

Quality Assessment of Exon Arrays

I. Introduction

Quality assessment can play an important role at many stages in the process of generating gene expression data for biological samples, from array design to sample processing to signal estimation and analysis. In this report we describe some quality assessment procedures that are based on CEL intensity data from the GeneChip® Human Exon 1.0 ST Array. The methods detailed here are described in Chapter 3 of the Bioconductor monograph [1]. The quality assessment procedures we consider entail computing some summary statistics for each array in a comparable set and comparing the level of the summary statistics across the arrays. We therefore assume that the user has a set of comparable arrays of the sort that would normally be analyzed together to address substantive biological questions. The quality assessment part of the procedure entails identifying outliers within the set. These could be flagged and excluded from the substantive biological analysis, or the downstream analysis could be adjusted to account of the outlier arrays, by weighting for example.

The procedures we describe can identify outliers, but it does not provide hard and fast rules (specific thresholds) as to which arrays to flag as outliers. These need to be developed in the context of particular applications with specific types of samples, with knowledge of the costs involved in repeating experiments and the costs of drawing wrong conclusions. The procedures also do not provide guidelines to assess the overall quality of a set of arrays. These can be arrived at by collecting data for many sets of arrays and using them to set target levels and variance limits for the various quality assessment measures on a user-by-user basis. Lastly, while the methods presented here can be used to identify outlier arrays, they will not indicate why the array is an outlier (ie input RNA quality problem, target prep problem, hybridization problem, etc...)

The rest of this report is organized as follows. Section 2 describes simple summaries of the intensity distributions and \log_2 ratios of cel intensities. Section 3 describes summaries of feature set model fit derivatives - signal estimates and residuals. In Section 4 we discuss the specific quality metrics reported in the Exon Array Computational Tool (ExACT) Quality Report and then end with some concluding remarks in Section 5.

II. Feature or cel level summaries

Note that the feature level graphs presented in this section are not directly obtainable from the Exon Array Computational Tool (ExACT). The “exact-probe-intensity-extraction” command line program can be used to extract probe level intensities for either the whole array or a subset of the features on the array. Normalized and un-normalized CEL files can be used with this tool. The

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

extracted probe intensities can then be evaluated with a statistical application such as R, S-Plus, or MatLab to generate the summary graphs.

II.A. Single array cel intensity distribution summaries

The simplest and most obvious thing to do with a set of cel files that have been collected for the purpose of addressing some questions of biological interest is to examine the distribution of cel intensities corresponding to each array in the set. As the array distribution of intensities is highly skewed, the \log_2 transformation helps to approximately normalize the distributions. Log base 2 is typically used as it facilitates the conversion back to the original scale: a difference of 1 on the log base 2 scale corresponds to a fold change of 2 on the original cel intensity scale. Array \log_2 cel intensity distributions can be summarized by a three point summary: 25th percentile (Q1), 50th percentile or median (Q2) and 75th percentile (Q3). To identify outlier arrays within a set, it helps to compare the distribution by means of boxplots. In the boxplot display, a box is formed with sides at the 25th and 75th percentiles of the distribution. A line is also drawn within the box at the level of the median. Whiskers are also drawn extending beyond each end of the box with points beyond the whiskers typically indicating outliers. As the purpose of the boxplots in this application is to compare the location and spread of the main body of the distribution of \log_2 cel intensities, we typically suppress the plotting of symbols for points beyond the whiskers. See [2] for guidelines regarding outlier detection.

Figure 1 shows an example set of boxplots of \log_2 cel intensities for a set of 20 arrays. The 20 arrays were hybridized with preparations made from RNA extracted from colon tissue samples. We observe a range of cel intensity distributions across the 20 arrays. A normalization transformation is typically applied to make the distribution of cel intensities more comparable. Based on cel intensity distributions alone it is difficult to assess the impact that observed differences will have on downstream analysis. It is recommended that cel intensity distributions be observed and recorded for future reference.

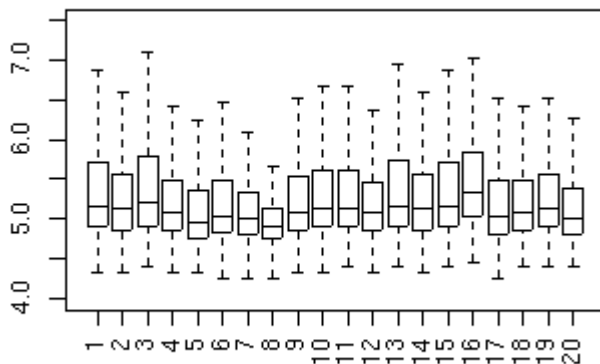


Figure 1: Boxplots of \log_2 cell intensities for a set of 20 arrays.

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

\log_2 intensity distributions can also be summarized by means of a histogram display. Whereas histograms provide more detail, but makes the comparison of several distributions more difficult. Histograms enable the detection of a secondary mode in the distribution corresponding to a bright spot on the array. Unless the array images are examined to identify locally bright or dim regions, it may be warranted to examine the histograms of the \log_2 cel intensity distributions for any evidence of bright spots. Figure 2 shows the histogram of \log_2 cel intensities for the first 8 arrays in the set of 20 displayed in Figure 1. As there is no evidence of bi-modality, these displays provide no additional information to what is captured by the boxplots in Figure 1.

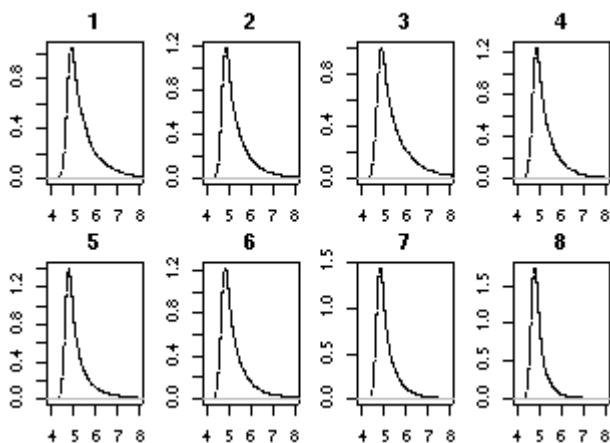


Figure 2: Histograms of \log_2 cell intensities for the first 8 arrays in the set of 20.

II.B. Multiple array feature level summaries

A significant characteristic of CEL intensity data is the large magnitude of feature specific effects. The typical range of \log_2 CEL intensities for arrays in Figure 1 is (lo, hi) (e.g. (2, 16)). The range of variability for any given feature, or CEL, across the set of arrays will be much lower – bright features tend to be bright across all arrays, and dim features tend to be dim across all arrays. The model used by PLIER (see PLIER Technote) to compute signal estimates, as well as the model used by RMA to compute feature set summaries, account for these feature effects. The MA-plot [3] provides a visual display that corrects for feature-specific effects and facilitates making comparative assessments of cell intensity distributions that are sensitive to the technical variability which we want to detect. Suppose that \mathbf{x} and \mathbf{y} are two vectors of \log_2 intensities corresponding to two arrays. The MA-plot is a scatter plot of $\mathbf{M}_i = (\mathbf{x}_i - \mathbf{y}_i)$ on the vertical axis vs. $\mathbf{A}_i = (\mathbf{x}_i + \mathbf{y}_i)/2$ on the horizontal axis. The \mathbf{M} values are relative \log_2 intensities, or \log_2 ratios of intensities. The median \mathbf{M} value gives a sense of relative shift in \log_2 intensity between two arrays, and the inter-quartile range (Q3-Q1) of \mathbf{M} values gives a sense of reproducibility of cell intensity values between arrays. The \mathbf{M} values also contain some real biological variability, but for practical

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

purposes that variability in M values provides a useful indicator to detect extra technical variability. The MA-plot as described above provides pair wise comparisons between any pair of arrays. When assessing quality in a set of arrays, pair wise comparisons produce much redundancy. To avoid this redundancy, each array can instead be compared to a synthetic array created by taking feature-wise medians across the set. For example application and interpretations of MA-plots see Dudoit et. al. [3].

The MA-plot captures the relationship between \log_2 ratio, or relative \log_2 intensity, on the vertical axis, and mean \log_2 intensity value on the horizontal axis. Figure 3 displays MA-plots for the first 8 chips in the set of 20 displayed in Figure 1. Note that the relationship is intensity dependent with both the center and the spread of the M values varying with \log_2 intensity. Because of the high and varying density of points in the MA-plots, it is difficult to visually detect subtle differences. Superimposing the scatter plot with smooth curves capturing the quartiles (median, lower and upper quartiles) in M values (relative \log_2 intensities) as a function of A values (median \log_2 intensities) is a helpful visual aid.

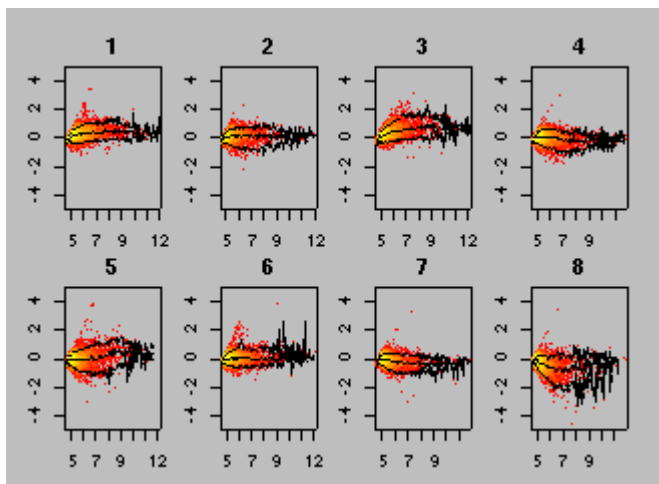


Figure 3: MA-plots for a set of 8 arrays: y-axis = M = relative \log_2 intensity = \log_2 intensity for array - median \log_2 intensity, x-axis = A = $(\log_2$ intensity for array + median \log_2 intensity)/2. Color indicates density of points. Black lines drawn at 5th, 50th, and 95th percentiles of local M value distributions.

Alternatively, boxplots of M values provide a summary of MA-plots which ignores the relationship between relative \log_2 intensity and mean \log_2 intensity but facilitates the comparative assessments of the distribution M values among many chips. Figure 4 shows boxplots of relative \log_2 intensity values for the 20 arrays whose \log_2 intensity values are displayed in Figure 1. The two figures, Figure 1

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

and Figure 4, deliver the same message, with latter being more sensitive to differences between arrays.

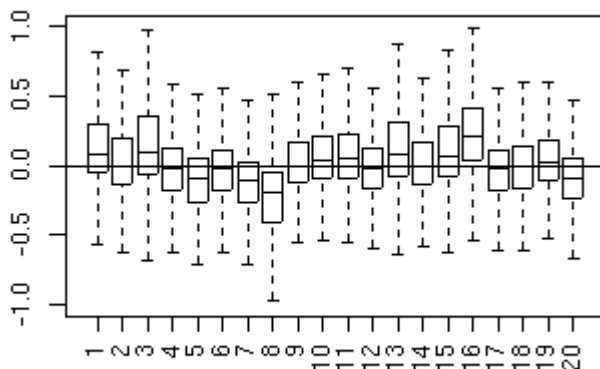


Figure 4: Boxplots of M values for a set of arrays: $M = \text{relative log}_2 \text{ intensity} = \text{relative log}_2 \text{ intensity} = \text{log}_2 \text{ intensity for array} - \text{median log}_2 \text{ intensity}$

At this point, a word on normalization is in order. Ideally, the distributions of log_2 cell intensities for a set of arrays to be analyzed together will be comparable – the three-point summary values (Q1, Q2, Q3) should be around the same level for all arrays, and the MA-plots should be flat, narrow and centered at zero. This ideal will rarely be achieved in practice due to the biological and technical variability that is an inherent characteristic of the cell intensity data being produced by a very elaborate process with multiple sources of variability. Normalization is the term used to refer to procedures that adjust cell intensity data in order to remove the technical variability without masking the biological variability. Some normalization methods will remove much of the variability that is detected by the quality assessment displays describe above. Assessing variability in the un-normalized data values is still important as the amount of adjustment that is required to normalize the cell intensities of an array is a good indicator of quality – the less adjustment required, the better the quality.

III. Summaries of feature set model derivatives

As mentioned above, feature-specific effects account for a large fraction of the variability in intensity between features. These feature effects are a function of sequence dependent probe affinities as well as the composition of the hybridization mixture hybridized to the array. Models have been developed to account for feature effects by estimating these effects in a multi-array context. The models are typically fitted to a biologically meaningful set of arrays – RNA extracted from liver samples; samples characterized as normal or cancerous, for example. In this section we discuss quality assessments based on derivatives, or by-products, of the PLIER model fits. The models are fitted feature set by

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

feature set and produce feature effect estimates, signal estimates and residuals. Feature effects are a nuisance parameter which yield improved signal estimates but can otherwise be ignored. Signal estimates are the expression indicators that will be used in the biological applications. They can also be used for quality assessment, as can the model residuals. The next two sections describe quality assessment measures that are based on these model-derived quantities. The analysis presented below can be performed in a variety of statistical analysis packages using the ExACT probeset summary output files as well as the option PLIER residual files.

III.A. Relative \log_2 signal summaries

After fitting the PLIER model to a set of arrays, we have signal estimates for each feature set and each array. We can assess data quality in terms of reproducibility of signal values across arrays as measured by \log_2 ratios of signal estimates or relative \log_2 signal. In a set of arrays hybridized to samples coming from distinct biological samples, some signal variability is expected due to real biological variability. Using a measure of variability based on the most reproducible relative \log_2 signal values – an inter-quartile range which in effect uses the 50% most reproducible signal values to assess reproducibility – produces a measure which can detect extra technical variability above and beyond biological variability. We can assess the reproducibility of signal estimates between two chips by means of an MA-plot, in which relative log signal is plotted on the vertical axis and average signal on the horizontal axis. As pair wise comparisons produce much redundancy, signal values for each array can instead be compared to signal values from a synthetic array constructed by computing for each feature set the median signal over all the arrays in the set. Superimposing the MA-plots with smooth curves capturing the quartiles (median, lower and upper quartiles) in relative \log_2 signal values as a function of average \log_2 signal is a helpful visual aid to detect subtle differences among MA-plots. Figure 5 displays MA-plots of relative log plier expression for the first 8 chips in the set of 20 displayed in Figure 1.

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

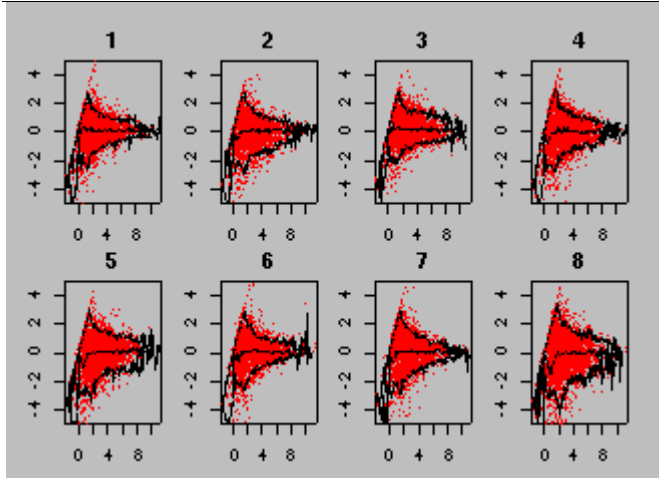


Figure 5: MA-plots for a set of arrays: $M = \text{relative } \log_2 \text{ plier signal for chip} = \log_2 \text{ plier signal for chip} - \text{median } \log_2 \text{ plier signal}$, $A = \log_2 \text{ plier signal for chip} + \text{median } \log_2 \text{ plier signal} / 2$.

Alternatively, boxplots of relative \log_2 signal values provide a succinct summary of MA-plots which ignores the relationship between relative \log_2 signal and mean \log_2 signal but facilitates the comparative assessments of the distribution relative \log_2 signal values among many chips. Figure 6 shows boxplots of relative \log_2 plier expression values for the set of arrays analyzed in Figure 1. We see that with the exception of array number 8, at the plier signal level the arrays are comparable in expression level – the center of the boxplots close to zero – and reproducibility – the size of the boxplots are comparable. Based on these assessments, one would be advised to exclude the data from array 8 from downstream analyses.

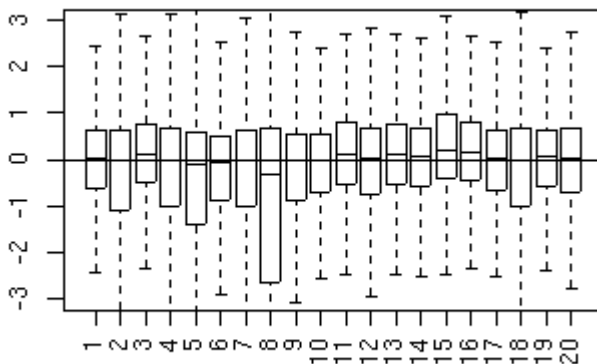


Figure 6: Boxplots of M values for a set of arrays: $M = \text{relative } \log_2 \text{ plier signal}$

In the assessment of reproducibility discussed so far we have implicitly assumed that the summaries would be computed based on all feature sets. We can

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

generalize the above assessments by summarizing over a various collections of feature sets. One may wish, for example, to look at signal reproducibility in the normalizing exon feature sets, or the normalizing intron feature sets, or a random sample of feature sets.

III.B. Summaries of residuals from fits

A by-product of the PLIER fit is the residuals from the model fit. These can be summarized in various ways to produce quality indicators. The median of the absolute value of residuals, pooled over all features sets, provides a simple array-level summary that will detect some differences in technical variability between arrays. Figure 7 displays a bar graph, with bar height proportional to $\text{mad}(\text{plier residual})$ for the 20 chips we have been using for illustration. Array 8 stands out as the chip with the largest $\text{mad}(\text{plier residual})$. $\text{Mad}(\text{plier residual})$ is highly correlated with the interquartile range of relative \log_2 signal.

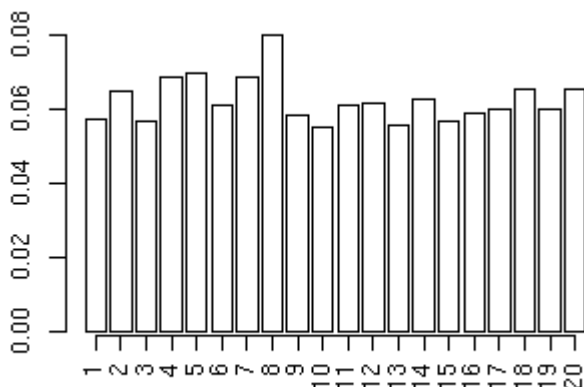


Figure 7: Bar chart of $\text{mad}(\text{plier residuals})$

Alternatively, the residuals can be combined into standard errors of signal estimates, and these can be summarized over all features sets on the array. The standard error can be normalized, by the median standard error for each feature set, for example, to provide a quality indicator that is slightly more sensitive to differences between arrays (see derivation of NUSE values in the Bioconductor monograph). As in the case of the relative \log_2 signal values summaries, the various residual summaries can be confined to various collections of features or feature sets.

IV. ExACT Quality Report

The probe summarization step (`exact-probe-summarize` command) will generate an optional report file. This text file contains various summary metrics from the CEL file data. It is important to note that the report file results are dependent on the particular analysis parameters used; thus different results can be generated

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

from the same set of CEL files depending on which intensity method, quantification method(s), and probesets are used.

The quality report file includes meta information about the particular analysis run. This information is encoded as lines starting with the pound symbol ('#'). For example, the line starting with '#%cmd=' contains the specific command line parameters used when the report file was generated.

Following the meta information is the summary metrics for each CEL file. The format is a column for each CEL file and a row for each metric with a tab separating each column. This file can be opened within Excel for instance as a tab separated text file. The various QC metrics reported are as follows with each metric potentially reported multiple times for different subsets of probesets. For example, the AFFX spike control probesets are reported separately (if included in the analysis) and as part of the total.

- Probe Count: This is the number of probes (features) analyzed.
- Probeset Count: The number of probesets analyzed.
- % Probes DABG Detected: The percent of probes with a DABG probe level p-value less than or equal to 0.01. This metric is only reported if the DABG quantification method is selected.
- % Probesets DABG Detected: The percent of probesets with a DABG probeset level p-value less than or equal to 0.01. This metric is only reported if the DABG quantification method is selected.
- Mean Probeset PLIER Target Response: The mean PLIER signal estimate.
- Mean Probeset Abs PLIER Residuals: The mean absolute PLIER probe level residuals.
- Mean Probeset Abs PLIER RLE: The mean absolute PLIER relative log expression (RLE). This metric is generated by taking the PLIER estimate for a given chip and calculating the difference in log base 2 from the median value of that probeset over all the chips. The mean is then computed from the absolute RLE for all the probesets for a given CEL file.
- Positive vs Negative ROC AUC: The area under the curve (AUC) for a receiver operator curve (ROC) comparing the intron controls to the exon controls by applying a threshold to the PLIER signal estimate. The ROC curve is generated by evaluating how well the PLIER signal estimate separates the intron controls from the exon controls. The assumption (which is only valid in part) is that the intron controls are a measure of false positives and the exon controls are a measure of true positives. An AUC of 1 reflects perfect separation whereas as an AUC value of 0.5 would reflect no separation. Note that the AUC of the ROC curve is equivalent to a rank sum statistic used to test for differences in the center of two distributions.

These metrics in general (and the RLE, ROC AUC, and Residual metrics specifically) have proven most useful in helping to identify outlier chip results relative to a given data set.

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

V. Conclusions

We have shown how some simple measures can be derived from array data, the feature or the feature set level, to assess the quality of array data in terms of comparability. It was argued that although comparing arrays in terms of feature intensity level is useful, decisions regarding which arrays to exclude from analysis should be based on signal level data – the data that will be used in downstream analysis. We have not exhausted all measures that can be used to assess quality.

Many of the quality assessment measures will be highly correlated. For the purpose of identifying gross outliers that should be excluded from downstream analysis any of the assessments discussed above will do a good job. Better methods – methods that are sensitive to departures in quality that have impact on particular downstream analyses – must by their very nature be developed on a case by case basis.

VI. Appendix sample R code

In this section we provide some example R code to produce the figures discussed in this note. For more information on the R language and it's use in bioinformatics applications, see [4] and [5].

```
#####
# Probe level assessments
# load cel intensity data
load('Data/Grp.cel.mtx')

# get Sample names
Sample.vec <- colnames(Grp.cel.mtx)

#####
# boxplot intensities
FigScale <- .5
x11(width=11*FigScale, height=8*FigScale)
par(mfcol=c(1,1), oma=c(0,0,0,0), mar=c(4,3,2,1))

boxplot(data.frame(log2(Grp.cel.mtx)), xlab='', xaxt='n',
          ylim=c(4, 7.5),outline=F)

axis(side=1, outer=F, at=1:length(Sample.vec), labels=Sample.vec, las=2)

#####
# Histogram cel intensities

MF <- c(2,4)
FigScale <- 0.5
x11(width=11*FigScale, height=8*FigScale)
par(mfrow=MF, oma=c(1,1,1,1), mar=c(2,2,2,1), cex.lab=1.25)

ndx.set <- 1:8

for (cc in ndx.set) {
```

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

```
    plot(density(log2(Grp.cel.mtx[,cc])),
          main=Sample.vec[cc], xlim=c(4,8), xlab='', ylab='')
  }

#####
# Now look at MVA plots
Grp.logcel.mtx <- log2(Grp.cel.mtx)
Grp.logcel.median.vec <- apply(Grp.logcel.mtx, 1, median)

MF <- c(2,4)
FigScale <- 0.5

x11(width=11*FigScale, height=8*FigScale)
par(mfrow=MF, oma=c(1,1,1,1), mar=c(2,2,2,1), cex.lab=1.25)

ndx.set <- 1:8

for (cc in ndx.set) {
  plot(x=Grp.logcel.median.vec, i
        y=Grp.logcel.mtx[,cc] - Grp.logcel.median.vec,
        main=Sample.vec[cc], ylim=c(-5,5), xlab='', ylab='', pch='.')
}

#####
# Repeat MVA plots using heatmap
Grp.logcel.mtx <- log2(Grp.cel.mtx)
Grp.logcel.median.vec <- apply(Grp.logcel.mtx, 1, median)

MF <- c(2,4)
FigScale <- 0.5

x11(width=11*FigScale, height=8*FigScale)
par(mfrow=MF, oma=c(1,1,1,1), mar=c(2,2,2,1), cex.lab=1.25)

ndx.set <- 1:8

for (cc in ndx.set) {
  mvaplot(y=Grp.logcel.median.vec, x=Grp.logcel.mtx[,cc], doLog=F,
           main=Sample.vec[cc], ylim=c(-5,5), xlab='', ylab='', pch='.')
}

#####
# Now look at M-boxplots
Grp.logcel.mtx <- log2(Grp.cel.mtx)
Grp.logcel.median.vec <- apply(Grp.logcel.mtx, 1, median)
Grp.RLE.mtx <- sweep(Grp.logcel.mtx, MARGIN=1,
                    STATS=Grp.logcel.median.vec, FUN='-')

MF <- c(1,1)
FigScale <- 0.5

x11(width=11*FigScale, height=8*FigScale)

par(mfcol=MF, oma=c(0,0,0,0), mar=c(4,3,2,1))

boxplot(data.frame(Grp.RLE.mtx), xlab='', xaxt='n', ylim=c(-1,1),
            outline=F)
abline(h=0)
axis(side=1, outer=F, at=1:length(Sample.vec), labels=Sample.vec, las=2)
```

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

```
#####  
# Probe set level assessments  
  
# load summary data  
load('Data/Grp.ProbeSetSum.mtx')  
  
# get Sample names  
Sample.vec <- colnames(Grp.ProbeSetSum.mtx)  
  
#####  
# Look at MA plots  
Grp.logSignal.mtx <- log2(Grp.ProbeSetSum.mtx)  
Grp.logSignal.median.vec <- apply(Grp.logSignal.mtx, 1, median)  
  
MF <- c(2,4)  
FigScale <- 0.5  
  
x11(width=11*FigScale, height=8*FigScale)  
par(mfrow=MF, oma=c(1,1,1,1), mar=c(2,2,2,1), cex.lab=1.25)  
ndx.set <- 1:8  
  
for (cc in ndx.set) {  
  plot(x=Grp.logSignal.median.vec,  
       y=Grp.logSignal.mtx[,cc] - Grp.logSignal.median.vec,  
       main=Sample.vec[cc], ylim=c(-5,5), xlab='', ylab='', pch='.')  
}  
  
Z#####  
# Repeat MA plots using heatmap mva  
Grp.logSignal.mtx <- log2(Grp.ProbeSetSum.mtx)  
  
MF <- c(2,4)  
FigScale <- 0.5  
  
x11(width=11*FigScale, height=8*FigScale)  
par(mfrow=MF, oma=c(1,1,1,1), mar=c(2,2,2,1), cex.lab=1.25)  
  
ndx.set <- 1:8  
  
for (cc in ndx.set) {  
  mvaplot(y=Grp.logSignal.median.vec, x=Grp.logSignal.mtx[,cc],  
          main=Sample.vec[cc], ylim=c(-5,5), xlab='', ylab='', pch='.')  
}  
  
#####  
# Now look at M-boxplots  
Grp.logSignal.mtx <- log2(Grp.ProbeSetSum.mtx)  
Grp.logSignal.median.vec <- apply(Grp.logSignal.mtx, 1, median)  
Grp.RLE.mtx <- sweep(Grp.logSignal.mtx, MARGIN=1,  
                    STATS=Grp.logSignal.median.vec, FUN='-')  
  
MF <- c(1,1)  
FigScale <- 0.5  
  
x11(width=11*FigScale, height=8*FigScale)  
  
par(mfcol=MF, oma=c(0,0,0,0), mar=c(4,3,2,1))  
  
boxplot(data.frame(Grp.RLE.mtx), xlab='', xaxt='n', ylim=c(-3,3),  
         outline=F)  
abline(h=0)
```

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

```
axis(side=1, outer=F, at=1:length(Sample.vec), labels=Sample.vec, las=2)
```

```
#####
```

```
# Bar plot median absolute deviation of residuals
```

```
load('Data/Grp.Residuals.mtx')
```

```
Grp.mad.vec <- apply(Grp.Residuals.mtx,2,mad)
```

```
MF <- c(1,1)
```

```
FigScale <- 0.5
```

```
x11(width=11*FigScale, height=8*FigScale)
```

```
par(mfcol=MF, oma=c(0,0,0,0), mar=c(4,3,2,1))
```

```
bar.out <- barplot(Grp.mad.vec , xlab='', xaxt='n', col=0)
```

```
axis(side=1, outer=F, at=bar.out, labels=Sample.vec, las=2)
```

```
#####
```

```
# Code for mvaplot function
```

```
mvaplot <- function (x, y, intensityPlot = TRUE, doLog = TRUE,
```

```
  nBin = 100,
```

```
  xbound = rep(NA, 2), ybound = rep(NA, 2),
```

```
  xlim = rep(NA, 2),
```

```
  ylim = rep(NA, 2), xlab = "Average", ylab = "Difference",
```

```
  probs = c(0.05, 0.5, 0.95), ...)
```

```
{
```

```
  a <- (x + y)/2
```

```
  m <- x - y
```

```
  if (intensityPlot) {
```

```
    op <- par(bg = "gray")
```

```
    if (any(is.na(xbound)))
```

```
      xbound <- range(a)
```

```
    aBinSize <- crossprod(xbound, c(-1, 1))/nBin
```

```
    aQuantized <- (xbound[1] + (floor((pmin(pmax(a, xbound[1]),  
      xbound[2] - 0.1 * aBinSize) - xbound[1])/aBinSize) +  
      0.5) * aBinSize)
```

```
    aBinCenter <- xbound[1] + ((0:(nBin - 1)) + 0.5) * aBinSize
```

```
    if (any(is.na(ybound)))
```

```
      ybound <- range(m)
```

```
    mBinSize <- crossprod(ybound, c(-1, 1))/nBin
```

```
    mQuantized <- (ybound[1] + (floor((pmin(pmax(m, ybound[1]),  
      ybound[2] - 0.1 * mBinSize) - ybound[1])/mBinSize) +  
      0.5) * mBinSize)
```

```
    mBinCenter <- ybound[1] + (0:(nBin - 1) + 0.5) * mBinSize
```

```
    dataTable <- matrix(0, nrow = length(aBinCenter),  
      ncol = length(mBinCenter))
```

```
    rownames(dataTable) <- as.character(aBinCenter)
```

```
    colnames(dataTable) <- as.character(mBinCenter)
```

```
    countTable <- table(aQuantized, mQuantized)
```

```
    dataTable[rownames(countTable),
```

```
      colnames(countTable)] <- countTable
```

```
    quantiles <- apply(dataTable, 1, myQuant, values = mBinCenter,  
      probs = probs)
```

```
    if (any(is.na(xlim)))
```

```
      xlim <- xbound
```

```
    if (any(is.na(ylim)))
```

```
      ylim <- ybound
```

Quality Assessment of Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

```
    if (doLog) {
      image(aBinCenter, mBinCenter, log(dataTable), xlim = xlim,
            ylim = ylim, xlab = xlab, ylab = ylab, ...)
    }
    else {
      dataTable[dataTable == 0] <- -Inf
      image(aBinCenter, mBinCenter, dataTable, xlim = xlim,
            ylim = ylim, xlab = xlab, ylab = ylab, ...)
    }
    apply(quantiles, 1, lines, x = aBinCenter, type = "l",
          lwd = 0.5)
    par(op)
  }
  else {
    plot(a, m, xlim = xlim, ylim = ylim, xla = xlab, ylab = ylab,
         ...)
  }
  iqr.diff <- quantile(m, prob = c(0.25, 0.75)) %*% c(-1, 1)
  return(iqr.diff)
}
```

VII. References:

- [1] Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, 2005.
- [2] Tukey, John W., *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass., 1977.
- [3] Dudoit S, Yang YH, Callow MJ and Speed TP. *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*. *Statistica Sinica*, 12:111-139, 2002.
- [4] R Development Core Team, *R: A Language and Environment for Statistical Computing*, 2005, <http://www.R-project.org>.
- [5] Gentleman, RC, et. al., *Bioconductor: Open software development for computational biology and bioinformatics*, *Genome Biology*, 5, 2004, <http://genomebiology.com/2004/5/10/R80>.