Quality and Statistical Analysis of AFGC Microarray Data

Finkelstein, D.B.[1], Sterky, F.1, Gollub, J.[1], Keegstra, K.[2], Miura, E.[1], Simon, V.[2], Somerville, S.[1], Cherry, J.M[1]. [1]*Department of Plant Biology, Carnegie Institute of Washington,*[2]*Plant Research Laboratory, Michigan State University*

Arabidopsis Functional Genomics Consortium

Abstract

The widespread use of microarray data depends on its reproducibility. The sources or variability in data derive from biological and technical sources. Biological variability is experiment dependent while technical variability is process dependent. This poster describes two large scale repetitive experiments that are designed to quantify technical variation of microarrays.

Site, Log_2Ratio and Rank Order Variability

Two sites perform microarray hybridizations for customers of the Arabidopsis Functional Genomics Consortium (AFGC) customers: one at Michigan State University (MSU), Plant Biology Laboratory and another at the Carnegie Institute of Washington (CIW), Department of Plant Biology. In order to verify the reproducibility of our methods at each site performed the same hybridization using 2 common pools of RNA experiment 6 times. Each site used the same 11 K element array. The complete data from these 12 experiments, as is true for all AFGC experiments, is public -- see http://afgc.stanford.edu for more information or email finkel@genome.stanford.edu. Graphical analysis is presented and statistical analysis is ongoing. In summary the variability of MSU was less than that of CIW and the variability within each site is less than the variability across sites. Most of this difference was due to a single array. Also, in general, rank order was more variable than log_2ratio. In all 12 cases the cy5 labeled cDNA was derived from cell culture while the cy3 labeled cDNA was derived from leaves.

Spot and Printing Variation

Also presented here are the experimental design and preliminary results of a statistical replication experiment. This factorial analysis was designed to test the reliability of ratios across a given slide, between slides printed in a given lot. Eight cDNA clones and salmon sperm DNA were spotted 36 times by each of 16 pins for a total of 576 replicate spots per array. Ultimately 60 slides will be tested, 30 each on polyamine slides and polylysine slides. Each slide is probed with the same cDNA derived from Arabidopsis RNA. Both the cy3 and cy5 dye labeled probes are identical in this case so that the expected expression ratio is always 1.

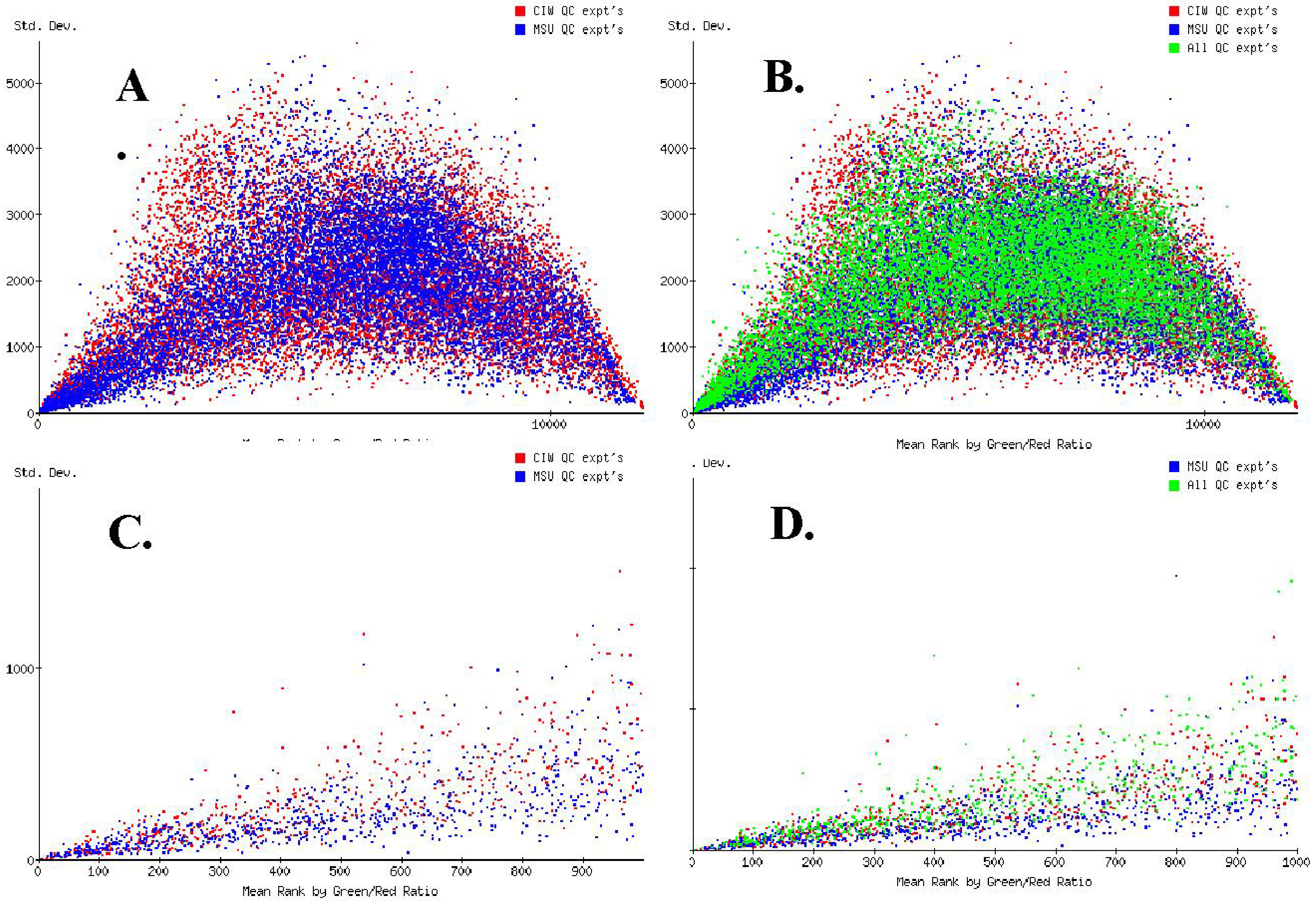

Figure 1. Variability of mean-rank order ratios at two sites: MSU (blue) CIW (red)

For each spot a log_2(channel2/channel1) was calculated and ranked from highest to lowest. The mean of this rank was then calculated as was the standard deviation of the mean rank. The graphs presented above are plots of each of the 11 K spots by mean rank (x axis) and std deviation of that mean rank (y axis). Blue spots are the average of 6 experiments performed at MSU, red spots are 6 Carnegie experiments, and green are the average of all 12 experiments. Figure 1A. compares Carnegie to MSU. Figure 2A. Includes the global averages in green. Figure1C. Is a close up of the most highly ranked spots which are also the least variant in rank order. Figure1D. Is a close up of the most highly ranked spots including the global average. In general, the variability of rank order is lowest at the two extremes, as expected. Statistical analysis of the differences between sites is ongoing, final results will be presented on our website http://afgc.stanford.edu. However, preliminary results indicate that the variability of rank within each site is significantly less than the variability of ranks between the 2 sites for the first 1000 most highly ranked spots.

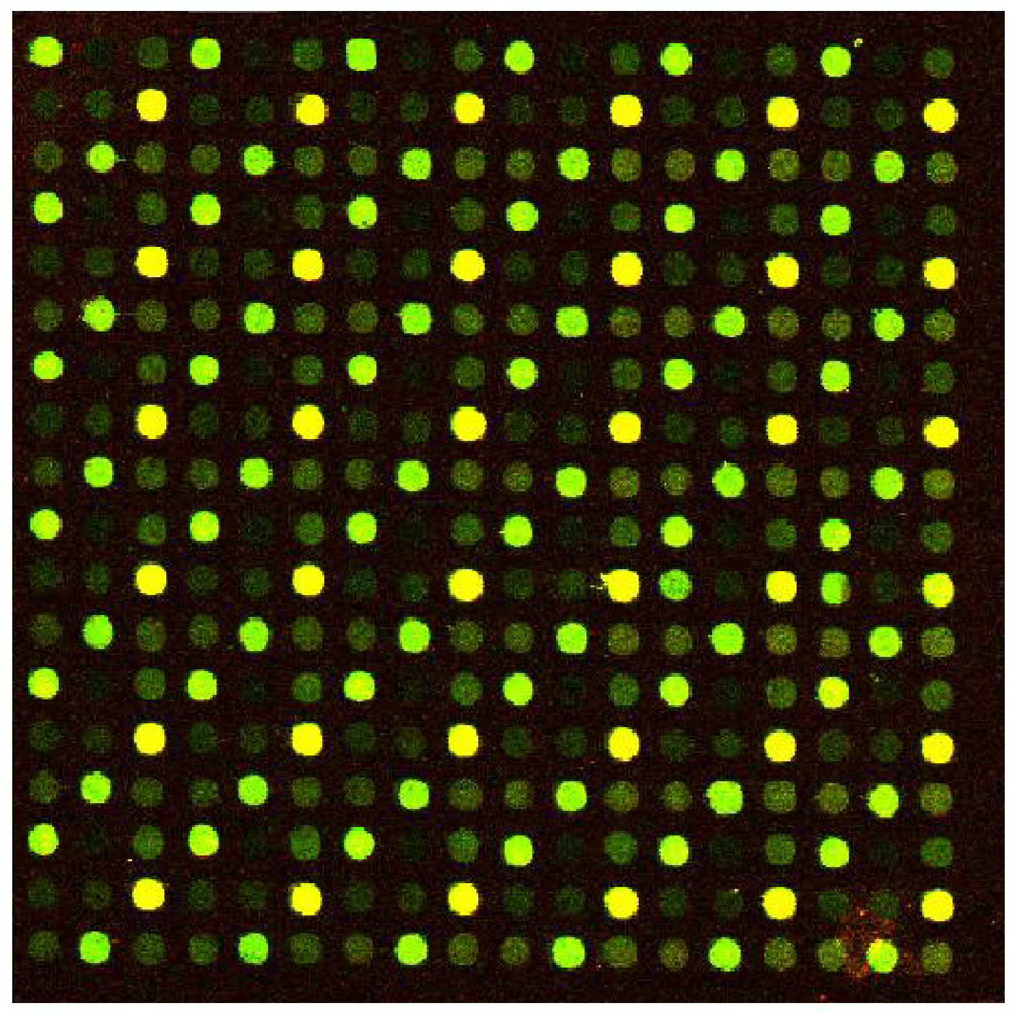

Figure 3. Statistical grid

This figure shows the 9 spot pattern repeated 18 times by a single printing pin. Note consistent intensities for repeated spots.

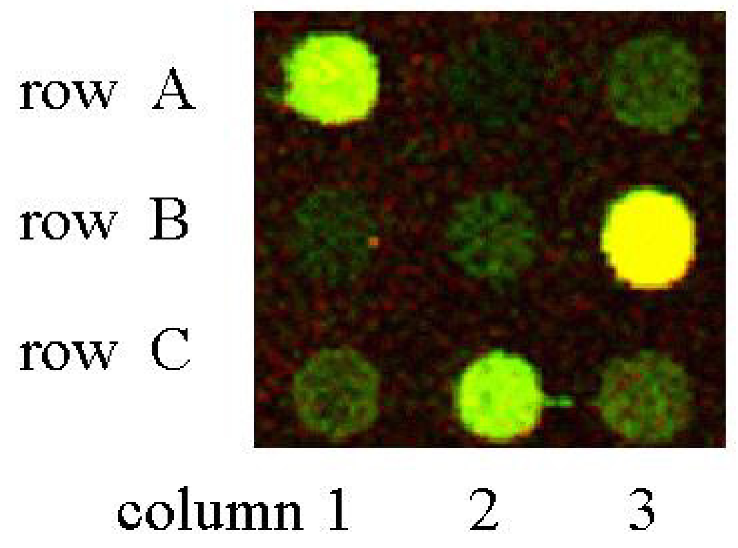

Figure 4. Spot identity

	column 1	2	3
rowA	ATPase subunit	triose translocator	heat shock transcription factor
rowB	AP2	salmon sperm DNA	Chlorophyll a/b binding
rowC	IAA11	Rubisco	Glutathione sulfur transferase

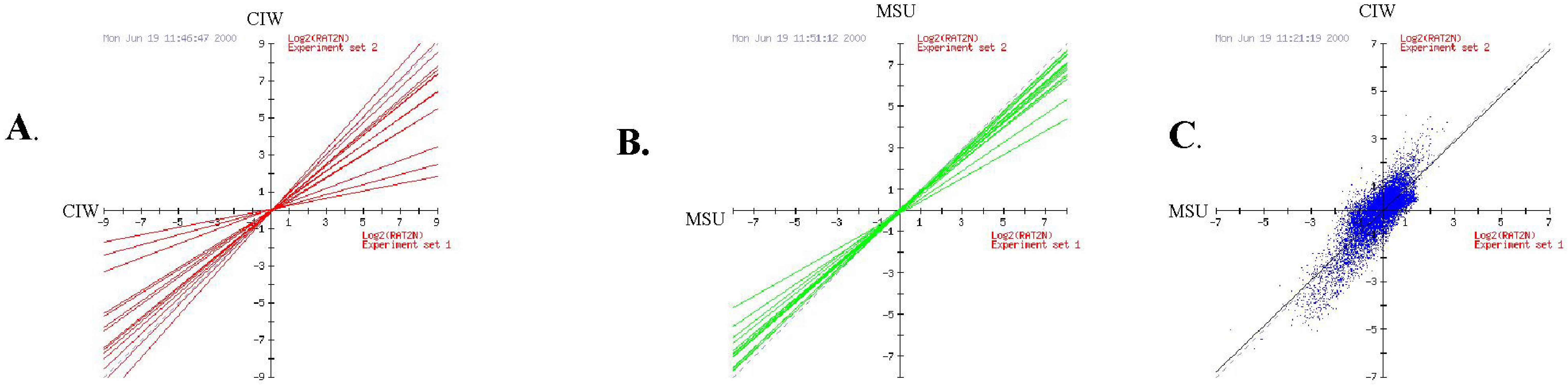

Figure 2. Linear regressions of all possible pairwise comparisons of experiments within each site and the mean log_2ratio values from MSU regressed against the mean log_2ratio values CIW.

Figure 2A. Each red line is a linear regression of one replicate experiment performed at the Carnegie against another. Ideally they should perfectly correlate around the y=x identity line. All 15 possible pairwise combinations are shown. Most of the variability in this plot is due to the influence on a single outlier experiment. Figure2B. These 15 pairwise regression lines were derived by comparing data from 2 MSU experiments. Figur2C. Is the regression of the mean of 6 log_2ratios from MSU experiments regressed against the mean log_2ratios of 6 CIW experiments. In this case the correlation of regression is 0.825 the slope is 0.966 and the intercept is -0.004. This means that results of CIW and MSU were nearly identical when viewed globally.

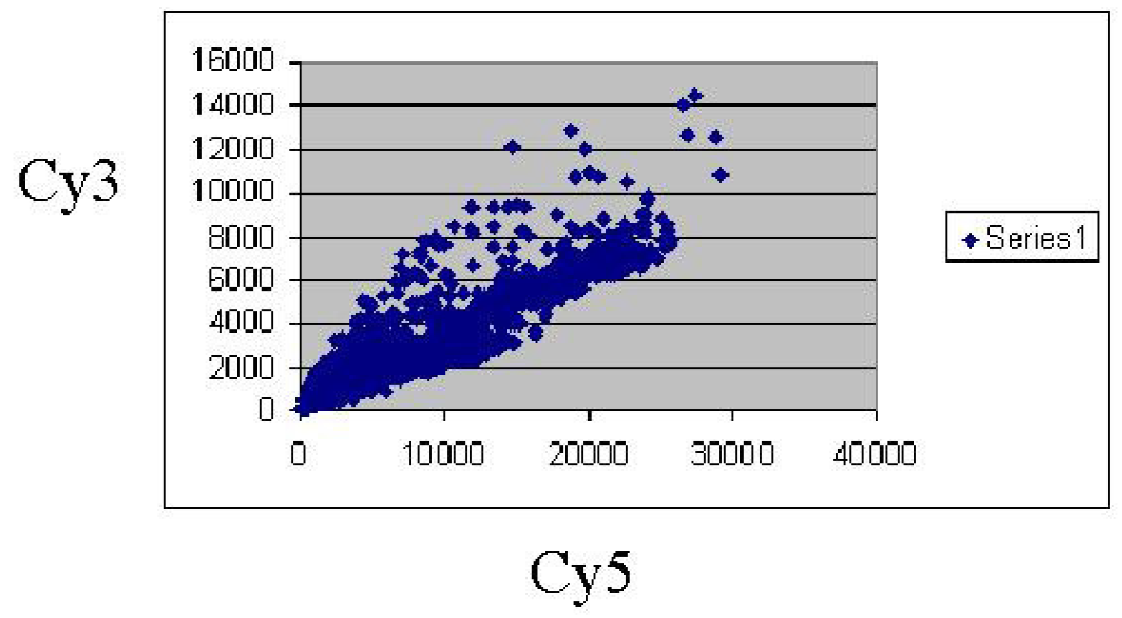

Figure 5. Scatter plot of replicate array data

For the linear regression model Cy3 = 2.8(Cy5)-298 the R^2 = 0.86

F stat= 1,892 (significant at 0.975 level). Note that most of the scatter is above the line. This is due to greater background in Cy3 than in Cy5.

This research was funded by the National Science Foundation. For more information and updates on this poster visit our website at: http://www.afgc.stanford.edu/