Preprocessing Methods for Two-Color Microarray Data

1/15/2011

Preprocessing Steps

- Background correction
- Transformation
- Normalization
- Summarization

What is background correction?

- Background correction involves an attempt to remove any portion of a raw fluorescence intensity measurement that is not attributable to fluorescence from target nucleic acid molecules hybridized to their complementary probe.
- Example sources of fluorescence other than hybridized target nucleic acid molecules include fluorescence in the microarray slide itself, fluorescence from neighboring probe spots, or fluorescence from unbound labeled nucleic acid sequences or other stray particles not washed from the slide.

What is transformation?

- Transformation refers to transforming the gene expression measures (usually after background correction).
- The most commonly used transformation is the log transformation.
- The base is irrelevant, but log base 2 is popular for microarray data.
- More complex transformations have been proposed that are linear for low values and logarithmic for high values.

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.
- Normalization does not necessarily have anything to do with the normal distribution that plays a prominent role in statistics.

Sources of Non-Biological Variation

- Variation across replicate microarray slides resulting from the manufacturing process
- Variation in the preparation of target samples
- Differences in the number of dyed target molecules hybridized for each target sample
- Dye variation: differences in heat and light sensitivity of dyes and differences in the efficiency of dye incorporation

Sources of Non-Biological Variation (continued)

- Variation across various steps in the measurement process, such as hybridization, washing, and microarray image acquisition
- Variation in laboratory conditions from day to day
- Variation among technicians doing the lab work
- etc.

7

What is summarization?

- If a gene is represented by multiple probes on a microarray, it may be desirable to combine the measures from multiple probes to obtain a single measure of the gene's expression level.
- Simply computing the mean or median is often reasonable.
- We will discuss more complex strategies in the context of Affymetrix GeneChip data.

8

Background Correction

- 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 molecules.

9

A Simple Background Correction Method

Subtract local background from the signal, e.g.,

signal mean – background mean

or

signal mean – background median

10

Drawbacks of the Simple Method

- Signal minus background may be more variable than the signal itself.
- Subtracting the background may produce a negative value that cannot be logarithmically transformed.

11

The Normal-Exponential Convolution Method

- Silver, J. D., Ritchie, M. E., Smyth, G. K. (2009). Microarray background correction: maximum likelihood estimation for the normal – exponential convolution. *Biostatistics* **2**, 352–363.
- The authors build upon an idea originally proposed for Affymetrix data by Irizarray et al. (2003) *Biostatistics* **4**, 249-264.

12

The Normal-Exponential Convolution Method

- Silver, J. D., Ritchie, M. E., Smyth, G. K. (2009). Microarray background correction: maximum likelihood estimation for the normal – exponential convolution. *Biostatistics* **2**, 352–363.
- The authors build upon an idea originally proposed for Affymetrix data by Irizarray et al. (2003) *Biostatistics* **4**, 249-264.

13

The Normal-Exponential Convolution Method

- Suppose there are n spots on a slide.
- For spot i=1,...n on any particular slide and for either dye, let D_i = spot signal$_i$ – spot background$_i$.
- Suppose $D_i = X_i + Y_i$, where

  $X_1,...,X_n$ iid Exponential(λ)

  independent of

  $Y_1,...,Y_n$ iid $N(\mu,\sigma^2)$.

14

The Normal-Exponential Convolution Method

- X_i represents the true signal.
- Y_i represents error and background that is present after subtraction of local background.
- It can be shown that $E(X_i \mid D_i)$ is

$$D_i - \mu - \lambda\sigma^2 + \frac{\sigma^2\varphi(0; D_i - \mu - \lambda\sigma^2, \sigma^2)}{1 - \Phi(0; D_i - \mu - \lambda\sigma^2, \sigma^2)}$$

  where $\varphi(\cdot; \text{mean}, \text{variance})$ and $\Phi(\cdot; \text{mean}, \text{variance})$ denote a normal density and cumulative density, respectively.

15

The Normal-Exponential Convolution Method

- Using $D_1,...,D_n$ as the observed data, find maximum likelihood estimates of μ, σ^2, and λ.
- Substitute these MLEs into $E(X_i \mid D_i)$ to get a background corrected signal.
- This entire process is repeated separately for each slide and dye combination.

16

Transformation of Background-Corrected Signals

- Silver, Ritchie, and Smyth (2009) recommend the transformation $\log_2$(BCS+50), where BCS denotes the background-corrected signal resulting from the normal-exponential convolution method.
- The addition of 50 prior to the log transformation reduces the variance of log ratios for genes with low signal intensities, i.e., rather than $\log_2$(R/G) use

  $\log_2\{(R+50)/(G+50)\}=\log_2(R+50)-\log_2(G+50)$,

  where R and G are red and green BCS, respectively.

17

Normalization

- Following background correction and transformation, the next preprocessing step is normalization.
- Recall that the goal of normalization is to reduce the impact of non-biological variation in measured signal intensity levels so that biological differences in gene expression can be appropriately detected.

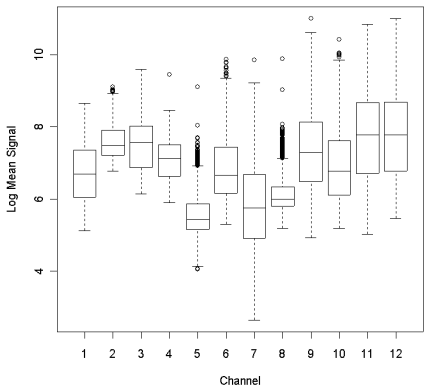
18

Normalization

- The next several slides show normalization of data from an actual microarray experiment.
- Although the normalization steps depicted would typically be carried out with $\log_2$(BCS) or $\log_2$(BCS+50) data, natural log signal means prior to background correction were used in most cases.
- The figures would look very similar if $\log_2$(BCS) or $\log_2$(BCS+50) data had been used instead of log signal means.

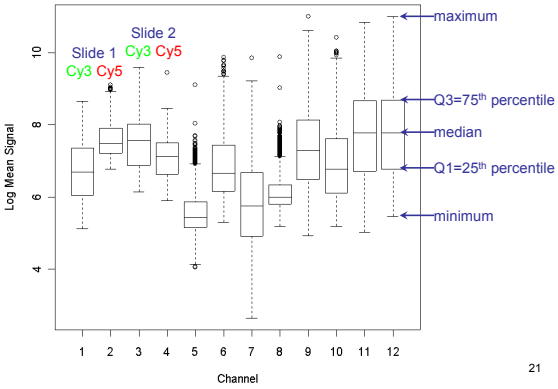
19

Side-by-side boxplots show variation across channels.



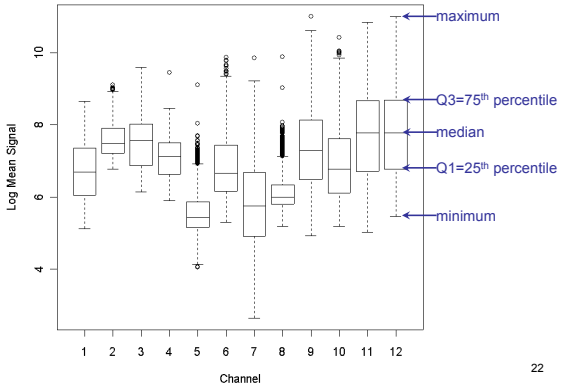
Here channel refers to a slide / dye combination.

20



21

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.

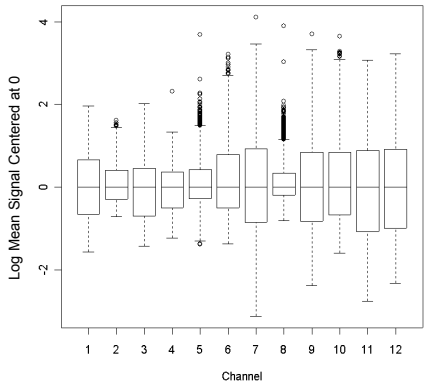


22

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

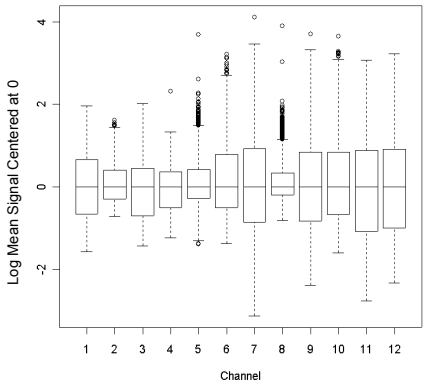
- The value of the common median is not important for subsequent analyses.
- A convenient choice is zero so that positive or negative values reflect signals above or below the median for a particular channel.
- If negative normalized signal values seem confusing, any positive constant may be added to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

23



24

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



25

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.

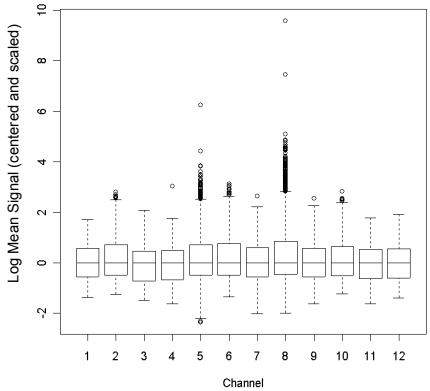
A common choice for C is the geometric mean of $MAD_1,...,MAD_J = \left(\prod_{j=1}^{J} MAD_j\right)^{1/J}$

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.

26

Data after Median Centering and Scale Normalizing



27

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

28

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

29

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.

30

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

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

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

Find Scaling Factors

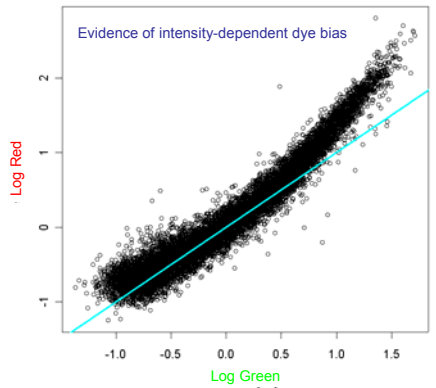
Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	9	3	2
2	0	-4	1	4
3	-4	0	-1	-3
4	-6	-1	-4	-2
5	2	7	0	0
Scaling Factors	$\frac{2}{2}$	$\frac{2}{4}$	$\frac{2}{1}$	$\frac{2}{2}$

Scale Normalize the Median Centered Data

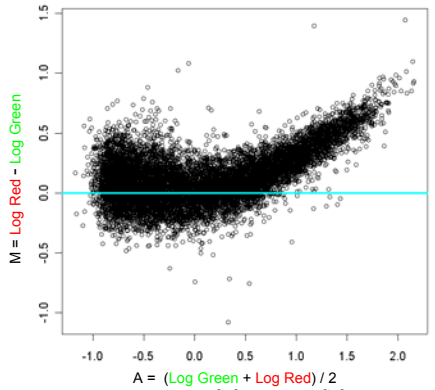
Gene	Slide1Cy3	Slide1Cy5	Slide2Cy3	Slide2Cy5
1	1	4.5	6	2
2	0	-2.0	2	4
3	-4	0.0	-2	-3
4	-6	-0.5	-8	-2
5	2	3.5	0	0

This is the data after median centering and scale normalizing.

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



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



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.

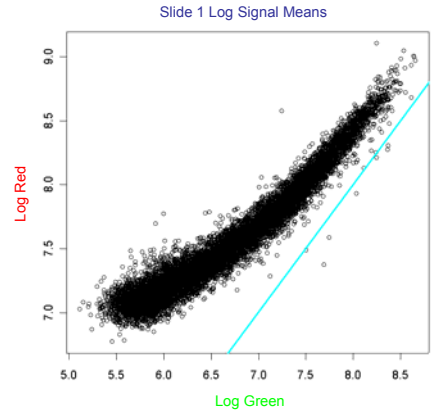
"lowess" stands for

LOcally WEighted polynomial regreSSion.

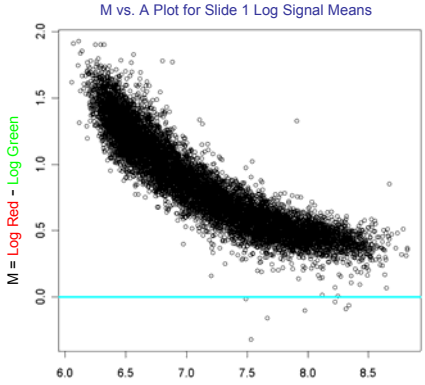
The original reference for lowess is

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

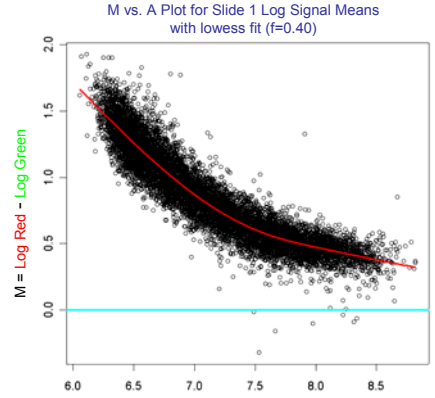
37



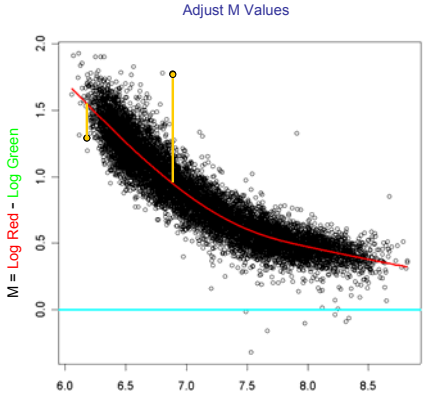
38



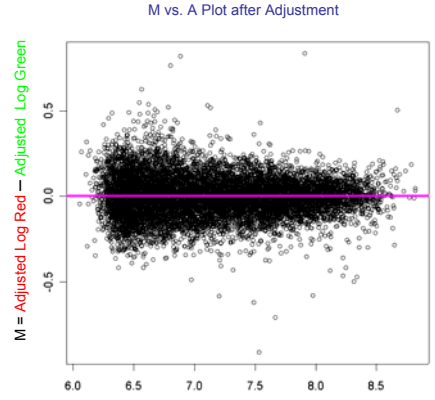
39



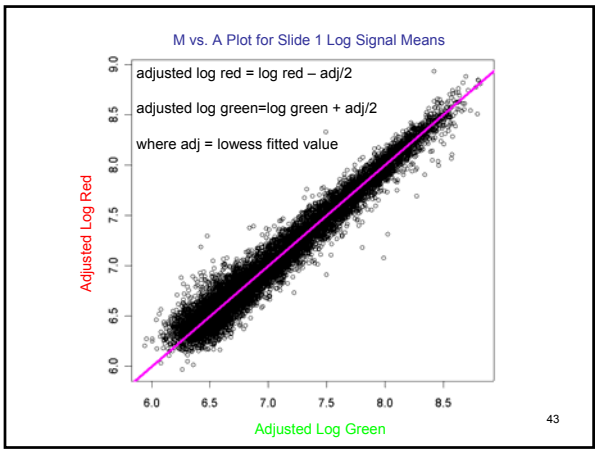
40



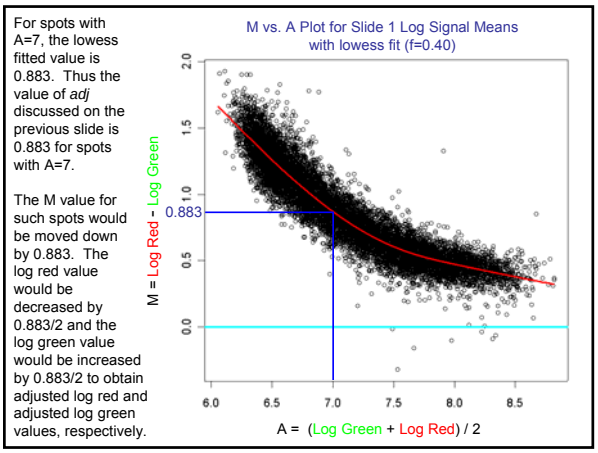
41



42

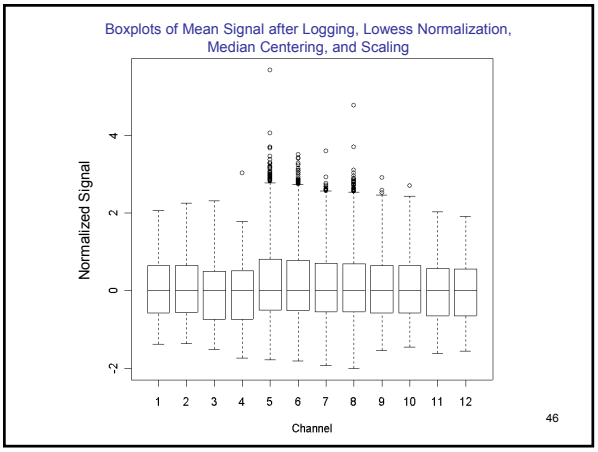


43



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.

45



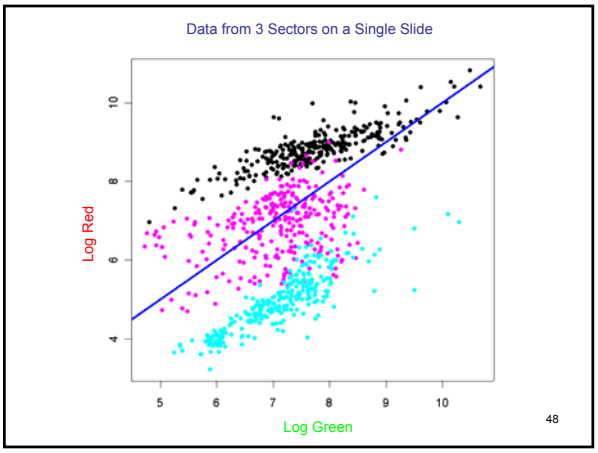
46

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.

47

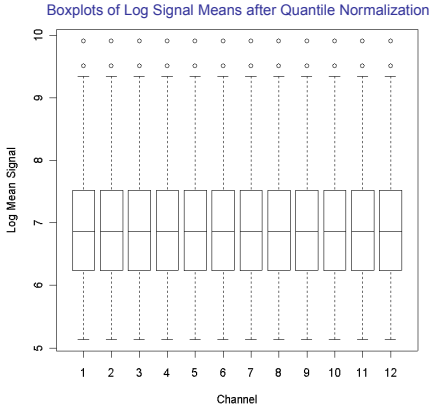


48

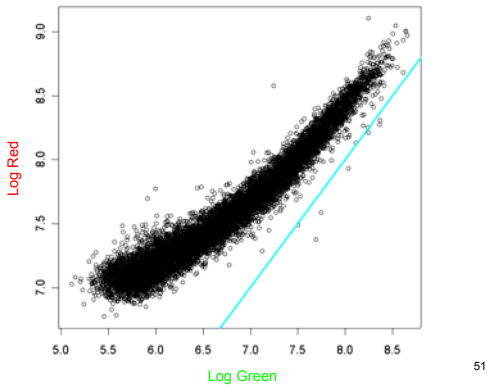
Bolstad, et al. (2003, *Bioinformatics* 19 2:185-193) propose *quantile normalization* for microarray data

- 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=$x_{0.75}$

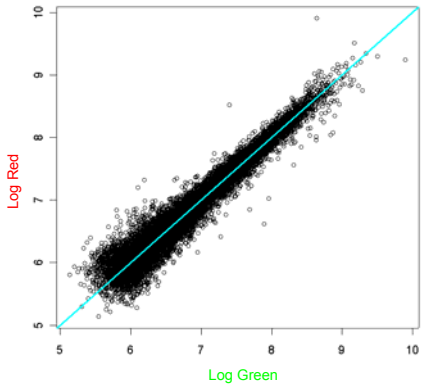
49



50



51



52

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.

53

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

54

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

55

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

56

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

57

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

58

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

59

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

60

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

61

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

62

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

63

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

64

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

65

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.

66

Miscellaneous Comments on Preprocessing

- We have only scratched the surface in terms of preprocessing methods. There are many variations on the techniques that we have described as well as other approaches that we won't discuss.
- Preprocessing affects the final results, but it is often not clear what strategies are best.
- It would be good to integrate preprocessing and statistical analysis, but it is difficult to do so. The most common approach is to preprocess data and then perform statistical analysis of the resulting data as a separate step in the microarray analysis process.

67