1. **A short tutorial of DAVID for project 2**
2. **Introduction**

In this project we use The Database for Annotation, Visualization and Integrated Discovery (DAVID, [http://david.abcc.ncifcrf.gov](http://david.abcc.ncifcrf.gov/)) to validate the clusters of genes you identified in project 2.
In general, DAVID provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes.
In this project, given a list of genes (genes in a particular cluster), we will use DAVID to identify the enriched biological processes, such that we can have a clue on the role of genes played in the underlying biological processes. An ideal clustering algorithm for genes can find cluster of genes which participate in the same biological pathway/processes.
In the following section, we will have a short step-by-step tutorial on how to use DAVID for this project purpose.

1. **Feature ID to Gene ID**

Somehow the feature ID (GI\_) we used in our dataset cannot be recognized by DAVID (PS: it is always hard for the mapping between biological entities). We need to translate/map our feature ID into something that DAVID can be used as input.

The table [here](http://www.cs.wustl.edu/~zhang/teaching/cs517/Spring12/CourseProjects/david_tut/annnot.tsv) provides such mapping. The first column **targetid** is the feature ID in our dataset, and you should use the second column **symbol**, as the input for DAVID. For example, if one of your cluster contains following feature ID:

* GI\_10047091-S
* GI\_10047093-S
* GI\_10047103-S
* GI\_10047133-A
* GI\_10092596-S
* GI\_10092600-S

Then you would like to map them into

* NP25
* HSP70-4
* SS18L2
* EB-1
* MCOLN1
* EIF4G3

And use them as the DAVID input gene list. It is worth to note that there might be the case that multiple feature IDs mapped to the same gene symbol. It is reasonable as our feature in the dataset actually is a probeset of a gene and for one gene there can be more than one probeset to detect its expression levels.

1. **Perform Function Annotation Test**

Follow the URL [http://david.abcc.ncifcrf.gov](http://david.abcc.ncifcrf.gov/) Click **submit** to upload your gene list.



This is what you get after you upload your gene list. There might be a popup says that there are multiple species involved. If it does show up, make sure to select the ¡°Homo sapiens¡± in the first selection-box and click ¡°Select Species¡± such that DAVID know your genes are from Human instead of some other organisms.



Then you need to tell DAVID what kind of functional annotations are you interested to investigate. For the purpose of this course, we should select the **GOTERM\_BP\_FAT** which is the summarized version of Biological Processes in the [Gene Ontology](http://www.geneontology.org/). You can definitely try other Annotation categories, for example, **KEGG\_PATHWAY** in the Pathways node is also a popular choice. You can learn more about those annotations by searching in the help and tool manual of DAVID.



Now you can click on the button **Functional Annotation Chart**. A window will be prompted to show the results of functional enrichment test. Basically this test exams the significance of enriched annotation (GOTERM\_BP\_FAT) in your gene list. Each row represents an enriched functional annotation and is ordered by their significance level. A term is enriched in the list of genes, if there are lots of genes are associated with this annotation.



To test it significance, DAVID used a [hypergeometric test](http://en.wikipedia.org/wiki/Hypergeometric_distribution) based method. In a nut shell, it test the null hypothesis that the enrichment of an annotation is purely by chance. The test is measured by the p-value, which is the second to last of the results table. The smaller the p-value is the more unlikely that its enrichment is purely by chance, indicating that the finding is significant (better). The last column is ¡°Benjaminin¡± correction of the p-value, which correct [multiple test error](http://en.wikipedia.org/wiki/Multiple_comparisons) when we perform multiple hypothesis tests.

The aim of your analysis is that you should find cluster of genes which show significant functional annotation enrichment and those enriched annotations can be related to the Alzheimer ¡¯s disease. In other words, you should try to make the result table contain more annotations whose ¡°Benjaminin¡± corrected p-value is less than a given threshold (You can set it to 0.01 or 0.05). The more significant annotations enriched in your gene list, the more we can infer the functional role of those genes. If you find enriched annotations, such as **G-protein coupled receptor protein signaling pathway**, **defense response**, and**neural**-related terms then you are probably on the right track.

Source: <http://www.cs.wustl.edu/~zhang/teaching/cs517/Spring12/CourseProjects/david_tut/david_tut.htm>