GENOME 560, Spring 2012
Problem Set #1

(Due May 10th 9:00am)

---

1. [**100 points**] **Computing Descriptive Statistics and Visualization**

   In this question, we will do a simple analysis on published microarray expression data using R. Download the data file from the course website `http://www.cs.washington.edu/homes/suinlee/genome560/RMA_Filtered.txt`, and do the following using R.

   (a) [**20 points**] Pick the first sample (named "CEU_1_1") by doing the commands: a1 < − exp[,2]. Then, use the "range", "quantile", "mean", "var" and "sd" command to obtain these descriptive statistics. Submit the results. For quantiles, let's try 2.5%, 5%, 25%, 50%, 75%, 95% and 97.5% points.

   The lower 25% and the upper 25% point in a Normal distribution will be -0.67448 and 0.67488 standard deviations from the mean. Is this approximately true for this dataset?

   (b) [**10 points**] Use the "hist" command with 50 bins classes to examine the distribution of the expression levels in a1 created in part (a) above. Print the resulting histogram as we learned in class and submit the commands and the figure.

   (c) [**20 points**] In class, we calculated simple descriptive statistics for the first probeset, including the overall mean, the CEU specific mean, and the YRI specific mean. Repeat these calculations for all 5194 probesets. First, average the replicates as we did in class; thus each probeset will consist of 8 CEU expression levels and 8 YRI expression levels.

   (Hint: There are many ways to do this including a simple for loop or more efficiently with a new R command related to "mean". A related command often appears in the help page under the section "See Also" at the bottom of the page.)

   Submit the commands that you used to obtain the desired means.

   (d) [**10 points**] Make a boxplot consisting of the 5194 overall mean expression levels, CEU specific means, and YRI specific means. Submit the commands you used and the figure.

   (e) [**10 points**] For each probeset, calculate the difference in average gene expression levels between the CEU and YRI samples. Calculate the standard deviation of the resulting 5194 values. Submit the commands that you used to generate the result.

(f) [**10 points**] Make a histogram of the results obtained in part (e) above. Modify the histograms x-label, y-label, and title to be more descriptive than the default values. Submit the figure.

(g) [**20 points**] Repeat (e) and (f) with “permuted” labels. First, consider the data created in (c) where each of the 5194 probesets consists of 8 CEU expression levels and 8 YRI expression levels. Create permuted labels by doing the commands:

```
labels < − c(rep(“CEU”, 8), rep(“YRI”, 8))
plabels < − permute(labels)
```

By doing the above, you just created “random” CEU and YRI populations. Based on plabels, redefine the CEU and YRI groups, and repeat (e) and (f).

Submit the commands you used and the figures. Based on the comparison with the results from part (e) and (f), what is your conclusion?