Microarray

Lecture 3

Outline

- Background
  - Biology Background
  - Introduction to useful packages in Bioconductor
- Preprocessing of oligonucleotide microarray
  - mas
  - rma
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

MAS 5.0

1. Background correction: weighted average of grid background
2. Normalization: trimmed mean scaling
3. PM-MM correction (optional): Ideal mismatch
4. Summarization: one step Tukey’s Biweight function

```
>  set <- expresso(Dilution, bgcorrect.method = "mas",
normalize.method = "constant", pmcorrect.method = "mas",
summary.method = "mas")
```

```
> expression <-mas5(Dilution)
```

MAS5: Summary

- Good
  - Usable with single chips (though replicated preferable)
  - Gives a p-value for expression data
- Bad:
  - Lots of fudge factors in the algorithm
  - Not *exactly* reproducible based upon documentation (source now available)
- Misc
  - Most commonly used processing method for Affy chips
  - Highly dependent on Mismatch probes

RMA

1. Background correction: RMA convolution
2. Normalization: quantile normalization
3. PM-MM correction (optional): none
4. Summarization: Fitting probe level model

- Under R

```
>  set <- expresso(Dilution, bgcorrect.method = “rma",
normalize.method = "quantiles", pmcorrect.method =
“pmonly", summary.method =“medianpolish")
```

```
> expression<-rma(Dilution)
```

RMA: Summary

- Good:
  - Results are $\log_2$ scaled from the raw intensity values
  - Rigidly model based method: defines model then tries to fit experimental data to the model. Fewer fudge factors than MAS5
- Bad
  - Does not provide "calls" as MAS5 does. MAS5 has p-values for each probe, and Present/Marginal/Absent calls are thresholded.
  - RMA cannot be applied to single chip.
- Misc
  - The input is a group of samples that have same distribution of intensities.
  - Requires multiple samples

Outline

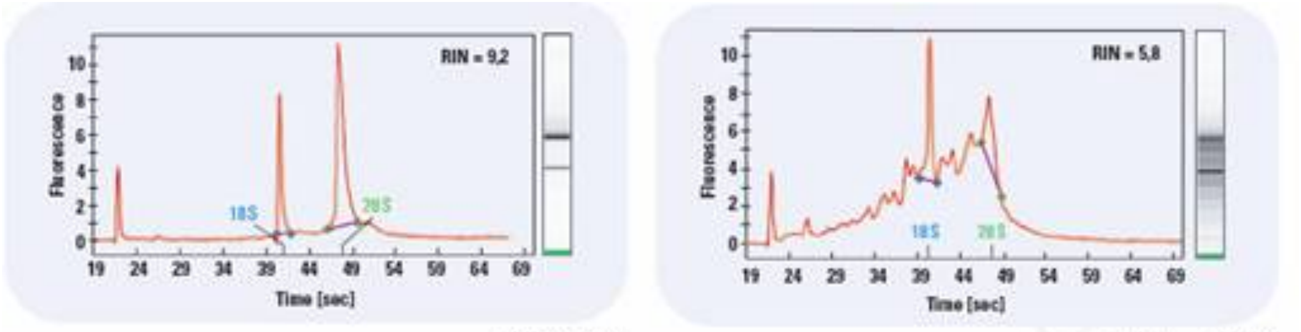
- Background
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

Some Causes of Technical Variation

- Temperature of hybridization differs
- Amount of RNA differs
- RNA degraded in some samples
- Yield of conversion to cDNA or cRNA differs
- Strength of ionic buffers differs
- Stringency of wash differs
- Scratches on some chips
- Ozone (affects Cy5) at some times

Quality Assessment

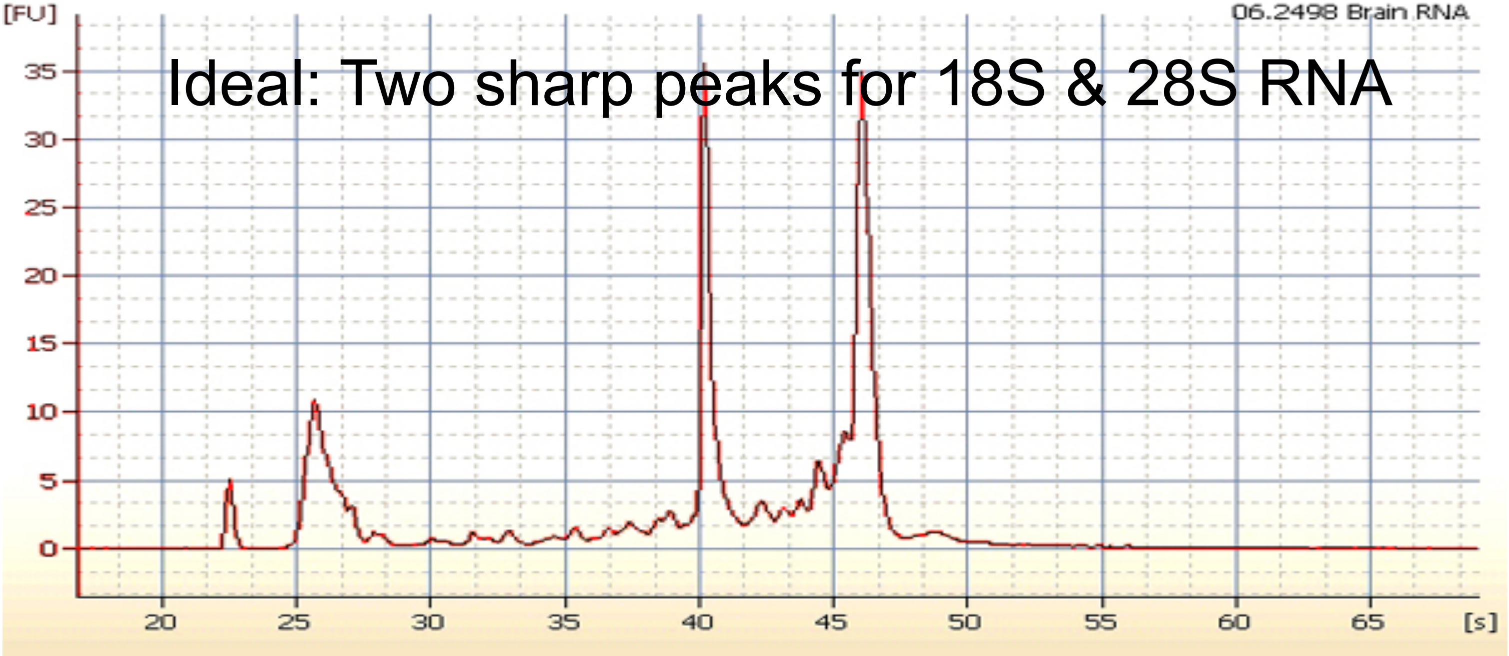
- Are there any factors that would lead you to doubt or distrust a particular array ?
- Quality of inputs – e.g. RNA quality



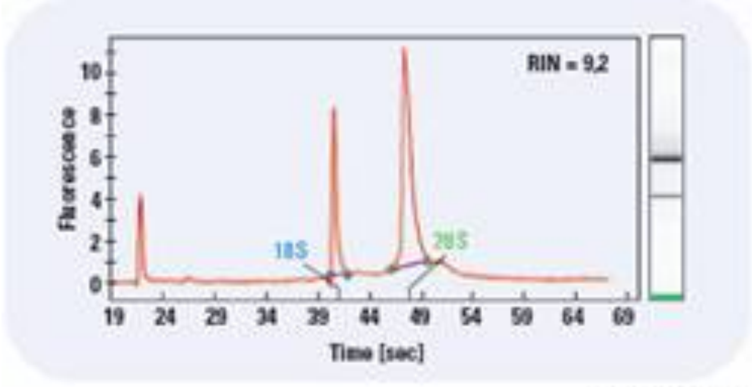
Intact total RNA

Partially degraded RNA

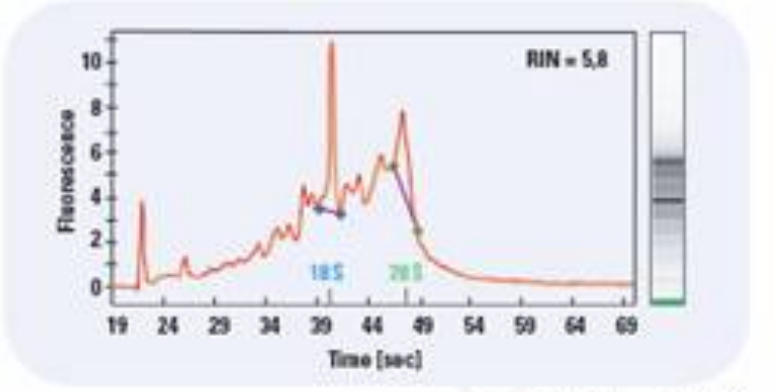
BioAnalyzer



Ideal: Two sharp peaks for 18S & 28S RNA



Intact total RNA



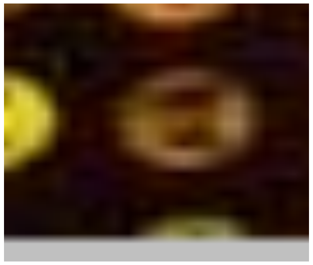
Partially degraded RNA

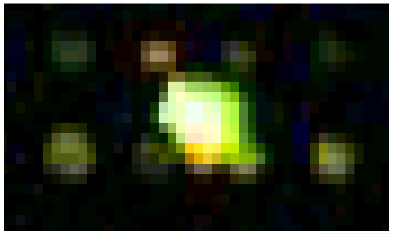
Spot QA for cDNA Spotted Arrays

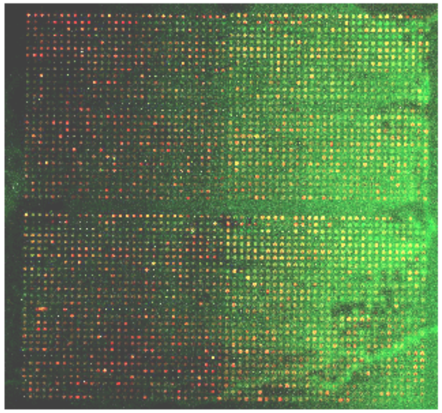
- Spot Measures
  - Uniformity
  - Spot Area

- Inspect images for artifacts

- Global Measures
  - Qualitative assessments
  - Averages of spot measures







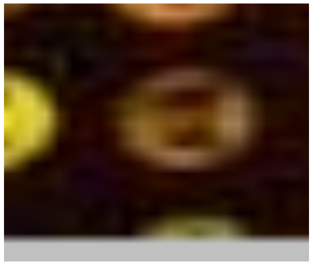
With commercial arrays we assume these issues are under control

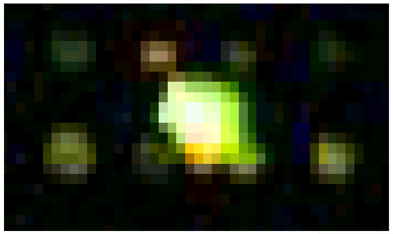
Spot QA for cDNA Spotted Arrays

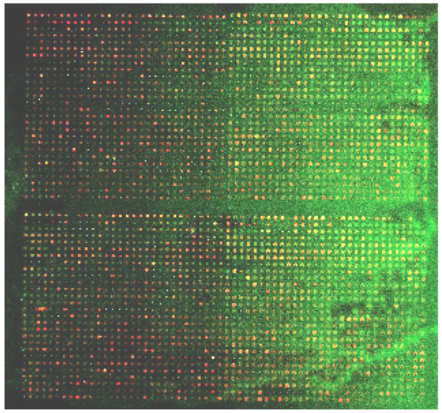
- Spot Measures
  - Uniformity
  - Spot Area

- Inspect images for artifacts

- Global Measures
  - Qualitative assessments
  - Averages of spot measures

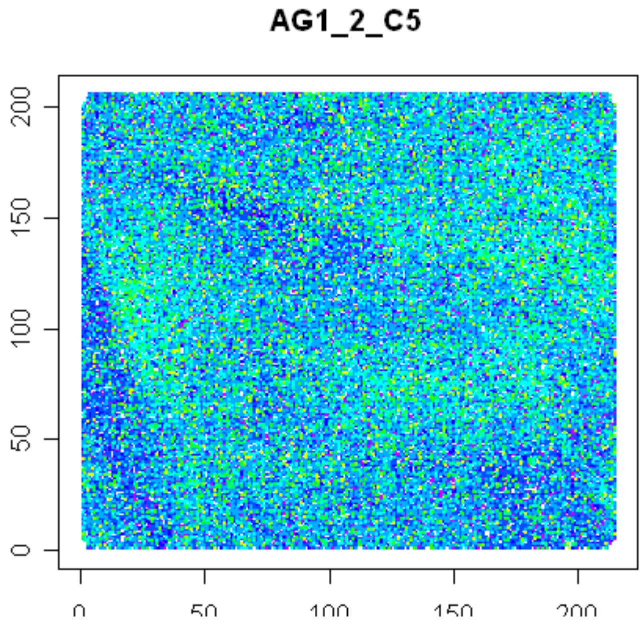


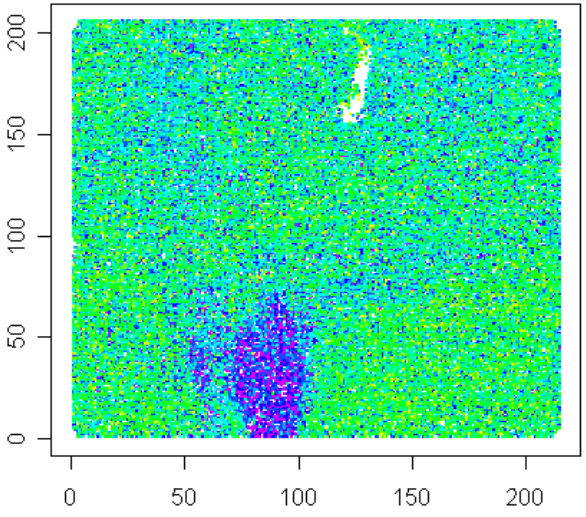




Spatial Artifacts in Agilent

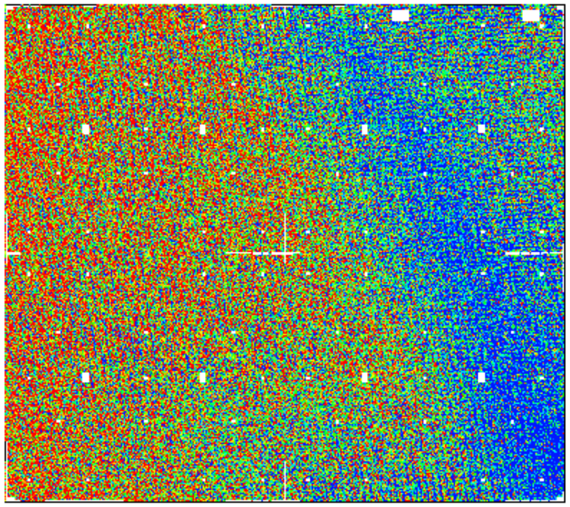
- Usually not so strong as on other array types
- More diffuse artifacts – probably reflecting washing irregularities

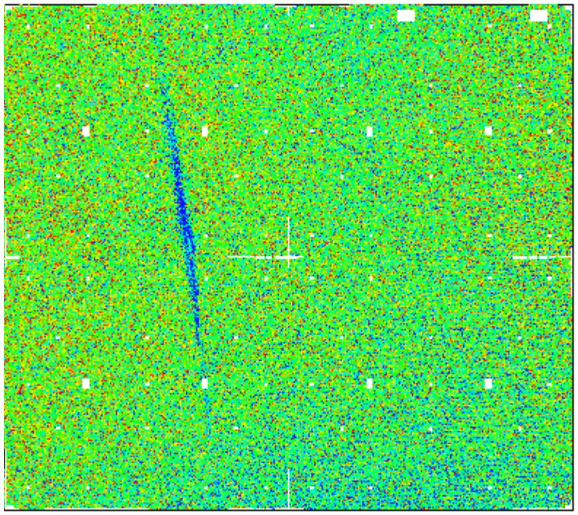




Spatial Artifacts in Nimblegen

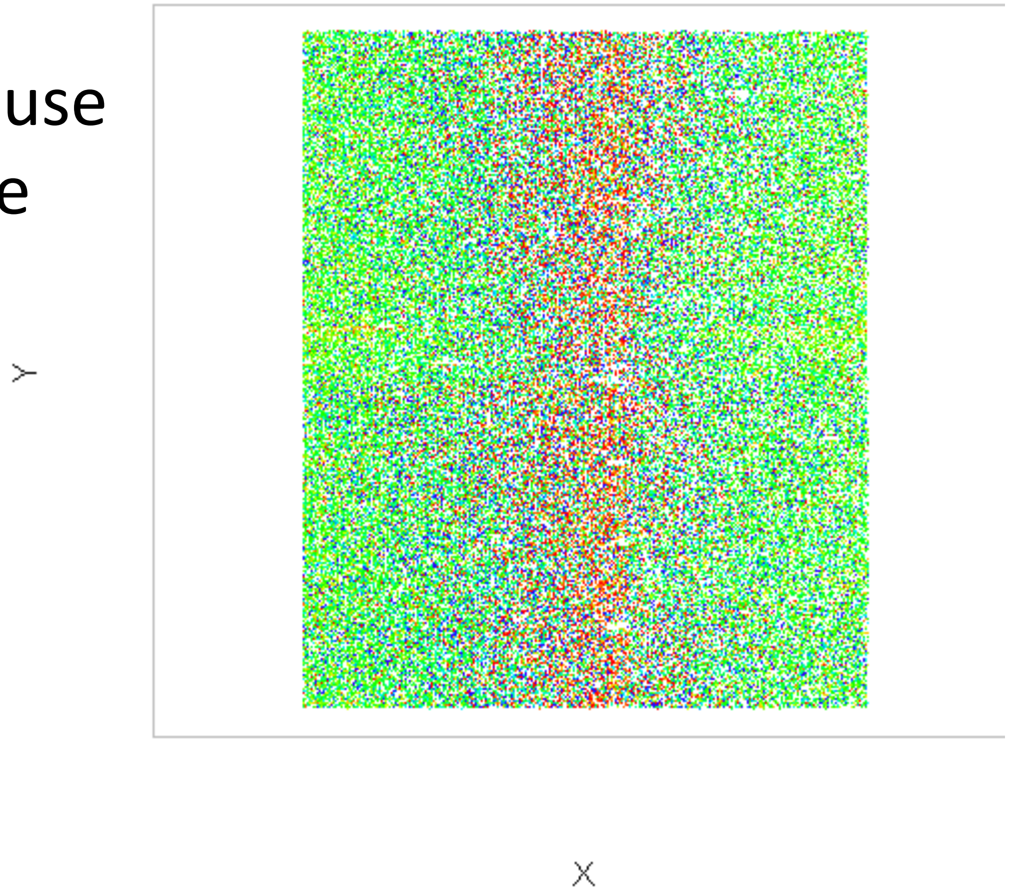
- More common than Agilent
- Usually more diffuse, probably reflecting washing
- Some sharp artifacts of unclear origin





Spatial Artifacts in Illumina Arrays

- Often bigger artifacts than Affy
- Less consequential because more beads, and all have same sequence

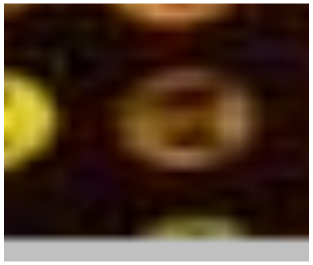


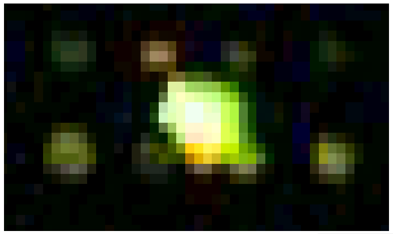
Spot QA for cDNA Spotted Arrays

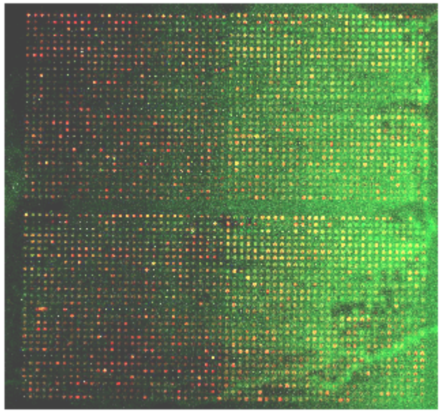
- Spot Measures
  - Uniformity
  - Spot Area

- Inspect images for artifacts

- Global Measures
  - Qualitative assessments
  - Averages of spot measures







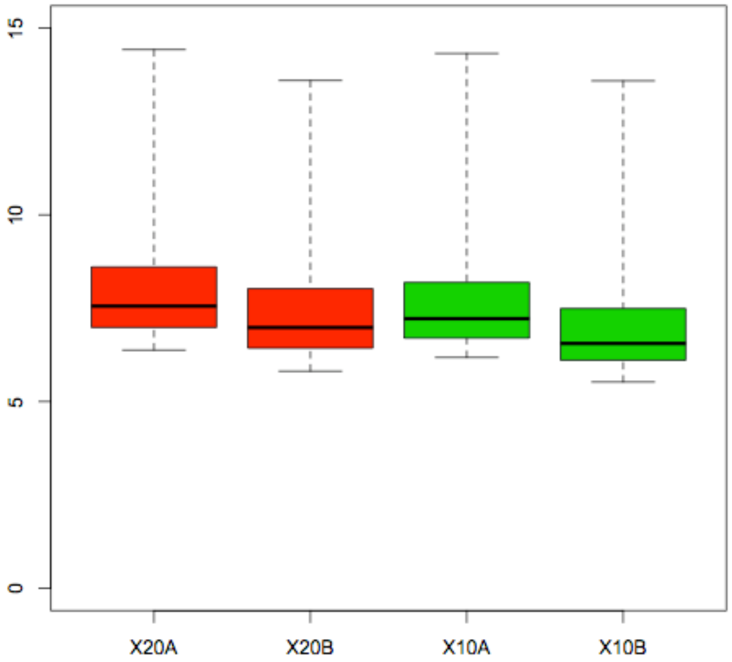
Quality Assessment for oligonucleotide Microarray

- Quality Assessment Plot
  - Box plot or Density plot
  - MA plot
  - …
- Quality Assessment Metric
  - 3'/5' ratio
  - Covariation with Probe Position

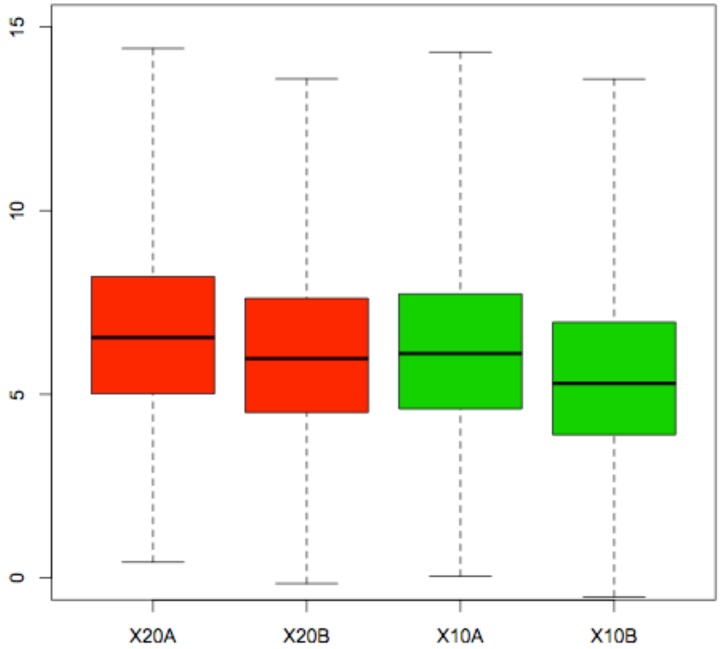
MAS 5.0: Background correction

- Under R:

```
> a=bg.correct (Dilution, method=“mas”)
```



Before

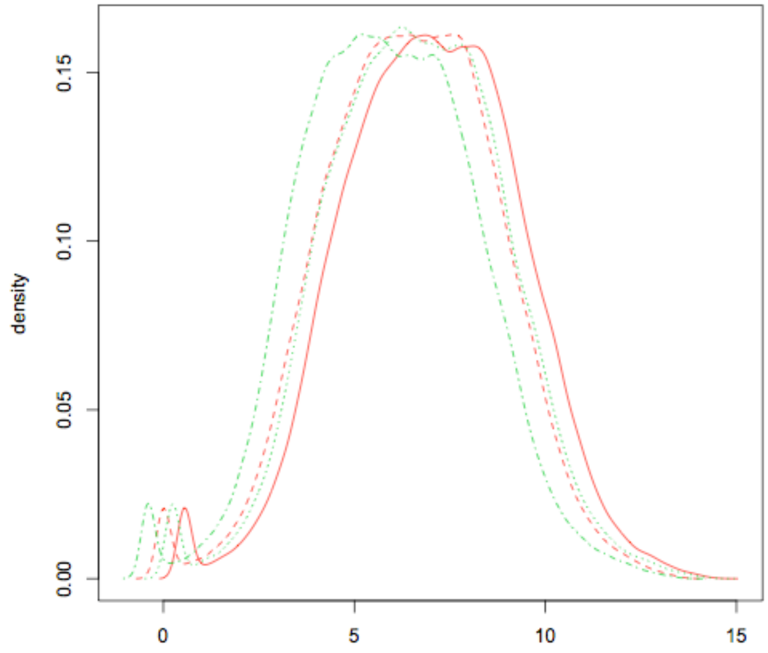


After

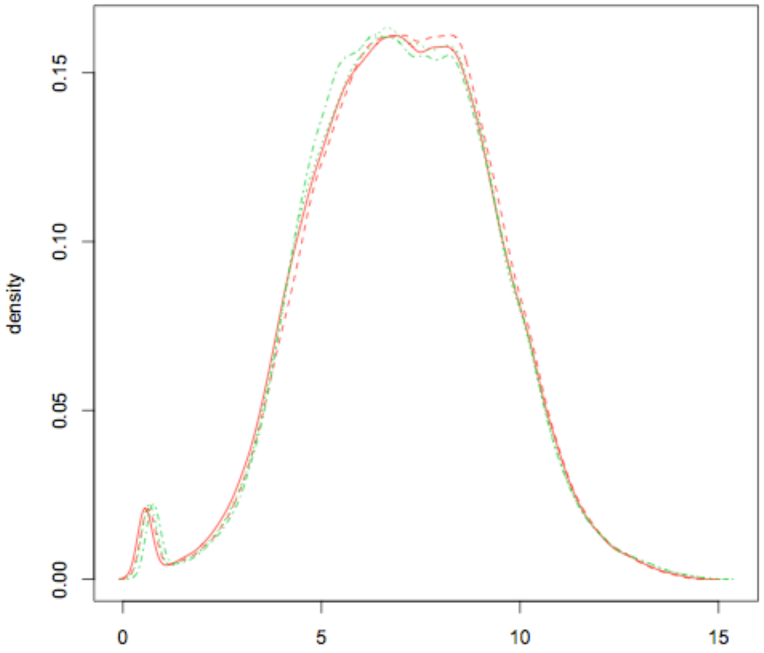
MAS 5.0: normalization

- Under R:

```
> b=normalize (Dilution, method=“constant”)
```



Before

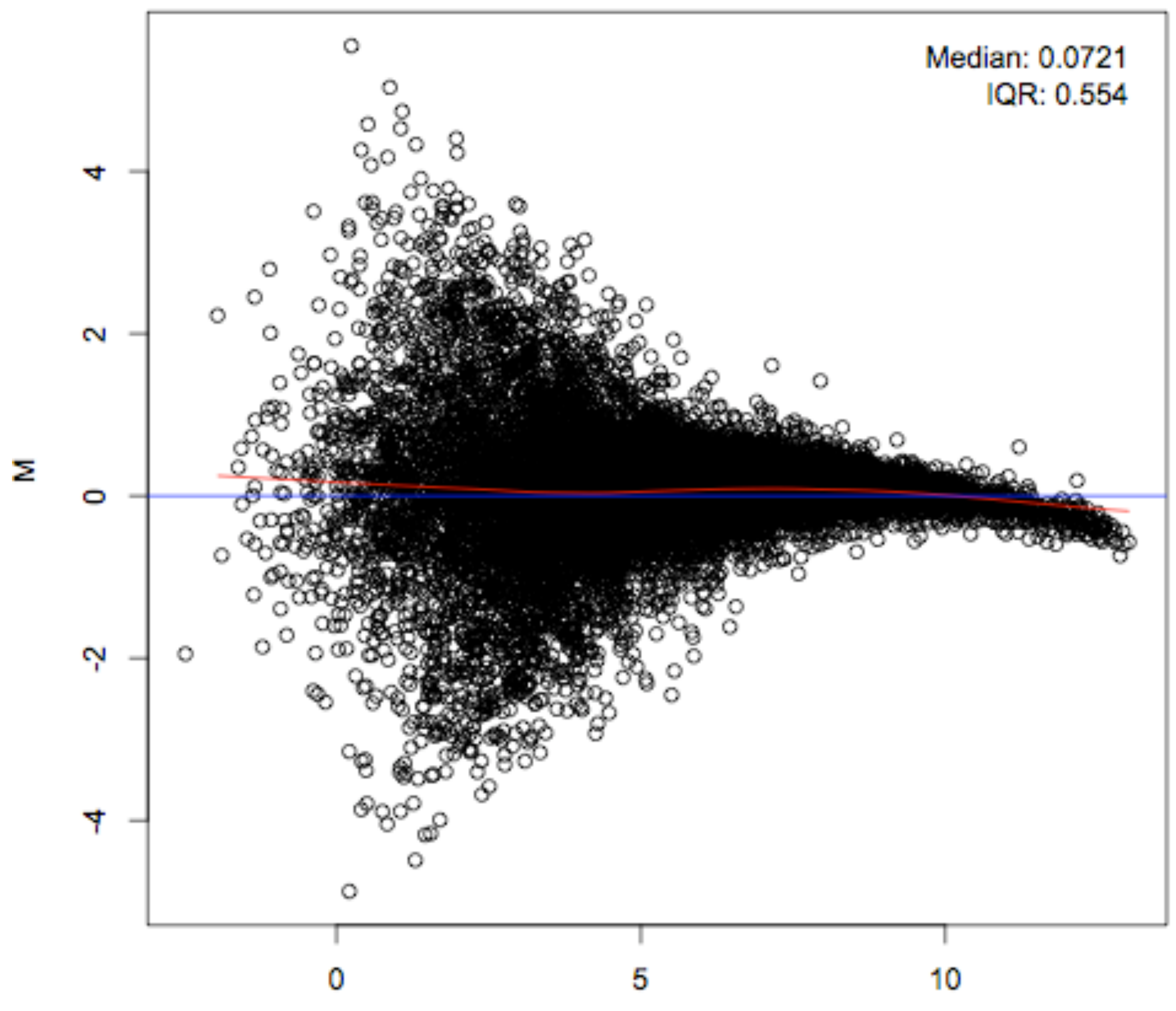


After

MA plot - Ratio vs Intensity Plots

- "M" is the **log2 intensity ratio** for a probe in the two chips
- "A" is the average log2 intensity for a probe in the two chips
- The MA plot gives a quick overview of the distribution of the data.
- The general assumption is that most of the genes would not see any change in their expression.
- Therefore the majority of the points on the y axis (M) would be located at 0, since Log(1) is 0.

MA plot

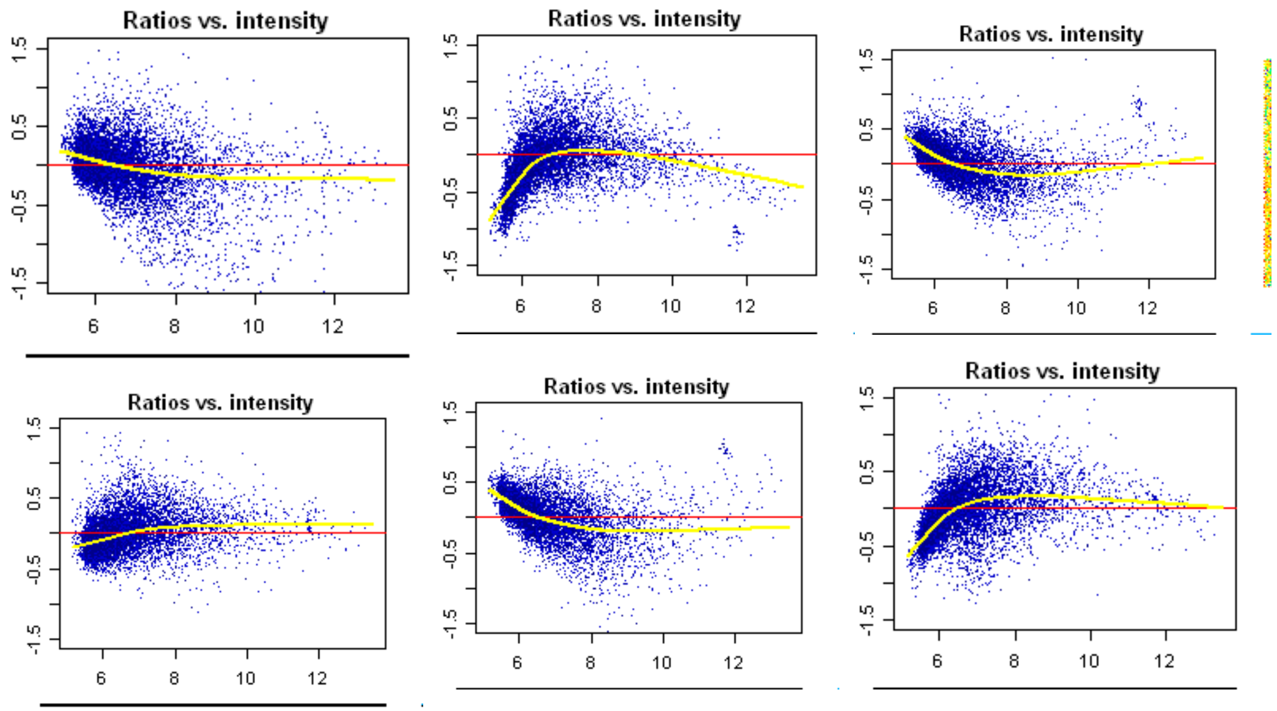


The general assumption is that most of the genes would not see any change in their expression.

MA Plots: Saturation & Quenching

- Saturation
  - Decreasing rate of binding of RNA at higher occupancies on probe
- Quenching:
  - Light emitted by one dye molecule may be re-absorbed by a nearby dye molecule
  - Then lost as heat
  - Effect proportional to square of density

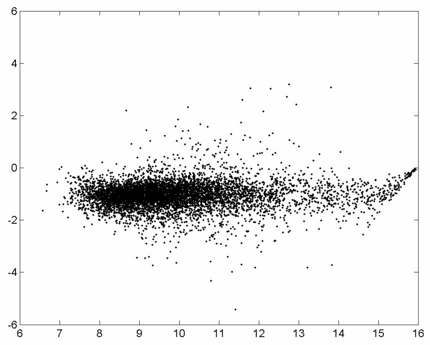
How Much Variability on MA?



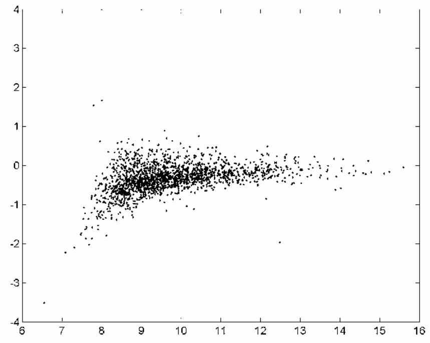
- MAplots for six arrays at random from Cheung et al ***Nature*** (2005)

Common problems diagnosed using MA-plots

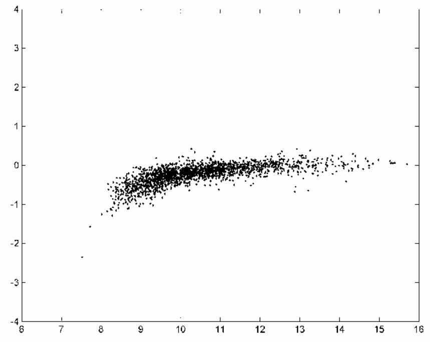
Saturation



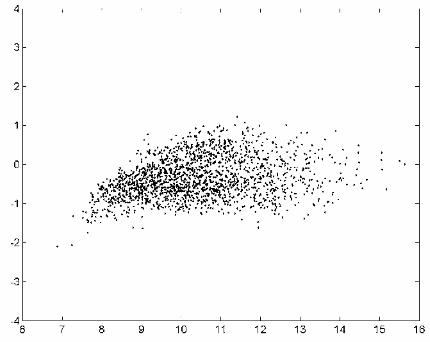
Low end variation



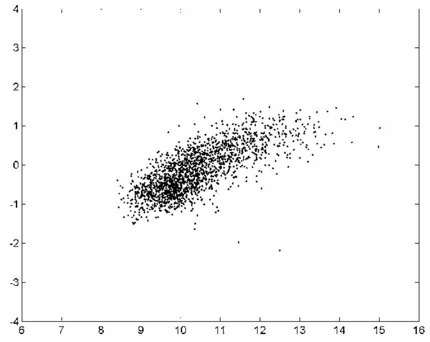
Curvature at Low intensity



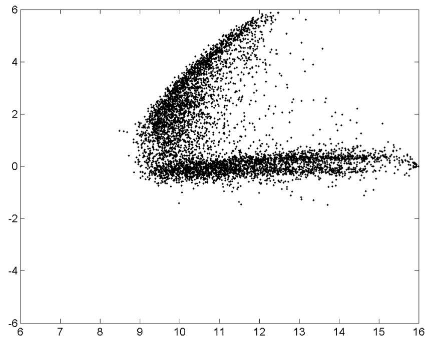
High end variation



Large curvature



Heterogeneity



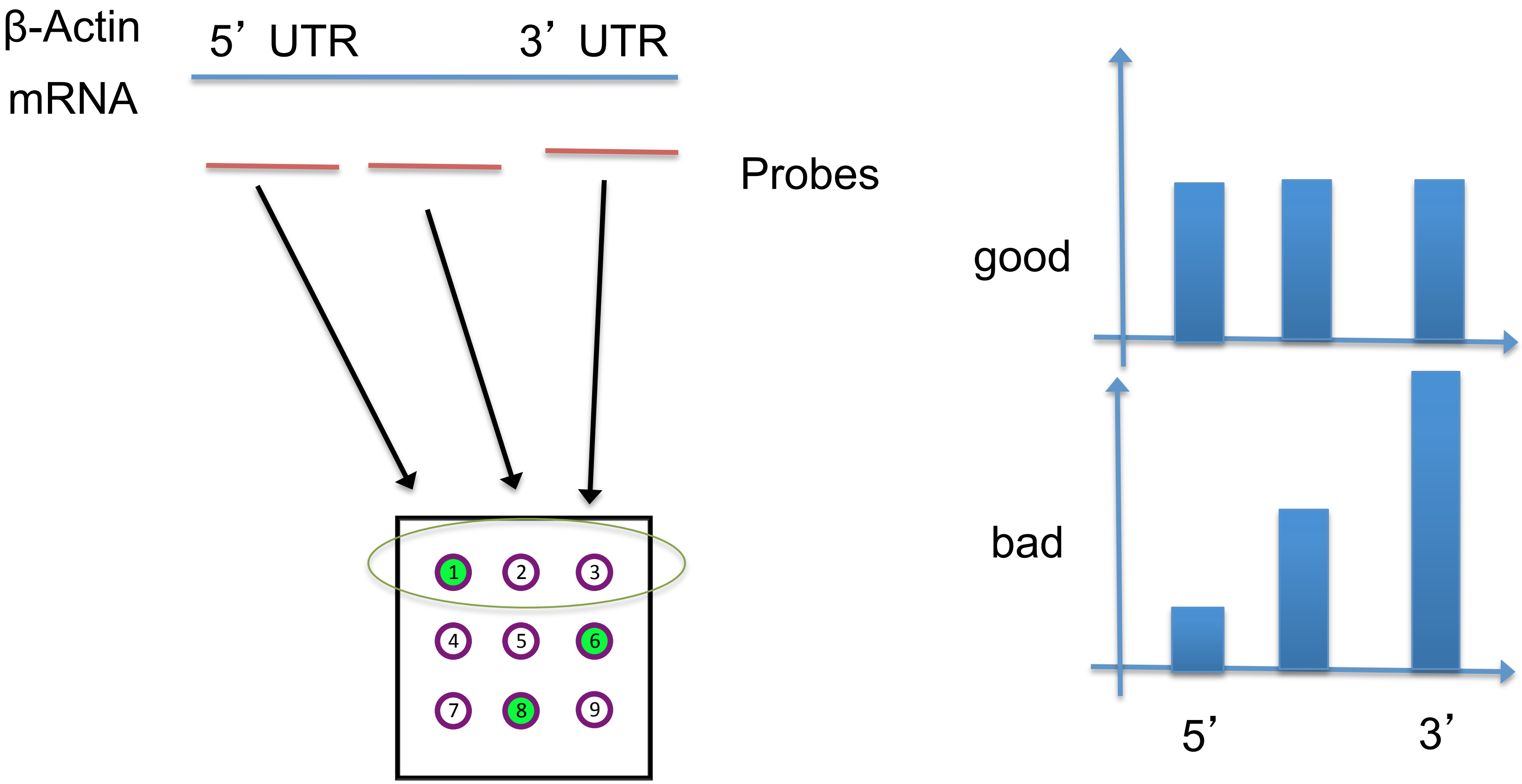
MA plot

```
> y <- (exprs(Dilution)[, c("20B", "10A")])

> ma.plot( rowMeans(log2(y)), log2(y[, 1]/y[, 2]),
cex=1 )

> title(“Pre-Norm Dilutions Dataset (array 20B v
10A)")
```

3'/5' ratio: RNA quality

- The assumption is that RNA degradation, or problems during labeling, can lead to under intensity representation at the 5′ end of RNA, allowing the ratio between signals from 5′ and 3′ probesets to be used to assess RNA quality and labeling.
- Affymetrix genechip include a few RNA quality genes, each represented by 3 probe-sets, one at 5' end of RNA, one at the middle, and one at the 3' end of expressed RNA.
- The intensity ratio of 3' probe-set to the 5' probe-set for these genes can be used as a measure of RNA quality (i.e., the severity of RNA degradation).

3’/5’ ratio: RNA quality



The assumption is that RNA degradation, or problems during labeling, can lead to under intensity representation at the 5′ end of RNA

3’/5’ ratio: RNA quality

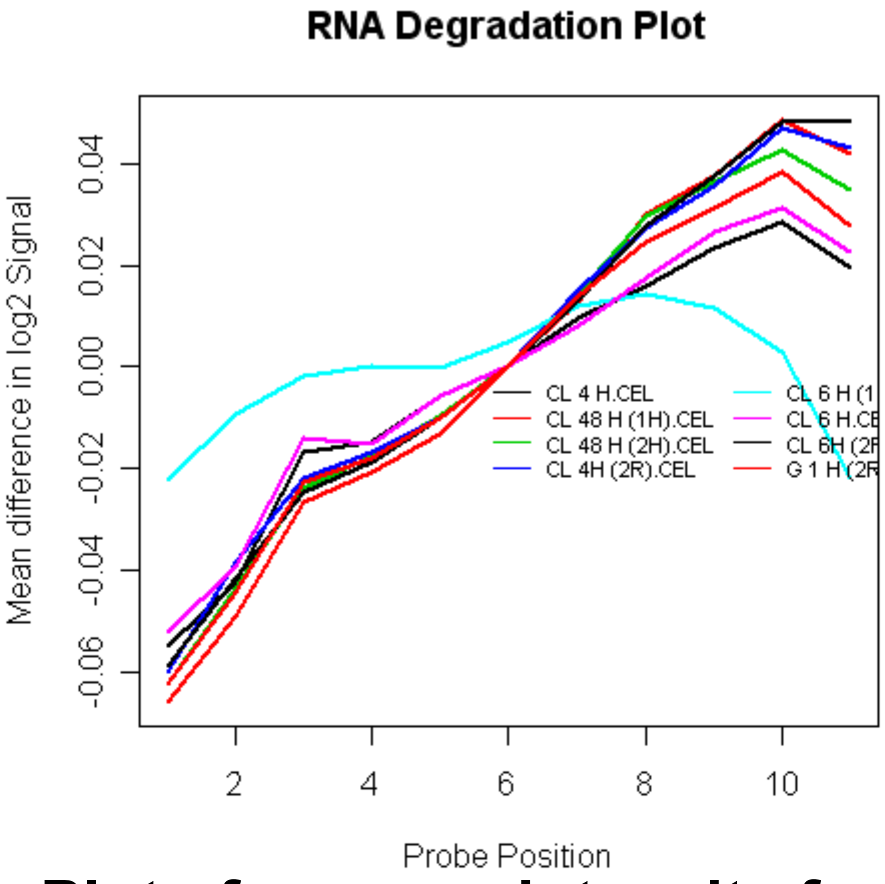
- Using Bioconductor, the “simpleaffy” package can compute these values.

```
> library("simpleaffy")
> d.qc=qc(Dilution)
> ratios(d.qc)
    actin3/actin5 actin3/actinM gapdh3/gapdh5 gapdh3/gapdhM
20A     0.6961423     0.1273385     0.4429746   -0.06024147
20B     0.7208418     0.1796231     0.3529890   -0.01366293
10A     0.8712069     0.2112914     0.4326566    0.42375270
10B     0.9313709     0.2725534     0.5726650    0.11258237
```

  - mostly β-Actin and GAPDH genes
  - 3 is the suggested safe threshold value for the 3’/5’ ratio.

Detect possible RNA degradation- Covariation with Probe Position

- RNA degrades from 5’ end
- Intensity should decrease from 3’ end uniformly across chips

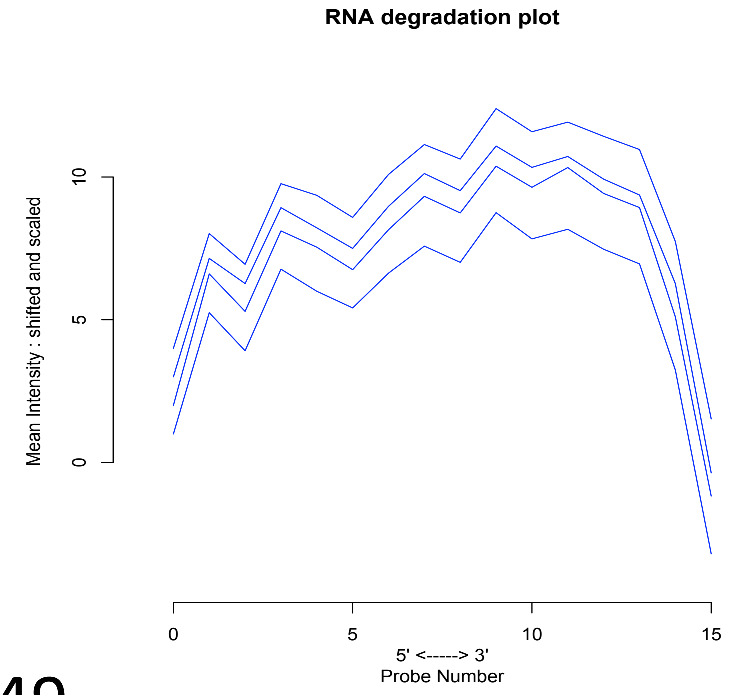


Plot of average intensity for each probe position across all genes against probe position

Covariation with Probe Position

- `AffyRNAdeg` plots in ***affy*** package

```
> RD<-AffyRNAdeg(Dilution)
> plotAffyRNAdeg(RD)


> summaryAffyRNAdeg(RD)
          20A    20B    10A    10B
slope  -0.0239 0.0363 0.0273 0.0849
pvalue  0.8920 0.8400 0.8750 0.6160
```



Homework Assignment 7

- To compare spikein133 data before and after RMA bgcorrection and normalization
- To compare MAS5() and RMA() for spikein133
- To plot MA-plot
- To plot AffyRNAdeg
- Due by March 10

Outline

- Background
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

The problem

- One the most common use of microarrays is to determine which genes are differentially expressed between pre-specified groups of samples.
- Most investigators want to use microarray to identify genes whose expression level changes across conditions under study
  - finding the genes affected by a treatment, or finding marker genes that discriminate diseased from healthy subjects

Type of problem

- To compare two groups
  - Treatment group vs. control group
- To compare multiple groups
  - Treatment A, Treatment B, Control group
- To consider multiple variables (factors) simultaneously
  - Treatment variable (Treatment vs. Control), age variables (>50 vs. <50), …

Two-group comparisons

- A typical example: To compare gene expression levels in breast tumor and normal tissues

- 6 affymetrix arrays (chips, samples) available
  - 3 independent tumor samples
  - 3 independent normal samples

Two-group comparisons

- Example Dataset: 20000 rows (genes) x 6 columns (samples)

	T 1	T 2	T 3	N 1	N 2	N 3
G 1						
G 2						
…						
G 20000						

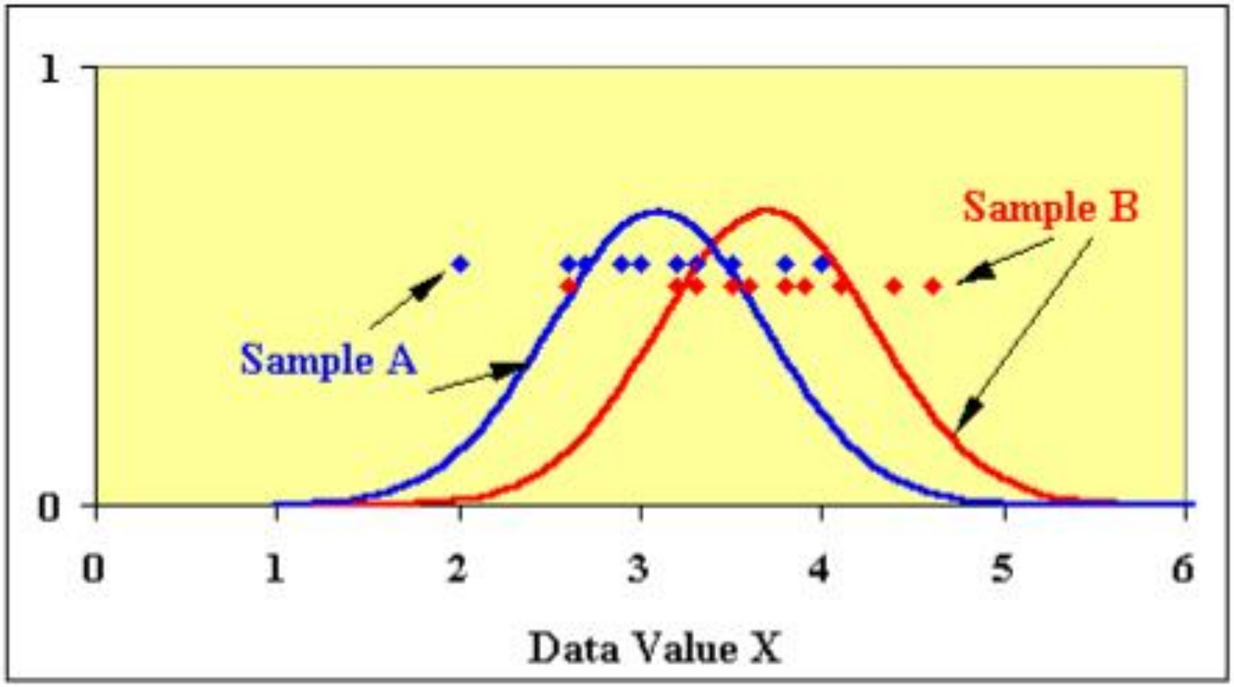
- Charactistics of dataset: many genes, only a few observations (chips, arrays, or samples) per gene.
- To find out which genes are differentially expressed, we need statistical analysis (e.g., applying separate statistical test for each gene).

Statistics methods for two-group comparisons

- T-test
  - Student' s t-test: assumes normally distributed data in each group, equal variance within groups
  - Welch t-test: as above, but allows unequal variance
- Univariate Linear model
- Nonparametric test
  - Wilcoxon, or rank-sums test: non-parametric, rank-based
  - Permutation test: estimate the distribution of the test statistics under the null hypothesis by permutations of the sample labels

Student' s T-test

– Student' s t-test: assumes normally distributed data in each group, equal variance within groups, fit the data.



Student’ s T-test

- Hypotheses:

$$H_0 : u_T = u_N \qquad H_a : u_T \neq u_N$$

- Test statistics:

$$T = \frac{\overline{Y_T} - \overline{Y_N}}{S} \qquad S = \sqrt{\frac{S_T^2}{n_T} + \frac{S_N^2}{n_N}}$$

Student’s T-test

- The standard error in this dataset is

$$S = \sqrt{\frac{S_T^2}{n_T} + \frac{S_N^2}{n_N}} \quad \text{where}$$

$$S_T^2 = \frac{1}{n_T - 1} \sum_{i=1}^{n_T} (Y_i - \overline{Y}_T)^2$$

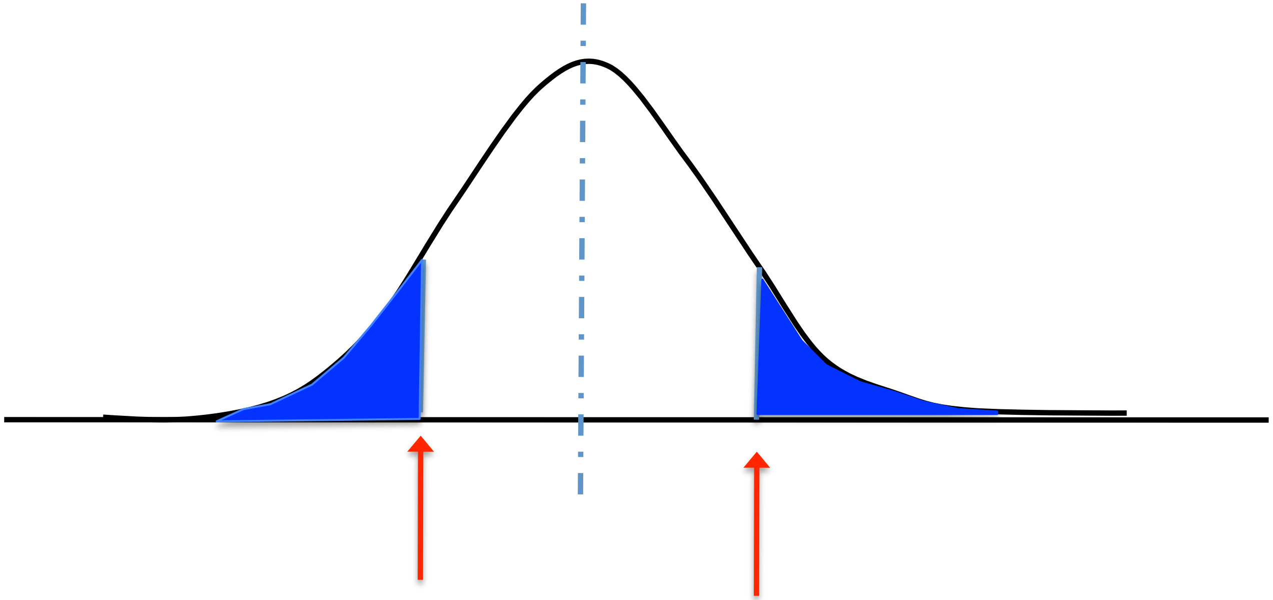
$$S_N^2 = \frac{1}{n_N - 1} \sum_{i=1}^{n_N} (Y_i - \overline{Y}_N)^2$$

Student' s T-test

- The null hypothesis H_0 is rejected when

$$\mathbf{p} = 2\mathbf{P}\{\mathbf{T}_\mathbf{i} \geq |\mathbf{T}|\}$$

it is smaller than a specified threshold (e.g., 0.05)



Absolute T value

Student’ s T-test

- Group 1: expression values of P53 gene in 10 breast tumor samples

  1.71 1.70 1.63 1.34 1.60 1.63 1.80 1.72 1.49 1.62

- Group 2: expression values of P53 gene in 10 breast normal samples

  1.57 1.40 1.50 1.49 1.25 1.44 1.57 1.53 1.50 1.51

Student' s T-test

- Questions: are P53' s expressions in tumor different from those in normal tissues? *i.e.*, differentially expressed?
- The two group means and their difference:

  Y_T = 1.624 Y_N= 1.476, $Y_T - Y_N$=0.148

- The variances and the sum of their variances.

  S_T=0.13023, S_N=0.09501, S=0.05098

T=0.148/0.05098=2.903

P-value = 0.0198

Student' s T-test

- P-value is the probability of seeing a t-statistic this extreme under the null hypothesis (i.e., area in both tails of the distribution).
  - Null hypothesis: The difference in mean expression between the two groups is zero.
  - Two-sided alternative hypothesis: The difference in mean expression is non-zero.

- Three ways to get a larger t-statistic (small p-value):
  - Bigger difference in means
  - Smaller standard deviation
  - More samples

Two-group comparisons: Student’s T-test

- Using student’s t-test to compare two groups, one for each gene

	T 1	T 2	T 3	N 1	N 2	N 3
G 1						
G 2						
…						
G 20000						

Two-group comparisons: Student’s T-test

- First, for gene1, calculated the difference between group means, divided by global standard error; obtain T1 and P1

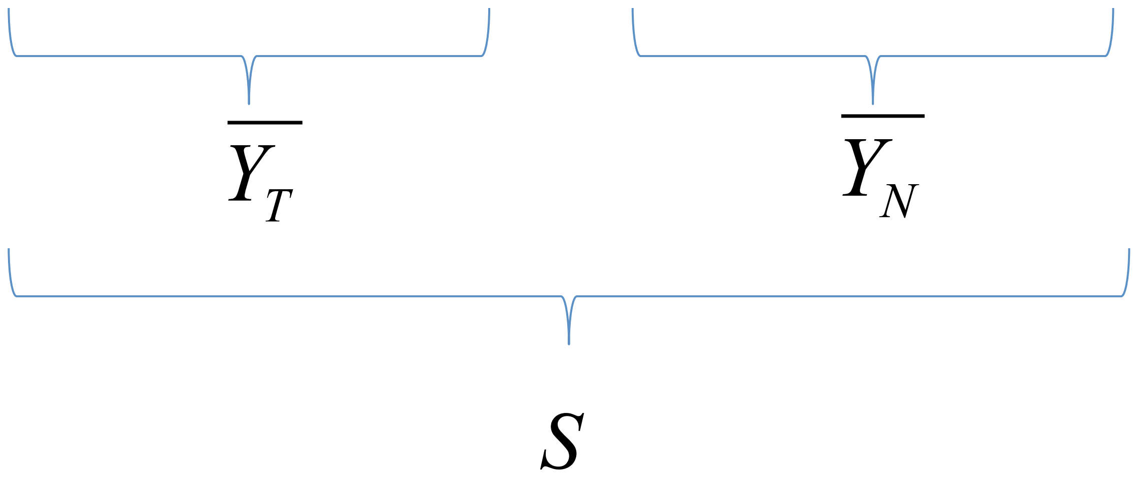
	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1	y1	y2	y3	y4	y5	y6	T1	P1
G 2								
…								
G 20000								

$$\underbrace{\underbrace{y1 \quad y2 \quad y3}_{\overline{Y}_T} \quad \underbrace{y4 \quad y5 \quad y6}_{\overline{Y}_N}}_{S}$$

Two-group comparisons: Student’s T-test

- Then, for gene2, calculated the difference between group means, divided by global standard error; obtain T2 and P2

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1							T1	P2
G 2	y1	y2	y3	y4	y5	y6	T2	p2
…								
G 20000								



Two-group comparisons: Student’s T-test

- Successively until the last gene, calculated the difference between group means, divided by global standard error; obtain T20000 and P2000

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1							T1	P2
G 2							T2	p2
…							…	….
G 20000	y1	y2	y3	y4	y5	y6	T20000	P20000

$$\underbrace{\underbrace{y1 \quad y2 \quad y3}_{\overline{Y}_T} \quad \underbrace{y4 \quad y5 \quad y6}_{\overline{Y}_N}}_{S}$$

Two-group comparisons: Student' s T-test

- The result is that we obtain one p-value for each gene

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1							T1	P2
G 2							T2	p2
…							…	….
G 20000							T20000	P20000

Problems of student' s T-test

- A drawback of the standard t-statistic for microarray datasets is that most experiments have only a few samples in each group (**n1** and **n2** are small), and so the standard error $\mathbf{s_i}$ is not very reliable.

- In a modest fraction of cases, $\mathbf{s_i}$ could be greatly under-estimated, and genes that are little changed give rise to extreme t-values, and therefore false positives.

Other t statistics

- Moderated t-statistics (G smith 2004, Limma package)

$$T^* = \frac{\overline{Y_T} - \overline{Y_N}}{\tilde{S}/\sqrt{n}}$$

where $\tilde{S}^2 = \frac{S^2 d + S_0^2 d_0}{d + d_0}$ is the shrinkage estimate of standard deviation

- Basically, it uses a way of Empirical Bayes to estimate the standard deviations by looking at all genes simultaneously

Problems of student' s T-test

- Student' s t-test assumes the data are normally distributed.
- However, the normality assumption might be violated in microarray study, especially when the sample size is small.
- To test if the data is from a normal distribution, you can use shapiro-wilks normality test.
  - Under R: > shapiro.test(x)

Two-group comparisons: Student' s T-test

- When the microarray data do not follow a normal distribution, we can use non-parametric tests to replace t-test.
- When the two groups are independent, we can use Wilcoxon Rank-sums test (Mann-Whitney test)
- When the two groups are paired, we can use wilcoxon signed-rank test
- We will mention these methods later.

Statistics methods for two-group comparisons

- T-test
  - Student' s t-test: assumes normally distributed data in each group, equal variance within groups
  - Welch t-test: as above, but allows unequal vairance
- Univariate linear model
- Nonparametric test
  - Wilcoxon, or rank-sums test: non-parametric, rank-based
  - Permutation test: estimate the distribution of the test statistics under the null hypothesis by permutations of the sample labels