Tutorial: Using DAVID bioinformatics resources

https://david.ncifcrf.gov/

The website DAVID bioinformatics resource 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 the same set of genes related to cardiovascular disease (CVDgenes) used in the WebGestalt tutorial.

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 Exercise 4 home page and click on the link for the CDDgenelDs.txt from the course website. This is some from a comparison of normal to ischemic heart. Either open them and copy and paste or use the Choose File button to upload the list of genes into the text box. Select Entrez_GENE_ID 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 menu will switch to the Gene List Manager tab and under the List Manager, there should be a line that reads List1. You can rename a list by clicking the Rename button and call it something else.

You can also upload a file using the Choose From a File if you have your identifiers in a text file. Using unique identifiers (e.g. Entrez Gene IDs reduces the issue with multiple species being recognized).

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 1 below.
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 Homo sapiens, set the Similarity Score (Kappa) to >=0.75 and then click the Rerun Using Options button. Now the search will find any genes in the human genome that are functionally related to the one in the list.

4) **Gene Ontology enrichment**

Return to the main page by clicking the Functional Annotation under the Shortcut to DAVID tools menu. Then expand the Gene_Ontology section. Here we will do the same gene ontology enrichment as you did with WebGestalt. There should be a check next to GOTERM_BP_DIRECT, GOTERM_CC_DIRECT and GOTERM_MF_DIRECT. Click on the Chart button to the right of GOTERM_BP_DIRECT and it should open a new window (rather than a tab) as shown in Figure 2. If you compare this with the results from WebGestalt, the first thing you should notice is that many more terms were identified. First change the stringency using the Options menu. Change

---

**Figure 1:** Gene List Report from DAVID.

**Figure 2:** GO enriched categories from DAVID
EASE to 0.01, check Fold Enrichment and FDR and uncheck Benjamini. Then click the Rerun using Options button. Now the window should display 38 Chart records. By default, they are listed in decreasing order of p-value. You can compare the two lists, but WebGestalt still does a better job of removing redundant terms than DAVID even at the higher stringency. If you want to see the genes for each enriched term, click on the blue bars under the Genes column. Which annoyingly opens it up in a Tab from your original list while the Chart remains its own window.

5) 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 10 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. Expand the Options section and here you can alter the stringency of the analysis. The default is a medium level. Switch it to High and then click the Rerun using options button. Now the number of clusters has decreased to 6.

Gene Group 1 contained 7 genes, but just looking at the gene names may not tell you why all of these genes are grouped together. 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 7 genes and the number of genes annotated with a given term. From this list of terms, it seems all 7 are annotated as secreted, containing a signal sequence and likely glycosylated. 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 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 annotated 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.
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 537 chart records. The chart lists the terms that are statistically enriched, in order of increasing Benjamini adjusted p-value, located in right-most column. You can adjust which stringency of the analysis and change which data columns are show. Expand the Options menu located above the list. Set the following:

- EASE = 0.01
- Check Fold Enrichment
- Uncheck Benjamini
- Check FDR

Now click the Rerun using Options button. It should show only 187 chart records, which is a bit more manageable. 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. For example, for the UP_SEQ_FEATURE term “signal peptide” there are 135 genes in the list of 344, or 41.7% of the total list that are likely to be secreted. 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. The source of the terms or category are indicated in the name of the category UP = Uniprot. GOTERM is a gene ontology term. INTERPRO and SMART are protein-protein interaction databases.

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, one category in UP_KEYWORDS is extracellular matrix and 35 genes and a second category of GOTERM_BP_DIRECT = extracellular matrix organization also has 35 genes. It is highly likely that most of the 35 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
Close the window and return to the main analysis window 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 58 clusters when I did this analysis. 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.

Changing the stringency to Highest reduced the number of clusters to 41. The algorithms are relatively opaque on this, so most of the time you are using your biological domain knowledge to identify which categories are important. Or you find something that sparks your interest, such as what is the role for LRR (leucine rich repeats) proteins in CVD? There is almost never a “one-size-fits-all” approach to analysis of gene lists. There will not be a single 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 likely to pinpoint the 2-5 genes that will form the core of a paper. That part of the analysis is 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, previous data generated in your lab 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. For example, from the functional annotation clustering (FAC) a medium stringency, the 4th cluster has genes related to angiogenesis. Clicking on the red G reveals that they are 20 genes in this list. You can use the Download File button to export these genes to a text file for later importing into Excel. Or, return to the FAC window and check the 4 terms in Annotation Cluster 4 and click Create Sublist button. It should now be available in the List Manager on the left of the main window. Under ListManager, select your submit and click the **Use** button. Then click the Gene Name Batch Viewer under the **Shortcut to DAVID tools**. You should get a chart with the 20 genes that can be exported. Or you can use the list in other analyses within DAVID.
7) **GeneID conversion**  
In case you didn’t get enough of this during the BioMart tutorial, here is another option for converting Gene identifiers. Select the original Genelist 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. If you got a bunch of duplicates, go back to the List Manager and make sure to select only the Homo sapiens species. Out of 314, all 314 were successfully converted. The output should give you a nice 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.  
DAVID also has links to Pathways and Protein-Protein interaction databases but you will use those next week.

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 in **Figure 5**. 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.