NORMALIZATION
(Spotted cDNA Microarrays)

MICROARRAY DATA

- Very noisy

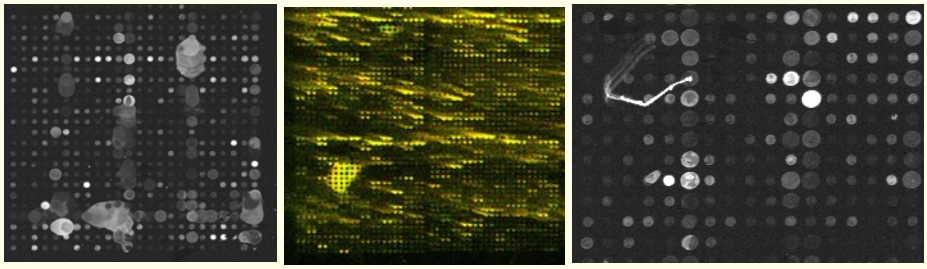
  (lot of factors interacting simultaneously)
- Expensive
- Lot of data from relatively few experimental units

  (large *p*, small *n* paradigm)

BEFORE ANALYSIS, VISUALIZE YOUR DATA !

- Unexpected problems: bubbles, dust, etc.
- No signal; or Saturation
- Spatial effects on slide
- Dye biases (intensity and spatial dependents)
- Other systematic effects

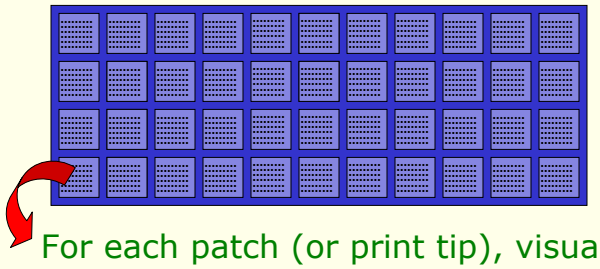
EXAMPLES: Bad quality arrays …



NO SIGNAL; SATURATION

		Array 1		Array 2		Array 3	
Address	Gene Name	Cy3	Cy5	Cy3	Cy5	Cy3	Cy5
B5.d2	MMP1 up	4062132	4062132	4062132	4062132	4062132	4062132
B5.g2	MMP1 up	1852018	1852018	1852018	1852018	1852018	1852018
B5.i9	MMP1 up	306319	306319	306319	306319	306319	306319
C7.c6	midkine	426076	426076	426076	426076	426076	426076
C7.f5	midkine	1287372	1287372	1287372	1287372	1287372	1287372
C7.i4	midkine	2238863	2238863	2238863	2238863	2238863	2238863
D1.a4	17a-hydroxylase	2009750	2009750	2009750	2009750	2009750	2009750
D1.d5	17a-hydroxylase	813810	813810	813810	813810	813810	813810
D1.g4	17a-hydroxylase	404578	404578	404578	404578	404578	404578
D5.a4	3BHSD	636744	636744	1	636744	636744	636744
D5.d5	3BHSD	1059708	1059708	1	1059708	1059708	1059708
D5.g4	3BHSD	1884075	1884075	1884075	1884075	1884075	1884075
C6.d2	INF gamma	5572910	5572910	5572910	5572910	5572910	5572910
C6.g2	INF gamma	3481462	3481462	3481462	3481462	3481462	3481462
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

SPATIAL EFFECTS ON SLIDE

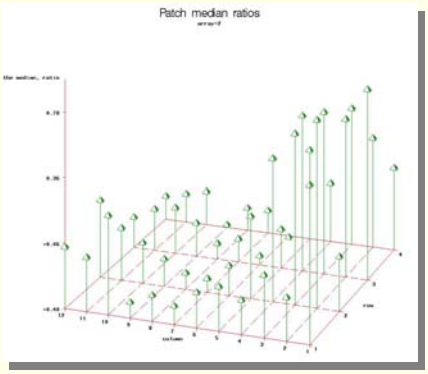


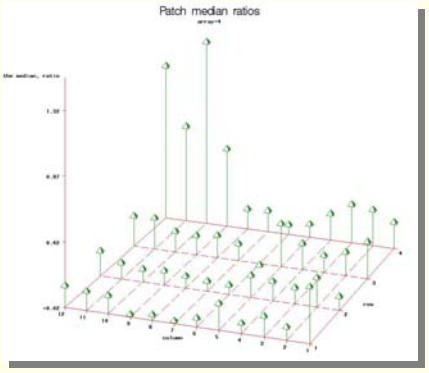
For each patch (or print tip), visualize intensities of control genes or use " robust" measures as median and trimmed means.

Trimmed mean ($\bar{x}_\alpha$): mean after eliminating the 100.α% of the smallest and biggest values.

i.e., mean of the 100.(1-2α)% of middle numbers.

SPATIAL EFFECTS ON SLIDE





SPATIAL EFFECTS ON SLIDE

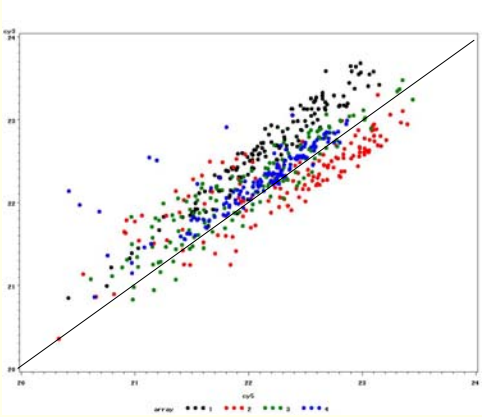
➡ We'll see later in the course some alternatives to take spatial effects into account (and correct for them) when analyzing the data…

➡ For now, imagine that we can consider the subdivisions of the slide (patches) as blocks into the statistical model.

➡ Other alternative would be continuous (smoothing) spatial modeling…

DYE BIAS

- Log(Cy3) *vs.* Log(Cy5) Scatterplot



DYE NORMALIZATION

- Cy3 and Cy5 are relatively unstable, and may present different incorporation efficiencies during labeling, different quantum efficiencies, and are detected by the scanner with different efficiencies.
- Normalization: to balance the fluorescence intensities of the two dyes, as well as to allow the comparison of expression levels across experiments (slides).

(Yang et al., 2002)

DYE NORMALIZATION

➡ Linear Regression ("Calibration")

$$\log(Cy3) = a + b.\log(Cy5)$$

Use estimates *a* and *b* to normalize the data:

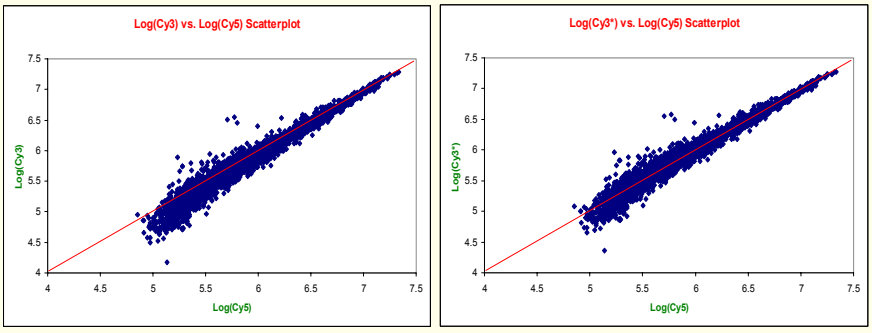
$$\log(Cy3)_{\text{normalized}} = \frac{\log(Cy3) - a}{b}$$

➡ Alternative: Regression through the origin

$$\log(Cy3) = b.\log(Cy5) \Rightarrow \log(Cy3)_{\text{normalized}} = \frac{\log(Cy3)}{b}$$

DYE BIAS NORMALIZATION

➡ Example (Data from S. Madsen)



Before After

DYE NORMALIZATION

➡ WHICH (SET OF) GENES TO USE ?

- Housekeeping genes
- Controls
- All genes on the array

β actin
RPL19
GAPDH
Lambda Q
DMSO
Blank
etc.

Selecting "Nondifferentially" Expressed Genes

Rank-invariat selection scheme

(Schadt et al., 2001; Tseng et al., 2001)

① Compute Rank($Cy5_g$) and Rank($Cy3_g$), the ranks of the Cy3 and Cy5 intensities of each gene *g* on the slide.

② Select the rank-invariat set (S):

$$S = \{g: Rank(Cy5_g) - Rank(Cy3_g) < d \;\&\; m < Rank[(Cy5_g + Cy3_g)/2] < G - m$$

(d and m are prespecified values)

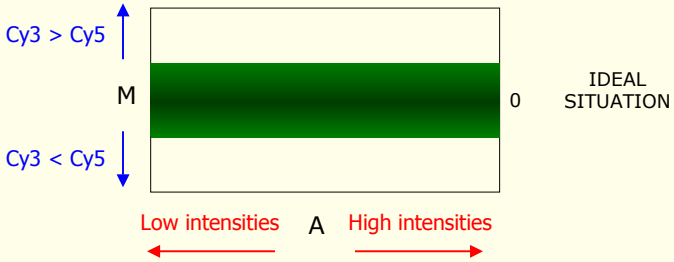
DYE-INTENSITY BIAS

M-A Plot (log intensity ratio *vs.* mean log-intensity)

45º counterclockwise rotation of the (logCy3, logCy5)-coordinate system, followed by scaling of the coordinates.

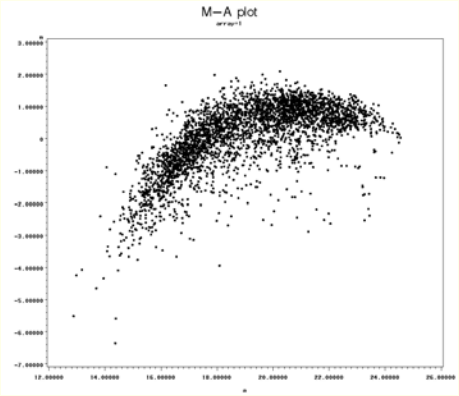
$$M = \log(Cy3/Cy5) = \log(Cy3) - \log(Cy5)$$

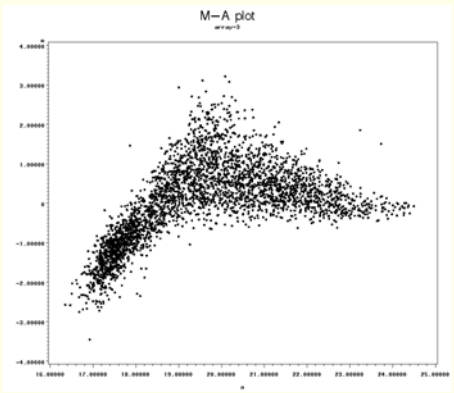
$$A = \log_{\sqrt{}}(Cy3 \times Cy5) = [\log(Cy3) + \log(Cy5)]/2$$



DYE-INTENSITY BIAS

Normalization (Data from Madsen et al., 2002)





LO(W)ESS

Locally-Weighted Regression and Smoothing Scatterplots

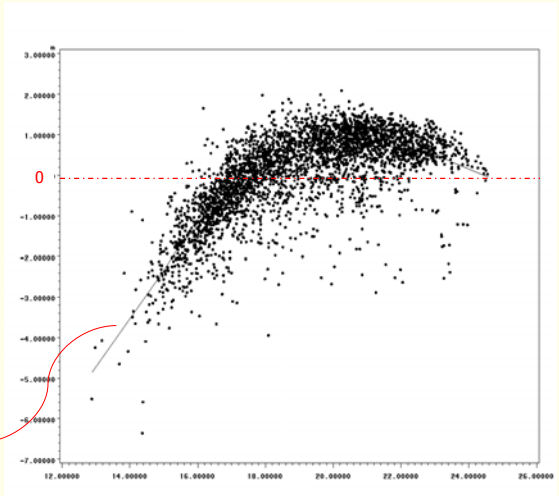
Basic Idea: ① For $x = x_0$, specify a neighborhood

- x_0: (center of the neighborhood)
- neighborhood: certain radius
  - radius: *smoothing parameter* (measured as a percentage of the data points)

② Weighted least squares to fit linear or quadratic functions at x_0

(by a decreasing function of the distances from x_0)

M-A Plot



Normalization

- Normalized intensities

$$M^* = M - \hat{M}$$

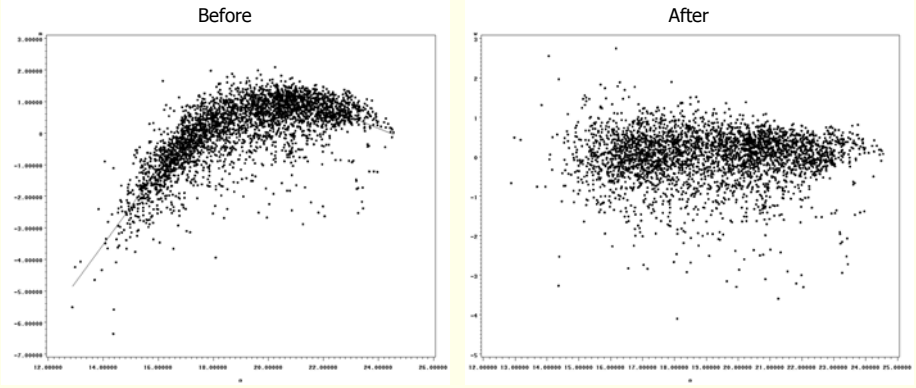
Normalized M — Loess-predicted M

Adjusted Values

$$\log(Cy3^*) = A + \frac{M^*}{2} \quad \text{and} \quad \log(Cy5^*) = A - \frac{M^*}{2}$$

Normalization

- LOESS (Local Regression)



LOESS SAS Code

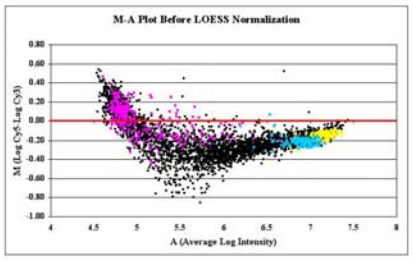
```
proc sort data=m_a;
  by array a;
run;

proc loess data=m_a;
  by array;
  model m=a;
  * smooth = 0.02 to .2 by 0.01;
  ods output OutputStatistics=resul;
run;
```

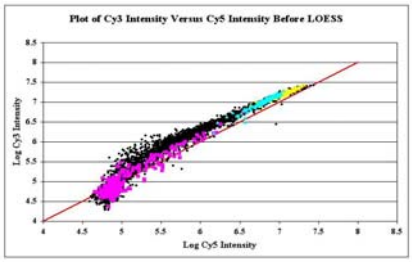
The use of Control Genes

- Important Assumption: Most of the genes are not differentially expressed;
- If this assumption seems not adequate (e.g. in aging studies), the use of control genes may be an alternative;
- But … Controversial. Control gene should be constant over all experimental situations, and should be representative of all intensities…

The use of Control Genes

(Data from Dr. P. Coussens)





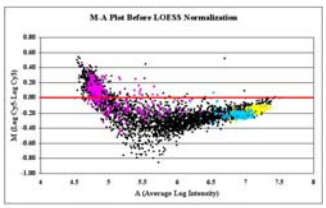
● Samples ● Blank ● GAPDH ● Lambda Q

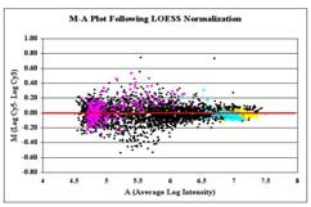
LOESS Normalized Data

(using all genes)

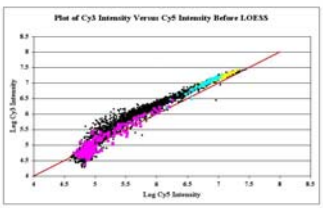
Before After

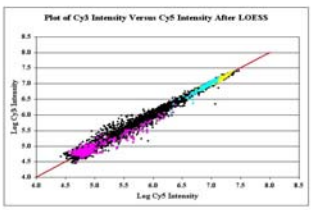
M-A



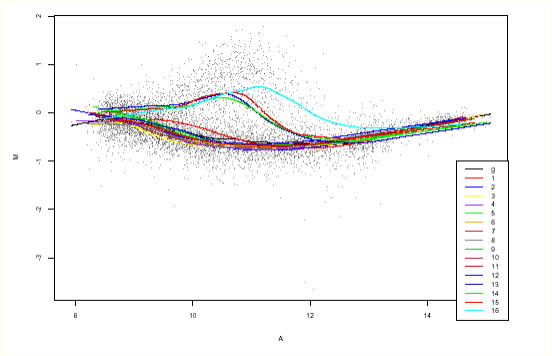


Cy3-Cy5





PRINT-TIP SPECIFIC NORMALIZATION



SCALE NORMALIZATION

➡ Some scale adjustments may be required so that the relative expression levels from one particular experiment (slide) do not dominate the average relative expression levels across replicate experiments.

(Yang et al., 2002)

WITHIN-SLIDE SCALE NORMALIZATION

Scale normalization across print-tips

$$M_{ij} \sim N(0, a_i^2 \sigma^2)$$

*j*th log-ratio in the *i*th print-tip

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\prod_{k=1}^{I} \sum_{j=1}^{n_i} M_{ij}^2}}$$

Estimate of the scale factor a_i

$$M_{ij}^* = \frac{M_{ij}}{\hat{a}_i}$$

Normalized values

REFERENCES

Afshari, C. (2002) Perspective: Microarray technology, seeing more than spots. *Endocrinology* 143: 1983-1989.

Cleveland, W. S. (1979) Robust locally-weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74: 829-836.

Cleveland, W. S., Grosse, E. (1991) Computational methods for local regression. *Statistics and Computing* 1: 47-62.

Madsen, S. A., Yao, J., Sipkovsky, S., Rosa, G. J. M., Coussens, P. M., Burton, J. L. (2002) cDNA microarray analysis of neutrophil gene expression around parturition. *Cold Spring Harbor Laboratories*, Cold Spring Harbor - NY, April 24-28.

Yang, Y. H., Buckley, M. J., Dudoit, S., Speed, T. P. (2002) Comparison of methods for image analysis on cDNA microarray data. *J. Comp. Graph. Stat.* 11: 1-29.

Yang, Y. H., Dudoit, S., Luu, P., Speed, T. P. (2001) Normalization for cDNA microarray data. Technical Report # 589. Department of Statistics, University of California – Berkeley.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nuclei Acids Research 30, No. 4 e15..