GS01 0163 Assignment 2 Fall 2004

1. Download the the data files from the course website. These should contain: the red and green binary image files from a single array and the corresponding quantification file for that array, red and green quantification files from a second array of a different type, and some of the R scripts that I used to assemble the notes for this week. The one thing that I know needs to be altered about the script is that currently it looks for files in a specific directory, and this will need to be altered depending on where you set things up.

   The R functions that you will need to learn about this week include `read.table`, `read.delim`, `names`, `dim`, `ceiling`, and `floor`. The actual assignment part is to read through the script lecture3.R and see if it makes sense. No written answer is required here.

2. Following the outline in the script provided, load in the binary files imageA.bin and imageB.bin, and the associated quantification file imagesAandBquant.txt. Compute the spot volume values in the four corners of the array (A - 1 : a - 1, A - 12 : a - 10, D - 1 : h - 1, and D - 12 : h - 10) for both the green (image A) and red (image B) channels.

3. Using the quantification file, arrange the spot background values from the A image into a 40 by 120 matrix (matching the geometry of the array) and produce an image plot of the result. The background value for A - 1 : a - 1 should be in the upper left, and that for D - 12 : h - 10 in the lower right. Sample rearrangement into matrix form is shown in the script.

4. We now shift to data from the quantification files of a second array: hA223-1_532.txt and hA223-1_635.txt (the wavelengths of green and red light used in the lasers are 532nm and 635nm respectively, giving the suffixes above). The array has 19200 spots, with a 12x4 subgrid layout and a 20x20 spot layout within subgrids. Rows are stored based on a unique identifier of the form w-x-y-z, where w = 1,...,12; x = 1,...,4; y = 1,...,20, and z = 1,...,20. The fun thing here is that all of the genes were printed in duplicate on the array, so we can check the agreement between replicates. The replication pattern is w-1-y-z == w-2-y-z, and w-3-y-z == w-4-y-z: the spots in grid column 2 are the same as those in grid column 1, and the spots in grid column 4 are the same as those in grid column 3. We want to produce 6 plots here. The first three are an M-A plot of log2(replicate 1 signal mean) versus log2(replicate 2 signal mean) for the green channel, for the red channel, and for the log ratios. The next three are the same, but replacing "signal mean" with "signal mean - background mean" and replacing all values less than 10 with 10 before taking logs (thresholding). Is there evidence that using log ratios is improving on using data from just a single channel? (side note – there are quite a few header rows in these files that will need to be skipped to let R load them nicely! The skip option in read.table or read.delim should work).