Normalization Methods for Two-Color Microarray Data

1/13/2009

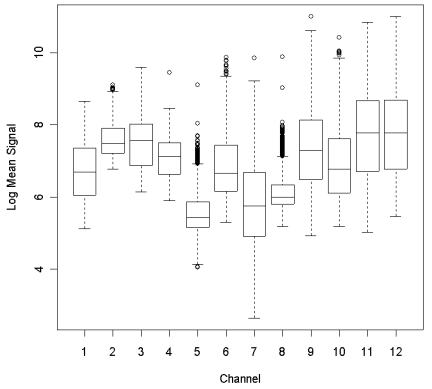
What is Normalization?

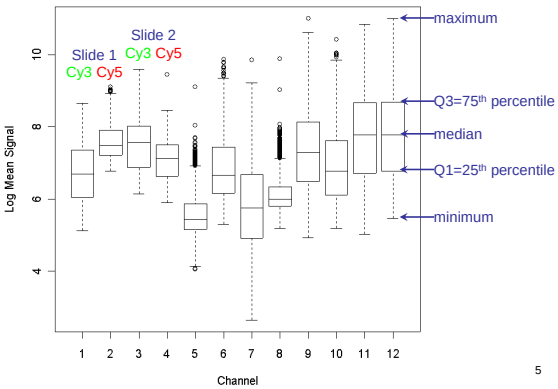
- Normalization describes the process of removing (or minimizing) non-biological variation in measured signal intensity levels so that biological differences in gene expression can be appropriately detected.
- Typically normalization attempts to remove global effects, i.e., effects that can be seen by examining plots that show all the data for a slide or slides.
- Normalization does not necessarily have anything to do with the normal distribution that plays a prominent role in statistics.

Sources of Non-Biological Variation

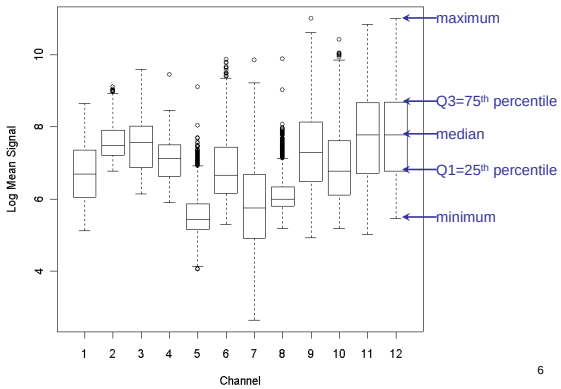
- Dye bias: differences in heat and light sensitivity, efficiency of dye incorporation
- Differences in the amount of labeled cDNA hybridized to each channel in a microarray experiment (here *channel* is used to refer to a particular slide/dye combination.)
- Variation across replicate slides
- Variation across hybridization conditions
- Variation in scanning conditions
- Variation among technicians doing the lab work
- etc.

Side-by-side boxplots show examples of variation across channels.





Interquartile range (IQR) is Q3-Q1. Points more than 1.5*IQR above Q3 or more than 1.5*IQR below Q1 are displayed individually.



The side-by-side boxplots were produced in R using the following commands.

```
boxplot(as.data.frame(log(x)),
        xlab="Channel",ylab="Log Mean Signal",
        axes=F)
axis(1,labels=1:ncol(x),at=1:ncol(x))
axis(2)
box()
```

x is a matrix with one column for each channel. Element i,j of the matrix is the signal mean for the i^{th} gene on the j^{th} channel.

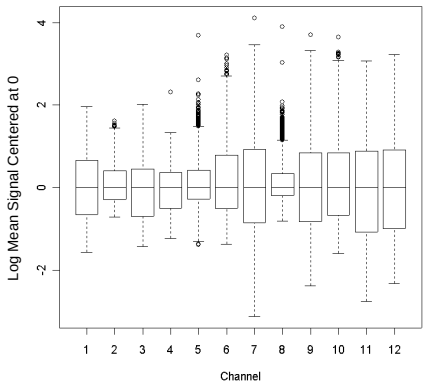
If the matrix x has other columns that you don't want to deal with, you may pick out the columns that you want or delete those you don't want. For example, x[,c(1,2,3,6)] (only work with columns 1,2,3 and 6) or x[,-1] (all columns except the first column).

7

One of the simplest normalization strategies is to align the log signals so that all channels have the same median.

- The value of the common median is not important for subsequent analyses.
- A convenient choice is zero so that positive or negative values reflect signals above or below the median for a particular channel.
- If negative normalized signal values seem confusing, any positive constant may be added to all values after normalization to zero medians.

8



9

Normalization to a median of 0 can be accomplished with the following R commands.

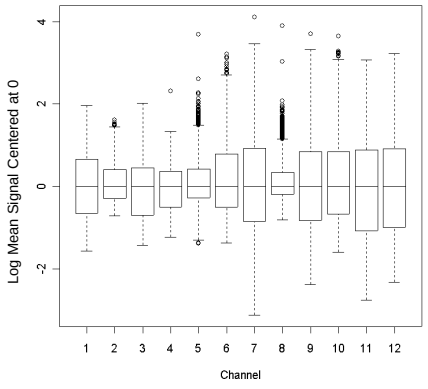
```
channel.medians=apply(log(x),2,median)

normalized.log.x=sweep(log(x),2,channel.medians)
```

x is a matrix with one column for each channel. Element i,j of the matrix is the signal mean for the i^{th} gene on the j^{th} channel.

If the matrix x has other columns that you don't want to deal with, you may pick out the columns that you want or delete those you don't want. For example, x[,c(1,2,3,6)] (only work with columns 1,2,3 and 6) or x[,-1] (all columns except the first column).

10

Note that medians match but variation seems to differ greatly across channels.



11

Yang, et al. (2002. *Nucliec Acids Research*, **30**, 4 e15) recommend scale normalization.*

Consider a matrix X with i=1,...,I rows and j=1,...,J columns.

Let x_{ij} denote the entry in row i and column j.

We will apply scale normalization to the matrix of log signal mean values that have already been median centered (each row corresponds to a gene and each column corresponds to a channel).

For each column j, let m_j=median(x_{1j}, x_{2j}, ..., x_{Ij}).

For each column j, let MAD_j=median($|x_{1j}-m_j|$,$|x_{2j}-m_j|$,...,$|x_{Ij}-m_j|$).

To scale normalize the columns of X to a constant value C, multiply all the entries in the j^{th} column by C/MAD_j for all j=1,...,J.

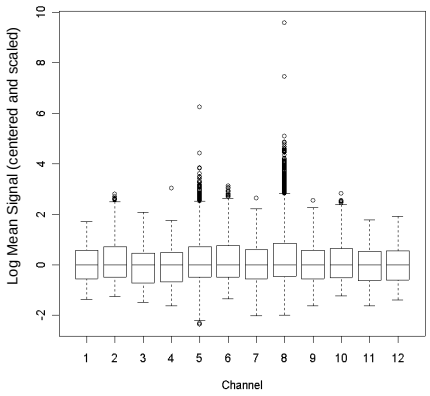
A common choice for C is the geometric mean of $MAD_1,...,MAD_J = \left(\prod_{j=1}^{J} MAD_j\right)^{1/J}$

The choice of C will not effect subsequent tests or *p*-values but will affect fold change calculations.

*Yang et al. recommended scale normalization for log R/G values.

12

Data after Median Centering and Scale Normalizing



13

Scale normalization can be accomplished with the following R commands.

```
medians=apply(X,2,median)
Y=sweep(X,2,medians)
mad=apply(abs(Y),2,median)
const=prod(mad)^(1/length(mad))
scale.normalized.X=sweep(X,2,const/mad,"*")
```

X is a matrix of logged (and usually median-centered) signal mean values. Element i,j of the matrix corresponds to the i^{th} gene on the j^{th} channel.

14

A Simple Example

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	2	7	15
3	3	6	5	8
4	1	5	2	9
5	9	13	6	11

15

Determine Channel Medians

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	2	7	15
3	3	6	5	8
4	1	5	2	9
5	9	13	6	11
medians	7	6	6	11

16

Subtract Channel Medians

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	9	3	2
2	0	-4	1	4
3	-4	0	-1	-3
4	-6	-1	-4	-2
5	2	7	0	0

This is the data after median centering.

17

Find Median Absolute Deviations

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	9	3	2
2	0	-4	1	4
3	-4	0	-1	-3
4	-6	-1	-4	-2
5	2	7	0	0
MAD	**2**	**4**	**1**	**2**

18

Find Scaling Constant

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	9	3	2
2	0	-4	1	4
3	-4	0	-1	-3
4	-6	-1	-4	-2
5	2	7	0	0
MAD	2	4	1	2

$C = (2*4*1*2)^{1/4} = 2$

19

Find Scaling Factors

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	9	3	2
2	0	-4	1	4
3	-4	0	-1	-3
4	-6	-1	-4	-2
5	2	7	0	0
Scaling Factors	$\frac{2}{2}$	$\frac{2}{4}$	$\frac{2}{1}$	$\frac{2}{2}$

20

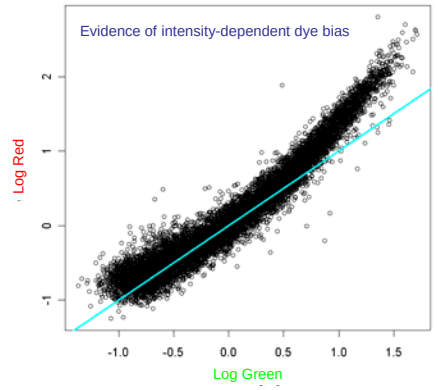
Scale Normalize the Median Centered Data

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	4.5	6	2
2	0	-2.0	2	4
3	-4	0.0	-2	-3
4	-6	-0.5	-8	-2
5	2	3.5	0	0

This is the data after median centering and scale normalizing.

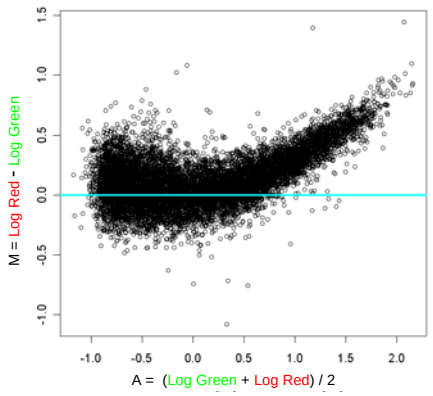
21

Slide 1 Log Signal Means after Median Centering and Scaling All Channels



22

M vs. A Plot of the Logged, Centered, and Scaled Slide 1 Data



23

To handle intensity-dependent dye bias, Yang, et al. (2002. *Nucliec Acids Research*, **30**, 4 e15) recommend "lowess" normalization prior to median centering and scale normalizing.

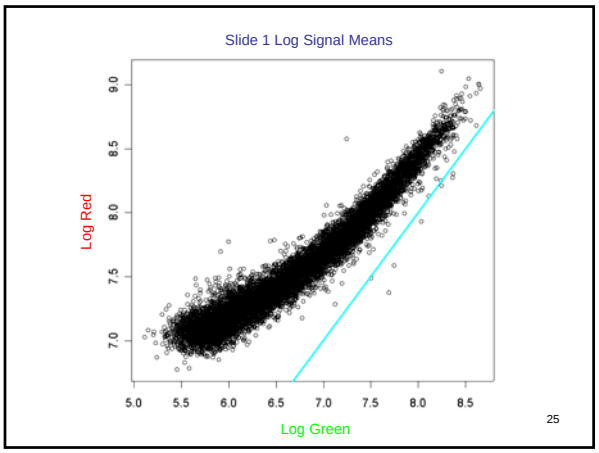
"lowess" stands for

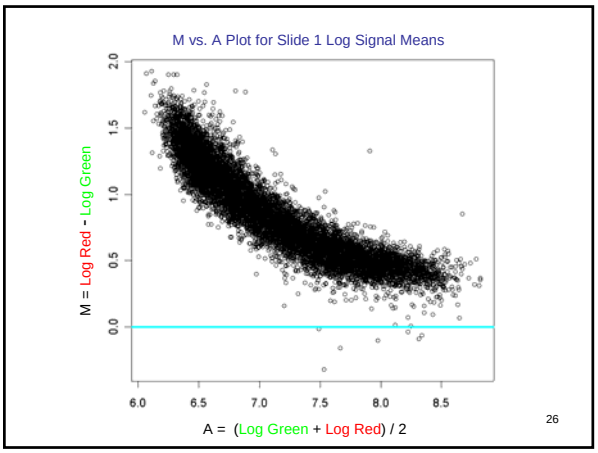
LOcally WEighted polynomial regreSSion.

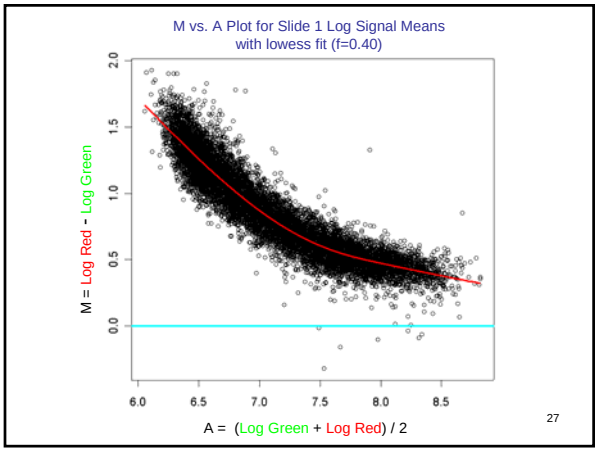
The original reference for lowess is

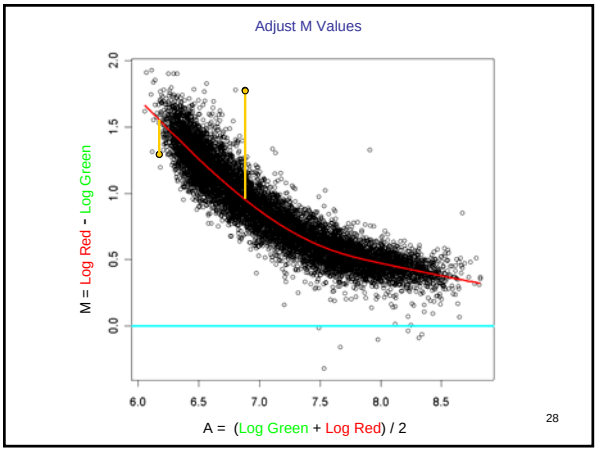
Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. JASA **74** 829-836.

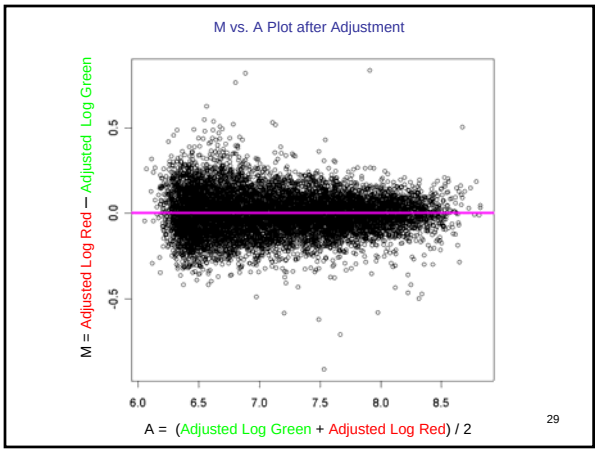
24

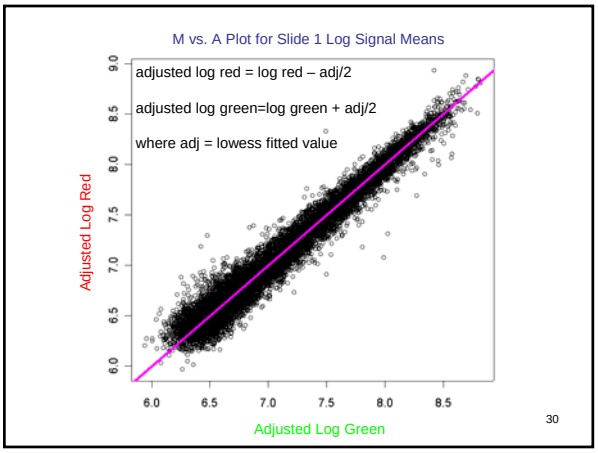

Slide 1 Log Signal Means
Log Red
Log Green
25


M vs. A Plot for Slide 1 Log Signal Means
M = Log Red - Log Green
A = (Log Green + Log Red) / 2
26


M vs. A Plot for Slide 1 Log Signal Means
with lowess fit (f=0.40)
M = Log Red - Log Green
A = (Log Green + Log Red) / 2
27

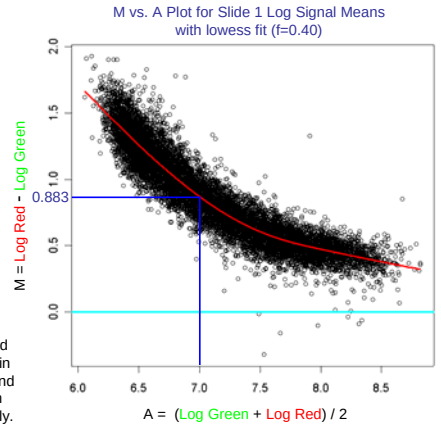

Adjust M Values
M = Log Red - Log Green
A = (Log Green + Log Red) / 2
28


M vs. A Plot after Adjustment
M = Adjusted Log Red – Adjusted Log Green
A = (Adjusted Log Green + Adjusted Log Red) / 2
29


M vs. A Plot for Slide 1 Log Signal Means
adjusted log red = log red – adj/2
adjusted log green=log green + adj/2
where adj = lowess fitted value
Adjusted Log Red
Adjusted Log Green
30

For spots with A=7, the lowess fitted value is 0.883. Thus the value of *adj* discussed on the previous slide is 0.883 for spots with A=7.

The M value for such spots would be moved down by 0.883. The log red value would be decreased by 0.883/2 and the log green value would be increased by 0.883/2 to obtain adjusted log red and adjusted log green values, respectively.



lowess in R

```
out=lowess(x,y,f=0.4)
plot(x,y)
lines(out$x,out$y,col=2,lwd=2)
```

out$x will be a vector containing the x values.

out$y will contain the lowess fitted values for the values in out$x.

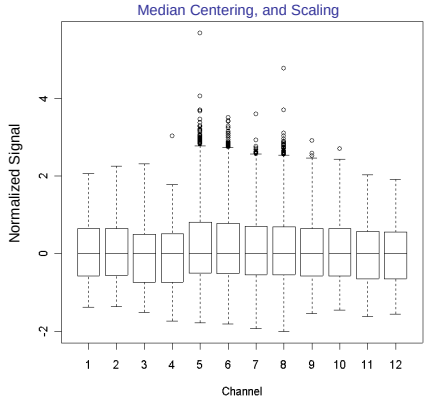
f controls the fraction of the data used to obtain each fitted value.

f = 0.4 has been recommended for microarray data normalization.

32

After a separate lowess normalization for each slide, the adjusted values can be median centered and (if deemed necessary) scale normalized across all channels using the lowess-normalized data for each channel.

33



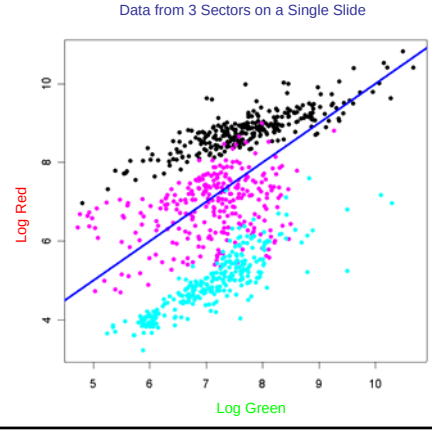
34

After a separate lowess normalization for each slide, the adjusted values can be median centered and scale normalized across all channels using the lowess-normalized data for each channel.

A *sector* represents the set of points spotted by a single pin on a single slide. The entire normalization process described above can be carried out separately for each sector on each channel.

It may be necessary to normalize by sector/channel combinations if spatial variability is apparent.

35

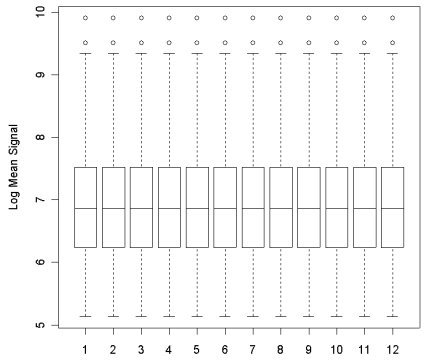


36

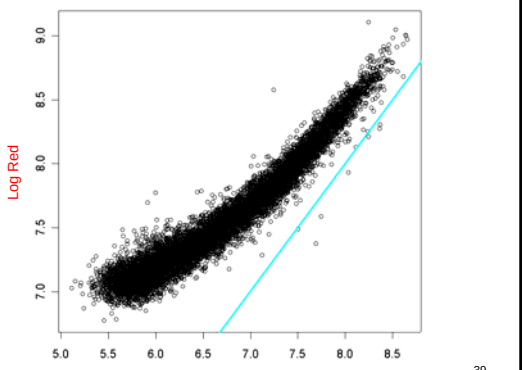
Bolstad, et al. (2003, *Bioinformatics* 19 2:185-193) propose *quantile normalization* for microarray data

- Quantile normalization is most commonly used in normalization of Affymetrix data
- It can be used for two-color data as well.
- Quantile normalization can force each channel to have the same quantiles.
- x_q (for q between 0 and 1) is the q quantile of a data set if the fraction of the data points less than or equal to x_q is at least q, and the fraction of the data points greater than or equal to x_q at least 1-q.
- median=$x_{0.5}$ Q1=$x_{0.25}$ Q3=$x_{0.75}$

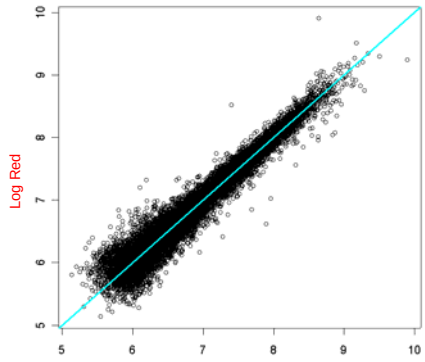
37



38



39



40

Details of Quantile Normalization

1. Find the smallest log signal on each channel.
2. Average the values from step 1.
3. Replace each value in step 1 with the average computed in step 2.
4. Repeat steps 1 through 3 for the second smallest values, third smallest values,..., largest values.

41

A Simple Example

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	2	7	15
3	3	6	5	8
4	1	5	2	9
5	9	13	6	11

42

Find the Smallest Value for Each Channel

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	[2]	7	15
3	3	6	5	[8]
4	[1]	5	[2]	9
5	9	13	6	11

43

Average These Values

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	[2]	7	15
3	3	6	5	[8]
4	[1]	5	[2]	9
5	9	13	6	11

(1+2+2+8)/4=3.25

44

Replace Each Value by the Average

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	3.25	7	15
3	3	6	5	3.25
4	3.25	5	3.25	9
5	9	13	6	11

(1+2+2+8)/4=3.25

45

Find the Next Smallest Values

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	3.25	7	15
3	[3]	6	[5]	3.25
4	3.25	[5]	3.25	[9]
5	9	13	6	11

46

Average These Values

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	3.25	7	15
3	[3]	6	[5]	3.25
4	3.25	[5]	3.25	[9]
5	9	13	6	11

(3+5+5+9)/4=5.5

47

Replace Each Value by the Average

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7	3.25	7	15
3	5.50	6	5.50	3.25
4	3.25	5.50	3.25	5.50
5	9	13	6	11

48

Find the Average of the Next Smallest Values

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	[7]	3.25	7	15
3	5.50	[6]	5.50	3.25
4	3.25	5.50	3.25	5.50
5	9	13	[6]	[11]

(7+6+6+11)/4=7.5

49

Replace Each Value by the Average

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	8	15	9	13
2	7.50	3.25	7	15
3	5.50	7.50	5.50	3.25
4	3.25	5.50	3.25	5.50
5	9	13	7.50	7.50

50

Find the Average of the Next Smallest Values

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	[8]	15	9	[13]
2	7.50	3.25	[7]	15
3	5.50	7.50	5.50	3.25
4	3.25	5.50	3.25	5.50
5	9	[13]	7.50	7.50

(8+13+7+13)/4=10.25

51

Replace Each Value by the Average

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	10.25	15	9	10.25
2	7.50	3.25	10.25	15
3	5.50	7.50	5.50	3.25
4	3.25	5.50	3.25	5.50
5	9	10.25	7.50	7.50

52

Find the Average of the Next Smallest Values

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	10.25	[15]	[9]	10.25
2	7.50	3.25	10.25	[15]
3	5.50	7.50	5.50	3.25
4	3.25	5.50	3.25	5.50
5	[9]	10.25	7.50	7.50

(9+15+9+15)/4=12.00

53

Replace Each Value by the Average

Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	10.25	12.00	12.00	10.25
2	7.50	3.25	10.25	12.00
3	5.50	7.50	5.50	3.25
4	3.25	5.50	3.25	5.50
5	12.00	10.25	7.50	7.50

This is the data matrix after quantile normalization.

54

Miscellaneous Comments on Normalization

- Data presented on previous slides are somewhat extreme. Many microarray data sets will require less normalization.
- We have only scratched the surface in terms of normalization methods. There are many variations on the techniques that were described previously as well as other approaches that we won't discuss at this point in the course.
- Normalization affects the final results, but it is often not clear what normalization strategy is best.
- It would be good to integrate normalization and statistical analysis, but it is difficult to do so. The most common approach is to normalize data and then perform statistical analysis of the normalized data as a separate step in the microarray analysis process.

55

Normalization for Specialized Arrays

- Sometimes researchers will construct an array with a set of probe sequences that represent a specialized set of genes.
- If the treatment effects are expected to cause changes of expression in the specialized set that are predominantly in one direction, the global normalization strategies that we discussed may remove the treatment effects of interest.
- One strategy for normalizing in such cases requires a set of control sequences spotted on each slide.

56

Normalization for Specialized Arrays (ctd).

For normalization purposes, good control sequences should represent genes that will not change expression in response to treatments of interest.

1. Housekeeping genes are genes involved in basic functions needed for sustenance of a cell. They are always expressed, but are they constant across conditions?
2. Random cDNA sequences can be used as a negative control (a control not expected to give biological signal).
3. cDNA sequences from an unrelated organism can be used as negative controls or positive spike-in controls (identical amounts of complementary labeled cDNAs added to each hybridized sample).

The idea is to determine the adjustment necessary to normalize the control genes and then make that same adjustment to **all** genes on the array.

57

Background Correction

- Background correction is often the very first step in microarray analysis
- Recall that *Spot signal* or simply *signal* is fluorescence intensity due to target molecules hybridized to probe sequences contained in a spot (what we would like to measure) plus background fluorescence (what we would rather not measure).
- *Background* is fluorescence that may contribute to spot pixel intensities but is not due to fluorescence from target molecules hybridized to spot probe sequences.
- The idea is to remove background fluorescence from the spot signal fluorescence because the spot signal is believed to be a sum of fluorescence due to background and fluorescence due to hybridized target cDNA.

58

Background Correction Strategies
(applied prior to logging signal intensity)

1. Subtract local background, e.g.,

signal mean – background mean

or

signal mean – background median

This can increase variation in measurements, especially for low expressing genes. Some believe that local background will overestimate the background contribution to spot fluorescence. Background fluorescence where cDNA has been spotted may be different than background where no cDNA has been spotted.

59

Background Correction Strategies
(applied prior to logging signal intensity)

2. For each spot, find the local background of the spot as well as the local backgrounds of all neighboring spots. Compute the median or mean of these local backgrounds. Subtract that summary of local backgrounds from the spot's signal.

This is similar to option 1 but can reduce some variation in background estimation.

60

Background Correction Strategies
(applied prior to logging signal intensity)

3. Find the median or mean of local backgrounds in a sector. Subtract the sector summary of local backgrounds from each signal in the sector.

4. Subtract the median or mean of blank spot signals or negative control signals in a sector from all other signals in a sector.

5. Estimate the background for each spot by fitting a model to the local background values.

61

Final Comments on Background Correction

- Subtracting background may result in a negative or zero adjusted-signal values. Such values cannot be logged. One simple approach is to replace all negative values by zero, add one to all values (whether zero or not), and log the resulting values.
- As technology improves and labs gain experience in carrying out microarray experiments, using signal with no background correction may be the best choice.

62