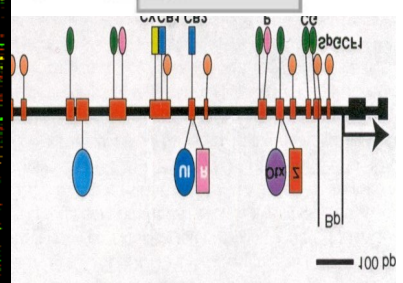
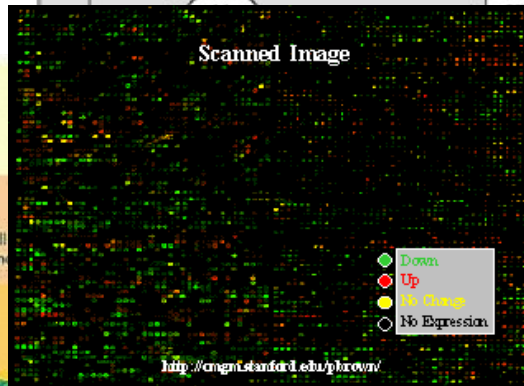
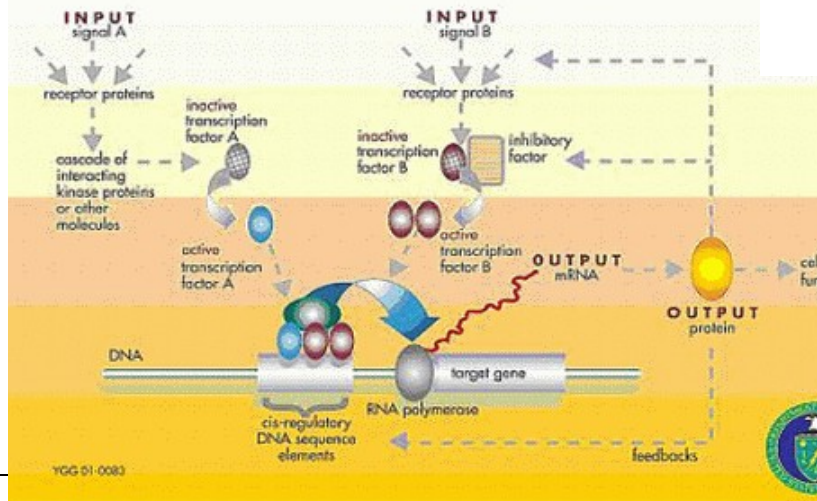
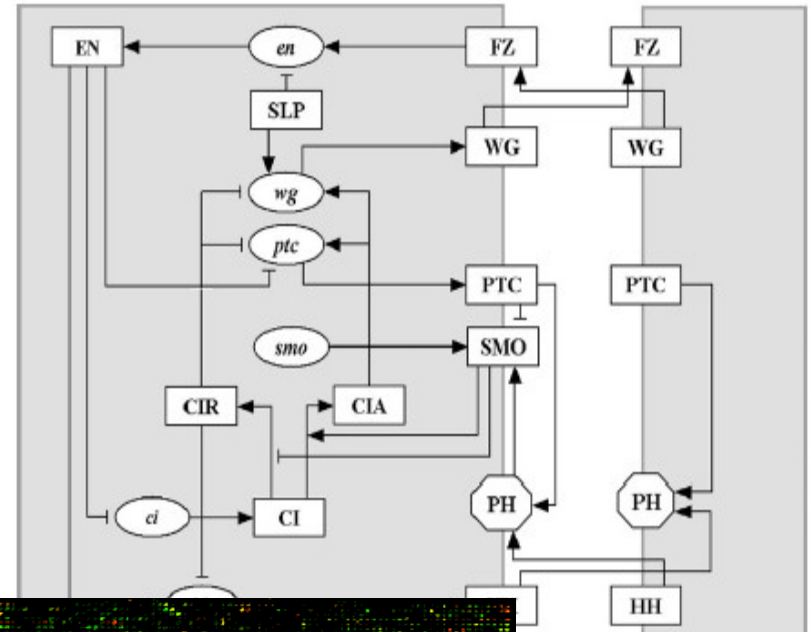
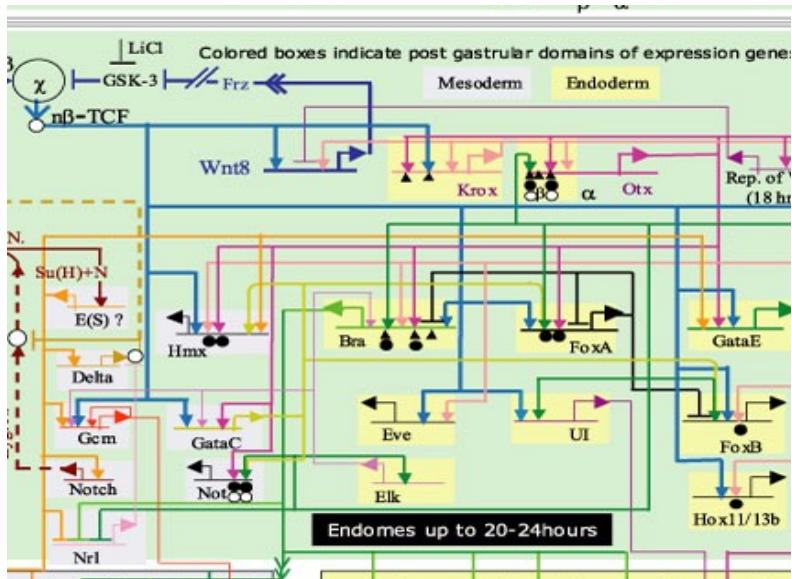


# ECS 234: Microarray Data Analysis



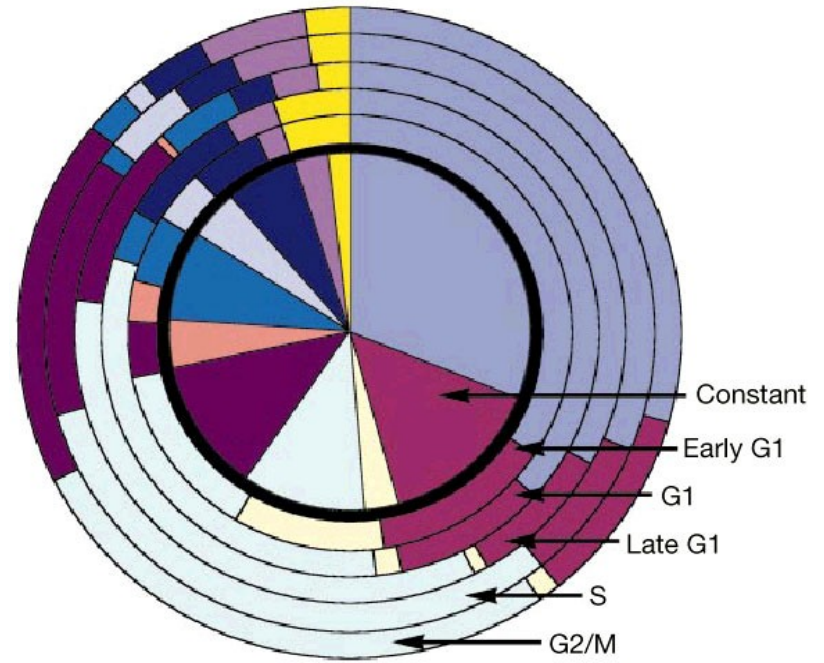
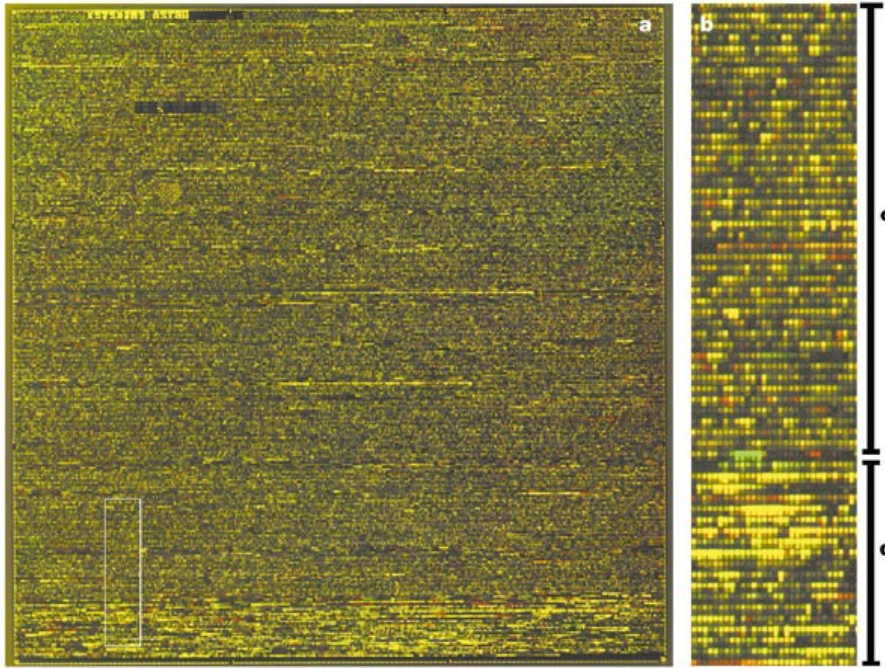


# Microarray Data Properties

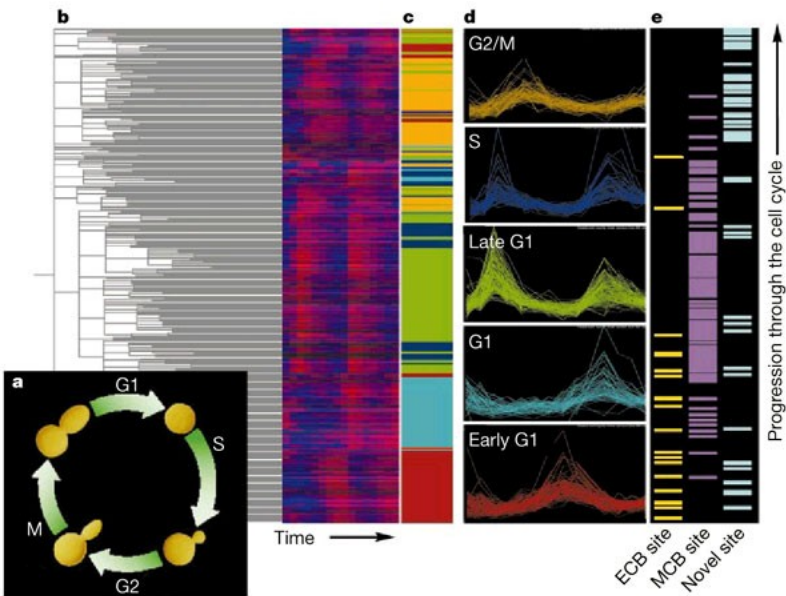
- A lot of data, but not enough!
- Many genes and few conditions (the dimensionality curse)
- Very few repeats (2, 3, 4, mainly)
- Data from different experiments difficult to compare: control conditions are different
- Inaccurate at low intensities

# What Can We Do With Microarray Data?

- Fishing Expeditions vs. Hypotheses: differentially expressed genes
- Part/Whole Genome Hypotheses: cell/tissue classification
- Gene Expression vs. Gene Function: guilt by association (co-regulation)
- Transcription Regulation
- Fingerprinting
- Genome analysis
- Gene Circuitry



- Signal transduction
- Cellular biogenesis
- Intracellular transport
- Transport facilitation
- Protein destination
- Protein synthesis
- Transcription
- Cell growth, division, DNA synthesis
- Energy
- Metabolism
- Cellular organization



Lochart and Winzler 2000

# How Do We Do Those Things?

- Single Gene Differential Expression
- Similarity in Expression Patterns of Genes and Experiments (Classification)
- Co-regulation of Genes: function and pathways (Clustering)
- Network Inference (Modeling)

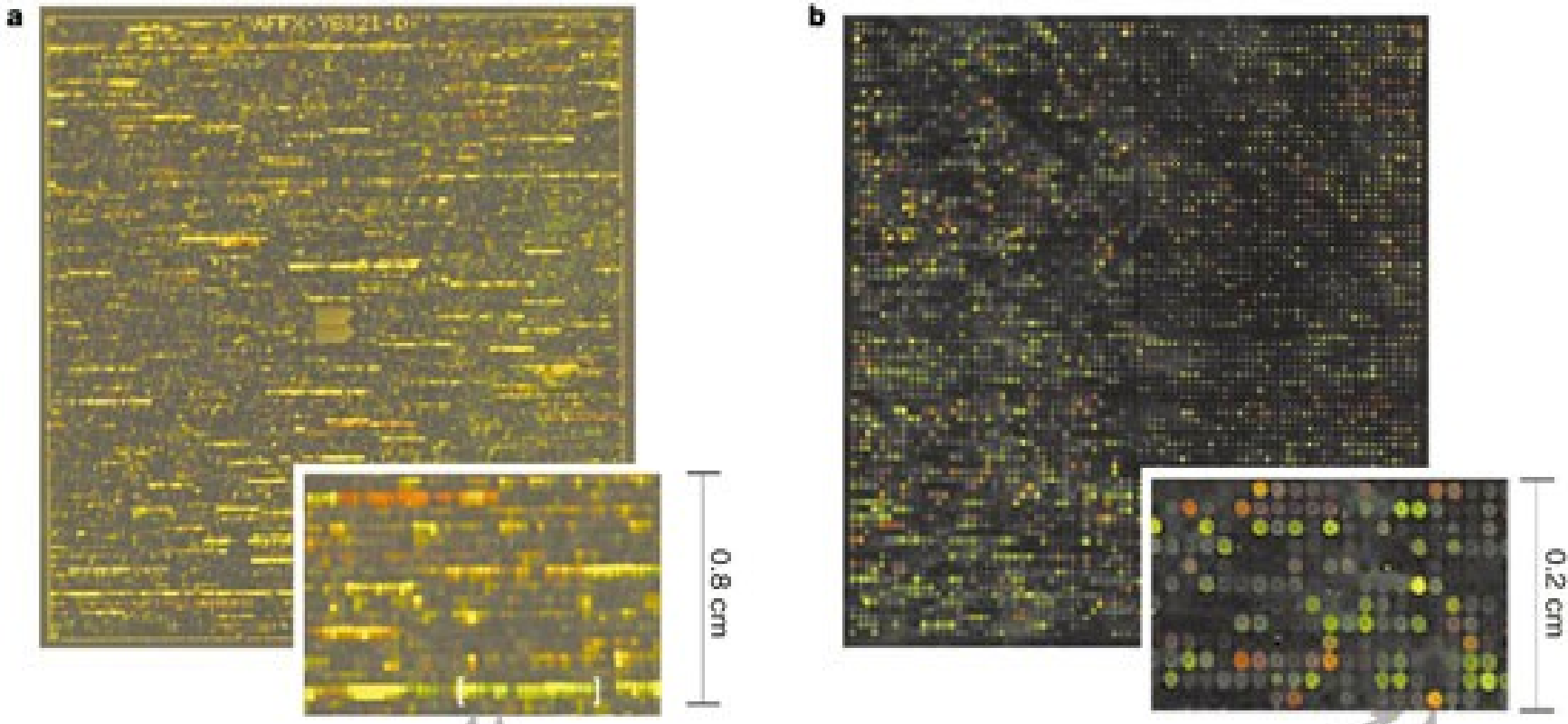
# Microarray Data Analysis I

1. Experimental Design
2. Normalization and Transformation
3. Identification of differentially expressed genes
  - Fold test
  - T-test
  - Correction for multiple testing

# Microarray Data Analysis II: Discovery

1. Classification
2. Clustering
3. Local Pattern Discovery
4. Projection Methods
  - PCA
  - SVD

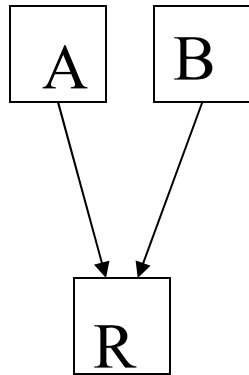
# 1. Choice of Technology



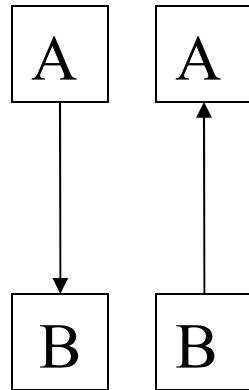
a) Oligonucleotide and b) Spotted Arrays

- Array Design
  - Affy
  - cDNA libraries and probes
- Controls
  - Affy: PM-MM, and others
  - Negative controls

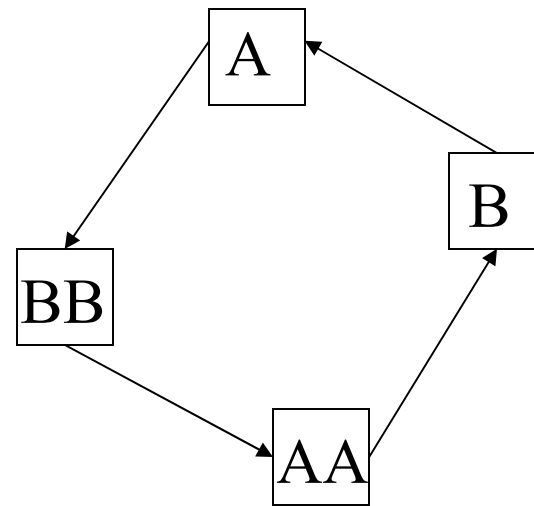
## 2. Experimental Design (two-color DNA microarrays)



Reference



Direct  
(with dye swap)

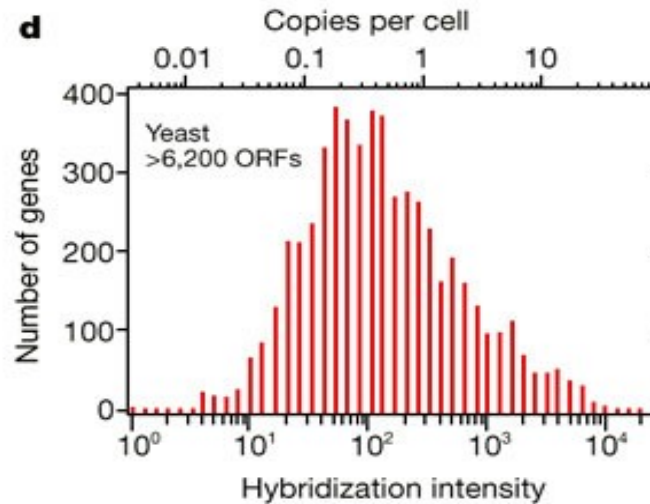
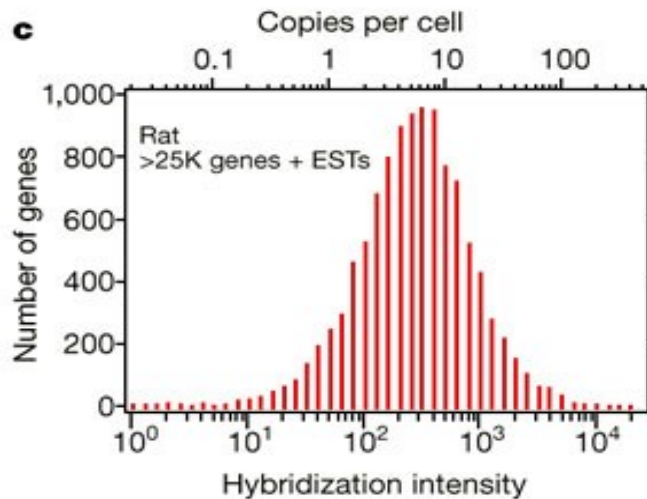
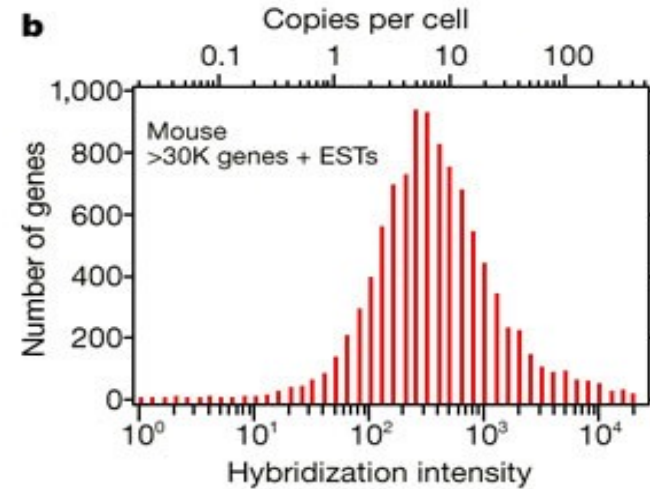
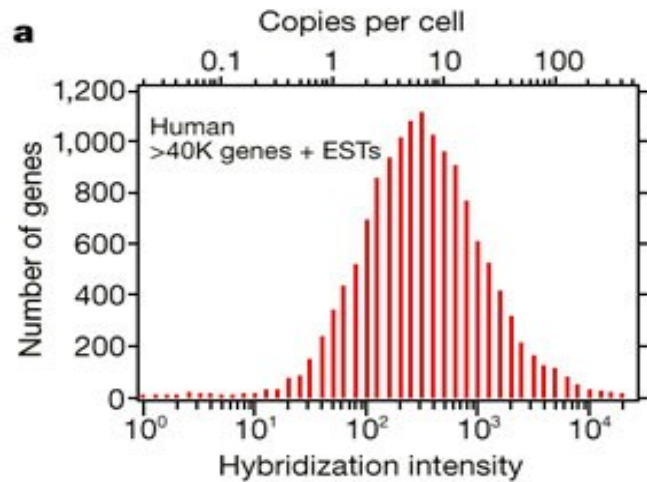


Loop

# Types of Microarray Data Experiments

- Control vs. Test
- Time-wise
  - Snapshots (each experiment is different conditions)
  - Time-Course Experiments (each experiment is a time-point)
- Gene-knockout (perturbation experiments)

# Distribution of Observed Values



# Distribution of Observed Values is $\sim$ log-normal

$\log$  (Color Intensity) or  $\log R/G$  is  
a good estimator of differential  
expression

But one can do better by properly accounting for all  
systematic sources of error

# Log-ing the Data

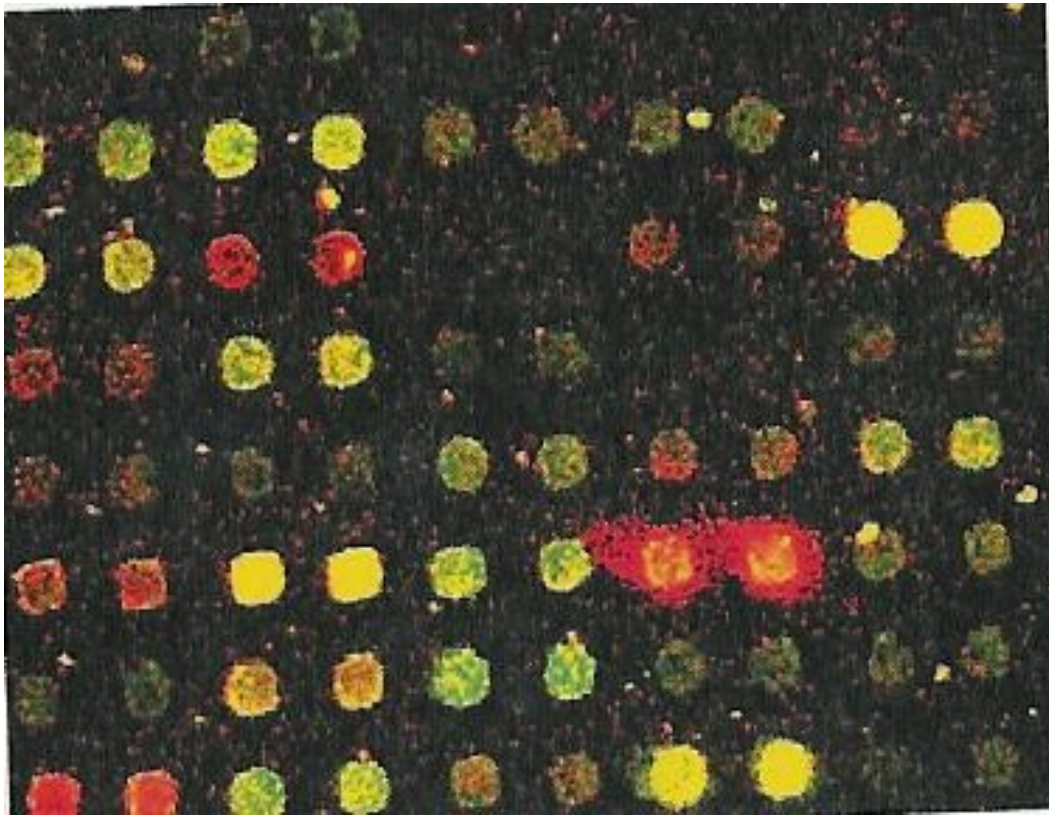
- It should always be done unless a more sophisticated analysis will follow
- Good for two symmetry between up- and down-regulated genes
- Simple comparison across experimental designs

# 3. Data Acquisition and Visualization

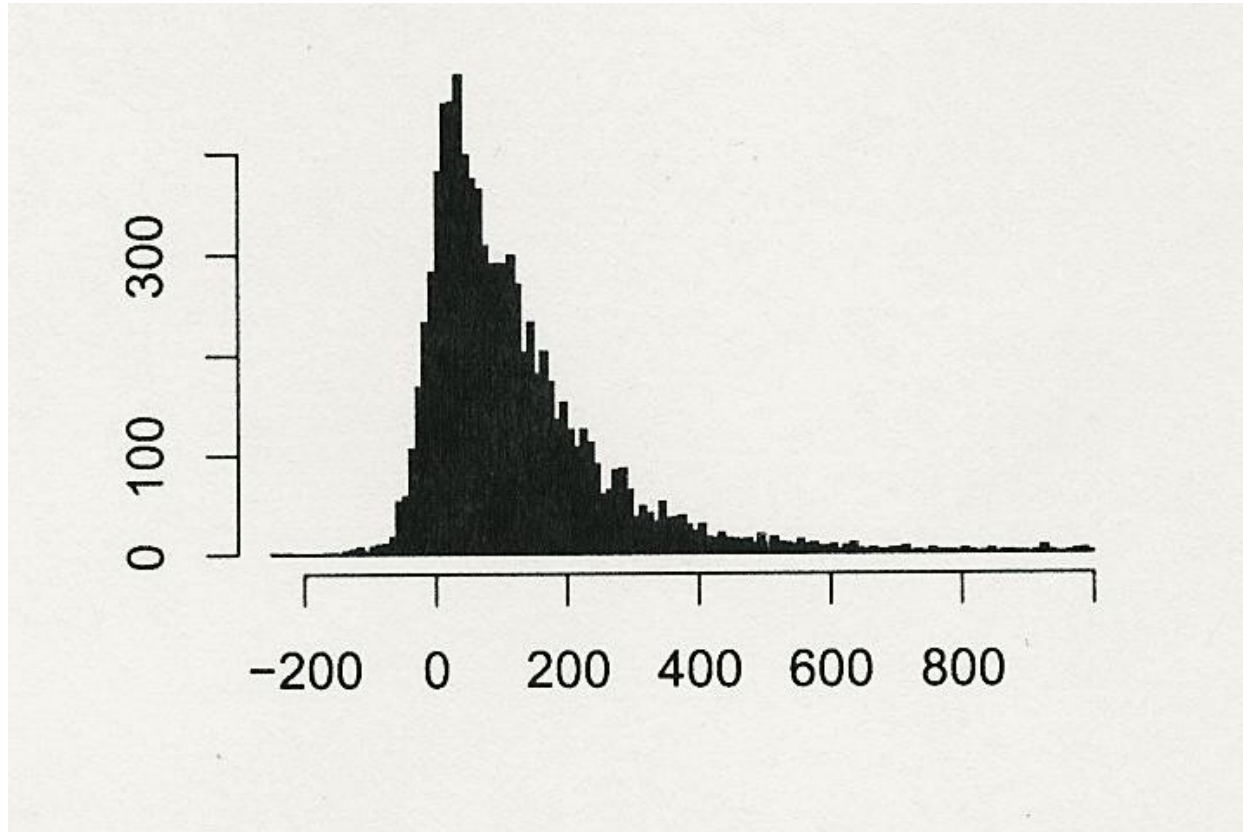
- Image quantification (spot reading)
- Dynamic Range and spatial effects
- Scatterplots
- Systematic sources of error

# 1. Data Visualization

Image quantification (spot reading)



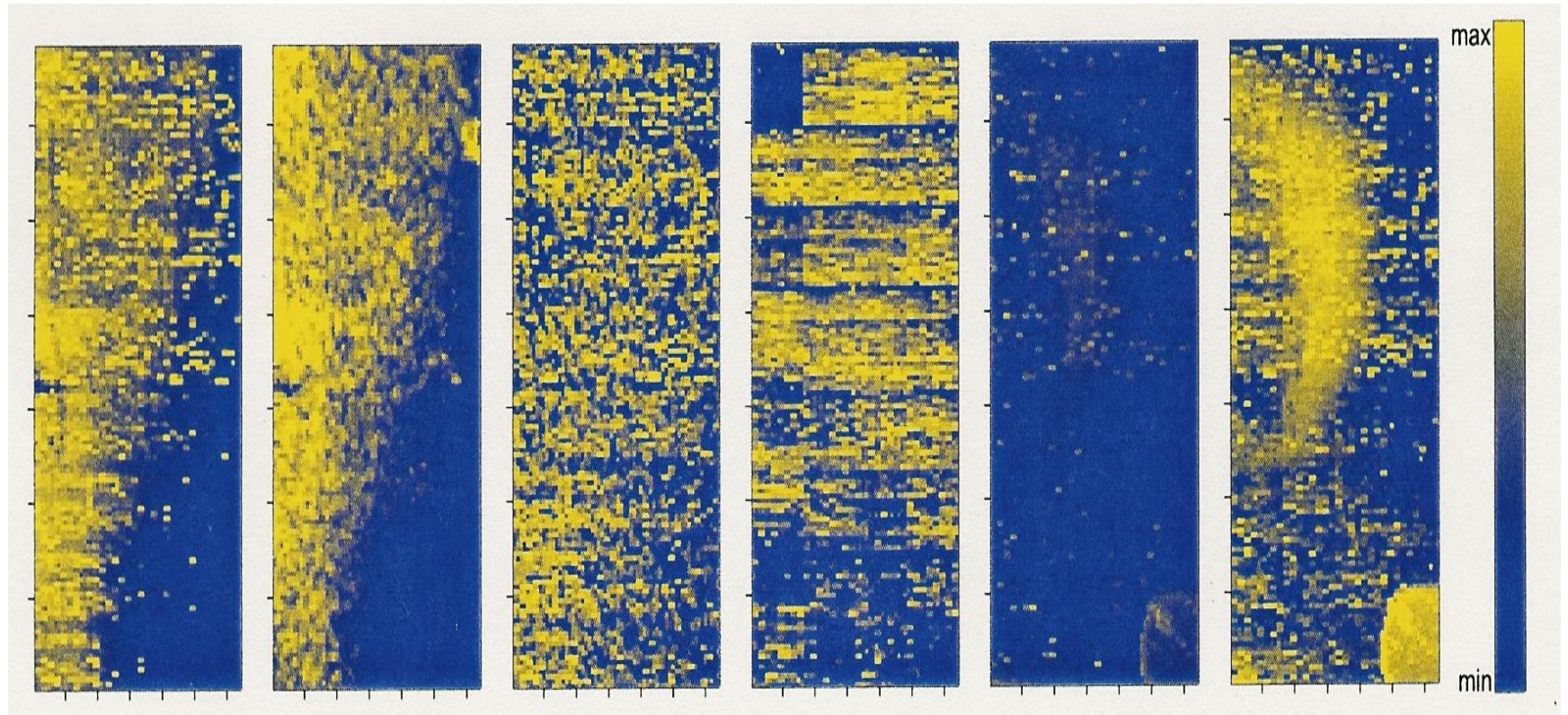
# Dynamic Range



Huber et al

Some values are negative after background subtraction

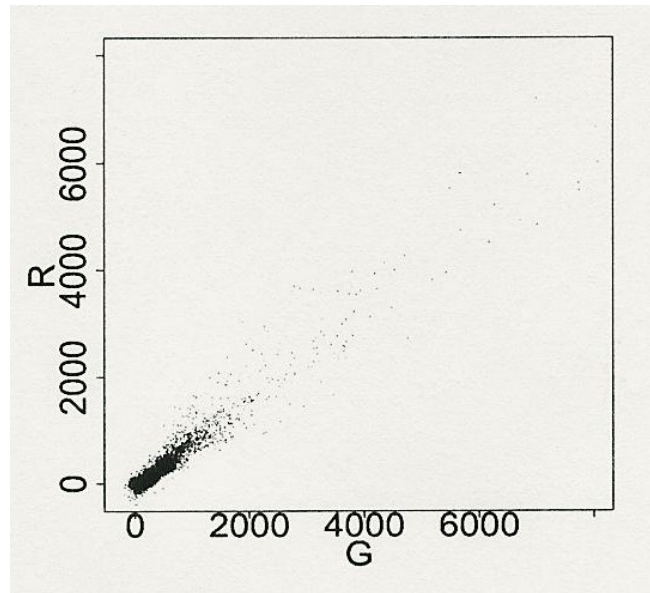
# Spatial Effects



Huber et al

# Checking the Data: Scatterplots

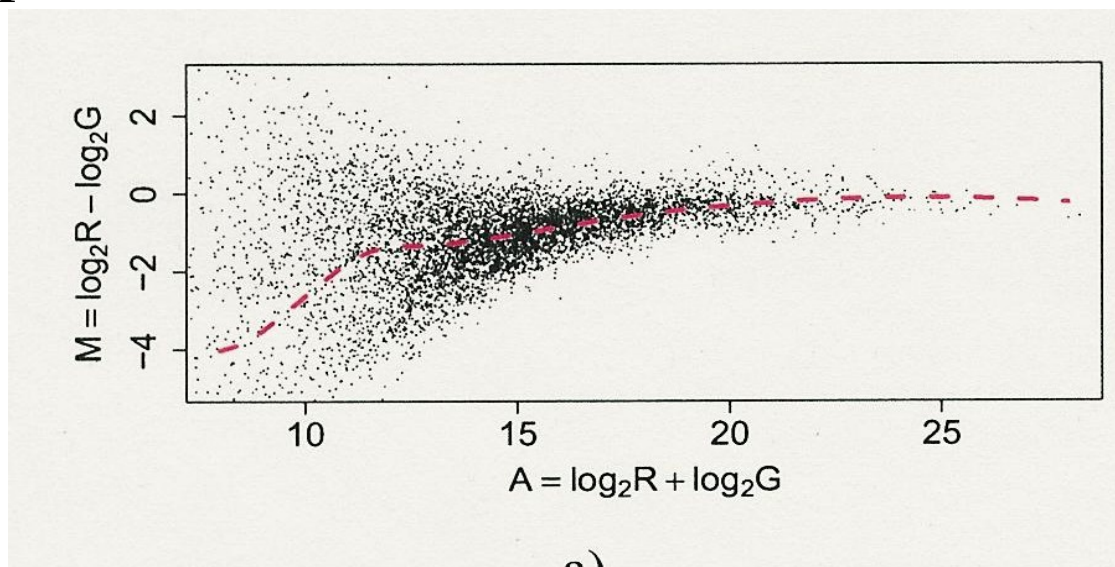
- Visual Aids for Data Calibration
- Plotting Red vs Green Expression



Huber et al

# Advanced Scatterplots: MA Plots

- Plotting Average vs. Differential Expression
  - $A = \log R + \log G$
  - $M = \log R - \log G$
  - Emphasize differences rather than similarities



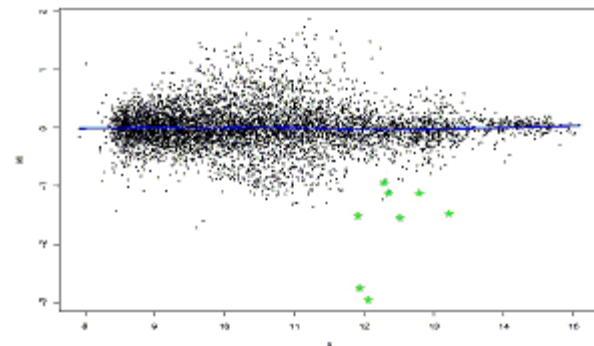
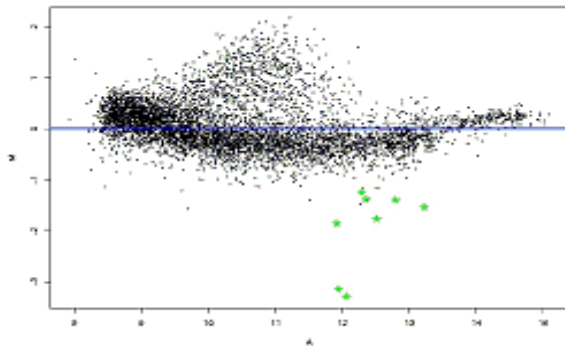
- Mean and Variance have to be corrected in general

# Sources of Error

- Spotting errors (tips, robot arm etc.)
- Imbalance in Red/Green Intensities
- PCR yield variance
- Preparation protocols (RNA degrading)
- Scanner and image analysis

## 4. Data Normalization

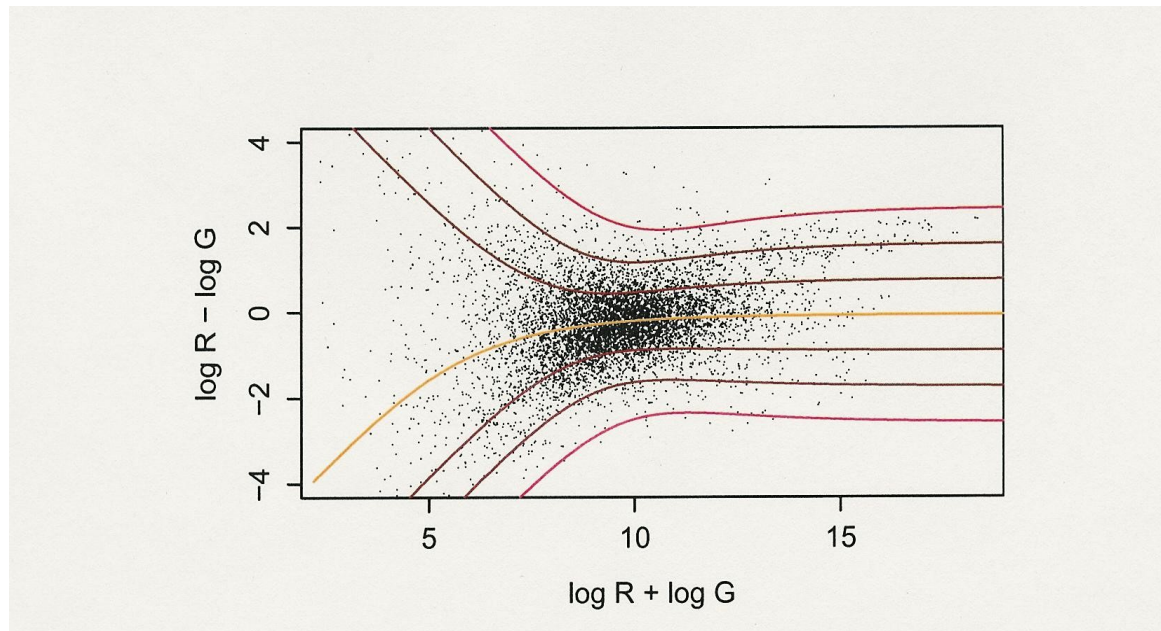
- Identification and removal of systematic sources of variation (adjusting the mean)
- To allow within slide and between slide data comparison



Before and after loess normalization of log data  
(locally weighted linear regression method)

# 5. Variance Stabilization

- Stochastic processes cause the variance of the log-values to be different at different intensities



# A Simple, Realistic Model for Reducing Systematic Error

$Y$  = Measured intensity,  $x$  = True abundance

$$Y = a + bx + \varepsilon$$

$a$  is an additive factor, corresponding to systemic effects stemming from the experimental medium and does not result from  $x$

$b$  is a gain factor resulting from the relationships between the abundance,  $x$ , and the rest of the experiment, i.e. color, detector gain, hybridization, etc.

$\varepsilon$  is a normally distributed random error

# Realistic Assumptions in the Model

## Yield Better Normalization

$Y$  = Measured intensity,  $x$  = True abundance

$$Y = a + bx + \varepsilon$$

$$b = e^\eta$$

$$\eta = N(0, \sigma_\eta), \varepsilon = N(0, \sigma_\varepsilon)$$

- The driving idea behind the model is to capture the variation of the variance at low intensities
- The normalcy assumptions are good approximations of real data

# Fitting the Data

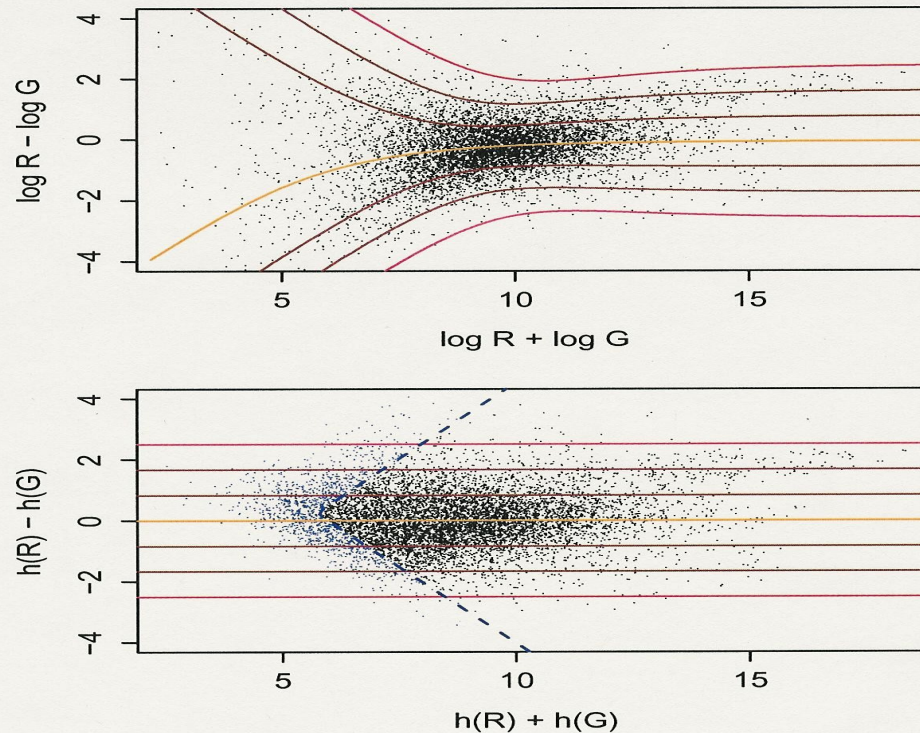
- Estimating the parameters of the model
- $a$ ,  $b$ , etc.
- Possible approaches:
  - least squares fit
  - Regression analysis

# Consequences of the model

- $\log Y_r/Y_g$  is no longer the best estimator for  $\log x_r/x_g$ .
- The appropriate measure of differential expression becomes

$$\Delta h = ar \sinh\left(\frac{\sigma_\varepsilon}{\sigma_\eta} \cdot \frac{Y_r - a}{b}\right) - ar \sinh\left(\frac{\sigma_\varepsilon}{\sigma_\eta} \cdot \frac{Y_g - a}{b}\right)$$

# This estimator has a constant variance across the range of intensities

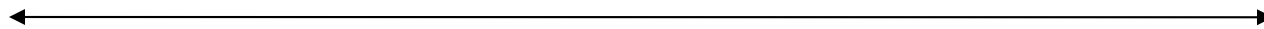


Huber et al

# 6. Identification of Differentially Expressed Genes in Replicated Microarray Experiments

Which genes are expressed differentially in different experiments?

	1,1	1,2	2,1	2,2
Gene 1	1	0	0	1
Gene 2	1	1	0	0



False Negatives  
(wrongly not identified)

False Positives  
(wrongly identified)

# 6.1. Statistical Tests

- Simple Fold Test
- Student t-test
- Wilcoxon rank sum

# Simple Fold Accounting

- A gene is differentially expressed up (down) if  $\log R/G > 2$  ( $< 0.5$ )
- Not good for low and high intensities (because the distribution of log-expression values has tails! )

# Student-t test

## Null Hypotheses Rejection:

- $H_j$  = mean expression levels are equal for control and treatment for gene  $j$ ,  $j=1, \dots, k$
- Let  $x_1, \dots, x_{n_c}$  and  $x_1, \dots, x_{n_t}$  be the normalized expression levels of  $n_c$  and  $n_t$  samples, respectively, in the control and test groups
- t-test for gene  $j$

$$t_j = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{\sigma_t^2}{n_t} + \frac{\sigma_c^2}{n_c}}}$$

where  $\bar{x}$  is the average and  $\sigma$  the standard deviation

# Alternatives for Student-t for Small Number of Replicates

- Regularized t-statistic
  - Estimate additional observations based on the overall data
- Full Bayesian Approaches

## 6.2. p-values

- $H_j$  is rejected if the significance of the t-test score is high, i.e. the probability of it happening at random is low (based on the Student-t distribution)
- Probability of happening at random:  
 $\alpha > 5\%$   
Rejection probability:  
 $\alpha < 0.5 \%$

# Correction for Multiple Hypotheses

- Even at small  $\alpha$ , say 0.5, when testing 1000 genes for differential expression we get 5 hits at random: high amount of false positives
- FWER, FDR
- Correcting for testing k hypothesis:
  - FWEER, FDR, Bonferoni

Bonferoni correction:

$$p = \min( k * p_t, 1 )$$

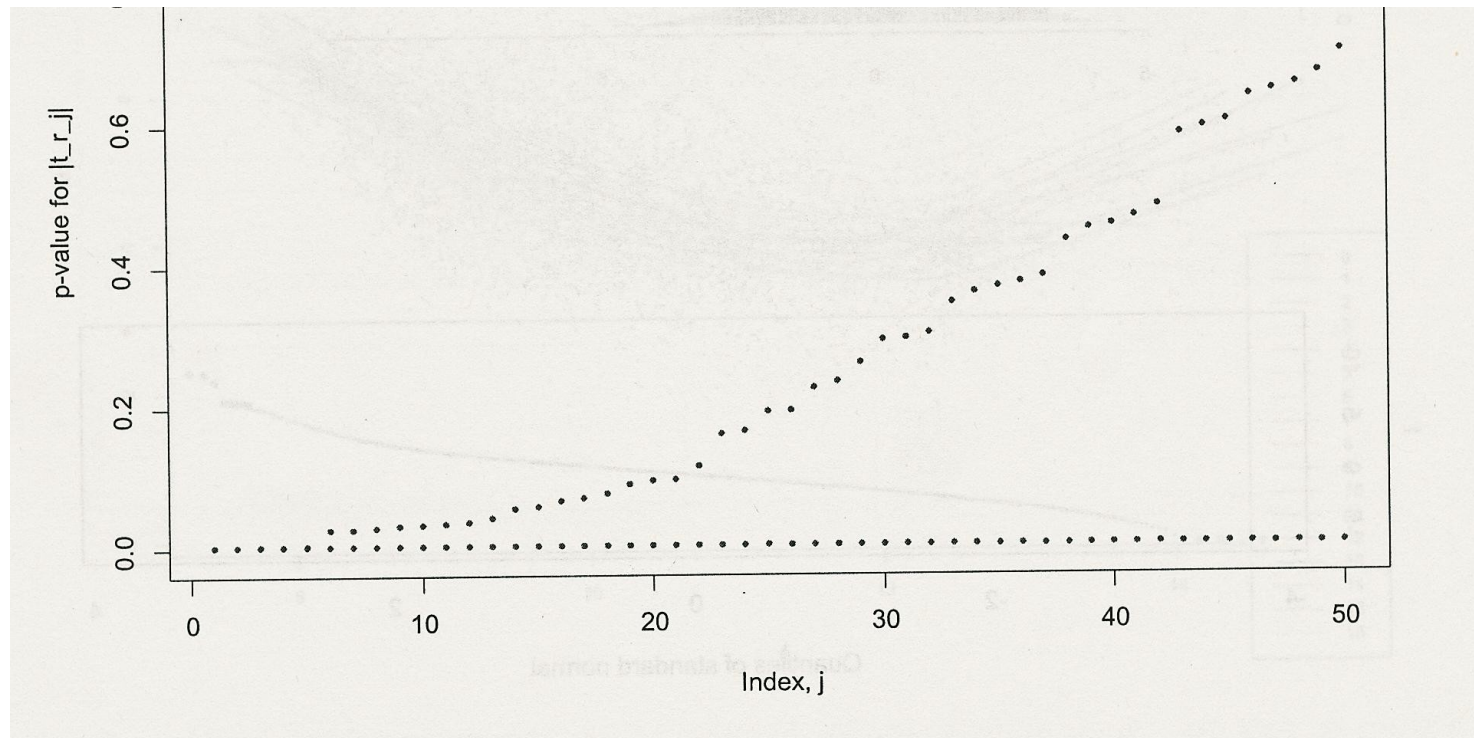
**Bonferoni doesn't work well for microarray data!!!**

# Alternatives to Bonferroni

Bonferroni is a very conservative correction, resulting in too many false negatives

- Westfall and Young step-down adjusted p-values
  - Not as conservative, but computationally intensive
- FDR (ex. SAM) most promising

# Adjusted vs. Unadjusted p-values



Dudoit et al

# 7. Microarray Data Standard

- Beyond systematic errors, microarray data from every experiment is different:
  - Environment
  - Experiment design
  - Data processing
- A Microarray Data standard is needed:  
MIAME: the minimal set of information about a microarray experiment

# Microarray Standard (MAIME)

- Environmental Conditions
- Control Conditions
- Test Conditions
- Data
- Data Processing (if any)

# 8. Verification and Validation

- Verify results in the lab
- Validation: check for enrichment
  - Functional categories
  - GO annotation