In-Class Lab – HSAvsNHP (Bioconductor Use Case Example)

Data Set: Two files (data.vsn.csv and annt.txt) on Sakai for a subset of the primate fibroblast gene expression from Karaman et al., Genome Research 2003. This study examines 3 groups, human, bonobo and gorilla expression profiles on Affymetrix HG U95Av2 chips. Note the full dataset contains 46 chips and is available in the Bioconducor library fibroEset
Script: LabWk4NHP.R (on Sakai)

Instructions: The R script is labeled in the Comments (Letters A-I)
Sections A-C: Logistics to load data and set up environment. Note the annotation file contains the cel filenames (Cels), shorter names for the arrays (short.names), information about the Donor (Gorilla, Bonobo, Human), Age (years), Gender (male/female), doubling time (DT) of the cell lines, and information about whether cells where established from the same cell lines (estb.same).

Section D: We will set this up a simple two-class problem to distinguish human from non-human primates (NHP). We will therefore set up our annotation for Human and NHP as two classes.

Section E: We next retrieve the gene annotation for the array.
To Do: **Before you start the next session, you should perform EDA to examine distributional assumptions, identify outliers etc.**
This is not provided for you and you should write this code.

Section F: We will use the Limma framework to set the design matrix and fit the gene-wise linear models. We will use topTable to select a subset of genes based on differential gene expression.
Question: How many genes met the p-value threshold that was set in the code?

Section G: Examine the heatmap, QQ and volcano plot.
Question: What is your assessment of these plots?

Section H: We will next use the RankProd library. This is a non-parametric method for identifying differentially expressed (up- or down- regulated) genes based on the estimated percentage of false predictions (pfp). Note: this method can combine data sets from different origins (meta-analysis) to increase the power of the identification (See the function RPadvance).

Section I: You will now load the full data set and re-run your analyses.
To Do: Examine each of the above genesets from Rank Products and Limma in the complete dataset. Generate the plots for each set.
Question: What is the overlap between Limma and RankProd? **How would this impact your assessment for feature selection?**