

Gene Expression Analysis and Visualization  
410.671  
HW #1

For this homework, we will be working with a study from Gene Expression Omnibus (GEO) with the accession GDS2880. This is an Affymetrix microarray experiment (HG133A array). The data researchers were investigating patient matched normal and stage 1 or stage 2 clear cell renal cell carcinoma (cRCC) tumors to provide insight into the molecular pathogenesis of cRCC. We will be conducting outlier analysis using various methods to identify aberrant samples, followed by missing value imputation to assess the accuracy of two different algorithms.

- 1.) Download and load the renal cell carcinoma data file into R. Make sure that the row names are in the correct location (Affymetrix fragment names). Look at the dimensions and verify that you have 22 arrays and 22,283 probesets. (2pts.)
- 2.) Label the header columns of your data frame maintaining the GSM ID, but adding the Normal/Tumor identity. (2pts.)
- 3.) Identify any outlier samples using the following visual plots:
- 4.) Correlation plot (heat map) (2pts.)  
Hierarchical clustering dendrogram (2pts.)  
CV vs. mean plot (2pts.)  
Average correlation plot (2pts.)  
For all plots, make sure you label the points appropriately, title plots, and label axes. You will also need to provide a legend for the correlation plot. You can use the `gplots` for a color gradient, or just use the default colors.
- 6.) Install and load the `impute` library. (1pt.)
- 7.) Remove the outlier samples you identified in the first part of this assignment. (2pts.)
- 8.) Now we are going to use a couple of transcripts that were determined in this study to be indicative of normal renal function. The genes we will assess are kininogen 1 (KNG1) and aquaporin 2 (AQP2). Using either NetAffx or Gene Cards websites (or other resources, if you like), extract the probesets for these two genes. Hint: KNG1 has two while AQP2 has one. Then plot a profile plot (expression intensity vs. samples) for each probeset for these two genes. You may have to convert the data frame row to a vector to plot it. Do the plots of these genes seem to indicate normal renal function? Explain. (6pts.)
- 9.) We want to assess the accuracy of missing value imputation. So assign the KNG1 probeset (206054\_at) an NA value, only for array GSM146784. Be sure to first save the original value before replacing it with an NA. Also cast the data frame to a matrix to run this function. (2pts.)
- 10.) Now estimate the missing values in the array using 6 nearest neighbors and Euclidean distance with the `impute.knn()` function. (2pts.)
- 11.) Look at the value that was imputed for your gene and calculate the relative error of this value using the actual value that you saved. (2pts.)
- 12.) Now impute the missing values using the SVD imputation method. This is in the `pcaMethods` package and the function is called `pca()` with method `svdImpute` and set `nPcs=9`. To retrieve the output matrix, see the help file. (2pts.)
- 13.) Finally, plot a gene profile plot of the probeset for this gene, where the two different imputed values are represented as different colored points and the actual value is a third point. (6pts.)

Generate the code and plots for each. Turn in the visuals, code, and an explanation of the questions asked. Paste all information into a PDF doc.