HW3 Finding Differentially Expressed genes
Due March 6 (tentative date)

Submit two files. One **pdf** document as a short report addressing each question (do not include your R code or directly copy-paste R output. For example, never report a value with 6 significant digits unless you can justify the need.). Separately, submit an **R** file that contains all your code that the instructor can run without error and generate your result. Keep your code well commented and it will become useful later.

In your report, briefly describe your analysis and explain your decisions. Use figures and tables to help illustrate your point, whenever you find them useful.

Using the dataset *estrogen* as you have seen in the lab material.

1. Perform exploratory analysis to check the quality of the data. Do you see any problem with the image? Do you think the data needs normalization? Choose your preprocessing method and convert the probe level data to expression measures. Do you spot any problem at the probe level data that may concern you in downstream analysis?

2. Follow the example in lab material,
    a. Find the top 20 genes that appear to be differentially expressed, comparing estrogen vs control at hour 10. (Always report the estimated FDR level, the amount of DE, in addition to the identity of the genes).
    Note: "top 20" means the most interesting 20 genes selected by you. Use your judgment.

    b. Find the top 20 genes that appear to be differentially expressed between estrogen and control groups at hour 48. Are there any overlaps between the two lists?
    c. Find the top 20 genes that show the strongest time effect in the controls
    d. Find the top 20 genes that show the strongest time effect in the estrogen group. Are these the same genes as in Question c?
    e. Is there an interaction between treatment and time in gene expression?

    Did you find some control genes (those that start with "AFFX" in their gene names) in some of these comparisons? Which control genes appear to have been added (spiked) in the samples, and which may have been spiked in inconsistently in the samples?

After you see the top 20 genes in each contrast, can you determine the size of the top list of genes by some measure of statistical significance? How many genes would you find interesting if not 20?

Write a short summary of your findings in the effect of estrogen on gene expression. Provide a visualization of your chosen genes in the context of the other genes' expression profiles (for example, you may use MA plot and/volcano plot).

2. Download the CEL files from the dataset GDS2938 on GEO.
Provide a short description of the experiment.
Use your choice of preprocessing to obtain gene expression levels.

Identify a list of genes whose expression level is changed by the IFN-gamma treatment, if there are any. Identify a list of genes whose expression level is changed by the IL1-beta treatment, if there are any.

Do IFN-gamma and IL1-beta cause similar changes in gene expression regulation?

Provide a summary of your analysis, use summary statistics, tables, and figures as you see fit.