PB HLTH C240C / STAT C245C

Assignment #1

Due: Tuesday, September 30th, 11:59pm

The purpose of this assignment is to get acquainted with basic and practical, yet crucial, statistical issues related to the analysis of high-throughput genomic data, by performing exploratory data analysis (EDA) and quality assessment/control (QA/QC) on microarray data from Schmidt *et al.* (2008).

Schmidt *et al.* (2008) used Affymetrix oligonucleotide HG-U133A arrays to measure the expression of 22,283 human genes in 200 tumors of patients with breast cancer. The data were pre-processed by MAS 5.0 software, using global-scaling normalization. The log-scale normalized expression levels comprise our dataset.

The dataset used by Schmidt *et al.* (2008) is available as a Bioconductor package at `http://www.bioconductor.org/packages/release/data/experiment/html/breastCancerMAINZ.html`.

Question 1. Getting Started. Read the article by Schmidt *et al.* (2008) for background on the biological question, details on the experimental design and pre-processing of the microarray gene expression levels. Read in and examine the dataset in R.

Question 2. Microarray-level EDA.

a) **Marginal distribution of gene expression measures for each microarray.** Provide numerical and graphical summaries of the marginal distribution of the gene expression measures for an arbitrary subset of 20 of the 200 microarrays, e.g., boxplots, density plots. Note that in this question microarrays are treated as variables and genes as observations.

b) **Bivariate distribution of gene expression measures for pairs of microarrays.** Provide numerical and graphical summaries of the bivariate distribution of the gene expression measures for pairs of microarrays, e.g., correlation, smoothed mean-difference scatterplots.

Question 3. Gene-subset-level EDA. Provide numerical and graphical summaries of the joint distribution of an arbitrary subset of 50 genes in the 20 arrays selected in question 2. Note that in this question genes are treated as variables and microarrays as observations.

Question 4. Gene-level EDA. Consider the gene with Affymetrix ID "203290_at." Provide graphical summaries that could reveal a possible role of this gene in breast cancer, e.g. the gene being differentially expressed between different subsets of tumors.

Question 5. Gene annotation. Consider the gene of question 4. Use the Bioconductor package *hgu133a.db* to provide its gene symbol, description and chromosome location. Verify that the information corresponds to that present in the `featureData` slot of the object `mainz`.

Question 6. (Bonus) S4 Methods.

Implement the method `boxplot` for the S4 class *ExpressionSet*. This method should retrieve the expression values of the experiments and draw a boxplot of the univariate distribution of each microarray. By default, the method should plot the distribution of all microarrays in the study, but optionally, it can display only a random subset of k microarrays, where k is chosen by the user.

References

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, **68**(13), 5405–5413.