Homework 2
BIOS 7659/CPBS 7659
Due 10/1/2020, 9AM

Please use one of the reproducible research templates (e.g., R Markdown, knitr, sweave, etc) and return your homework in report format.

1. Quality Control

   - Install the `affy` and `simpleaffy` packages from Bioconductor.
   - Download the `celfiles` directory from Canvas.
   - These data are from human cell lines treated with low or high levels of a treatment. The meta-data is provided in `targets.txt` in the `celfiles` directory.

   (a) Read the 8 CEL files in the directory `celfiles` using the following functions from the `affy` package:

   ```
   >pd = read.AnnotatedDataFrame("targets.txt", header=TRUE, row.names=1,
   as.is=TRUE)
   >Data = ReadAffy(filenames=pData(pd)$FileName, phenoData=pd, sampleNames=
   sampleNames(pd))
   ```

   Practice extracting information from `Data` and the sample file information `pd` using the following functions from the `Affy` package:

   `exprs(), sampleNames(), probeNames(), mm(), pm(), pData()`

   Hint: Use `head()` to only display the first few entries of your objects.

   (b) Plot the raw microarray images using `image()` on the object `Data`. Comment on what you see in these plots.

   (c) Plot quality control metrics using `qc.affy()` and `plot.qc.stats()` from the `simpleaffy` package. Comment on what you see in the plot. A description of `simpleaffy` is in Wilson & Miller (2005):

   `http://bioinformatics.oxfordjournals.org/content/21/18/3683.full`

   (d) Plot the mean intensity from 3' to 5' end of the target mRNA using `AffyRNAdeg()` and `plotAffyRNAdeg()`. Comment on what you see in the plot.

   (e) Use `boxplot()` and `plotDensity.AffyBatch()` to examine the distribution of intensity values for the perfect-match and mis-match probes separately. What patterns do you see? (Make sure you are plotting the `log` transformed data)

   (f) Based on the summaries and figures you generated, would you recommend that one or more chips be removed from the analysis?

2. Normalization

- Use the data set from problem 1 and the articles from Journal Club 2.

(a) Create log transformed data and plot the density before and after log transforming using `plotDensity.AffyBatch`. Comment on these plots.

(b) Plot MA plots using `MAplot()` and summarize your observations.

(c) Using `expresso()`, try different `normalize.method` options (`quantiles, loess, constant`), keeping the arguments `summary.method, bgcorrect.method, pmcorrect.method`) set to one selection. Plot boxplots of the un-normalized and normalized intensities for the different normalization methods.

Then, try different `summary.method` options (`avgdiff, mas, medianpolish`), keeping the other arguments set.

Finally, try different `pmcorrect.method` options (`pmonly, mas, subtractmm`), keeping the other arguments set.

Describe the options you set and compare and contrast the boxplots for the different normalization and summary methods.

(d) Get present and absent calls Mas 5.0 using `mas5calls()`. How many probesets have at least one present call in each of the two groups? Use this as your filter for part e).

(e) If you think there are any problematic chip(s) based on your quality control analysis in both problems 1 and 2, remove them. Justify why you removed or did not remove any chip(s). Normalize (using `rma()`) and filter the remaining chips. Save files of the intensities and calls locally. Show the top few lines of the intensities and calls in your homework report.