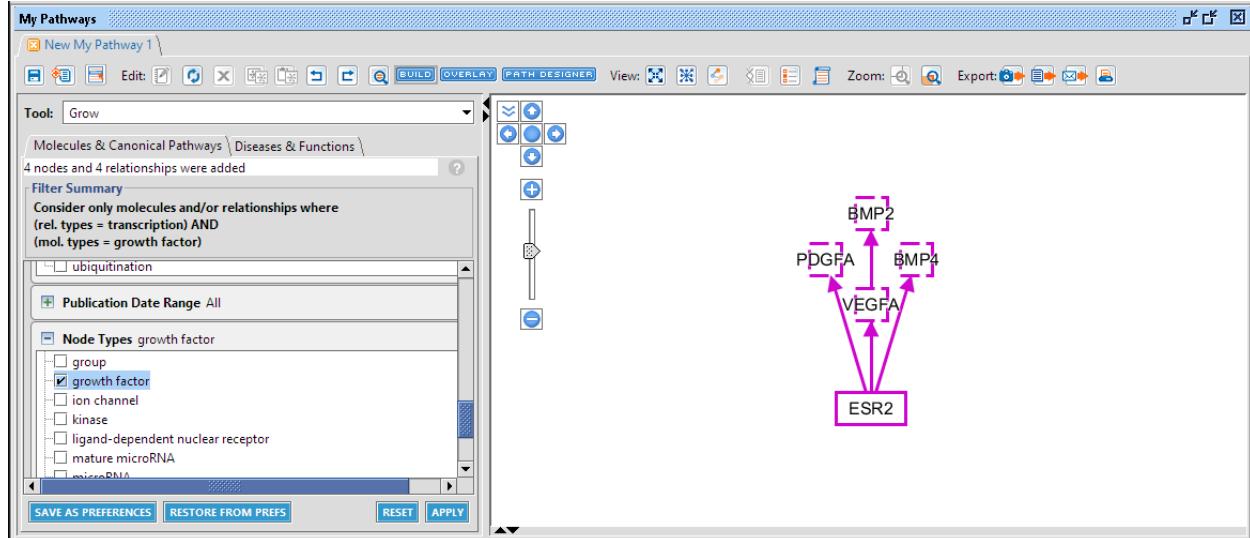
A screenshot of the IPA software showing the search results for 'ER-beta'. The search bar at the top has 'ER-beta' entered. Below it, the 'Project Manager' section shows a list of items under 'My Projects'. One item, 'ESR2', is highlighted with a yellow box. The 'Search' tab is active, and the 'ADD TO MY PATHWAY' button is highlighted. A table below lists the search results for 'ER-beta', with one item shown: ESR2, which is highlighted with a yellow box. The table columns include: #, Symbol, Matched Term, Synonym(s), Entrez Gene Nr, Location, Type(s), Biomarker, Drug(s), Target(s), and Sp. The 'Synonym(s)' column lists beta ESTROGEN, Erb, ER[b], ERB2, ER-beta, ER  $\beta$ , ESRB, ESR-BETA, ESR- $\beta$ , and ESTRB.

3. In the *Network* view click on the **ESR2** node (rectangle) and click **BUILD** button.



4. Select *Grow* as the Tool.

- a. In *General Settings*, set Grow out... ...that are “Downstream of selected molecules”
- b. In *Relationship Types*, Click “Select all” to unselect all types and scroll down to select “transcription”
- c. In *Node Types*, Click “Select all” to unselect all types and scroll down to select “growth factor”
- d. Click **APPLY**



5. Export molecules and relationships



- Export All Molecules

The screenshot shows an Excel spreadsheet titled "IPA\_grow\_network\_nodes [Compatibility Mode] - Excel". The columns are labeled A through J. The data includes the following rows:

	A	B	C	D	E	F	G	H	I	J
1	© 2000-2020 QIAGEN. All rights reserved.									
2	Symbol	Synonym(	Entrez Gei	Location	Family	Drugs	Entrez Gene ID for Human	Entrez Gene ID for Mouse	Entrez Gene ID for Rat	
3	BMP2	AI467020;	bone morph	Extracellul	growth factor		650	12156	29373	
4	BMP4	bone morph	bone morph	Extracellul	growth factor		652	12159	25296	
5	ESR2	beta ESTR	estrogen r	Nucleus	ligand-dependent nuc	suilindac/tamoxifen,	2100	13983	25149	
6	PDGFA	Pdgf a-ch	platelet dei	Extracellul	growth factor		5154	18590	25266	
7	VEGFA	Gd-vegf,	V vascular ei	Extracellul	growth factor	bevacizumab/caper	7422	22339	83785	

- Export All Relationships

The screenshot shows an Excel spreadsheet titled "IPA\_grow\_network\_edges [Compatibility Mode] - Excel". The columns are labeled A through O. The data includes the following rows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	© 2000-2020 QIAGEN. All rights reserved. USE OF THIS CONTENT IS SUBJECT TO THE TERMS OF YOUR IPA END USER LICENSE AGREEMENT														
2	From Mole	Relationship T	Y To Molecule(s)		Catalyst(s)										
3	ESR2	transcription	BMP2												
4	ESR2	transcription	BMP4												
5	ESR2	transcription	PDGFA												
6	ESR2	transcription	VEGFA												

## 2.8.2 Find connections between molecules

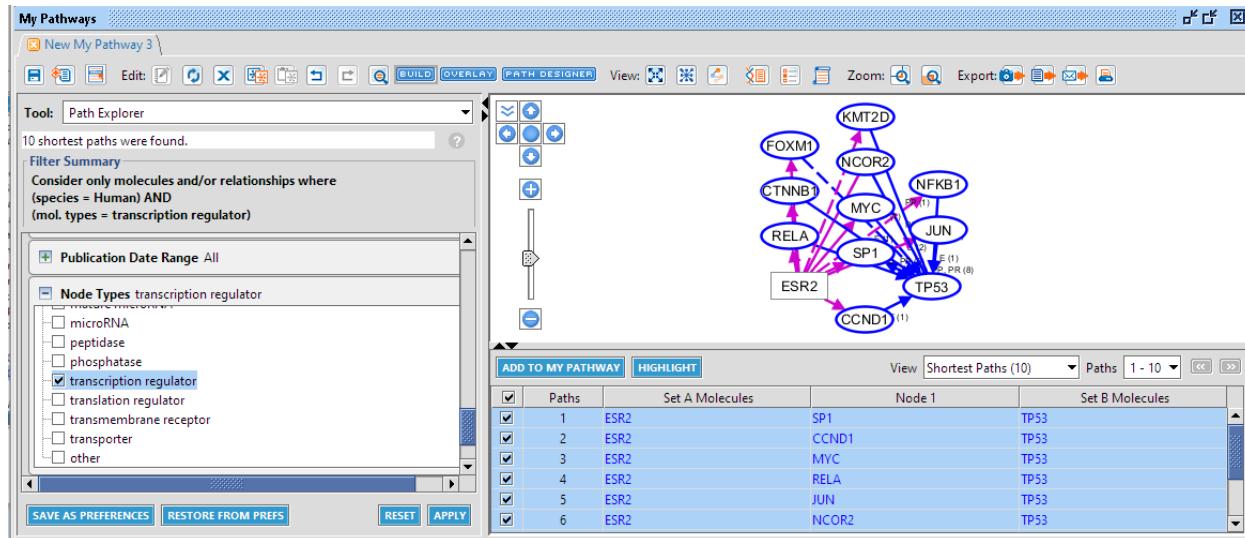
1. Enter “ER-beta” in search dialog and click **Search**.
2. Select “ESR2” and click **Add To My Pathway** and select *New My Pathway*.
3. Enter “p53” in search dialog and click **Search**
4. Select “TP53” and click **Add To My Pathway** and select bottom *New My Pathway*.

Symbol	Synonym(s)	Entrez Gene Name	Location	Type(s)	Biomarker Application
TP53	p53,P53 cellular tumour antigen,p53 tumor suppressor,tumor protein p53,tumour protein p53	tumor protein p53	Nucleus	transcription	diagnosis, disease progression, efficacy, prognosis, response to therapy, unspecified application
TP53BP2	p53-Binding,tumor protein p53 binding protein 2,tumor protein p53 binding protein 2,tumour protein p53 binding protein 2,transformation related protein Trp53bp2,	tumor protein p53 binding protein 2	Nucleus	other	

5. In the *Network* view click **BUILD** button.
6. Select *Path Explorer* as the Tool.
  - a. Select “ESR2” in the view and click the **ADD** button for *Set 1*
  - b. Select “TP53” in the view and click the **ADD** button for *Set 2*
  - c. Set the *Direction* as “From Set A to Set B”
  - d. In *Species*, click “Select all” to unselect all species and select “Human”
  - e. In *Node Types*, click “Select all” to unselect all types and scroll down to select “transcription regulator”
  - f. Click **APPLY** button.

Paths	Set A Molecules	Node 1	Set B Molecules
1	ESR2	SPI	TP53
2	ESR2	CCND1	TP53
3	ESR2	MYC	TP53
4	ESR2	RELA	TP53
5	ESR2	JUN	TP53
6	ESR2	NCOR2	TP53

7. Click on the checkbox in the header row and click **ADD TO MY PATHWAY**



5. Export molecules and relationships

- Export *All Molecules*

IPA_path_network_nodes [Compatibility Mode] - Excel								
C11	A	B	C	D	E	F	G	H
1 © 2000-2020 QIAGEN. All rights reserved.								
2 Symbol	Synonym(	Entrez Gei	Location	Family	Drugs	Entrez Gene ID for Human	Entrez Gene ID for Mouse	Entrez Gene ID for Rat
3 CCND1	AI327039, cyclin D1		Nucleus	transcription regulato	arsenic trioxide/cytarabini	595	12443	58919
4 CTNNB1	armadillo, catenin be		Nucleus	transcription regulato	PRI-724	1499	12387	84353
5 ESR2	beta ESTR estrogen r		Nucleus	ligand-dependent nuc	sulindac/tamoxifen, ethiny	2100	13983	25149
6 FOXM1	AA408308 forkhead b		Nucleus	transcription regulator		2305	14235	58921
7 JUN	Activator p Jun proto-		Nucleus	transcription regulator		3725	16476	24516
8 KMT2D	AAD10, Al lysine mett		Nucleus	transcription regulator		8085	381022	100362634
9 MYC	AU016757 MYC prot		Nucleus	transcription regulato	AVI-4126, MYC-targeting	4609	17869	24577
10 NCOR2	N-Cor, nu nuclear re		Nucleus	transcription regulator		9612	20602	360801
11 NFKB1	CVID12, Enuclear fa		Nucleus	transcription regulato	triflusal, bortezomib/dexan	4790	18033	81736
12 RELA	CMCU, NFRELA prot		Nucleus	transcription regulato	NF-kappaB decoy	5970	19697	309165
13 SP1	1110003E Sp1 transcr		Nucleus	transcription regulator		6667	20683	24790
14 TP53	bbl, BCC7 tumor prot		Nucleus	transcription regulato	BI 907828, kevetrin, cene	7157	22059	24842

- Export *All Relationships*

IPA_path_network_edges [Compatibility Mode] - Excel								
B3	A	B	C	D	E	F	G	H
1 © 2000-2020 QIAGEN. All rights reserved. USE OF THIS CONTENT IS SUBJECT TO THE TERMS OF YOUR IPA END USER LICENSE AGREEMENT								
2 From Molecule(s)	Relationship Type	To Molecule(s)		Catalyst(s)				
3 CCND1	expression	TP53						
4 CTNNB1	expression	TP53						
5 ESR2	activation	JUN						
6 ESR2	activation	SP1						
7 ESR2	expression	CCND1						
8 ESR2	expression	FOXM1						
9 ESR2	expression	MYC						
10 ESR2	localization	KMT2D						
11 ESR2	phosphorylation	JUN						
12 ESR2	protein-DNA interaction:	CCND1						
13 ESR2	protein-DNA interaction:	FOXM1						