Gene Set Analysis –Methods and Tools
Exercise 2.1

Antonio Mora

Xie Chengshu

December, 2018

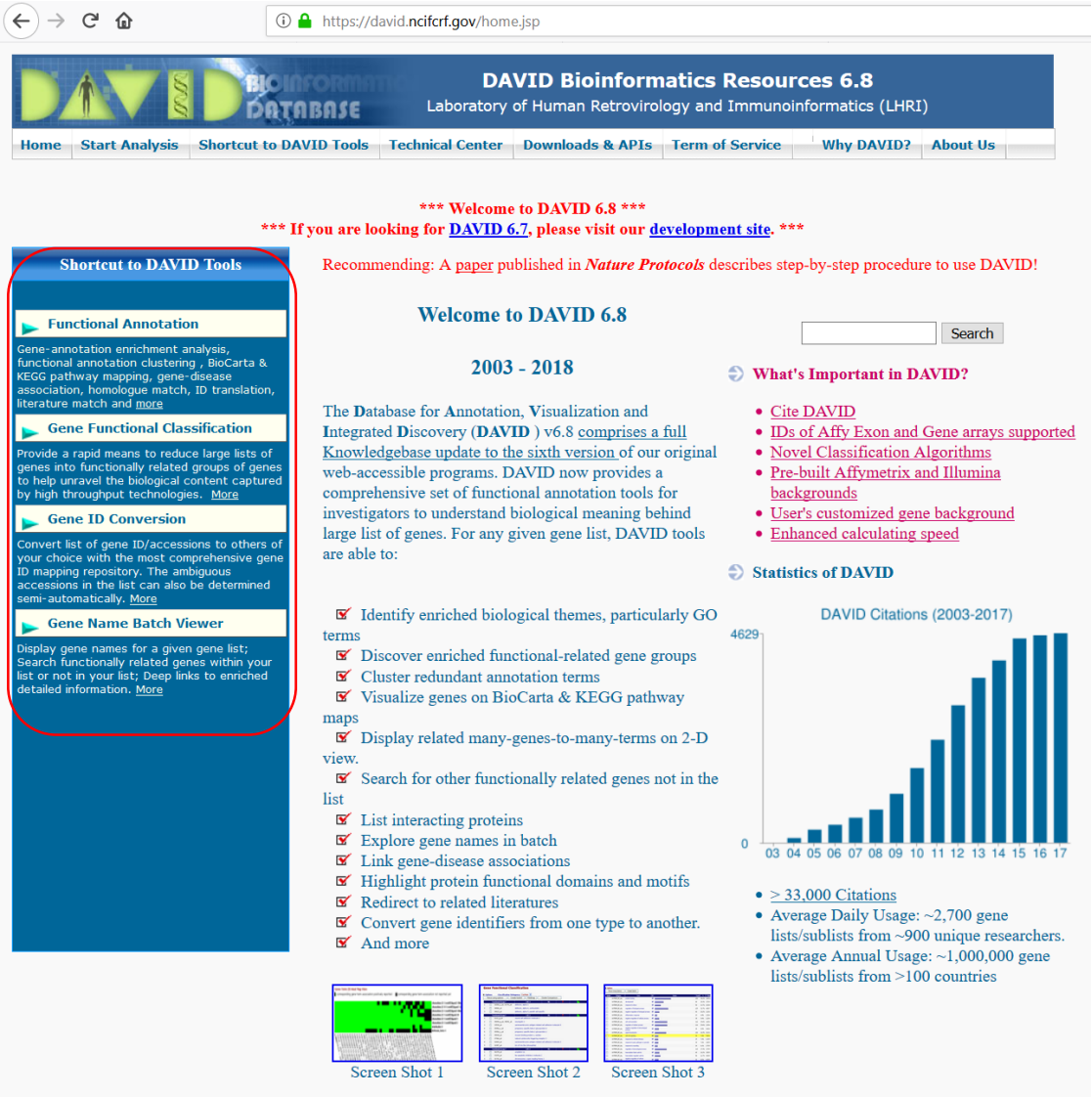
MORA LAB
FUNCTIONAL BIOINFORMATICS

Exercise 1. Using DAVID

DAVID is the acronym for “The Database for Annotation, Visualization and Integrated Discovery”. You can find it at: https://david.ncifcrf.gov/home.jsp

The picture below is its main page, which contains some general information about this platform. DAVID provides four main tools (details on the website):

1. Functional Annotation
2. Gene Functional Classification
3. Gene ID Conversion
4. Gene Name Batch Viewer



1. Upload datasets

Click on the “Start Analysis” button.

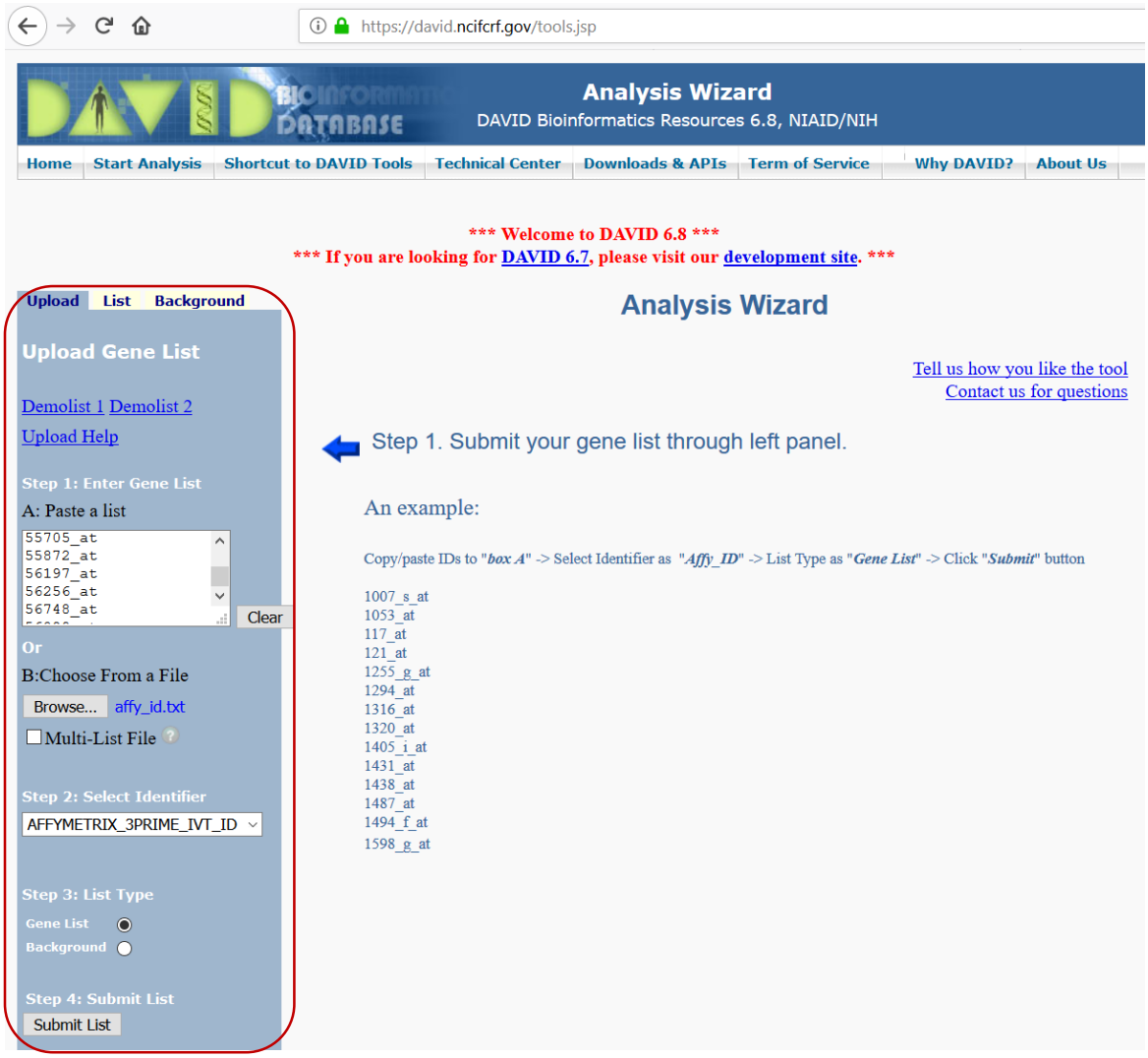
On the left panel of the page, there will be 3 steps:

1. Paste the gene list or choose a gene list file to upload. There are two ways to upload your gene list. One is to load a gene list from a file, another is to paste a gene list to the text box. Here we can upload the “affy_id.txt” file. Regarding the limitations of gene lists, please see DAVID FAQs. (https://david.ncifcrf.gov/content.jsp?file=FAQs.html).

2. Select the ID format, according to the format of the gene list. Here we use “affymetrix ID”.

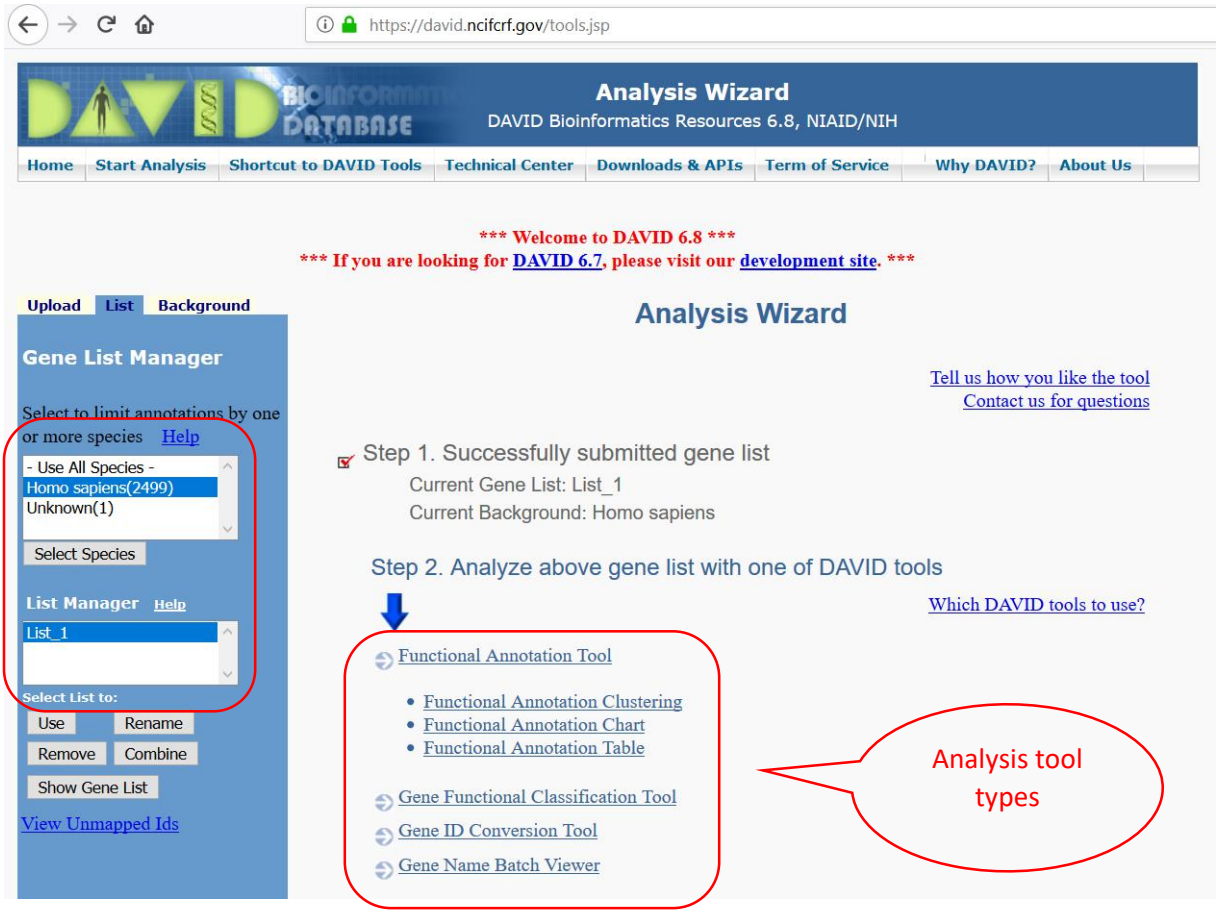
3. The list type may be a gene list or using a list as background. We choose the “gene list”.

At last, click on the “Submit List” button.



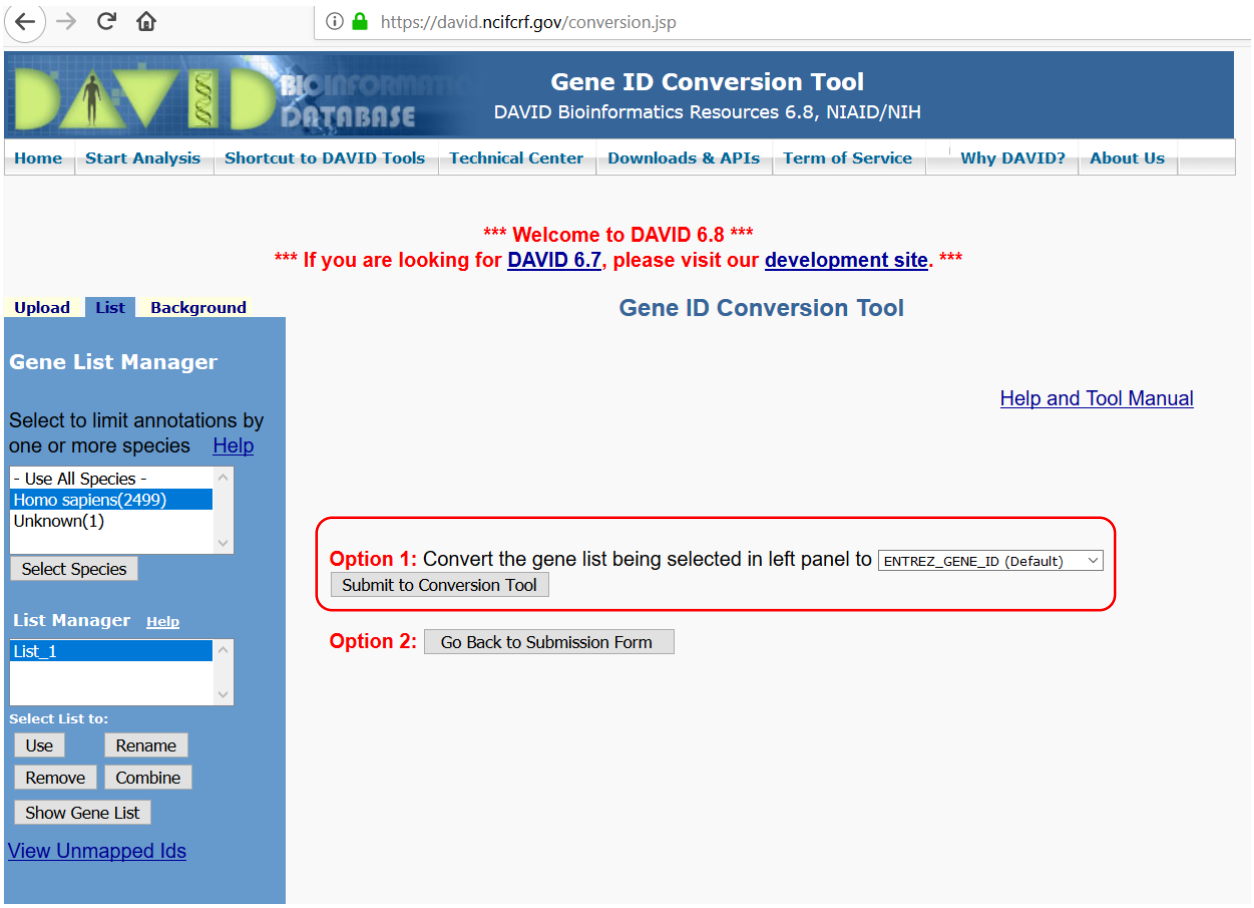
2. Use DAVID tools

After task submission, the left panel shows the summary of the submitted gene list. The different available tools can be found under “Step 2”.

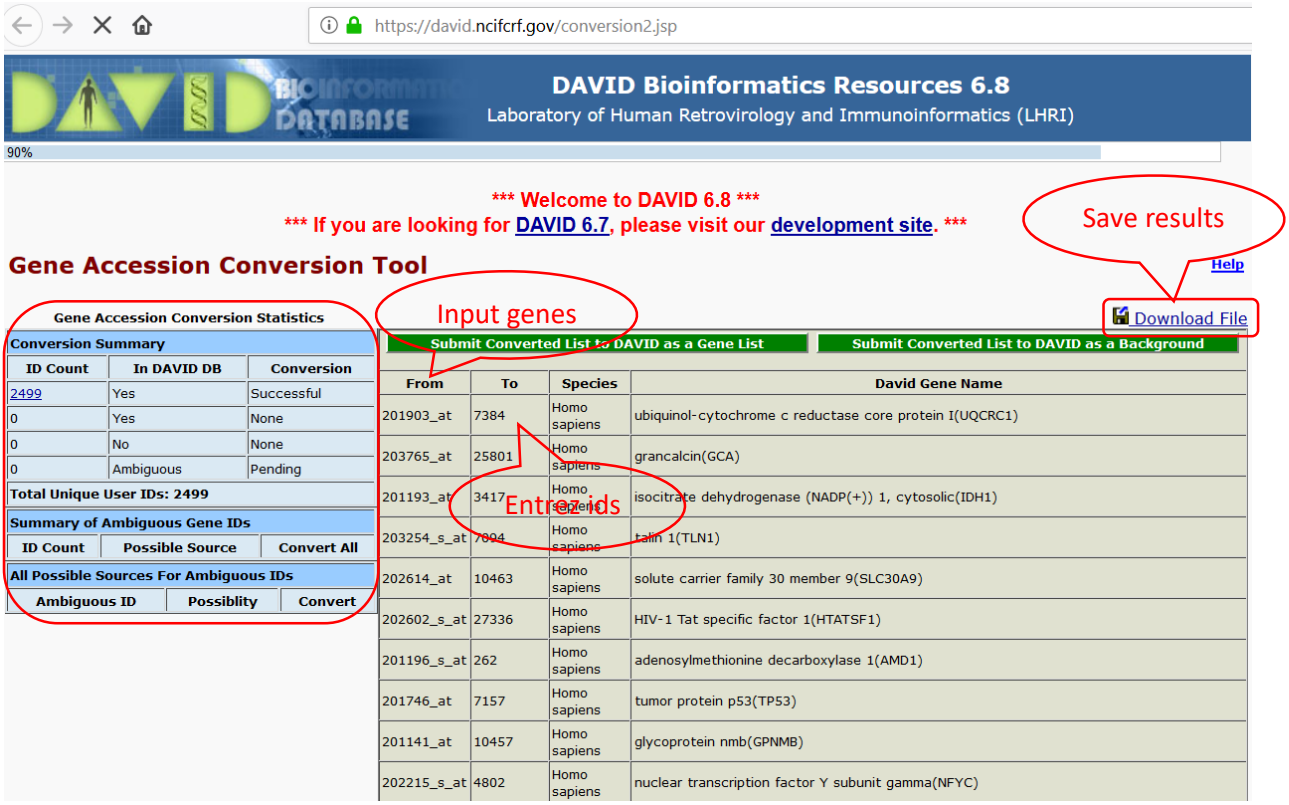


Tool 1. ID conversion

We can click on the “Gene ID Conversion Tool”, go to the new page, and select a new ID format (Entrez_Gene_ID). In the left panel we find that there are 2499 genes from our uploaded gene list that can be found in the DAVID database, and 1 that cannot be found. Click on the “Submit to conversion tool” button.

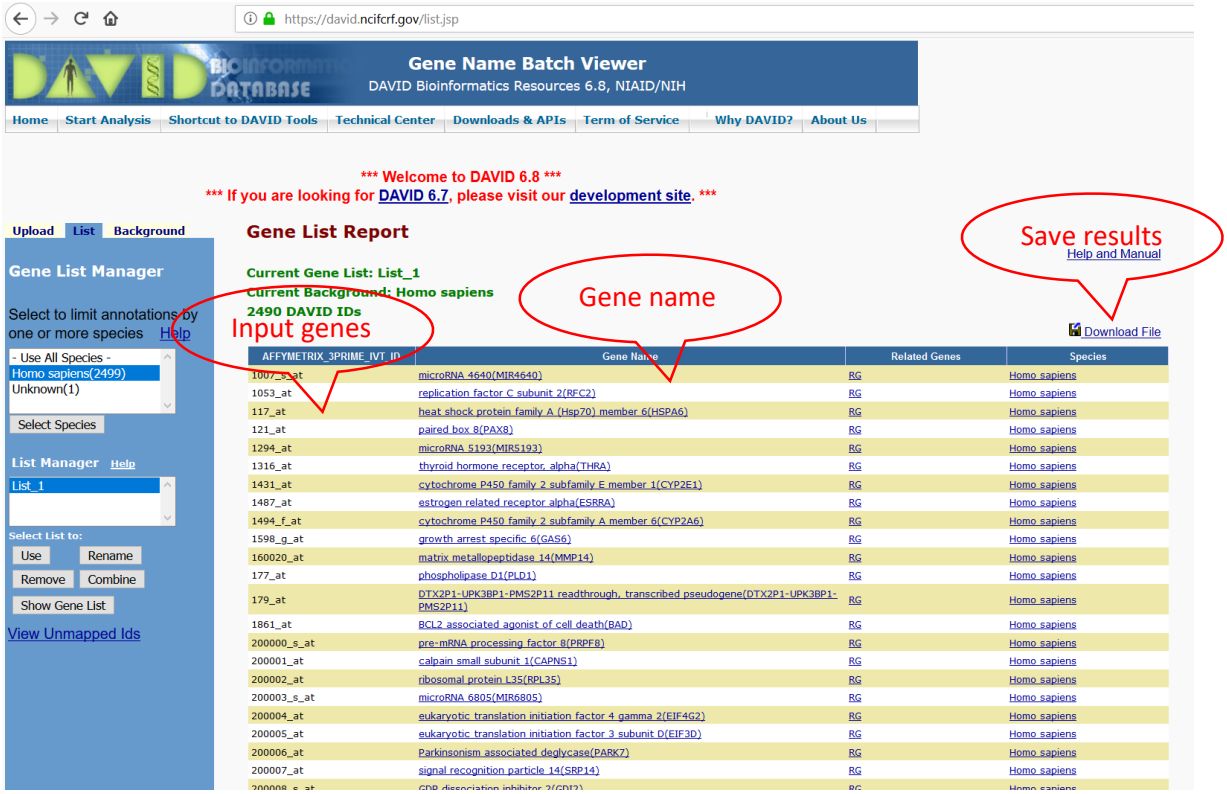


On the left side of the ID conversion result page, there is a summary table of the gene list conversion. In the table, there are 2499 affymetrix IDs converted into Entrez Gene IDs in DAVID database. On the right top corner, a “download file” option allows to download the whole conversion file.



Tool 2. Gene Name Batch Viewer

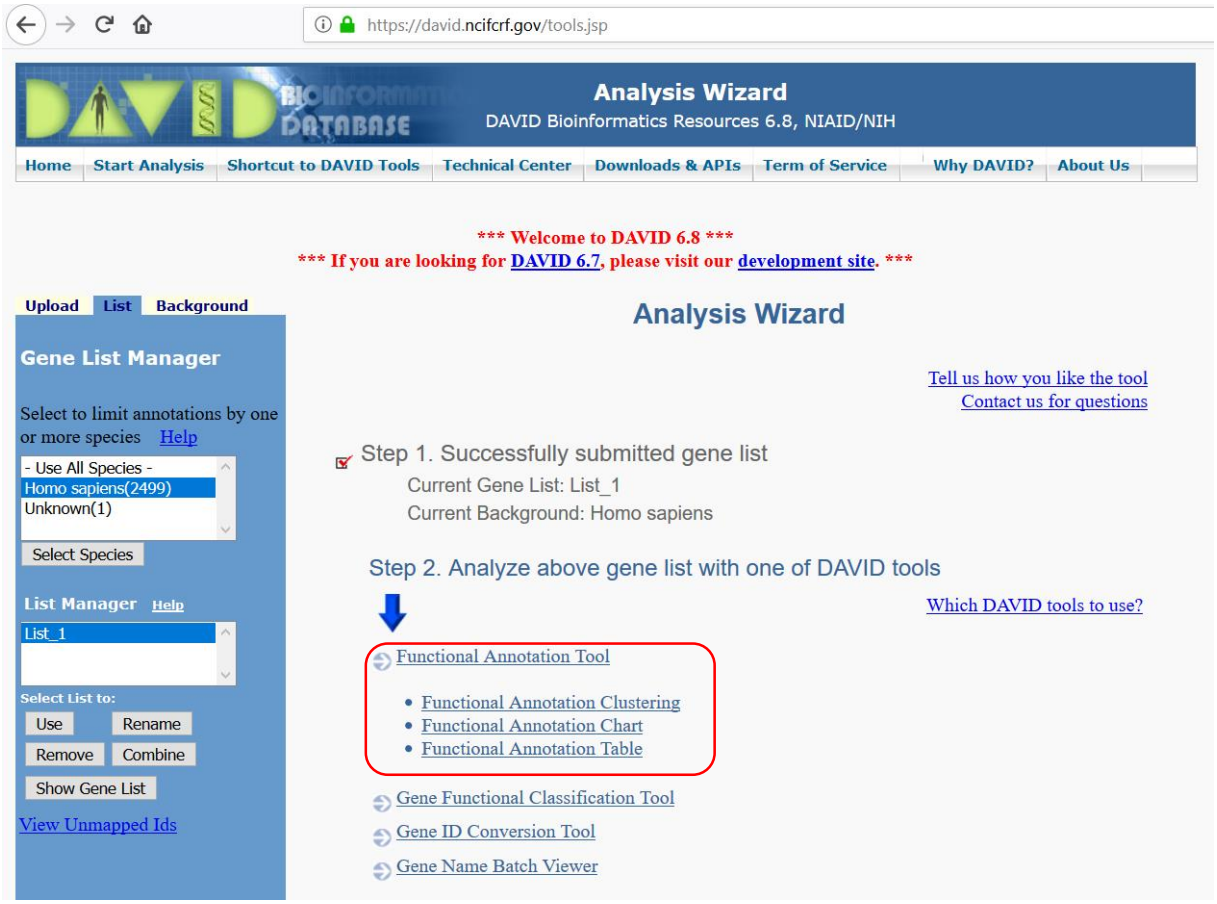
This tool converts gene list IDs into gene names directly. Click the “Gene Name Batch Viewer” under the list of “Shortcut to DAVID tools”.



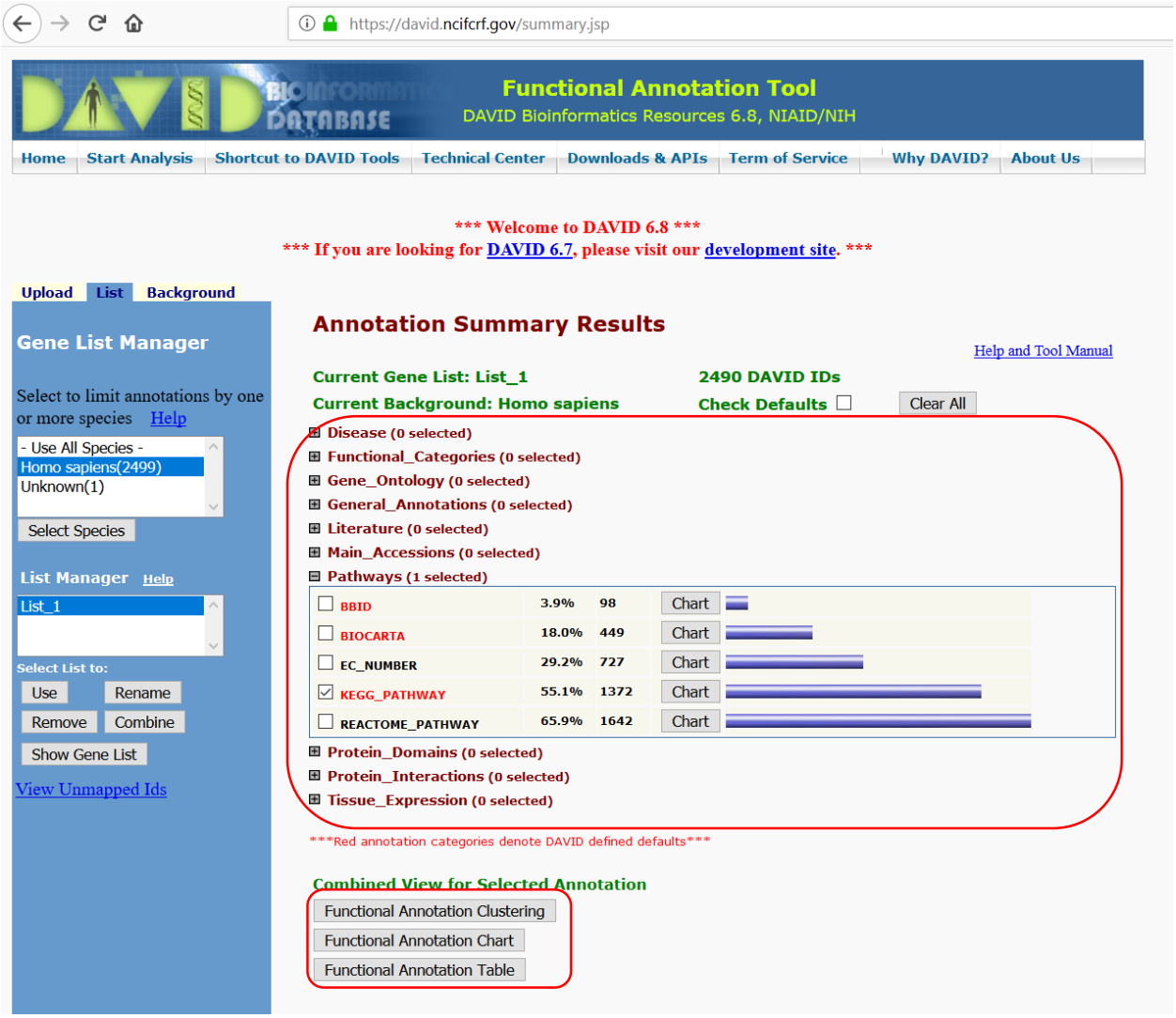
Q1: What are the gene names of the genes with Affy_id : “1053_at” and “200010_at”?

Tool 3. Functional Annotation Tool

Go back to the previous page or choose “shortcut to DAVID Tools”—“Functional Annotation Tool”.



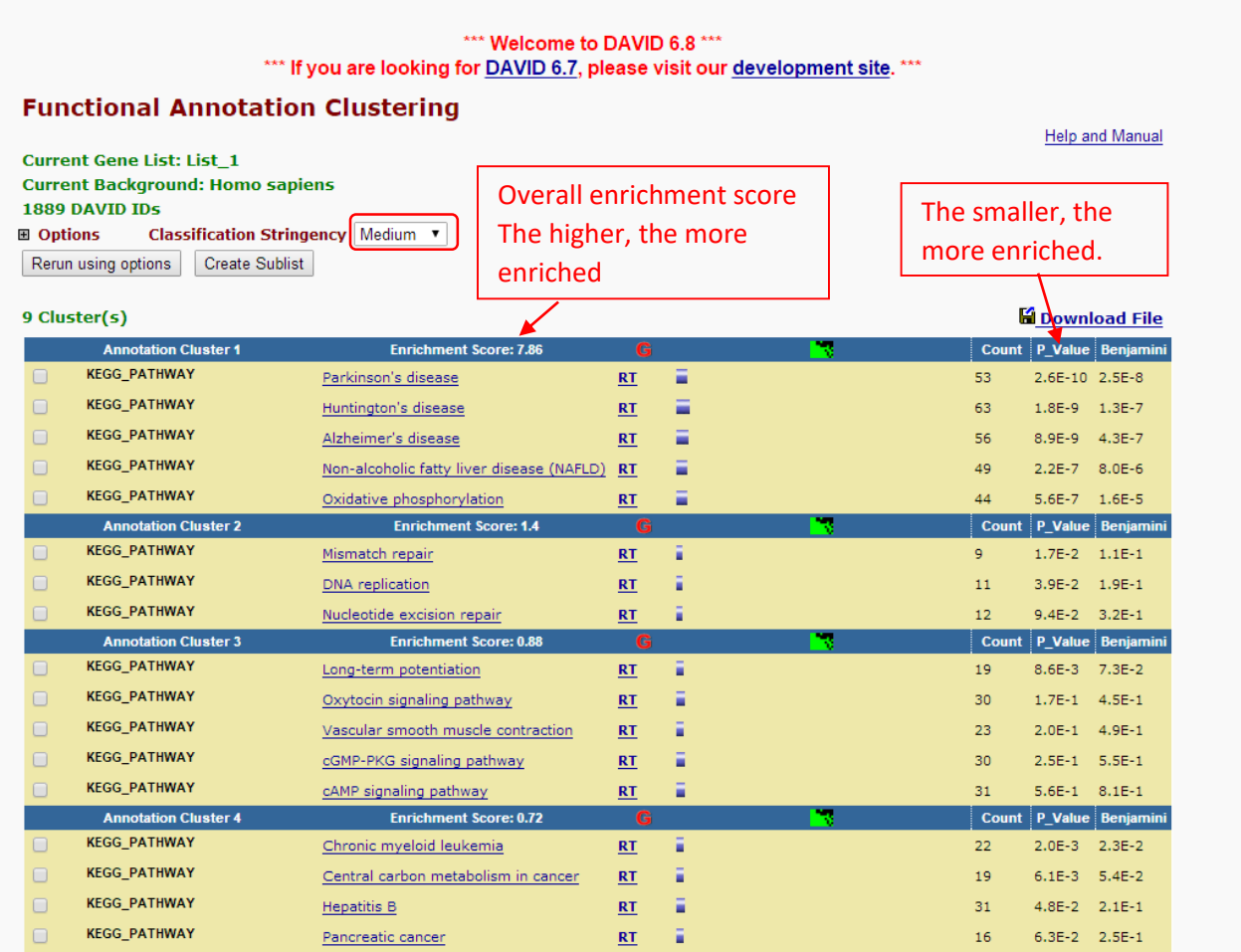
The functional annotation tool includes three options: Functional Annotation Clustering, chart and table. Click the “Functional Annotation Tool”, go to the new page, and choose the annotation we want (Gene Ontology and KEGG pathway for this exercise).



Choose the KEGG pathway analysis only, and open the KEGG pathway chart.

Q2: What are the 3 most significant KEGG pathways? What are their p-values? Open them in the KEGG website.

Now choose “Functional Annotation Clustering“. The results show that pathways can be combined into 9 clusters.



*** Welcome to DAVID 6.8 ***

*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1

Current Background: Homo sapiens

1889 DAVID IDs

Options Classification Stringency: Medium

Rerun using options Create Sublist

Overall enrichment score
The higher, the more enriched

The smaller, the more enriched.

9 Cluster(s)

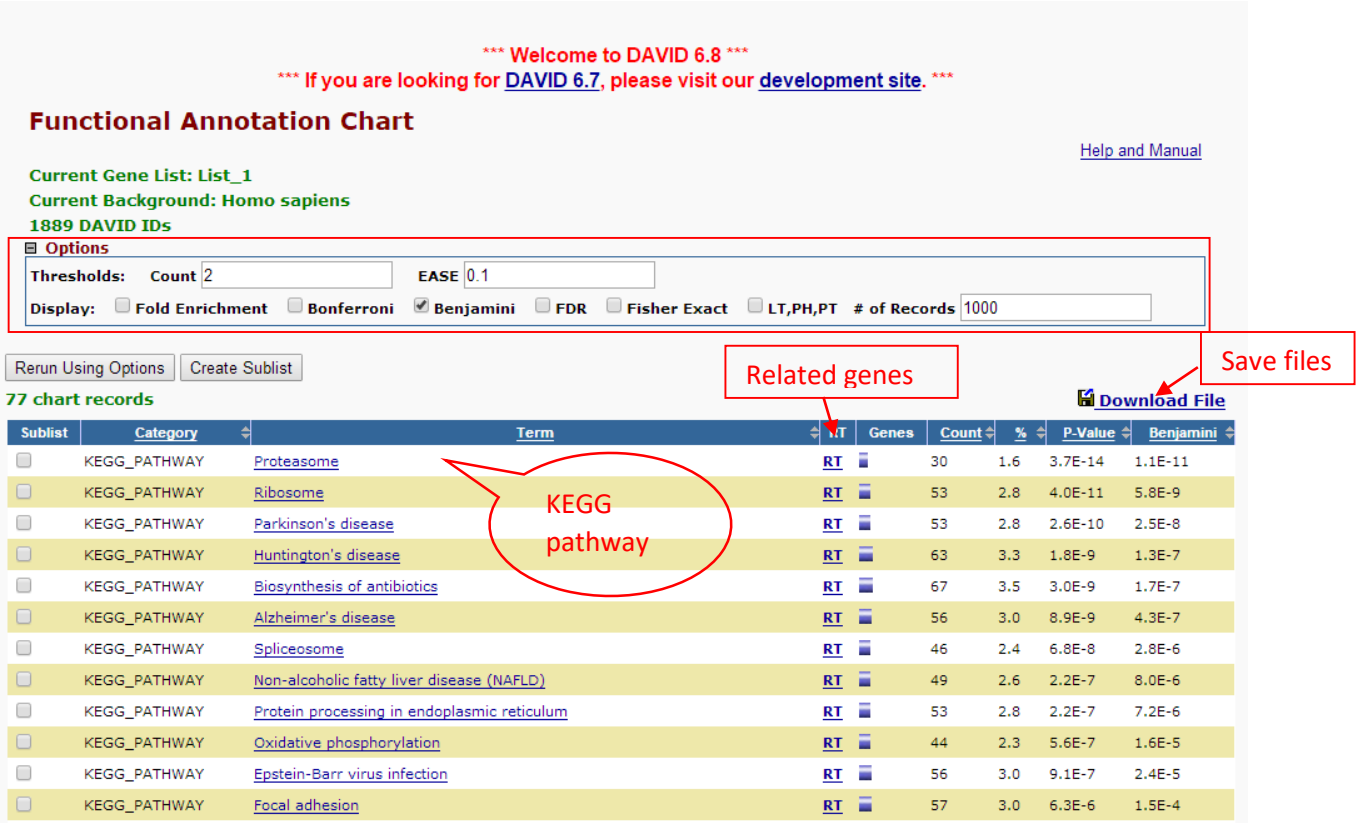
Download File

Annotation Cluster 1	Enrichment Score: 7.86					Count	P_Value	Benjamini
KEGG_PATHWAY	Parkinson's disease	RT				53	2.6E-10	2.5E-8
KEGG_PATHWAY	Huntington's disease	RT				63	1.8E-9	1.3E-7
KEGG_PATHWAY	Alzheimer's disease	RT				56	8.9E-9	4.3E-7
KEGG_PATHWAY	Non-alcoholic fatty liver disease (NAFLD)	RT				49	2.2E-7	8.0E-6
KEGG_PATHWAY	Oxidative phosphorylation	RT				44	5.6E-7	1.6E-5
Annotation Cluster 2	Enrichment Score: 1.4					Count	P_Value	Benjamini
KEGG_PATHWAY	Mismatch repair	RT				9	1.7E-2	1.1E-1
KEGG_PATHWAY	DNA replication	RT				11	3.9E-2	1.9E-1
KEGG_PATHWAY	Nucleotide excision repair	RT				12	9.4E-2	3.2E-1
Annotation Cluster 3	Enrichment Score: 0.88					Count	P_Value	Benjamini
KEGG_PATHWAY	Long-term potentiation	RT				19	8.6E-3	7.3E-2
KEGG_PATHWAY	Oxytocin signaling pathway	RT				30	1.7E-1	4.5E-1
KEGG_PATHWAY	Vascular smooth muscle contraction	RT				23	2.0E-1	4.9E-1
KEGG_PATHWAY	cGMP-PKG signaling pathway	RT				30	2.5E-1	5.5E-1
KEGG_PATHWAY	cAMP signaling pathway	RT				31	5.6E-1	8.1E-1
Annotation Cluster 4	Enrichment Score: 0.72					Count	P_Value	Benjamini
KEGG_PATHWAY	Chronic myeloid leukemia	RT				22	2.0E-3	2.3E-2
KEGG_PATHWAY	Central carbon metabolism in cancer	RT				19	6.1E-3	5.4E-2
KEGG_PATHWAY	Hepatitis B	RT				31	4.8E-2	2.1E-1
KEGG_PATHWAY	Pancreatic cancer	RT				16	6.3E-2	2.5E-1

Q3: What do pathways have in common for annotation cluster1? What about annotation cluster 2?

Now go back and select “Functional Annotation Chart”.

The “Functional Annotation Chart” provides the clustering of genes’ annotations (KEGG pathway or others). It shows 77 chart records, which means that all the 1889 genes are included in 77 KEGG pathways.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Proteasome	RT		30	1.6	3.7E-14	1.1E-11
	KEGG_PATHWAY	Ribosome	RT		53	2.8	4.0E-11	5.8E-9
	KEGG_PATHWAY	Parkinson's disease	RT		53	2.8	2.6E-10	2.5E-8
	KEGG_PATHWAY	Huntington's disease	RT		63	3.3	1.8E-9	1.3E-7
	KEGG_PATHWAY	Biosynthesis of antibiotics	RT		67	3.5	3.0E-9	1.7E-7
	KEGG_PATHWAY	Alzheimer's disease	RT		56	3.0	8.9E-9	4.3E-7
	KEGG_PATHWAY	Spliceosome	RT		46	2.4	6.8E-8	2.8E-6
	KEGG_PATHWAY	Non-alcoholic fatty liver disease (NAFLD)	RT		49	2.6	2.2E-7	8.0E-6
	KEGG_PATHWAY	Protein processing in endoplasmic reticulum	RT		53	2.8	2.2E-7	7.2E-6
	KEGG_PATHWAY	Oxidative phosphorylation	RT		44	2.3	5.6E-7	1.6E-5
	KEGG_PATHWAY	Epstein-Barr virus infection	RT		56	3.0	9.1E-7	2.4E-5
	KEGG_PATHWAY	Focal adhesion	RT		57	3.0	6.3E-6	1.5E-4

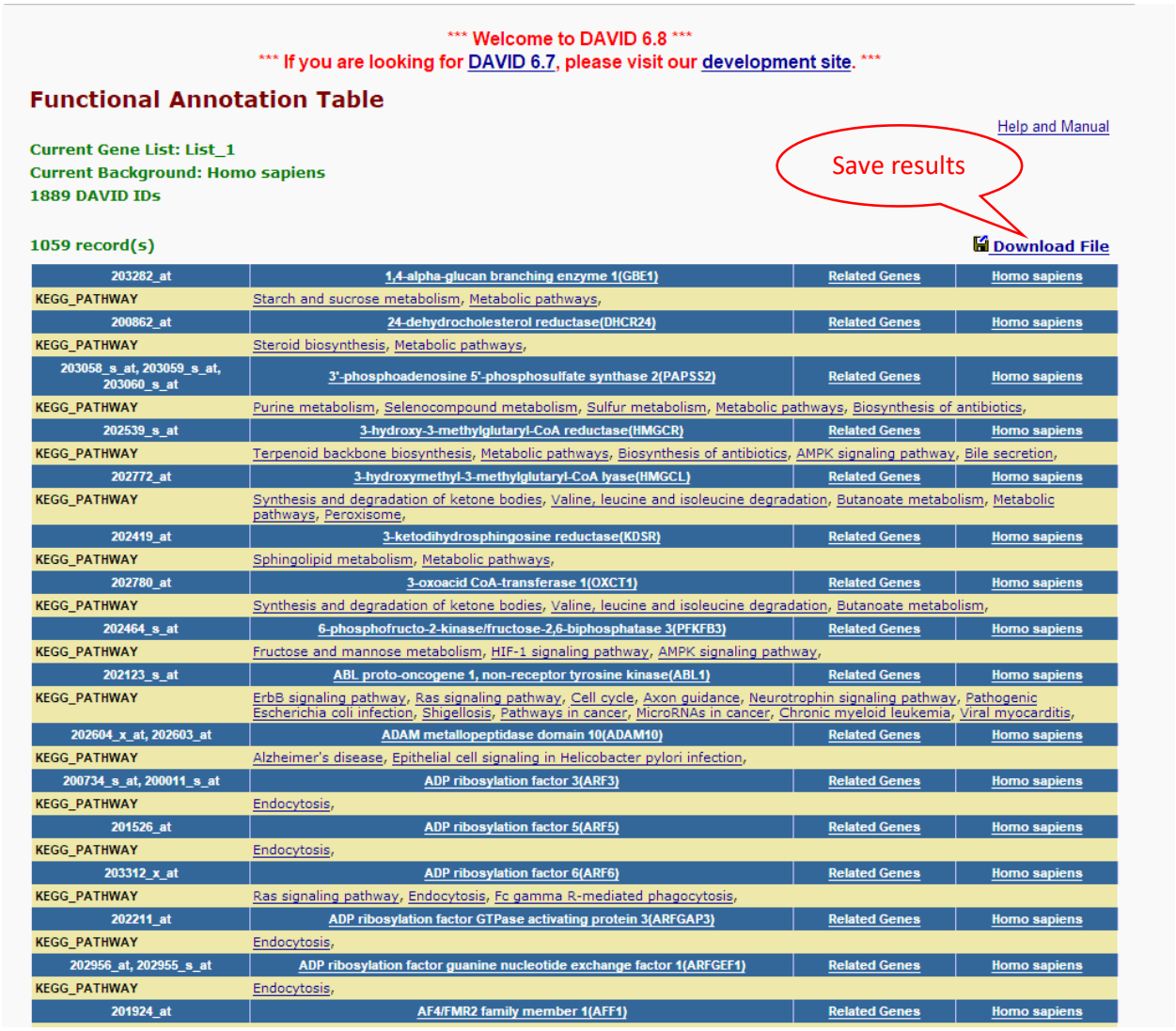
Count Threshold (Minimum Count): The threshold of minimum gene counts belonging to an annotation term. Default value is 2. In short, you do not trust the term only having one gene involved.

Pathways are ordered by ascending p-value but can be ordered by any other column by clicking on the header of the column.

Q4: What is the pathway with a higher gene count?

Now choose the “Functional Annotation Table”.

The “Functional Annotation Table” shows that 1059 genes are annotated with one or more annotations (here, KEGG pathways).



*** Welcome to DAVID 6.8 ***

*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Table

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
1889 DAVID IDs

Save results

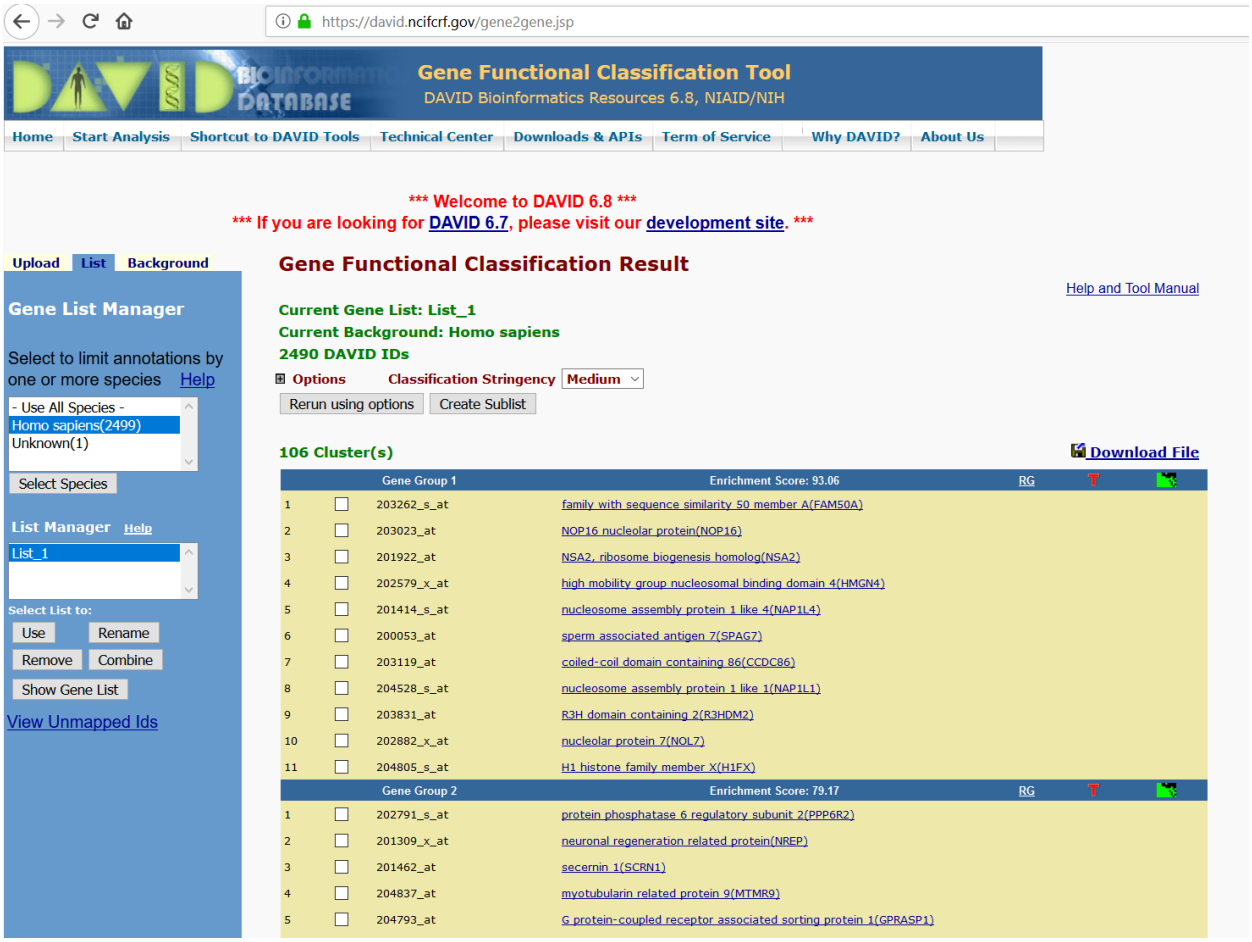
1059 record(s)

Download File

203282_at	1,4-alpha-glucan branching enzyme 1(GBE1)	Related Genes	Homo sapiens
KEGG_PATHWAY	Starch and sucrose metabolism, Metabolic pathways,		
200862_at	24-dehydrocholesterol reductase(DHCR24)	Related Genes	Homo sapiens
KEGG_PATHWAY	Steroid biosynthesis, Metabolic pathways,		
203058_s_at, 203059_s_at, 203060_s_at	3'-phosphoadenosine 5'-phosphosulfate synthase 2(PAPSS2)	Related Genes	Homo sapiens
KEGG_PATHWAY	Purine metabolism, Selenocompound metabolism, Sulfur metabolism, Metabolic pathways, Biosynthesis of antibiotics,		
202539_s_at	3-hydroxy-3-methylglutaryl-CoA reductase(HMGCR)	Related Genes	Homo sapiens
KEGG_PATHWAY	Terpenoid backbone biosynthesis, Metabolic pathways, Biosynthesis of antibiotics, AMPK signaling pathway, Bile secretion,		
202772_at	3-hydroxymethyl-3-methylglutaryl-CoA lyase(HMGCL)	Related Genes	Homo sapiens
KEGG_PATHWAY	Synthesis and degradation of ketone bodies, Valine, leucine and isoleucine degradation, Butanoate metabolism, Metabolic pathways, Peroxisome,		
202419_at	3-ketodihydrosphingosine reductase(KDSR)	Related Genes	Homo sapiens
KEGG_PATHWAY	Sphingolipid metabolism, Metabolic pathways,		
202780_at	3-oxoacid CoA-transferase 1(OXCT1)	Related Genes	Homo sapiens
KEGG_PATHWAY	Synthesis and degradation of ketone bodies, Valine, leucine and isoleucine degradation, Butanoate metabolism,		
202464_s_at	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3(PFKFB3)	Related Genes	Homo sapiens
KEGG_PATHWAY	Fructose and mannose metabolism, HIF-1 signaling pathway, AMPK signaling pathway,		
202123_s_at	ABL proto-oncogene 1, non-receptor tyrosine kinase(ABL1)	Related Genes	Homo sapiens
KEGG_PATHWAY	ErbB signaling pathway, Ras signaling pathway, Cell cycle, Axon guidance, Neurotrophin signaling pathway, Pathogenic Escherichia coli infection, Shigellosis, Pathways in cancer, MicroRNAs in cancer, Chronic myeloid leukemia, Viral myocarditis,		
202604_x_at, 202603_at	ADAM metallopeptidase domain 10(ADAM10)	Related Genes	Homo sapiens
KEGG_PATHWAY	Alzheimer's disease, Epithelial cell signaling in Helicobacter pylori infection,		
200734_s_at, 200011_s_at	ADP ribosylation factor 3(ARF3)	Related Genes	Homo sapiens
KEGG_PATHWAY	Endocytosis,		
201526_at	ADP ribosylation factor 5(ARF5)	Related Genes	Homo sapiens
KEGG_PATHWAY	Endocytosis,		
203312_x_at	ADP ribosylation factor 6(ARF6)	Related Genes	Homo sapiens
KEGG_PATHWAY	Ras signaling pathway, Endocytosis, Fc gamma R-mediated phagocytosis,		
202211_at	ADP ribosylation factor GTPase activating protein 3(ARFGAP3)	Related Genes	Homo sapiens
KEGG_PATHWAY	Endocytosis,		
202956_at, 202955_s_at	ADP ribosylation factor guanine nucleotide exchange factor 1(ARFGEF1)	Related Genes	Homo sapiens
KEGG_PATHWAY	Endocytosis,		
201924_at	AF4/FMR2 family member 1(AFF1)	Related Genes	Homo sapiens

Tool 4. Gene Functional Classification

Click the “Gene Functional Classification tool” under the list of “Shortcut to DAVID tools”. The results show 106 clusters of annotations. This tool is used to cluster the functionally related genes as a group and give a score to this cluster.



Q6: What differences can you see between gene groups?

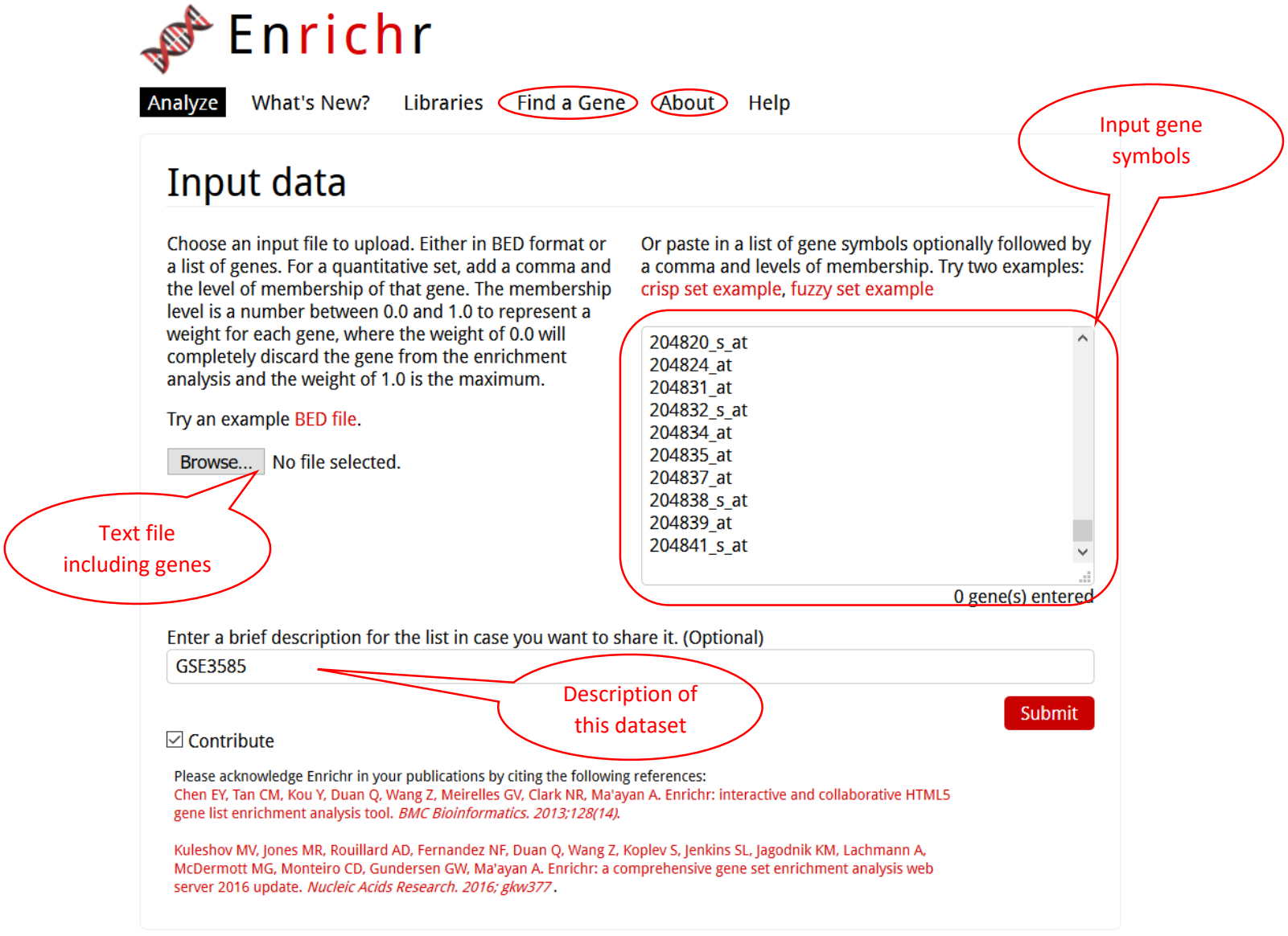
Exercise 2. Using Enrichr

Enrichr (http://amp.pharm.mssm.edu/Enrichr/) accepts either BED format or a list of genes with gene symbols.

1. Upload your gene list

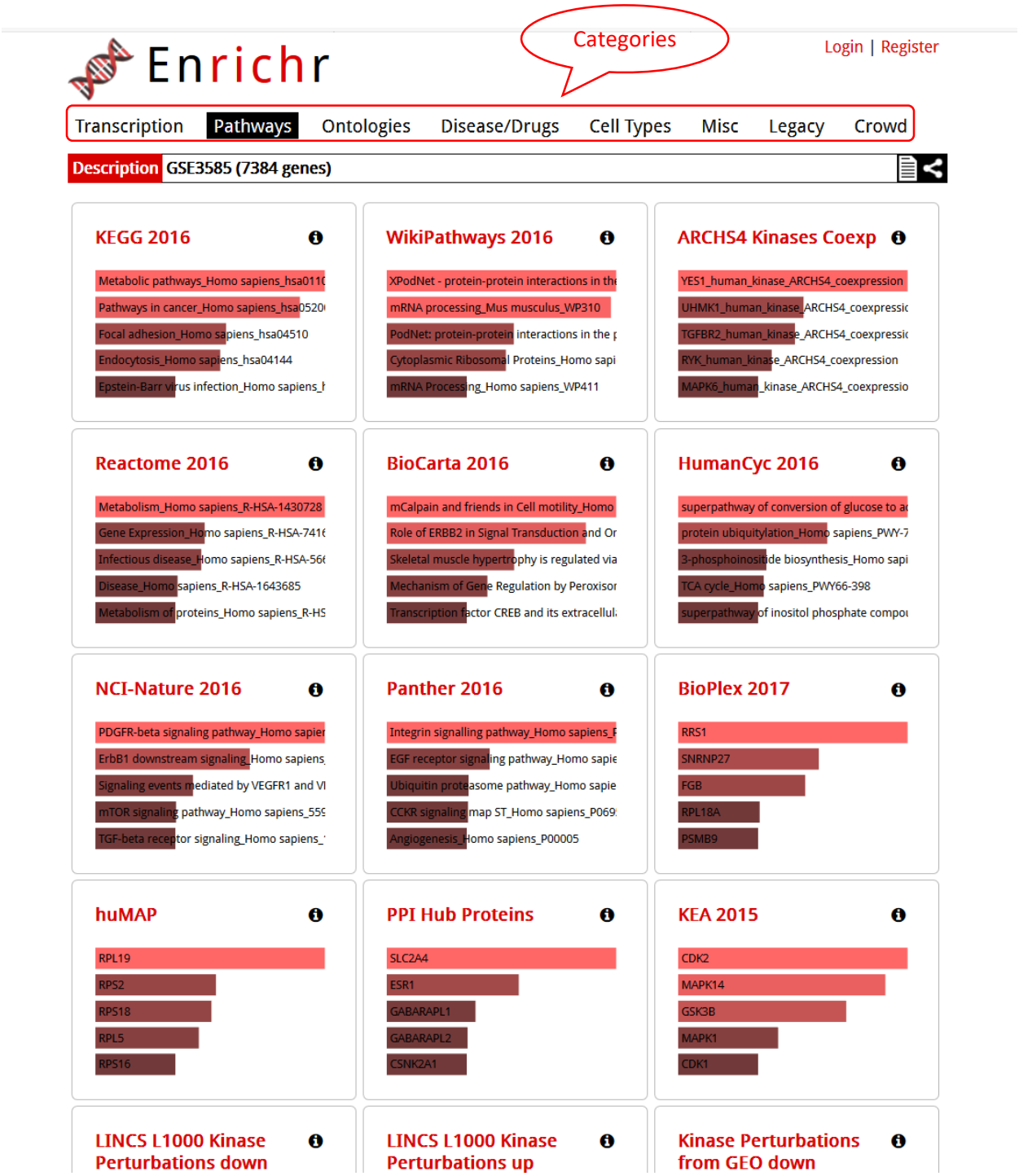
Enrichr uses a list of gene symbols as input data. You can upload the list by either selecting the text file that contains the list or just simply pasting the list into the text box. It is better to enter a description for the gene list so that multiple lists can be differentiated from each other.

We will use the same genes from the previous exercise.



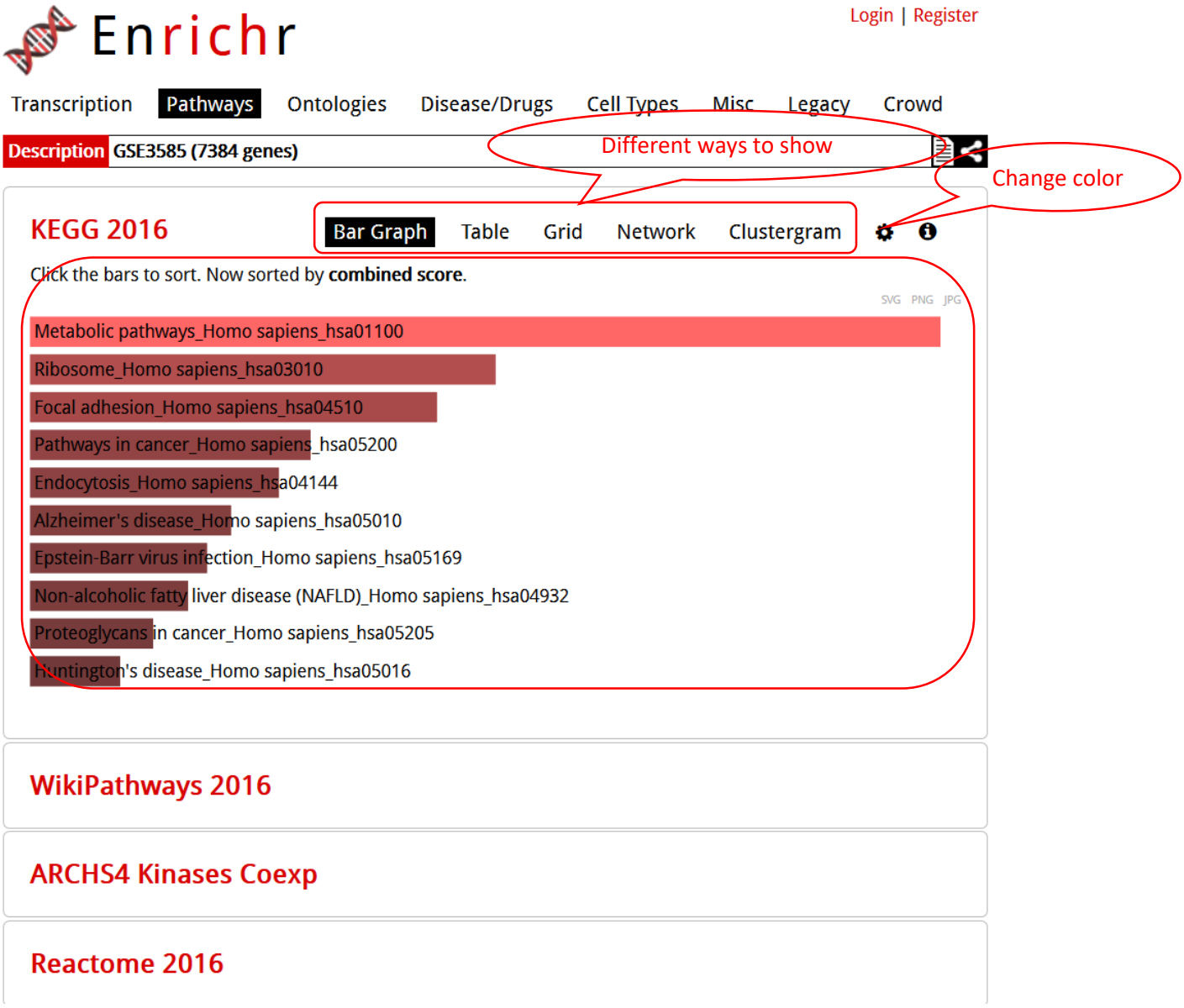
2. Results page

On the results page, the analysis is divided into different categories of enrichment (Transcription, Pathways, ontologies and so on). The first category is shown. Within each category, the enrichment analyses of various gene-set libraries are listed. We open the pathway analysis as an example, presenting a multitude of visualizations. If you want to change the category, just tap the other category name.



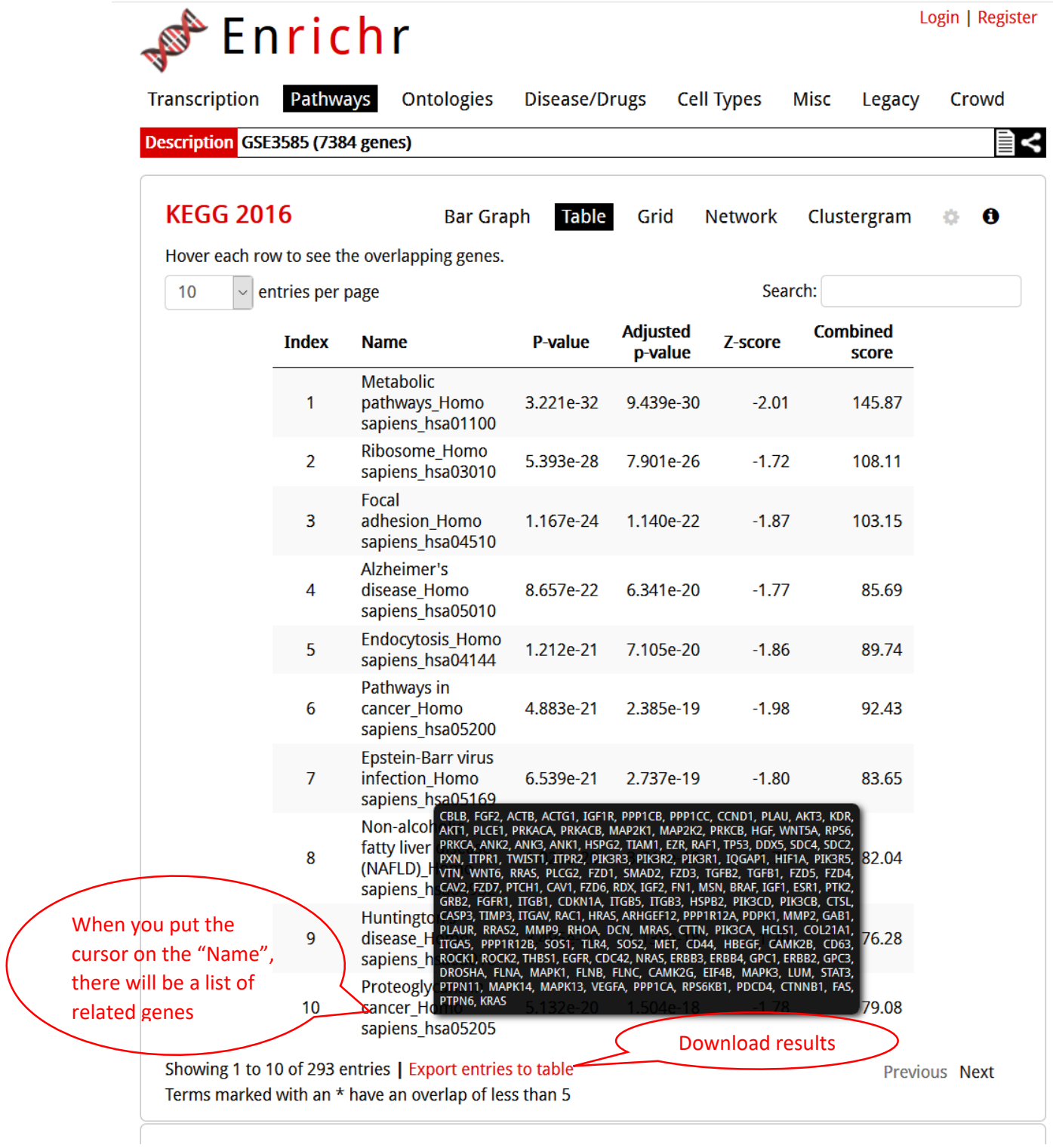
Click on “KEGG 2016” to view the detailed results. They include: “Bar Graph”, “Table”, “Grid”, “Network”, and “Clustergram”. When you click on the bars, you get different ranks by other score methods. Notice that it takes longer time to open “Clustergram”.

Bar Graph:



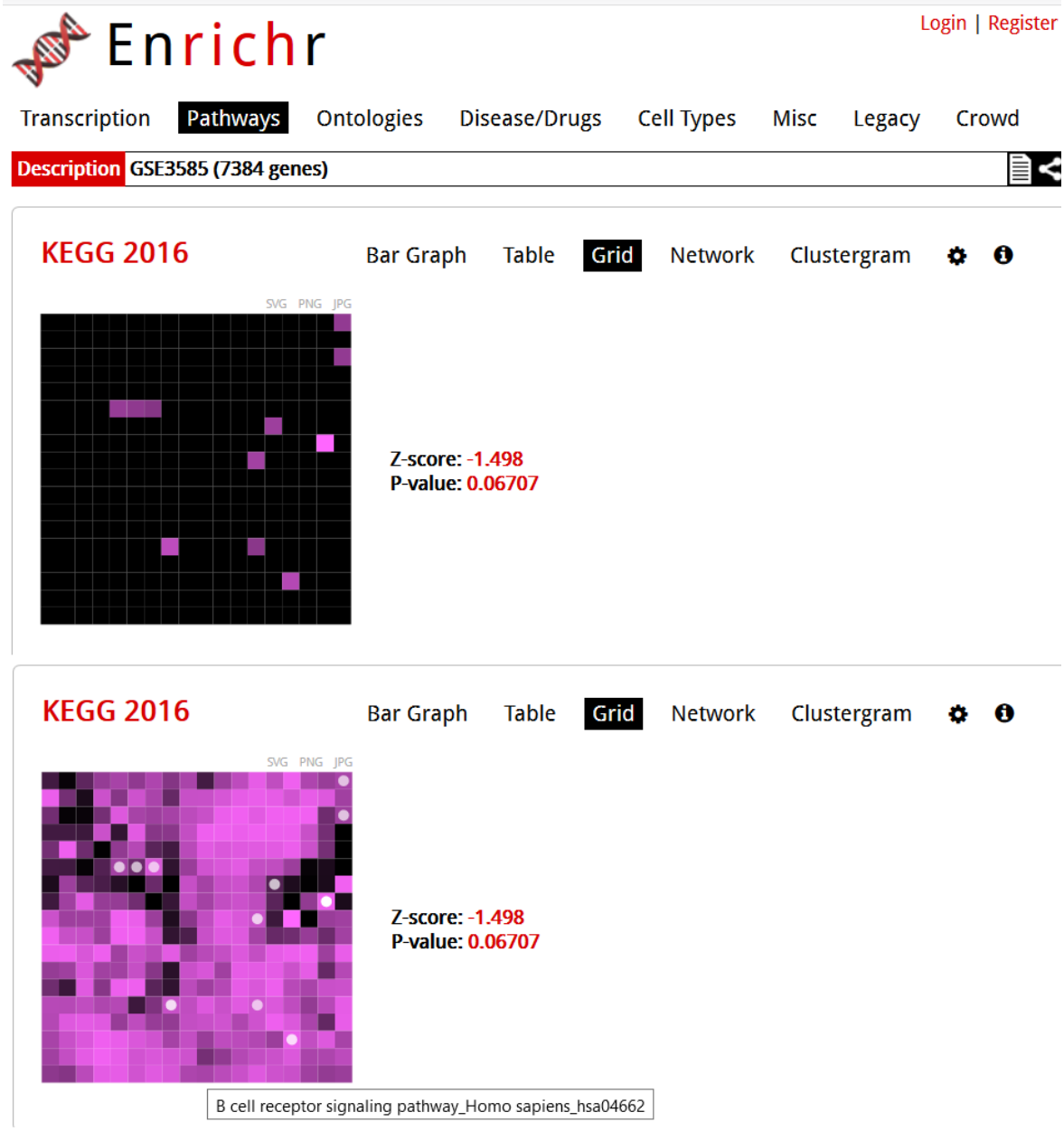
The length of the bar represents the significance of that specific gene-set or term. In addition, the brighter the color, the more significant that term is.

Table:



By clicking on the column header, you can sort the table by the term, p-value, z-score, or combined score. You can also download the table information by clicking on the "Export entries to table" button.

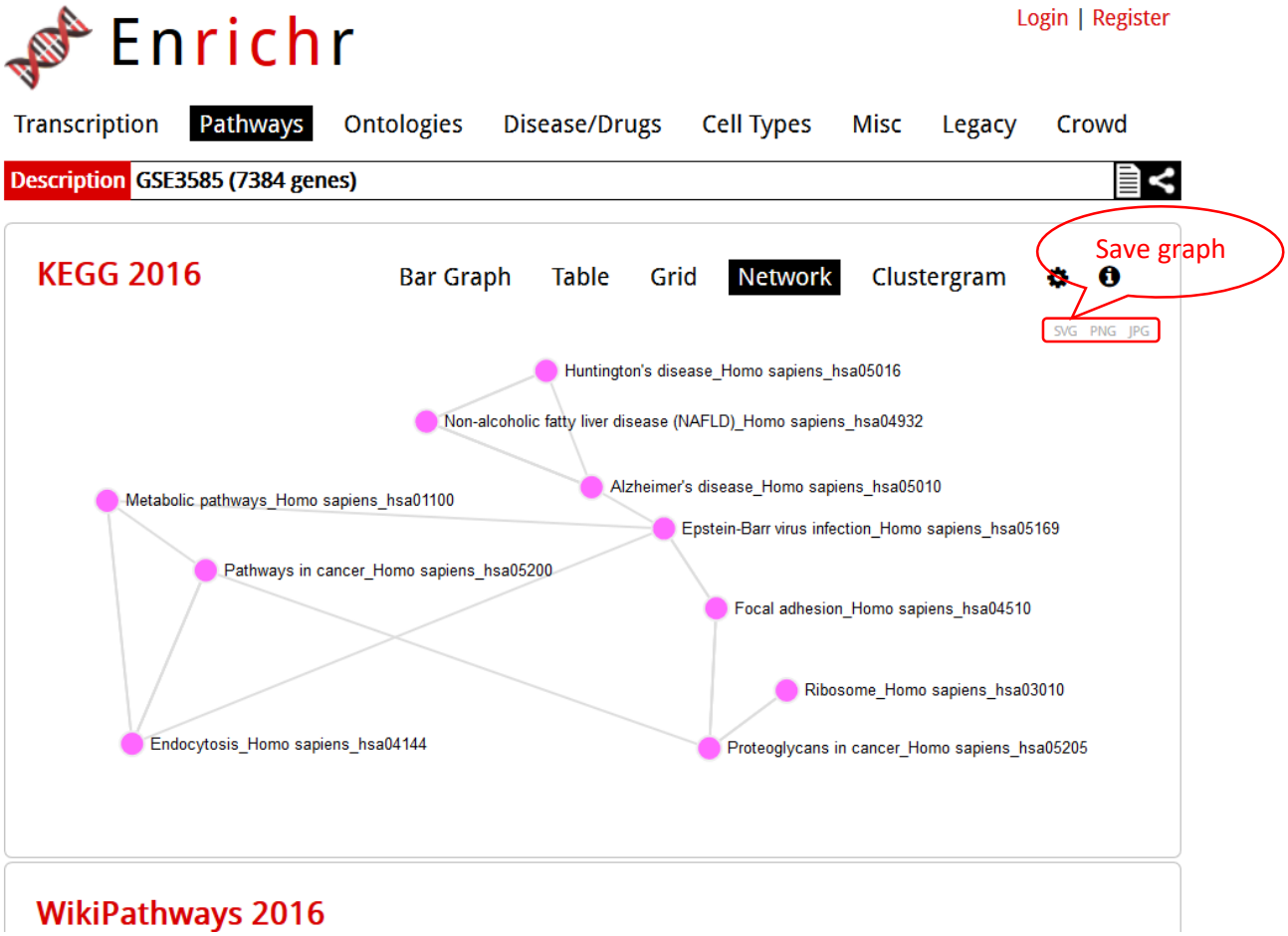
Grid:



Each grid square represents a term and is arranged based on its gene-set similarity with other terms. It shows only the top 10 terms sorted by combined score. The brighter the square, the more significant that term is. Clicking on the grid allows you to another view that colors the grid based on its correlation score with neighbors with white dots representing the significant terms. The z-score and p-value is a measure of how clustered the top 10 terms are on the grid.

Q7: Where can we find the significant gene terms?

Network:

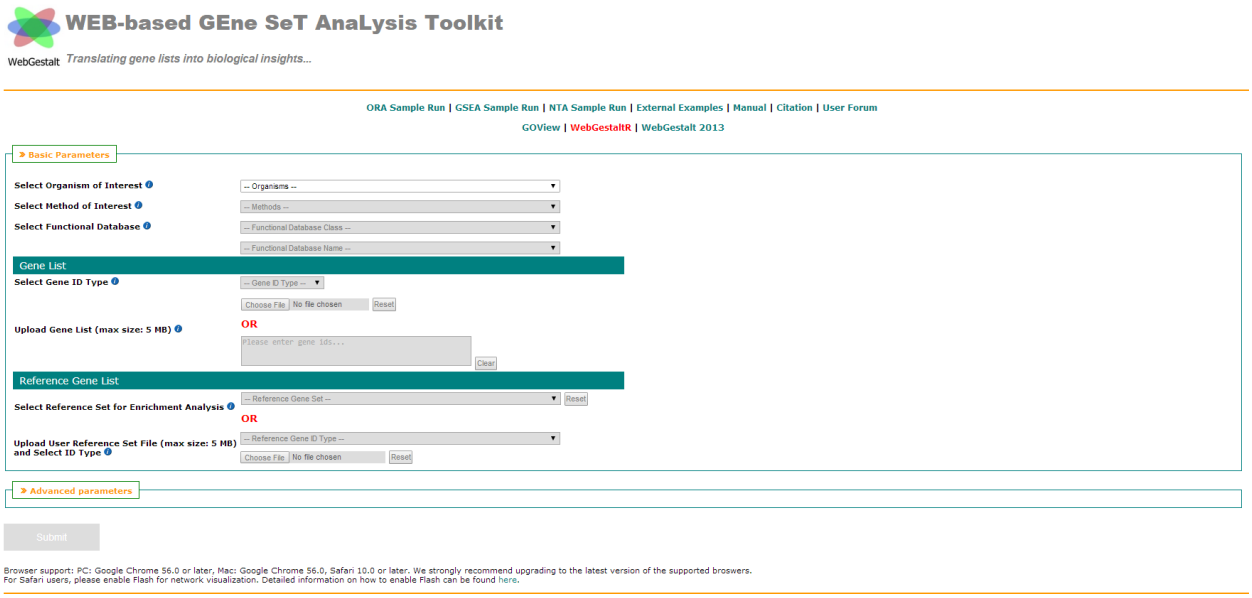


Each node represents a term and a link between two nodes means that the two terms have some gene content similarity.

Q8: How to find the pathway with your genes of interest?

Exercise 3. Using WebGestalt

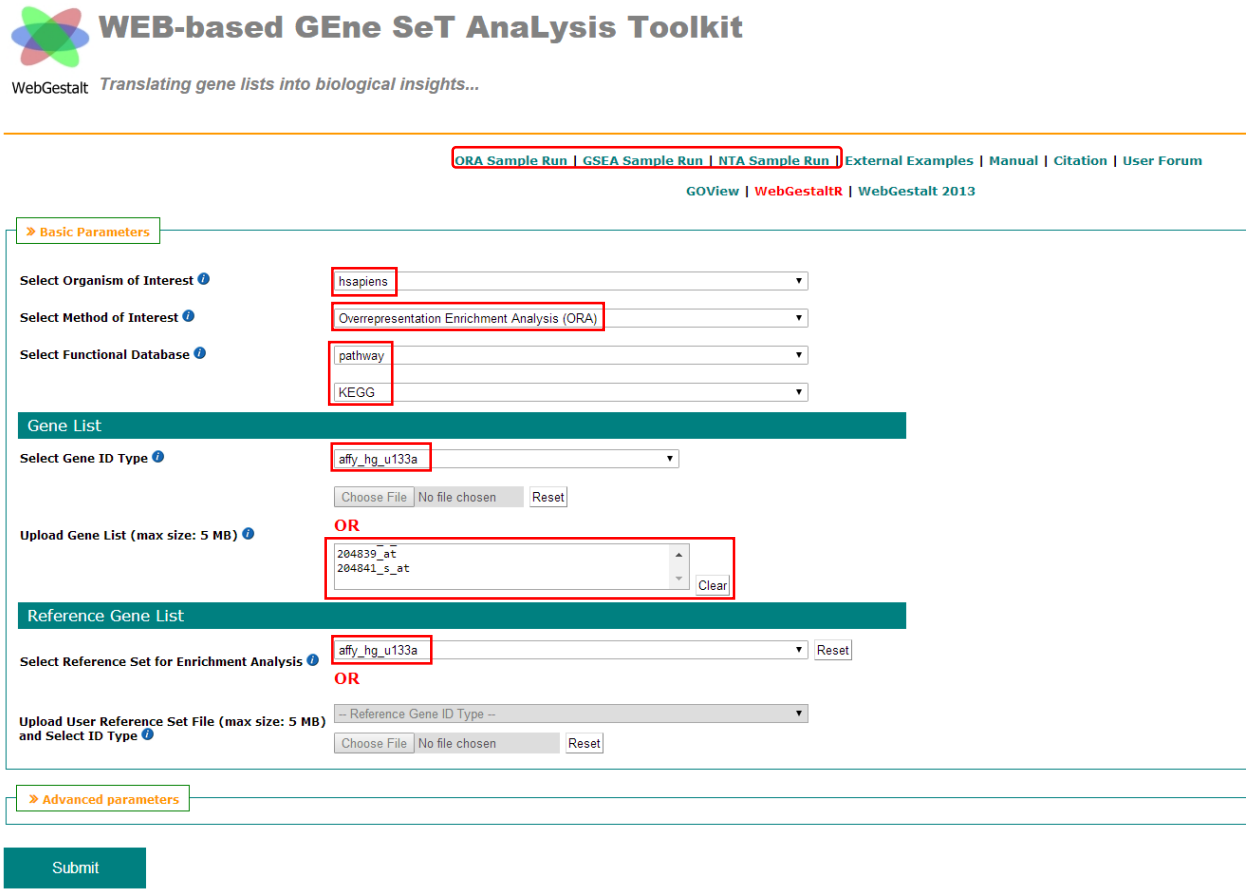
WebGestalt(http://www.webgestalt.org/option.php) is a functional enrichment analysis web tool that supports three well-established and complementary methods for enrichment analysis: Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA), and Network Topology-based Analysis (NTA).



1. Setting parameters

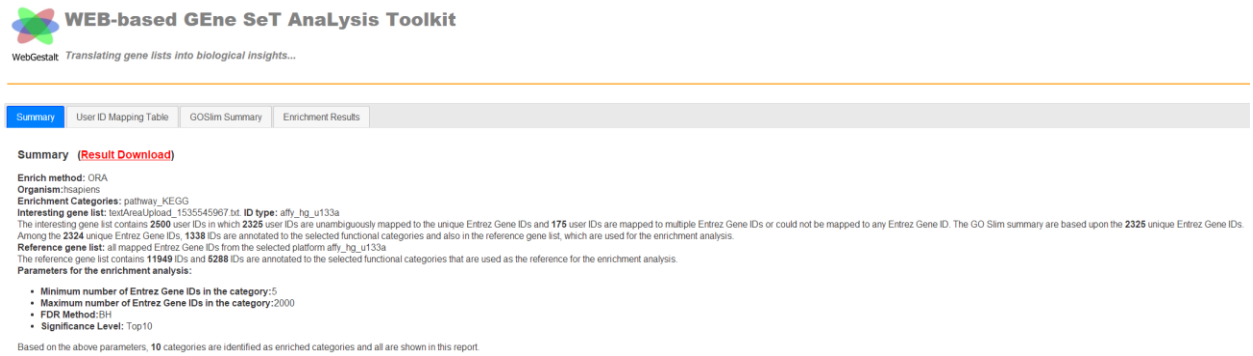
Set the parameters and upload the gene list, as in the following picture, and click the “Submit” button.

We are using ORA. If we change the method to “GSEA”, then we need a ranked gene list.



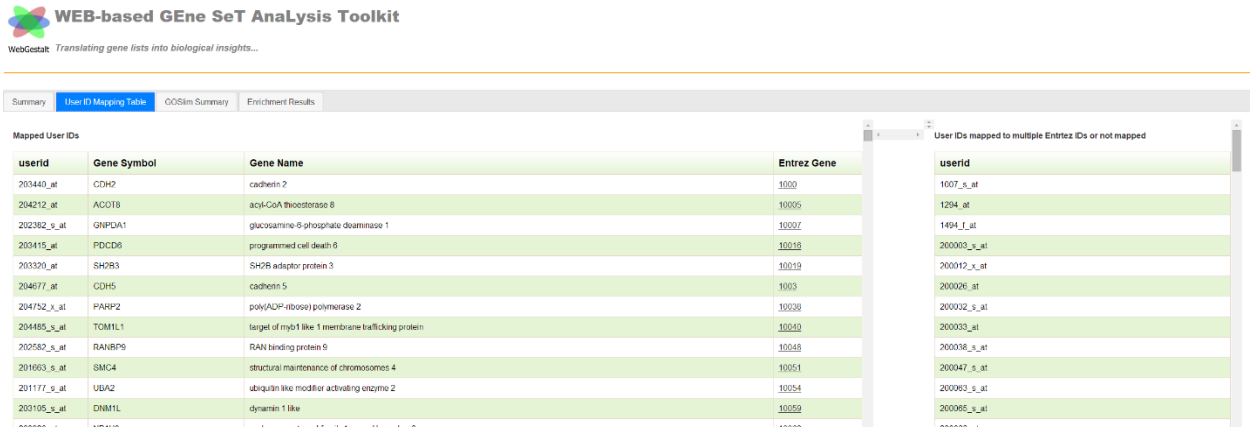
2. Results

After we submit the task, the summary comes into being at first. It contains enrichment method, organism, enrichment category, gene list with ID type, reference gene list, and parameters for enrichment analysis. We also get: “User ID Mapping Table”, “GOSlim Summary” and “Enrichment Results”.



2.1 User ID Mapping Table

In the table, the left contains the mapped ID, gene symbol, gene names, and Entrez gene ID. The right contains the “User IDs mapped to multiple Entrez IDs or not mapped”.

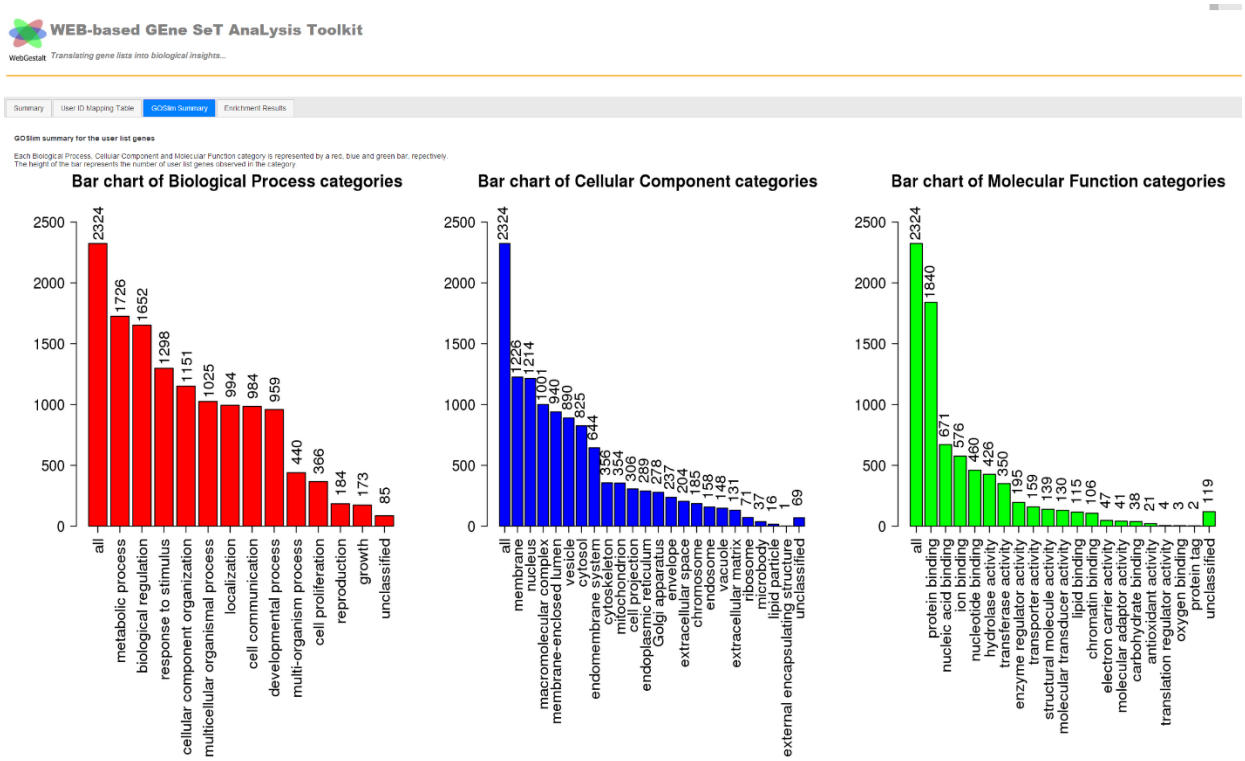


userid	Gene Symbol	Gene Name	Entrez Gene
203440_at	CDH2	cadherin 2	1000
204212_at	ACOT8	acyl-CoA thioesterase 8	10005
202382_s_at	GNPDA1	glucosamine-6-phosphate deaminase 1	10007
203415_at	PDCD6	programmed cell death 6	10016
203320_at	SH2B3	SH2B adaptor protein 3	10019
204677_at	CDH5	cadherin 5	1003
204752_x_at	PARP2	poly(ADP-ribose) polymerase 2	10038
204485_s_at	TOM1L1	target of myb1 like 1 membrane trafficking protein	10040
202582_s_at	RANBP9	RAN binding protein 9	10048
201663_s_at	SMC4	structural maintenance of chromosomes 4	10051
201177_s_at	UBA2	ubiquitin like modifier activating enzyme 2	10054
203105_s_at	DNM1L	dynamin 1 like	10059

userid
1007_s_at
1294_at
1494_f_at
200003_s_at
200012_x_at
200026_at
200032_s_at
200033_at
200038_s_at
200047_s_at
200063_s_at
200065_s_at

2.2 GOSlim Summary

The three charts represent Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) categories, in red, blue and green bars, respectively. The height of the bar represents the number of user list genes observed in the category.



Q9: Based on the pictures, how would you describe the genes in your dataset in your own words?

2.3 Enrichment Results

The left table is the summary table, and the right one is the detailed information table.

In the right table:

“C”: the number of reference genes in the category

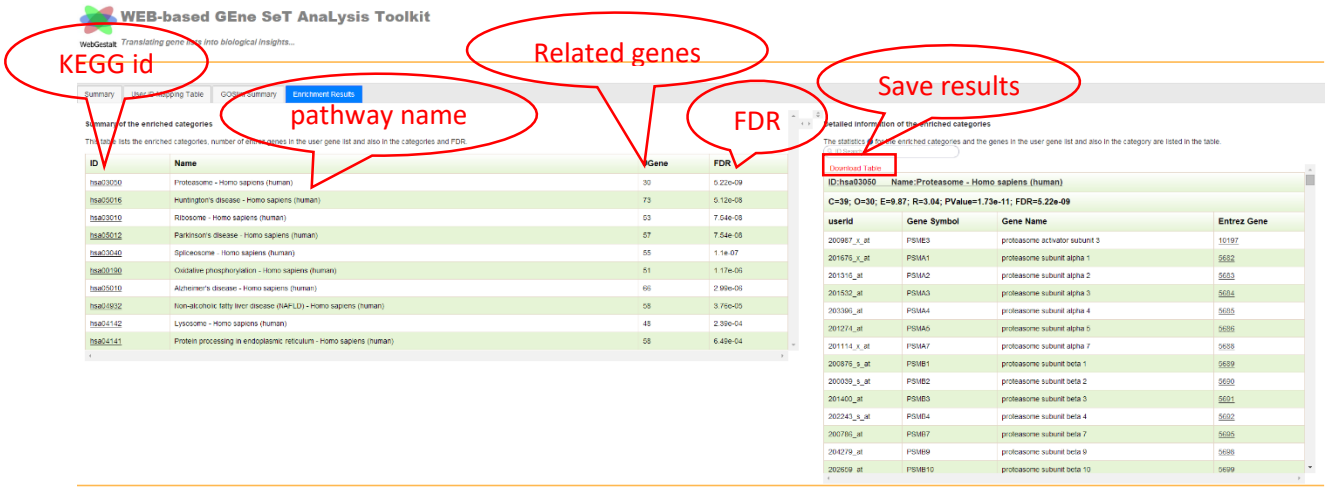
”O”: the number of genes in the uploaded gene list and also in the category

”E”: the expected number in the category

”R”: ratio of enrichment

”P-Value”: p-value from hypergeometric test

”FDR”: FDR from BH



Q10: What are the top 10 significant pathways?

Exercise 4. Compare the three websites in terms of KEGG pathways enrichment

What are the most significant pathways in each of the GSA websites?

How well do they agree?

Which website uses more databases? Which website uses more GSA methods?

Which website gives you better summary tables and figures?

What was your favorite GSA website?


MORA LAB
FUNCTIONAL BIOINFORMATICS