Example of a Statistical Analysis

James W. MacDonald

August 11, 2007

This analysis is based on microarrays that were processed in the microarray facility in August 2004. Filenames and samples were as follows:

	Filenames	Samples
1	sample01.CEL	A
2	sample02.CEL	A
3	sample03.CEL	A
4	sample04.CEL	B
5	sample05.CEL	B
6	sample06.CEL	B
7	sample07.CEL	C
8	sample08.CEL	C
9	sample09.CEL	C
10	sample10.CEL	D
11	sample11.CEL	D
12	sample12.CEL	D

The goal of this analysis is to see if there is a difference between the A and B samples, as well as between the C and D samples.

The first step in my analysis was to make some quality control plots that can be used to check the overall quality of the raw data.

Figure 1 shows the distribution of the PM probes for each chip. One of the underlying assumptions for the normalization procedure I use is that the PM probe data all come from the same distribution, with the only differences being the location and scale. Basically, this means that we want the shape of the curves to be very similar, and we want the curves to be fairly close to each other. There appears to be a large difference between the A/B and C/D samples that may have an impact on our analysis. Since we are only concerned with the comparison of A/B and C/D it might make sense to pre-process these samples separately.

Figure 2 is designed to show differences between samples due to mRNA degradation or from the *in vitro* translation step. Any differences between samples will be indicated by a different slope. Again, the only differences are between

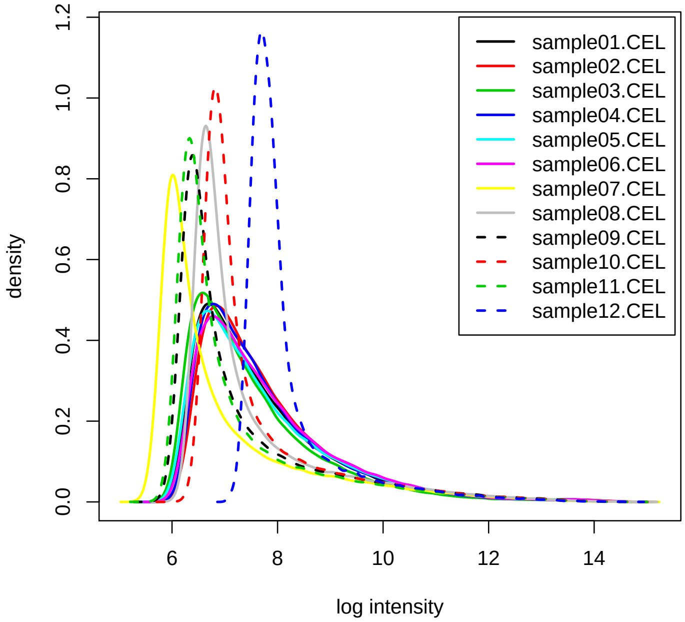

Figure 1: Plot of perfect match (PM) chip densities

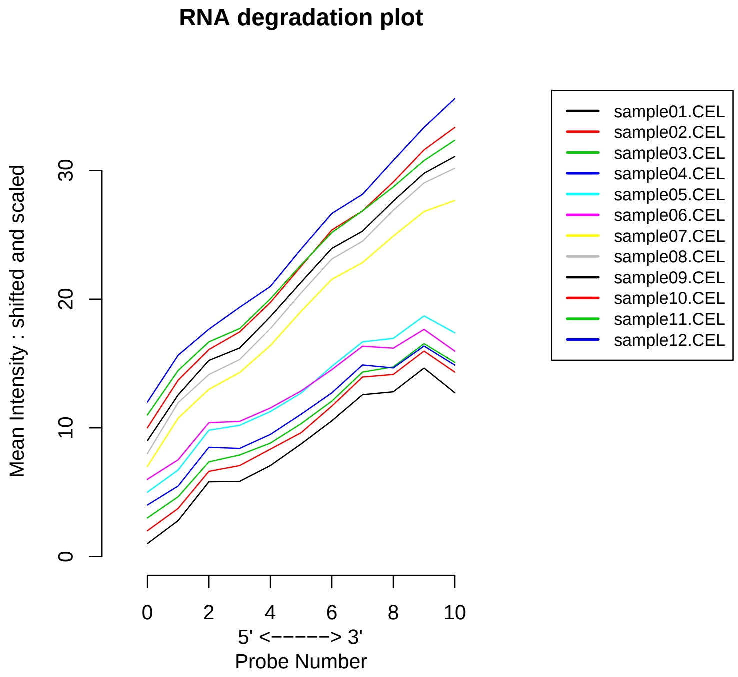

Figure 2: RNA degradation plot

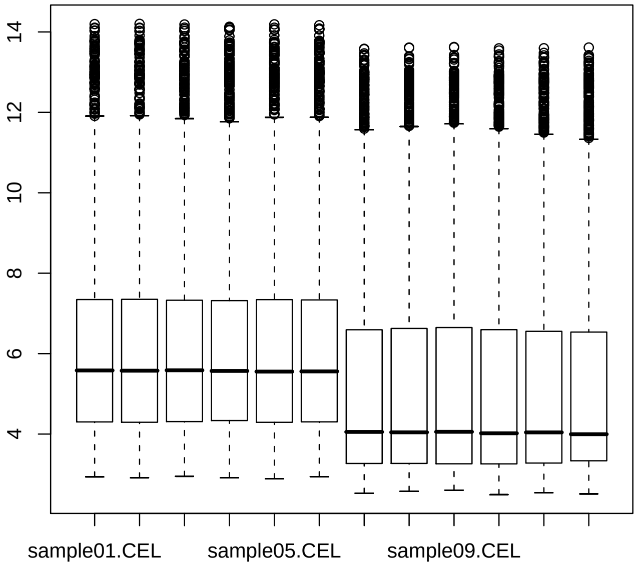

Figure 3: Boxplot of Expression values from both sets

the two sample sets. These two plots indicate that we should probably process each set of samples separately and then combine later.

I calculated expression values for each gene using a robust multi-array average (RMA) Irizarry et al. (2003). This is a modeling strategy that converts the PM probe values into an expression value for each gene. Note that the expression values are log_2 transformed data. These data can be converted to the natural scale by exponentiating (e.g., convert by using 2^x, where x is the expression value). Figure 3 shows a boxplot of the expression values for each set of data. It appears here that the data are fairly well normalized.

I then fit a principal components analysis (PCA) on the expression values and then plotted the first two principal components (PCs). PCA can be used to visualize the overall structure of high dimensional data; in this case, we are using it to see if the replicated samples are grouping together, which would indicate that the replicates are relatively similar in their gene expression profiles.

Figure 4 shows the PCA plot. Here again we can see that there is a very large difference between the A/B and C/D samples.

The PCA plot indicates that the replicated samples are quite similar, but doesn't tell us much about the quality of the RMA model fit. For this we can do boxplots of the normalized unscaled standard errors (NUSE) from the model fit.

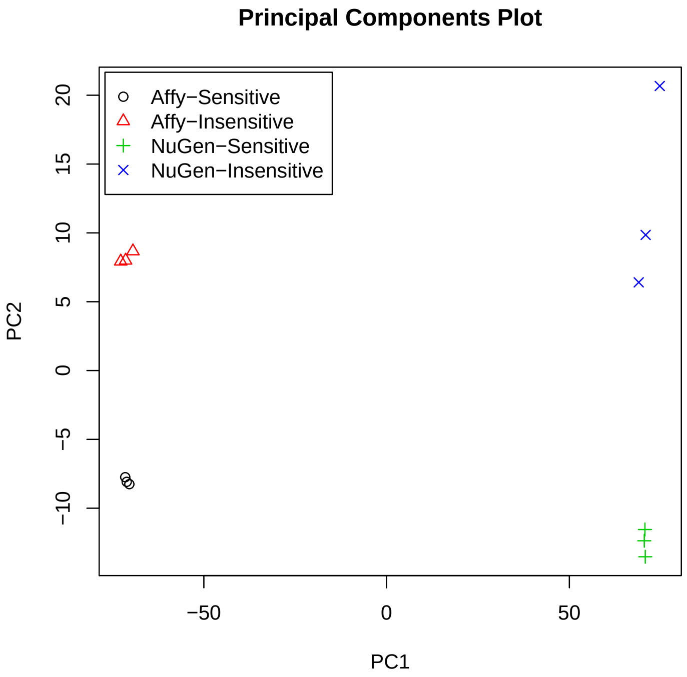

Figure 4: PCA plot

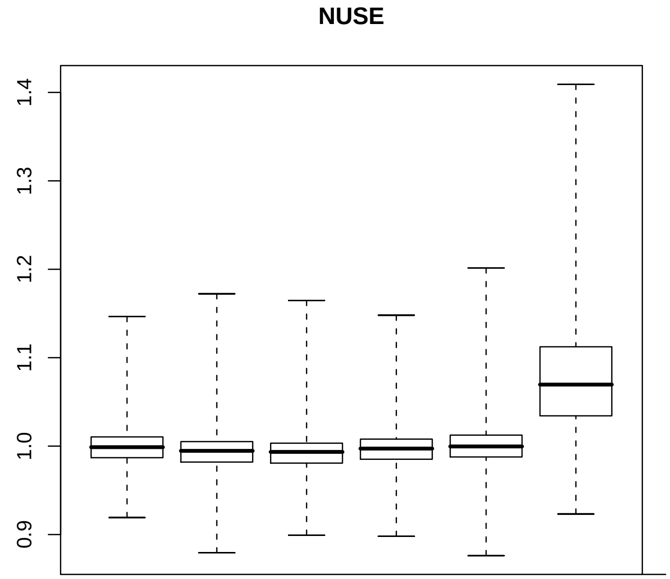

Figure 5: NUSE plot

Any chip that has large standard errors in comparison to the others is probably of lower quality, as the model isn't fitting the data from that chip very well.

Figure 6 shows the NUSE plot for the first six chips. The standard errors are all quite similar for each chip, indicating that the model fits the data similarly on each chip.

One last QA plot that might be of interest is a relative log expression (RLE) plot. For this plot we compute the relative log expression for each probeset on each chip, relative to the median value for that probeset. Any chip that is very different from the others in this plot is typically of lower quality.

```
connected to:  ensembl
Reading database configuration of: hsapiens_gene_ensembl
Checking attributes and filters ... ok
Checking main tables ... ok
Error: ' affy 'is not an available annotation source for this biomaRt or this species.
Available choices are listed below:
```

Prior to making comparisons, I filtered the genes to remove any that do not appear to be differentially expressed in any samples, based on at least samples having an expression value of or greater. This resulted in a total of 4195 genes.

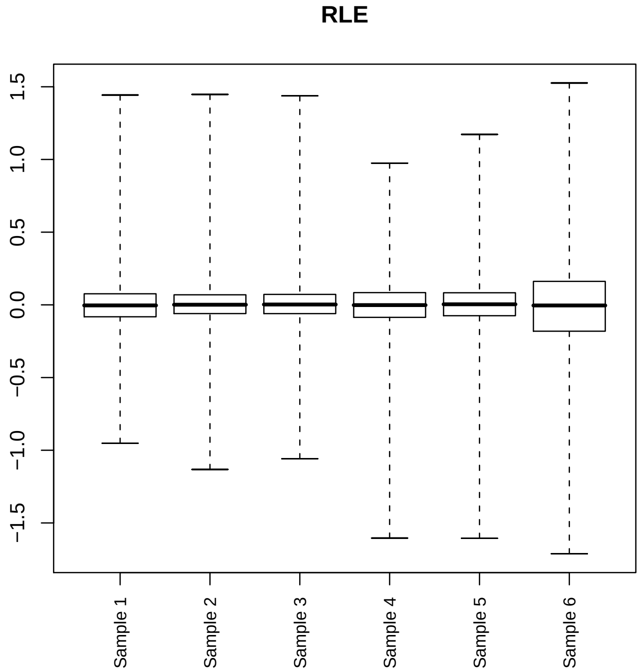

Figure 6: NUSE plot

I then made the requested comparisons using a modeling strategy developed for microarray analyses (Smyth (2004)), selecting those probesets with an adjusted p-value less than 0.05 and a two-fold difference (adjusting for multiple comparisons using false discovery rate (Benjamini and Hochberg (1995))). This resulted in the following number of probesets:

	Comparisons	Probesets
1	A vs B	52
2	C vs D	73

I output these data in HTML and text tables that include various statistics, as well as annotation for the different probesets. I also output all the expression values in a text table that can be opened using a spreadsheet and used for ongoing analyses, or to look for probesets that might not appear in the HTML tables.

Please note that I used the affy, and limma packages of Bioconductor for this analysis. If you publish these results, please use the following citations.

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

Rafael A. Irizarry, Bridget Hobbs, Francois Collin, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–64, 2003.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 3, 2004.