RNA Degradation and NUSE Plots

Austin Bowles

STAT 5570/6570

April 22, 2011

References

- Sections 3.4 and 3.5.1 of Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Gentleman et al., 2005)

Notes 2.1 – Quality Checks

- Even after preprocessing to remove “noise” and make arrays comparable, some arrays may be “beyond correction.”

- Visual checks of microarray quality (from Notes 2.1)
  - Array Images
  - Boxplots/Histograms
  - MA Plot
  - PLM Images/Residual Images
- Also:
  - RNA Degradation
  - Normalized Unscaled Standard Error (NUSE) Plots

Why RNA Degradation?

- Once RNA has reached the end of its useful life (i.e. it has participated in protein synthesis) it is “degraded” by cellular enzymes.

- Some arrays might have been prepared using a sample with “bad” RNA
  - –RNA that has been degraded past the point of providing useful information.

- We need a way to check whether or not our arrays have “good” RNA.

5’ to 3’ Ordering of Probesets

- A gene is represented by a *probeset* on a microarray.
- Each probeset consists of about 11 different PM probes.
- For each probeset on an array, the individual probes are numbered sequentially from the 5’ end of the transcript to the 3’ end.

Gene B (DNA)

Transcription

RNA Sequence

5' CACGCACCCCG AUGCCCUUGACGUUAGUGUGUCCAG UACCUGUGUAGCGUAUUUACAUGCCUUCUCGACUC AAUGCACUAACCACUGAACUACCCG UACUCAAAGA

Probe 0

Probe 1

G

C

UGCACGACUAUCUAUGUGUAGAACAUUG.....UCAGCAGCCCGCCACCCUAUAGCGGACCCCUGAGCACUCUGGCUUUACUGCACGGUAUUGGUUCCAUGUUGGCA

G

A

Probe 10

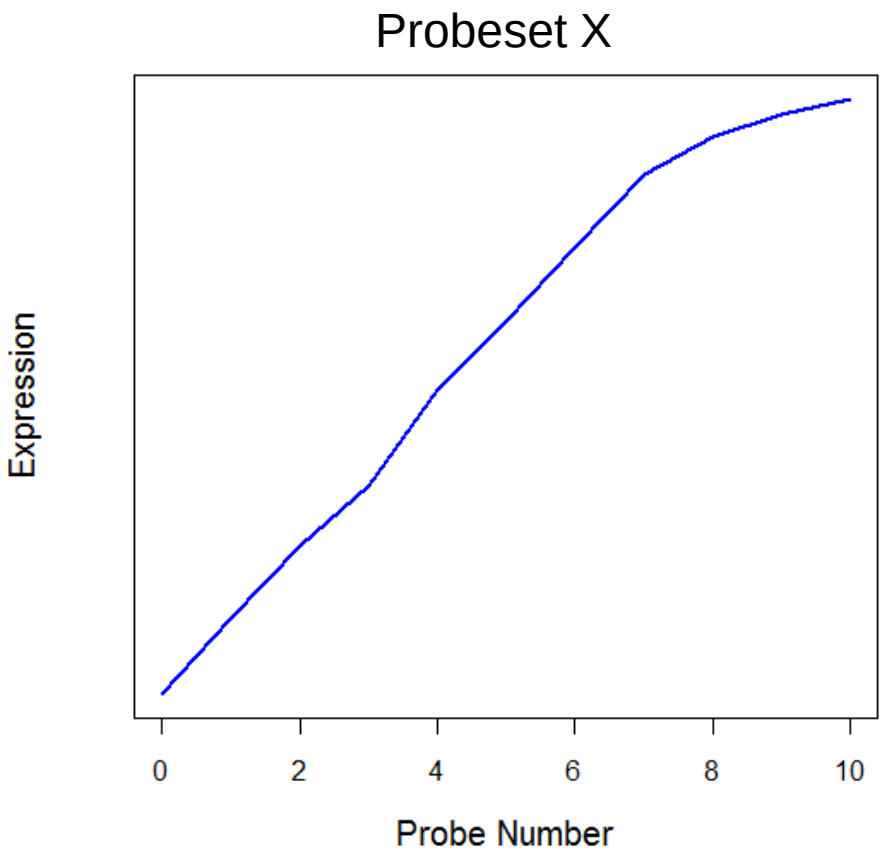
CCAGAGUACUAGGCCACUUAUAUUACUCAAG GAAGGACGGUGGUCCGAGGCAUUCG CAUACAGACCUAGUCCUCG 3'

RNA Degradation in Probesets

- Due to the specific mechanisms of RNA degradation, probe intensities should be systematically lower towards the 5’ end of a transcript than towards the 3’ end.

- This fact can be exploited for analyzing expression array data.

RNA Degradation for One Probeset



- Even with minimal degradation, we should see an upward trend in expression levels as probe number increases.
- The slope of this trend depends on the probeset and the extent of the degradation.
- We need a way to look at all probesets on an array.

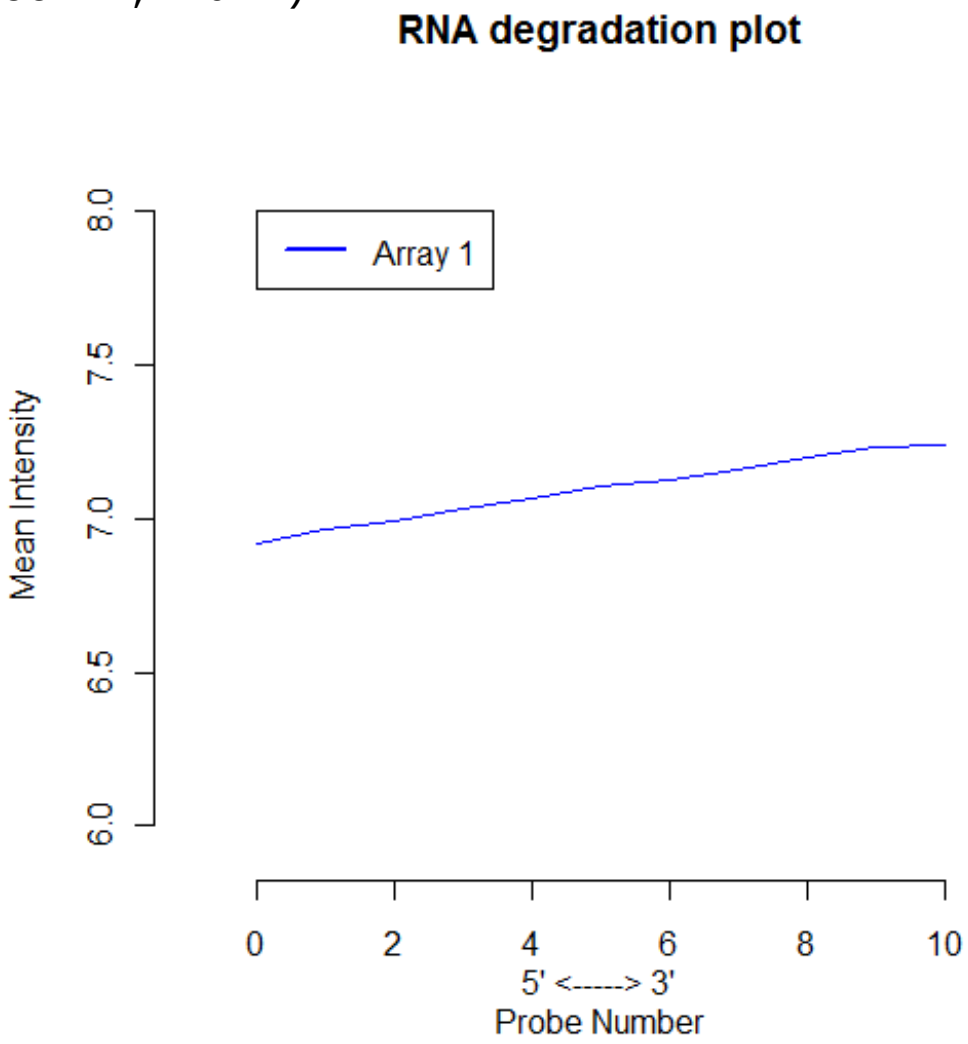
The "**AffyRNAdeg**" Function

Y_{ij} = the log transformed intensity for the j^{th} probe in the i^{th} probeset.

$Y_{\cdot j}$ = the average log intensity at the j^{th} position, taken over all probesets in an array.

- Plotting the $Y_{\cdot j}$ vs. j shows a linear 3'/5' trend, even in an experiment with "good" RNA.

```
library(affy);library(affydata)
abatch.raw <- ReadAffy(celfile.path="F:\\R\\Data\\GSE5425")
hw.rnadeg.1 <- AffyRNAdeg(abatch.raw[,1])
plotAffyRNAdeg(hw.rnadeg.1,”neither”)
legend(0,40,"Array 1",lty=1,col=4,lwd=2)
```



Rescaling

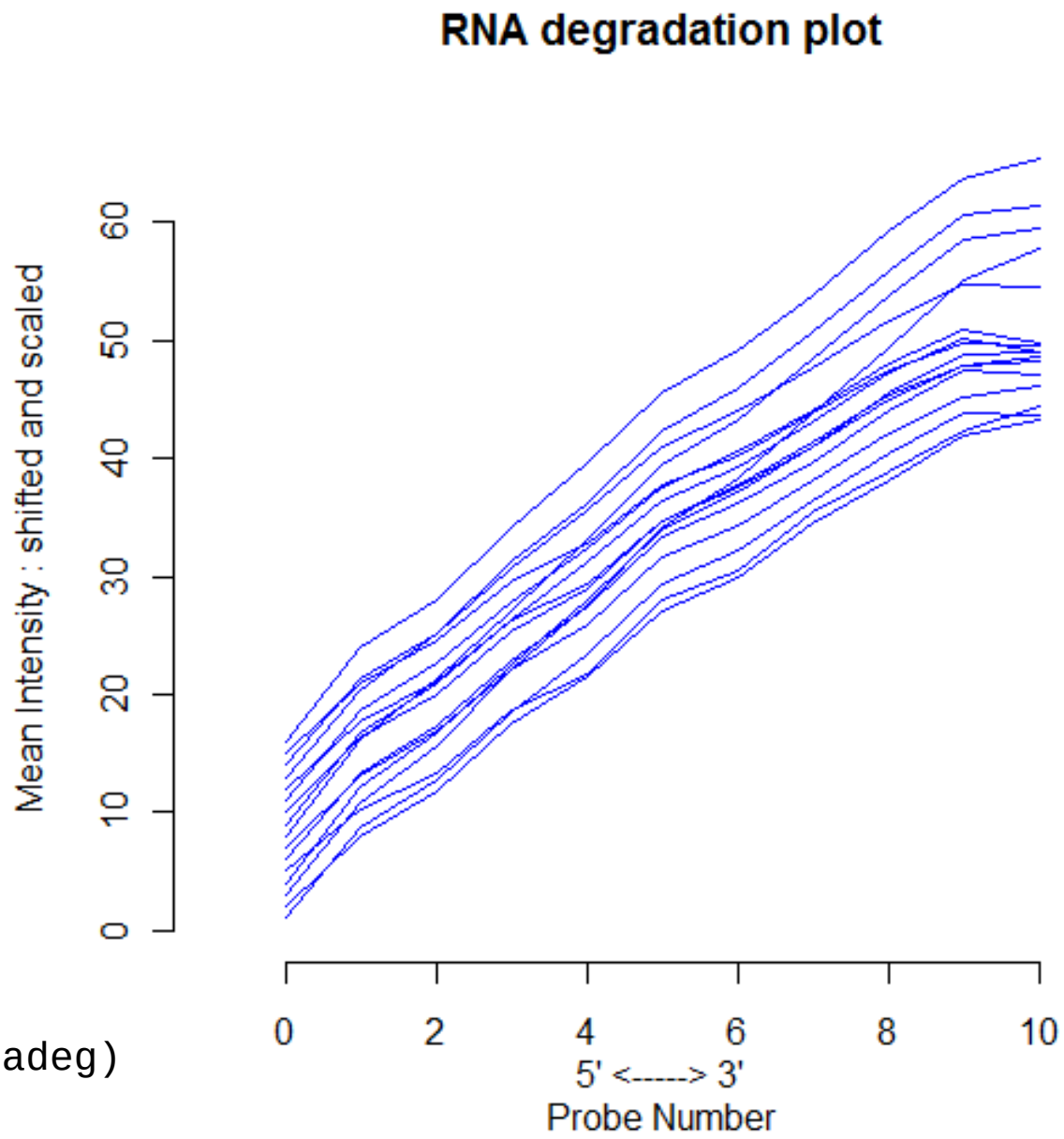
- To make these plots comparable across arrays, we first rescale them so that the standard errors of the rescaled means are approximately 1.

$\hat{\sigma}_j$ = standard deviation for probes at position j in all N probesets

$\hat{\sigma}_j/\sqrt{N} \approx$ standard error for mean intensity at position j

$$\frac{Y_{\cdot j}}{\hat{\sigma}_j/\sqrt{N}} = \text{rescaled mean with standard error} \approx 1$$

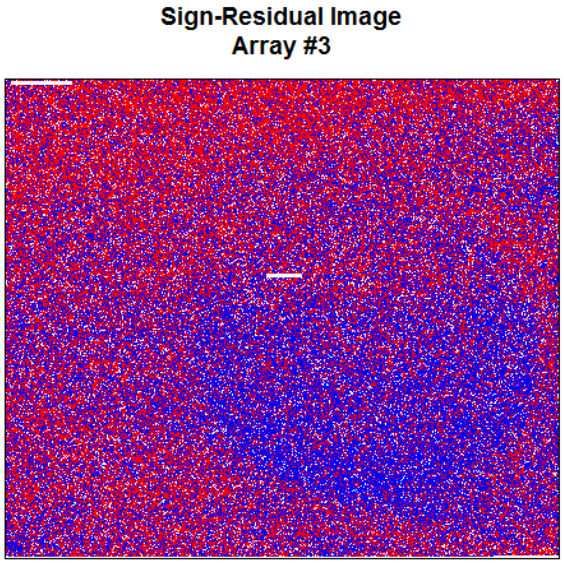
The lines have been shifted from the original data for a clearer view, but the slopes remain unchanged.

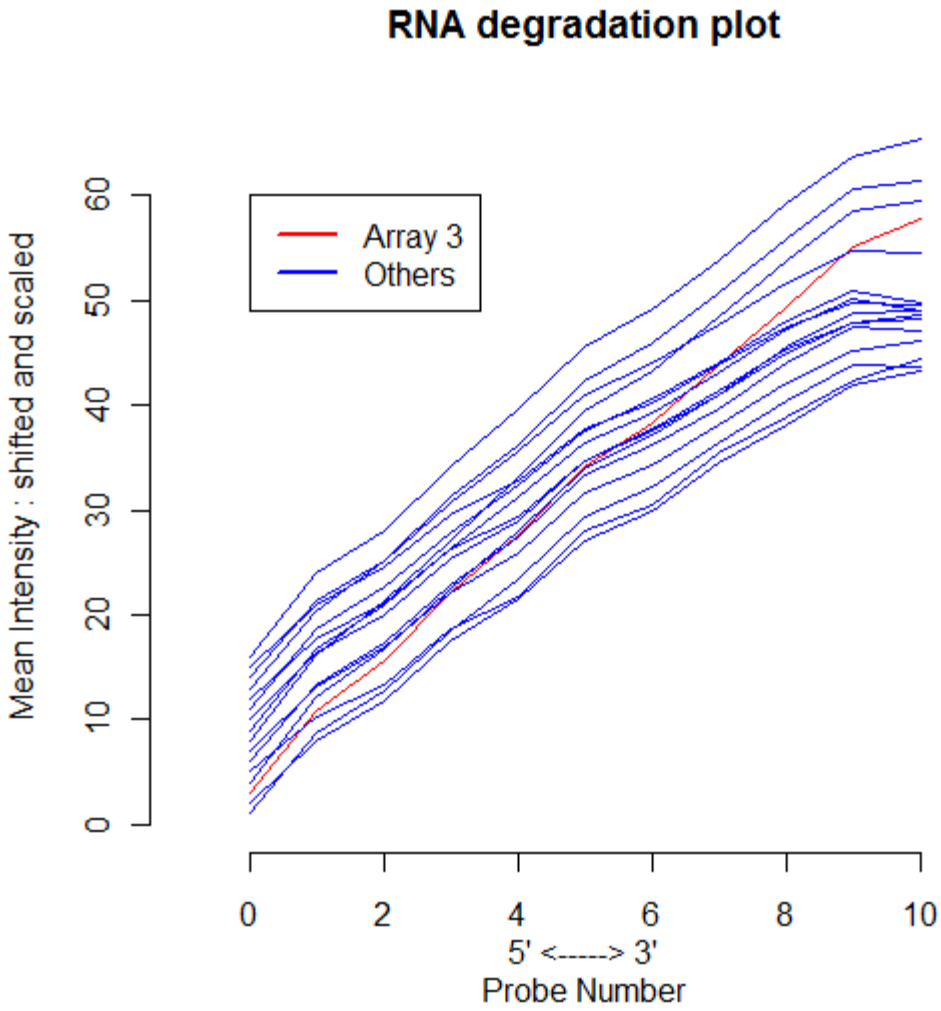



```
plotAffyRNAdeg(hw.rnadeg)
```

What to look for:

- Slopes deviating from the group pattern.
- Remember Array #3 from our homework?



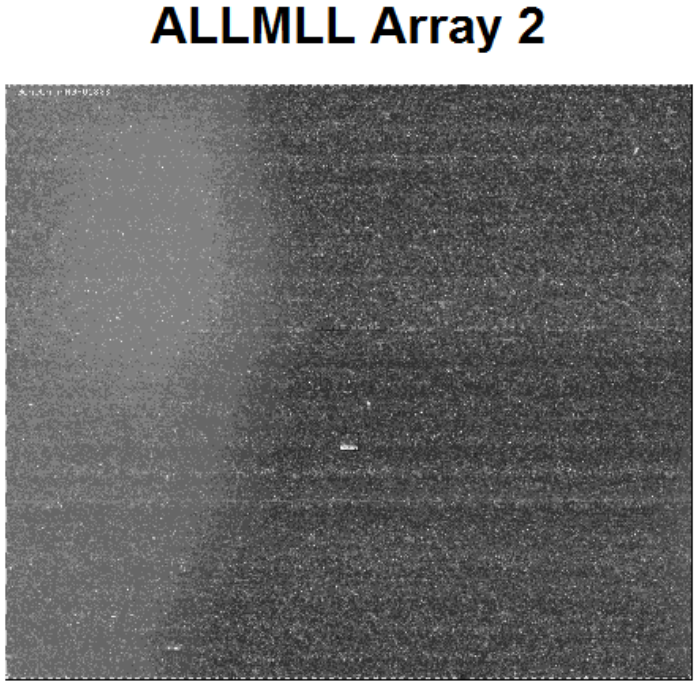


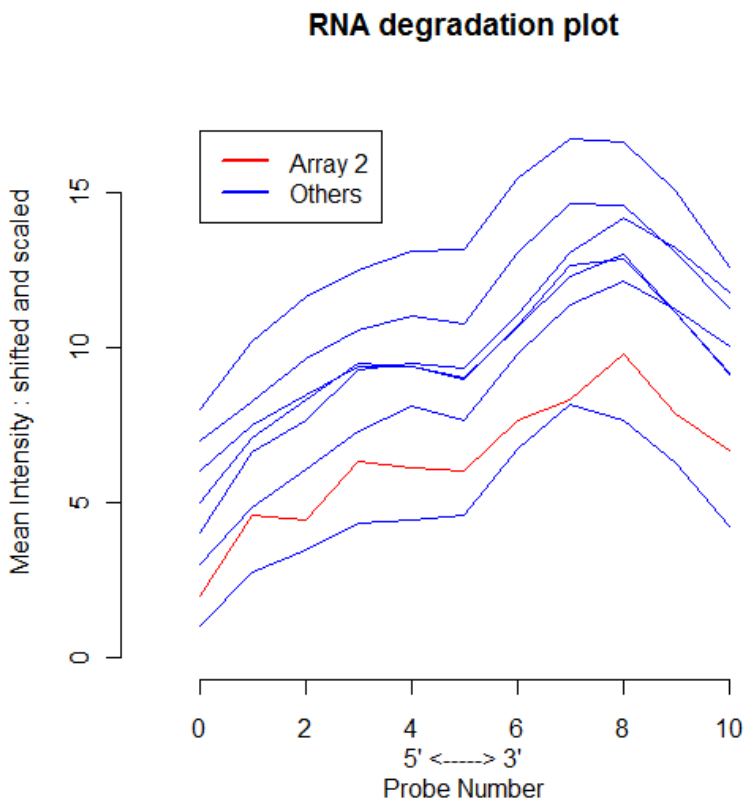
```
cols <- rep(4,16);cols[3]<-2
plotAffyRNAdeg(hw.rnadeg,cols=cols)
legend(0,60,c("Array 3","Others"),lty=1,col=c(2,4),lwd=2)
summaryAffyRNAdeg(hw.rnadeg)
```

	GSM118665.CEL	GSM118666.CEL	GSM118667.CEL	GSM118668.CEL	GSM118669.CEL	GSM118671.CEL
slope	4.38e+00	4.21e+00	5.52e+00	4.58e+00	4.04e+00	4.22e+00
pvalue	1.12e-08	2.52e-10	4.43e-12	1.31e-08	3.07e-11	1.17e-08
	GSM118673.CEL	GSM118674.CEL	GSM118677.CEL	GSM118679.CEL	GSM118681.CEL	GSM118682.CEL
slope	4.01e+00	5.25e+00	4.14e+00	3.92e+00	4.00e+00	3.77e+00
pvalue	1.07e-09	3.62e-10	3.96e-08	3.70e-09	4.53e-08	1.01e-09
	GSM118684.CEL	GSM118686.CEL	GSM118687.CEL	GSM118689.CEL		
slope	4.95e+00	4.16e+00	3.58e+00	4.98e+00		
pvalue	2.61e-10	1.64e-08	9.56e-09	1.27e-10		

Array #3

Example: ALLMLL Data





- RNA Degradation Plots can help identify physical artifacts, but should not be used exclusively.

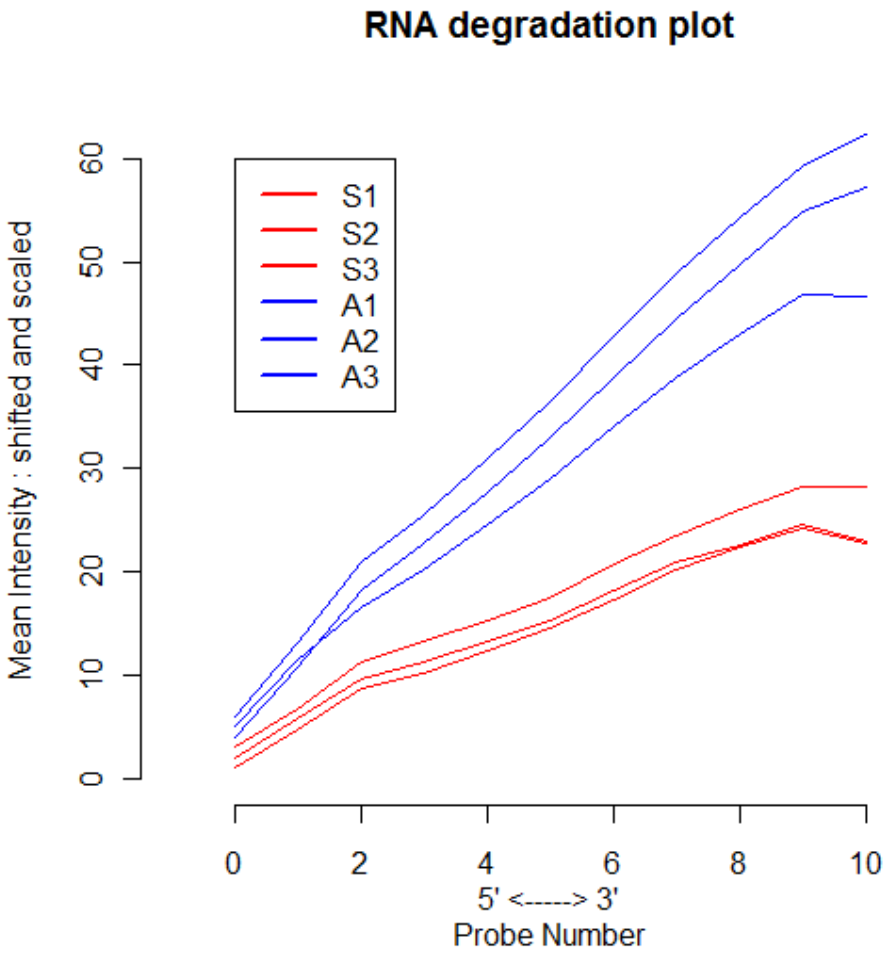
```
library(ALLMLL); data(MLL.B)
ALLdata <- MLL.B[,c(1:6,13,14)]
ALL.rnadeg <- AffyRNAdeg(ALLdata)
cols <- rep(4,8);cols[2]=2
plotAffyRNAdeg(ALL.rnadeg, cols=cols)
legend(0,17,c("Array 2","Others"),lty=1,col=c(2,4),lwd=2)
```

- Instead, we should look for derivations that suggest an array was prepared with “bad” RNA.

```
library(AmpAffyExample)
data(AmpData)
AmpData
#Array S3 had issues in lecture notes
sampleNames(AmpData) <- c("S1","S2","S3","A1","A2","A3")
AmpData.rnadeg <- AffyRNAdeg(AmpData)
par(mfrow=c(1,1))
plotAffyRNAdeg(AmpData.rnadeg, col=c(2,2,2,4,4,4))
legend(0,60,sampleNames(AmpData),lty=1,col=c(2,2,2,4,4,4),lwd=2)
```

Example: AmpAffyExample Data

- The steeper slopes for the A samples suggest that the RNA degradation has advanced further than for the S samples.



How steep is too steep?

- Unfortunately, different array types have different “normal” slopes. Experience will give the user a sense of what is typical for the different arrays.

- Instead, it is important to check for agreement within studies.
  - If the arrays in a study have similar slopes, then within gene comparisons may still be valid.
  - If the arrays have different degrees of RNA degradation (i.e. different slopes), extra bias is introduced into the experiment.

NUSE Plots

- Normalized Unscaled Standard Error Plots.

- Remember the Probe Level Model:

$$Y_{ijk} = \mu_{ik} + \alpha_{jk} + \varepsilon_{ijk}$$

Log-scale expression level for gene k on array i

Probe Affinity Effect

- The R function `fitPLM` fits the Probe Level Model and other functions are available for accessing its output.

Standard Error Estimates

- We're interested in the Standard Error Estimates, SE($\hat{\mu}_{ik}$), obtained by the PLM fit.

```
Pset1 <- fitPLM(abatch.raw)
##NOTE: rmaPLM(abatch.raw) will not return standard errors
se(Pset1)[1:6,1:3]
               GSM118665.CEL GSM118666.CEL GSM118667.CEL
1415670_at        0.10637111    0.11596710    0.11225729
1415671_at        0.11275118    0.12974944    0.11359738
1415672_at        0.11465433    0.13025307    0.11655682
1415673_at        0.12740979    0.14139465    0.12500363
1415674_a_at      0.08401818    0.09562223    0.08614656
1415675_at        0.12846968    0.14957198    0.12846968
```

- Since variability differs between genes, we standardize these standard errors so that the median SE across arrays is 1 for each gene.

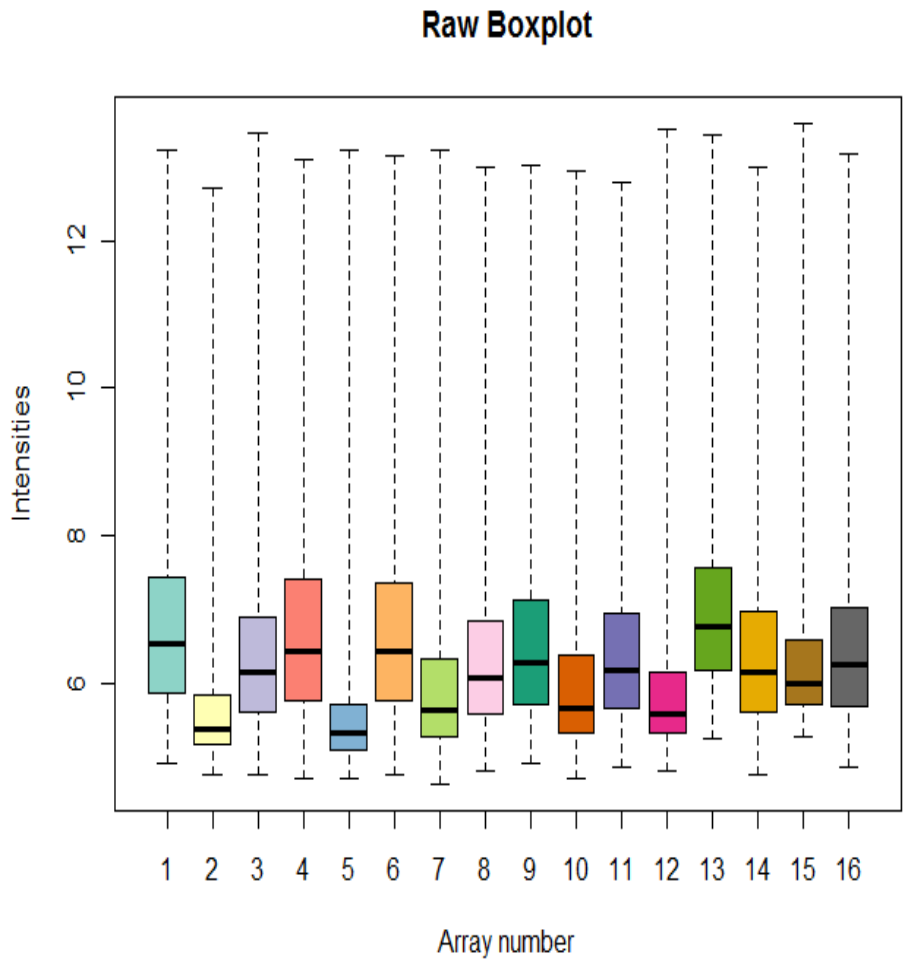
- $$\mathrm{NUSE}(\hat{\mu}_{ik}) = \frac{\mathrm{SE}(\hat{\mu}_{ik})}{\mathrm{med_i}(\mathrm{SE}(\hat{\mu}_{ik}))}$$

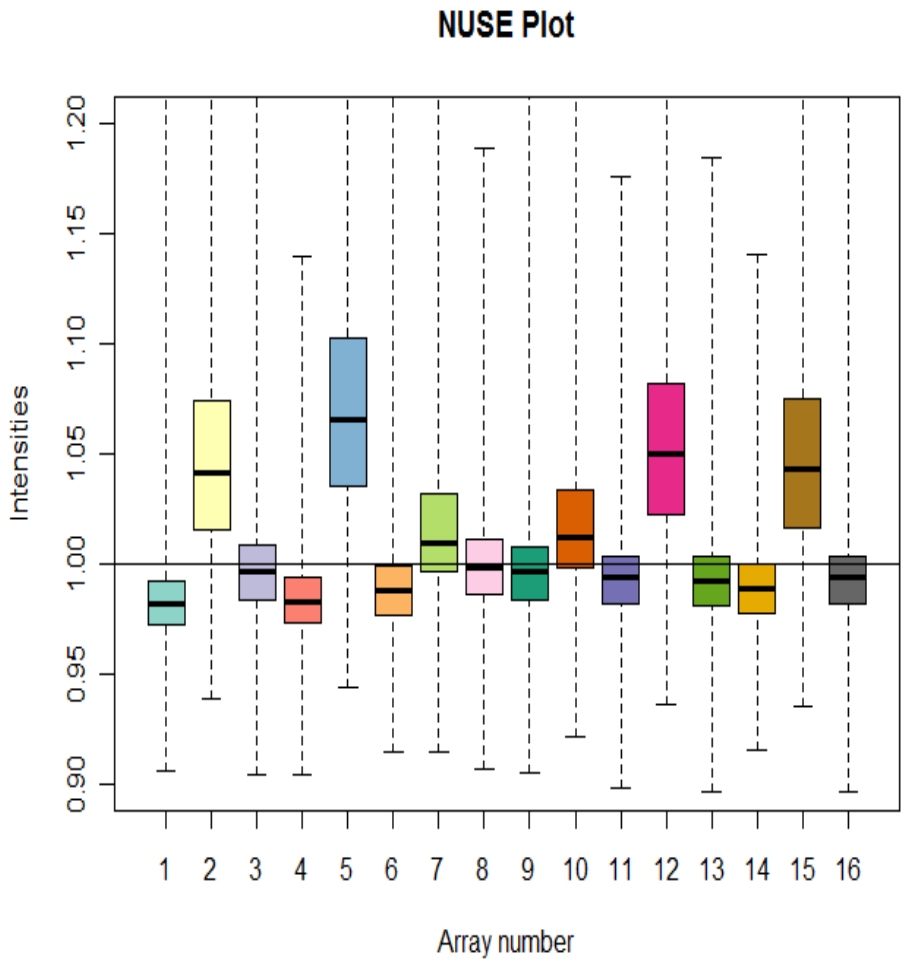
- We visualize these NUSE values with a boxplot. This can be done with `boxplot( )` or `NUSE( )`.

- When examining these plots, look for arrays with boxes that are significantly elevated or more spread out than other arrays. These indicate lower quality arrays.

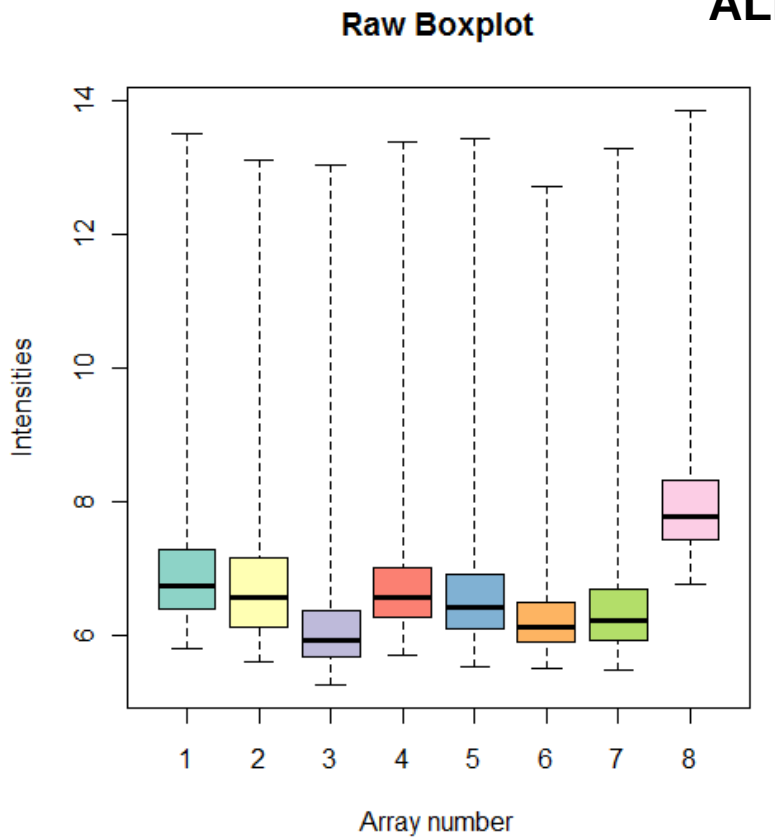
```
par(mfrow=c(1,2))
cols <- c(brewer.pal(8, "Set3"),brewer.pal(8, "Dark2"))
boxplot(abatch.raw[,1:16],col=cols,names=1:16,
        xlab= "Array number", ylab= "Intensities",
        main="Raw Boxplot")
NUSE(Pset1,col=cols,names=1:16,xlab= "Array number",
        ylab= "Intensities",main="NUSE Plot")
```

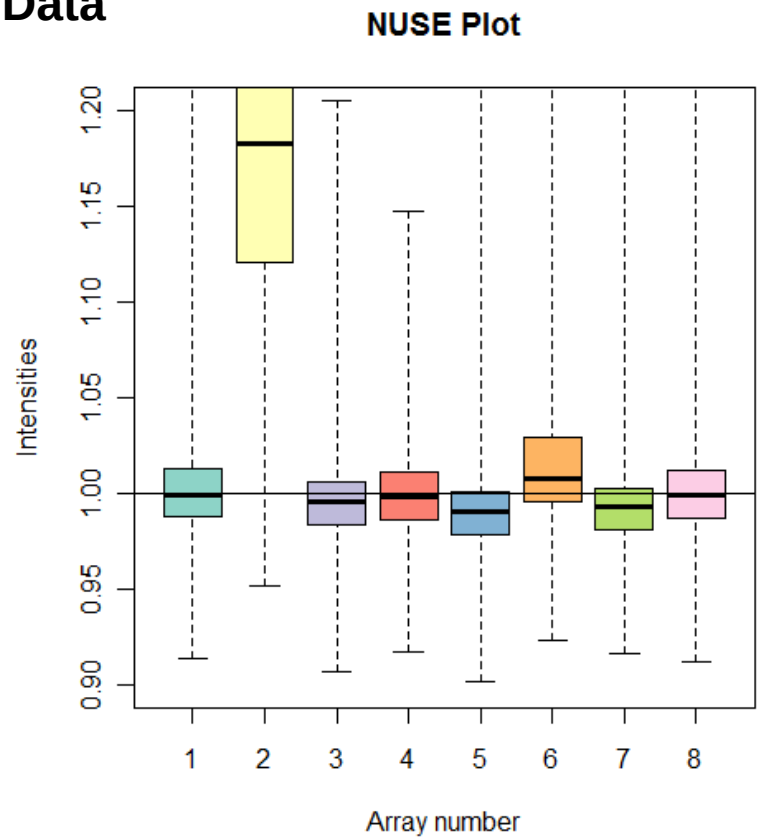
HW Data


Raw Boxplot
Intensities
6
8
10
12
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Array number

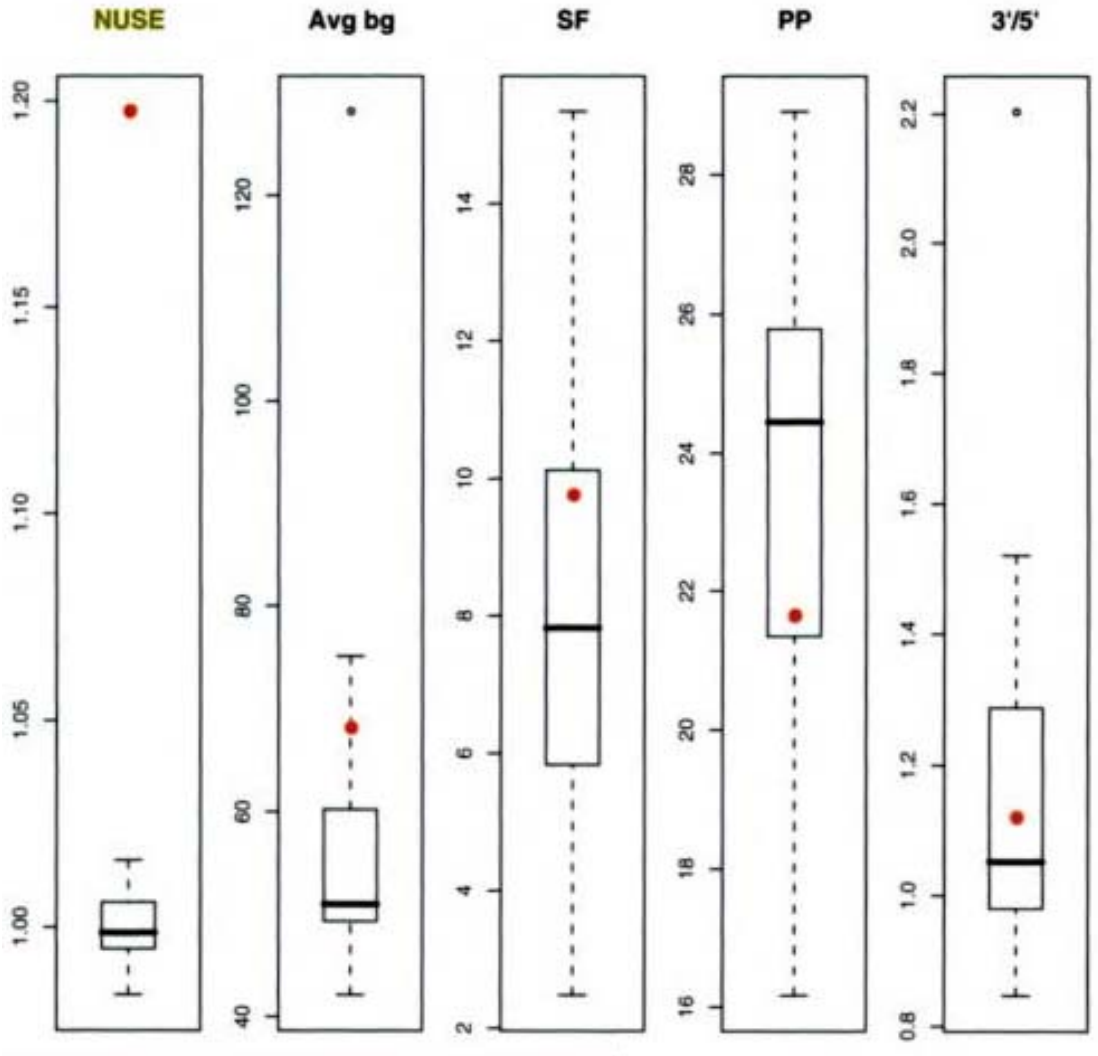

NUSE Plot
Intensities
0.90
0.95
1.00
1.05
1.10
1.15
1.20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Array number

ALL Data






```
Pset2 <- fitPLM(ALLdata)
cols <- brewer.pal(8,"Set3")
boxplot(ALLdata,col=cols,names=1:8,xlab="Array number",
ylab="Intensities",
   main="Raw Boxplot")
NUSE(Pset2,col=cols,names=1:8,xlab="Array number",
ylab="Intensities",
   main="NUSE Plot")
```



Comparing median NUSE with Affymetrix quality metrics for the ALLMLL dataset. Array 2 is indicated on each plot.

Figure 3.8 courtesy of Course Text (Gentleman et al., 2005)

Summary

- RNA Degradation

  -Helps identify arrays with “bad” RNA.

  -Can sometimes help identify arrays with physical artifacts.

  -Samples with similar degradation can be used for valid comparisons.

- NUSE Plots

  -Can help identify lower quality arrays.

  -Easier than Affymetrix quality standards to distinguish problematic arrays.