

Begin with the sheet called “Sample Data” which contains all the data for 16 of the 22,000 spots on the tested array. This sheet will familiarize you with the headings and entries you can expect to see when you examine your own data. Some important landmarks are

- **Columns C and D** the address for each gene on the array
- **Column K** the gene name associated with each spot
- **Column M** some useful information about each gene
- **Column E** some of genes do not have their systematic names listed, however their NCBI accession numbers can be found in column E. The NCBI website (<http://www.ncbi.nlm.nih.gov/>) will be useful for these spots. Be sure to search “nucleotide” when you use these accession numbers at NCBI.

Please answer the following questions about this data:

1. Where on the array is the spot for the PRKCL2 gene (column and row)?
2. What gene is April 3rd? What is its cellular function?
3. What is Pro25G? (hint: google with “Agilent” to find some information).

To look at the signal intensities for these 16 spots, you should scroll right. **Column AG** reports the signal from the green fluorescent molecule Cy3 and **Column AH** reports the signal from the red fluorescent molecule Cy5. Use the “Format” menu to convert the values in these columns from scientific notation to numbers with two decimal places, and then answer the following questions.

4. What are the green and red signal intensities for PRKCL2?
5. What genes give the highest and lowest values in **Column AG**? Are these the same genes that give the highest and lowest values in **Column AH**?
6. What percentage of the genes in this set has a larger value in **Column AG** than in **Column AH**?
7. Using the values in **Column AG**, find the mean and the median value for the three Pro25G signals. What does the agreement of the mean and median tell you about the three values?
8. Find the mean and the median for all the values in **Column AG**. What is the significance that these values are not identical?

Now you are ready to look at a bigger data set and practice some analytical methods. Look at the second sheet called “Test Array” in the Excel file. This sheet has a subset of the data (9 of the 61 columns) for a subset of the spots (1,853 of the 22,000) from a single microarray experiment.

Some of the data analysis you will perform is

- **Normalization** to correct for the physical and chemical differences in Cy3 and Cy5
- **Background subtraction** to correct for signal intensity in areas of the array that do not have DNA spots, and
- **Log₂ Transformations** to avoid fractions when expressing signal ratios

You will begin by “normalizing” the data. Many normalization methods have been suggested since microarray technology was introduced. We will practice a “global normalization” method that assumes the Cy3 and Cy5 fluorescent intensities differ by a constant factor,

$$R = kG$$

where R = red (Cy5) and G = green (Cy3)

One way to determine k is to label the same RNA sample with either Cy3 or Cy5 and then compare the mean signal intensities observed on an array. Since microarray experiments are expensive to perform, this direct comparison is not often done. Instead it is assumed that arrays have the same amount of total mRNA for two samples and the difference in overall intensity is k .

9. Use the mean signal intensities (data in **Columns B** and **C**) from the Test Array to calculate the average intensity for the green and red signals. What is k ?
10. Now use the median signal intensity (data in **Columns G** and **H**) to calculate k . Is there a difference when you calculate k using the mean and the median signal intensities?

Because of microarrays are physically small, signal artifacts routinely arise. These artifacts come from tiny droplets with fluorescent molecules that remain on the array, and from scratches on the surface of the slide. Even the light that leaks into some scanners can make parts of the array appear more green or more red. The column headings in your spreadsheet that include “BG” have background measurements and these values can be used to correct the signal intensities for background artifacts.

11. Determine the average red and green background signals. Do this for **Column D** and **E** (the mean signals) as well as for **Column I** and **J** (the median signals).
12. Do the differences in the average background signal mirror the differences in the signal itself (**Columns B** and **C** vs **D** and **E** for example)?
13. Find one green background measurement that is considerably different from the average. Is the red background measurement also different? How could you explain this?
14. Insert two new columns after the background signal columns and calculate the “background corrected” values for the green and red signals. These corrected values are determined by subtracting the background measurement for each spot from the signal measurement.

So far you’ve seen that microarray data must be normalized to correct for Cy3 and Cy5 differences as well as “background subtracted” to correct for artifacts on the slide. Recall that microarray experiments are designed to simultaneously compare the expression of many genes in two samples. The corrected intensities can be expressed as a ratio between

the corrected signals for the two samples (Green/Red). A ratio of 4 means 4-fold gene induction and a ratio of 0.25 means four-fold repression of that gene.

To avoid the decimals associated with gene repression, the \log_2 of the ratios is useful.

Four-fold induction is reported at $\log_2(4) =$ the power of 2 needed to get $4 = 2^2$.

Four-fold repression is reported as $\log_2(0.25) =$ the power of 2 needed to get $1/4 = 2^{-2}$. $\log_2(1) - \log_2(4) = -2$. \log_2 transformed data makes more sense graphically since a 4-fold induction and a 4-fold repression have the same value but different signs (i.e. +2 and -2).

15. Add another column to the Test Array called "Net Green/Red" and calculate the ratio of the background-corrected green signal to the background-corrected red signal. What is the average value for the column?
16. Add another column to the Test Array sheet called "Log₂ Green/Red" and transform the "Net Green/Red" data to \log_2 values. What is the average of this column? Draw a histogram that plots these values. Sort the data. Which 5 genes in this data set are most strongly induced and which are most strongly repressed?