Last updated: June, 2015

Tutorial: Using DAVID bioinformatics resources

http://david.abcc.ncifcrf.gov

DAVID bioinformatics resources is an integrated biological knowledgebase and analytical tool aimed at extracting biological meaning from large gene/protein lists generated from a variety of high-throughput genomic experiments. This tutorial will be conducted starting with a set of 387 probe IDs from an experiment done with peripheral blood mononuclear cells treated with an HIV envelope protein (gp120) and genome-wide expression changes were observed using Affymetrix U95A microarray chips.

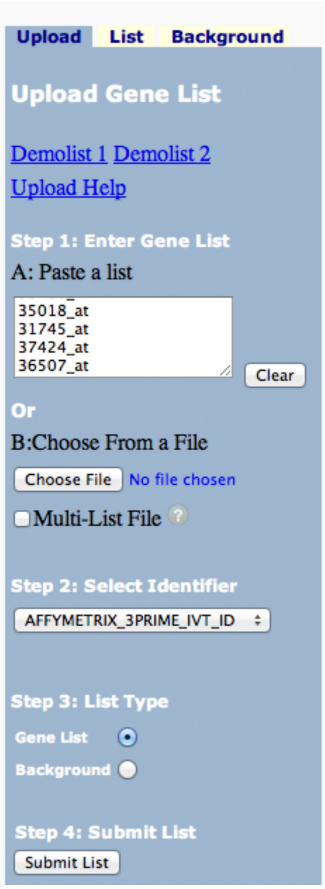
Protocol 1: Uploading a list of gene or probe identifiers

1) Starting DAVID

Open the home page for the DAVID resources site and then click on the link at the top that says **Start Analysis**

2) Uploading gene lists into DAVID.

In a second tab, open the Week 2 home page and click on the link for the GenesforGO from the course website. This is some from a colon cancer study. Either open them and copy and paste or use the Choose File button to upload the list of genes into the text box below **A: Paste a List** as shown in **Figure 1**.



Step 1: Enter Gene List

A: Paste a list

```
35018_at
31745_at
37424_at
36507_at
```

Or

B:Choose From a File

No file chosen

□ Multi-List File

Step 2: Select Identifier

AFFYMETRIX_3PRIME_IVT_ID

Step 3: List Type

Gene List ●

Background ○

Step 4: Submit List

Figure 1: Upload gene list menu

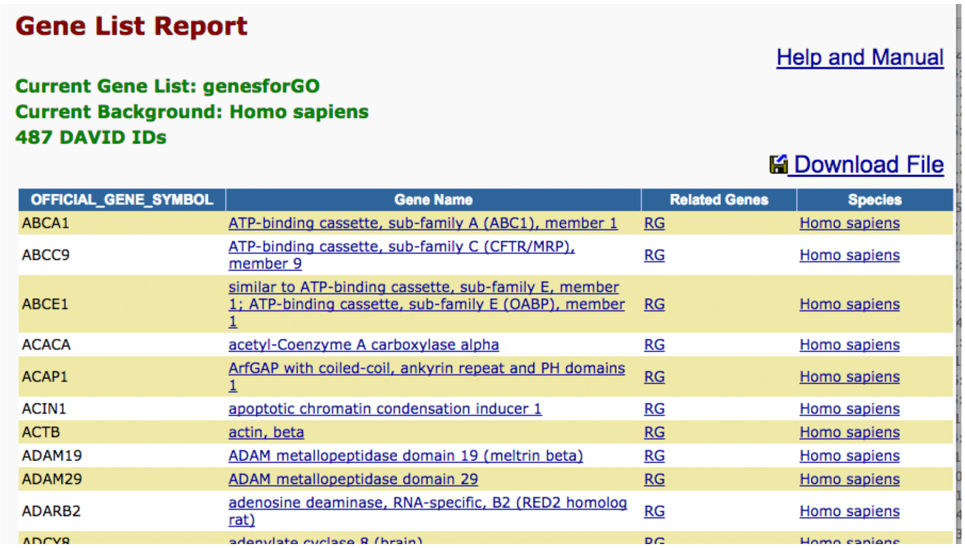
Select OFFICIAL_GENE_SYMBOL under the drop down menu options for **Select Identifier** and select Gene List under the **List Type** options. Then click the **Submit List** button. The Analysis Wizard should say that multiple species have been detected and and the menu on the left-hand side will switch to the **Gene List Manager** tab. Highlight Homo Sapiens (487) and click Select Species. Under the List Manager, there should be a line that reads genesforGO. You can rename a list by clicking the **Rename** button and call it something else.

You can also upload a file using the **B: Choose From a File** if you have your identifiers in a text file.

3) Gene name Batch Viewer.

Click on the **Shortcut to DAVID tools** menu option at the top of the page and select the menu option Gene Name Batch Viewer.

This should change the page to a Gene List Report as shown in **Figure 2** below.



Gene List Report

Current Gene List: genesforGO
Current Background: Homo sapiens
487 DAVID IDs

OFFICIAL_GENE_SYMBOL	Gene Name	Related Genes	Species
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	RG	Homo sapiens
ABCC9	ATP-binding cassette, sub-family C (CFTR/MRP), member 9	RG	Homo sapiens
ABCE1	similar to ATP-binding cassette, sub-family E, member 1; ATP-binding cassette, sub-family E (OABP), member 1	RG	Homo sapiens
ACACA	acetyl-Coenzyme A carboxylase alpha	RG	Homo sapiens
ACAP1	ArfGAP with coiled-coil, ankyrin repeat and PH domains 1	RG	Homo sapiens
ACIN1	apoptotic chromatin condensation inducer 1	RG	Homo sapiens
ACTB	actin, beta	RG	Homo sapiens
ADAM19	ADAM metallopeptidase domain 19 (meltrin beta)	RG	Homo sapiens
ADAM29	ADAM metallopeptidase domain 29	RG	Homo sapiens
ADARB2	adenosine deaminase, RNA-specific, B2 (RED2 homolog rat)	RG	Homo sapiens
ADCY8	adenylate cyclase 8 (brain)	RG	Homo sapiens

Figure 2: Gene List Report from DAVID.

The gene names are hot links to pages with information about the genes. The **RG** links will display, in a new tab, a list of other genes WITHIN the submitted list that are related to the gene you clicked on. These relationships are based on a statistical measure of functional similarity. **Note:** Not all genes will have functionally related genes within the submitted list. Within the RG tab, you can expand the scope of related genes search under the Options menu. Change it from User to *Mus musculus* and then click the **Rerun Using Options** button. Now the search will find any genes in the mouse genome that are functionally related to the one in the list.

3) Gene functional classification.

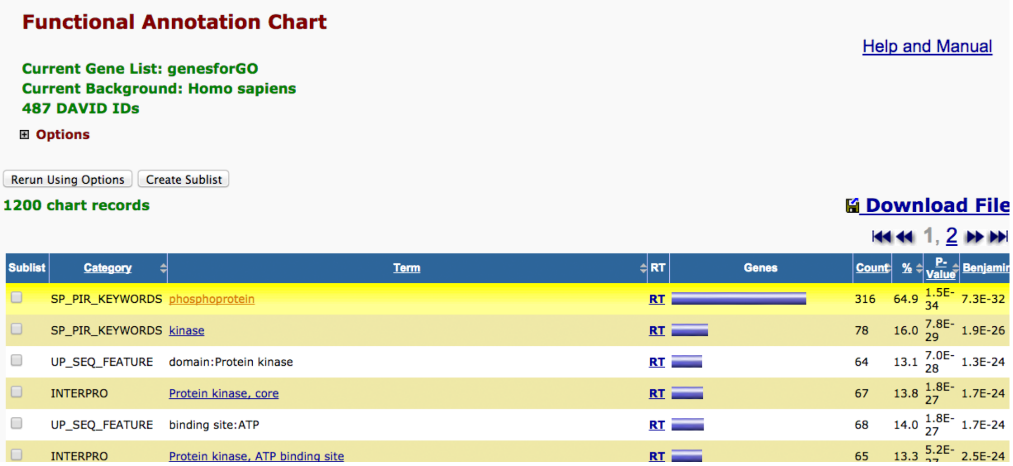
Return to the tab with your gene list. Under the **Shortcut to DAVID tools** menu option select the option Gene Functional Classification

This should open a window displaying 18 clusters of gene groups. Each cluster is an attempt by DAVID to classify the submitted list of genes into functionally related gene groups. These groups are identified by annotation term co-occurrence. The blue bars at the top of each cluster contain a number of links that allow you to learn more about the clusters were identified. Gene Group 1 contained 76 genes, but just looking at the gene names may not tell you why all of these genes are grouped together. Although in this case, My gene group 1 had kinesisns, cyclins and kinases. Click on the red **T** located in the blue bar at the top of the cluster. This will open a new tab called **Term Report**. This report lists all of the terms annotated to these 76 genes and the number of genes annotated with a given term. From this list of terms, it seems most of these genes are kinases or related to kinases. The category column lists the source provenance for the term. One advantage of DAVID to some other gene ontology analysis programs is that it uses annotation information from a wider variety of sources. Clicking on the blue bar in the column titled **Genes in Group** will open a tab with a list of the genes annotate with a given term.

Return to the tab Gene Functional Classification results tab. You may want to close some of the tabs that DAVID opened, as the browser can get fairly cluttered.

4) Functional Annotation Chart

Under the **Shortcut to DAVID tools** menu option select the option Functional annotation. This will open a window titled **Annotation Summary Results** and has a lot of different options. This is the main window by which you can do an enrichment analysis with DAVID. Use the default options and click the **Functional Annotation Chart** button located at the bottom of the page. This will open a new window (rather than a new tab) as shown in **Figure 3** below.



Functional Annotation Chart

Help and Manual

Current Gene List: genesforGO
Current Background: Homo sapiens
487 DAVID IDs

⊞ Options

Rerun Using Options | Create Sublist

1200 chart records

Download File

1, 2

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamir
□	SP_PIR_KEYWORDS	phosphoprotein	RT		316	64.9	1.5E-34	7.3E-32
□	SP_PIR_KEYWORDS	kinase	RT		78	16.0	7.8E-29	1.9E-26
□	UP_SEQ_FEATURE	domain:Protein kinase	RT		64	13.1	7.0E-28	1.3E-24
□	INTERPRO	Protein kinase, core	RT		67	13.8	1.8E-27	1.7E-24
□	UP_SEQ_FEATURE	binding site:ATP	RT		68	14.0	1.8E-27	1.7E-24
□	INTERPRO	Protein kinase, ATP binding site	RT		65	13.3	5.2E-	2.5E-24

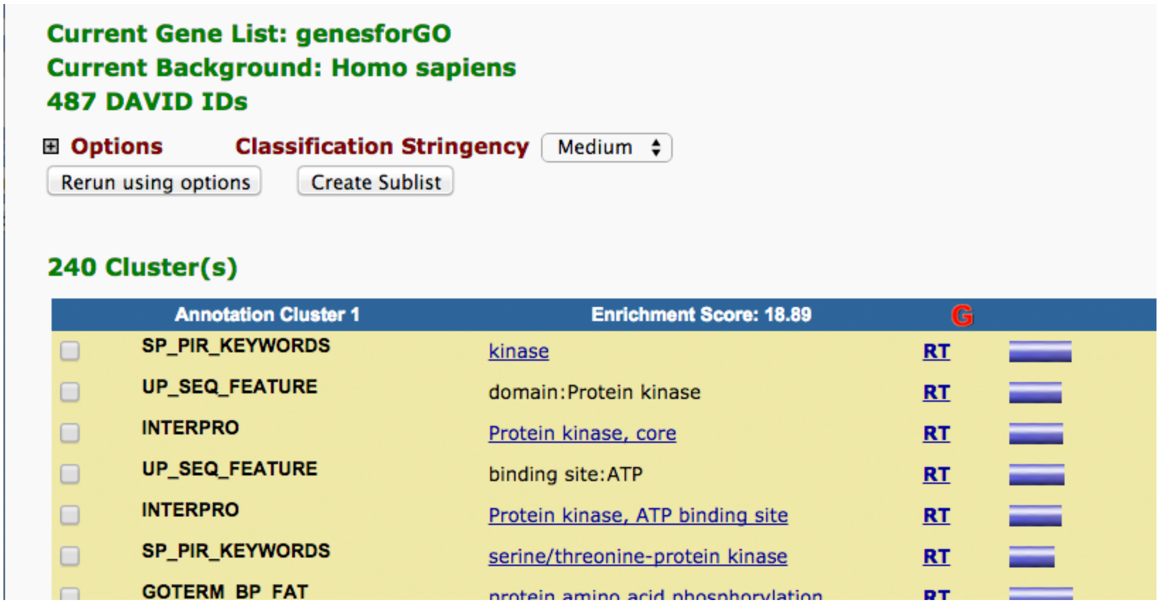
Figure 3: Functional annotation chart

In the functional annotation chart, each annotation term associated with the list of genes is tested for enrichment relative to the background set of genes. In this case there are 1200 chart records. The chart lists the terms that are statistically enriched, in order of increasing Benajmini adjusted p-value, located in right-most column. You can adjust which columns of data are shown under the **Options** menu located above the list. The entire file can be downloaded in a format compatible with Excel. The Count and % columns represent the number of genes in the list annotated with the term and percent of genes in the list annotated with that term. So for the SP_PIR_KEYWORD term "phosphoprotein" there are 316 genes in the list of 487, or 65% of the total list. If you want to see what genes are annotate with that term, click on the purple bar under the Genes column and it will open a tab with a list of all the genes annotated with that term in the list.

This type of term enrichment is very similar to the results you would get from other GO analysis programs, with the added benefit of additional annotation sources. One thing you have probably already noticed is that the different terms have a fair amount of overlap with regard to their function or role in the cell. For example two INTERPRO terms are "protein kinase, core" with 67 genes and "Protein kinase, ATP binding site" with 65 genes. It's likely that most of the 6 genes in these two terms overlap. This is a frustrating aspect of conducting a gene ontology analysis. That is knowing when and how to collapse terms to remove some of the redundancy. DAVID has attempted to address this issue with another tool called Functional Annotation Clustering.

5) Functional Annotation Clustering

Return to the window and tab Functional Annotation Tools and click on the **Functional Annotation Clustering** button located at the bottom of the page. This will open another new window as shown in **Figure 4.** I got 240 clusters when I did this analysis.



Current Gene List: genesforGO
Current Background: Homo sapiens
487 DAVID IDs

⊞ Options Classification Stringency: Medium

240 Cluster(s)

	Annotation Cluster 1	Enrichment Score: 18.89	G	
□	SP_PIR_KEYWORDS	kinase	RT	
□	UP_SEQ_FEATURE	domain:Protein kinase	RT	
□	INTERPRO	Protein kinase, core	RT	
□	UP_SEQ_FEATURE	binding site:ATP	RT	
□	INTERPRO	Protein kinase, ATP binding site	RT	
□	SP_PIR_KEYWORDS	serine/threonine-protein kinase	RT	
□	GOTERM_BP_FAT	protein amino acid phosphorylation	RT	

Figure 4: Functional annotation clustering window

Now instead of a long list of single terms, you see clusters of terms that are related to each other. The clusters are shown in order of decreasing enrichment scores. If you click on the red G button located on the blue bar at the top of each cluster, you should get a list of the genes in that cluster that are annotated with one or more of the terms. **NOTE:** the number of genes is not going to be a sum of the Count for each term because many of the genes will be annotated with more than one term listed in the cluster. You can alter the options and stringency of the cluster generation and rerun the analysis. Increasing the stringency will usually decrease both the number of clusters and the number of genes in the clusters. In this case, when I changed the stringency from medium to high, it returned 191 clusters but with many few terms per cluster. The first cluster went from 170 to 111 genes.

This is where the analysis becomes much more involved. There is NOT going to be one answer or even one approach to this type of analysis. This type of analysis provides a guide to new ways of thinking about the data, but is not going to pinpoint the 2-5 genes that will form the core of a paper. That's up to you and your ability to synthesize information from this as well as other sources. That is, here is where you need to really think about your data and what it means within the context of your experimental model and in the larger context of other published data.

6) Creating a sublist

Suppose you find a functional cluster that is of particular interest. You may want to follow up on the genes in that cluster in more depth. To do that, check the boxes located to the left of each term that you want to include in the sublist. Then click the **Create Sublist** button. It will give you an option to give it a more descriptive name, then click **OK**. After that, the new list should show up in the List Manager. You can select it and click the Use button to do a more in-depth analysis on the new list.

7) GeneID conversion

In case you didn't get enough of this during the BioMart tutorial, here's another opportunity. Select the original GenesforGO under List Manager. Under the **Shortcut to DAVID tools**, choose the Gene ID Conversion tool option. For option 1, select the conversion output to

ENTREZ_GENE_ID then click the **Submit to Conversion Tool** button. Out of 487, all 486 were successfully converted. The output should give you a ice list of the Gene symbols (in the From column) and Entrez GeneID numbers (To column) and the gene name. The option exists to download this file in a format compatible with Excel.

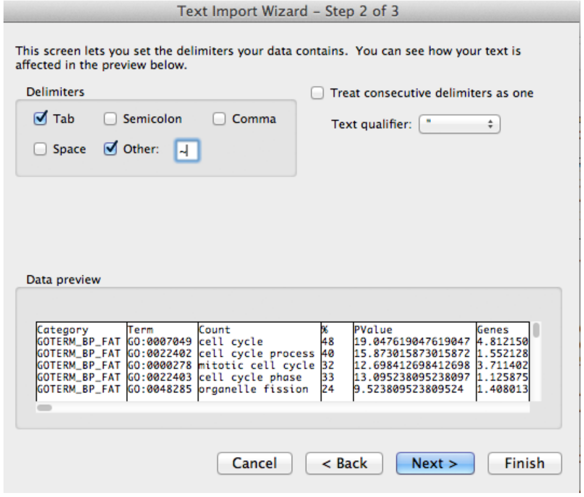
8) General annotation information

You can also use the Functional Annotation tools to obtain more general information about each of your genes. Return to the Functional Annotation Tool tab. Expand the menu under General Annotations. Click on the blue bar to the right of ENTREZ_GENE_SUMMARY. You should see a new window pop up with data for 356 of the genes. There is a summary paragraph for each probe ID that was matched in the Entrez Gene database that also contained summary data. This is probably more useful for a smaller set of genes, but is a quick way to assemble some basic information about the genes in whatever list you are working with.

DAVID also has links to Pathways and Protein-Protein interaction databases but we'll work on those later this week and next.

9) Saving Charts from DAVID & importing into Excel.

After you've done a functional annotation cluster or generated a chart for gene ontology, the data returned can be downloaded in a tab-delimited text file. It can be easily imported into excel. Click on the **Download File** link at the top right of the page. It will open a new page or tab with the data in text format. Click on the page and then under the browser File menu, select Save Page As… and save it as a text file to your hard drive. Open Excel, then open the text file from within Excel. It should import nicely, with the data distributing itself into columns. There is one caveat with the gene ontology data. It is presented in DAVID as "GOtermID~GOname" and when you import it into Excel, that comes in as a single entity. To split the GOtermID and GOname into two adjacent cells, when you get to Step 2 of the Text Import Wizard, add the tilde (~) character as another delimiter. That is, check the Other box under Delimiters and type a ~ in the box as shown below:



This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☑ Tab ☐ Semicolon ☐ Comma

☐ Space ☑ Other: ~

☐ Treat consecutive delimiters as one

Text qualifier: "

Data preview

Category	Term	Count	%	PValue	Genes
GOTERM_BP_FAT	GO:0007049	cell cycle	48	19.047619047619047	4.812150
GOTERM_BP_FAT	GO:0022402	cell cycle process	40	15.873015873015872	1.552128
GOTERM_BP_FAT	GO:0000278	mitotic cell cycle	32	12.698412698412698	3.711402
GOTERM_BP_FAT	GO:0022403	cell cycle phase	33	13.095238095238097	1.125875
GOTERM_BP_FAT	GO:0048285	organelle fission	24	9.523809523809524	1.408013

This will split your GOterm ID from the GOterm name. You will need to adjust the column headings as once you've split the GO data, the columns headings are off by one.