Homework 1 (100 points, Due date: Wednesday, March 13th, 11:59 PM):

For our first homework assignment, we will apply the statistical methods we learned in class to identify differentially expressed genes between 179 lymph-node negative metastasis-free patients and 107 lymph-node negative patients that developed a subsequent metastasis. This gene expression data was generated using the Affymetrix Human Genome U133A microarray. The detailed description of the microarray data and the authors' original motivation in generating it can be found in Wang et al. (https://www.sciencedirect.com/science/article/pii/S0140673605179471).

The raw data is publicly available from the Gene Expression Omnibus (GEO) database, which will provide you with links to the raw data, but we have also preprocessed the data for you so you can instead just focus on the downstream analysis. We have used an R script to preprocess the data, but please feel free to implement your solution in any language you choose.

Note about working in teams: As described in class, you have the option of completing this homework individually or as part of a team of **at most** 2 students. Teams must consist of students with complementary background/expertise (e.g. one student with a primarily computational background, one with a primarily biology background). If you are unsure whether you are a complementary pairing, please confirm with Prof. Myers in advance of completing the homework. Teams will need to answer a few additional questions in the homework (marked with **) and will need to submit a single copy of the report/code implementing the solution. Both students will be assigned the same grade for the assignment.

1. **Acquiring the gene expression data:** Go to the GSE2034 repository page and read the descriptions of the dataset (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034). We have already downloaded the raw data file (`GSE2034-22071.txt`) for you and included it in the HW1 folder. You can also view the clinical information associated with the patients here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GSE2034&id=40089&db=GeoDb_blob26. We have already downloaded these data and included them in the file `clinical_data.txt`. The clinical data links each patient to the GEO accession associated with each array and includes several pieces of information about each patient, including the relapse status which is the focus of this homework. The final piece of data is the mapping between probes on the array and the corresponding genes in the human genome. We have downloaded this information for the Affymetrix Human Genome U133A Array from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL96 (click the "Download full table" button), which we have included as the file `GPL96-39578.txt`.

   NOTE: we have already merged all of these files together into a single matrix using the R script, `import_wang_data.r`, to form the final data file, `wang_data.txt`. This is a tab-delimited file with the sample IDs across the first row, the relapse status for each patient across the second row, and the probe values in the following rows. Each probe row contains the Affymetrix probe ID in the first column, and the corresponding human gene name in the second column. You can execute `import_wang_data.r` and regenerate `wang_data.txt`, but you are not required to regenerate this data file yourself as we have already provided this file in the HW1 folder. Please just use the text file we have provided.

2. **Exploring the dataset (10 points):** Process the data file using whatever programming language you have chosen and answer the following questions:
   (a) How many probes are included in the dataset?
   (b) How many patient samples are in the dataset? We will divide patients into two groups based on their relapse status for our analysis. How many patients were relapse free (relapse=0, i.e. no metastases)? How many patients had relapses (relapse=1)?
   (c) How many unique genes are represented by the probes in the dataset? Note: we would typically average multiple probes mapping to the same gene, but to keep the lab relatively simple, we will analyze each probe independently.

   ** **Additional question for teams:** Read Wang et al. (https://www.sciencedirect.com/science/article/pii/S0140673605179471), the paper that originally published this data. What was the goal of their study? Briefly summarize what they did and what they concluded from this gene expression dataset.

3. Data processing and normalization (30 points):

(a) Plot a histogram (x-axis: expression levels, y-axis: probe counts) of the complete dataset. Describe the distribution—what is the overall shape? Replace any values <= 0 with a value of 1, and log-transform (log 10) the entire data matrix. Plot the log-transformed data.

(b) Plot individual histograms of the log-transformed data for the first four arrays (GSM36777-GSM36780). How do the distributions compare from sample to sample?

(c) Perform quantile normalization **on your log-transformed data** (from part b) across all arrays (samples) such that each has the same empirical distribution. Use the mean of each probe across all samples as the reference distribution for this normalization. Plot a histogram of the normalized data for each of the first four samples (GSM36777-GSM36780). Use the normalized data for the remaining problems.

4. Analysis of differential expression (30 points): Use the *t*-test and Wilcoxon rank-sum statistics to identify differentially expressed probes with a per-probe significance level of $p < 0.05$. You should divide the gene expression data into two groups (metastasis vs. non-metastasis), using the relapse variable in the clinical data, and test each probe independently.

(a) For each test, list the top 10 probes, the corresponding gene names, and the p-values associated with them. Pick 1-2 of these genes and discuss what is known about their function and how it might relate to cancer metastasis. There are many databases such as GeneCards (http://www.genecards.org), NCBI (http://www.ncbi.nlm.nih.gov/gene), and WikiGenes (http://www.wikigenes.org) that you can use to learn more about a gene.

** **Additional question for teams:** Comment on how well established the genes that you discovered are in terms of their relevance to breast cancer or metastasis. Are there

other genes that are well-characterized as playing a major role in breast cancer metastasis, but that do not appear in your list?

(b) Report the number of selected probes for each test and plot the histogram of the log-transformed (negative log10) *p*-values of all probes.

(c) What is the overlap between the set of probes deemed significant by the two different approaches (*t*-test and Wilcoxon rank-sum statistics)? You do not need to report the entire list associated with each approach. You only need to report the total number of probes that overlap and a list of those overlapping probes.

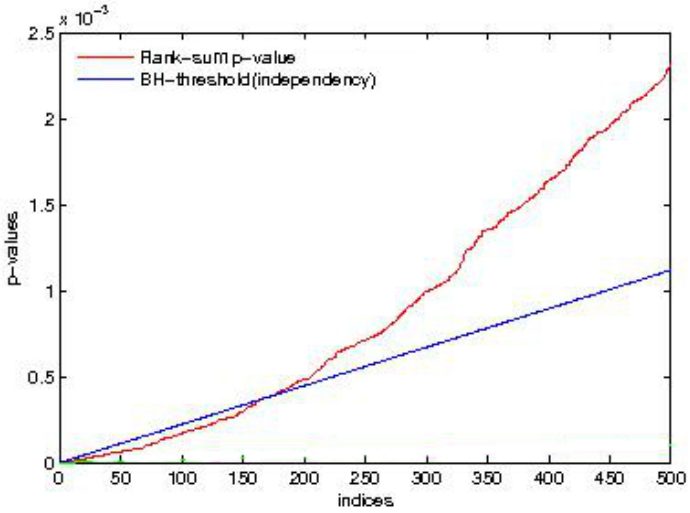
5. Multiple hypothesis correction (30 points):

(a) Use Bonferroni correction with the rank-sum test to identify differentially expressed probes at a global significance level $p < 0.05$ and report the number of selected probes.

(b) Use the Benjamini-Hochberg step-up procedure to control the False Discovery Rate (FDR) with the rank-sum statistic to identify differentially expressed probes at an FDR < 0.05.Report the number of selected probes by the BH procedure (specify which independence assumption you are using for the BH procedure).

(c) Rank the probes by their *p*-values in ascending order and plot the *p*-values and the threshold used for adjusted *p*-values as a function of the index of the ranked probes. More specifically, with the x-axis as the index of the ranked probes from 1 to the number of probes, plot the first 500 probes (*i* on the x-axis, the *p*-value of the i^{th} probe on the y-axis for each probe). As a separate color, plot the adjusted p-value threshold at each point (*i* on the x-axis, threshold for adjusted p-value on the y-axis), where *i* is the index of the genes and alpha=0.05.

HINT: Your plot should look like this:



6. *Extra Credit* (5 points) Use the data from this paper: van't Veer *et al.* "Gene expression profiling predicts clinical outcome of breast cancer", Nature 415, 530 - 536 (2002), which also studied gene expression signatures associated with breast cancer metastases and perform a rank-sum test with FDR (as in problem 5; you do not need to do data processing and normalization). You can find the original dataset in the `hw1_extra.zip` file. Select the top 100 genes. What is the overlap between this list and the one you created earlier (from the Wang *et al.* study)? Discuss what you find.

Submission Instructions:

Zip all files into a single lastname.zip (individual submission) or lastname1_lastname2.zip (group submission) file and submit your .zip file on the Moodle site. Please avoid including the raw data files we provided in your .zip file. Your homework submission should only include:

1. Any source code you used to complete the assignment.
2. Report.pdf: A file with all of your plots and answers to questions.