Bioinformatics 525 – Module 3 – Homework 2
Due 4/6/2017, 5 pm

Work through the exercises below, and provide your answers in either paper or electronic form to Lauren Jepsen (ljepsen@umich.edu) by the deadline.

Choose a disease or chemical perturbation that interests you, and find a GEO Curated Data Set that allows you to compare expression profiles of cells that are and are not affected by that disease or perturbation. (If your first idea doesn't have data sets available, you can search GEO as a whole if you're feeling adventurous, or pick a different target).

Describe the condition pair that you plan to look at, and based on your background knowledge, formulate three hypotheses regarding the expression of different genes in affected vs. unaffected cells.

Now, record the expression levels (using percentile and/or value units) from all replicates from the affected and unaffected conditions that you are comparing.

Apply a statistical test of your choice to each of these three cases to determine whether the expression changed substantially in the direction that you expected, and record the results and likely implications for the hypotheses that you formulated. Also state the test that you applied, and justify your choice.

For more untargeted analysis of the expression levels in your data set, go back to the initial GEO page for your GDS entry and download the "DataSet full SOFT file" from the download menu on the right side of the screen. Open the SOFT file in a text editor (after uncompressing it) and delete all lines up to and including the one indicating "!dataset_table_begin". Also delete everything at and after the line containing "!dataset_table_end".

At this point, you have a simple tab-delimited data table that can be loaded into R, MS Excel, or anything in between. Load the data table into your analysis platform of choice, and identify the genes showing the largest (in magnitude) increase and decrease in expression between your conditions of interest. (Note that you will likely have to look up which of the GSM identifiers in the table correspond to each of your conditions).

What genes showed the largest increase and decrease in expression between the conditions? If you apply a test similar to the one that you did above to look for differential expression, to those genes show significant changes?