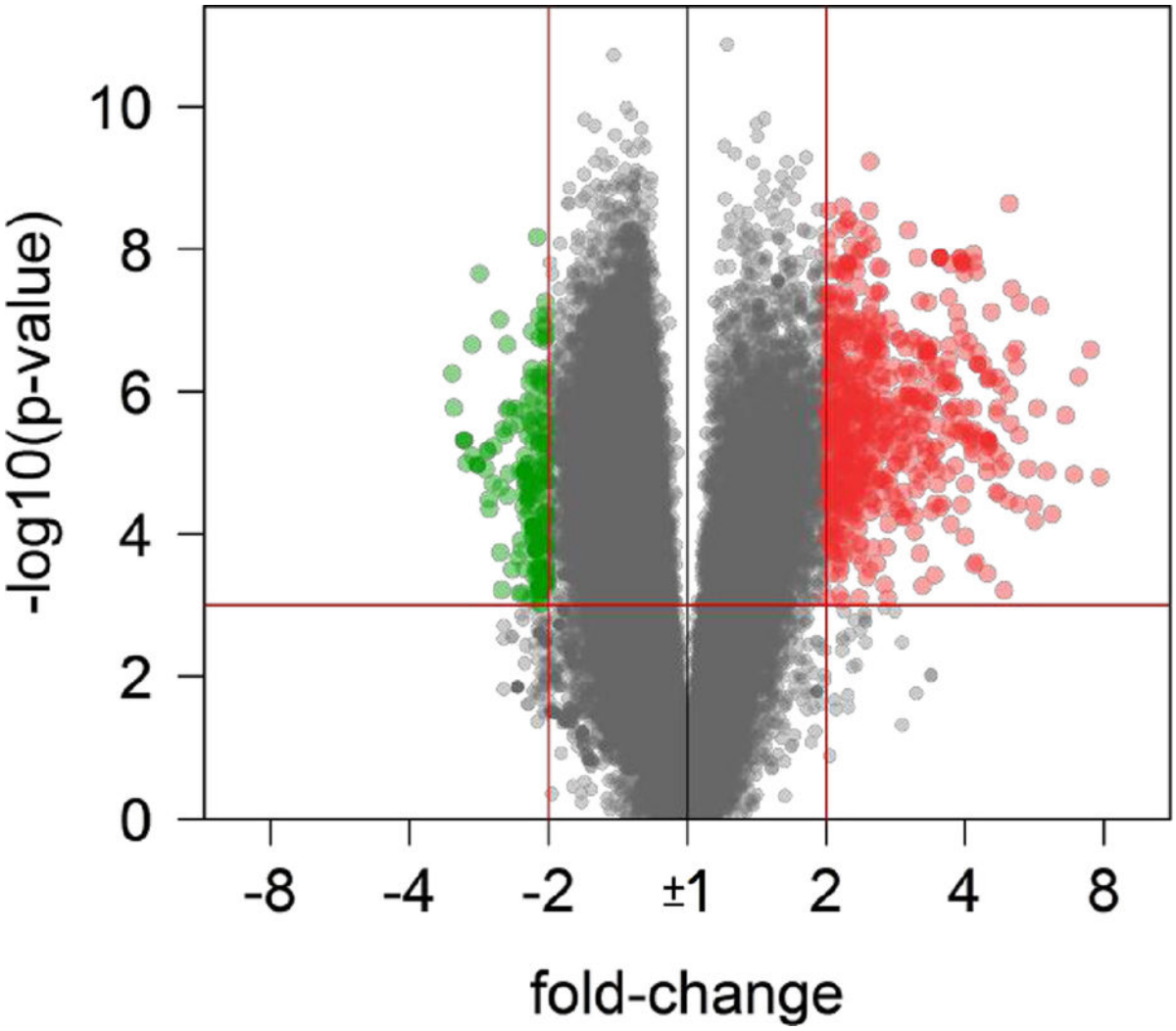
Lab 17

Microarray analysis of squamous cell lung cancer

BACKGROUND: Lung cancer is a heterogeneous disease with several well-defined subtypes. One of these, squamous cell cancer (SCC), generally arises from the bronchial airway epithelium, and has features reminiscent of squamous epithelium, which includes flat, layered cells that line cavities such as the mouth or throat, or epidermal skin cells (aka keratinocytes).

In today's lab you will use three web resources to see which changes in gene expression occur in squamous lung tumors. The first of these, the Gene Expression Omnibus (GEO), is the largest repository of public gene expression data in the world. The second of these, DAVID, is a free set of tools from the National Institues of Health (NIH).

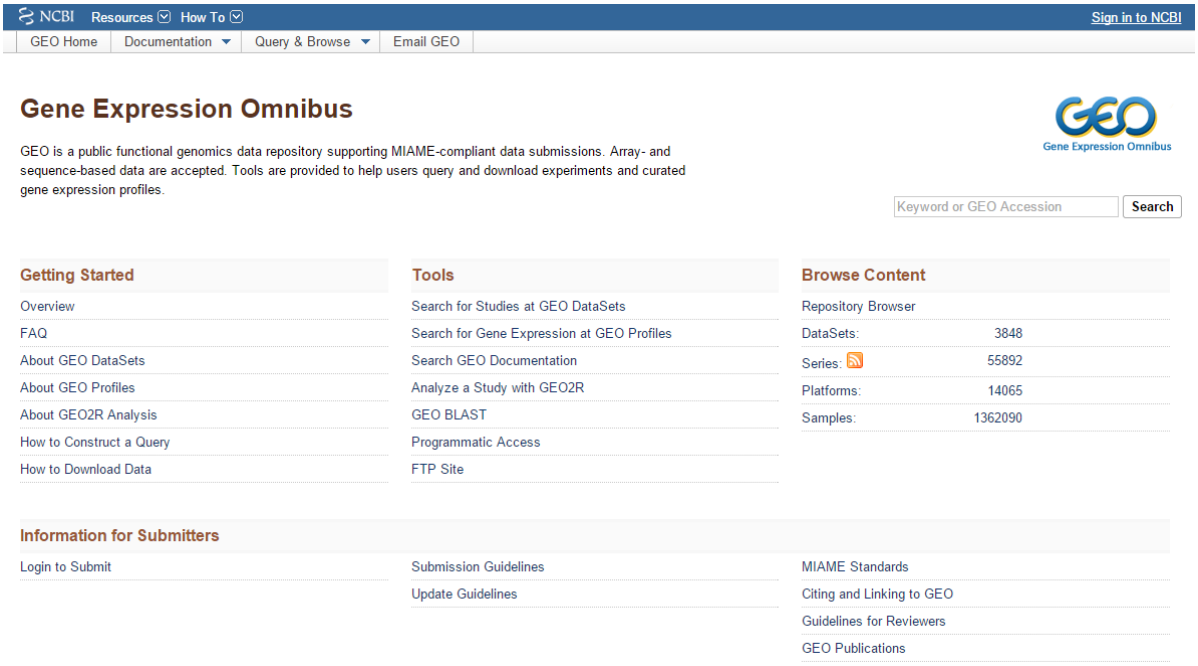
Some Jupyter. Embed images into Jupyter notebook from anywhere! Change this cell from Raw NBConvert (a.k.a. simple text) and see what happens.



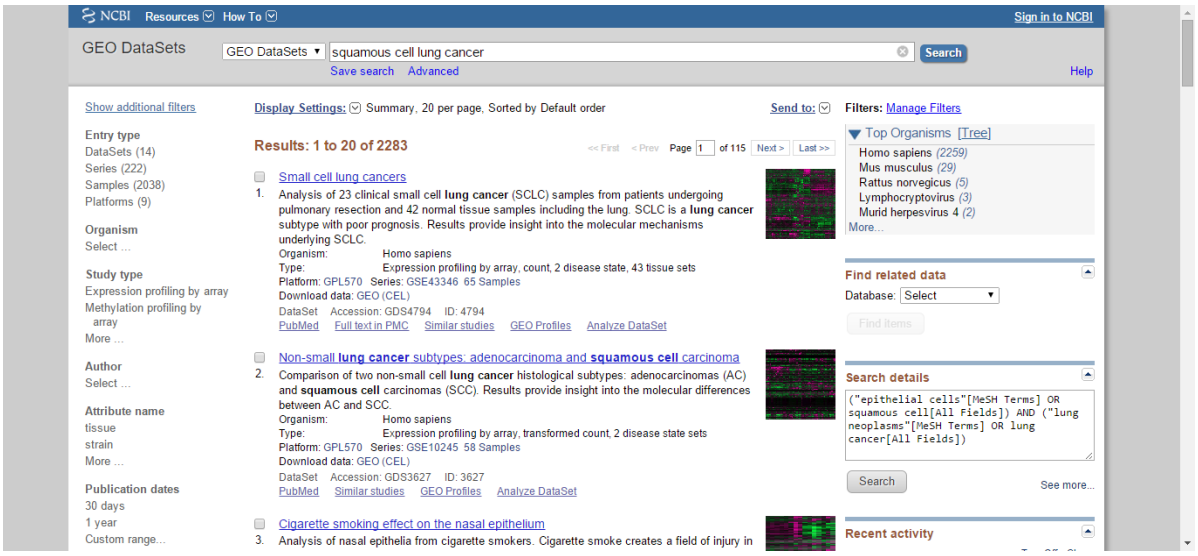
PROCEDURES, Part 1 (GEO):

Today you'll be performing analysis on a GEO DataSet. GEO DataSets are sets of raw gene expression data (GEO Series) that have been manually curated so that analysis can be performed quickly between two groups of samples (e.g., normal vs. disease, drug A vs. drug B).

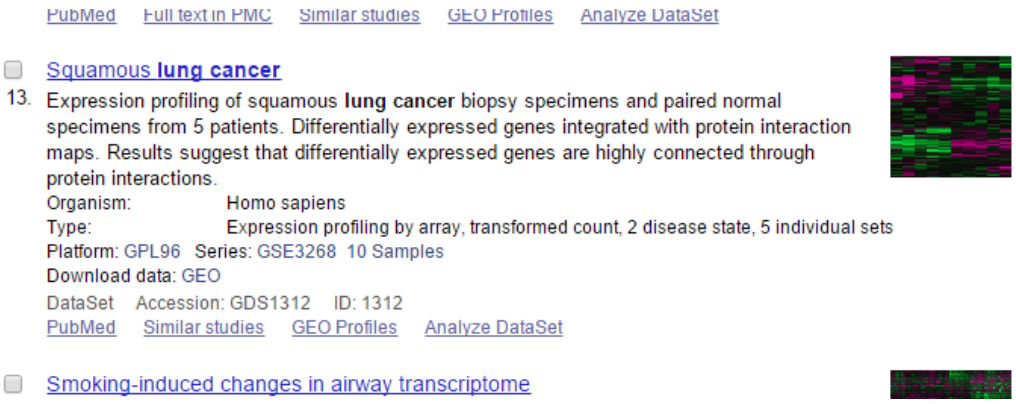
Start by opening the GEO start page at: http://www.ncbi.nlm.nih.gov/geo/



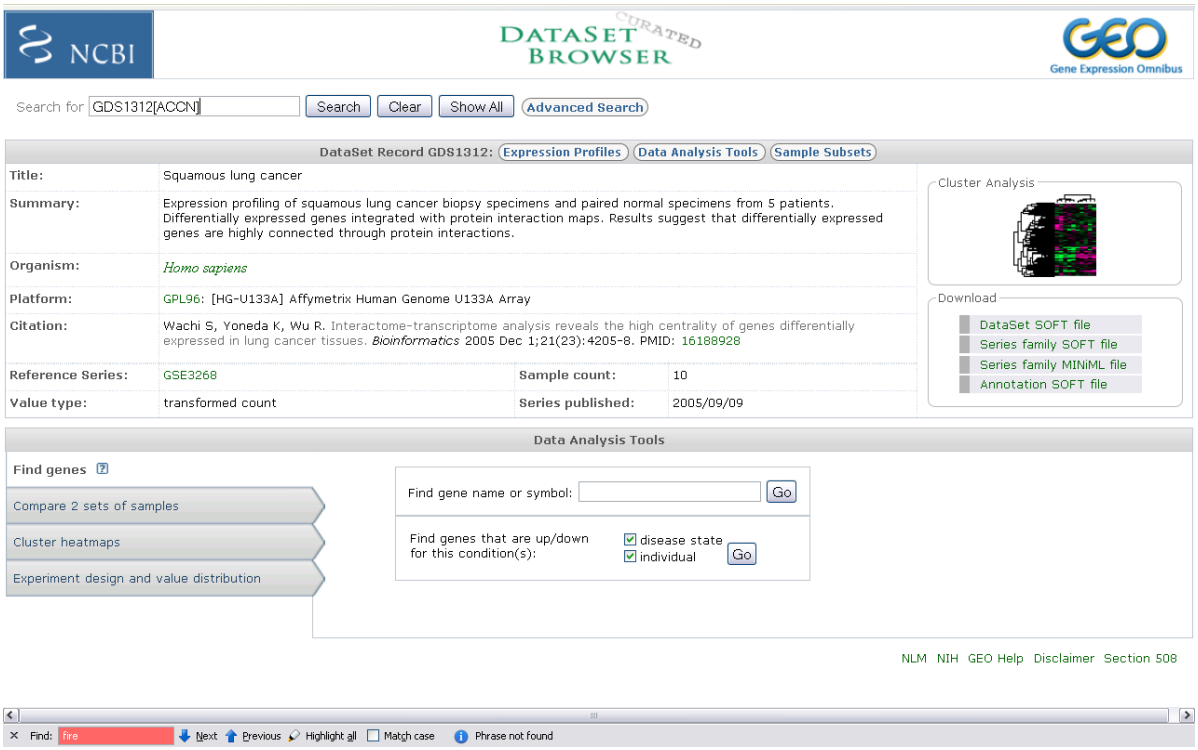
The GEO Query interface is organized according to the different types of results. Click Search for Studies under Tools and then type “squamous cell lung cancer" into the DataSets text box and click SEARCH.



Several kinds of search results (DataSets, Platforms, Samples, and Series) are returned. Click on DataSets at the left to restrict the search to curated DataSets.

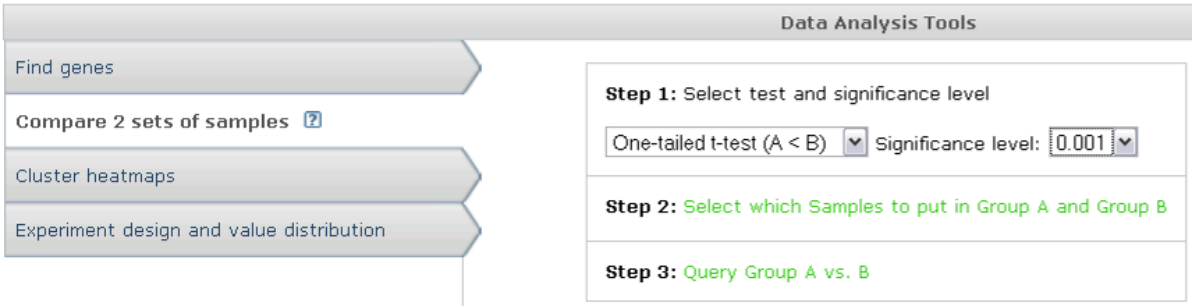


Scroll down and click on DataSet GDS1312. This is a set of lung tumor biopsies and paired normal samples from 5 individuals, and is a good example of a typical microarray experiment.

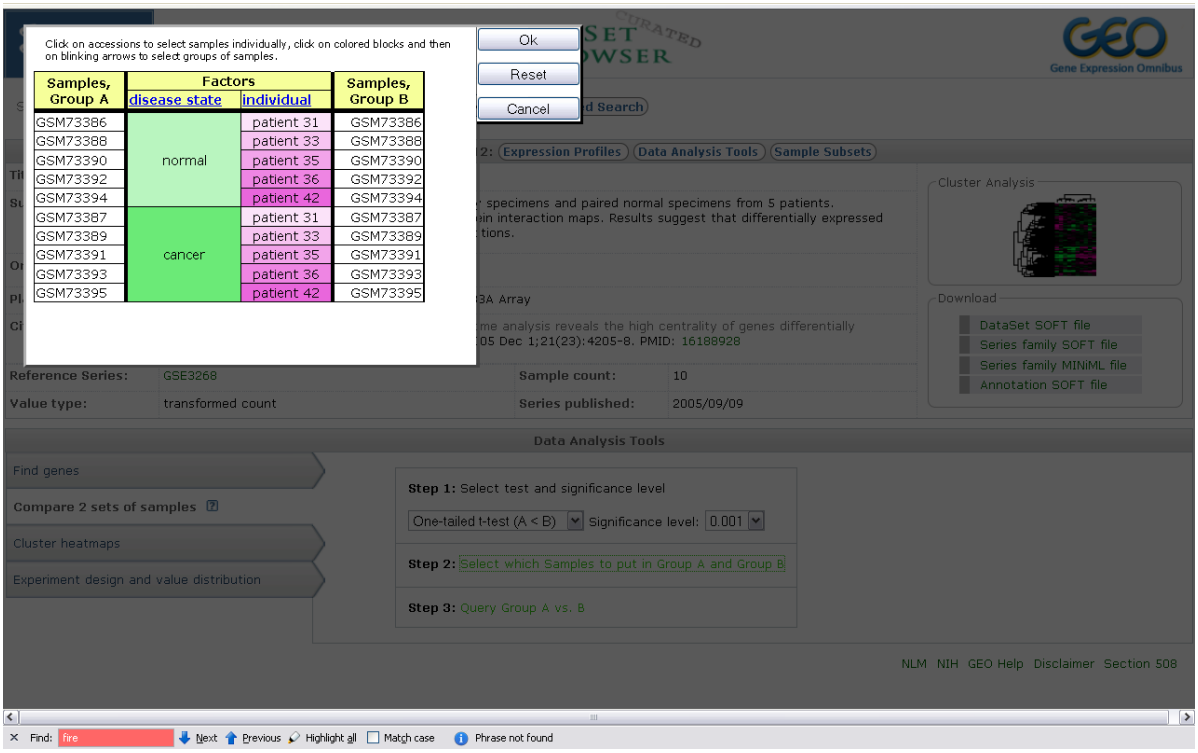


In this lab, your goal is to find groups of genes that have significantly higher expression in squamous cell tumors than in normal lung tissue. One way to do this is with a t-test.

Click on the **Compare 2 sets of samples** tab at the bottom left.



Step 1: Choose the **One-tailed t-test (A < B)** and a **Significance level (p-value) of 0.001**.



Step 2: Click **Select which Samples to put in Group A and Group B.** Click "normal" and click the left arrows to assign that group to Group A. In the same way, assign "cancer" to Group B. Click **Ok.**

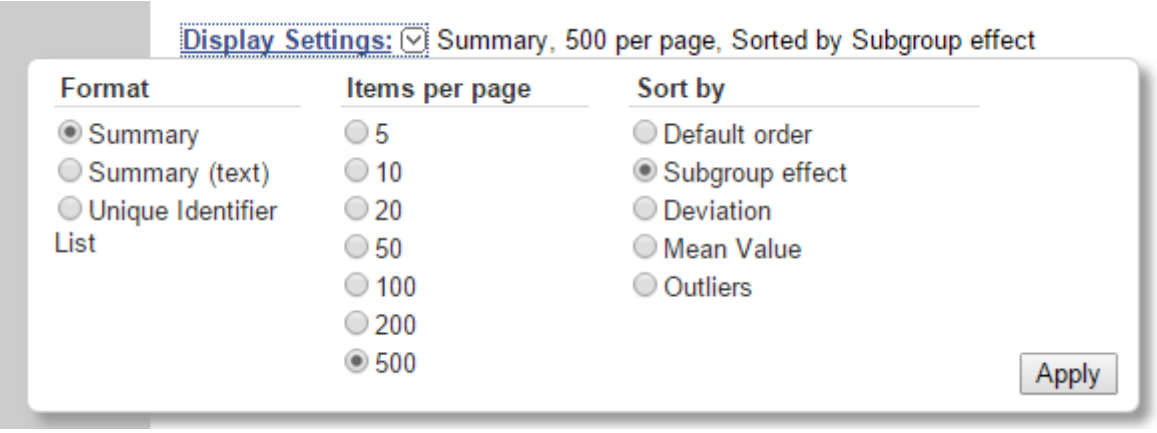
Step 3: Click **Query Group A vs. B** to perform the t-test. A new window will open with the results. *Question: How many genes were significantly increased in tumor vs normal at this significance level? There are 22,283 genes on this chip. About how many genes would you expect to find differentially expressed by chance alone at p < 0.001? How many would you expect to change in each direction (higher in tumor vs. higher in normal)?*

There are 327 genes upregulated in tumor vs normal at this significant level. 45 genes can be expected to differentially expressed by chance. Aprroximate 22 genes can be expected to be upragulated in tumor, and about 22 genes can be expected be higher in normal cells.

Because this data set is small, and we want to find genes with strong changes in expression, we might benefit from using a more robust measure instead: ranked fold change. To run this test, first close the current window and return to the DataSet record for GDS1312.

Step 1: Choose **Rank means difference** and a fold-change cutoff of 3 (A < B).

Step 2: You already did this.

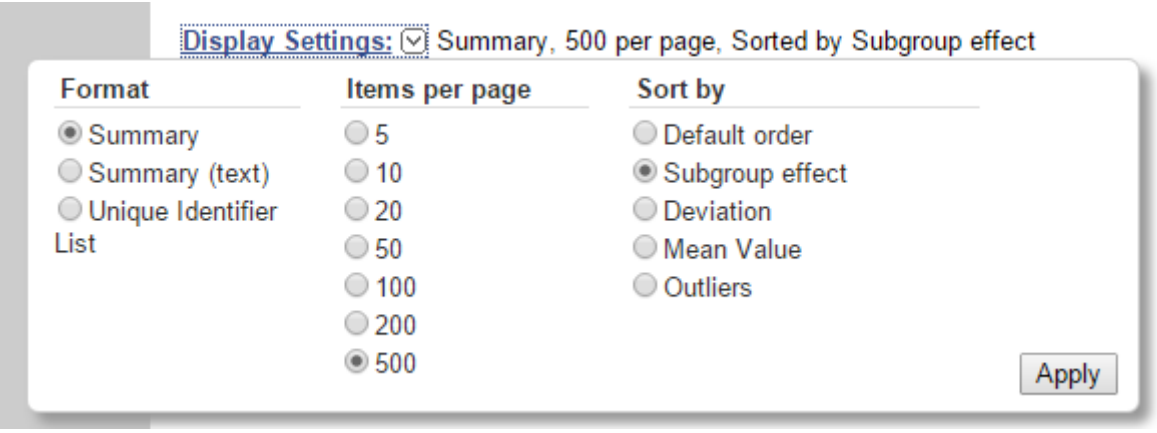


Step 3: Click **Query Group A vs. B.** A new window will open with the results. Change the **Sort By** dropdown box to read **Subgroup effect.** This will sort genes in order from greatest to least fold change.

Question: What is the symbol of the gene at the top of the list? Click on the link in the Annotation line to go to its Entrez Gene page. What's the gene's common name?

KRT6C keratin 6C

Go back to the page with the results of the ranked fold change analysis.



Under **Display Settings** Change the **Show** button to **500**, to show the first 500 of the genes that are at least 3-fold higher in squamous cell cancer than normal tissue.

Click on the **Download profile data** button to retrieve a data matrix for the entire list of genes.

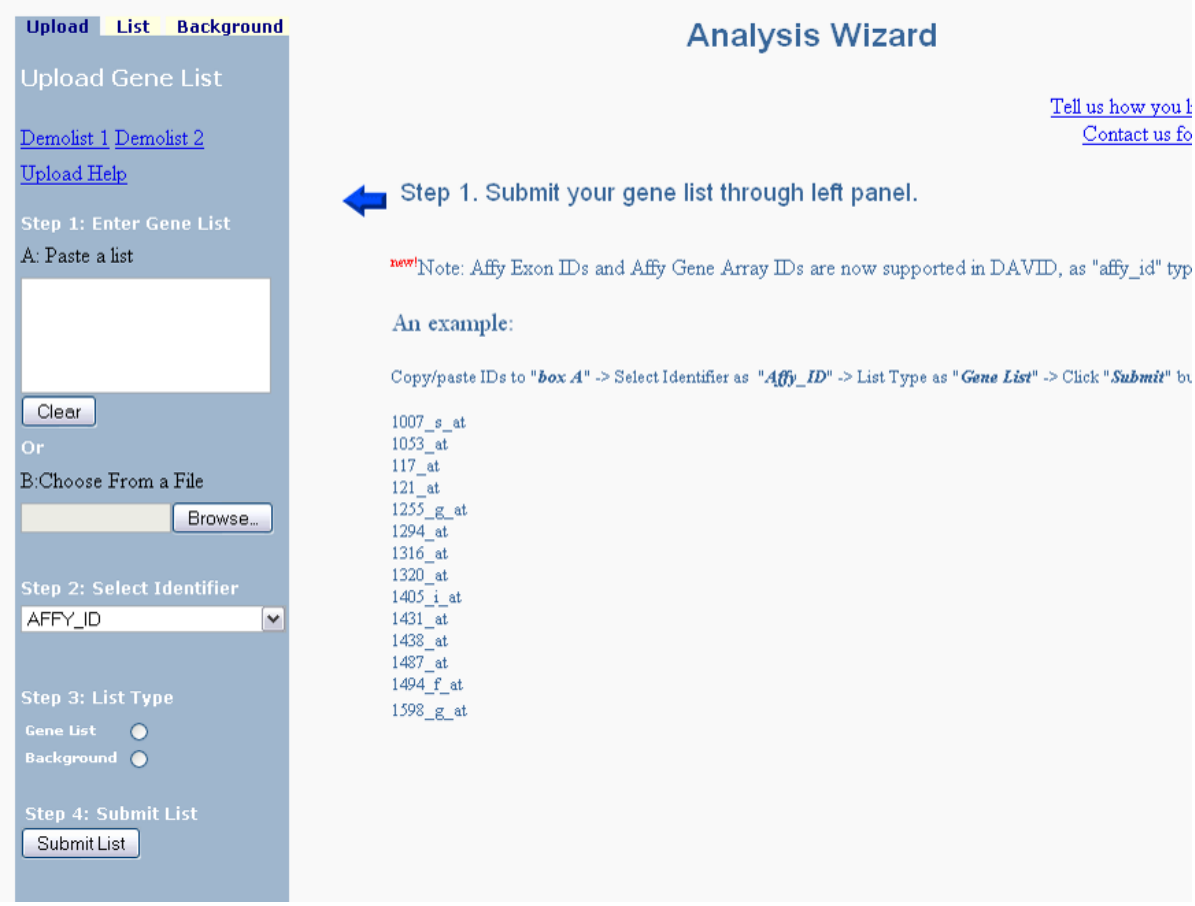
Open the file with Microsoft Excel or OpenOffice Calc. Select the Affymetrix probeset identifiers (they should look something like 2xxxxx_at) in the first column (starting in row 5) and copy them to the clipboard (Ctrl-C or Edit->Copy).

Alternatively, try selecting the Affy IDs using UNIX!
Hint: you should use 'less', 'cut', and 'grep' commands, connected through pipes.

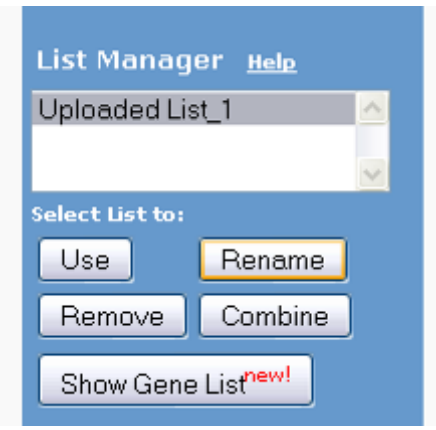
PROCEDURES, Part 2 (DAVID):

To analyze the functions of the genes in the list we just generated, we need some better tools. DAVID is a free public resource that allows rapid functional annotation of lists of genes. Open a new browser window to: http://david.abcc.ncifcrf.gov/

Click **Start Analysis** to begin.



To upload the list of Affymetrix probeset IDs you just copied from the file you downloaded from GEO, click in the **Paste a list** box under **Step 1** on the left side of the screen, and paste them in (Ctrl-V or right click-> Paste). Leave Step 2 as **AFFY_ID**. Select **Gene List** under **Step 3**, and click **Submit List**.



The next window of the Wizard will appear. Rename the list to something more useful with the **Rename** button.

The next step is to run a clustering analysis that will group genes into sublists based on their shared functional annotation (GO terms, KEGG pathway terms, etc.)

Click on **Gene Functional Classification Tool.** A results page will appear with a set of clusters of genes.

To see which terms are enriched in each cluster, click on the red **T**erms symbol at the top right of each cluster.

Question: What do the functions of the top 5 or 6 clusters seem to be? Describe each in a few words using the Terms feature. Why do you think some of these clusters might be upregulated in tumor cells?

The functions of top 1,3, and 4 clusters seem to be related to activities of cell division, mitosis, the formation of cytoskeleton,etc. These functions are upregulated in tumor cells may because the tumor cells often reproduce very quickly. The top 2 cluster is related to tumor antigen, which might be a common feature in tumor cells. The top 5 cluster is related to protease inhibitor function, which may be related to inflammatory responses in tumors.