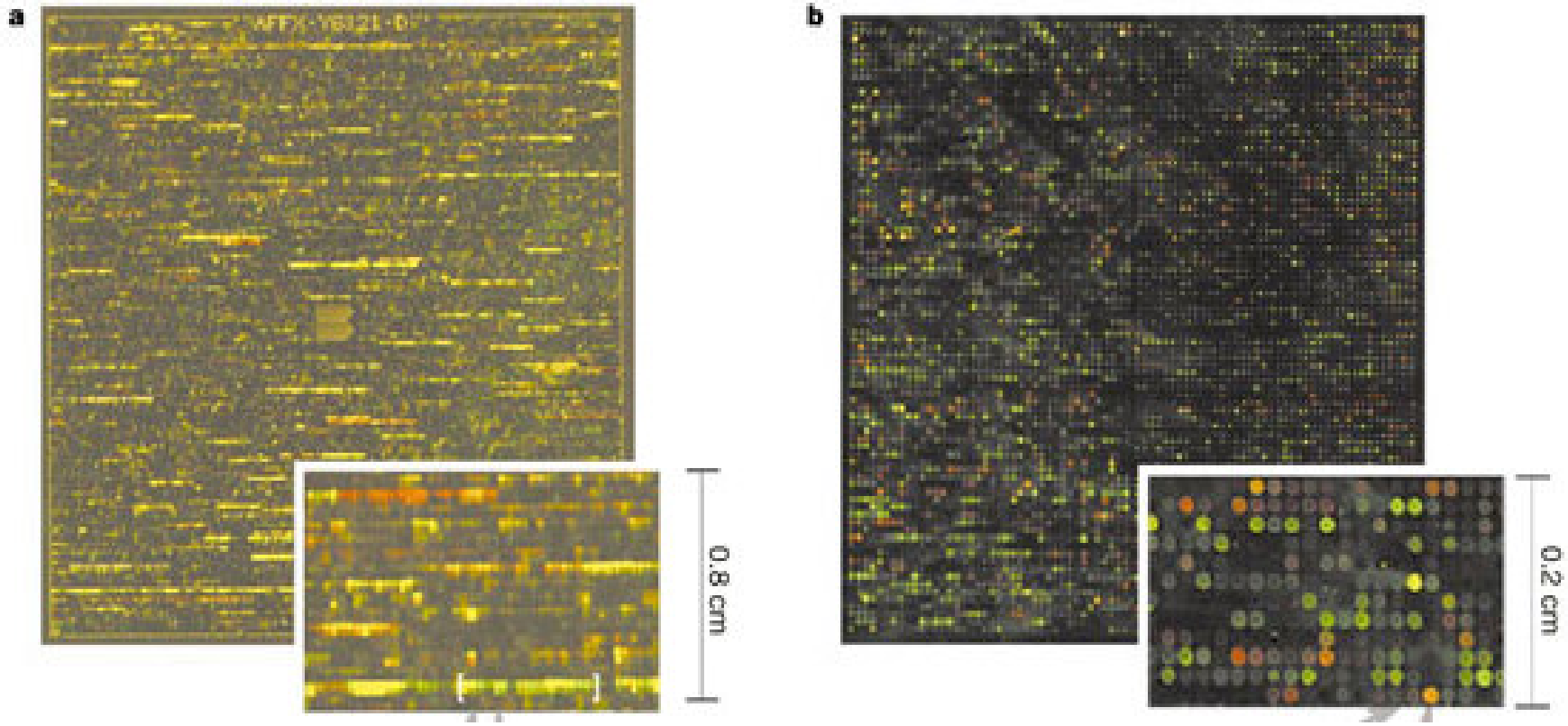
Microarray Data Analysis

ECS 289A



a) Oligonucleotide and b) Spotted Arrays

Lochart and Winzeler 2000

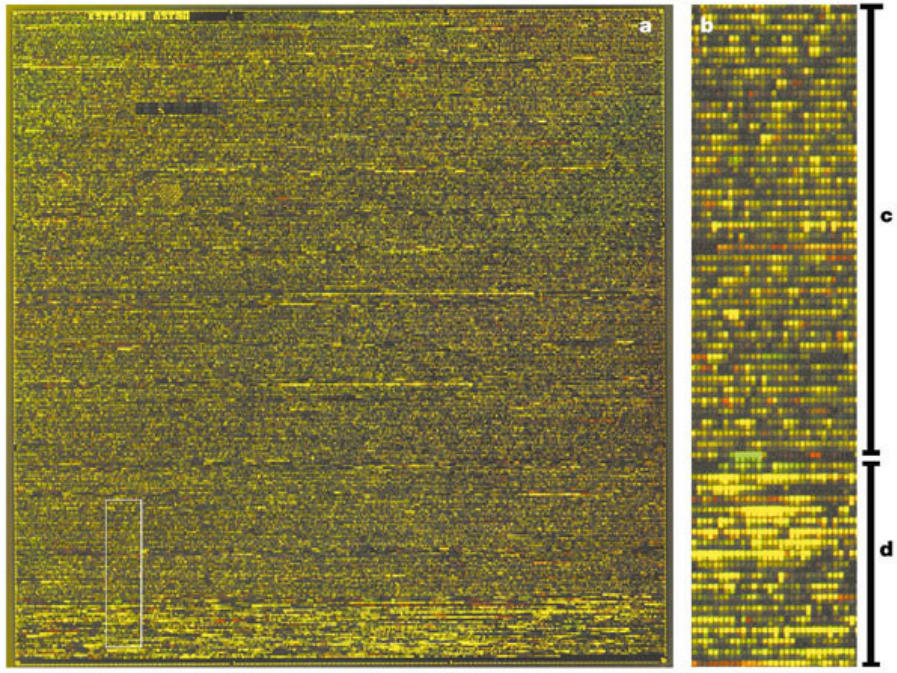
Microarray Data

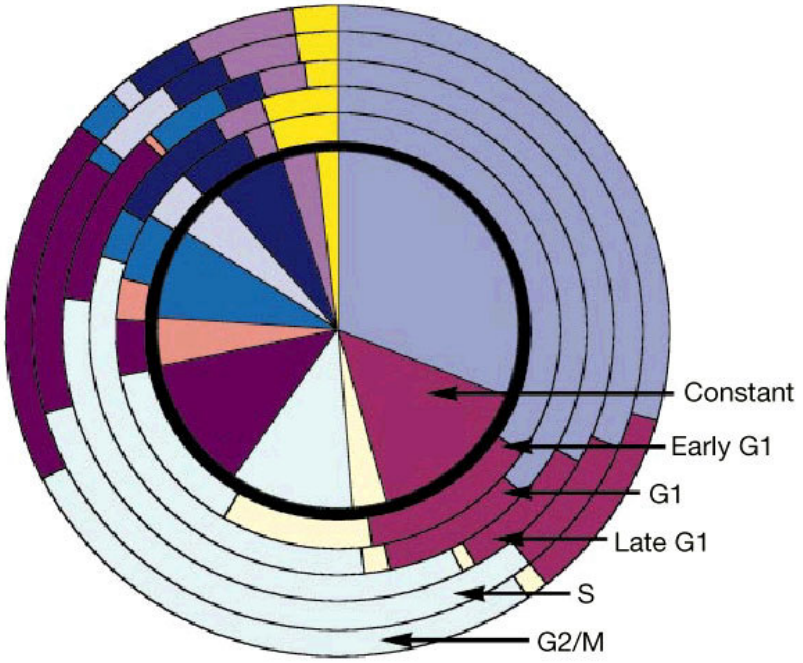
	Plate 1	Plate 2	…	Plate 10
Gene 1	0.013	2.14		
Gene 2	…	…		
Gene 3	…	…		
…				
…				
…				
…				
Gene 6200				

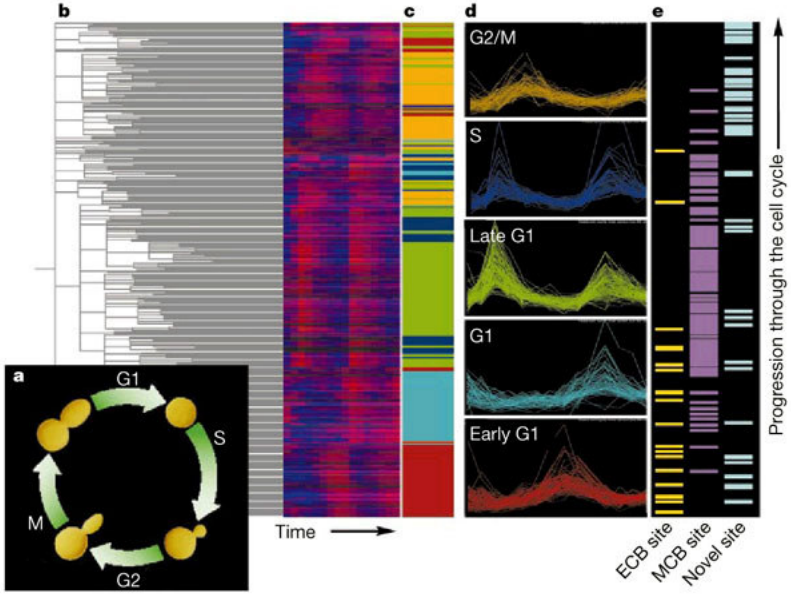
- Each entry is the relative expression of a gene in test vs. control.
- Ratio of the color intensities green/red (Cy3/Cy5) (spotted)
- Single color intensity (Affy)

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

Lochart and Winzeler 2000

ECS289A

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)

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)

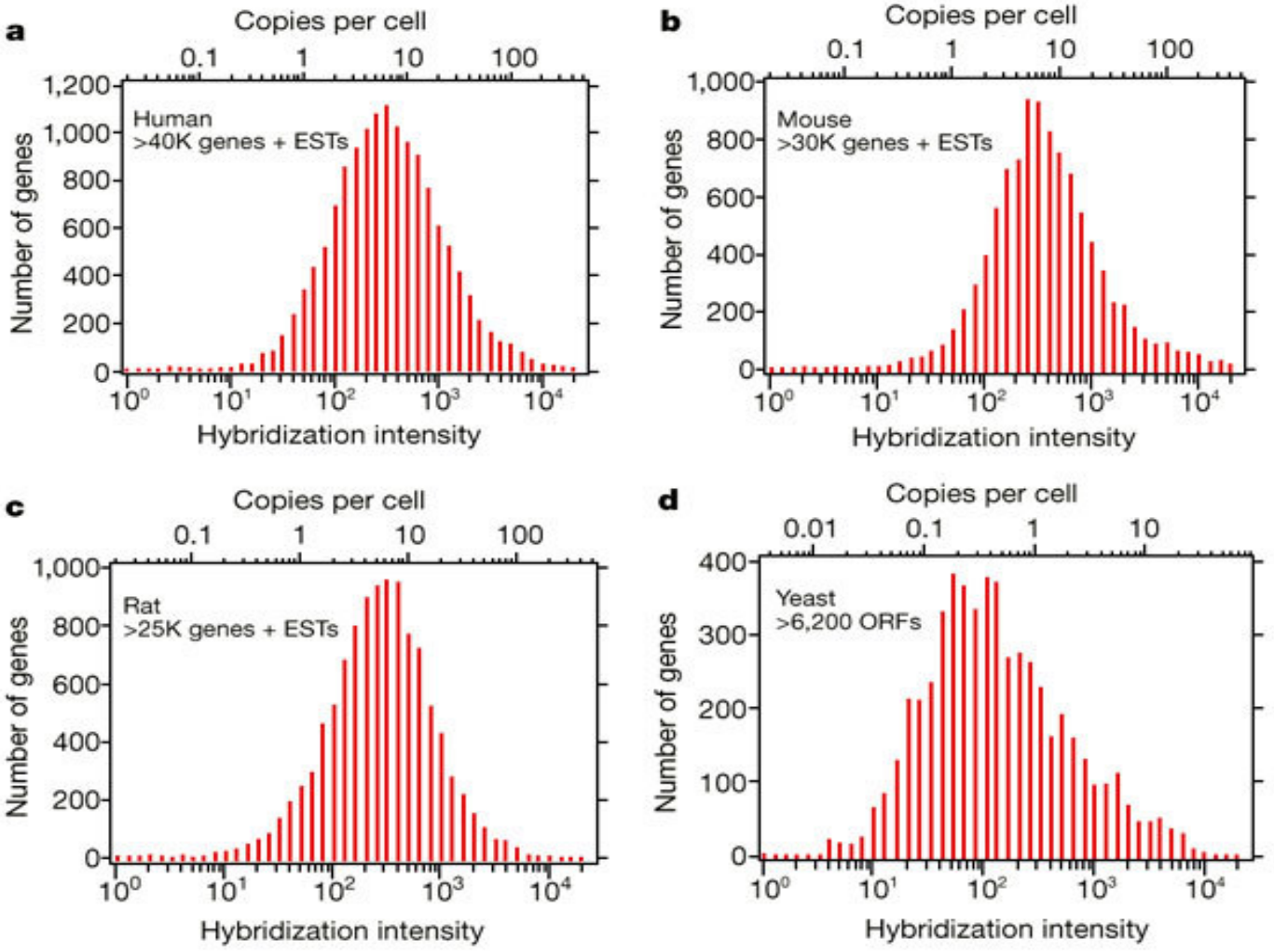
Microarray Data Properties

- 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

Microarray Standard (MAIME)

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

Distribution of Observed Values



Lochart and Winzeler 2000

Distribution of Observed Values is ~ 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

Microarray Data Analysis (stats)

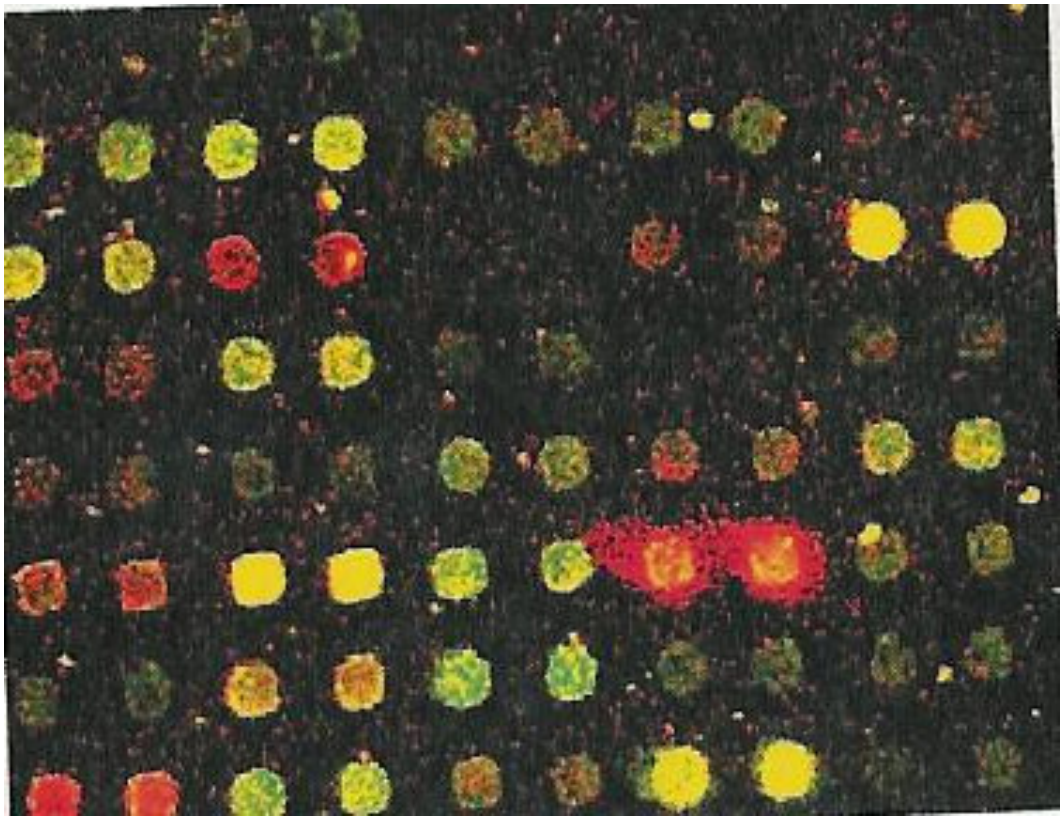
1. Data Acquisition and Visualization
   - Image quantification (spot reading)
   - Dynamic Range and spatial effects
   - Scatterplots
   - Systematic sources of error
2. Error models and data calibration
3. Identification of differentially expressed genes
   - Fold test
   - T-test
   - Correction for multiple testing

Microarray Data Analysis (discovery, next classes)

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

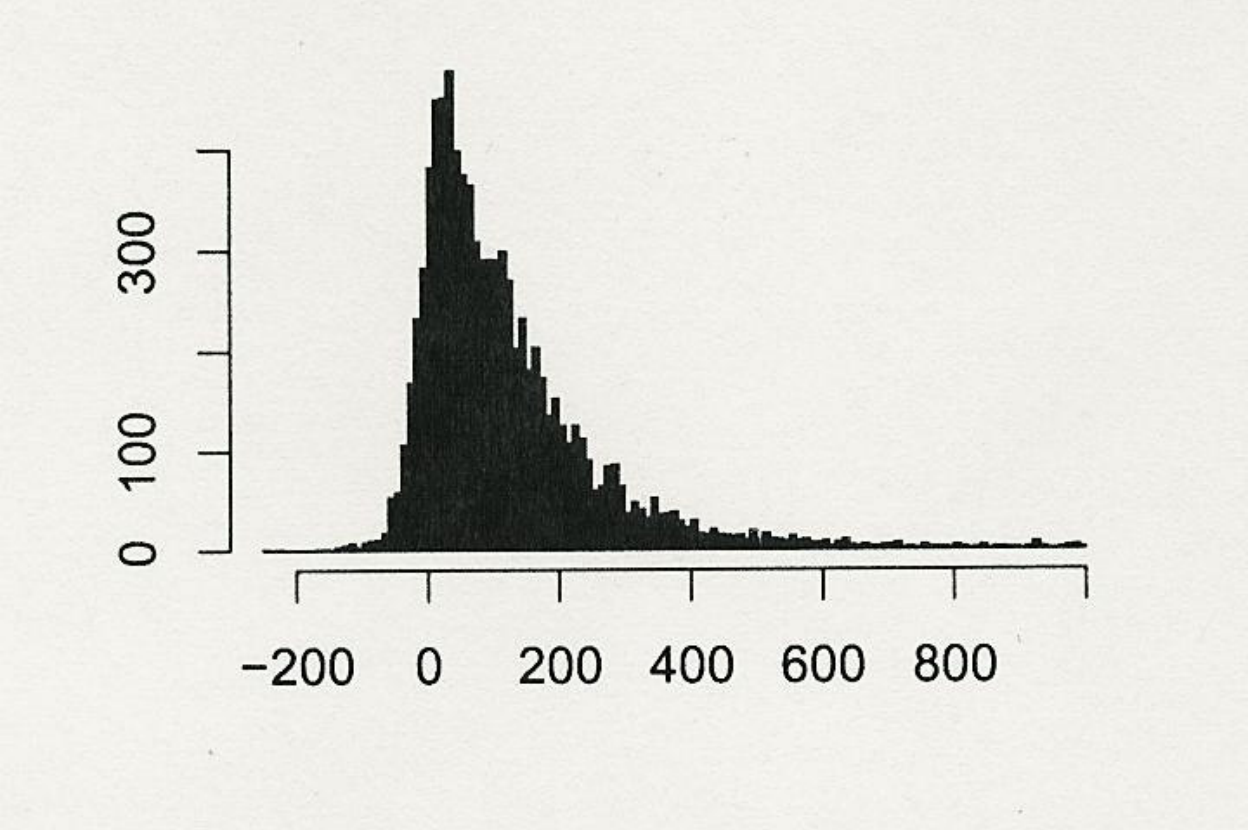
1. Data Visualization

- Image quantification (spot reading)

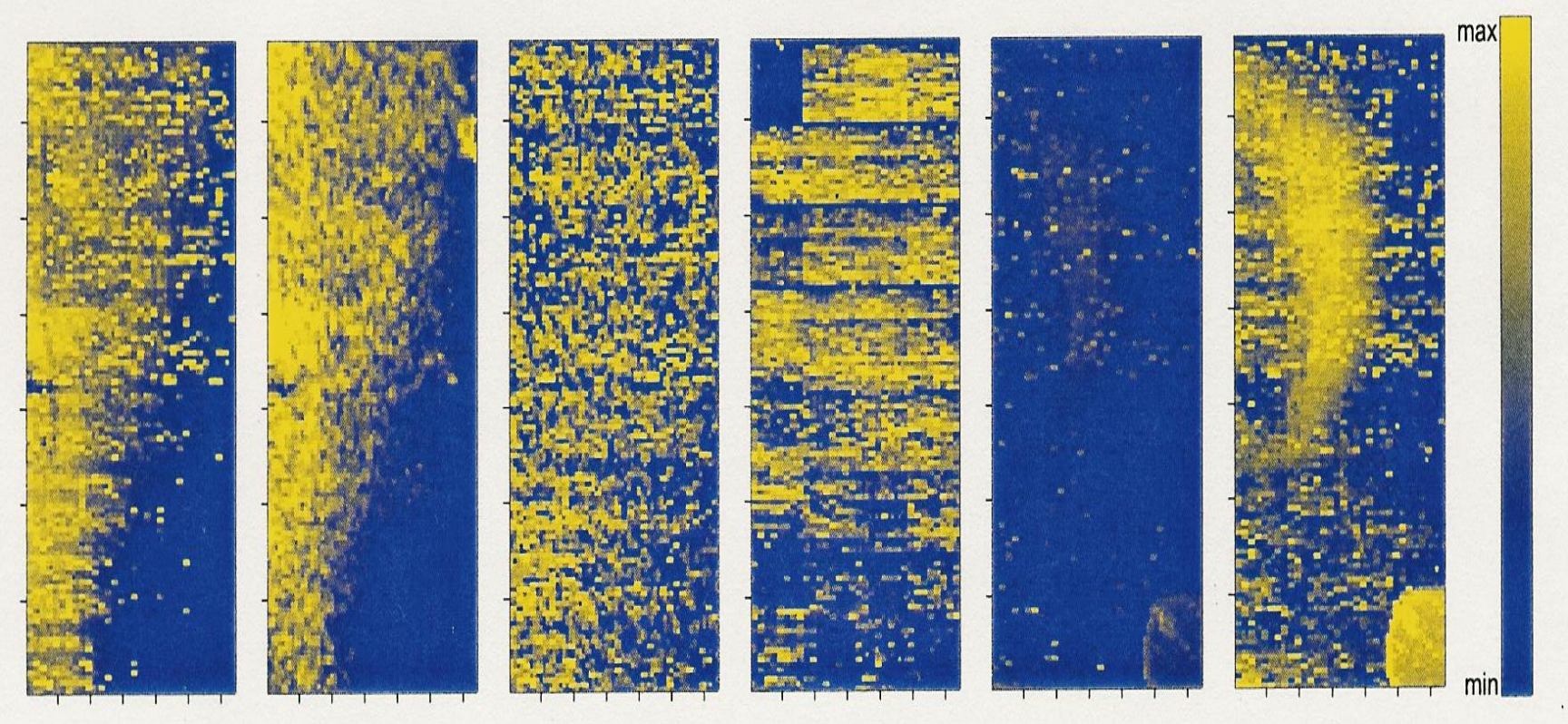


Huber et al

- Dynamic Range and spatial effects



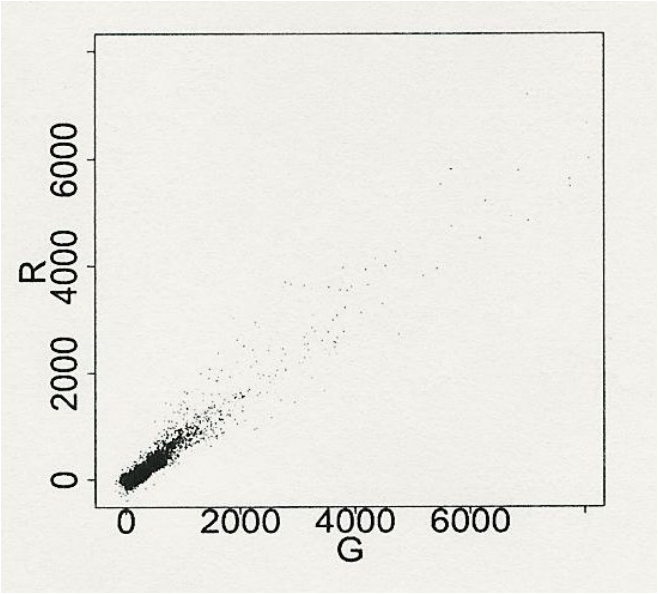
Huber et al



Huber et al

Scatterplots

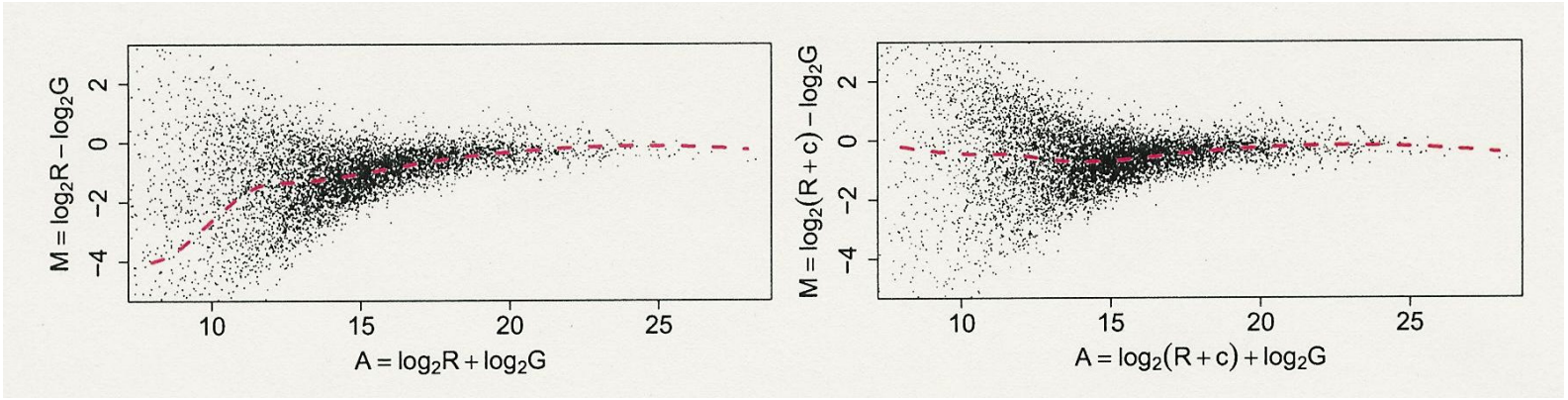
- Visual Aids for Data Calibration
- Plotting Red vs Green Expression



Huber et al

Scatterplots

- Plotting Average vs. Differential Expression
  - A = log R+log G
  - M = log R - log G
- Variance is increasing for low intensities, consequently it is difficult to capture lowly expressed genes



Huber et al

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

2. Error Models for Data Calibration (normalization)

- Identification and removal of systematic sources of variation
- Constant Variance across all intensities
- To allow within slide and between slide data comparison

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.

ε is a normally distributed random error

Realistic Assumptions in the Model Yield Better Normalization

$Y = \text{Measured intensity},\ x = \text{True abundance}$

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

$$b = e^{\eta}$$

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

- 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

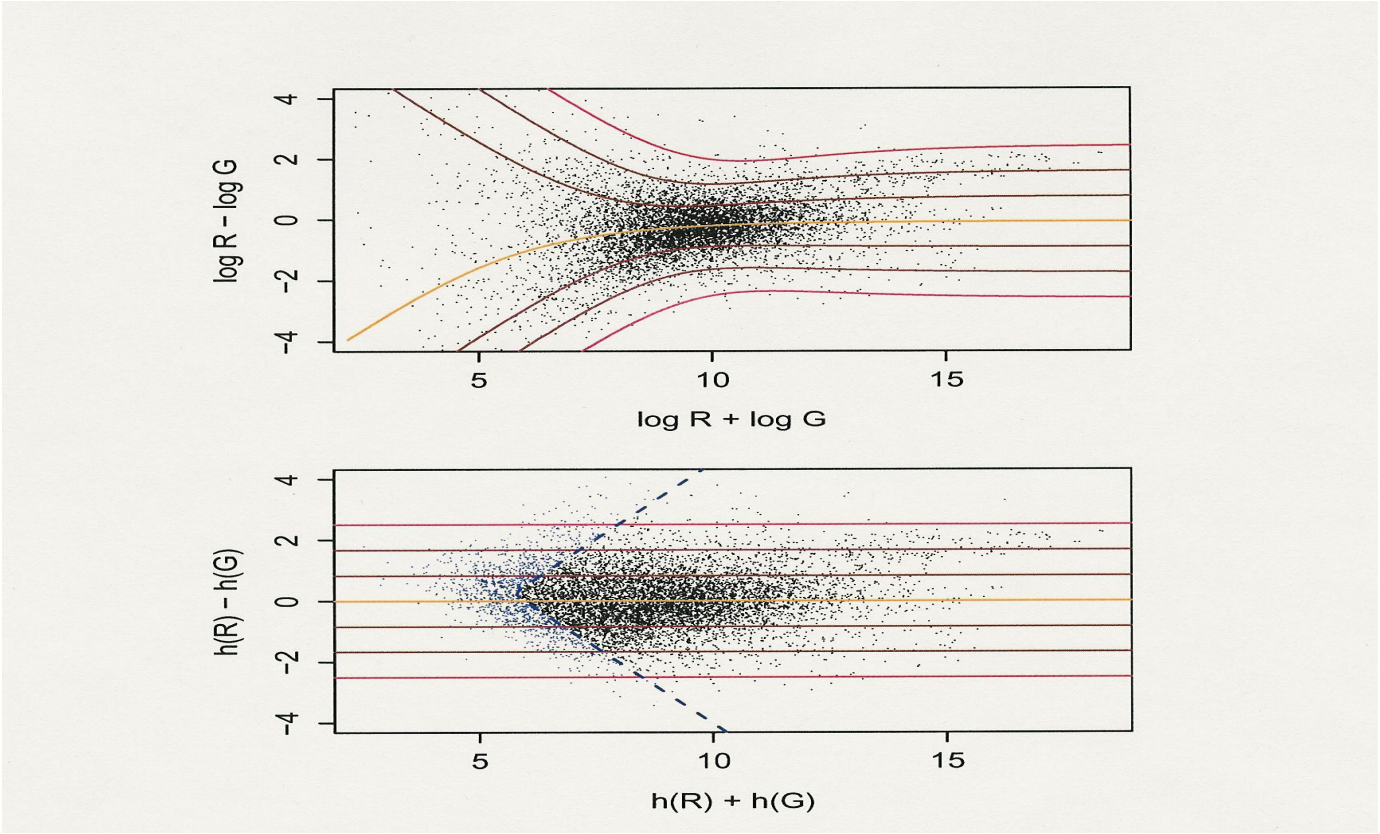
- Estimating the parameters of the model
- a, b, etc.
- Possible approaches:
  - least squares fit
  - Regression analysis

Consequences of the model

- *log Yr/Yg* is no longer the best estimator for *log* x_r/x_g.
- The appropriate measure of differential expression becomes

$$\Delta h = ar\sinh(\frac{\sigma_\varepsilon}{\sigma_\eta}\cdot\frac{Yr-a}{b}) - ar\sinh(\frac{\sigma_\varepsilon}{\sigma_\eta}\cdot\frac{Yg-a}{b})$$

This estimator has a constant variance across the range of intensities

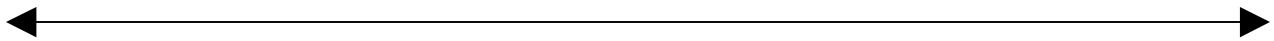


Huber et al

3. 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)

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, $j=1,\ldots,k$
- Let $x_1^c,\ldots,x_{nc}^c$ and $x_1^t,\ldots,x_{nt}^t$ be the normalized expression levels of n_c and n_t samples, respectively, in the control and test groups
- t-test for gene j

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

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

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 α, say 0.5, when testing 1000 genes for differential expression we get 5 hits at random: high amount of false positives
- Correcting for testing k hypothesis:

  Bonferoni correction:

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

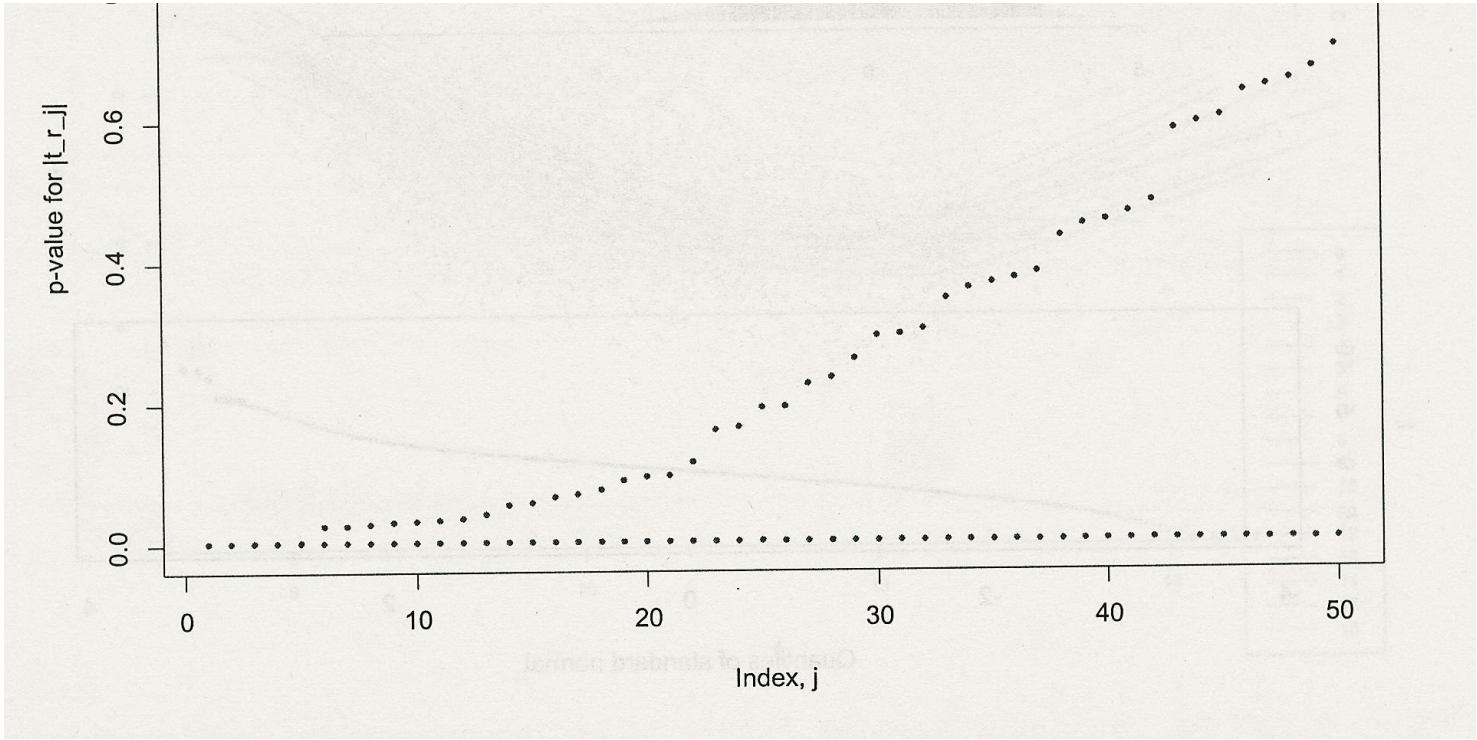
Alternatives to Bonferoni

- Bonferoni 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

Alternatives for Student-t for Small Number of Replicates

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

Adjusted vs. Unadjusted p-values



Dudoit et al

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

References:

- Lochart, Winzeler. “Genomics, gene expression and DNA arrays, Nature, 2000, v.405, 827-836
- Huber, et al. “Analysis of Microarray Gene Expression Data”, from http://www.dkfz-heidelberg.de/abt0840/whuber/publicat/hvhv.pdf
- Terry Speed’s Microarray Data Analysis Page: http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html
- David Rocke’s web page: http://www.cipic.ucdavis.edu/~dmrocke/