CSC 314, Bioinformatics **Name:**______________________________

The Gene Expression Omnibus

The Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) is a public functional genomics data repository for gene expression (microarray) and sequence-based data. Like GenBank and GenPept, GEO is hosted by the National Center for Biotechnology Information (NCBI) under the National Library of Medicine (NLM) at the National Institutes of Health (NIH).

There are four kinds of records on GEO (summarized at: http://www.ncbi.nlm.nih.gov/geo/info/overview.html):

1. A **GEO sample** (GSM*) describes an individual sample, including the experimentally conditions in which it was collected, and the gene expression value for each element on the array.
2. A **GEO platform** (GPL*) is a summary of the array used, and links the array probe to a gene
3. A **GEO series** (GSE*) links together a collection of samples with one or more platforms for a particular experiment or study (such as profiling gene expression from 100 patients with lung cancer)
4. A **GEO dataset** is a curated collection of samples that allows for user-friendly analysis. Not all series exist as datasets.

Quick searching

1. Go to the GEO homepage (http://www.ncbi.nlm.nih.gov/geo/), search for *bladder cancer*, and look at the results in the GEO Datasets database (this includes the 4 kinds of records above)

2. How many series are there? How many samples are there?

Using the Dataset Browser

1. Let's look at the *Bladder tumor stage classification dataset.* Stage is a classification of how far the tumor has spread, which in bladder is either non-muscle invasive (Ta-T1) or muscle invasive (T2-T4). You can see a a figure here: http://www.cancerresearchuk.org/cancer-help/type/bladder-cancer/treatment/bladder-cancer-stage-and-grade

2. Click on *Experiment design and value distribution* – notice that the samples can be grouped by stage and grade, and note the color codes used.

3. Click on *Cluster heatmaps*. A heatmap is a visualization of the data, with each row corresponding to a gene (or probe) and each column corresponding to a sample. Colors are used to represent the amount of gene expression. Clustering involves grouping together genes with similar expression patterns and patients with similar expression profiles

4. Click on *Compare 2 sets of samples*, select the appropriate test and let's compare NMI (Ta-T1) tumors with MI (T2+) tumors. One of the genes will be RPLP2. Note: this analysis show you significant genes, but does not give you p-values.

5. Click on Find Gene. Let's look at FGFR3.

Looking at GEO series and analysis using GEO2R

1. Let's look at the GEO series GSE3167.
2. What platform were these samples profiled on?
3. Look at the sample GSM71028. What type of tumor did this patient have?
4. For this sample, what is the expression value for the first probe, 1007_s_at, and what gene does this probe correspond to? (Note: you need to look at the platform data to find the gene).
5. From the main series, page, click on Analyze with GEO2R. Let's find the top 250 differentially expressed genes for tumor vs. normal samples. What is the p-value, adjusted-p.value and fold change (FC) for SH3GL3. Is it upregulated (higher) or downregulated (lower) in tumors compared to normal samples?
6. The adjusted p-value indicates that the probability that this probe is a false positive is _______________?
7. The p-value for this gene corresponds to the following probability: if there really was no difference between the tumor and normal samples, the probability of observing a log2 fold change of at least _____________ is equal to ____________________?