Pathway Analysis

Research Informatics Core

November 02, 2023

Contents

1 Morning

1.1 Gene ID conversion using DAVID

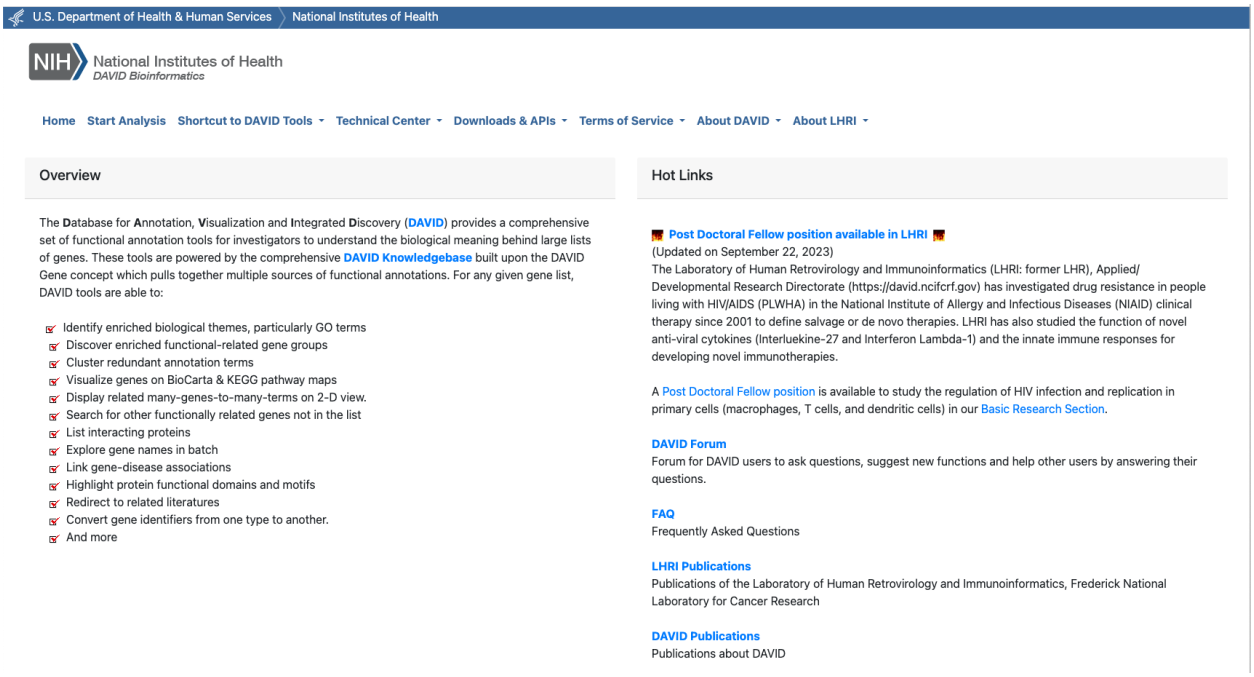
1.1.1 Prepare input data

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

`https://wd.cri.uic.edu/pathway/GeneID_Conversion_list.txt`

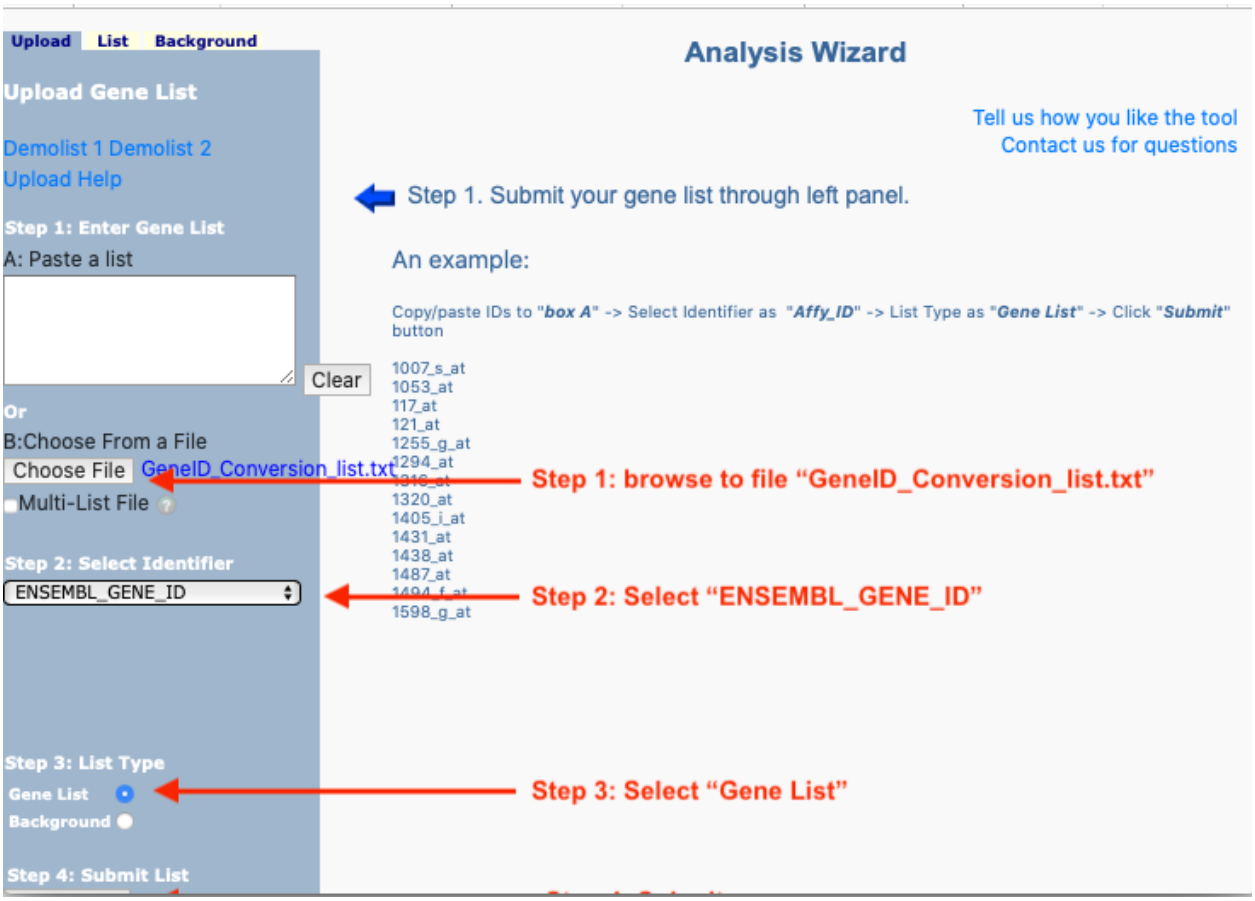
1.1.2 Navigate to DAVID

Navigate to DAVID `https://david.ncifcrf.gov`, and choose "Start Analysis" from the top menu.



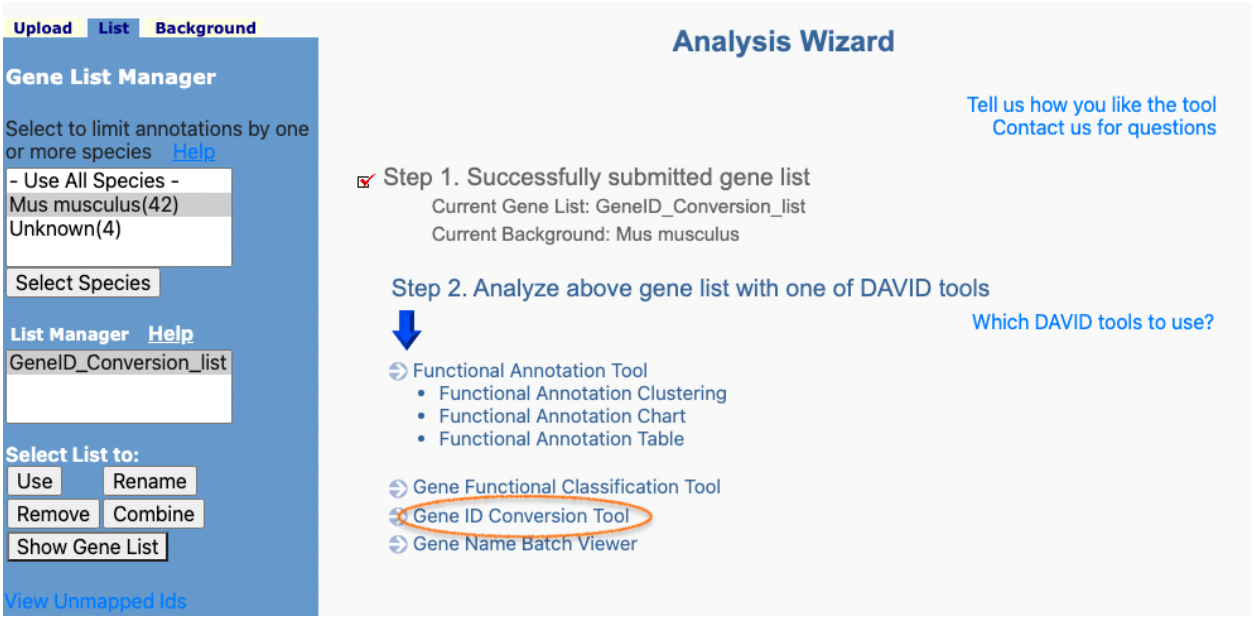
1.1.3 Upload to DAVID

- **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.

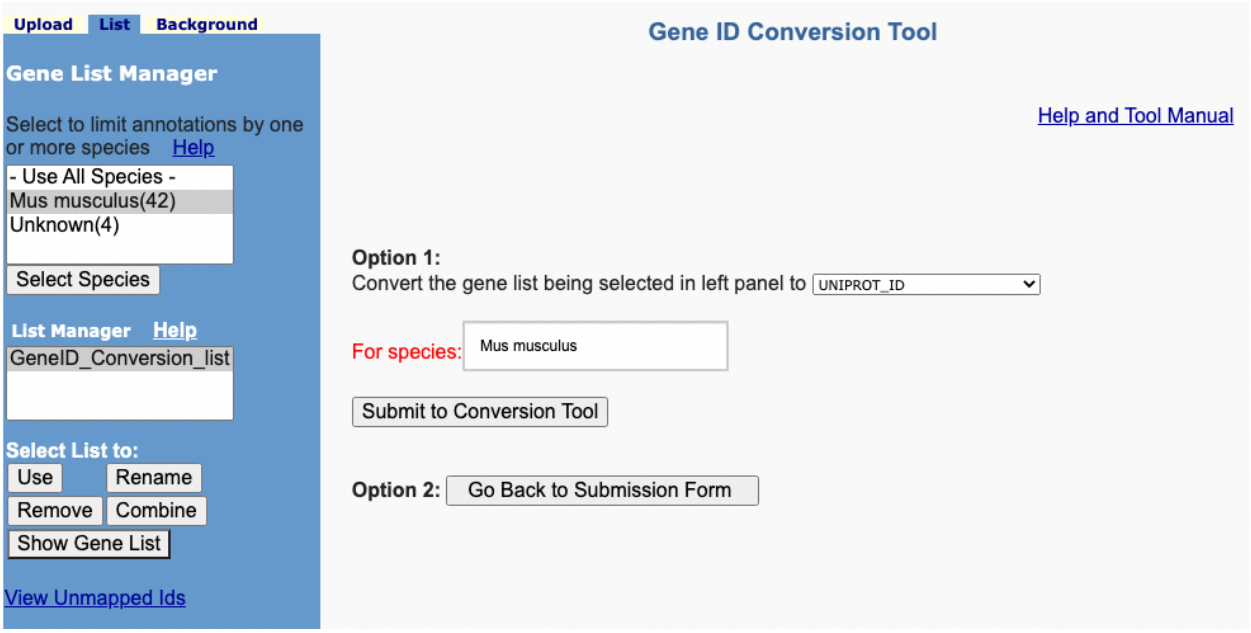


1.1.4 Submit to conversion tool

- **Step 5:** Choose "GeneID Conversion Tool"

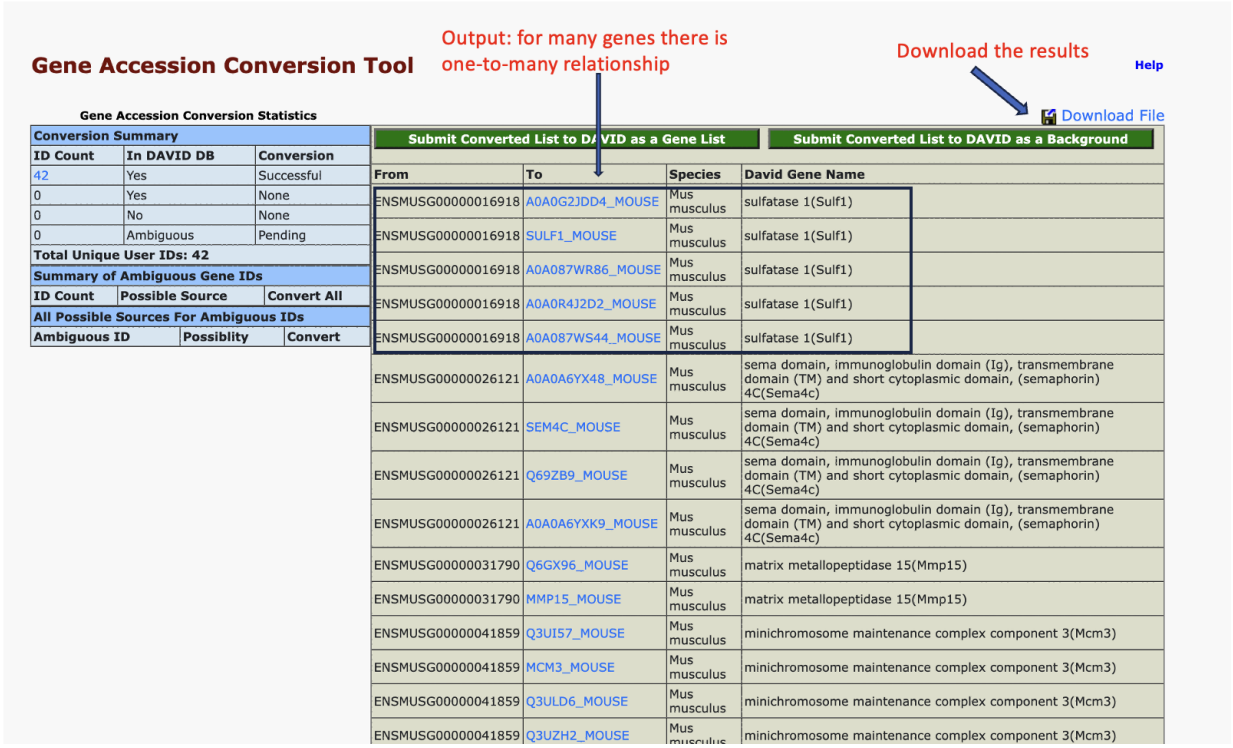


- **Step 5:** Choose "UNIPROT_ID" from the drop down menu
- **Step 6:** Enter "Mus musculus" for the species
- **Step 7:** Click on "Submit to Conversion Tool" to submit the gene list for conversion*



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



- 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
1	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	Q69ZB9_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(Hecw2)									
18	ENSMUSG00000042807	A3KPB7_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hecw2)									
19	ENSMUSG00000042807	Q8BQD5_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hecw2)									
20	ENSMUSG00000042807	Q8C8B5_MOUSE	Mus musculus	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2(Hecw2)									
21	ENSMUSG00000026042	CO5A2_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)									

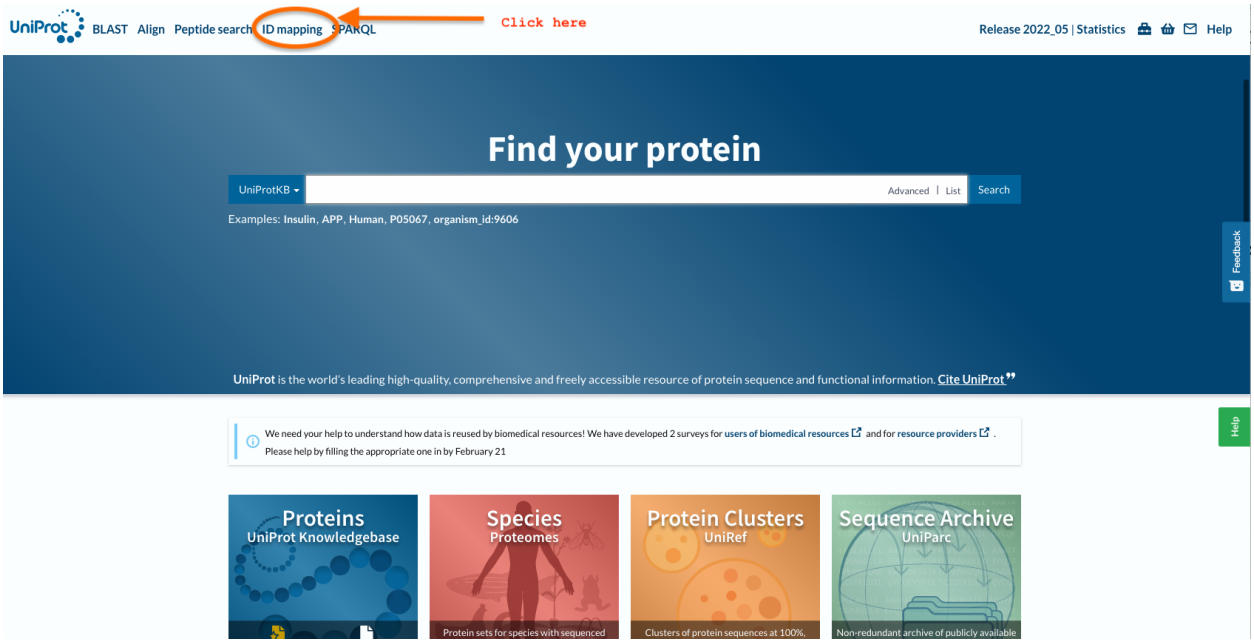
1.2 Gene ID conversion using UniProt

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`

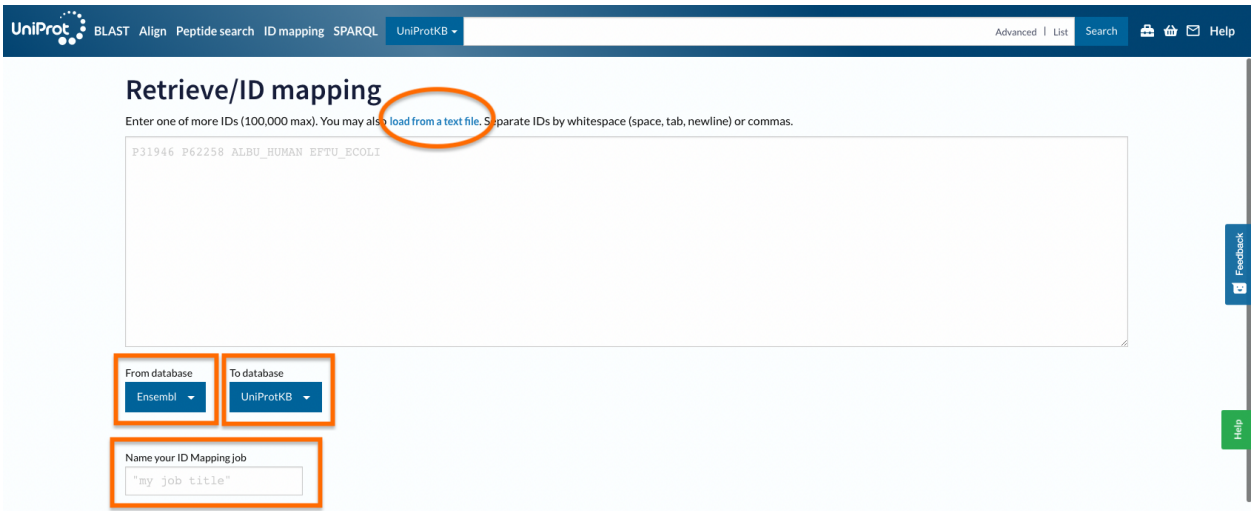
1.2.1 Navigate to UniProt

Navigate to UniProt `https://www.uniprot.org`, and choose "ID mapping" from top menu bar



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*



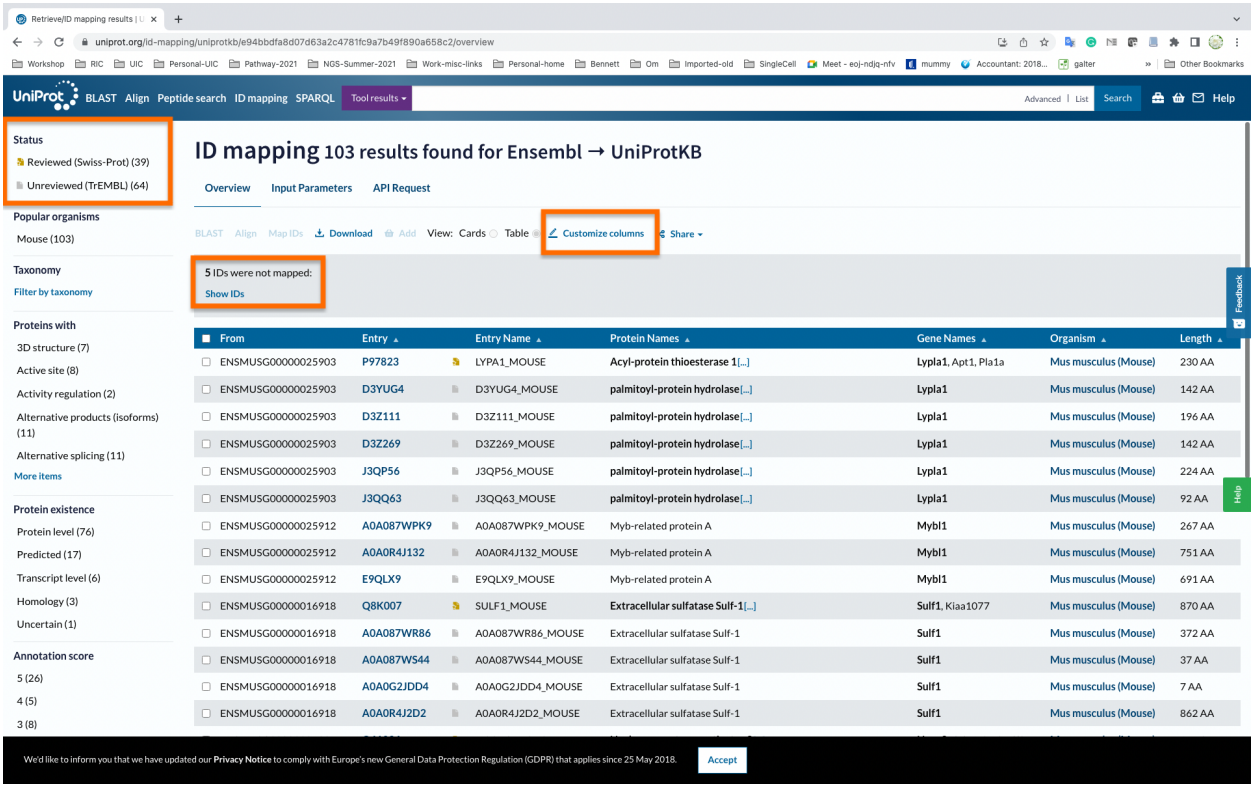
1.2.3 Results from UniProt

UniProt Knowledgebase has two databases

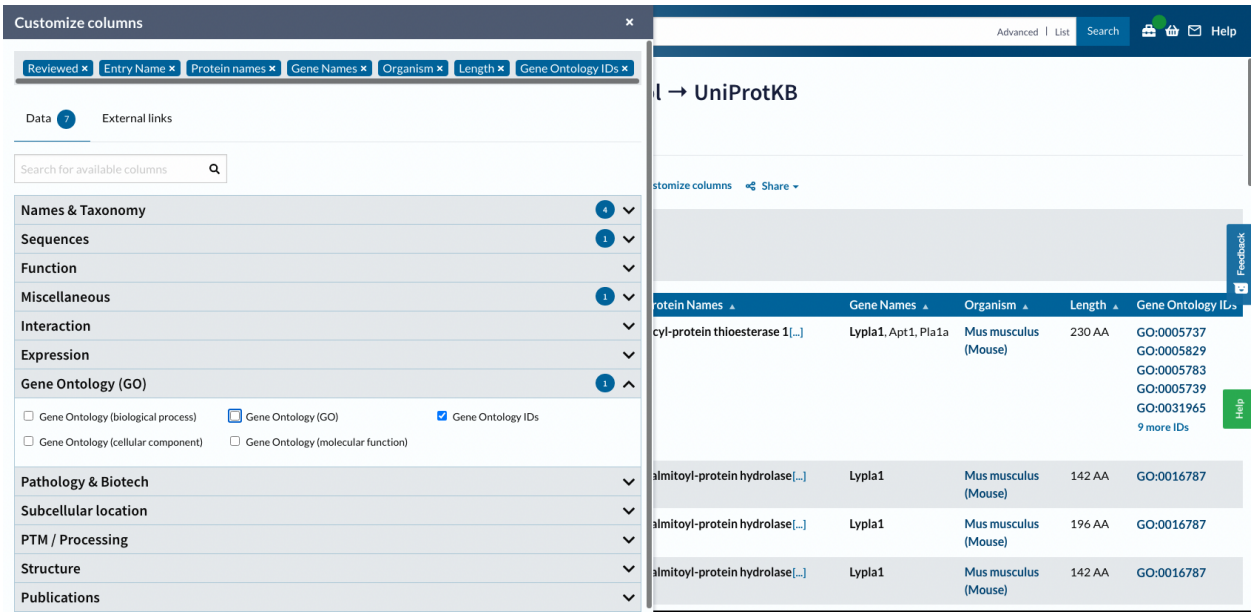
- **Swiss-Prot:** Proteins that are manually annotated and reviewed
- **TrEMBL:** Proteins that are automatically annotated and not reviewed

Observations:

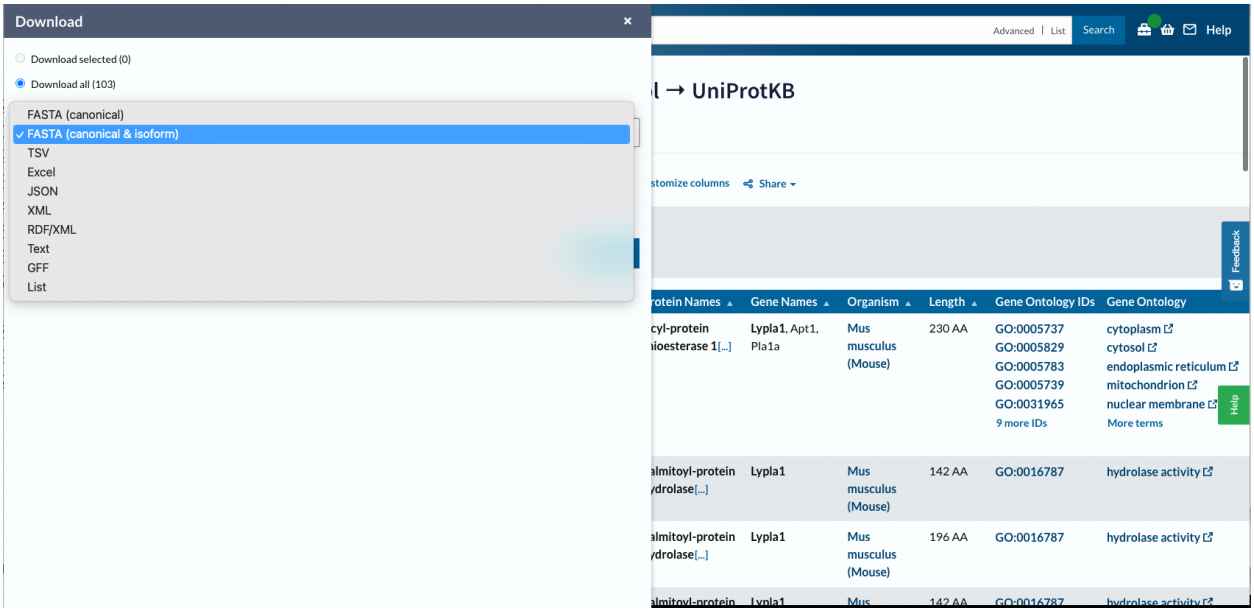
- 41 out 46 Ensemble identifiers are successfully mapped to 103 UniProtKB IDs
  - For many Ensembl gene IDs we have multiple UniProt IDs
  - Out of 103 records, 39 records have Swiss-Prot Ids and 64 have TrEMBL IDs
- Click pencil icon by 'Customize columns' - to review other available annotations.



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



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



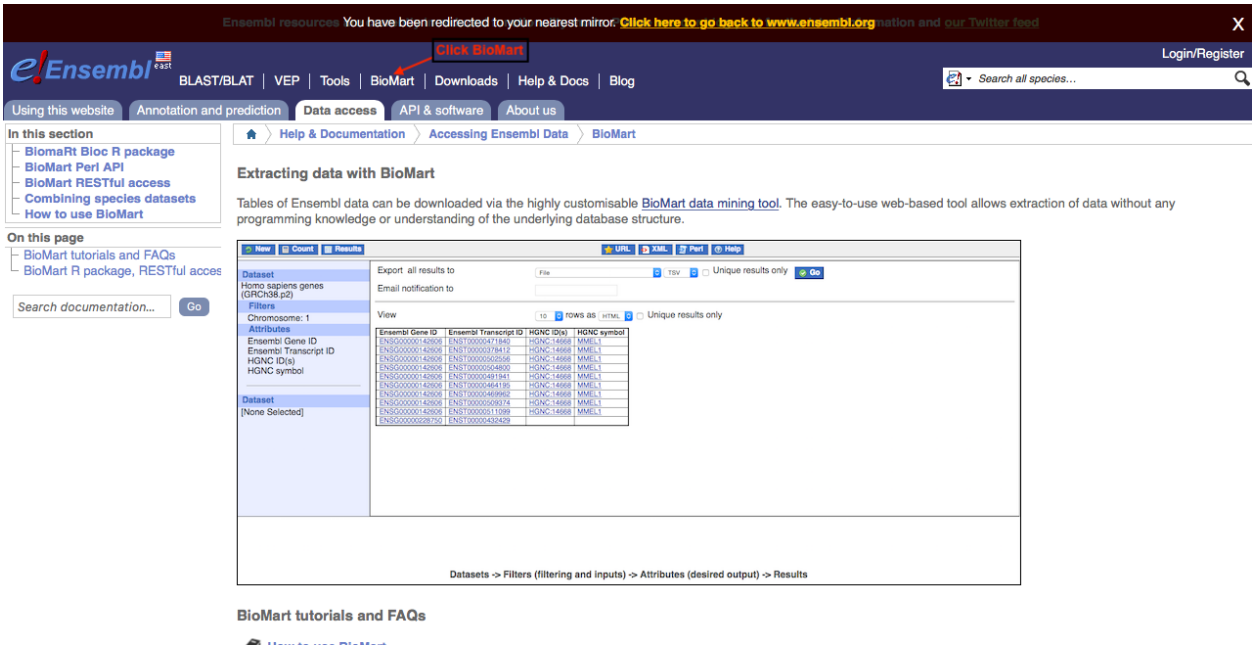
1.3 Gene ID conversion using BioMart

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`

1.3.1 Navigate to BioMart

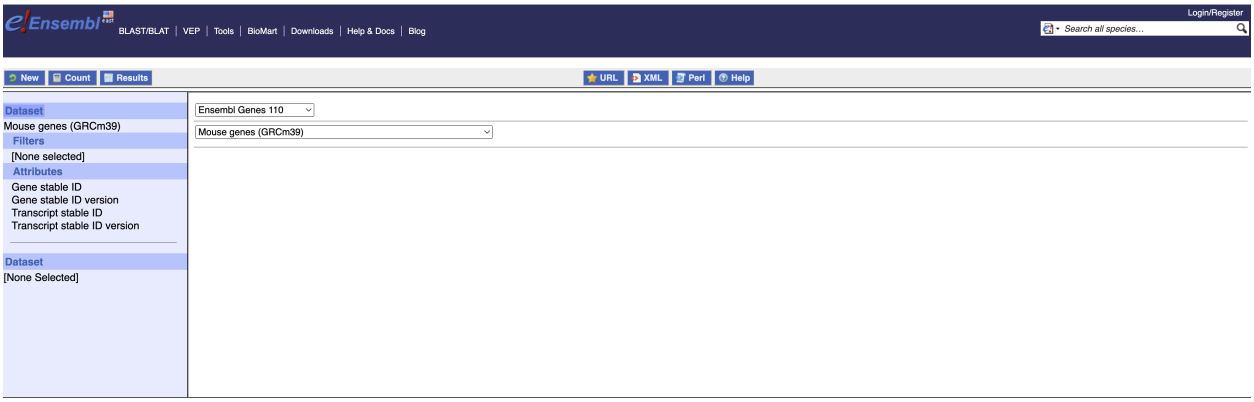
Navigate to Ensembl BioMart `https://useast.ensembl.org/info/data/biomart/index.html`, and choose **BioMart** from top menu bar.



1.3.2 Choose Database(s)

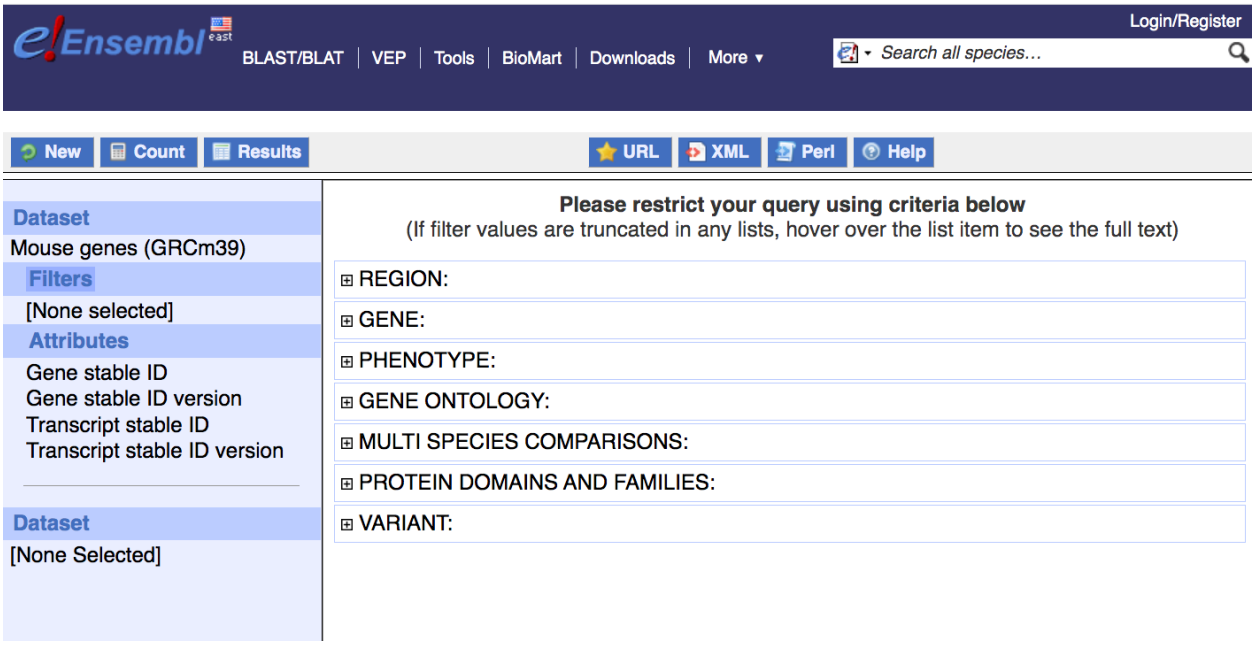
- **Step 1** Choose “Ensembl Genes 110” from “CHOOSE DATABASE”
- **Step 2** Choose “Mouse genes (GRCm39)” from “CHOOSE DATASET”

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

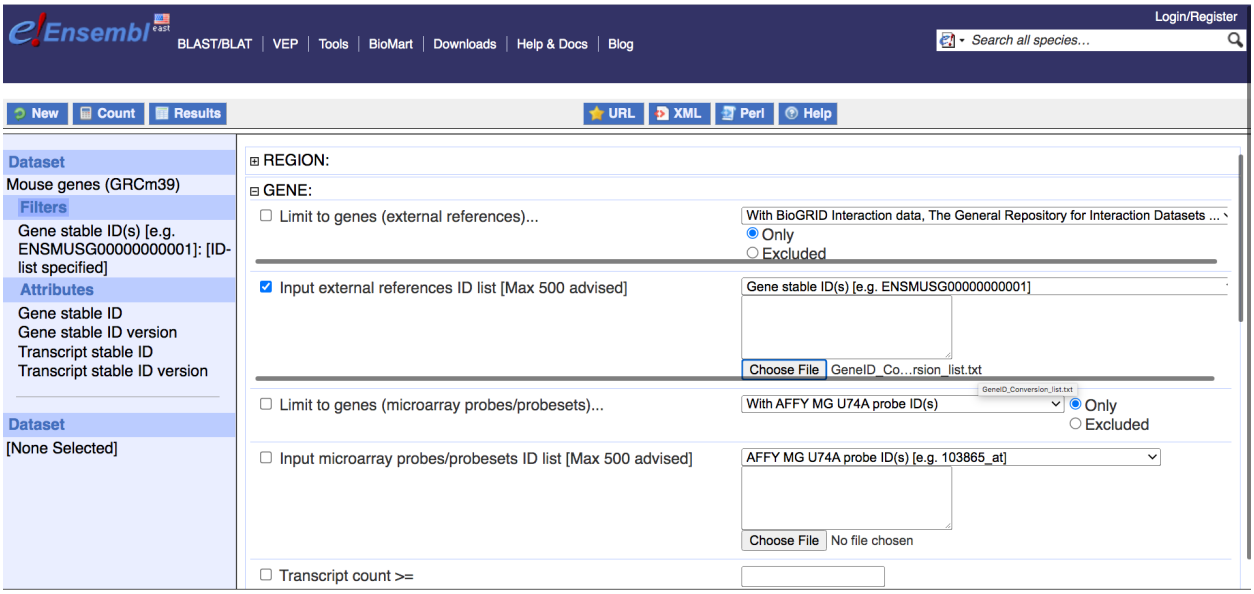


1.3.3 Upload to BioMart

- **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.

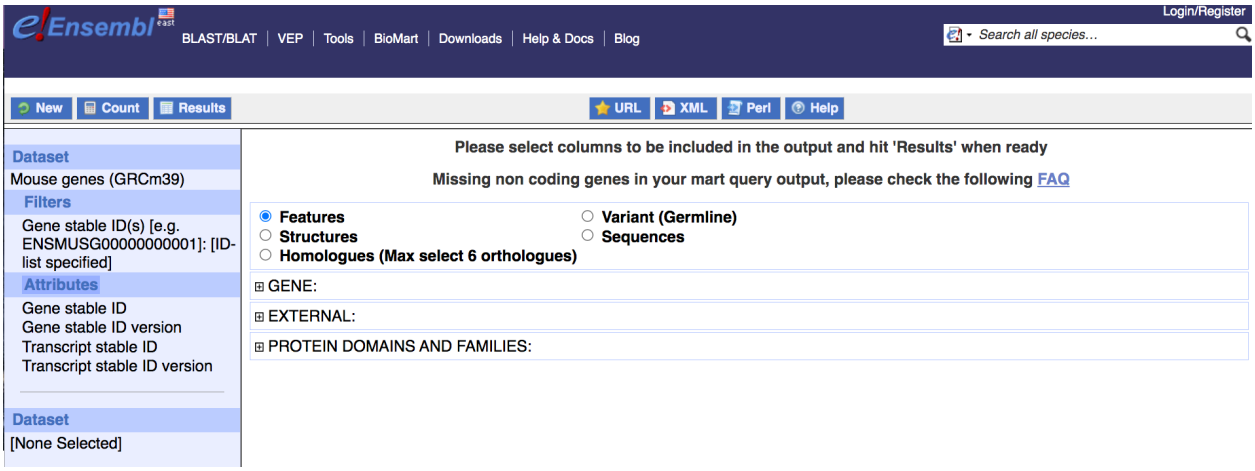


- **Step 3**
  - Clicking on "Gene" will expand and will show option to upload the file.
  - Click "Choose File" and browse to the gene list text file that we downloaded `GeneID_Conversion_list.txt`. Note we are using Ensembl Gene IDs as our input so we do not need to change anything in select identifier dropdown.

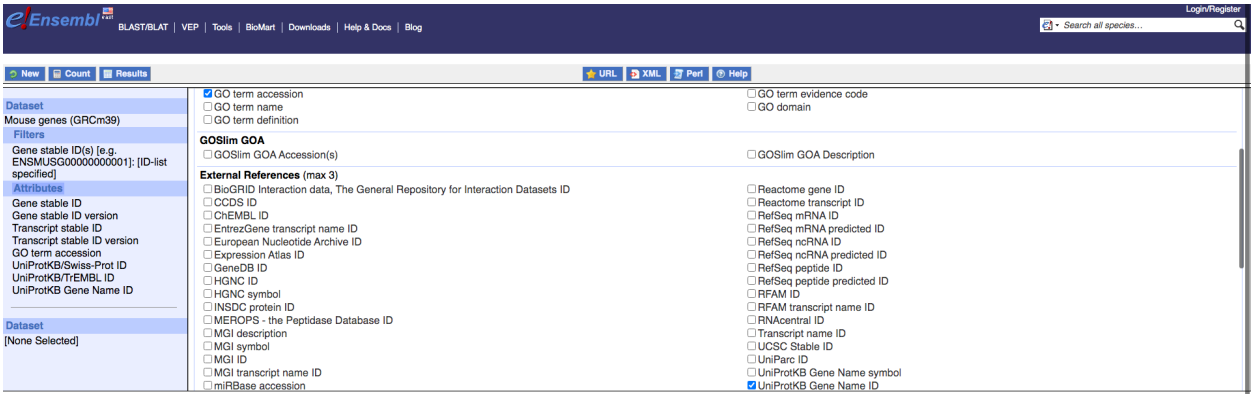


Select Attributes: Click "Attributes" in the left panel. This will open "Attributes" panel in the middle.

- **Step 1:** "Features" is selected by default



- **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)



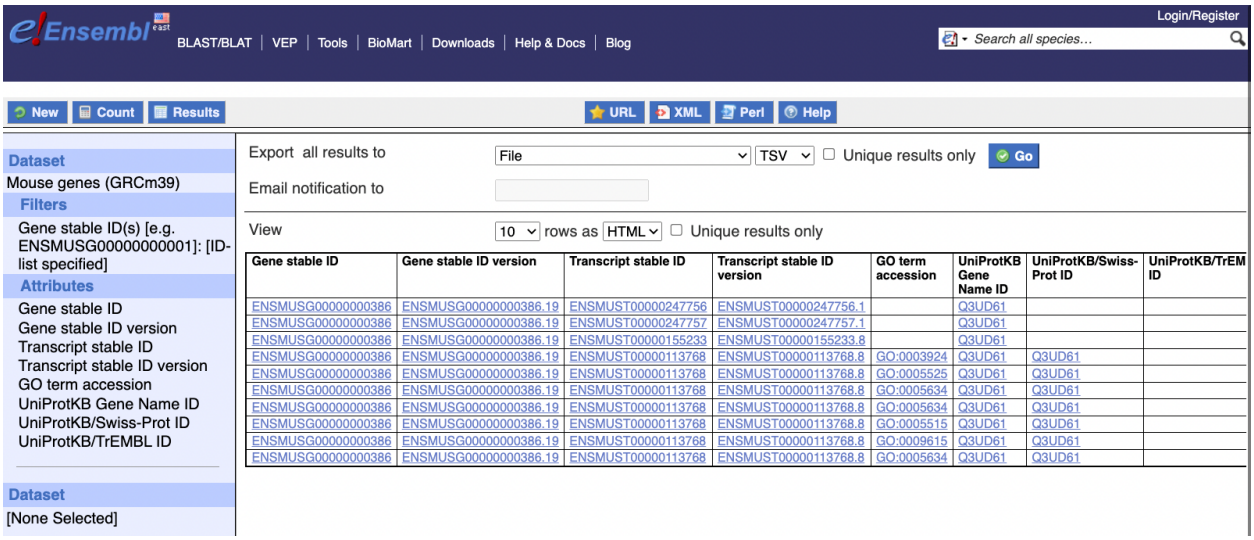
1.3.4 Results from BioMart

- 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.



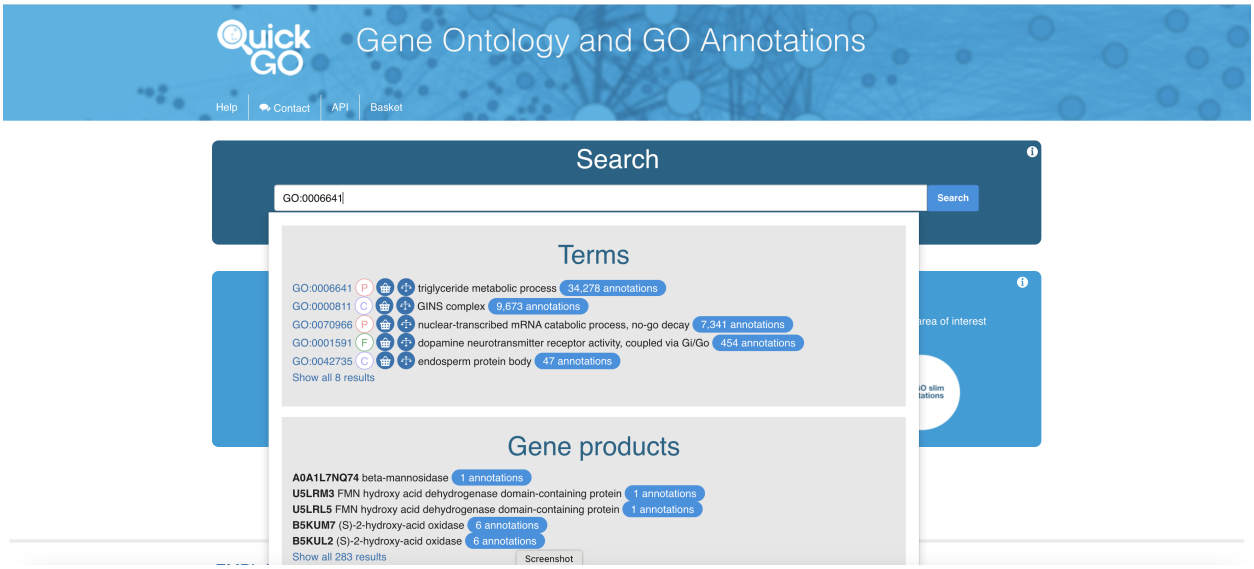
1.4 Using QuickGO

1.4.1 Navigate to QuickGO

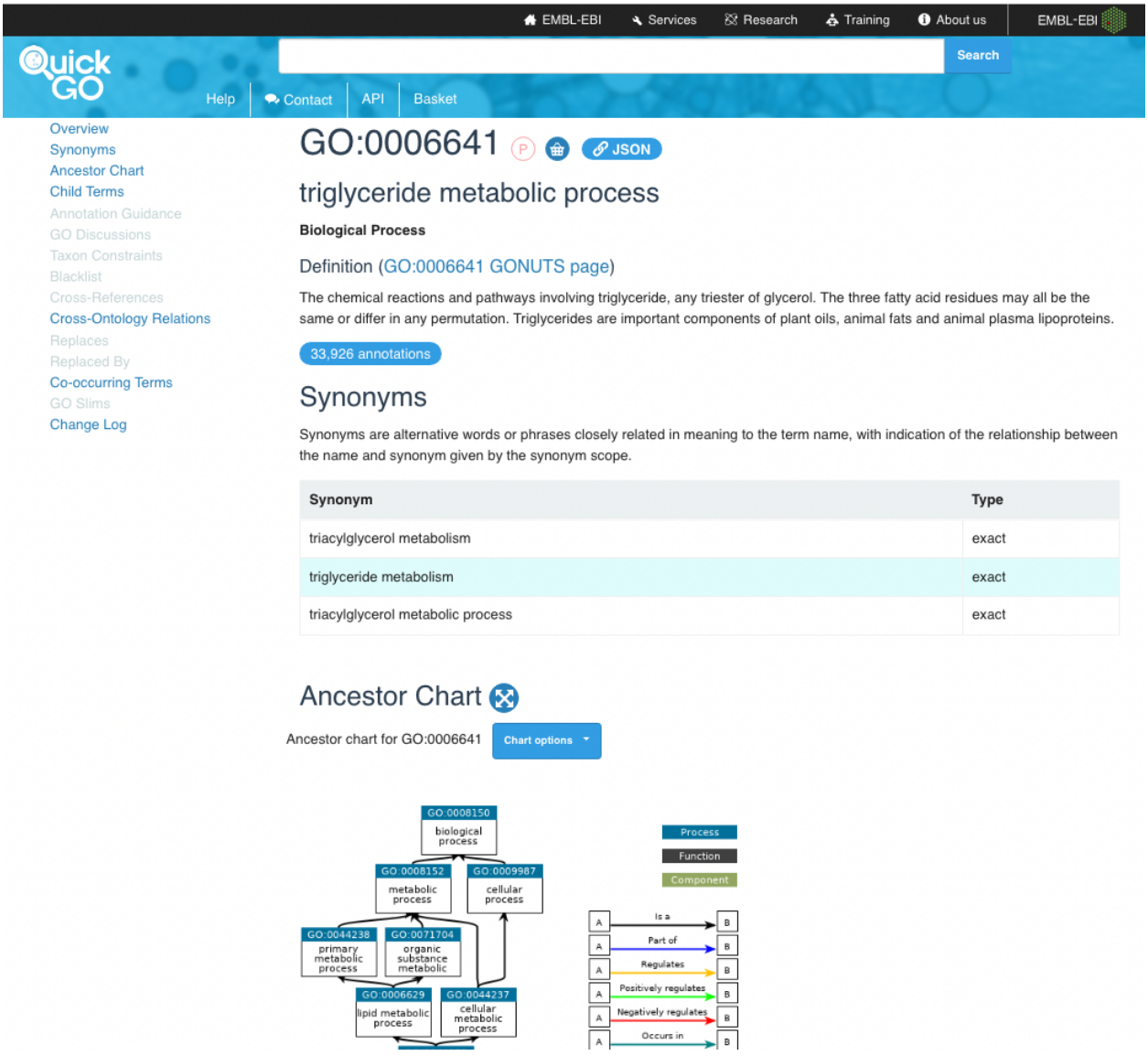
Navigate to QuickGO `https://www.ebi.ac.uk/QuickGO/`

1.4.2 Search using GO ID

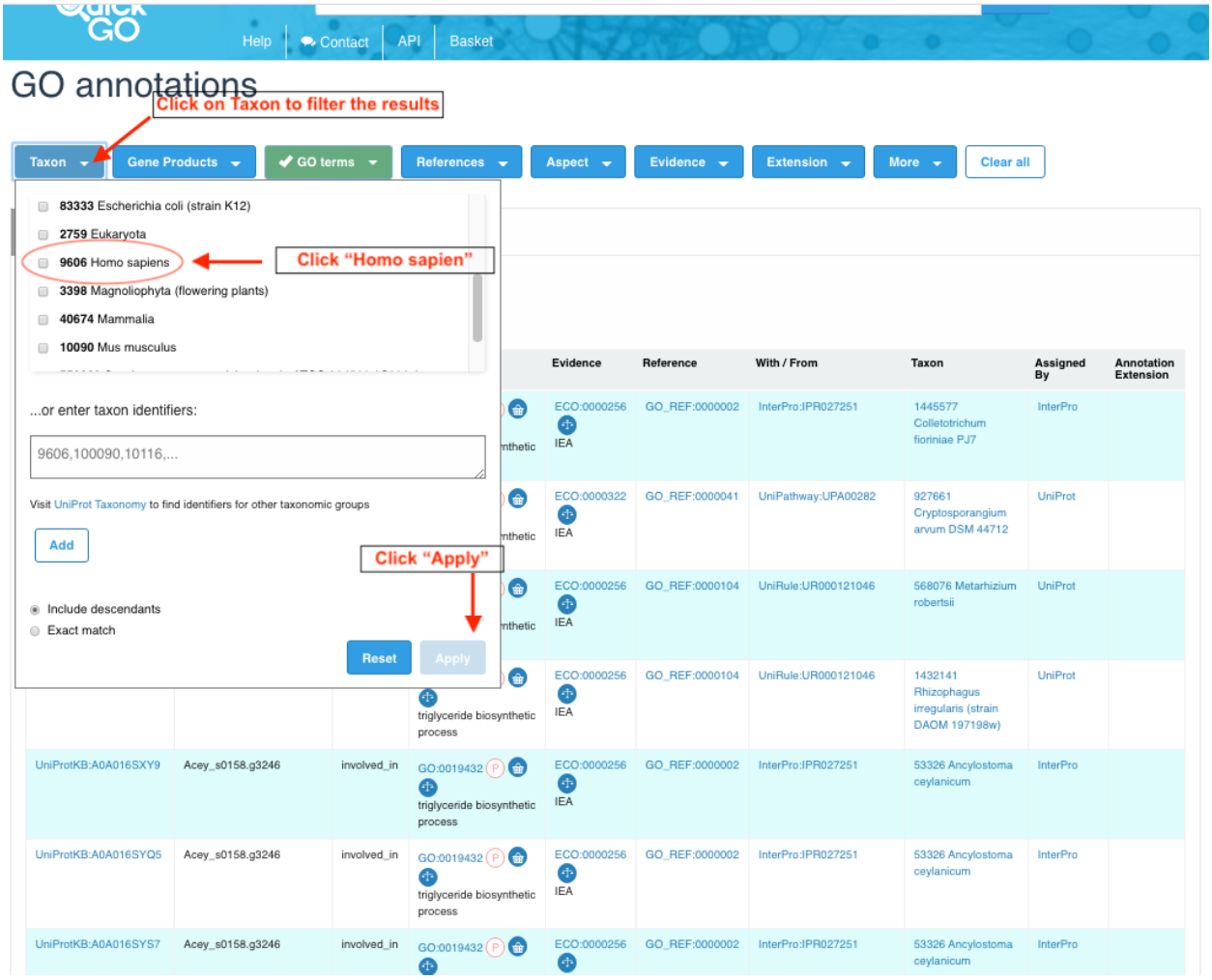
- Type "GO:0006641" in the search box.
- Click on "GO:0006641" from the returned results



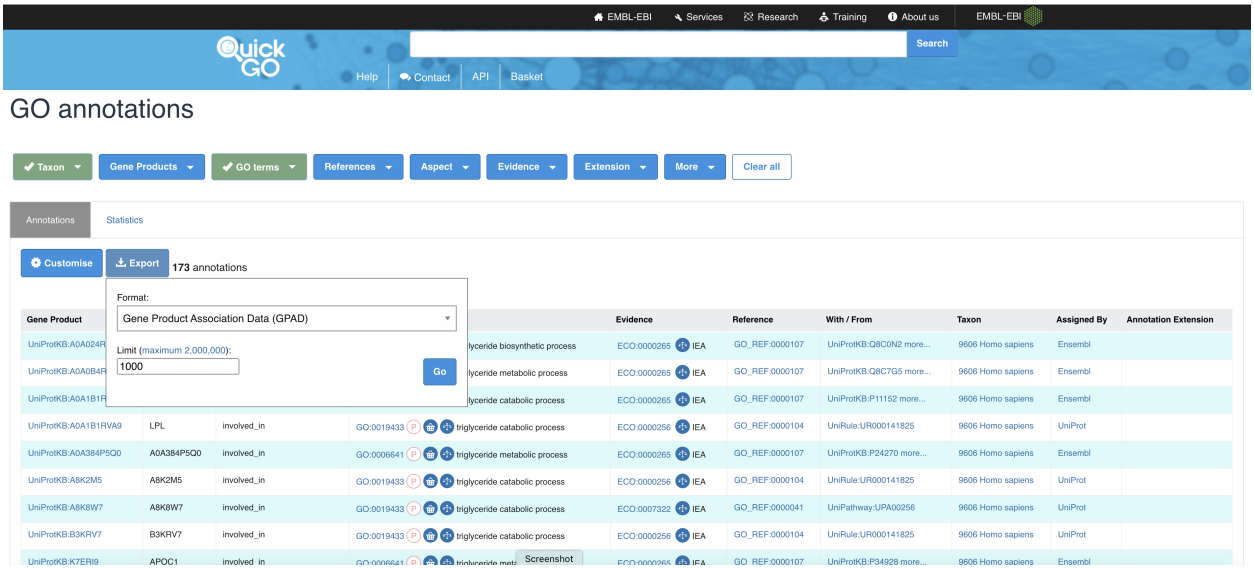
- View GO ID Results
  - Explore results by clicking on left menu for tabs, e.g. “Synonyms”, “Ancestor Chart”, or “Child terms”



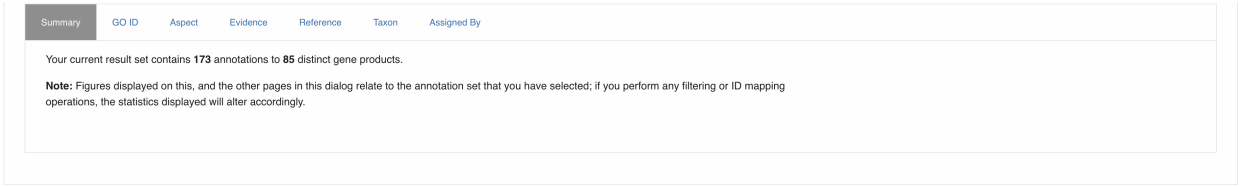
- Filter data
  - Click on “34,278 annotations” as shown in the previous image
  - Explore different tabs like “Taxon”, “Go terms”
  - Click “Taxon” to filter the results
    - Click “Homo sapiens”
    - Click “Apply”



- Click “Export” to download the data

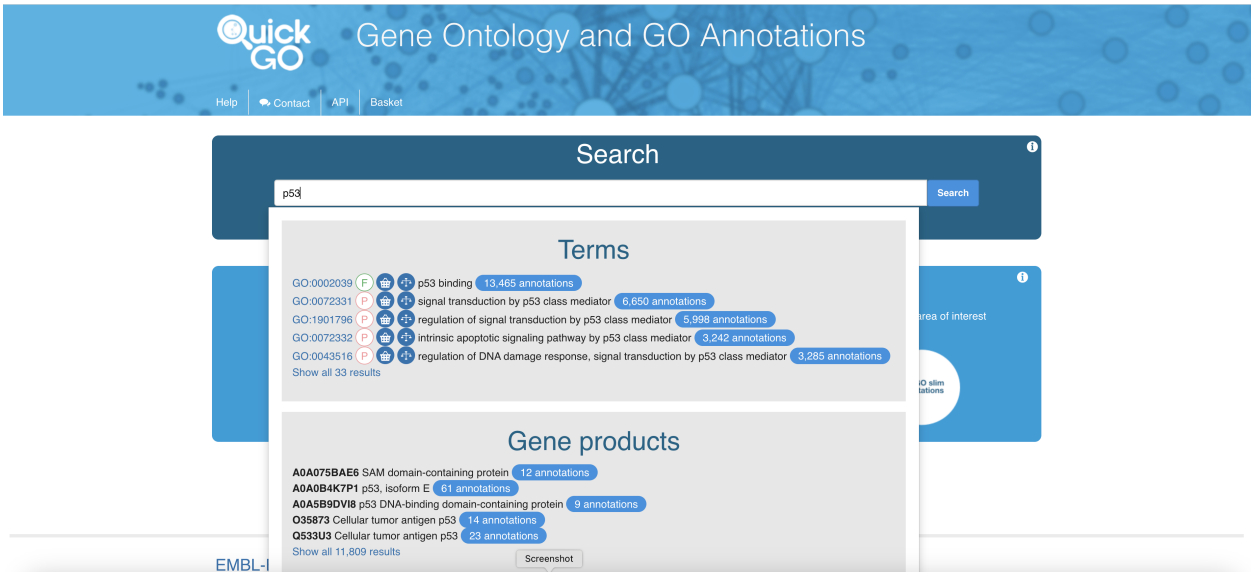


- Click “Statistics” to get Summary



1.4.3 Search using keyword

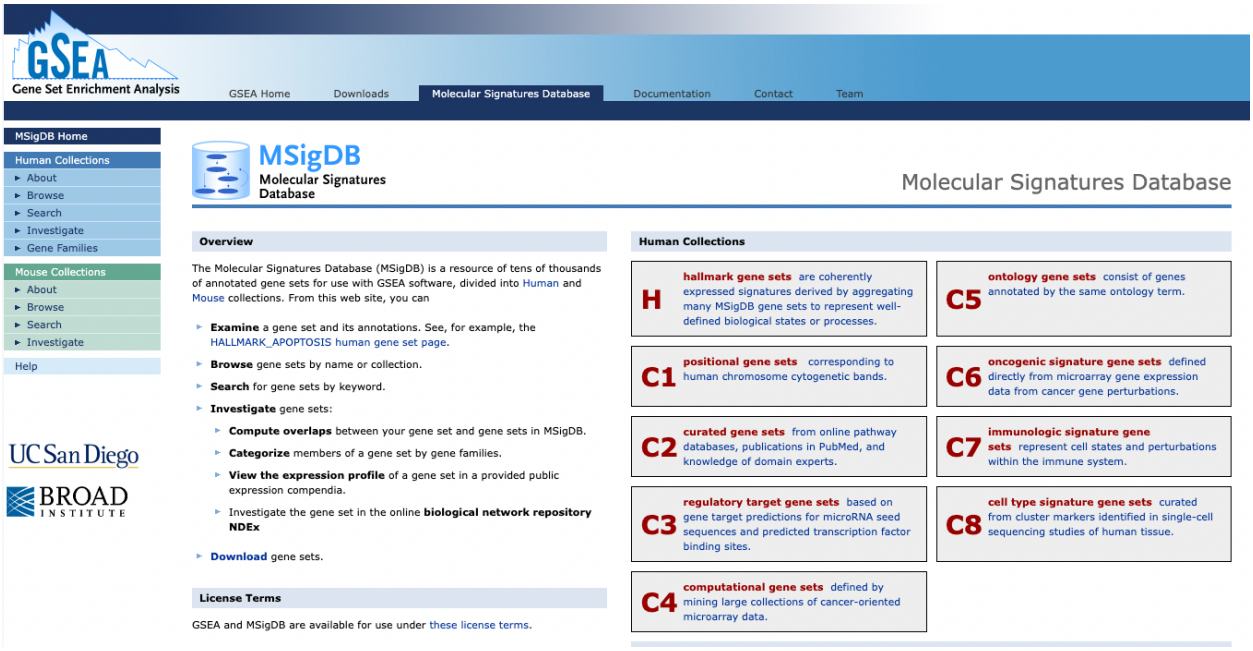
- Navigate to QuickGO `https://www.ebi.ac.uk/QuickGO/`
- Type “p53” in the search box
- Explore the results



1.5 Using MSigDB

1.5.1 Navigate to MSigDB

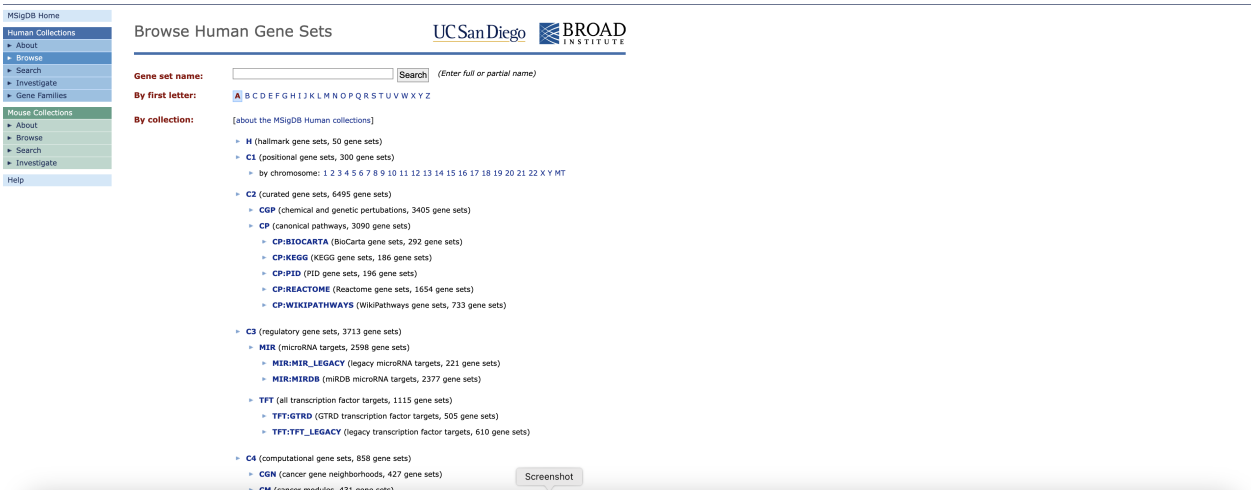
Navigate to MSigDB `https://www.gsea-msigdb.org/gsea/msigdb/index.jsp`



1.5.2 Explore different collections

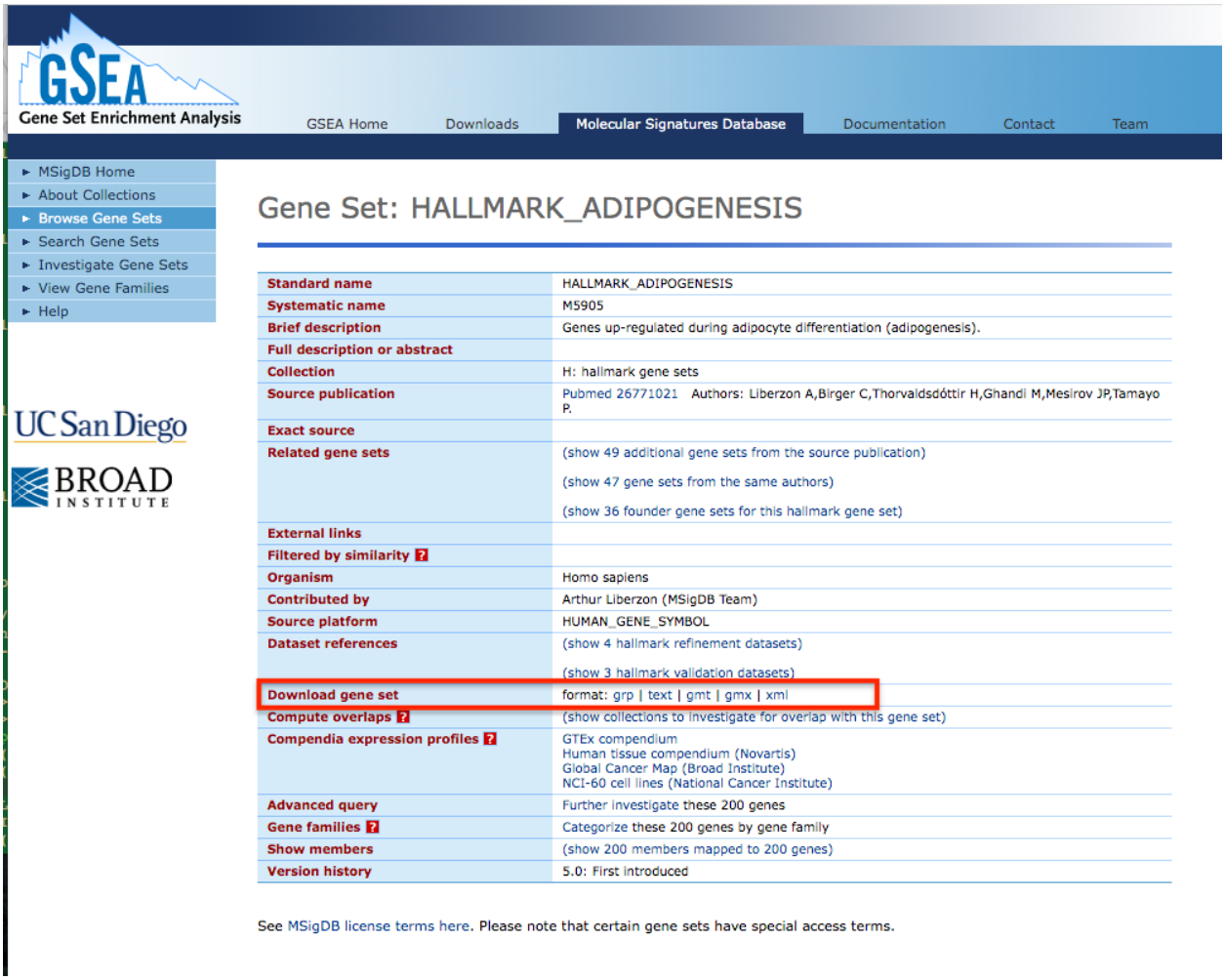
For this exercise we will highlight several different collections.

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

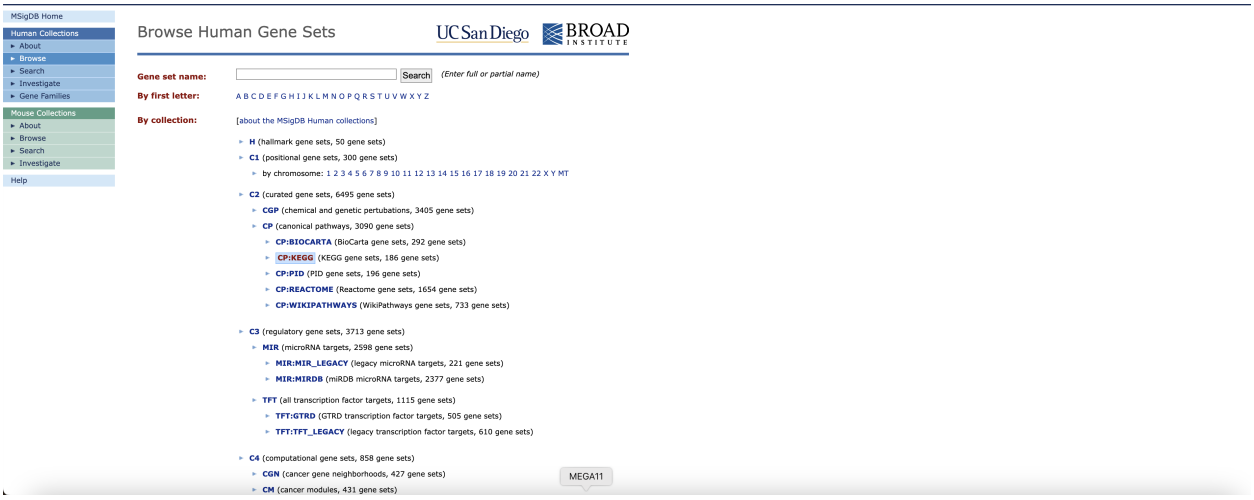


- **Hallmark Collection:** Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. Currently it has total of 50 gene sets.
  - Click "Browse Gene Sets" on the left panel
  - Click "H" to expand the Hallmark collection
    * Scroll 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.

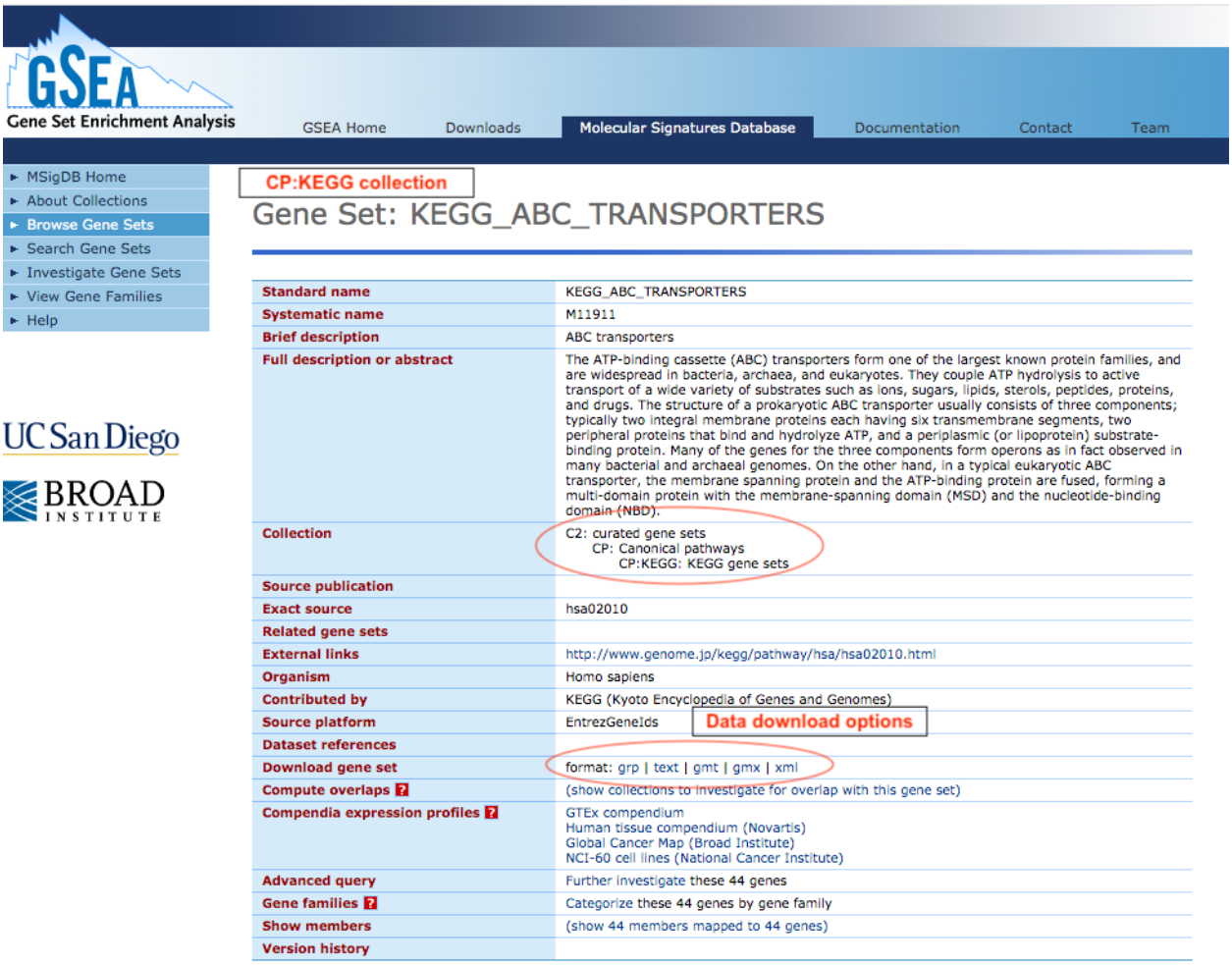
- Review the details for the given gene set collection.



- **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

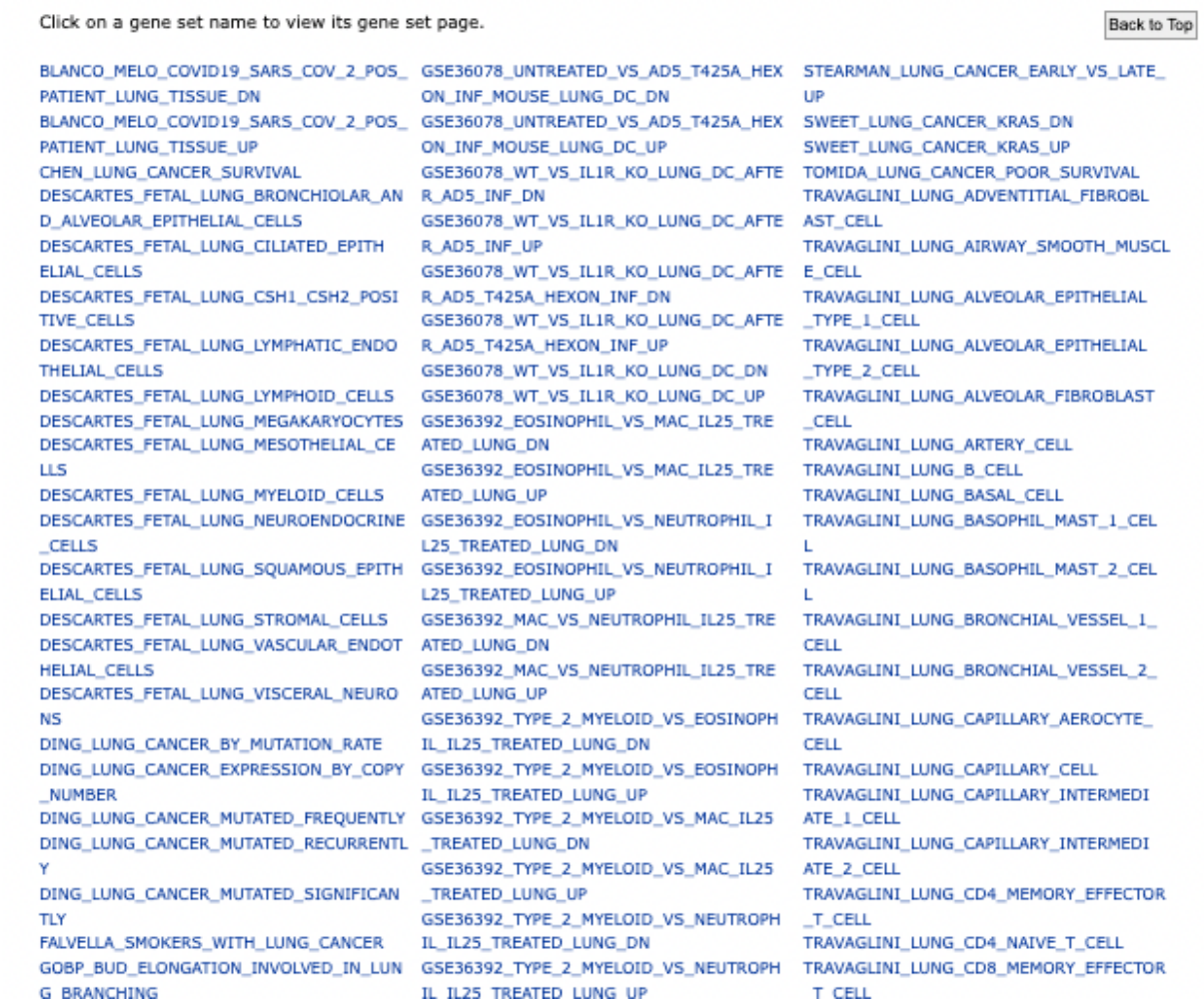


- Review the details for the given gene set collection.



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



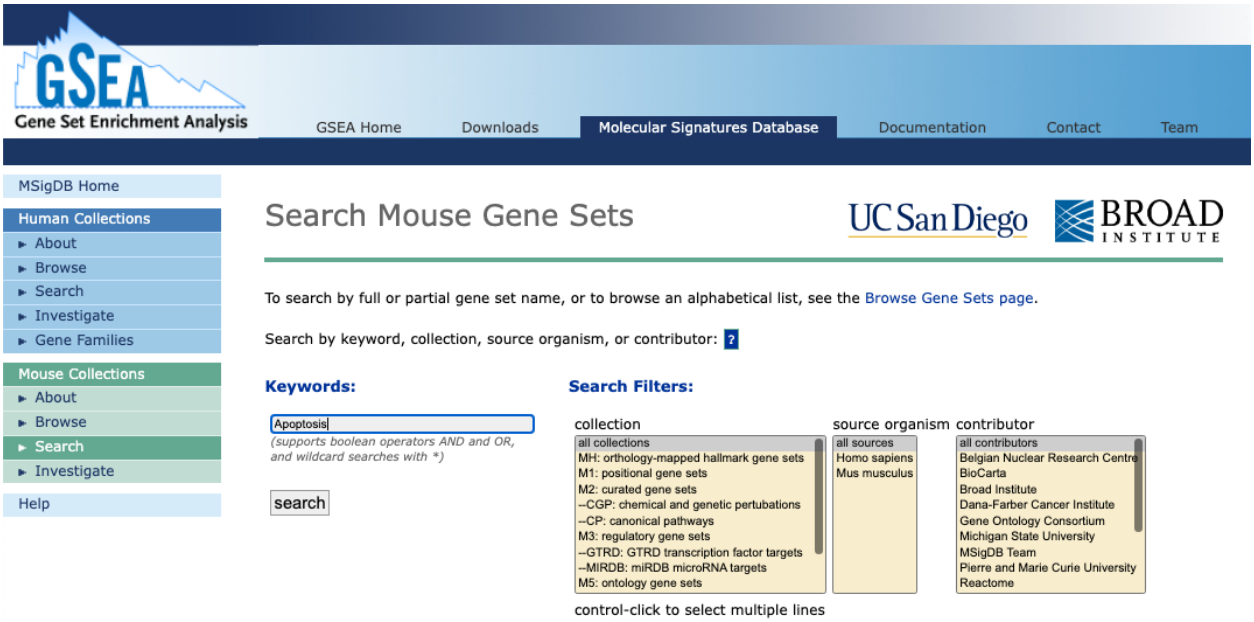
- Review results for “BLANCO_MELO_COVID19_SARS_COV_2_POS_PATIENT_LUNG_TISSUE_DN”
- Download gene set of interest

Human Gene Set: BLANCO_MELO_COVID19_SARS_COV_2_POS_PATIENT_LUNG_TISSUE_DN

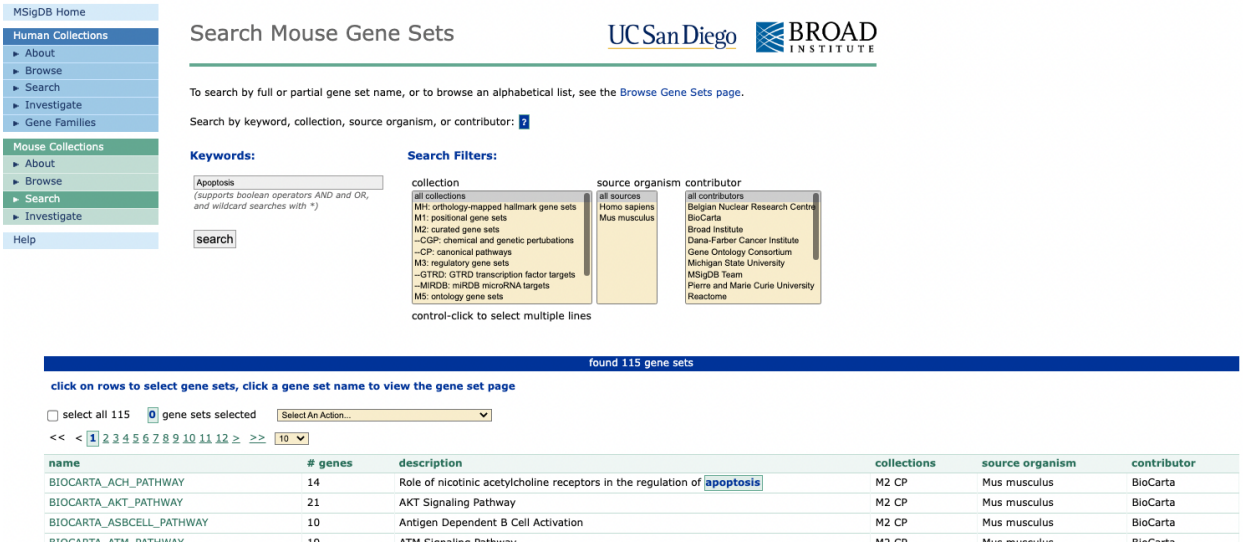
Standard name	BLANCO_MELO_COVID19_SARS_COV_2_POS_PATIENT_LUNG_TISSUE_DN
Systematic name	M34029
Brief description	Genes strongly down-regulated (log2(FC)<-3.58, padj<0.05) in post mortem lung tissue from COVID-19 patients vs uninfected biopsy.
Full description or abstract	Transcriptional profiling of post-mortem lung samples from COVID-19-positive patients (N=2) compared with biopsied healthy lung tissue from uninfected individuals.
Collection	C2: Curated CGP: Chemical and Genetic Perturbations
Source publication	Pubmed 32416070 Authors: Blanco-Melo D,Nilsson-Payant BE,Liu WC,Uhl S,Hoagland D,Møller R,Jordan TX,Oishi K,Panis M,Sachs D,Wang TT,Schwartz RE,Lim JK,Albrecht RA,tenOever BR
Exact source	Table S4. Differential Gene Expression Analysis of COVID-19 Patients, Related to Figure 4. log2(FC)<-3.58, padj<0.05
Related gene sets	(show 29 additional gene sets from the source publication) (show 22 gene sets from the same authors)
External links	
Filtered by similarity	
Organism	Homo sapiens
Contributed by	Anthony Castanza (MSigDB Team)
Source platform	HUMAN_GENE_SYMBOL
Dataset references	(show 1 datasets)
Download gene set	format: grp | gmt | xml | json | TSV metadata
Compute overlaps	(show collections to investigate for overlap with this gene set)
Compendia expression profiles	GTEx compendium Human tissue compendium (Novartis) Global Cancer Map (Broad Institute) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 175 genes
Gene families	Categorize these 175 genes by gene family
Show members	(show 175 members mapped to 175 genes)
Version history	

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

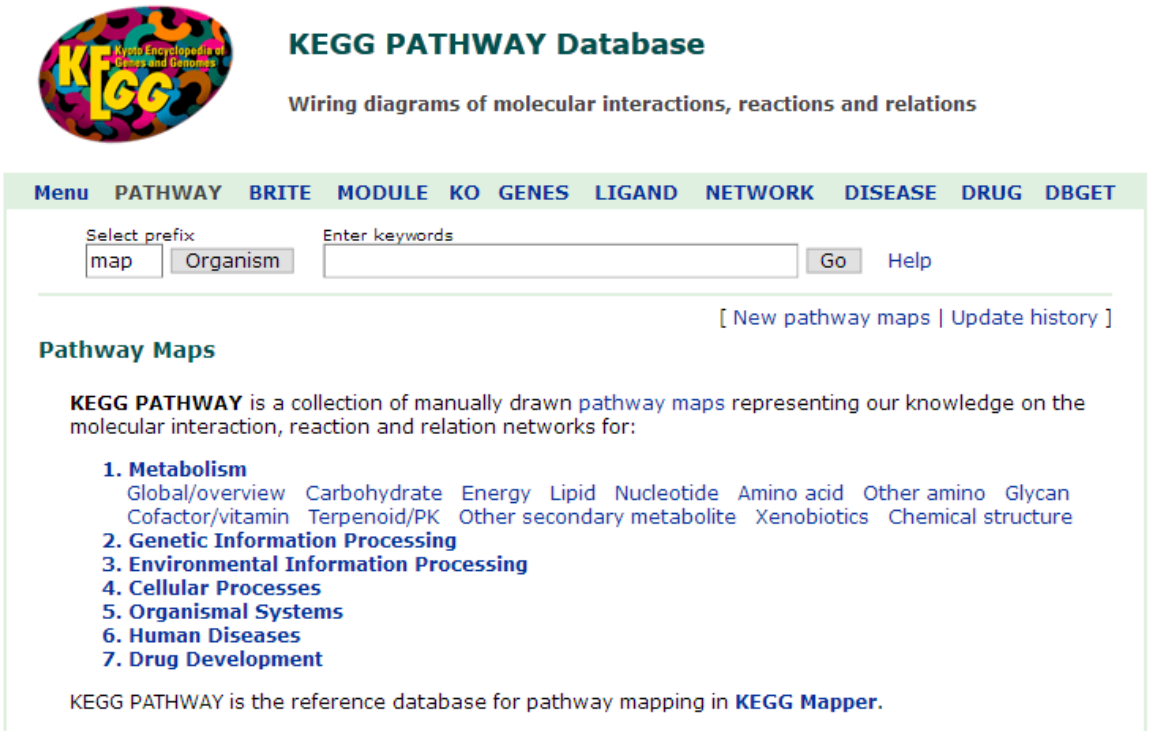


Review results

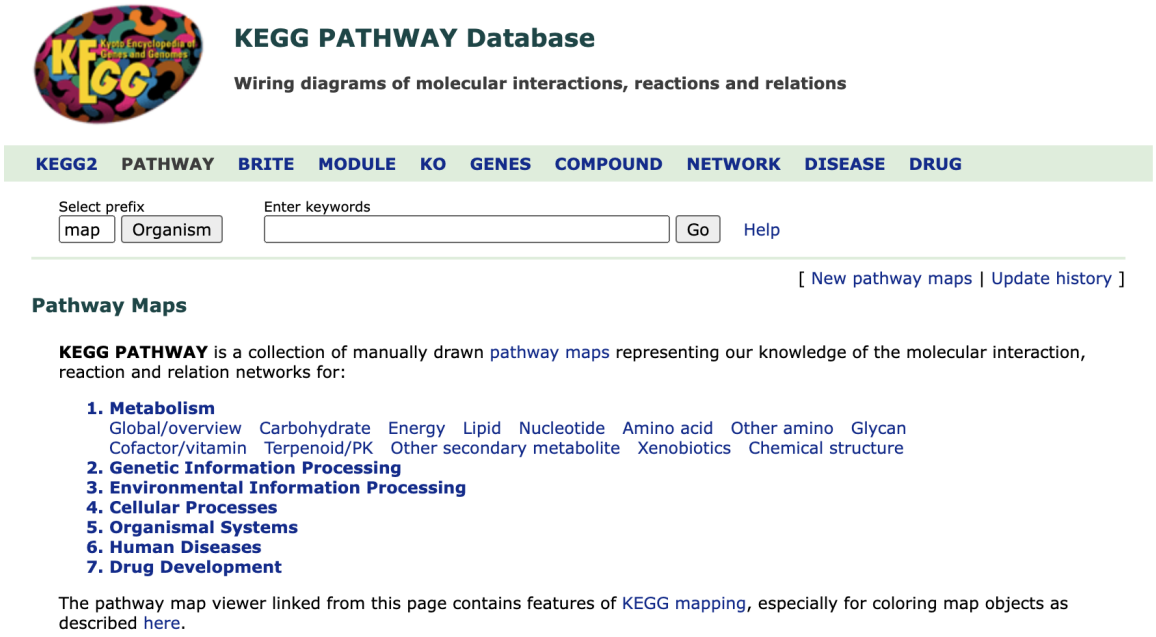


1.6 Browse KEGG

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*.

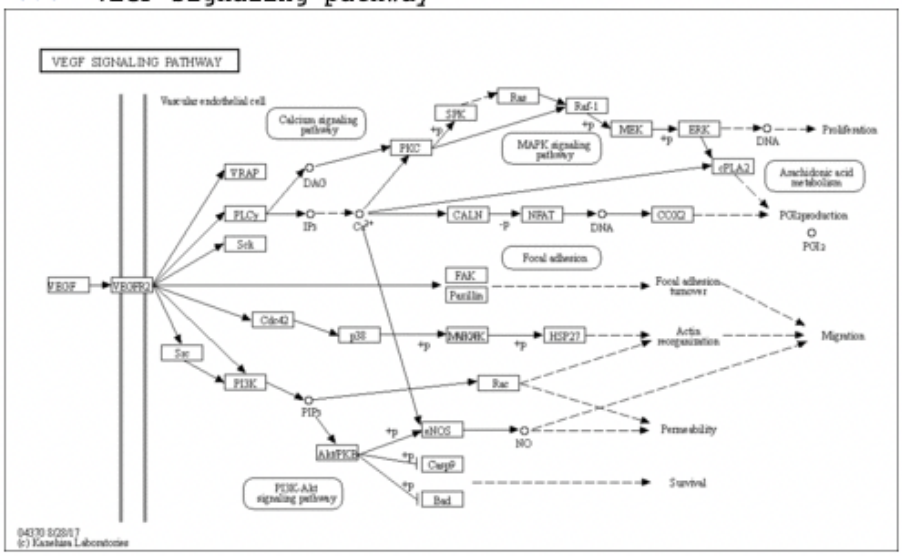


1.6.1 Pathway entry view

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

PATHWAY: map04370

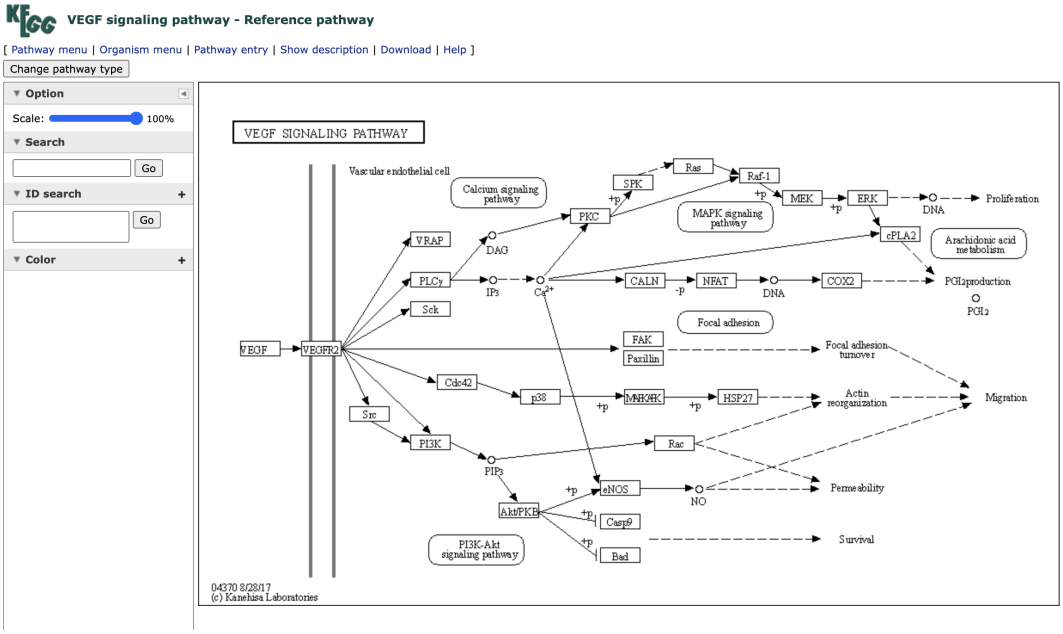
Help

Entry	map04370 Pathway
Name	VEGF signaling pathway
Description	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.
Class	Environmental Information Processing; Signal transduction BRITE hierarchy
Pathway map	map04370 VEGF signaling pathway  Ortholog table
Other DBs	GO: 0048010
Reference	PMID:13678960
Authors	Cross MJ, Dixelius J, Matsumoto T, Claesson-Welsh L.
Title	VEGF-receptor signal transduction.
Journal	Trends Biochem Sci 28:488-94 (2003) DOI:10.1016/S0968-0004(03)00193-2

- **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

1. 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.

KEGG **ORTHOLOGY: K05448** Help

Entry	K05448 KO
Symbol	VEGFA
Name	vascular endothelial growth factor A
Pathway	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
Disease	H01456 Diabetic nephropathy H01457 Diabetic retinopathy H01459 Diabetic neuropathy H01529 Avascular necrosis of femoral head H01709 Glucocorticoid-induced osteonecrosis H02559 Microvascular complications of diabetes
Brite	KEGG Orthology (KO) [BR:ko00001] 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.

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%

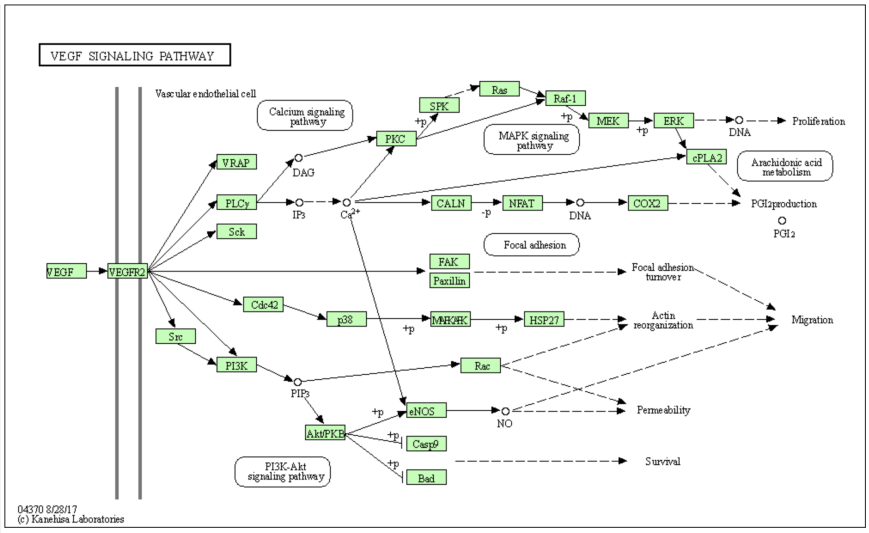
Search

Go

ID search

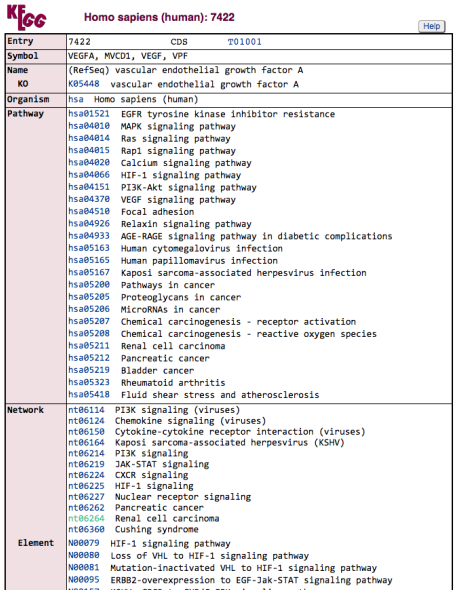
Go

Color



1.6.5 Gene Details Page

1. 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	hsa01521 EGFR tyrosine kinase inhibitor resistance hsa04010 MAPK signaling pathway hsa04014 Ras signaling pathway hsa04015 Rap1 signaling pathway hsa04020 Calcium signaling pathway hsa04066 HIF-1 signaling pathway hsa04151 PI3K-Akt signaling pathway hsa04370 VEGF signaling pathway hsa04510 Focal adhesion hsa04926 Relaxin signaling pathway hsa04933 AGE-RAGE signaling pathway in diabetic complications hsa05163 Human cytomegalovirus infection hsa05165 Human papillomavirus infection hsa05167 Kaposi sarcoma-associated herpesvirus infection hsa05200 Pathways in cancer hsa05205 Proteoglycans in cancer hsa05206 MicroRNAs in cancer hsa05207 Chemical carcinogenesis - receptor activation hsa05208 Chemical carcinogenesis - reactive oxygen species hsa05211 Renal cell carcinoma hsa05212 Pancreatic cancer hsa05219 Bladder cancer hsa05323 Rheumatoid arthritis hsa05418 Fluid shear stress and atherosclerosis
Network	nt06114 PI3K signaling (viruses) nt06124 Chemokine signaling (viruses) nt06150 Cytokine-cytokine receptor interaction (viruses) nt06164 Kaposi sarcoma-associated herpesvirus (KSHV) nt06214 PI3K signaling nt06219 JAK-STAT signaling nt06224 CXCR signaling nt06225 HIF-1 signaling nt06227 Nuclear receptor signaling nt06262 Pancreatic cancer nt06264 Renal cell carcinoma nt06360 Cushing syndrome
Element	N00079 HIF-1 signaling pathway N00080 Loss of VHL to HIF-1 signaling pathway N00081 Mutation-inactivated VHL to HIF-1 signaling pathway N00095 ERBB2-overexpression to EGF-Jak-STAT signaling pathway

- **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

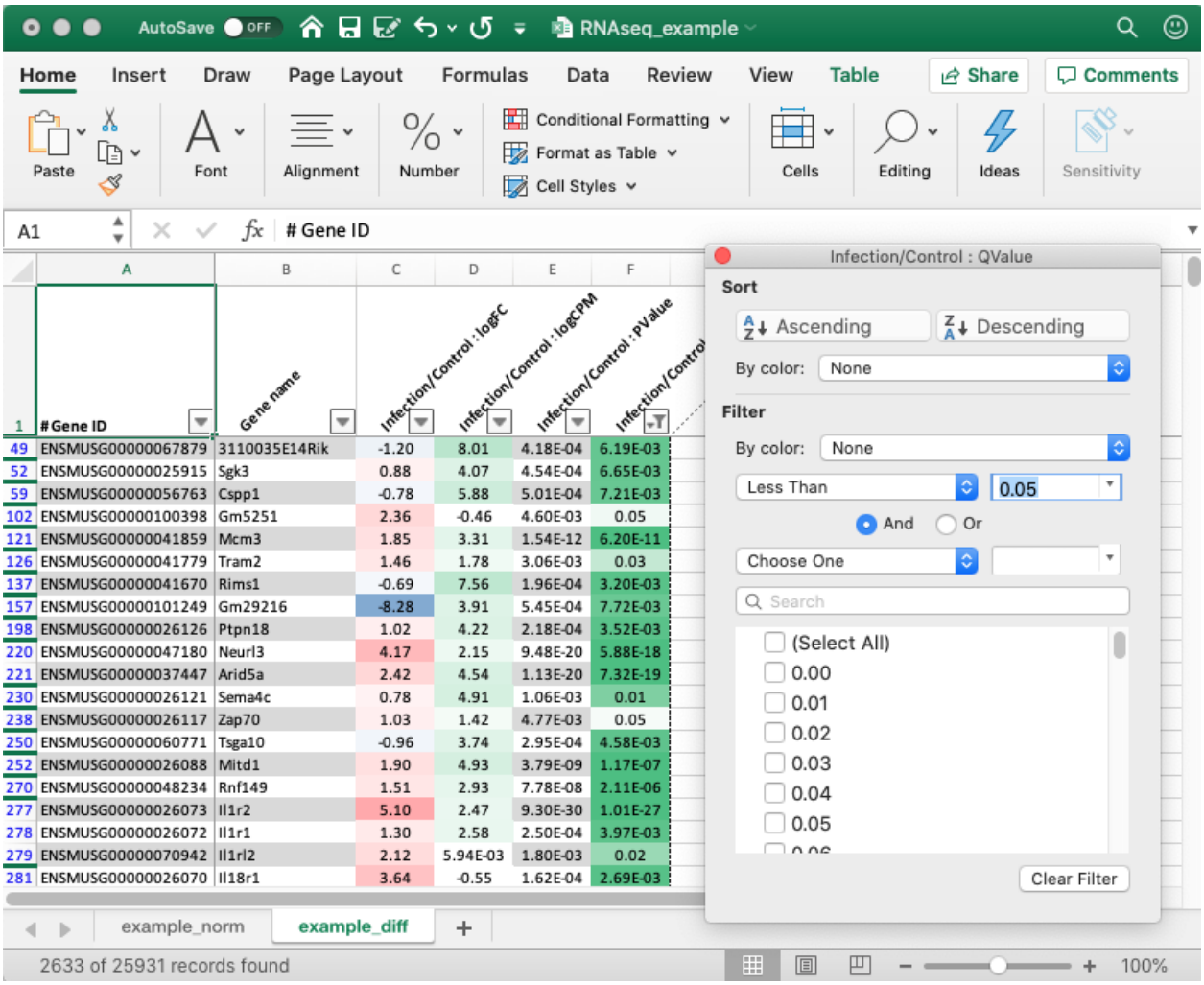
1.7.1 Prepare input data

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

`https://wd.cri.uic.edu/pathway/RNAseq_example.xlsx`

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



1	# Gene ID	Gene name	Infection/Control : logFC	Infection/Control : logCPM	Infection/Control : PValue	Infection/Control
49	ENSMUSG00000067879	3110035E14Rik	-1.20	8.01	4.18E-04	6.19E-03
52	ENSMUSG00000025915	Sgk3	0.88	4.07	4.54E-04	6.65E-03
59	ENSMUSG00000056763	Cspp1	-0.78	5.88	5.01E-04	7.21E-03
102	ENSMUSG00000100398	Gm5251	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`

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:

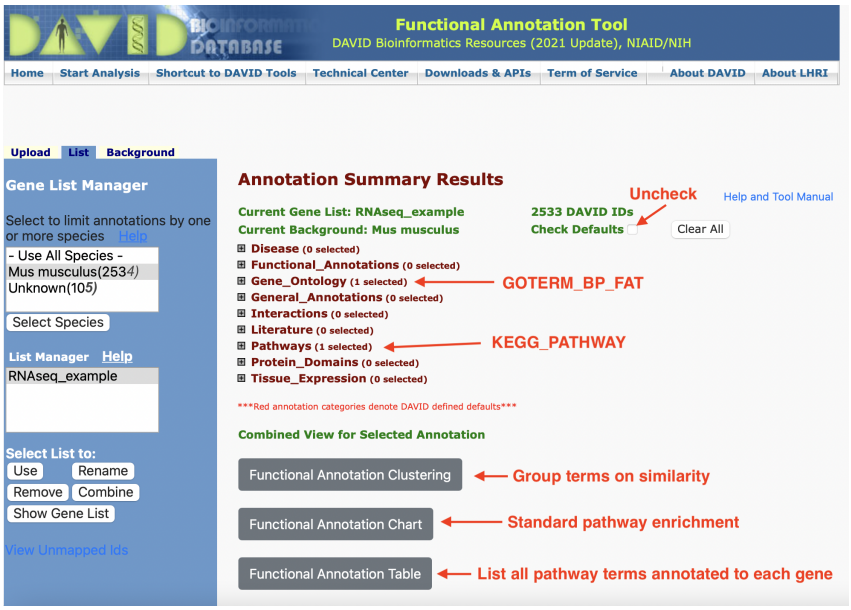
- Click "Browse..." and browse to the gene list text file that downloaded
- Choose "ENSEMBL_GENE_ID" from the Select Identifier drop-down
- Select the "Gene List" radio button
- Submit List to DAVID.

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.

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:

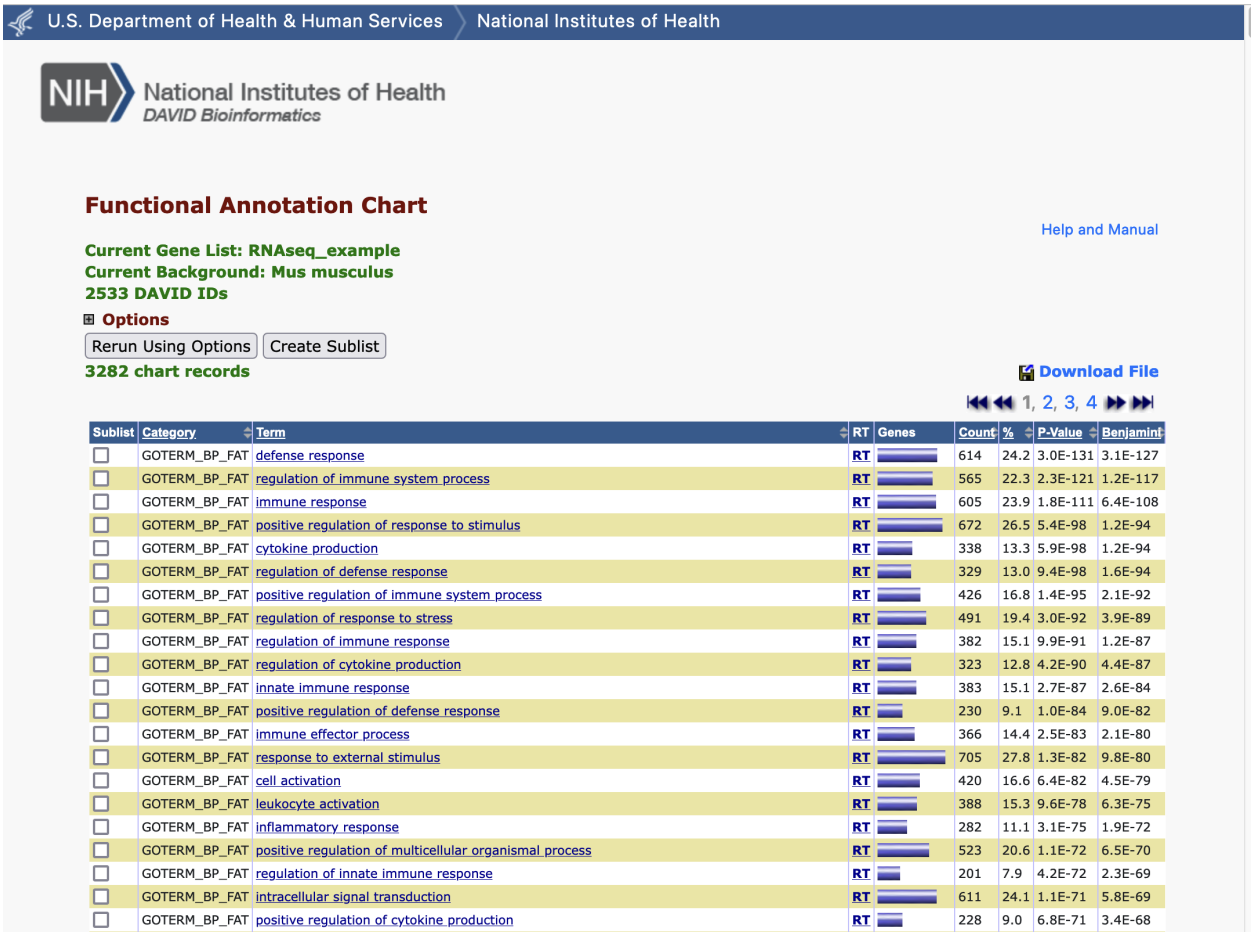
- GO MF has types of molecular functions, such as cytokine binding or kinase activity, which can be useful if you're trying to characterize classes of molecules.
- GO CC has localization of molecules in the cell. This can be useful if, for example, you were looking for cell surface markers in a set of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

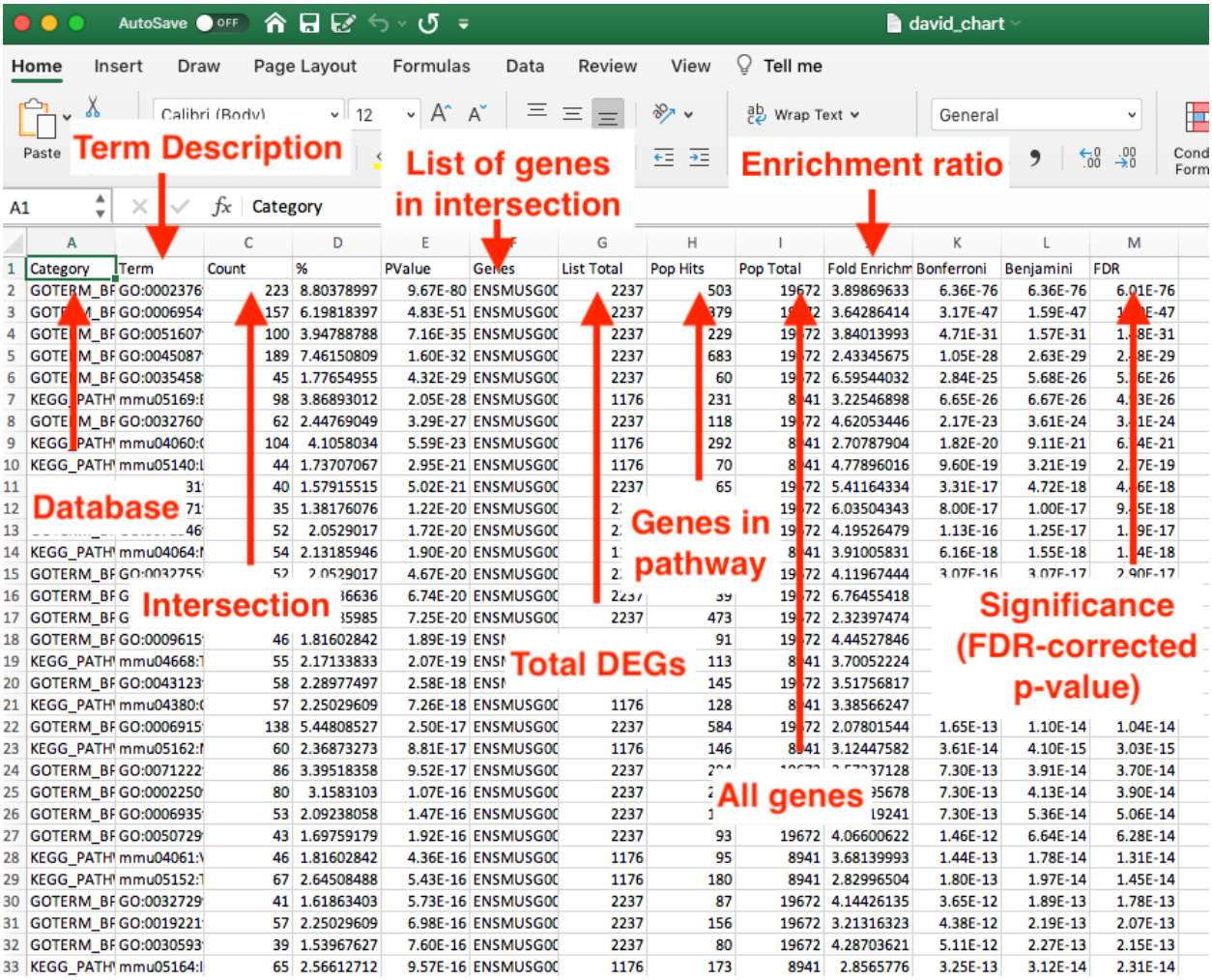
- NOTE: by default DAVID only downloads terms with at an intersection of at least 2 molecules and a nominal p-value < 0.1. If you're running a pathway comparison, as we will in the afternoon, you may want to relax these thresholds for the purposes the comparison. You can do this in the *Rerun Using Options*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



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	B	C	D	E	F	G	H	I	J	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	ENSMUSG0(	2237	503	19672	3.89869633	6.36E-76	6.36E-76	6.01E-76
3	GOTERM_BF	GO:0006954	157	6.19818397	4.83E-51	ENSMUSG0(	2237	379	19672	3.64286414	3.17E-47	1.59E-47	1.[illegible]E-47
4	GOTERM_BF	GO:0051607	100	3.94788788	7.16E-35	ENSMUSG0(	2237	229	19672	3.84013993	4.71E-31	1.57E-31	1.[illegible]8E-31
5	GOTERM_BF	GO:0045087	189	7.46150809	1.60E-32	ENSMUSG0(	2237	683	19672	2.43345675	1.05E-28	2.63E-29	2.[illegible]8E-29
6	GOTERM_BF	GO:0035458	45	1.77654955	4.32E-29	ENSMUSG0(	2237	60	19672	6.59544032	2.84E-25	5.68E-26	5.[illegible]6E-26
7	KEGG_PATH	mmu05169:E	98	3.86893012	2.05E-28	ENSMUSG0(	1176	231	8941	3.22546898	6.65E-26	6.67E-26	4.[illegible]3E-26
8	GOTERM_BF	GO:0032760	62	2.44769049	3.29E-27	ENSMUSG0(	2237	118	19672	4.62053446	2.17E-23	3.61E-24	3.[illegible]1E-24
9	KEGG_PATH	mmu04060:(	104	4.1058034	5.59E-23	ENSMUSG0(	1176	292	8941	2.70787904	1.82E-20	9.11E-21	6.[illegible]4E-21
10	KEGG_PATH	mmu05140:I	44	1.73707067	2.95E-21	ENSMUSG0(	1176	70	8941	4.77896016	9.60E-19	3.21E-19	2.[illegible]7E-19
11		31	40	1.57915515	5.02E-21	ENSMUSG0(	2237	65	19672	5.41164334	3.31E-17	4.72E-18	4.[illegible]6E-18
12		71	35	1.38176076	1.22E-20	ENSMUSG0(	2[illegible]	[illegible]	19672	6.03504343	8.00E-17	1.00E-17	9.[illegible]5E-18
13		46	52	2.0529017	1.72E-20	ENSMUSG0(	2[illegible]	[illegible]	19672	4.19526479	1.13E-16	1.25E-17	1.[illegible]9E-17
14	KEGG_PATH	mmu04064:I	54	2.13185946	1.90E-20	ENSMUSG0(	1[illegible]	[illegible]	8941	3.91005831	6.16E-18	1.55E-18	1.[illegible]4E-18
15	GOTERM_BF	GO:0032755	52	2.0529017	4.67E-20	ENSMUSG0(	2[illegible]	[illegible]	19672	4.11967444	3.07E-16	3.07E-17	2.90E-17
16	GOTERM_BF	G	[illegible]6636		6.74E-20	ENSMUSG0(	2237	39	19672	6.76455418			
17	GOTERM_BF	G	[illegible]5985		7.25E-20	ENSMUSG0(	2237	473	19672	2.32397474			
18	GOTERM_BF	GO:0009615	46	1.81602842	1.89E-19	ENSI		91	19672	4.44527846			
19	KEGG_PATH	mmu04668:T	55	2.17133833	2.07E-19	ENSI		113	8941	3.70052224			
20	GOTERM_BF	GO:0043123	58	2.28977497	2.58E-18	ENSI		145	19672	3.51756817			
21	KEGG_PATH	mmu04380:(	57	2.25029609	7.26E-18	ENSMUSG0(	1176	128	8941	3.38566247			
22	GOTERM_BF	GO:0006915	138	5.44808527	2.50E-17	ENSMUSG0(	2237	584	19672	2.07801544	1.65E-13	1.10E-14	1.04E-14
23	KEGG_PATH	mmu05162:I	60	2.36873273	8.81E-17	ENSMUSG0(	1176	146	8941	3.12447582	3.61E-14	4.10E-15	3.03E-15
24	GOTERM_BF	GO:0071222	86	3.39518358	9.52E-17	ENSMUSG0(	2237	[illegible]	[illegible]	[illegible]37128	7.30E-13	3.91E-14	3.70E-14
25	GOTERM_BF	GO:0002250	80	3.1583103	1.07E-16	ENSMUSG0(	2237	[illegible]	[illegible]	[illegible]5678	7.30E-13	4.13E-14	3.90E-14
26	GOTERM_BF	GO:0006935	53	2.09238058	1.47E-16	ENSMUSG0(	2237	1	[illegible]9241		7.30E-13	5.36E-14	5.06E-14
27	GOTERM_BF	GO:0050729	43	1.69759179	1.92E-16	ENSMUSG0(	2237	93	19672	4.06600622	1.46E-12	6.64E-14	6.28E-14
28	KEGG_PATH	mmu04061:V	46	1.81602842	4.36E-16	ENSMUSG0(	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	ENSMUSG0(	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	ENSMUSG0(	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	ENSMUSG0(	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	ENSMUSG0(	2237	80	19672	4.28703621	5.11E-12	2.27E-13	2.15E-13
33	KEGG_PATH	mmu05164:I	65	2.56612712	9.57E-16	ENSMUSG0(	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) **Download File**

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

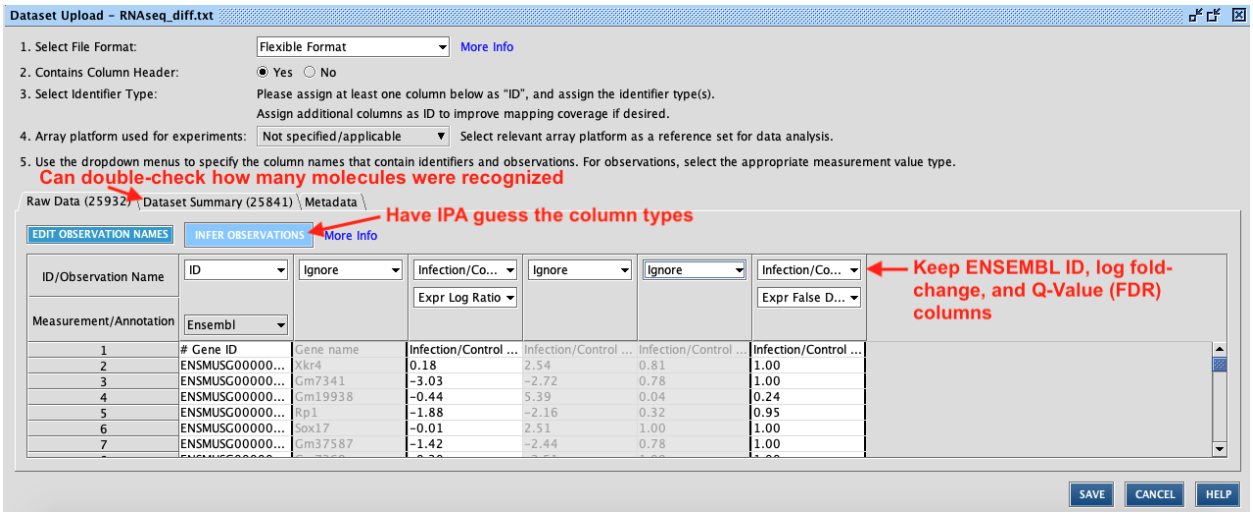
1.8.1 Upload data to IPA

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



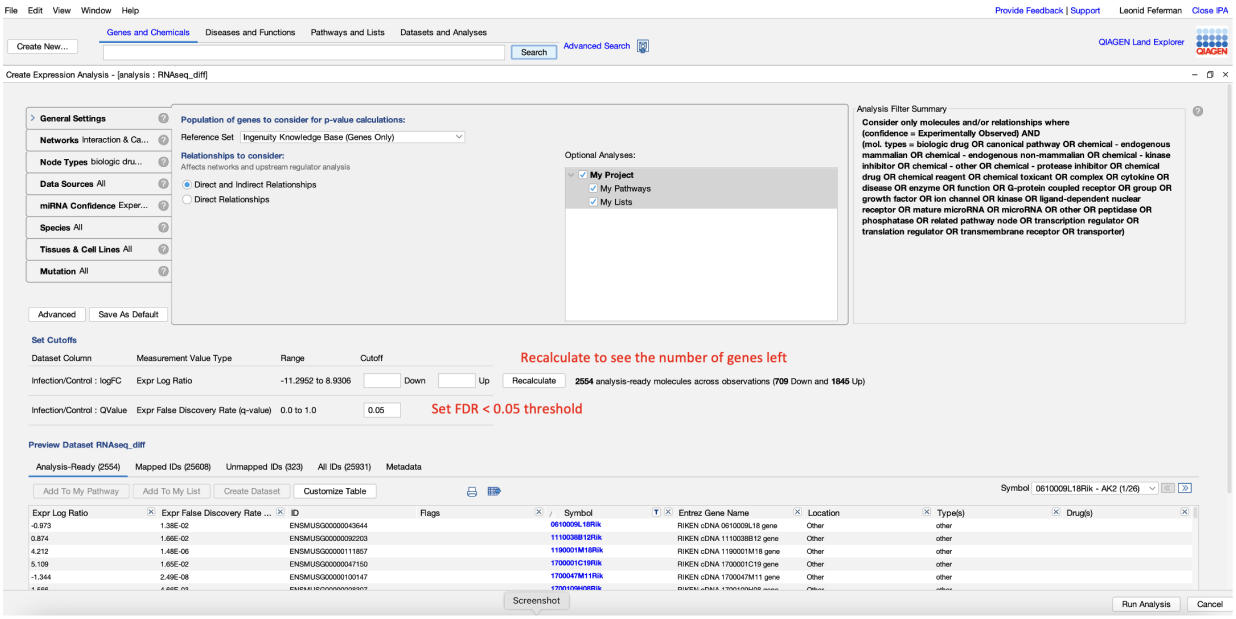
After loading, IPA opens a annotated dataset view, which lets you know how IPA is interpreting all of the genes. We can close this window.

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.



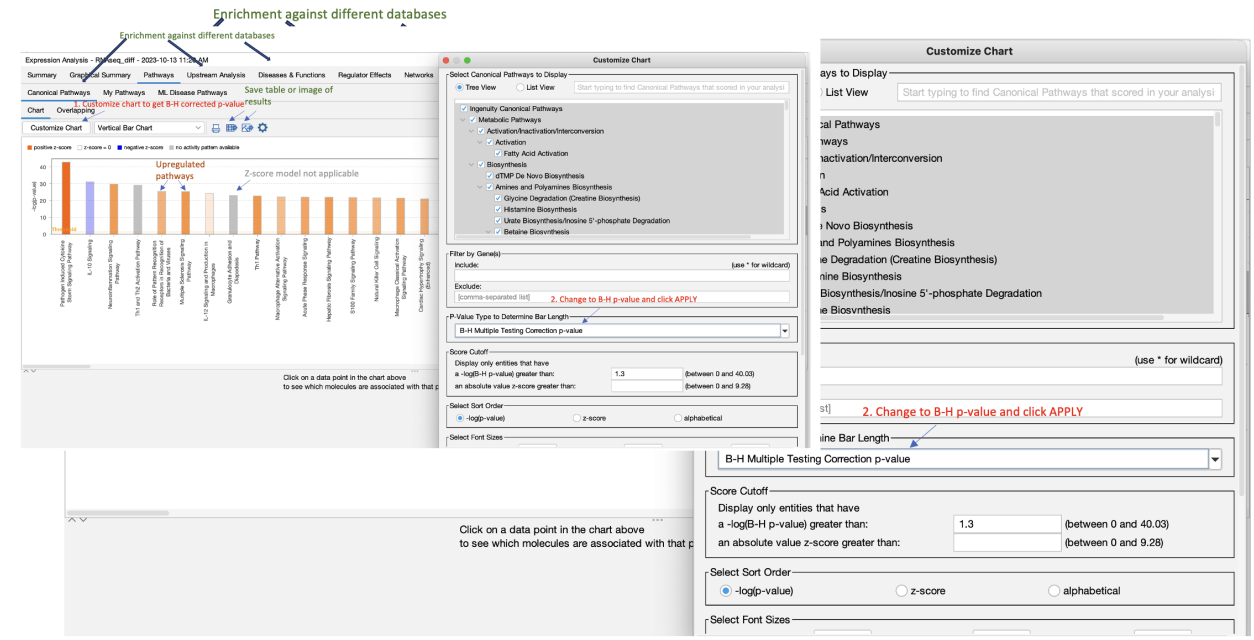
1.8.3 View IPA results

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.

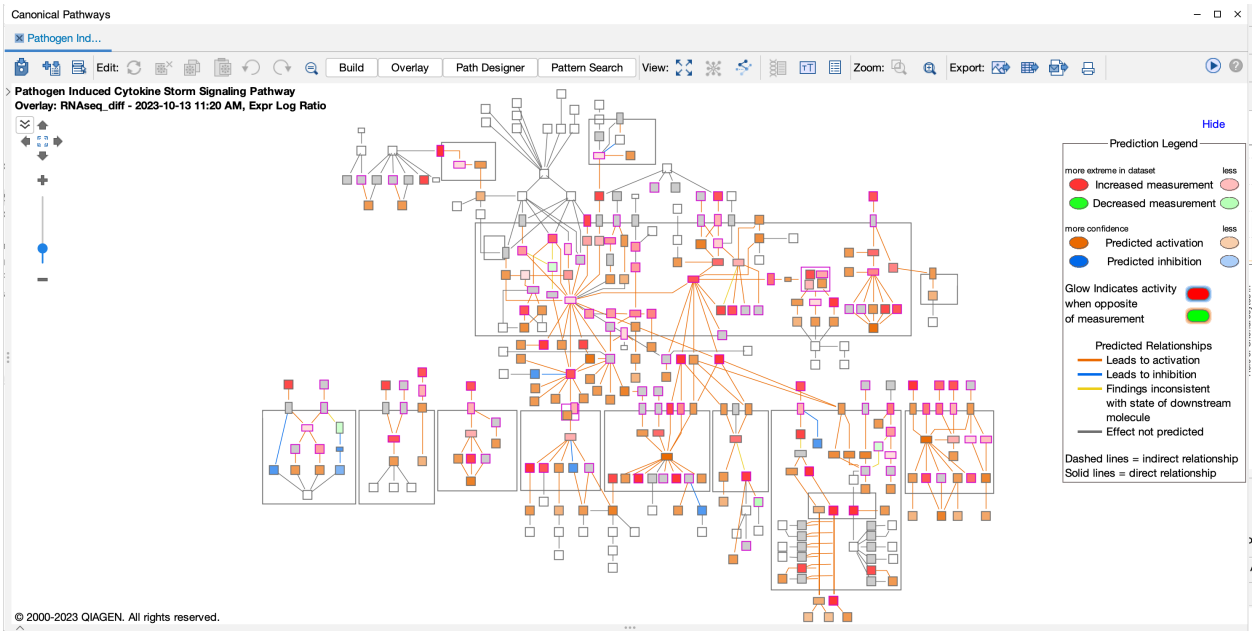
- The bar plot shows significance in -log10 scale (i.e., 2 = p-value of 0.01). Bars are colored by z-score; gray bays indicate that the z-scoring model could not be computed for that pathway.
- **IMPORTANT:** IPA reports nominal p-value by default. **This is wrong**: it guarantees many false positives. 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`

- 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
```

```r
table(degs_list)
```

```
## degs_list
## FALSE  TRUE
## 23298  2633
```

```r
# and we can use table to get the 2x2 contingency table
fet.table <- table(data.frame(antiviral_list, degs_list))
fet.table
```

```
##               degs_list
## antiviral_list FALSE  TRUE
##          FALSE 23293  2594
##          TRUE      5    39
```

```r
# run fisher's exact test
fisher.test(fet.table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  fet.table
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   27.51973 229.17905
## sample estimates:
## odds ratio
##   70.10235
```

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:*

```r
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`

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`

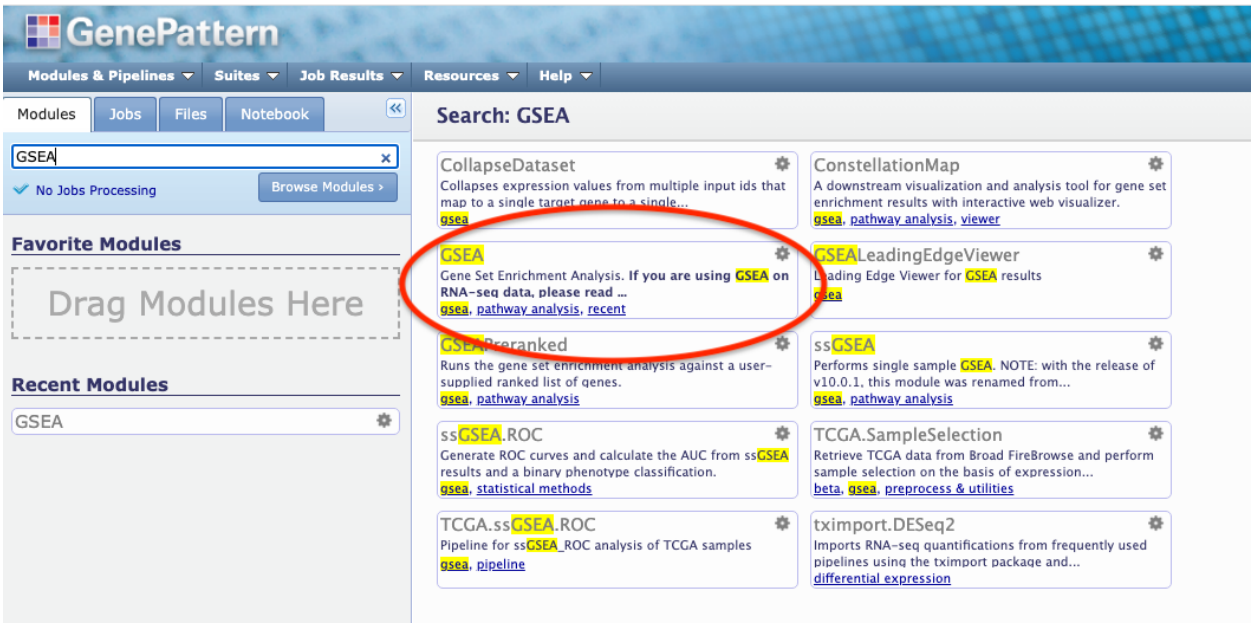
`https://wd.cri.uic.edu/pathway/RNAseq_for_gsea.cls`

1.10.2 Go to GSEA

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

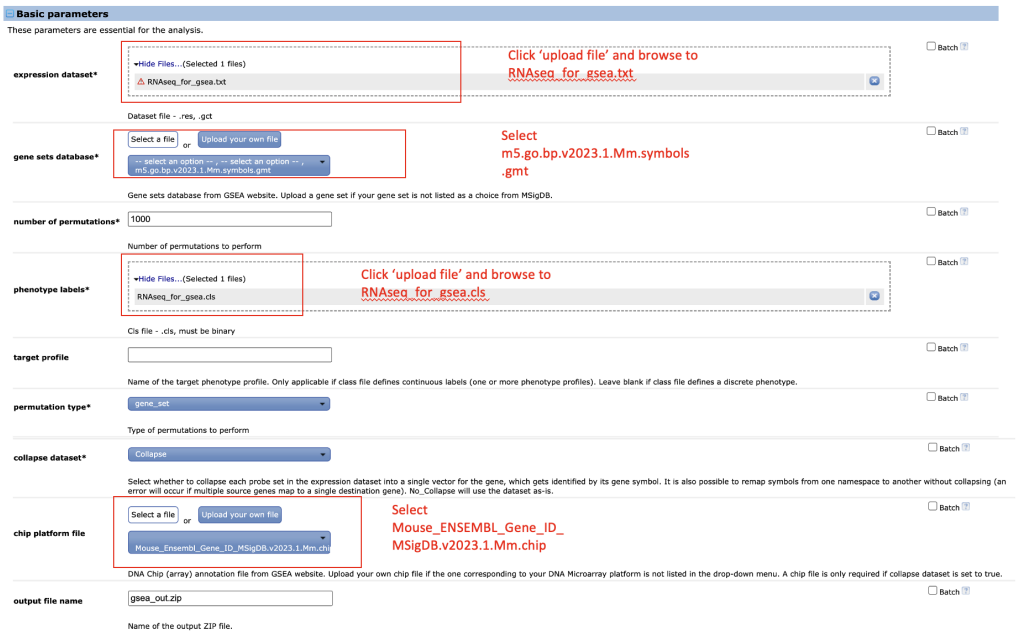
After logging in, enter "GSEA" under the Modules search box, and pick the first option (GSEA).

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



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.



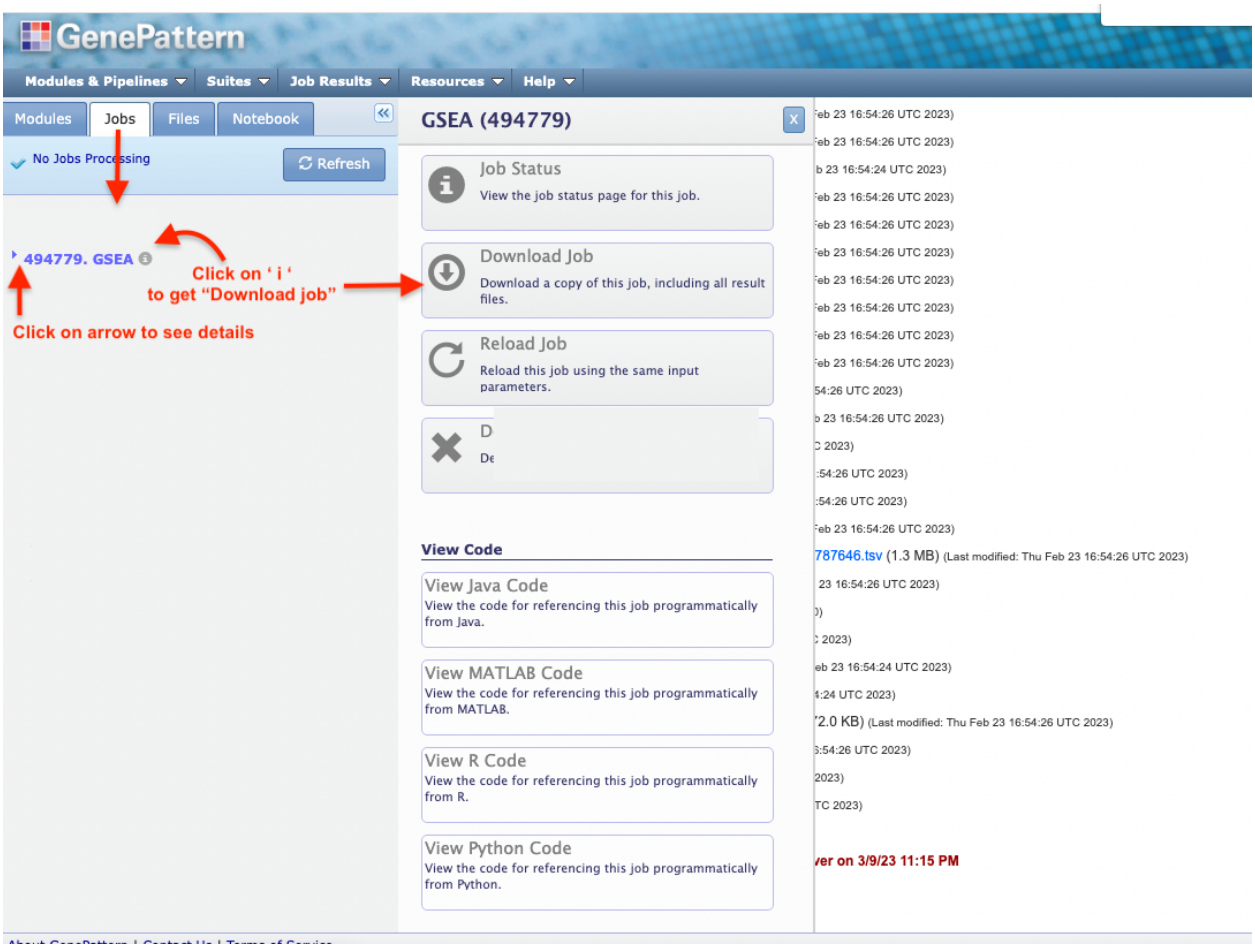
You can close your browser while the job runs. When you log back in, you can check the status or see results.

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...

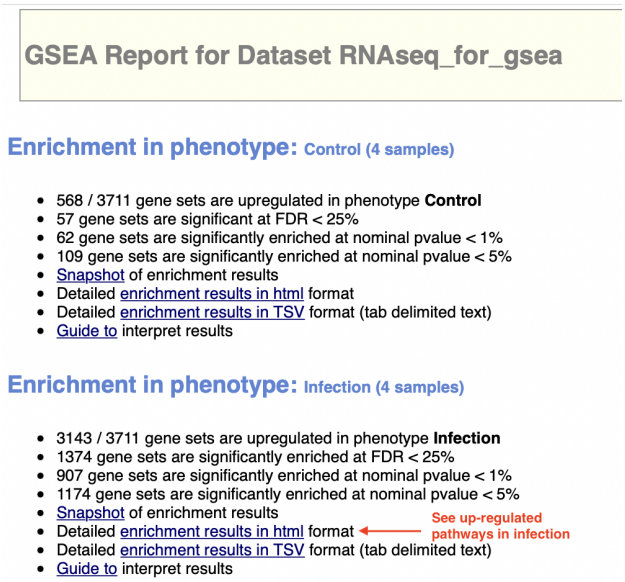
`https://wd.cri.uic.edu/pathway/gsea_out.zip`



1.10.5 View results

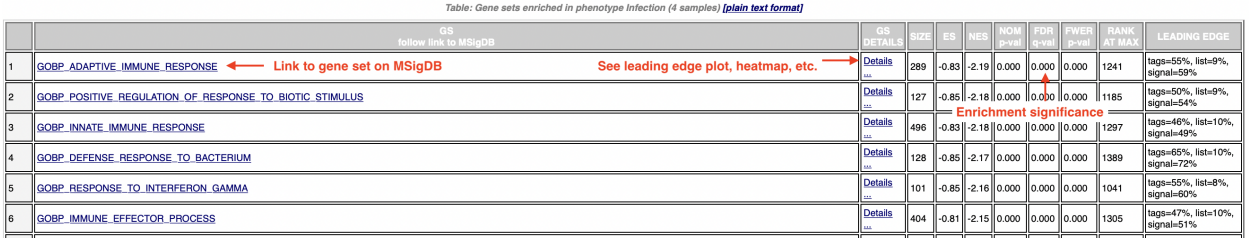
Unpack the zip folder, and find the main result file, **index.html**, double-click to open in a browser.

- 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.



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



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"
```

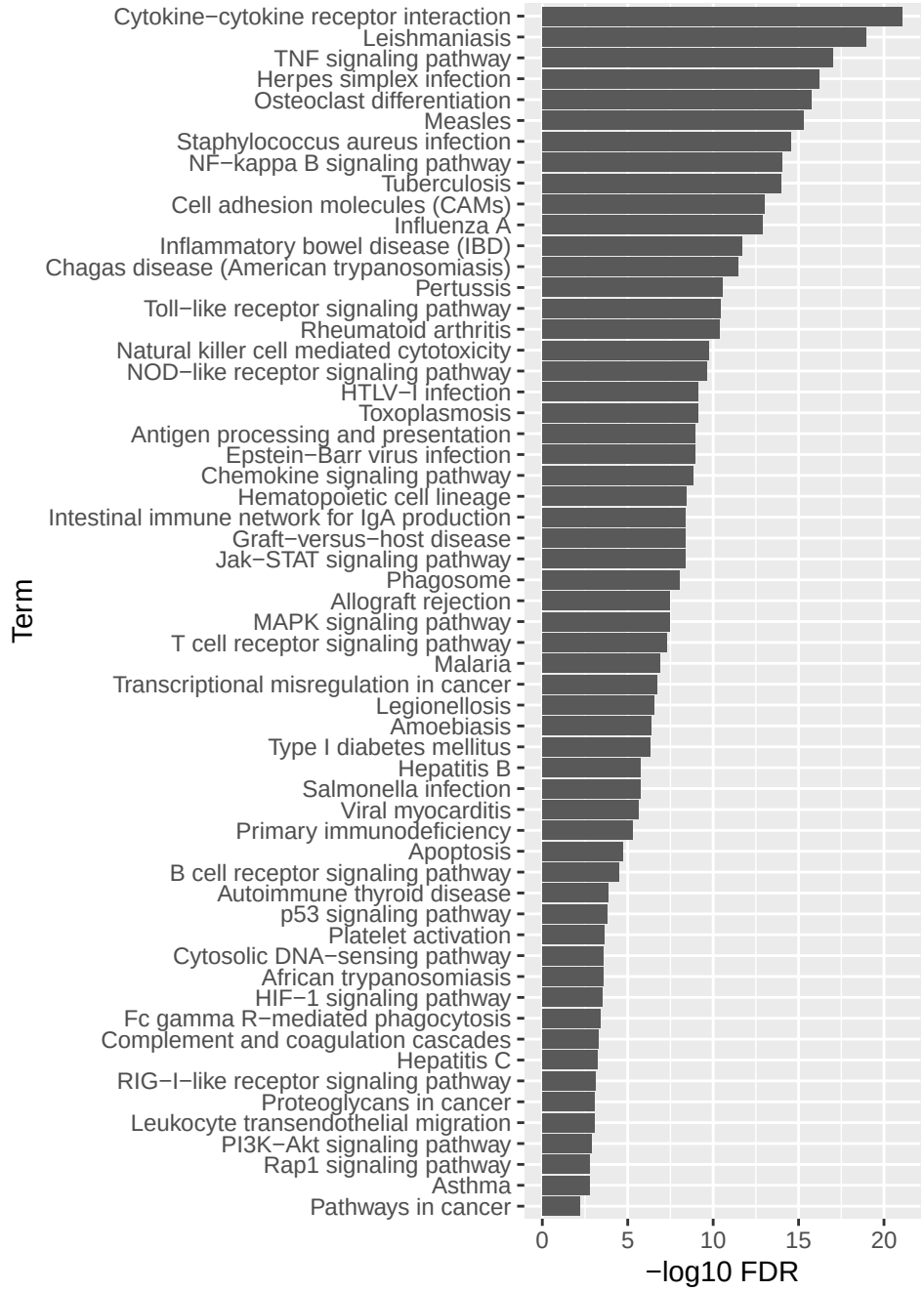
```
# load the ggplot2 library
library(ggplot2)
```

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

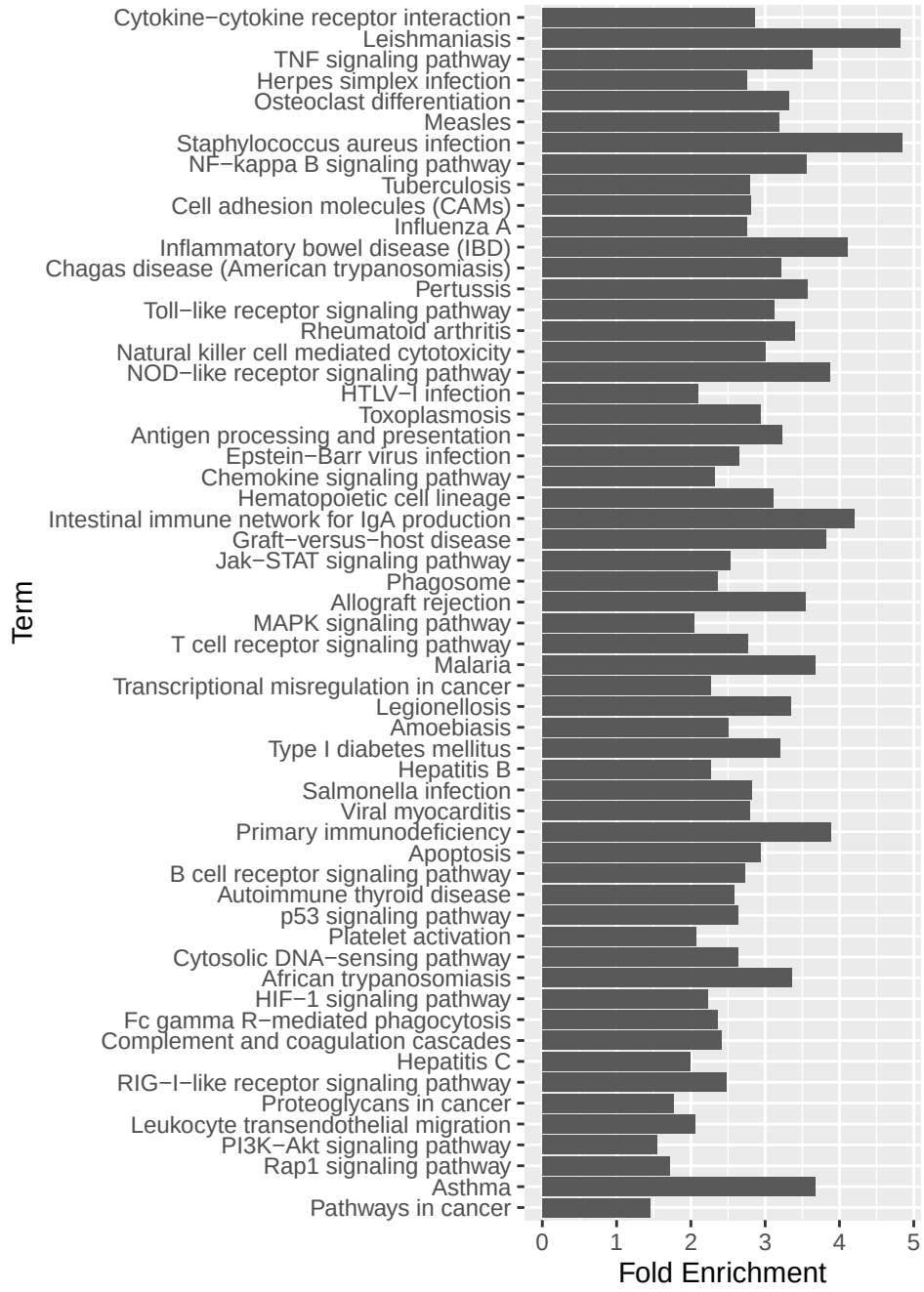
2.1.1 Plot $-\log_{10}$ FDR

```
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 `tidyr` 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(tidyr)
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
```

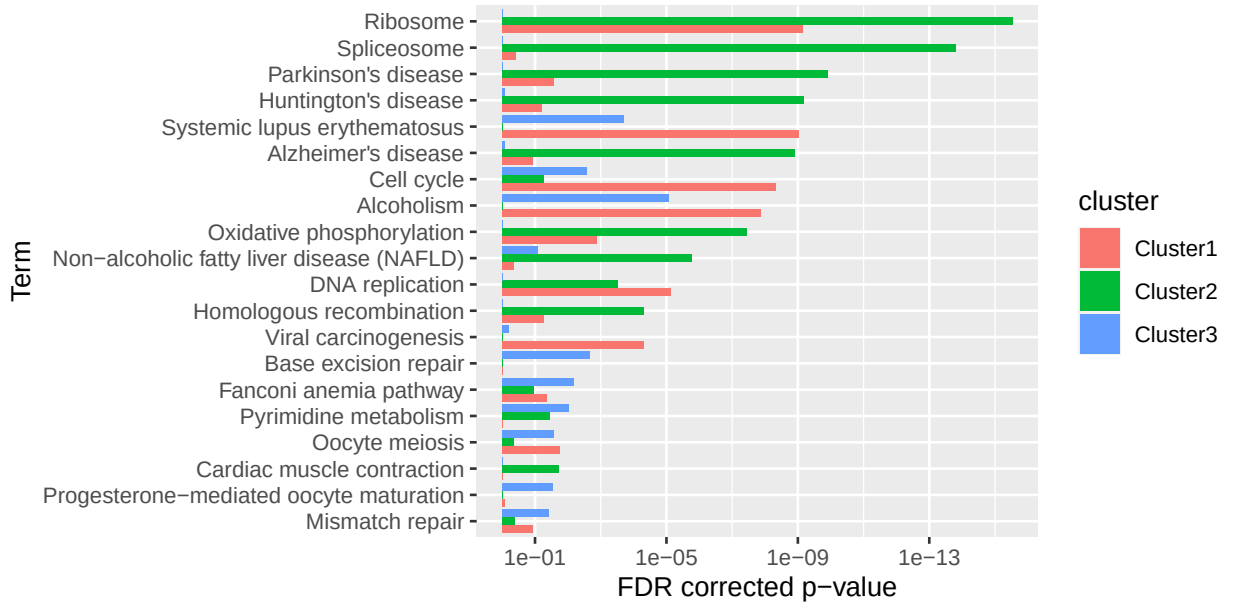
```
## Ribosome                      2.96408e-16 1.000000000
## Spliceosome                   1.55431e-14 0.999992409
## Parkinson's disease           1.23347e-10 0.981089407
## Huntington's disease          6.58956e-10 0.851762867
## Systemic lupus erythematosus 1.00000e+00 0.000195325
## Alzheimer's 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
```

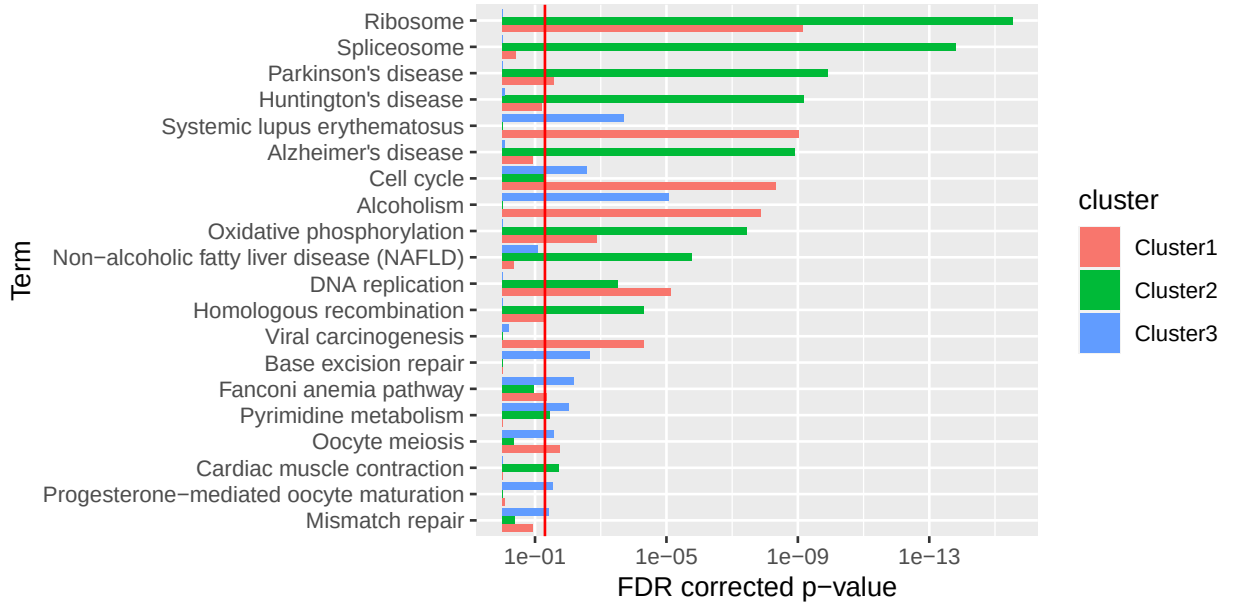
3. Create barplot using *ggplot2*
    - `fill=variable` will color based on the different clusters in the variable column
    - `position=dodge` will put the bars next to each other
    - `scale_x_discrete`, To order the items using the kegg_subset2 list.
    - `scale_y_continous`, This will use our 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

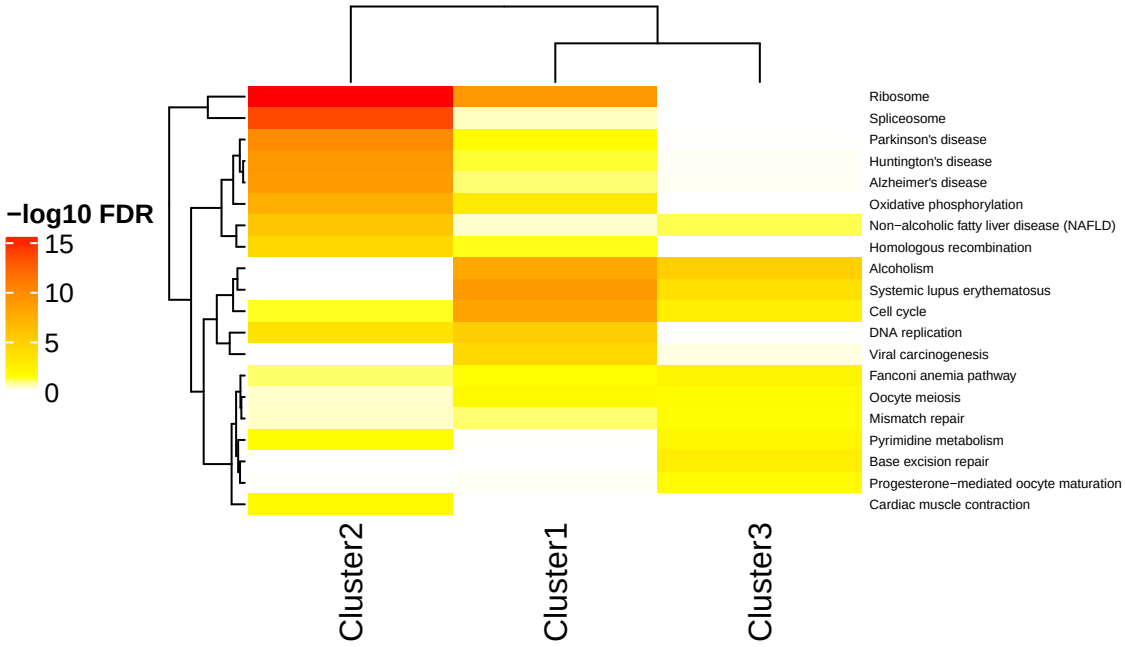
```
library(circlize)
```

```
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

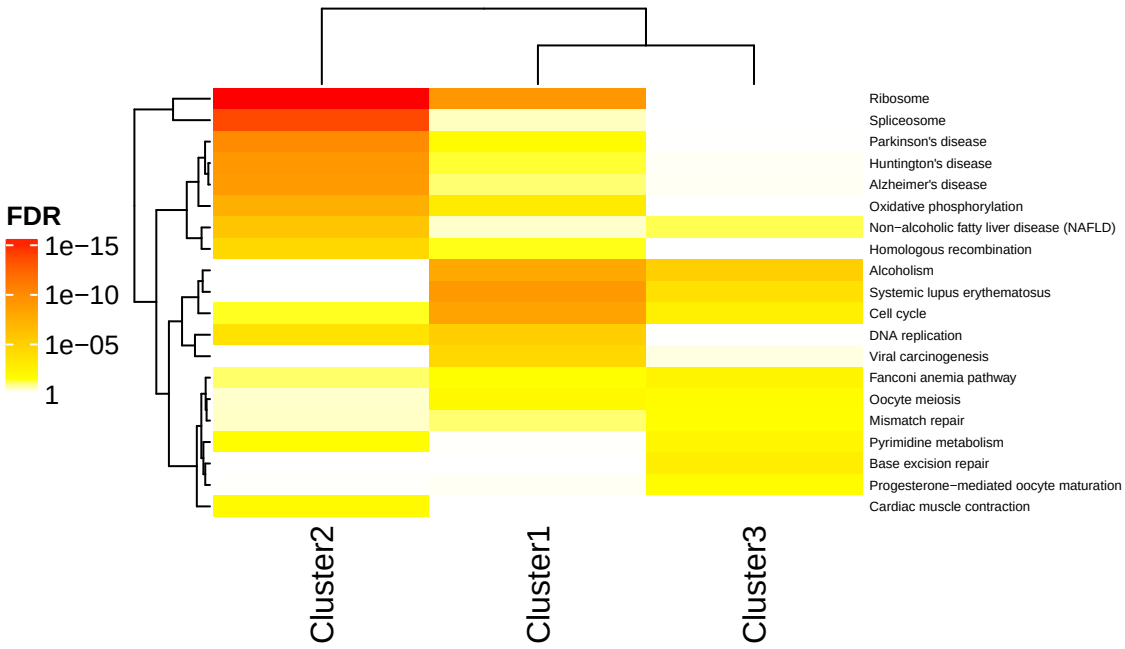
```
library(ComplexHeatmap)
```

```
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")
```


FDR
1e−15
1e−10
1e−05
1
Ribosome
Spliceosome
Parkinson's disease
Huntington's disease
Alzheimer's disease
Oxidative phosphorylation
Non−alcoholic fatty liver disease (NAFLD)
Homologous recombination
Alcoholism
Systemic lupus erythematosus
Cell cycle
DNA replication
Viral carcinogenesis
Fanconi anemia pathway
Oocyte meiosis
Mismatch repair
Pyrimidine metabolism
Base excision repair
Progesterone−mediated oocyte maturation
Cardiac muscle contraction
Cluster2
Cluster1
Cluster3

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)
```

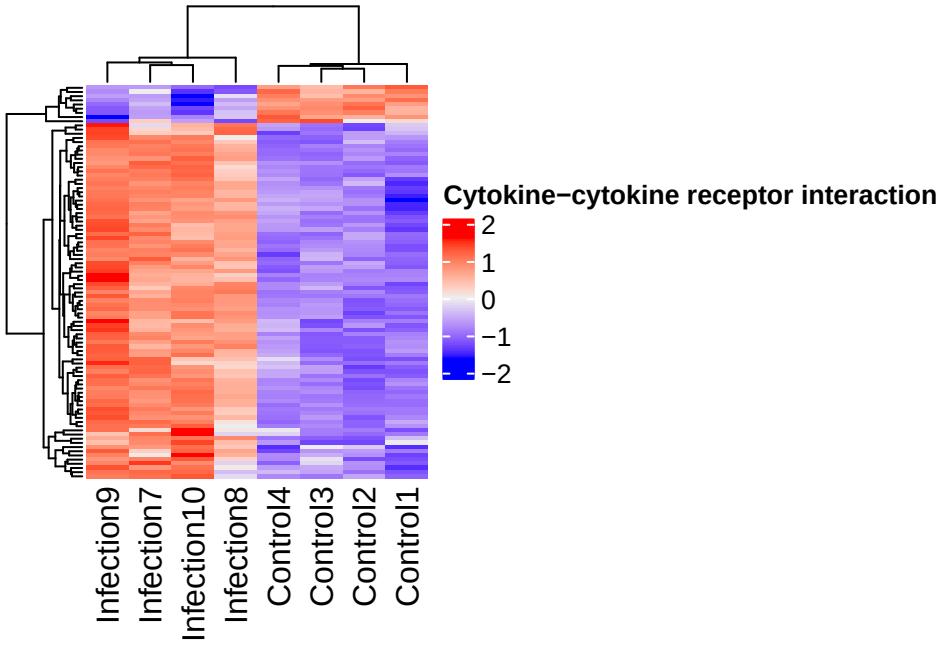
```
## [1] 94
```

```
# 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 Interlant 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
##                                          AAChange 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    AACS 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:

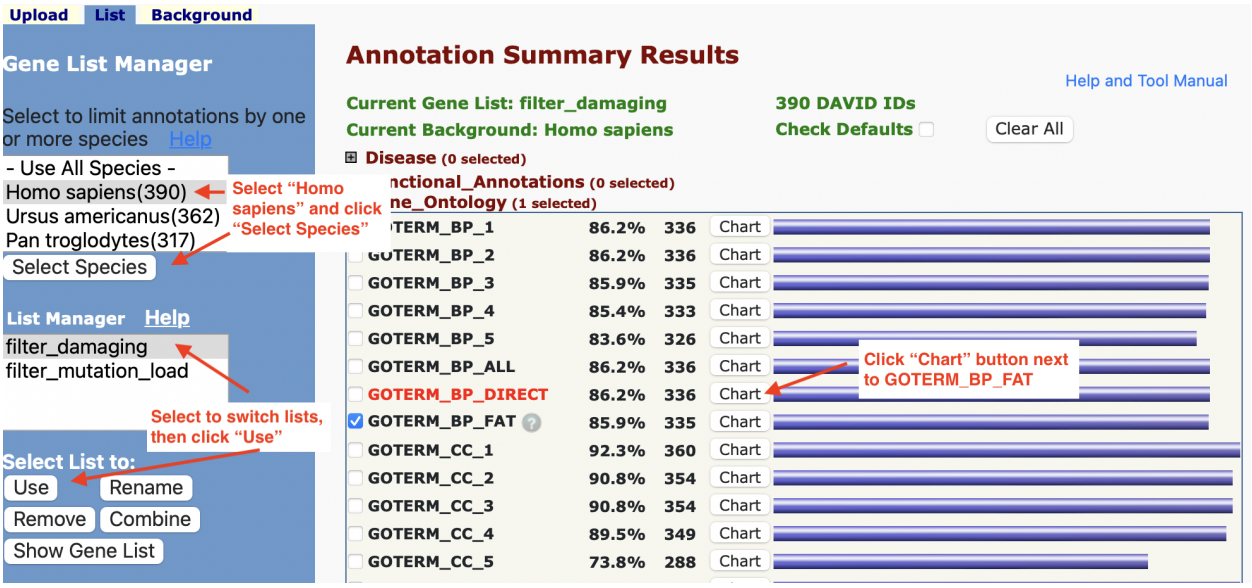
- https://wd.cri.uic.edu/pathway/filter_damaging.txt
- https://wd.cri.uic.edu/pathway/filter_mutation_load.txt

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.

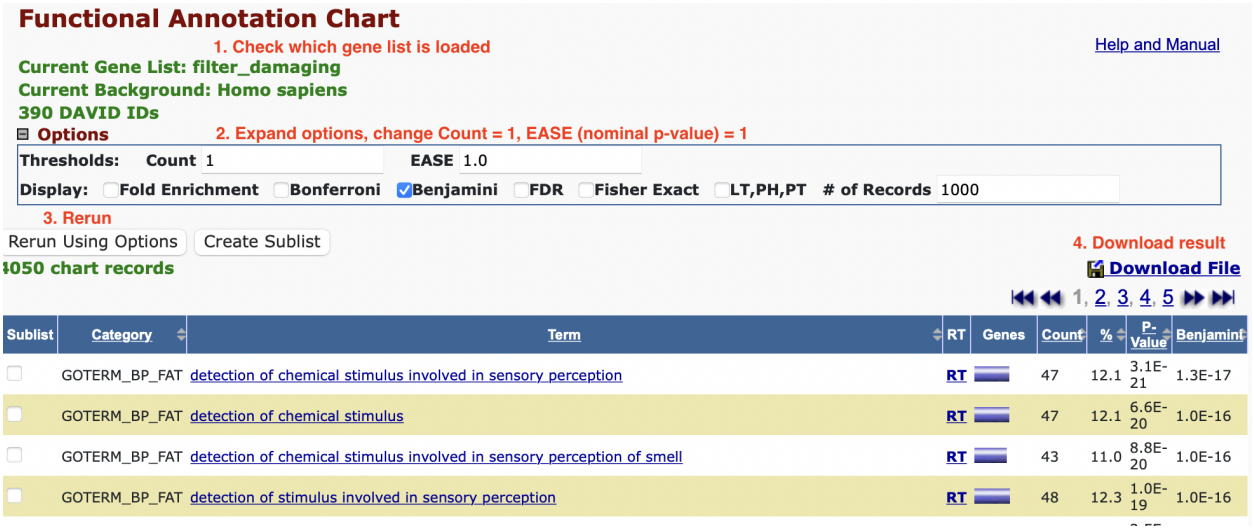
Go to DAVID `https://david.ncifcrf.gov`, and follow similar steps as before:

- 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.



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**

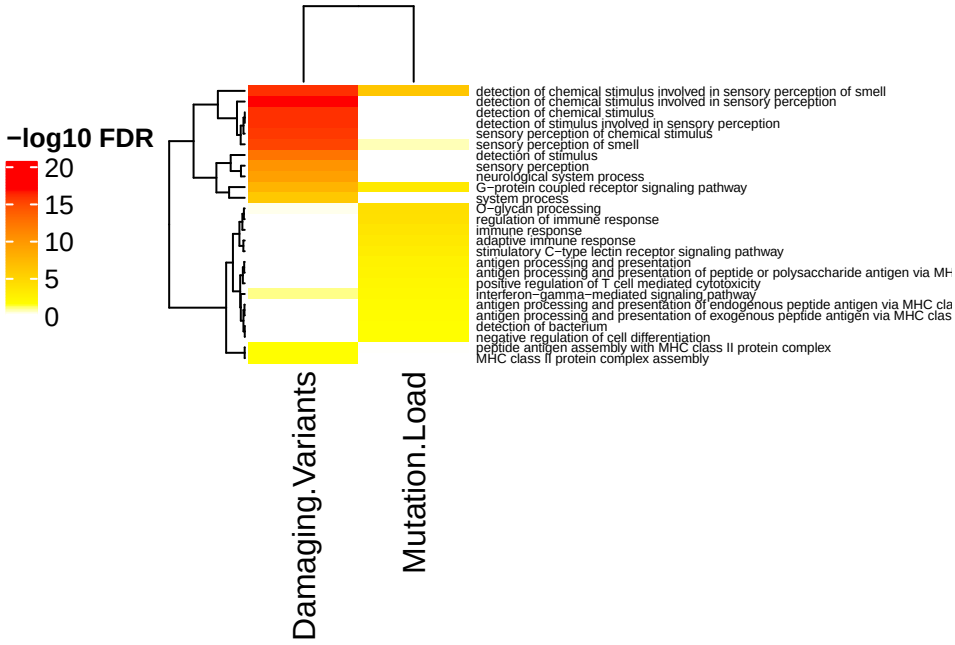


2.4.3 Comparison visualization in R

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]
```

```r
# 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")
```



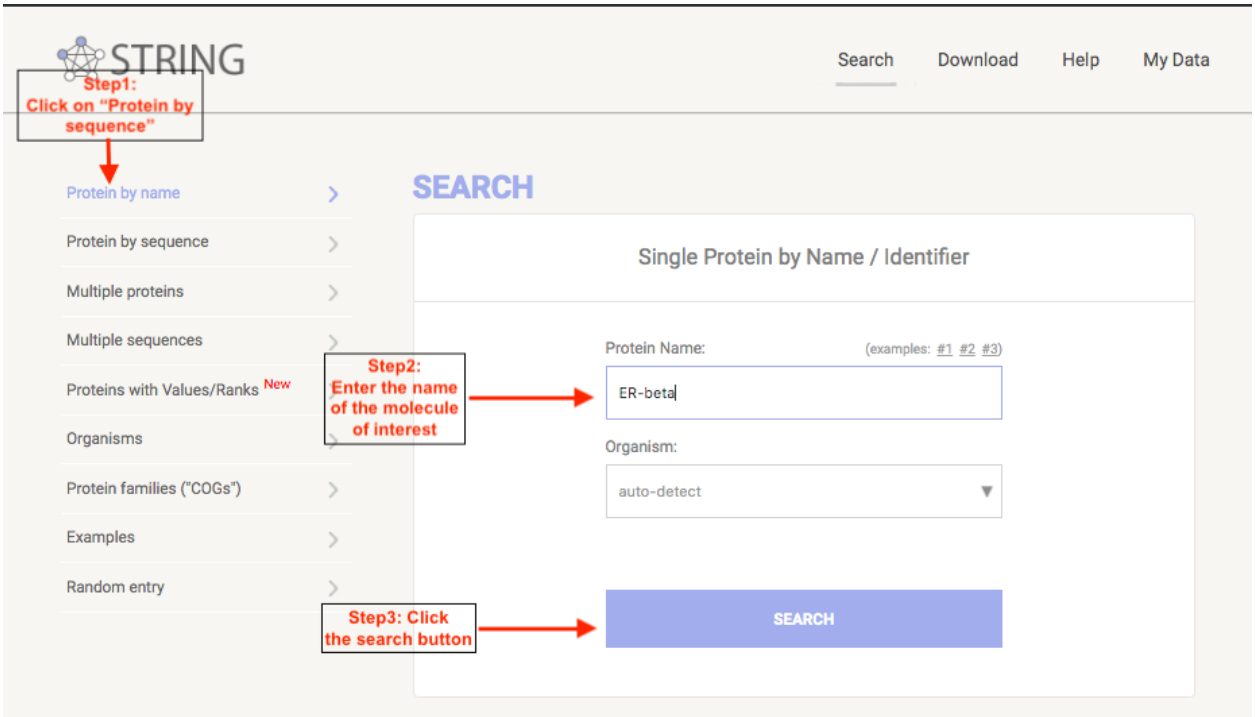
2.5 Using STRING

2.5.1 Navigate to STRING

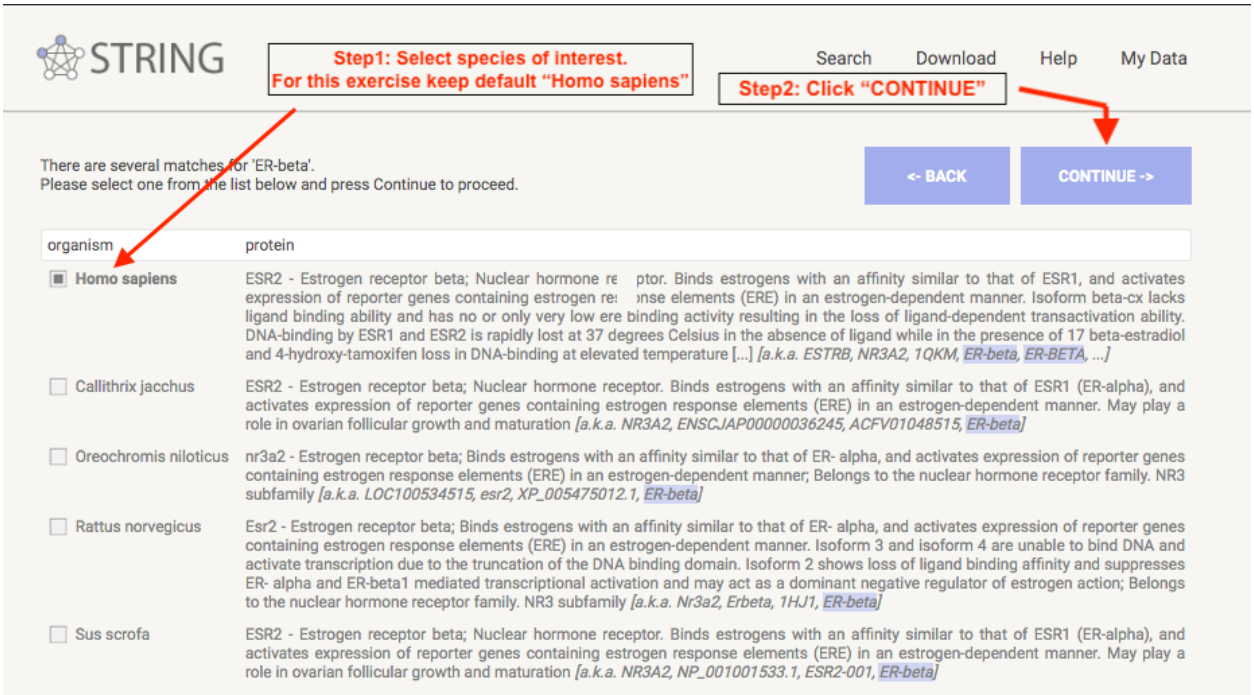
Navigate to STRING `https://string-db.org/` Click on Search Button on Right hand side.

2.5.2 Search by molecule of interest

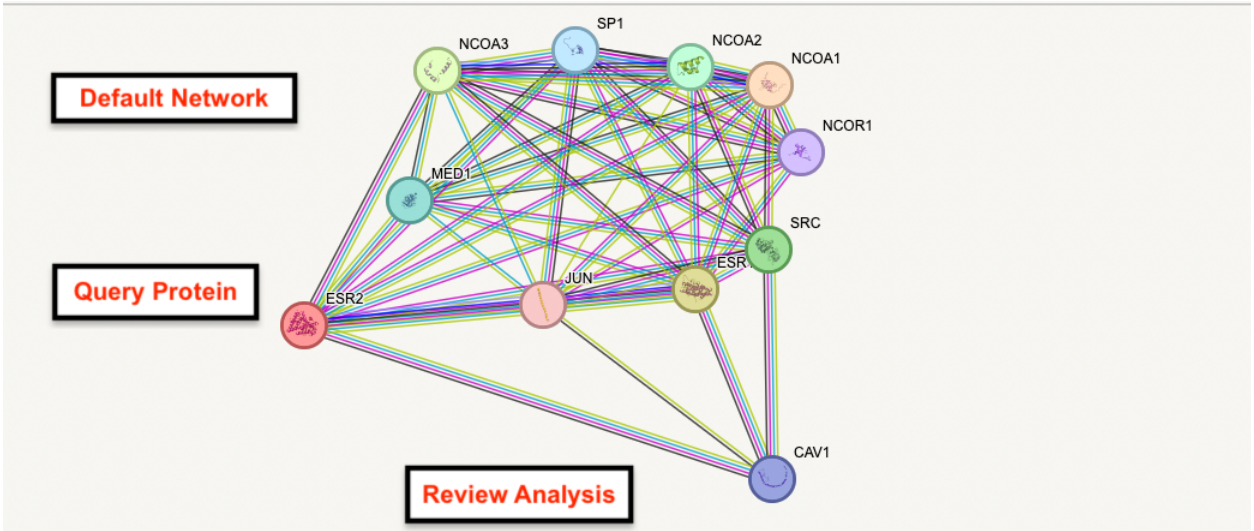
- **Step 1:** Click search button top right side of the menu.
- **Step 2:** Select "Protein by name" in the left side menu
- **Step 3:** Type "ER-beta" in the "Protein Name" search box
- **Step 4:** Click "SEARCH" button



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



- 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.



Viewers | Legend | Settings | Analysis | Exports | Clusters | More | Less

Network Stats

number of nodes:	11	expected number of edges:	22
number of edges:	48	PPI enrichment p-value:	8.18e-07
average node degree:	8.73	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.932		

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:1904017	Cellular response to Thyroglobulin triiodothyronine	2 of 3	3.08	0.00063
GO:0086098	Angiotensin-activated signaling pathway involved in heart pr...	2 of 3	3.08	0.00063
GO:0060750	Epithelial cell proliferation involved in mammary gland duct ...	2 of 4	2.95	0.00086
GO:0035357	Peroxisome proliferator activated receptor signaling pathway	3 of 9	2.78	1.85e-05
GO:0060745	Mammary gland branching involved in pregnancy	2 of 6	2.78	0.0014

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0030284	Nuclear estrogen receptor activity	2 of 3	3.08	0.0014
GO:0034056	Estrogen response element binding	2 of 4	2.95	0.0019
GO:0017162	Aryl hydrocarbon receptor binding	2 of 9	2.6	0.0049
GO:0046966	Nuclear thyroid hormone receptor binding	3 of 29	2.27	0.00035
GO:0030331	Nuclear estrogen receptor binding	4 of 41	2.24	7.63e-06

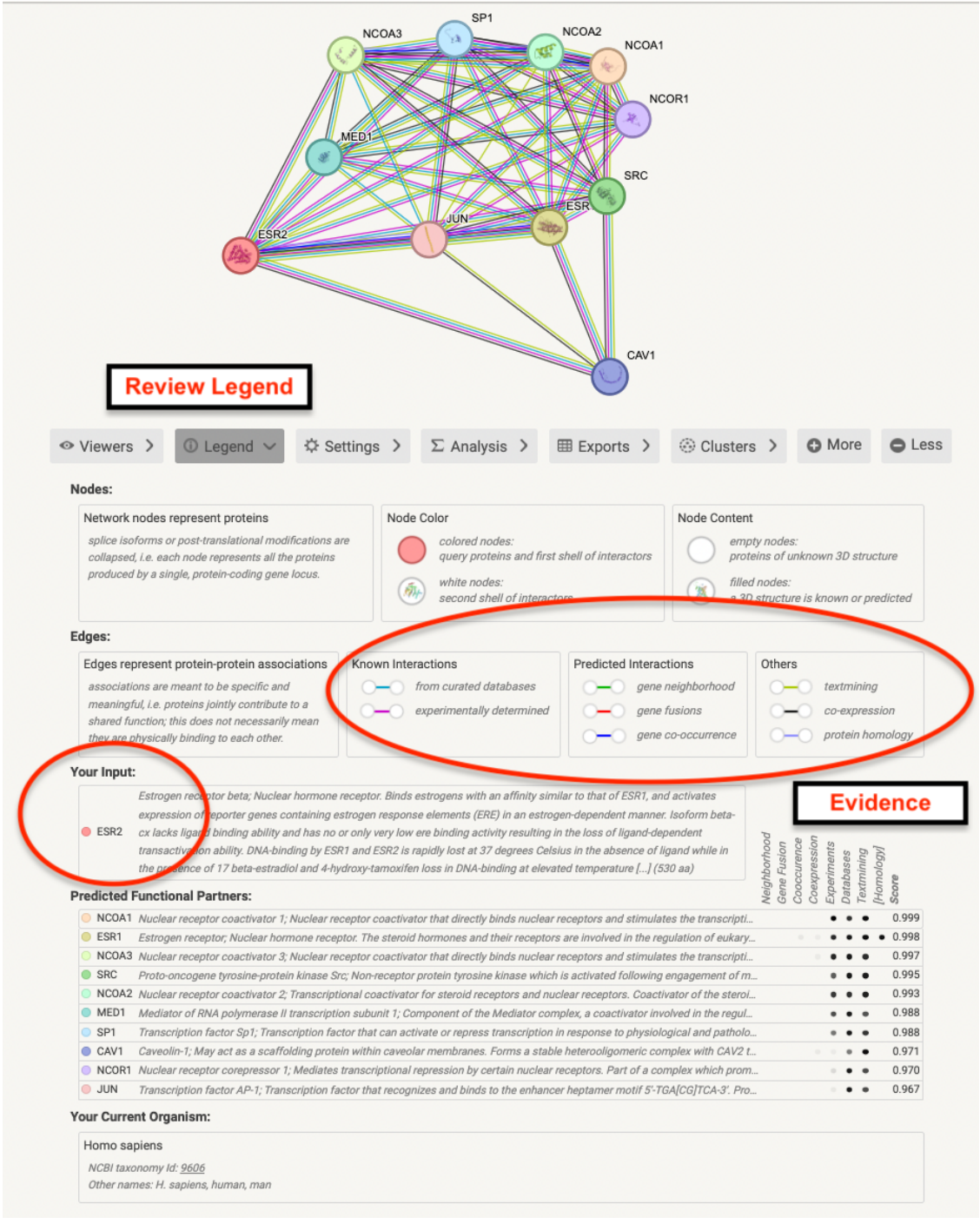
(more ...)

Cellular Component (Gene Ontology)

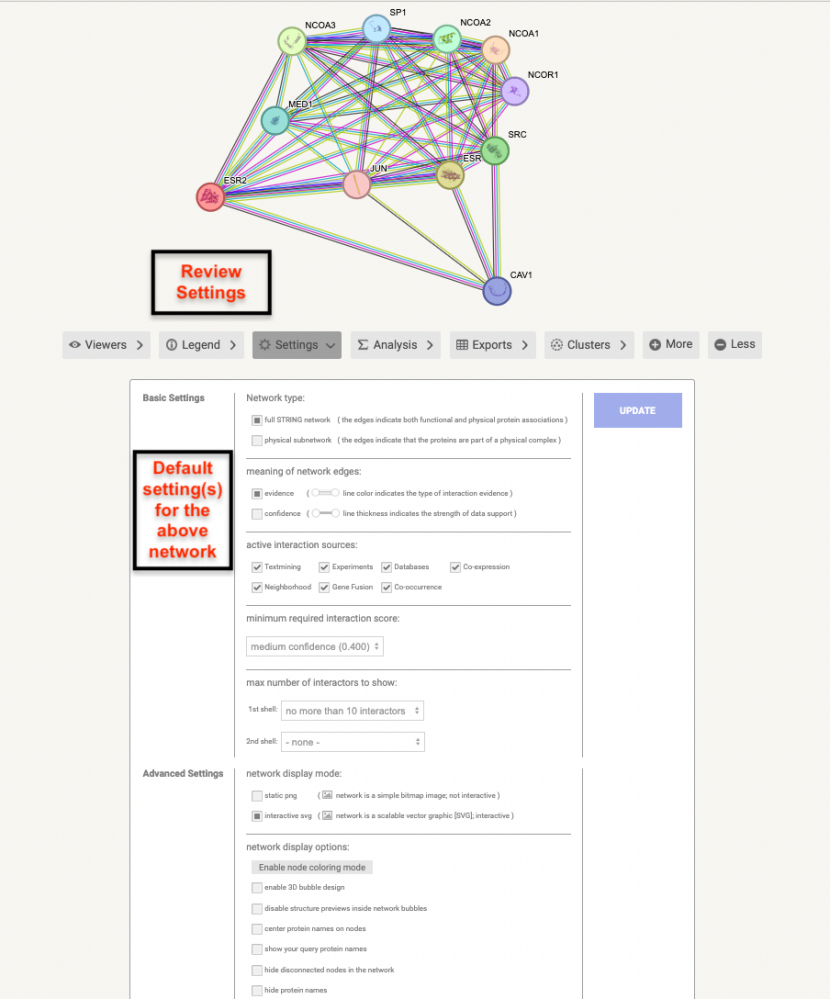
GO-term	description	count in network	strength	false discovery rate
GO:0000791	Euchromatin	3 of 60	1.95	0.0013
GO:0017053	Transcription repressor complex	3 of 77	1.84	0.0021
GO:0090575	RNA polymerase II transcription regulator complex	4 of 256	1.45	0.0021
GO:0005667	Transcription regulator complex	7 of 517	1.38	5.56e-06
GO:0000785	Chromatin	8 of 1285	1.05	4.75e-05

(more ...)

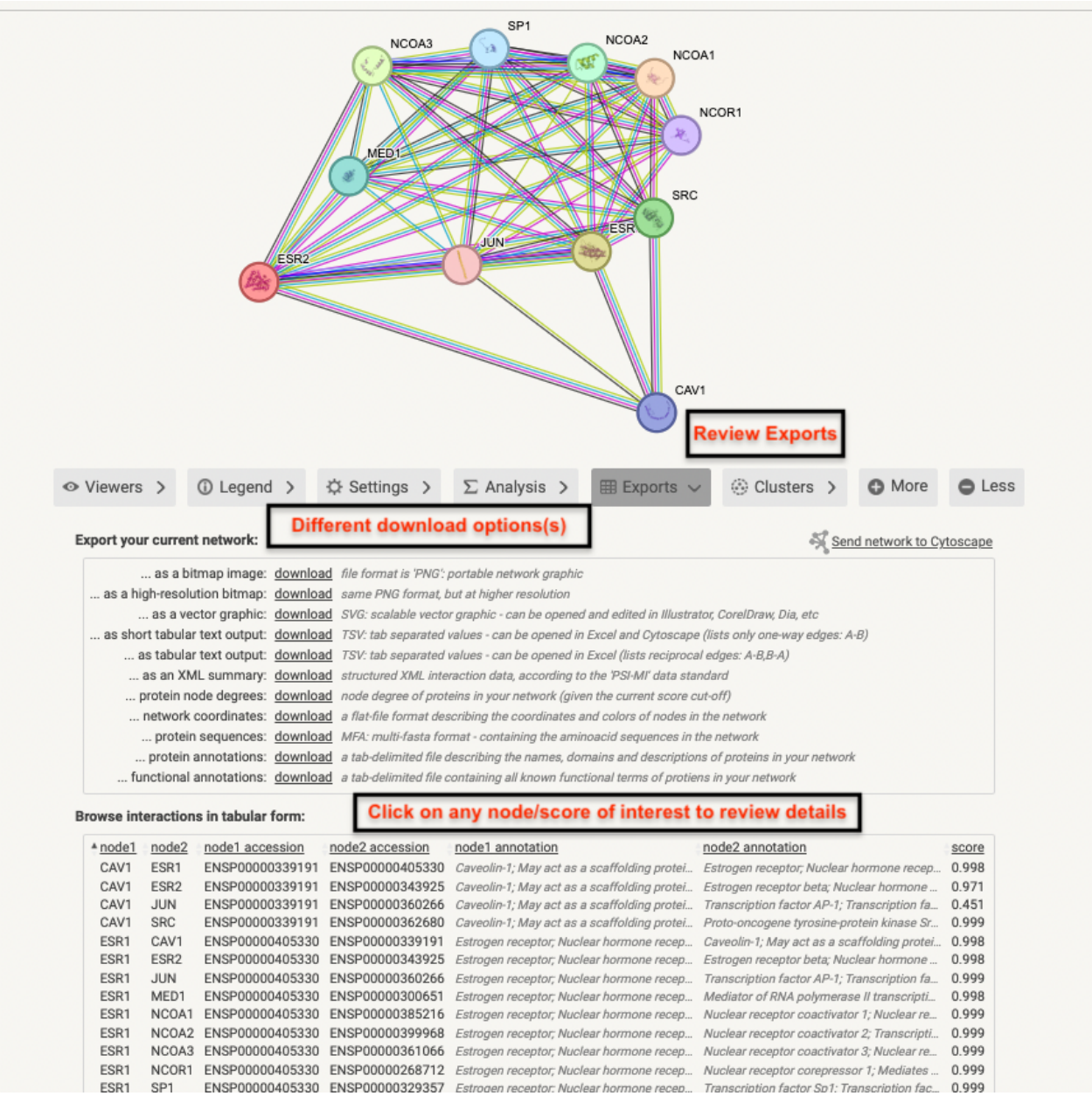
- Click on “Legend”
  - Review edges colors for interaction type, e.g. pink = Experimental Evidence
  - Click on node/edge to get more details for a given interaction.



- 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 iteractors are the proteins directly associated with input protein(s)/query. 2nd shell of iteractors 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.



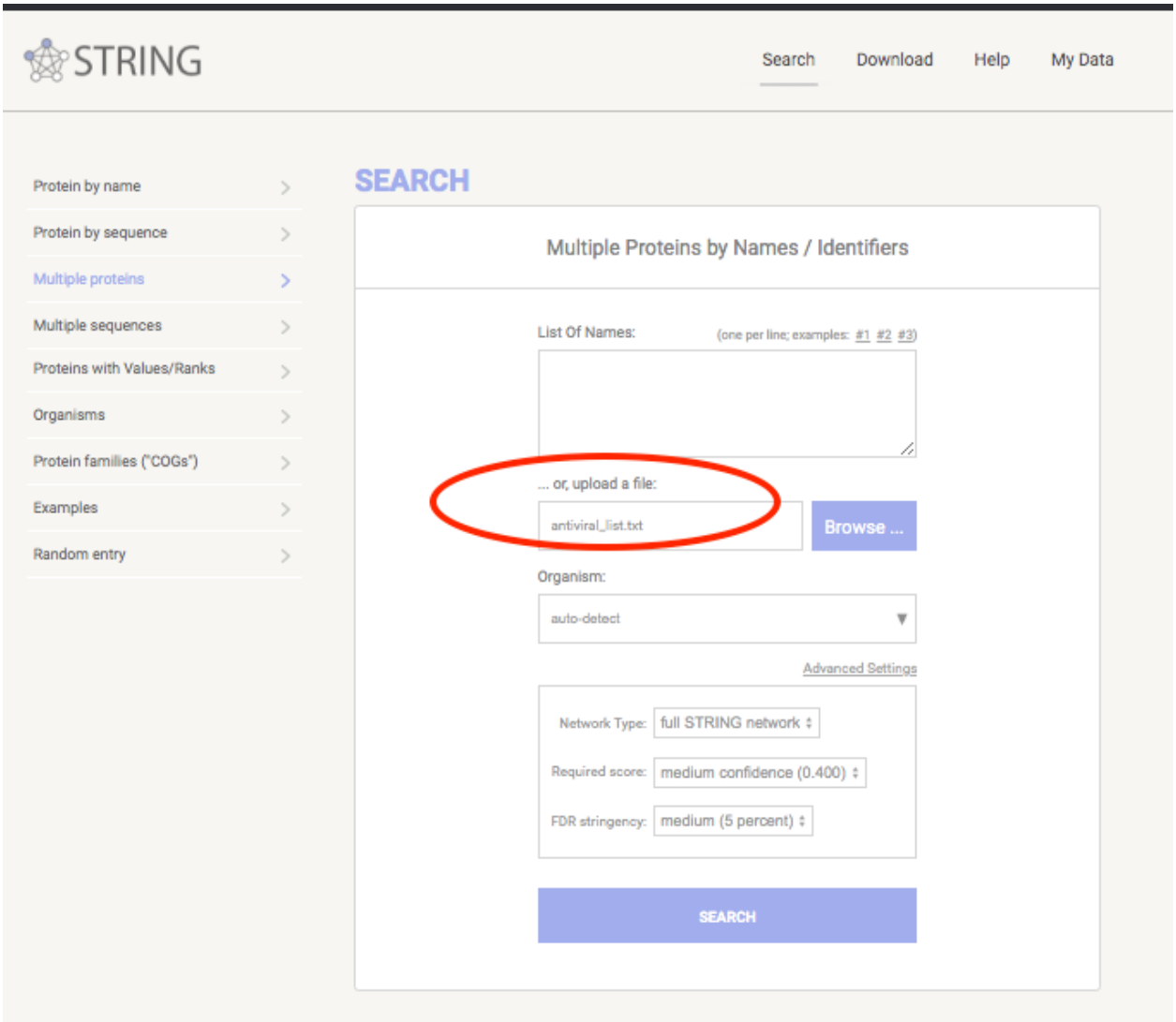
2.5.3 Search by multiple proteins

Download antiviral gene list.

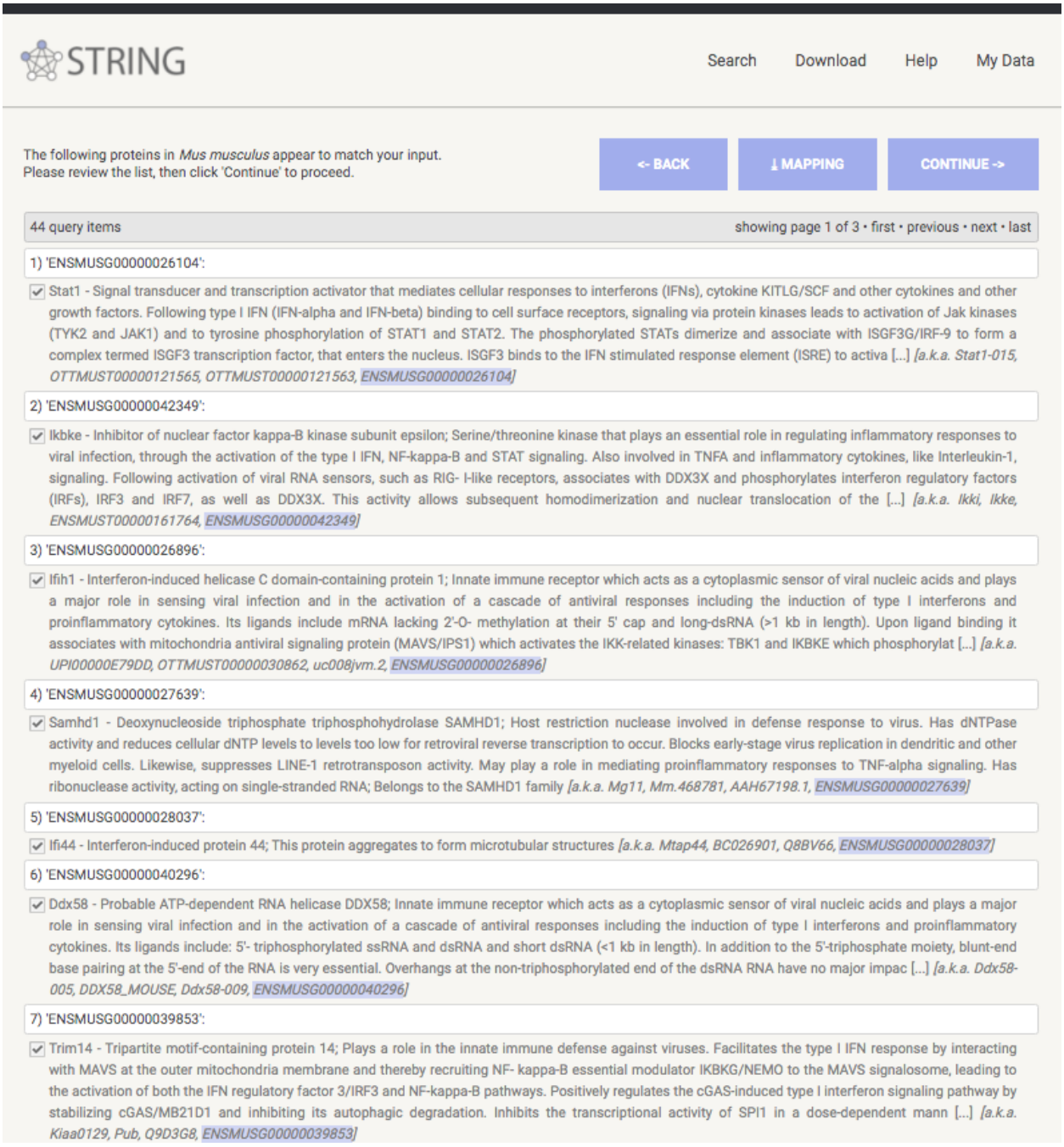
`https://wd.cri.uic.edu/pathway/antiviral_list.txt`

- **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

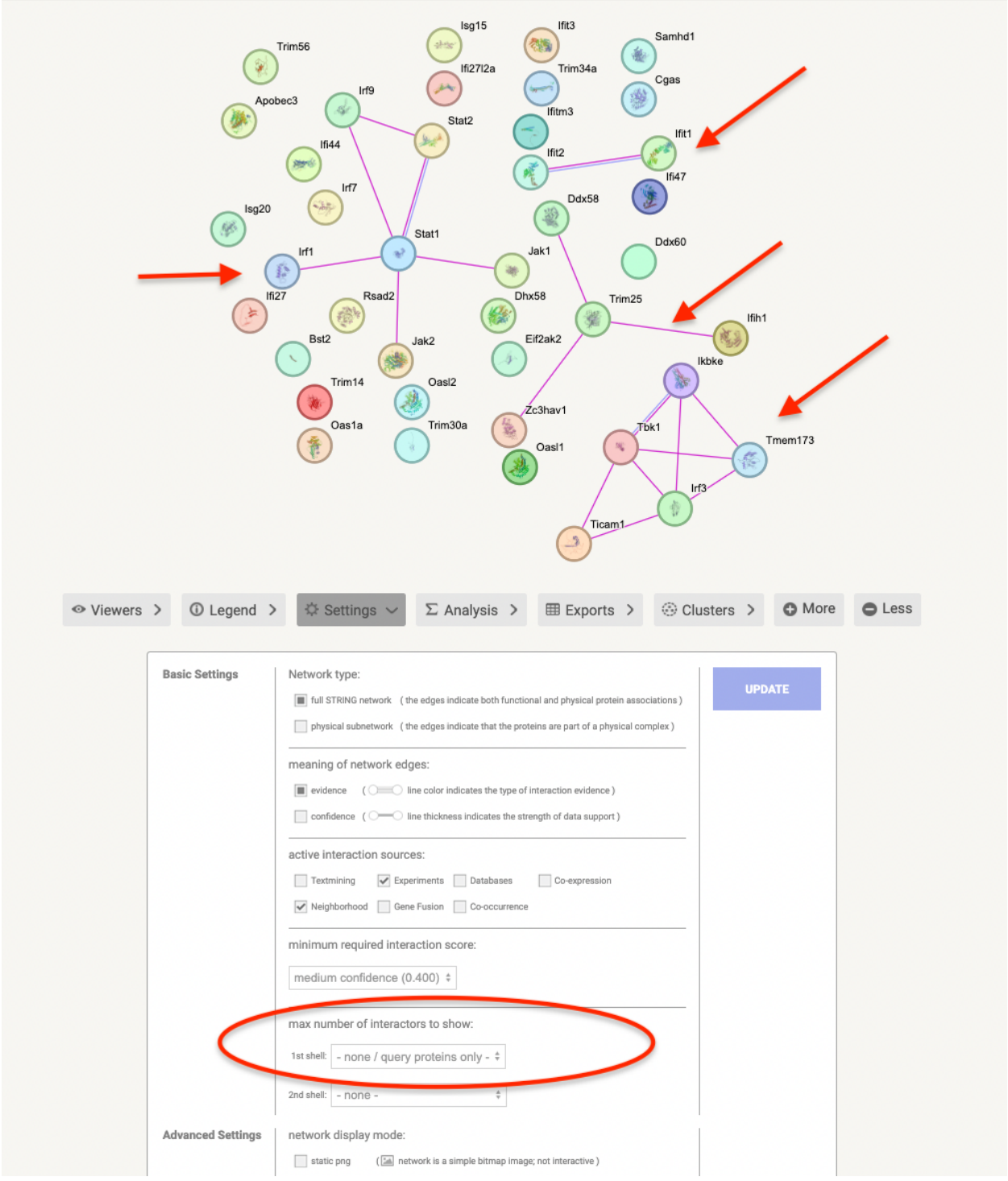
NOTE: If STRING asks you to select an organism, type in "Mus musculus".



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

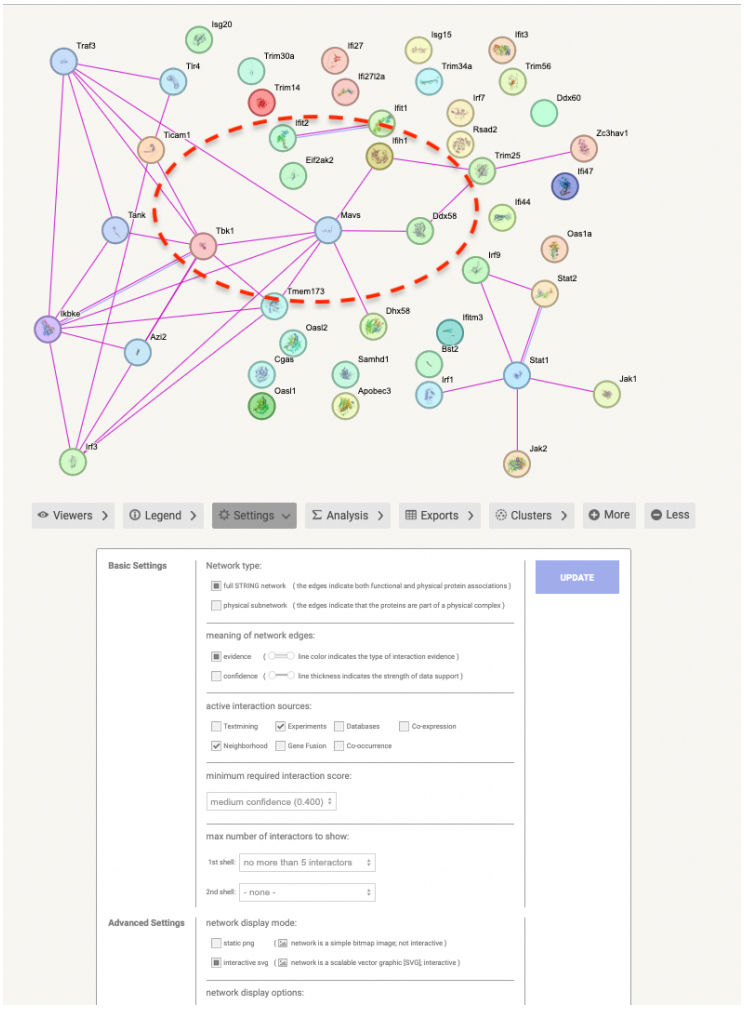


- Adjustments to view settings (**The network is shown using these settings**):
  - Please choose following options if your network looks different than the figure.
    - Step 1: Under "Active interaction sources:" Select "Experiments" and "Neighborhood""
    - Step 2: Under "minimum required interaction 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


Isg15
Ifit3
Samhd1
Trim56
Ifi27l2a
Trim34a
Cgas
Irf9
Apobec3
Ifitm3
Stat2
Ifit1
Ifi44
Ifit2
Ifi47
Irf7
Ddx58
Isg20
Stat1
Ddx60
Irf1
Jak1
Ifi27
Rsad2
Dhx58
Trim25
Ifih1
Bst2
Jak2
Eif2ak2
Ikbke
Trim14
Oasl2
Zc3hav1
Oas1a
Trim30a
Tbk1
Tmem173
Oasl1
Irf3
Ticam1
Viewers
Legend
Settings
Analysis
Exports
Clusters
More
Less
Basic Settings
Network type:
full STRING network (the edges indicate both functional and physical protein associations)
physical subnetwork (the edges indicate that the proteins are part of a physical complex)
meaning of network edges:
evidence (line color indicates the type of interaction evidence)
confidence (line thickness indicates the strength of data support)
active interaction sources:
Textmining
Experiments
Databases
Co-expression
Neighborhood
Gene Fusion
Co-occurrence
minimum required interaction score:
medium confidence (0.400)
max number of interactors to show:
1st shell:
- none / query proteins only -
2nd shell:
- none -
UPDATE
Advanced Settings
network display mode:
static png
(network is a simple bitmap image; not interactive)

- 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 iteractors are the proteins directly associated with input protein(s)/query. 2nd shell of iteractors 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.

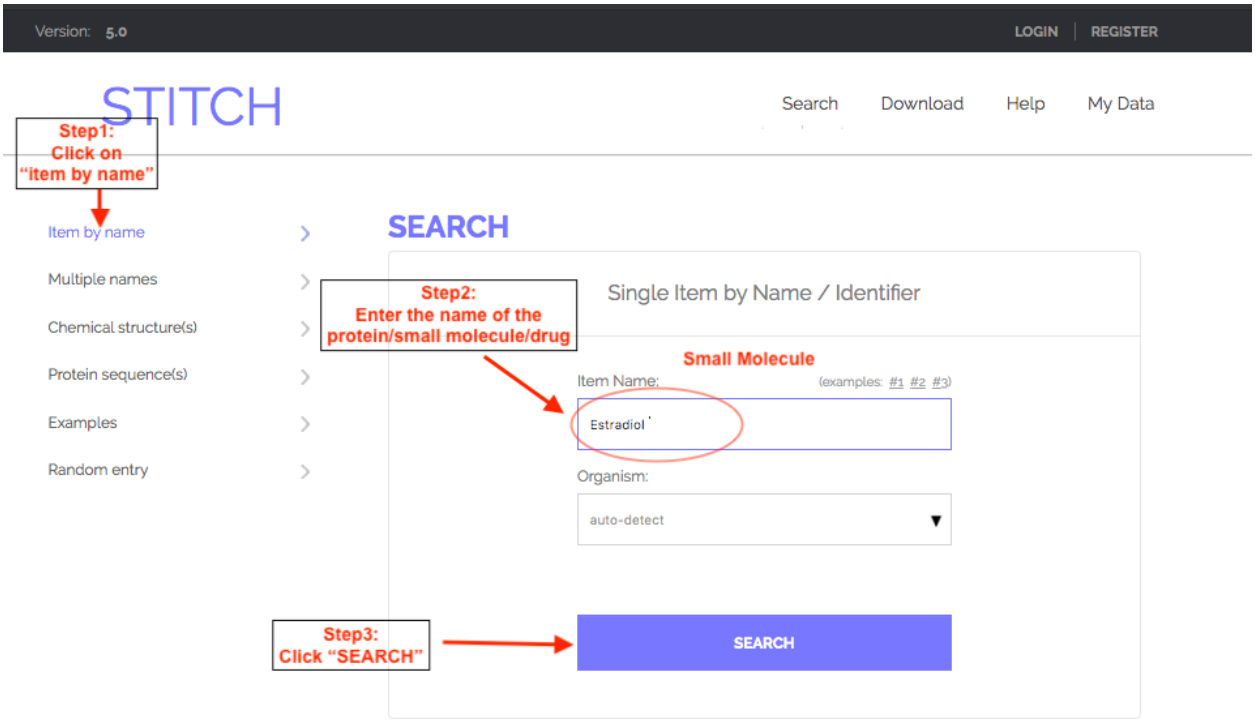
2.6 Using STITCH

2.6.1 Navigate to STITCH

Navigate to STITCH `http://stitch.embl.de/`

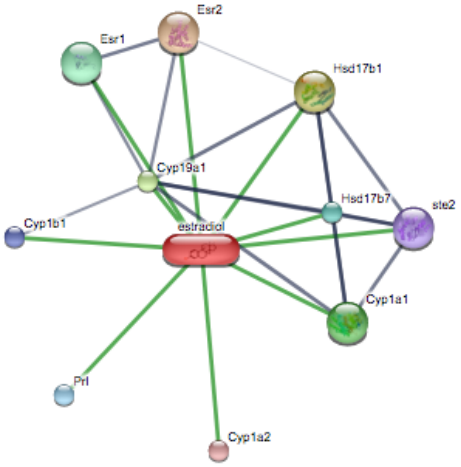
2.6.2 Search by molecule name

- 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



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

- 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.



This is the **confidence view**. Stronger associations are represented by thicker lines.
Protein-protein interactions are shown in grey, chemical-protein interactions in green and interactions between chemicals in red.

Viewers › | Legend › | Settings › | Analysis ⌄ | Tables / Exports › | More | Less

Protein-Protein Network Stats

number of nodes:	10	expected number of edges:	10
number of edges:	15	PPI enrichment p-value:	0.107
average node degree:	3	*your network does **not** have significantly more interactions than expected (what does that mean?)*	
clustering coefficient:	0.875		

Functional enrichments in your network *Note: some enrichments may be expected here (why?)*

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0008210	estrogen metabolic process	5	1.99e-10
GO:0034754	cellular hormone metabolic process	6	2.96e-10
GO:0008202	steroid metabolic process	6	2.27e-08
GO:0032355	response to estradiol	6	8.6e-08
GO:0006629	lipid metabolic process	7	1.15e-06

(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0070330	aromatase activity	4	3.9e-06
GO:0020037	heme binding	4	8.56e-05
GO:0046914	transition metal ion binding	6	8.56e-05
GO:0005506	iron ion binding	4	0.000116
GO:0030284	estrogen receptor activity	2	0.000233

(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0012505	endomembrane system	8	9.48e-05
GO:0005737	cytoplasm	10	0.000313
GO:0044444	cytoplasmic part	9	0.000313
GO:0043231	intracellular membrane-bounded organelle	9	0.00125
GO:0005783	endoplasmic reticulum	5	0.00218

(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
00140	Steroid hormone biosynthesis	6	5.3e-11
04913	Ovarian steroidogenesis	4	7.82e-07
04917	Prolactin signaling pathway	3	0.000474
01100	Metabolic pathways	5	0.00932
00380	Tryptophan metabolism	2	0.0111

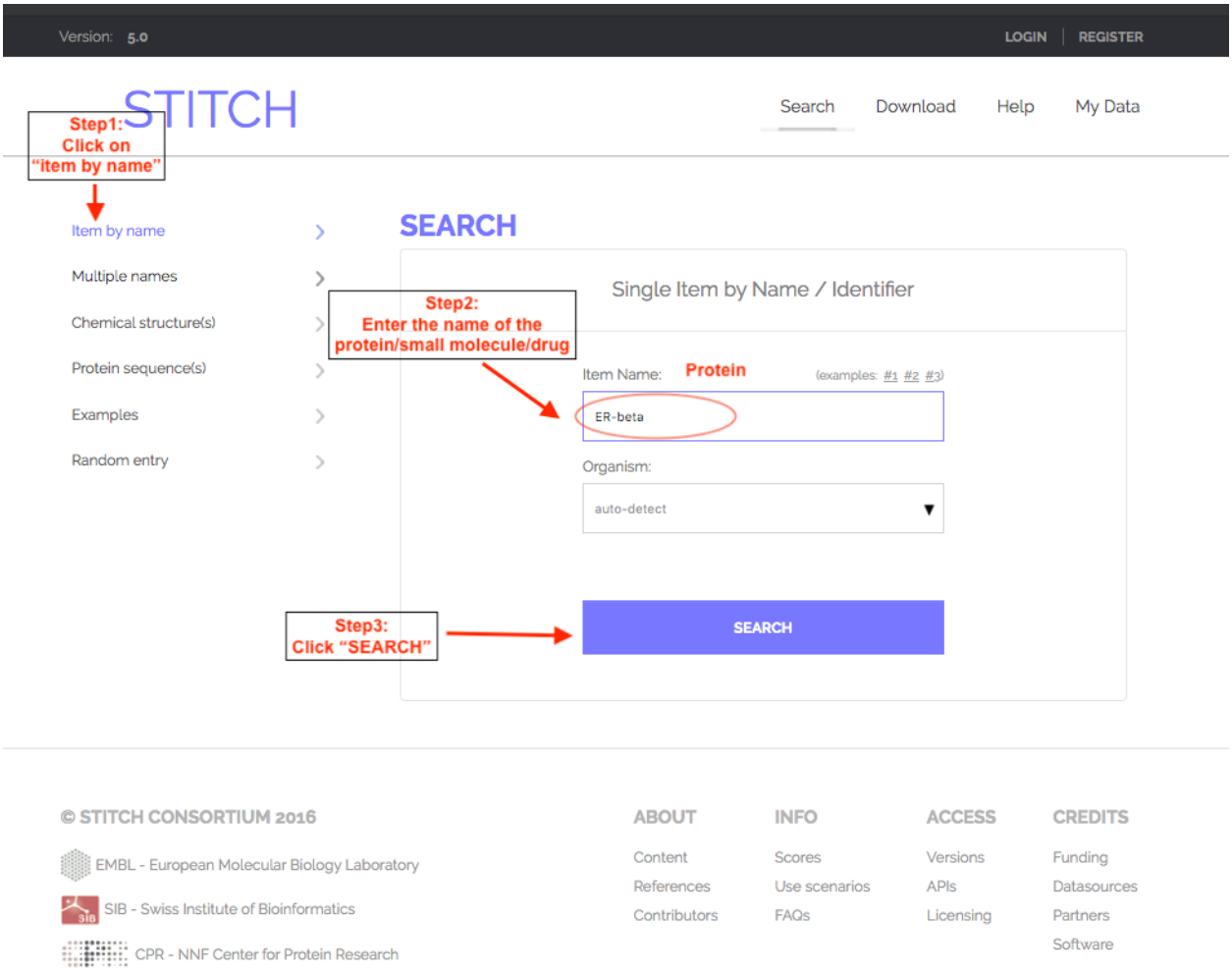
(more ...)

Statistical background

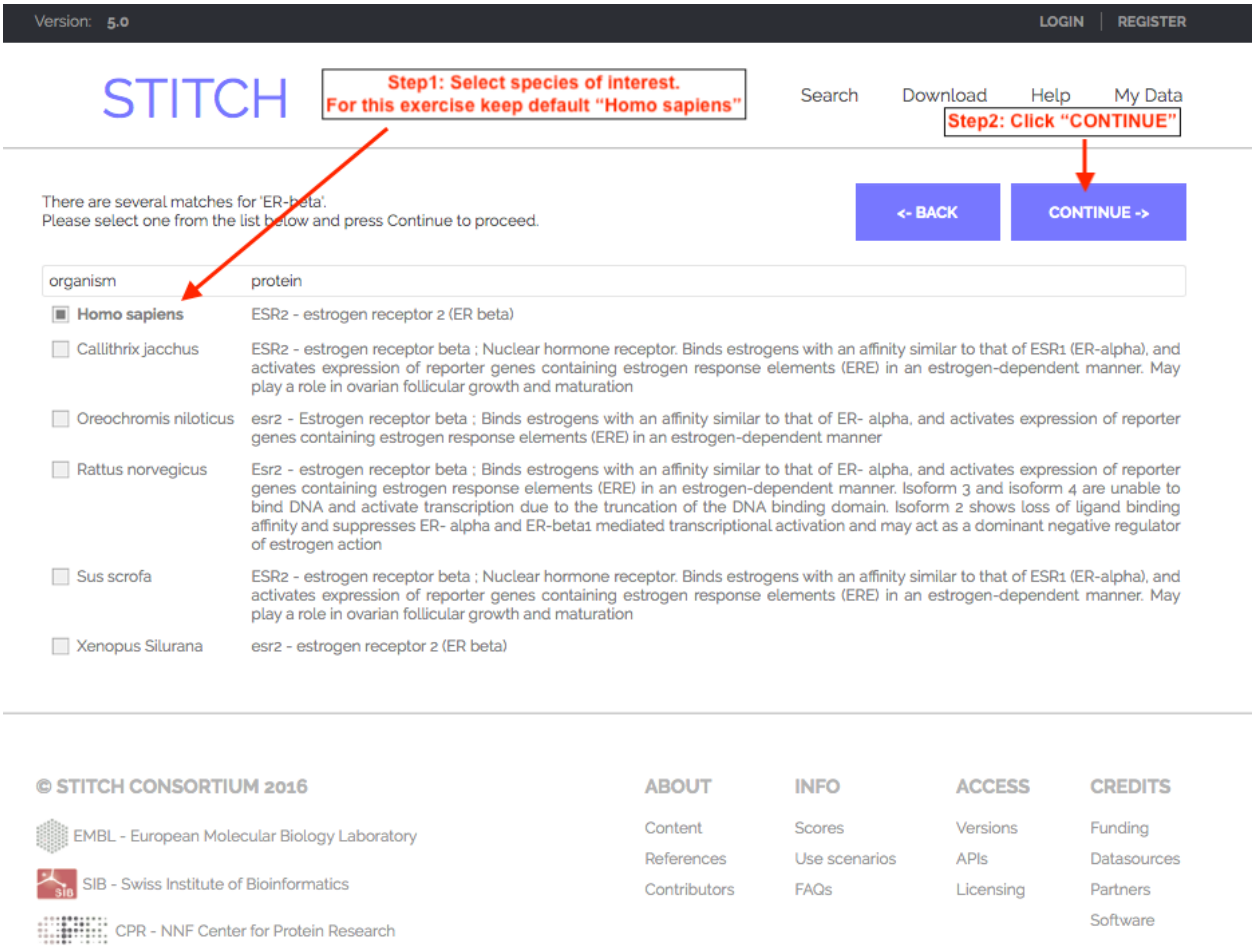
For the above enrichment analysis, the following statistical background is assumed: Whole Genome ⌄ UPDATE

2.6.3 Search by protein name

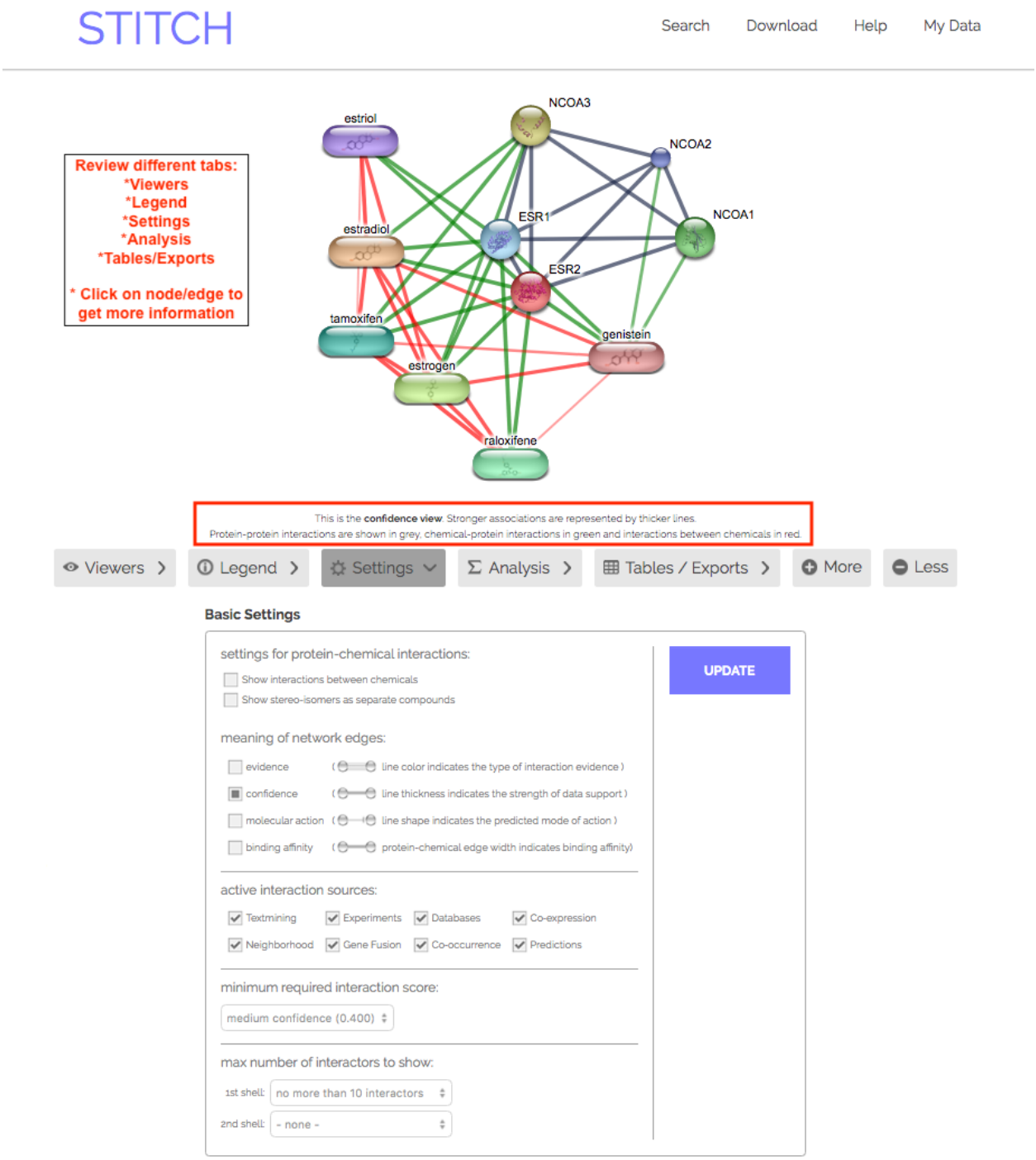
- ER-beta (protein)
  - Select “Item by name” in the left side menu
  - Type “ER-beta” in the “Item Name” search box (leave default “Organism: auto-detect”)
  - Click “SEARCH” button



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



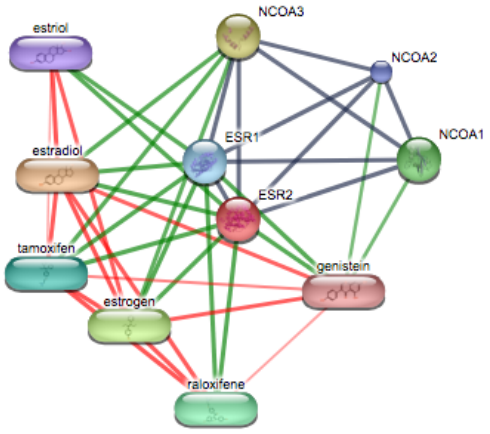
- Click on “Settings”
  - Review different settings for protein-chemical interaction



- 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.

STITCH

Search Download Help My Data



This is the **confidence view**. Stronger associations are represented by thicker lines.
Protein-protein interactions are shown in grey, chemical-protein interactions in green and interactions between chemicals in red. Links between chemicals are not used to extend the network.

Viewers › Legend › Settings › Analysis ⌄ Tables / Exports › More Less

Protein-Protein Network Stats

number of nodes: 5
number of edges: 10
average node degree: 4
clustering coefficient: 1

expected number of edges: 5
PPI enrichment p-value: 0.0134
your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0030518	intracellular steroid hormone receptor signaling pathway	4	0.000562
GO:0032870	cellular response to hormone stimulus	5	0.0181
GO:0010469	regulation of receptor activity	3	0.035
GO:0016573	histone acetylation	3	0.035
GO:0030879	mammary gland development	3	0.035

(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0016922	ligand-dependent nuclear receptor binding	3	0.000173
GO:0004402	histone acetyltransferase activity	3	0.000858
GO:0090595	acetyl-CoA:L-lysine N6-acetyltransferase	3	0.000858
GO:0034056	estrogen response element binding	2	0.00162
GO:0030284	estrogen receptor activity	2	0.00176

(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04919	Thyroid hormone signaling pathway	4	8.71e-05

INTERPRO Protein Domains and Features

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR024178	Oestrogen receptor/oestrogen-related receptor	2	0.0101

Statistical background

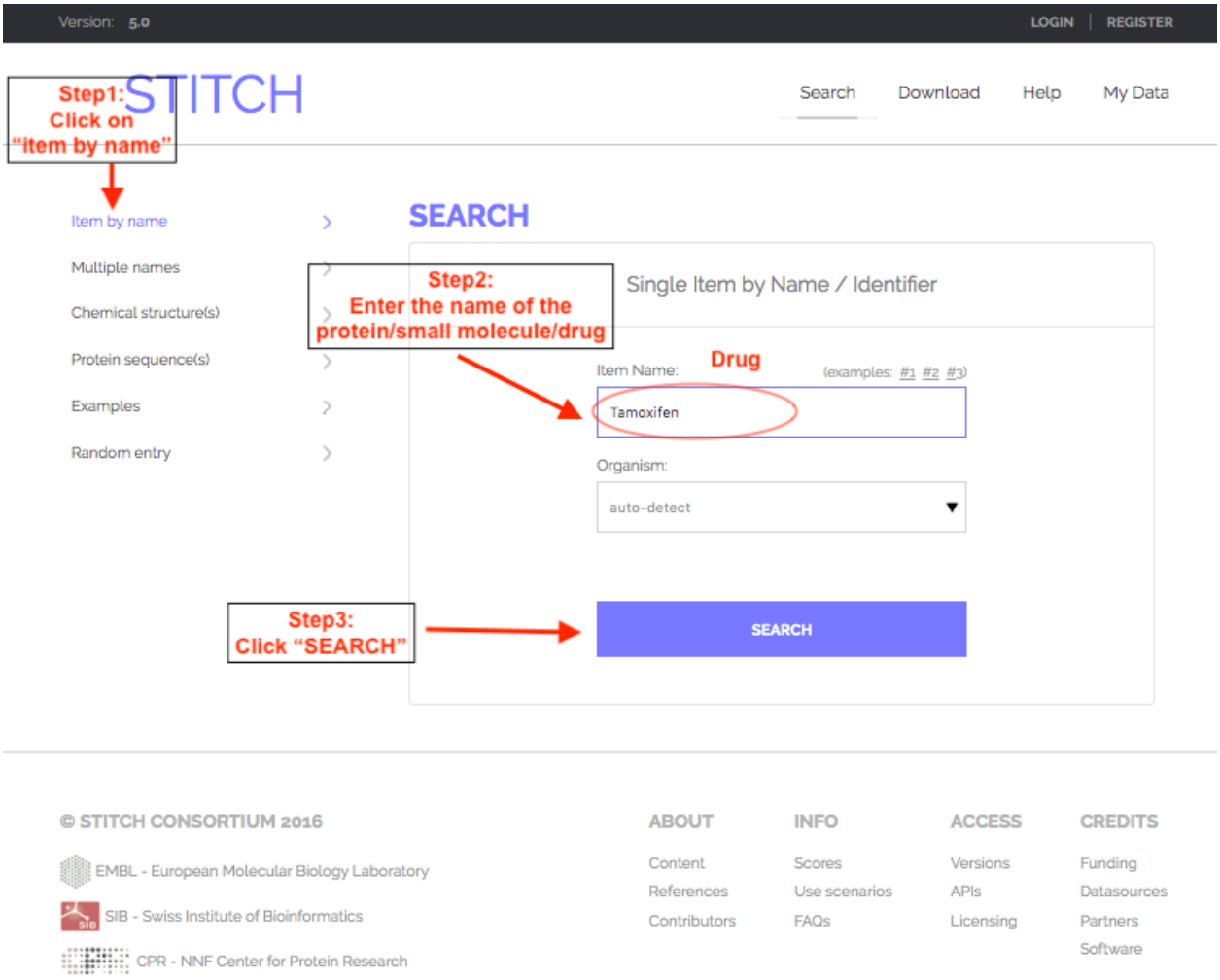
For the above enrichment analysis, the following statistical background is assumed: Whole Genome UPDATE

Save / Export

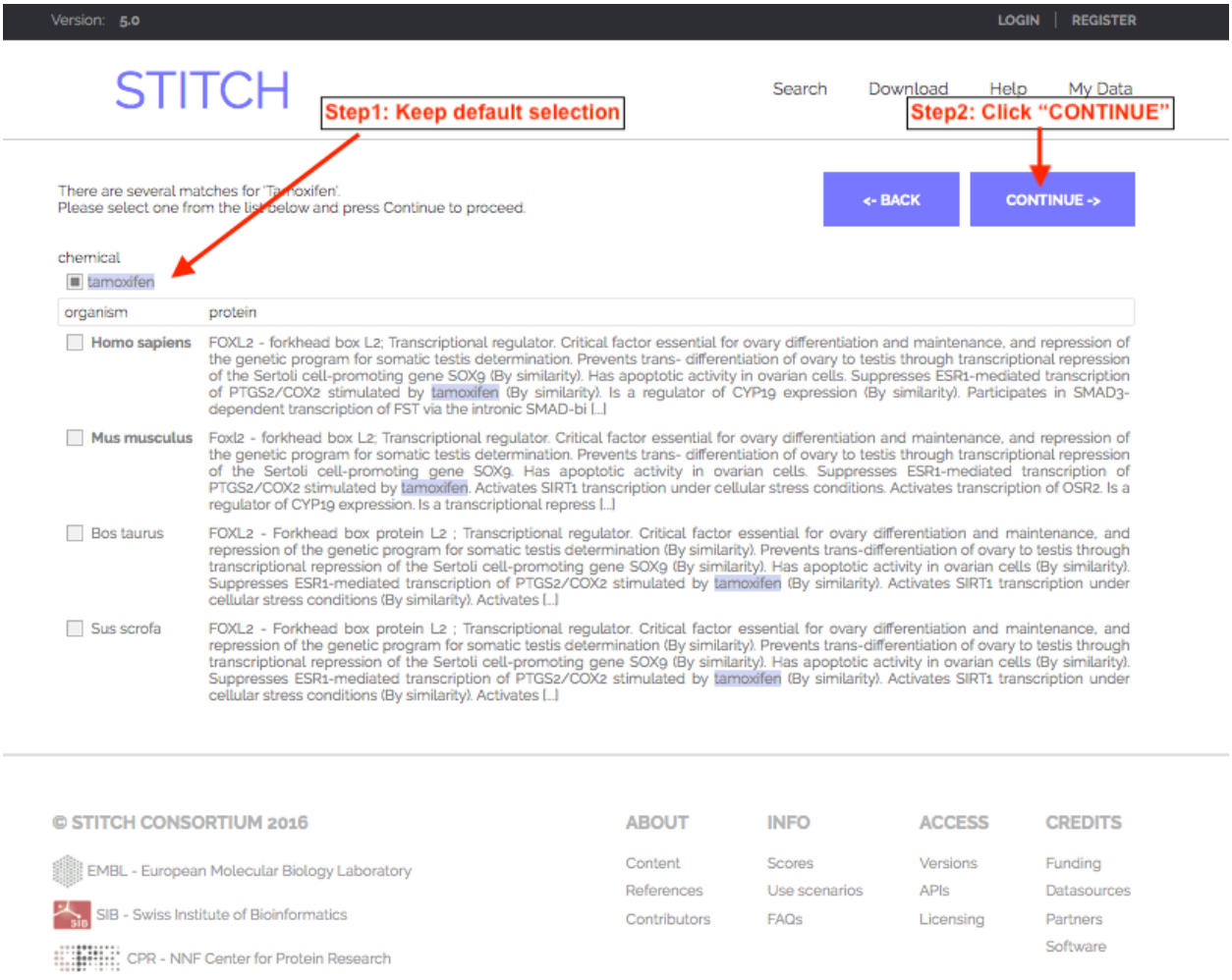
Biological Process (GO) download *14 GO-terms significantly enriched; file-format: tab-delimited*
Molecular Function (GO) download *11 GO-terms significantly enriched; file-format: tab-delimited*

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



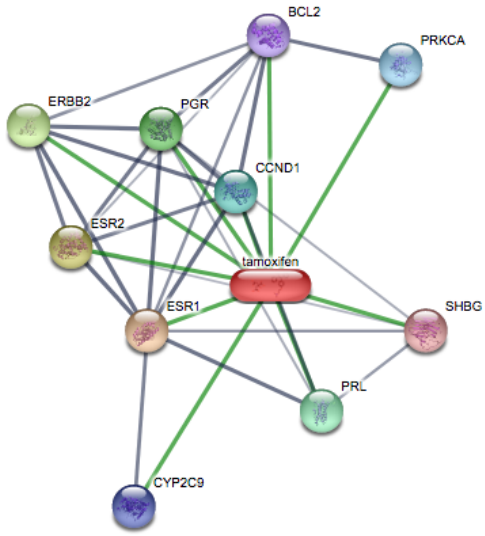
- Leave default species to “Homo sapiens”
- Click “CONTINUE” button



- 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

STITCH

Search Download Help My Data



This is the **confidence view**. Stronger associations are represented by thicker lines.
Protein-protein interactions are shown in grey, chemical-protein interactions in green and interactions between chemicals in red.

Viewers › Legend › Settings › Analysis ⌄ Tables / Exports › More Less

Protein-Protein Network Stats

number of nodes:	10	expected number of edges:	15
number of edges:	24	PPI enrichment p-value:	0.0188
average node degree:	4.8	*your network has significantly more interactions than expected (what does that mean?)*	
clustering coefficient:	0.818		

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

Biological Process (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0007565	female pregnancy	4	0.00293
GO:0009719	response to endogenous stimulus	7	0.00293
GO:0009725	response to hormone	6	0.00293
GO:0030879	mammary gland development	4	0.00293
GO:0048545	response to steroid hormone	5	0.00293
			(more ...)

Molecular Function (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0005496	steroid binding	4	0.000303
GO:0004879	RNA polymerase II transcription factor activity, ligand-activated sequence-specific DNA binding	3	0.00162
GO:0019899	enzyme binding	7	0.00162
GO:0034056	estrogen response element binding	2	0.00162
GO:0030284	estrogen receptor activity	2	0.0018
			(more ...)

2.7 Using miRNet

2.7.1 Prepare input data

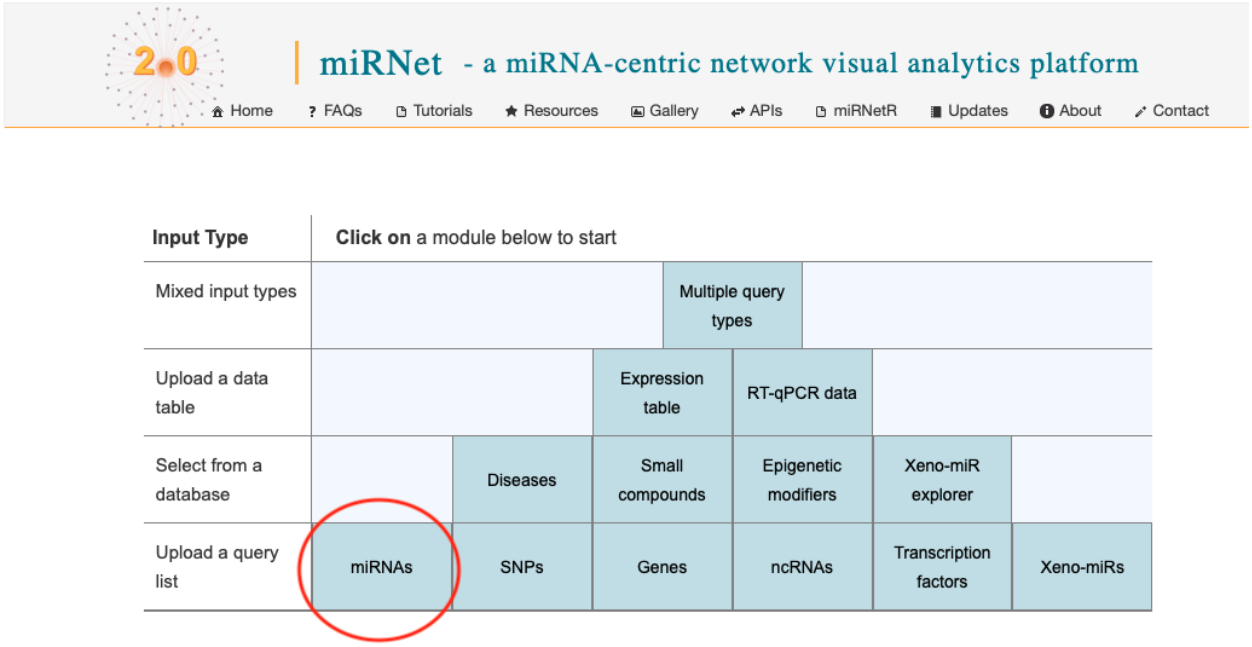
Download the miRBase IDs

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

2.7.2 Navigate to miRNet

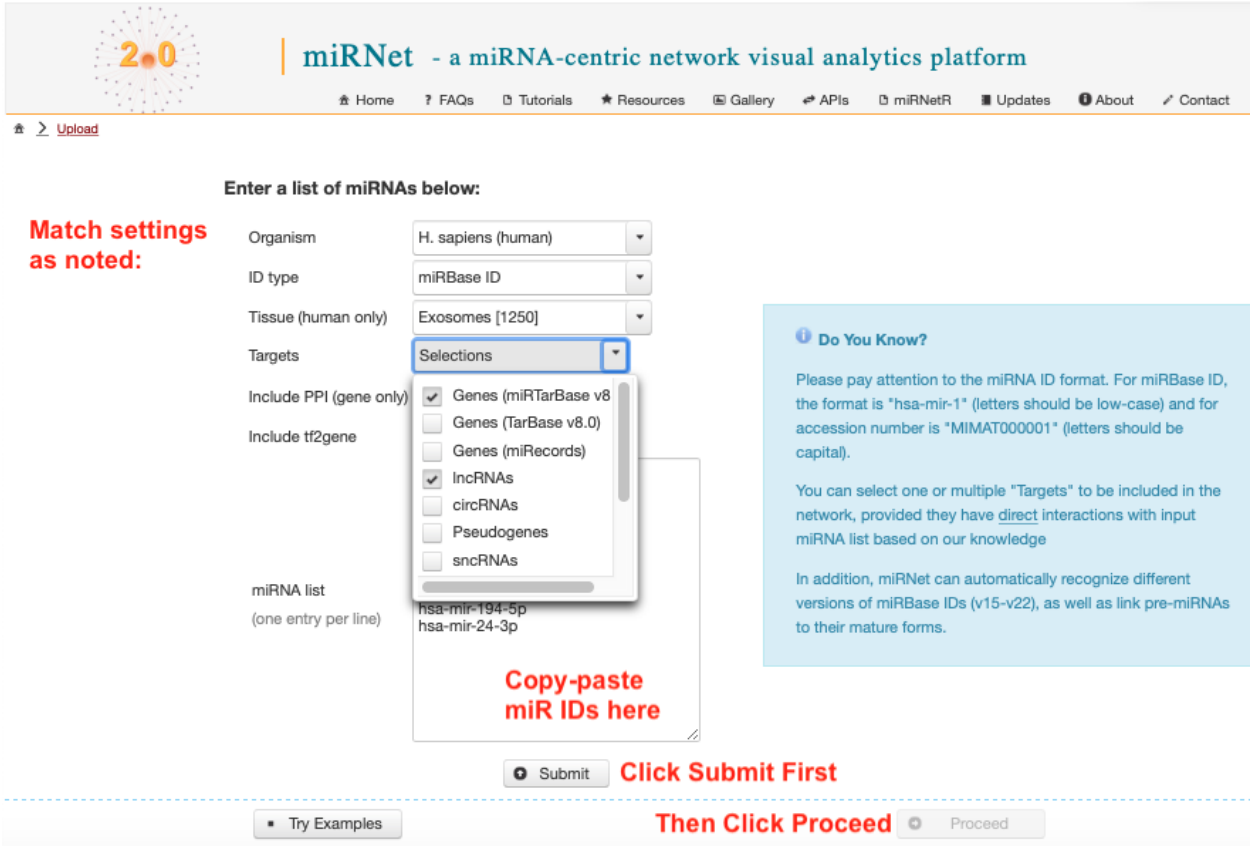
`https://www.mirnet.ca/miRNet/home.xhtml`

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



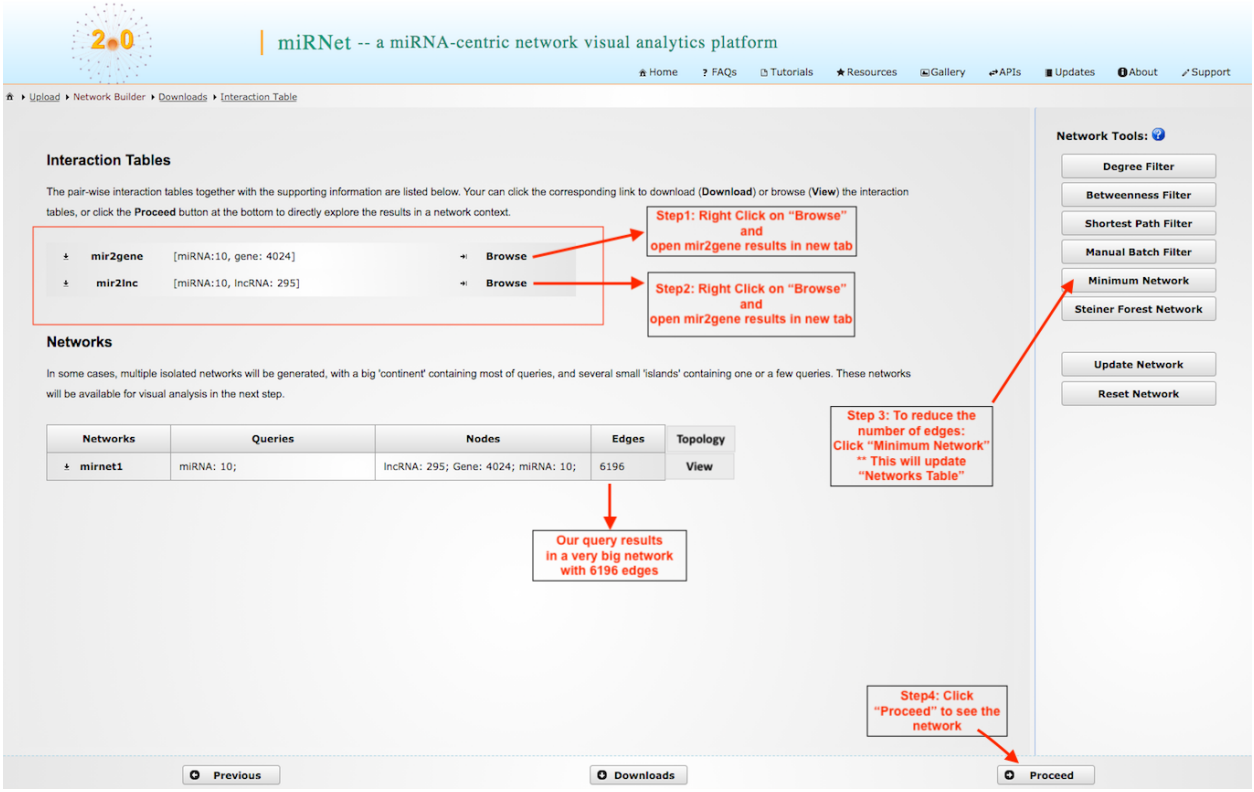
2.7.3 Perform network analysis for 10 microRNAs

- **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**



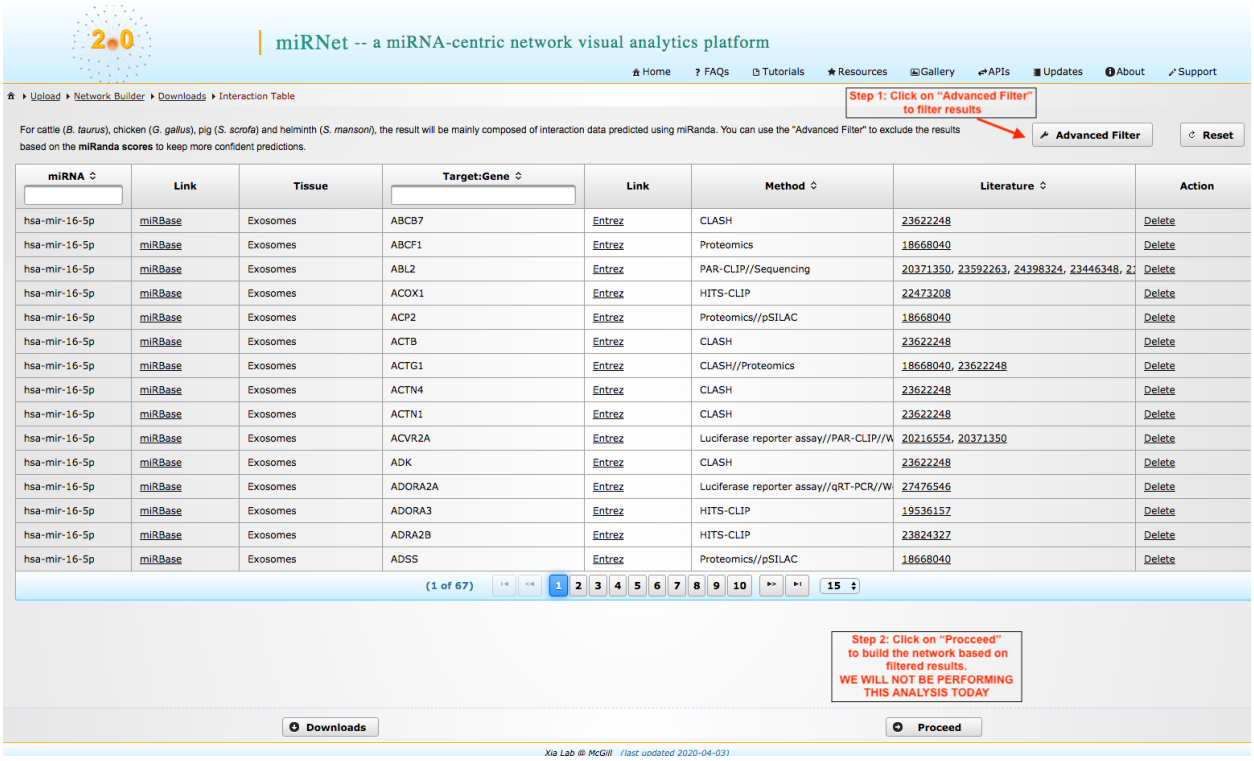
2.7.4 Results overview

- **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 minumum 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”



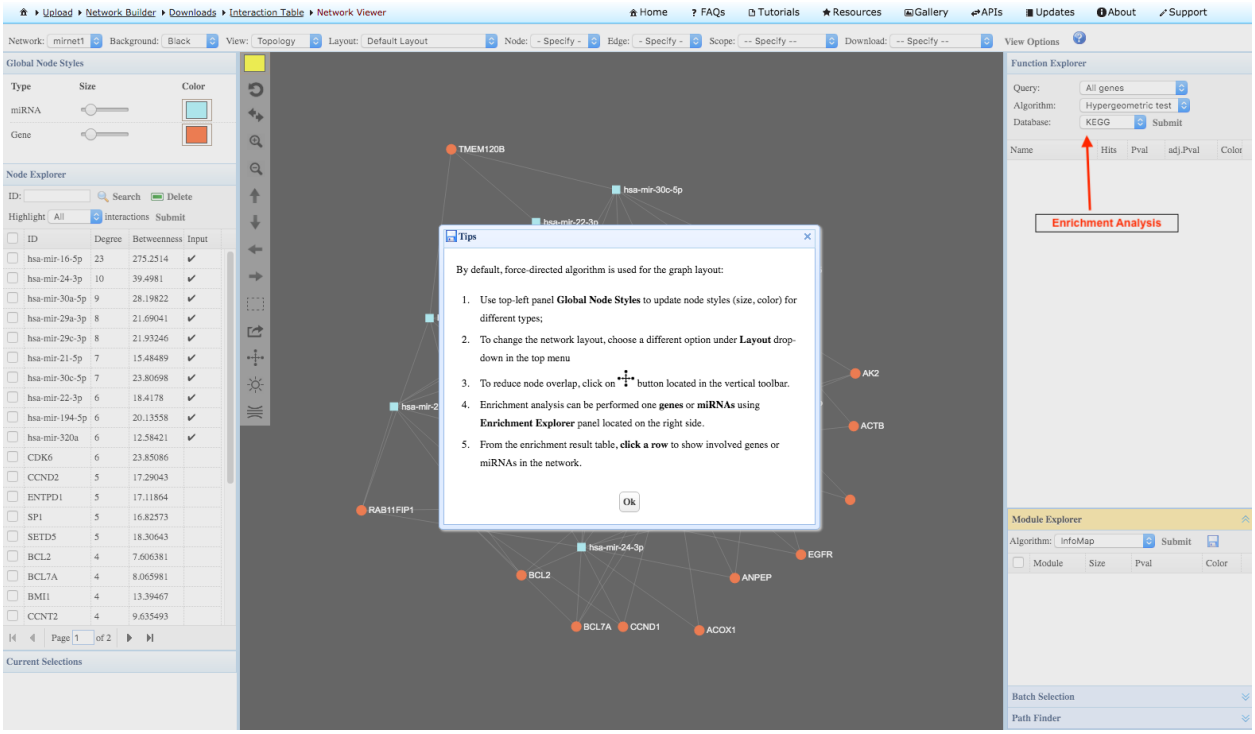
2.7.5 Review Interaction Tables results: microRNA - Gene interactions

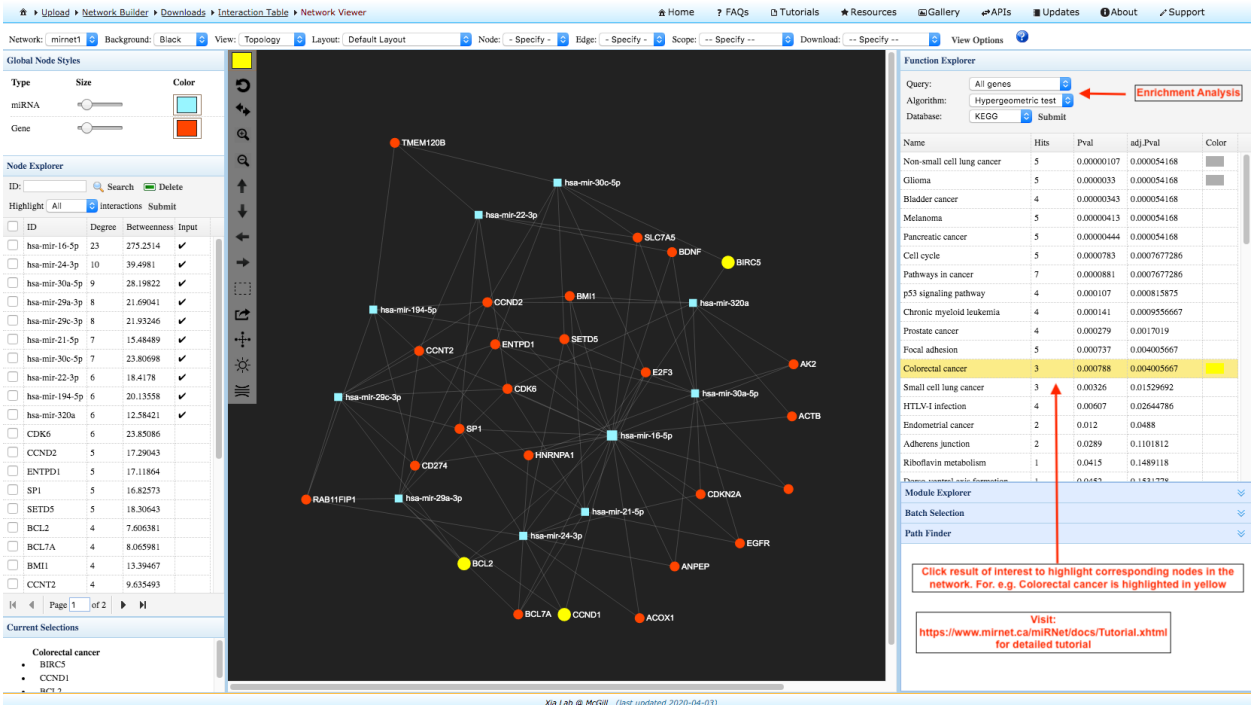
- microRNA - Gene interactions (from "Browse" button)
  - Review Advanced filter for different options to filter based on Target, Method, Literature, etc.
  - Filtered results can be used further downstream to do network analysis



2.7.6 Review network

- Review “minimum network”
- Enrichment analysis options

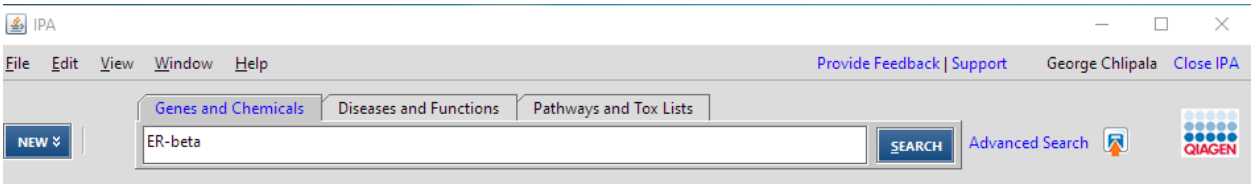




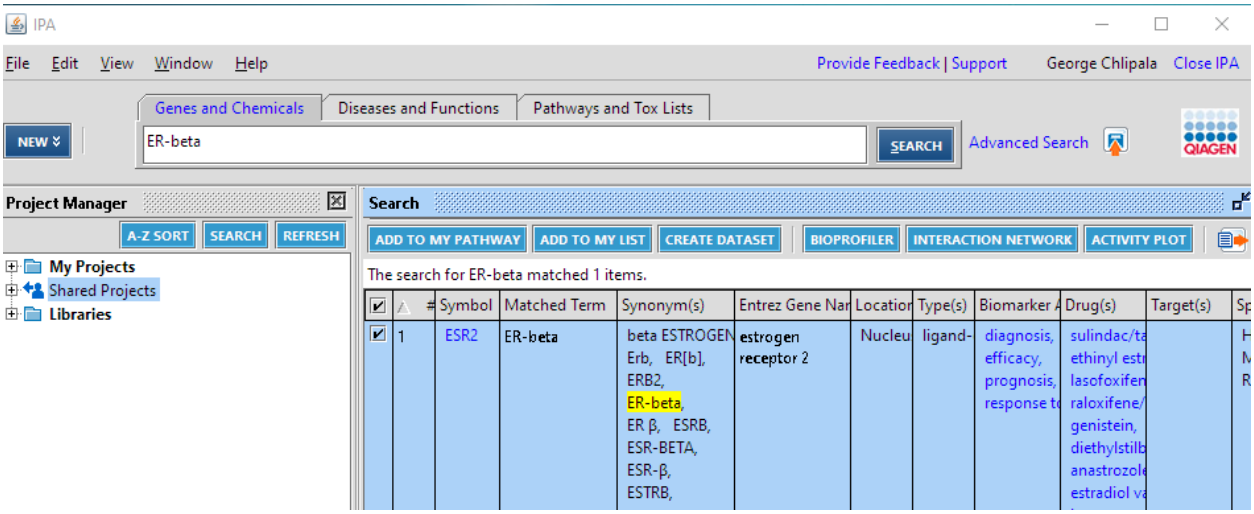
2.8 *INSTRUCTOR DEMONSTRATION*: Network analysis using IPA

2.8.1 Build (Grow) a network from a single molecule

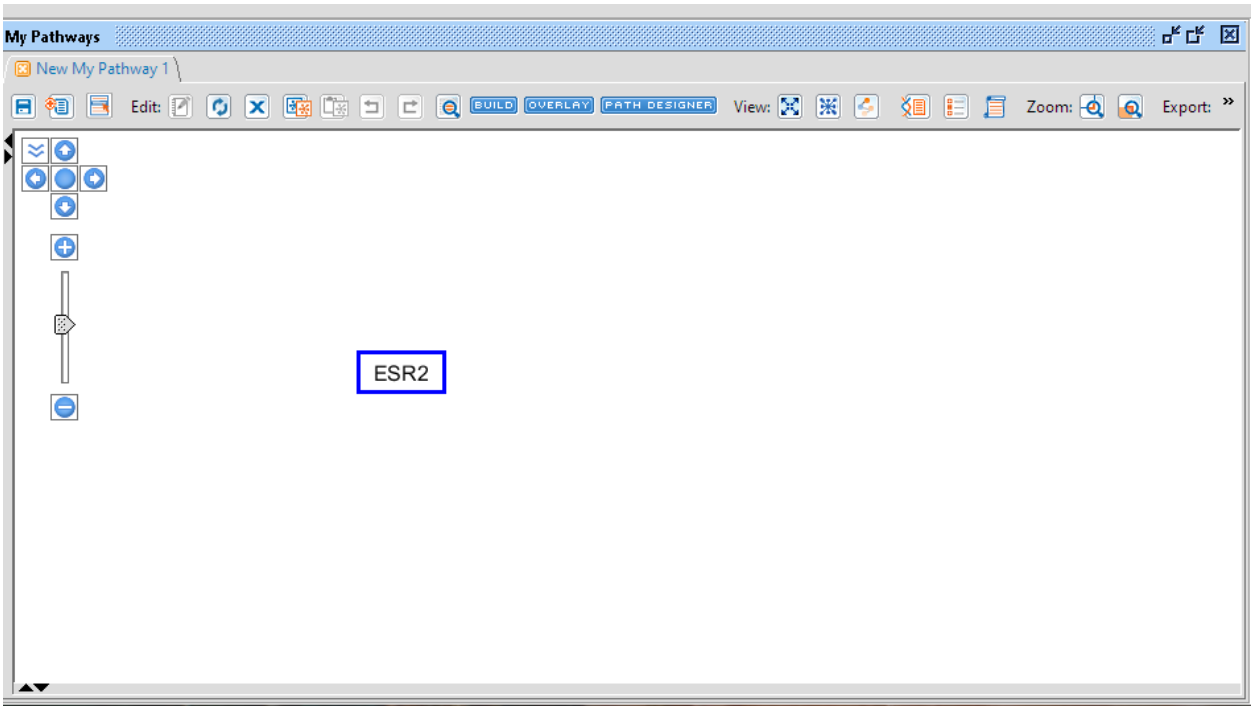
1. Enter "ER-beta" in search dialog and click **Search**.



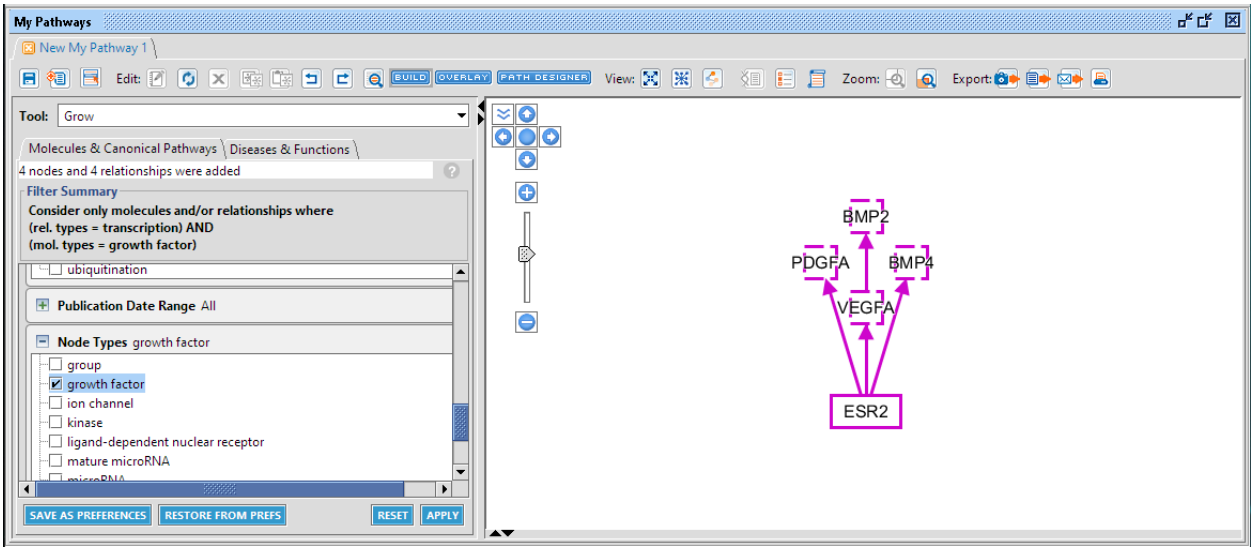
2. Select "ESR2" and click **Add To My Pathway** and select *New My Pathway*.



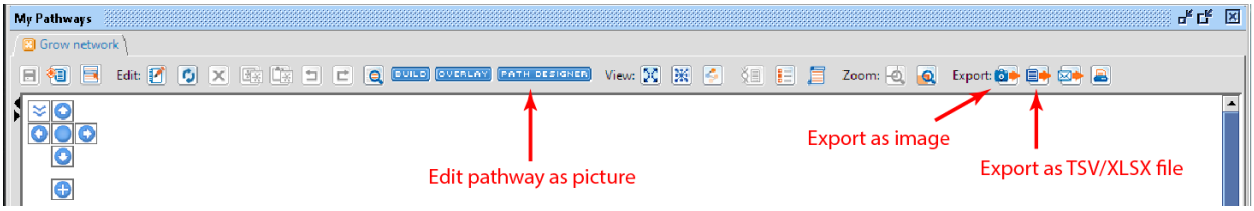
3. In the *Network* view click on the **ESR2** node (rectangle) and click **BUILD** button.



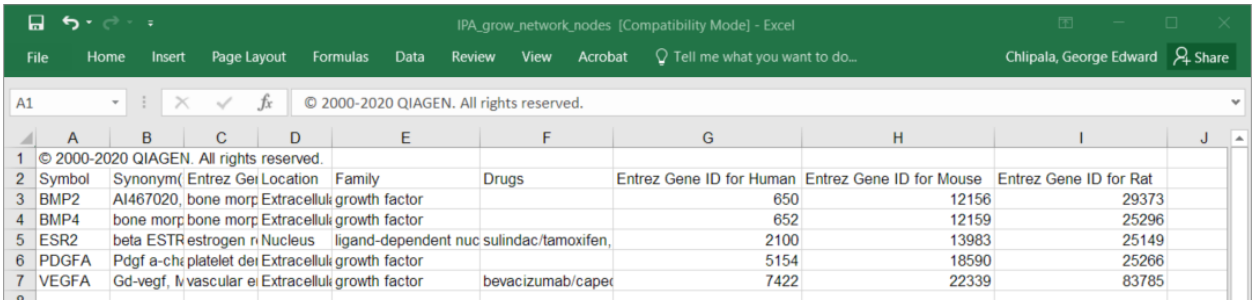
4. Select *Grow* as the Tool.
    a. In *General Settings*, set Grow out… …that are "Downstream of selected molecules"
    b. In *Relationship Types*, Click "Select all" to unselect all types and scroll down to select "transcription"
    c. In *Node Types*, Click "Select all" to unselect all types and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



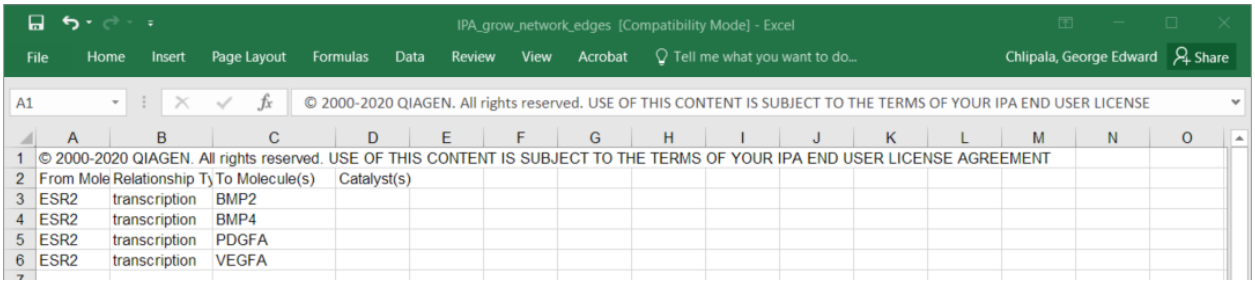
5. Export molecules and relationships



- Export *All Molecules*

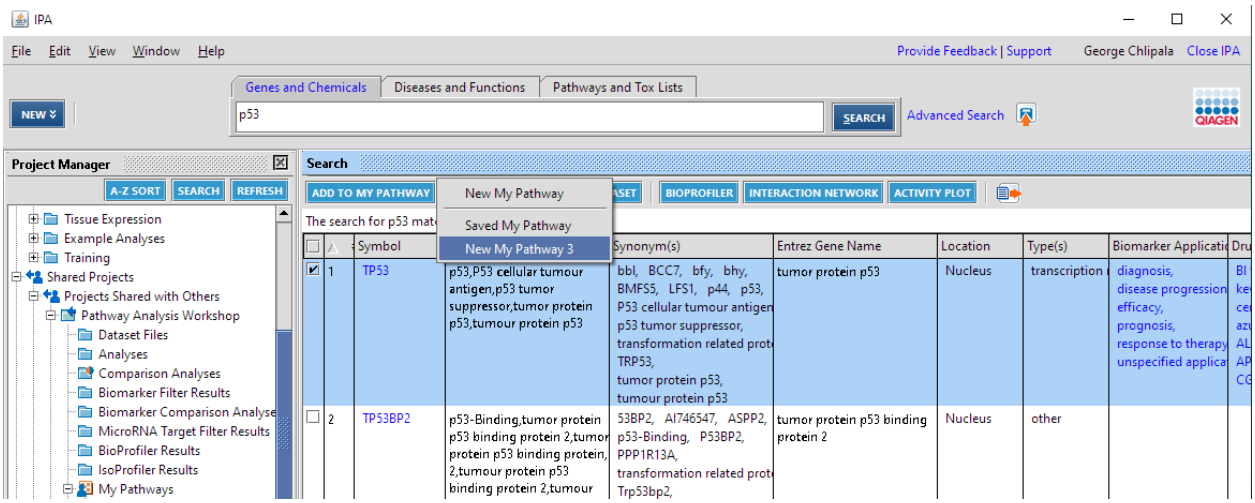


- Export *All Relationships*

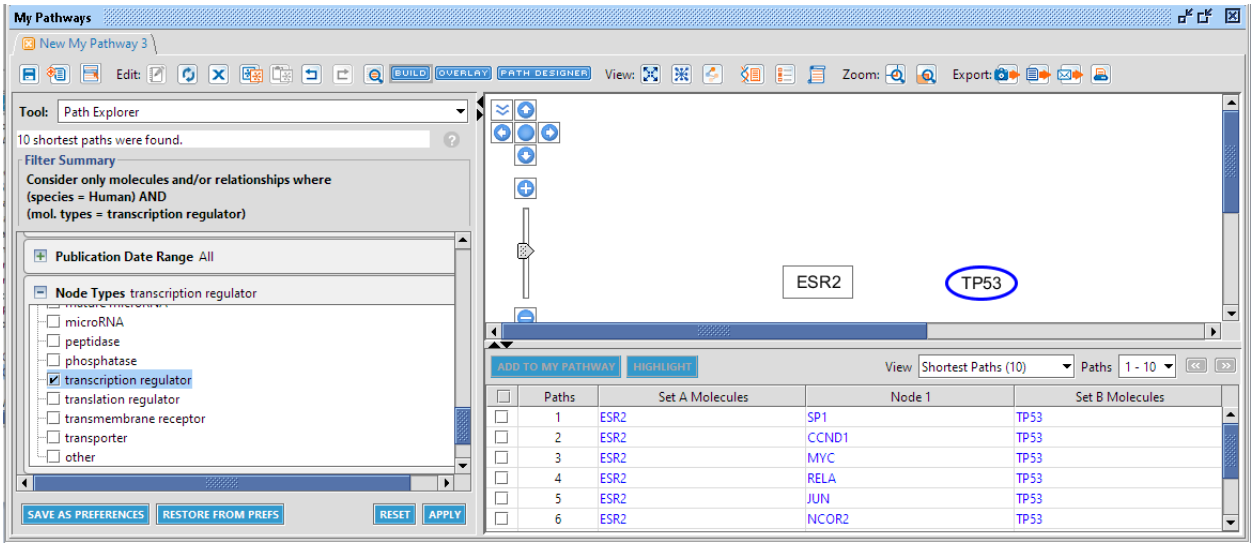


2.8.2 Find connections between molecules

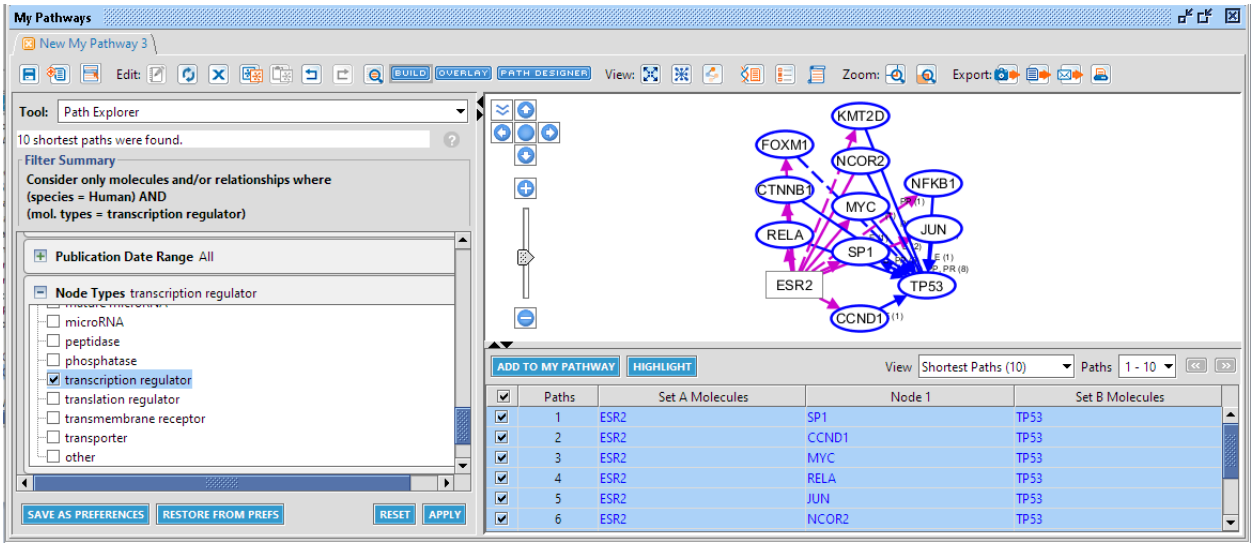
1. Enter “ER-beta” in search dialog and click **Search**.
2. Select “ESR2” and click **Add To My Pathway** and select *New My Pathway*.
3. Enter “p53” in search dialog and click **Search**
4. Select “TP53” and click **Add To My Pathway** and select bottom *New My Pathway*.



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.

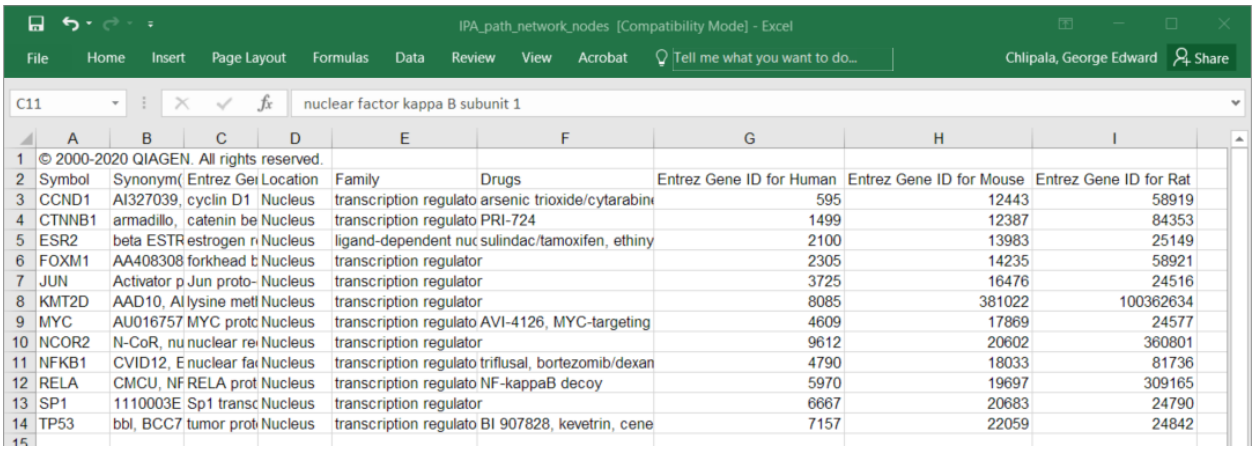


7. Click on the checkbox in the header row and click **ADD TO MY PATHWAY**



5. Export molecules and relationships

- Export *All Molecules*



- Export *All Relationships*