**Introduction to Genomic Databases and Analysis Tools**

**Hands-On Exercises: Step-by-step instructions**

Goal: Get comfortable browsing through a genomic dataset repository, and analyzing the data from a complete microarray gene expression experiment. It’s an easy way to get your feet wet with a big dataset, see some of the data analysis methods in practice, view the programming code that is used, and explore the biological meaning of all those gene expression changes. We will all work on this together.

**Part I – Discovery and review of a dataset from a single experiment**

1. Go to the GEO website <http://www.ncbi.nlm.nih.gov/geo/> and search with the term “sjogren”.
2. Click on the 26 results in GEO DataSets Database. 
3. On the left side of the screen, use the metadata for these records to:
	1. Filter for Series
	2. Filter for Expression Profiling by Array



1. Click on the Series Record entitled “Gene expression data of parotid tissue from Primary Sjogren’s Syndrome and controls”. Together we will review the GEO record to understand:
	1. where the samples came from
	2. how the data were analyzed before they were deposited in GEO
	3. what type of microarray was used
	4. which samples are from patients with the disease, which from normal patients, etc.
2. Click on an individual sample (for example, GSM997868). Together we will review the metadata for this individual sample (a single microarray)
	1. Tissue that the mRNA was isolated from
	2. Characteristics of the sample
	3. Table of data from the microarray
3. Go back to the full record for this Series by clicking the Back button on your browser
4. Scroll down the page and click on “Analyze with GEO2R”



**Part II – We will compare gene expression levels in patients with Sjogren’s Disease to gene expression levels in normal healthy controls. We will create a comparison between these 2 groups of samples and apply statistical analysis of the data using GEO2R. The result from this will be an Excel spreadsheet with genes whose expression is significantly different in patients with the disease, relative to normal controls.**

1. Define the groups you will be comparing.
	1. Type “Controls” into the box and hit return.
	2. Type “Sjogrens” into the box and hit return.



1. Assign your samples to the appropriate group by clicking and selecting samples to highlight them, then click on the name of the group the sample belongs to (control or patient).
2. Scroll down the page to the GEO2R tabs and click on Value distribution. We are going to look at how “clean” or “noisy” this data might be by viewing the distribution of the data across all samples. To do this, Click View in the Value Distribution tab. This will take a few minutes to complete.



1. We will review the expression data across samples to see if the samples look uniform, or if there are big differences between the microarrays. This is data that has been normalized (“smoothed out”). What does it look like?
2. Next, click on the GEO2R tab, then click on the Top 250 button. This will start the gene expression comparison analysis.
3. While this is running, click over to the R script tab, and we will take a look at the programming code that is running in the background.
4. Click back on the GEO2R tab and wait for the analysis to finish. We will review the results of the data analysis together:
	1. Identifiers (IDs) are now translated into their gene names (Gene.symbol).
	2. Adj.P.Val tells you whether or not the gene expression levels are significantly different between patient samples and control samples. If the value in this column is <0.05, the gene expression change is significant.
	3. logFC tells you the magnitude of the gene expression change. It is in log2 scale. So a value of 3.67 is 2 to the power of 3.67 (2^3.67), or ~12 fold. In other words, the levels of gene expression for this gene are 12 times higher in patients than in normal healthy controls.
5. Click on the name of the gene at the top of the list, and we will review the graph that appears below it together.
6. Now we are going to export all of this data into an Excel file so we can explore it biologically and understand what it all means. To do this, click on “Save all results”



1. When the job finishes, the data will appear in a new tab in your browser. Right click on the webpage, click Select All, Select Copy.
2. Open up a blank worksheet in Microsoft Excel.
3. Right click in the top left empty cell on that blank Excel worksheet.
4. Select Paste Special, then select Unicode Text. Click OK.



Congratulations! You now have a fully analyzed microarray dataset from actual patient data.

In the last hands-on exercise in class, we will figure out what it all means by exploring the data in the context of biological pathways. And what – if anything – those pathways tell us about our patients with Sjogren’s Disease.