

# Microarray Data Preprocessing

## cDNA Spotted Microarray

國立臺灣大學 資訊所

Course: 生物資訊之統計與計算方法

2006/03/28

吳漢銘

[hmwu@stat.sinica.edu.tw](mailto:hmwu@stat.sinica.edu.tw)

<http://www.sinica.edu.tw/~hmwu/>

Institute of Statistical Science, Academia Sinica

中央研究院 統計科學研究所

# Outlines



2/41

## ■ Overview

## ■ Image Processing

- Addressing
- Segmentation
- Information extraction

## ■ Diagnostic Plots

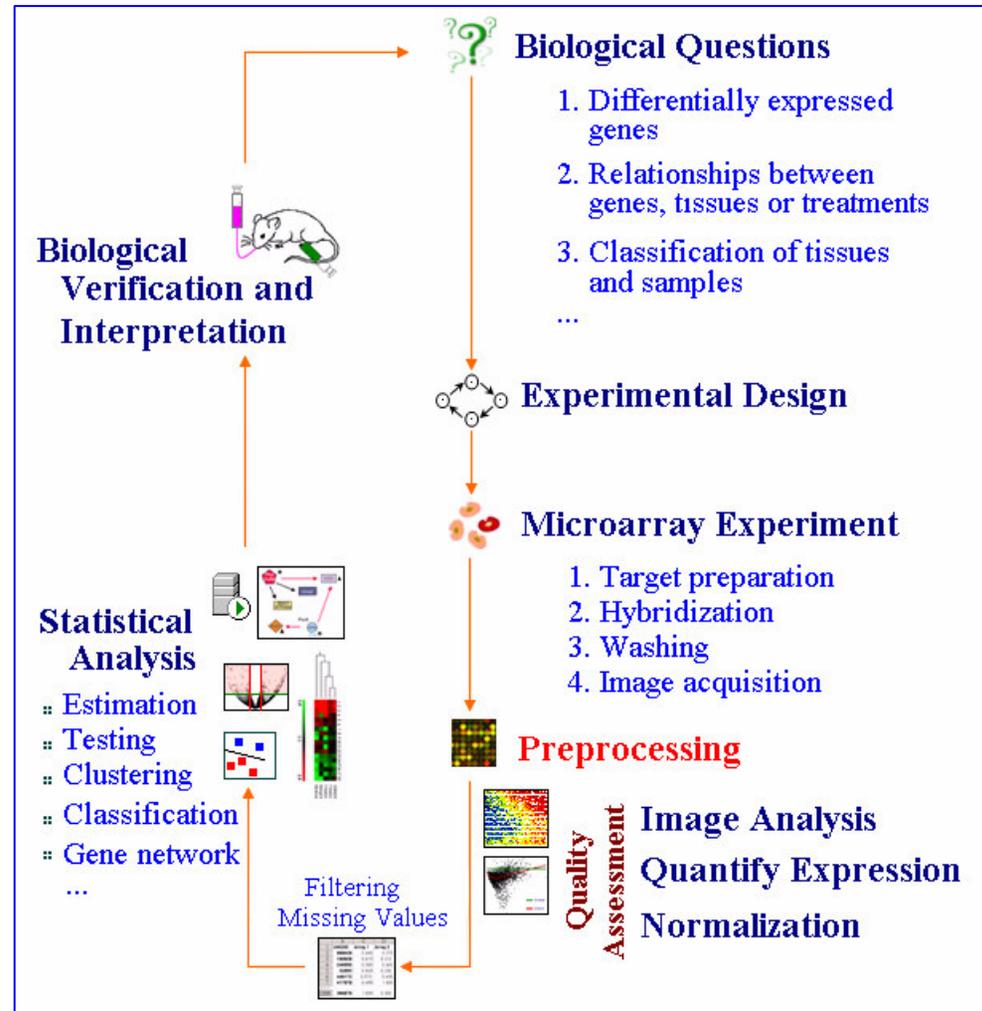
- Box plots, Histogram
- Array plot, MA plot

## ■ Normalization

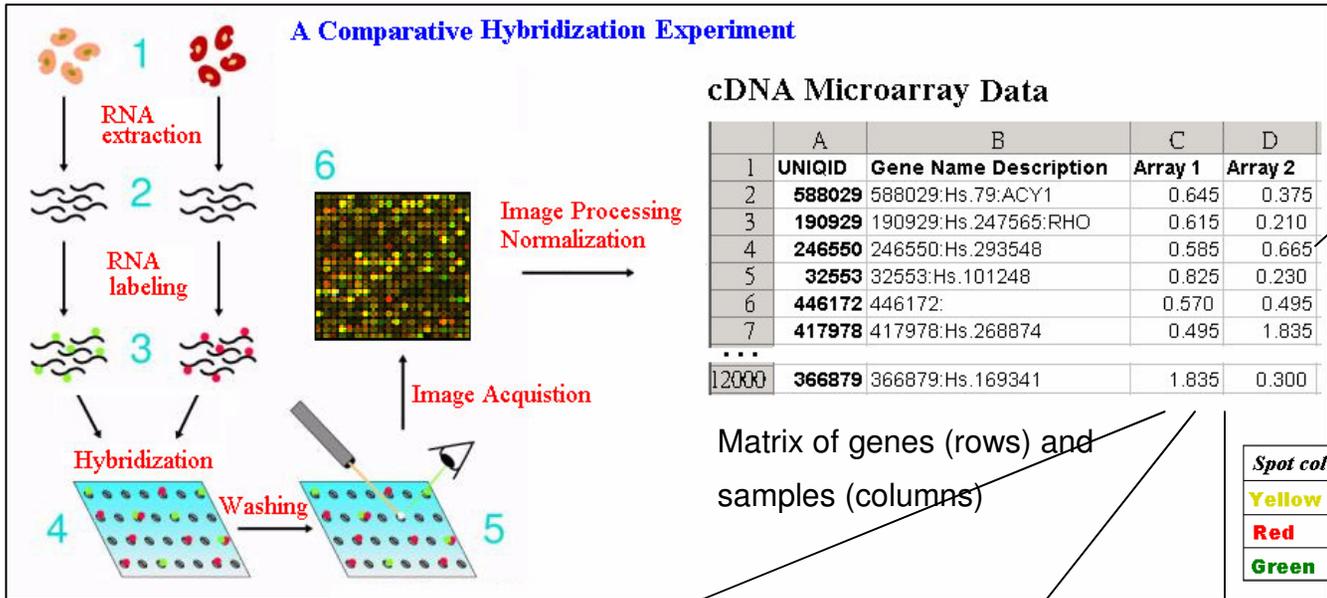
- Within-slides
- Between-slides
- Dye-swap

## ■ Software

## Microarray Life Cycle



# Overview



$$\log_2(\text{Cy5}/\text{Cy3})$$

$$R = R_f - R_b$$

$$G = G_f - G_b$$

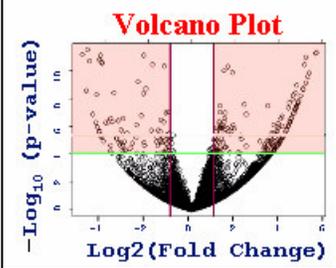
$$M = \log_2 R/G$$

$$A = 1/2 \log_2 RG$$

Spot color	Signal strength	Gene expression
Yellow	Control = Treated	Unchanged
Red	Control < Treated	Induced
Green	Control > Treated	Repressed

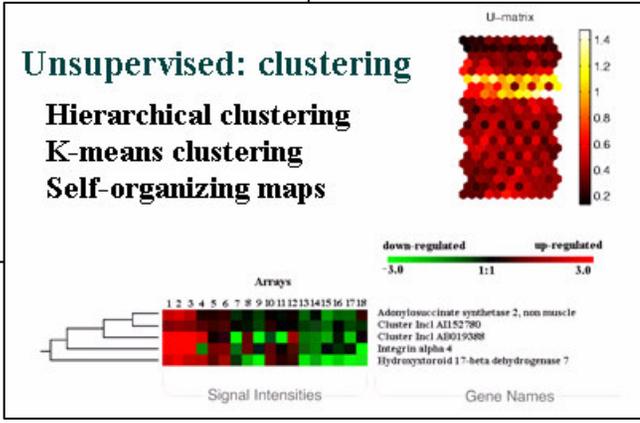
### Discovery of differentially expressed genes

- Parametric: t-test
- Non-parametric: Wilcoxon, Mann-Whitney test



### Unsupervised: clustering

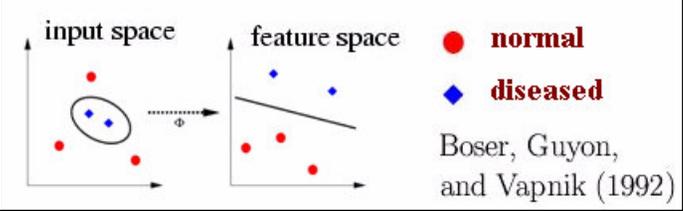
- Hierarchical clustering
- K-means clustering
- Self-organizing maps



### Supervised: classification

- Linear discriminants
- Decision trees
- Support vector machines

### Support Vector Classifiers



# An Example Image of a Complete Microarray

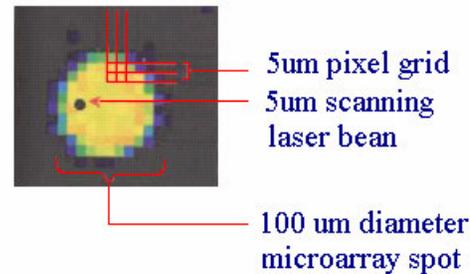
**Probe:** DNA spotted on the array, immobile substrate.

**Target:** DNA hybridized to the array, mobile substrate.

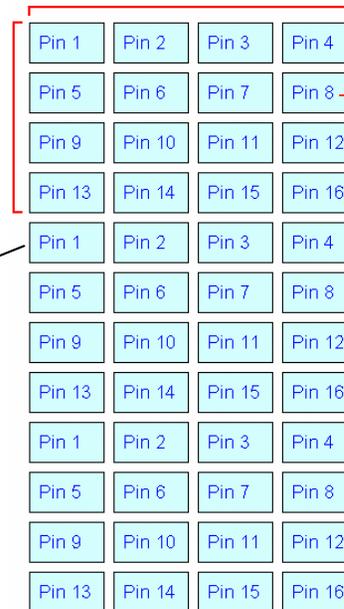
**Sector:** collection of spots printed using the same print-tip (or pin), print-tip-group, pin-group, spot matrix, grid.

**Batch:** collection of slides with the same probe layout.

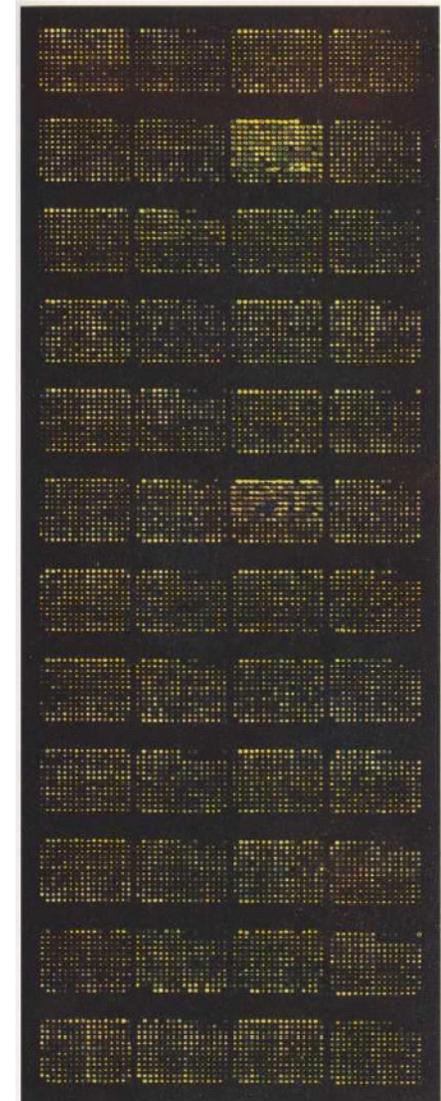
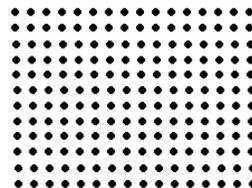
The terms slide or array are often used to refer to the printed microarray.



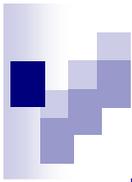
4x4 pattern corresponding to cassette is repeated 3 times on the array



12x16 pattern of features in each grid



1. 48 grids in a 12x4 pattern.
2. Each grid has 12x16 features.
3. Total 9216 features.
4. Each pin prints 3 grids.



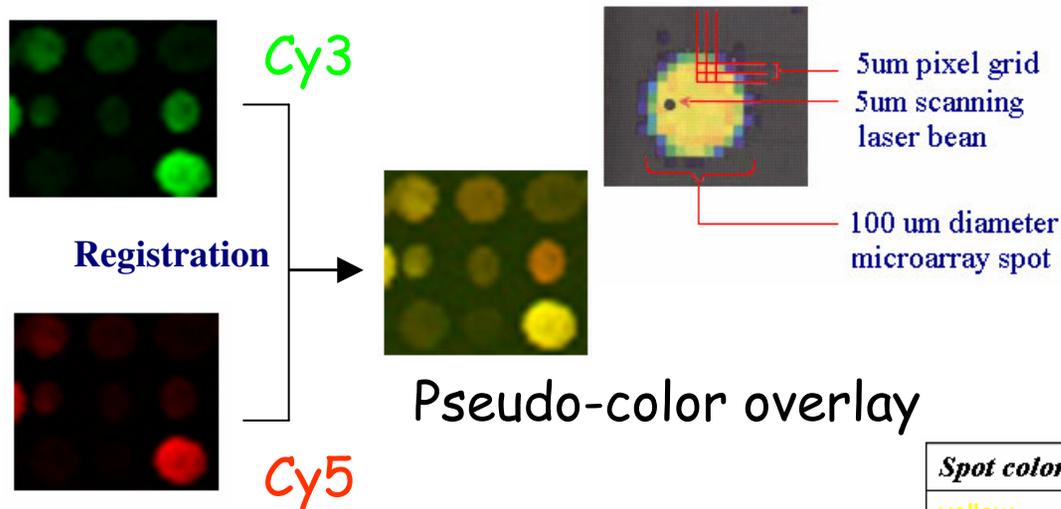
# Image Processing



The image of the array is the raw data.

- The total amount of hybridization for a spot is proportional to the *total fluorescence* generated by the spot.
- Spot intensity = sum of pixel intensities within the spot mask.

1. Identifying the Positions of the Features: **Addressing**
2. Identifying the Pixels That Comprise the Features: **Segmentation**
3. Identifying the Background Pixels.
4. Calculation of Numerical Information: **Information extraction**



## Resolution

- standard 10µm [currently, max 5µm]
- 100µm spot on chip = 10 pixels in diameter

## Image Format

- TIFF (tagged image file format) 16 bit (65,536 levels of gray)
- 1cm x 1cm image at 16 bit = 2Mb (uncompressed)
- other formats exist e.g. SCN (used at Stanford)

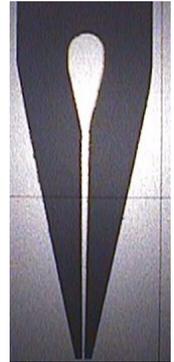
635 (F635, B635) refers to Cy5 (red channel)  
 532 (F532, B532) refers to Cy3 (green channel)

Spot color	Signal strength	Gene expression
yellow	Control = perturbed	Unchanged
red	Control < perturbed	Induced
green	Control > perturbed	Repressed

# 1. Identifying the Positions of the Features

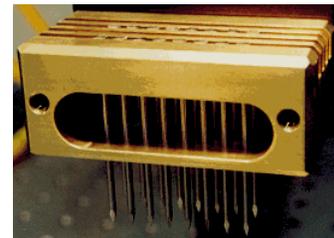
## Addressing/Gridding

- The features on most microarrays are arranged in a rectangular pattern.
- **grid**: larger spaces between the grids than between the features within each grid.
- The grids come about because there are several pins on the cassette on the spotting robot.
- The parameters associated with the grids:
  - How many grids in each direction (x and y).
  - How many features per grid in each direction (x and y).
  - The spacing between the grids.



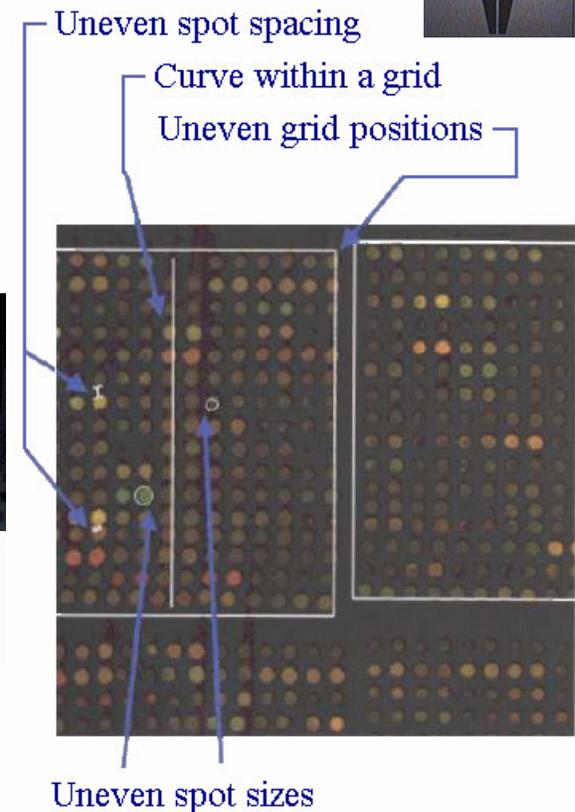
## The positions and sizes of the features within each grid may not be uniform:

- **Uneven grid positions**: this can happen if the pins are not perfectly aligned in the cassette.
- **Curve within a grid**: the glass slide is not completely horizontal or the pin has moved slightly in the cassette.
- **Uneven feature spacing**: the pins have moved slightly in the cassette or the surface of the glass is not completely flat.
- **Uneven feature size**: more or less fluid has been deposited on the glass during the manufacture of the array.



Statistics  
UC Berkeley

Images Source:



# 2. Identifying the Pixels: Segmentation

Classification of pixels as either foreground (signal) or background.

## ■ Fixed Circle Segmentation

- places a circle of fixed size over the region of the feature and uses all the pixels in the circle as those that form part of the feature.
- Problem: gives inaccurate results if the features are of different size. (most common)

## ■ Variable Circle Segmentation

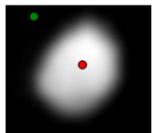
- fits a circle of variable size onto the region containing the feature.
- resolve features of different sizes, but perform less well on irregularly shaped features.

## ■ Histogram Segmentation

- fit a circle over the region of the feature and background and then looks at a histogram of the intensities of the pixels in the feature.
- The brightest and dimmest pixels are not used in the quantification of feature intensity.
- produces reliable results for irregularly shaped features.
- Histogram methods can be unstable for small features if the circular mask is too large.

## ■ Adaptive Shape Segmentation

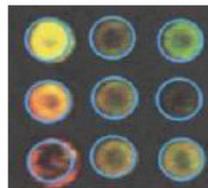
- A more sophisticated algorithm that can also resolve features with irregular shapes.
- The algorithm required a smaller number of seed pixels in the center of each feature to start.
- It then extends the regions of each feature by adjoining pixels that are similar in intensity to their neighbors.



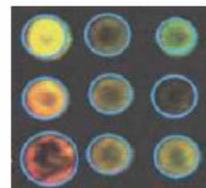
Histogram Segmentation



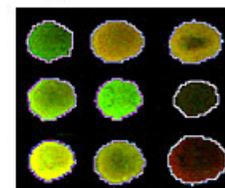
Fixed circle segmentation



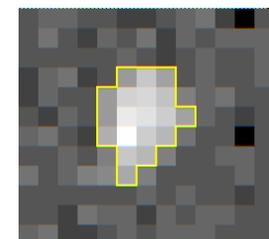
Variable circle segmentation



Adaptive shape segmentation



Limitation of circular segmentation



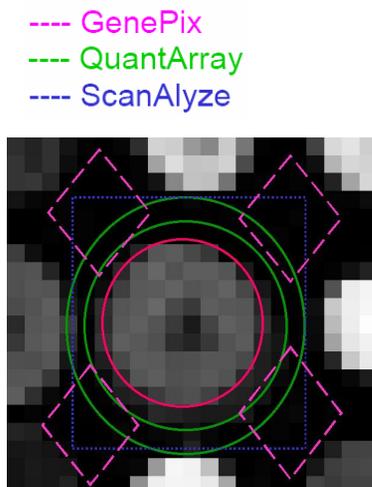
# 3. Identifying the Background Pixels



- The **signal intensity** of a feature includes contributions from non-specific hybridization and other fluorescence from the glass.
- It is usual to estimate this fluorescence by calculating the background signal from pixels that are near feature but are not part of any feature.

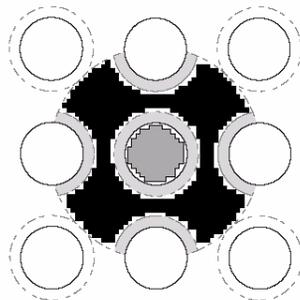
## 1. Local Background

- Focusing on small regions surrounding the spot mask.
- Median (mean) of pixel values in this region.
- By not considering the pixels immediately surrounding the spots, the background estimate is less sensitive to the performance of the segmentation procedure



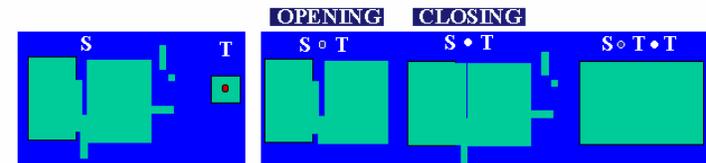
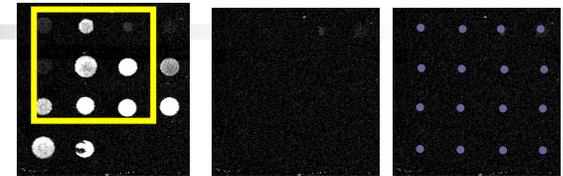
### GenePix Pro 6.0

- background pixels
- 2-pixel exclusion region
- feature pixels



## 2. Morphological Opening

- The image is probed with a structuring element, here, a square with side length about twice the spot to spot distance.
- Morphological opening: **erosion followed by dilation**.
- Erosion (Dilation): the eroded (dilated) value at a pixel  $x$  is the minimum (maximum) value of the image in the window defined by the structuring element when its origin is at  $x$ .
- Done separately for the red and green images.
- Produces an image of the estimated background for the entire slide.



## 3. Constant Background

- Global method which subtracts a constant background for all spots.
- Some evidence that the binding of fluorescent dyes to 'negative control spots' is lower than the binding to the glass slide
- More meaningful to estimate background based on a set of negative control spots
- If no negative control spots : approximation of the average background = third percentile of all the spot foreground values

## 4. No Background Adjustment

- Probably not accurate, but may be better than some forms of local background determination!

# Morphological Opening and Closing



## 外形影像處理

### 浸蝕運算 (Erosion)

《作用》浸蝕運算會將原影像去掉一層邊。只要影像中某一點  $x$  的周圍有一點是 0 則刪除  $x$  這一點，其運算如下：

$$G(x, y) = x \cap (x_0 \cap x_1 \cap x_2 \cap x_3 \cap x_4 \cap x_5 \cap x_6 \cap x_7)$$

$G(x, y)$ ：像素  $(x, y)$  的灰度值

點  $x$  和  $x_1$  至  $x_7$  的相對關係如下：

$x_3$	$x_2$	$x_1$
$x_4$	$x$	$x_0$
$x_5$	$x_6$	$x_7$

註： $x = 1$  代表物體； $x = 0$  代表背景。

### 增長運算 (Dilation)

《作用》增長運算會將原影像長一層邊。只要影像中某一點  $x$  的周圍有一點是 1 則輸出 1，其運算如下：

$$G(x, y) = x \cup (x_0 \cup x_1 \cup x_2 \cup x_3 \cup x_4 \cup x_5 \cup x_6 \cup x_7)$$

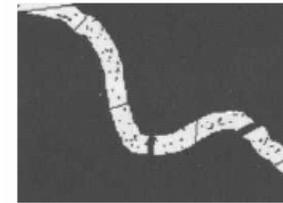
$G(x, y)$ ：像素  $(x, y)$  的灰度值

註： $x = 1$  代表物體； $x = 0$  代表背景。

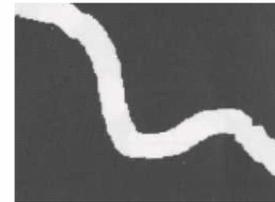
### 開閉運算 (Opening and Closing)

《作用》開運算為連續  $N$  次浸蝕運算後再加上  $N$  次增長運算；閉運算為連續  $N$  次增長運算後再加上  $N$  次浸蝕運算。

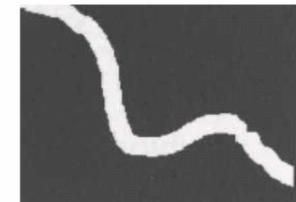
開運算可將影像中藕斷絲連的物體切開；閉運算則可補滿影像物體之間的細縫。



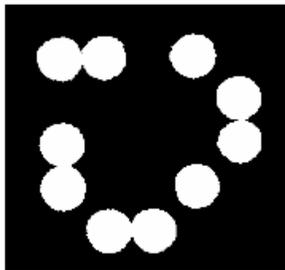
(a) original image



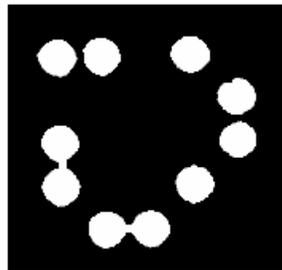
(b) 2 Dilation



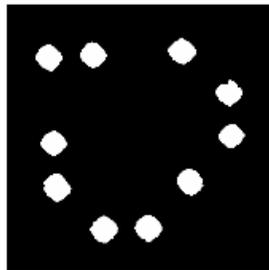
(c) 2 Erosion of (b)



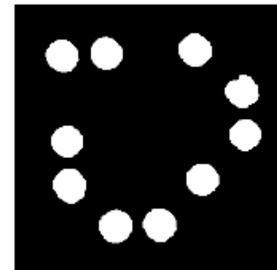
(a) original image



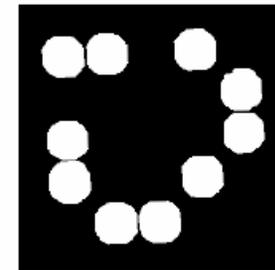
(b) 2 Erosion



(c) 5 Erosion



(d) 2 Dilation of (c)



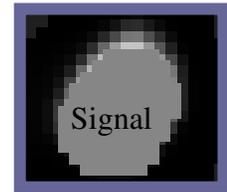
(e) 2 Dilation of (c)

# 4. Calculation of Numerical Information

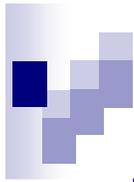


10/41

- Image-processing software will typically provide a number of measures:
  - Signal mean, Background mean, Signal median, Background median.
  - Signal standard deviation, Background standard deviation.
  - Diameter, Number of pixels.
  - **Flag**: a variable that is 0 if the feature is good, and will take different values if the feature is not good.
- **First important**: the measure of hybridization intensity for each feature.
  - It is preferable to use the median over the mean.
  - Median is more robust to outlier pixels than the mean: a small number of very bright pixels (arising from noise) have the potential to skew the mean, but will leave the median unchanged.
- **Second important**: signal standard deviation is used as a quality control for the array in two different ways.
  - **As measure of quality control of the feature**. If the standard deviation of a feature is greater than say 50% of the median intensity, the feature could be rejected as substandard.
  - **To determine whether an array is saturating**. The problem with saturated features is that we do not know the true intensity of the feature, and so it is not possible to use such features as part of a quantitative analysis of gene expression.
- **Third important**: flag
  - **Bad feature**: the pixel standard deviation is very high relative to the pixel mean.
  - **Negative feature**: the signal of the feature is less than the signal of the background.
  - **Dark feature**: the signal of the feature is very low.
  - **Manually flagged feature**: the user has flagged the feature using the image-processing software.



Background



# Quality Measurements



11/41

## ■ Spot Quality

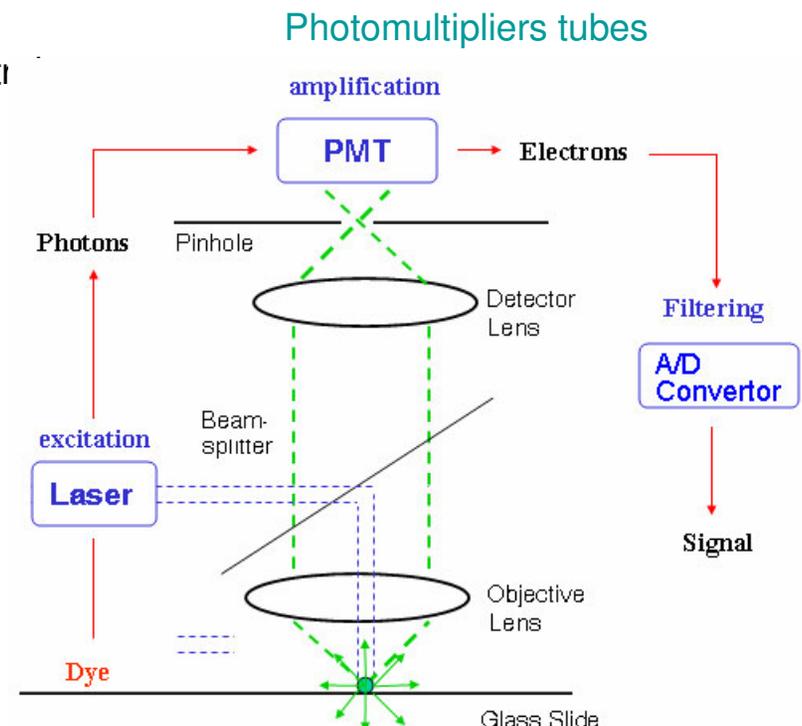
- **Brightness**: foreground/background ratio.
- **Uniformity**: variation in pixel intensities and ratios of intensities within spot.
- **Morphology**: Spot size, area, perimeter, circularity.
- Identification of “**bad spot**” (spots with no signal). Percentage of automatically flagged spots.
- **Background spatial variation**: variance of background median
- **Median signal intensity** for (1) empty, negative and positive controls spots. (2) difference between negative and positive spots.

## ■ Array Quality

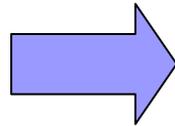
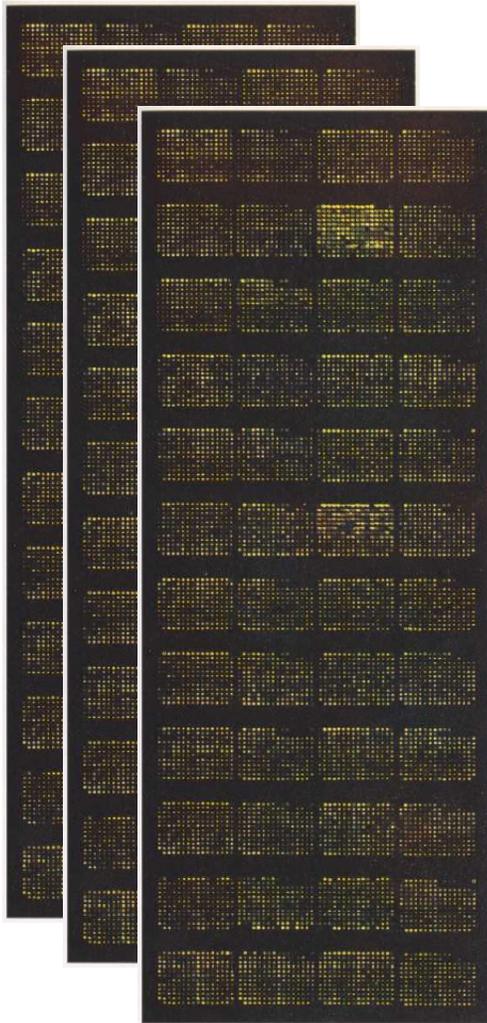
- **Controls variation**: variance of replicated controls
- Correlation between spot intensities.
- Percentage of spots with no signals.
- Distribution of spot signal area.
- Range of intensities

## ■ Ratio (2 spots combined)

- **Signal to Noise ratio**:  
 $\log_2(\text{fg.median} / \text{bg.median})$   
**GenePix**:  
 $(\text{Cy5FG.median} - \text{Cy5BG.median}) / \text{Cy5BG.SD}$
- Difference in PMT value
- Amount of normalization adjustment:  
M value adjustment at 25, 50, 75 percentile of A values.



# Example from GenePix



gal

```
ATF 1.0
27 82
"Type=GenePix Results 2"
"DateTime=ATF 1.0"
"Settings=27 82"
"GalFile="
"PixelFormat=ATF 1.0"
"Wavelength=Settings 27 82"
"ImageFile="GalFile="Type=GenePix Results 2"
"Normalization=DateTime=2004/09/16 10:43:58"
"JpegImage=Wavelength=Settings=D:\sher\UPV1\UPV1_gride.gps"
"RatioForm=ImageFile=GalFile=D:\sher\UPV1\UPV1.gal"
"Barcode=Normalization=PixelFormat=5"
"Background=Wavelengths=635 532"
"ImageOrigin=JpegImage=ImageFiles=D:\sher\UPV1\F22.tif 0 D:\sher\UPV1\F22.tif 1"
"NormalizationMethod=ImageOrigin=Barcode=NormalizationMethod=None"
"RatioFormulations=JpegImage=JpegOrigin=JpegOrigin=1100, 2300"
"Temperature=Creator=GenePix Pro 4.0.0.54"
"LinesAveraged=Scanner=GenePix 4000B [81990]"
"Comment=FocusPosition=0"
"ScanPower=PMTGain=Temperature=28.1"
"LaserOnTime=ScanPower=LinesAveraged=1"
"Supplier=LaserPower=Comment="
"Block" "Column" "Row" "Name" "ID" "X" "Y" "Dia." "F635 Median" "F635 Mea
1 1 1 "LHC-1" "control" 1340 2500 135 1086 1114 316 3
1 1 2 "Lambda13-2" "control" 1560 2530 110 352 358 1
1 1 3 1 "high-mobility group (nonhistone chromoso" "NM_002129" 1760 2
1 1 4 1 "KIAA0628 gene product" "NM_014789" 1965 2530 110 443 4
1 1 5 1 "protein kinase (cAMP-dependent, catalyti" "NM_032471" 2165 2
1 1 6 1 "Homo sapiens, clone IMAGE:3627860, mRNA," "BC006132" 2370 2
```

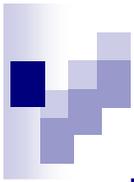
\*.gpr

\*.tif

- **F635 Median:** the median feature pixel intensity at wavelength #1 (635 nm).
- **B635 Median:** the median feature background intensity at wavelength #1 (635 nm).
- **F532 Median:** the median feature pixel intensity at wavelength #2 (532 nm).
- **B532 Median:** the median feature background intensity at wavelength #2 (532 nm).

## Quantification of Expression

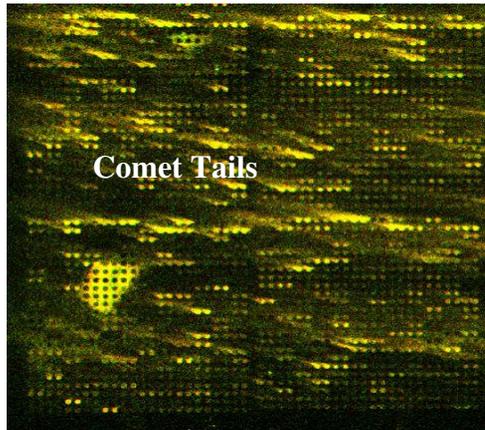
$$\begin{aligned} \text{Red intensity (Cy5)} &= R_{fg} - R_{bg} \\ \text{Green intensity (Cy3)} &= G_{fg} - G_{bg} \end{aligned} \Rightarrow \log_2 \text{ratio} = \text{Log}_2(\text{Cy5}/\text{Cy3})$$



# Practical Problems

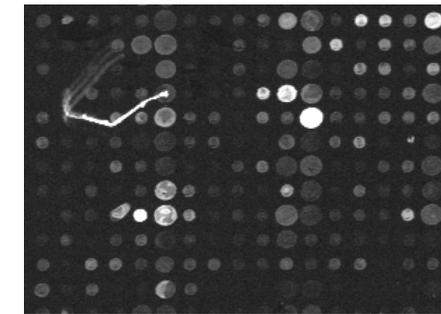
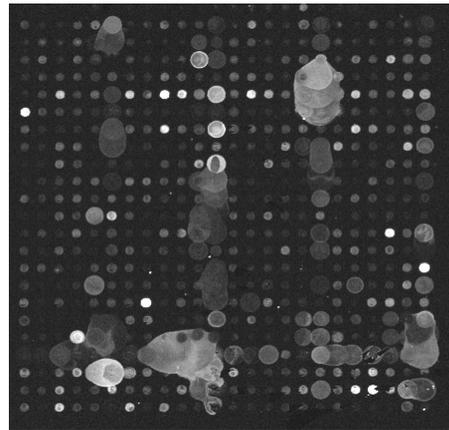


13/41



Comet Tails

**Comet Tails:** Likely caused by insufficiently rapid immersion of the slides in the succinic anhydride blocking solution.

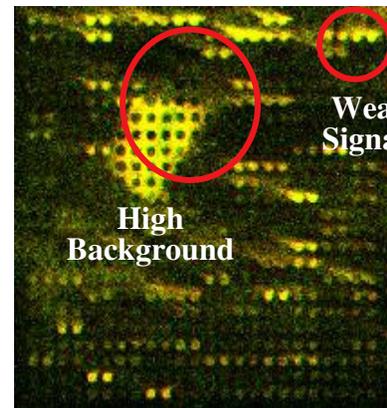


Dust



Spot overlap

**Spot overlap:** Likely cause: too much rehydration during post - processing.



High Background

Weak Signal

2 likely causes:

- Insufficient blocking.
- Precipitation of the labelled probe.

Source: Yang et al. (2000), TR 584, Statistics, UC-Berkeley.

More...

[http://www.corning.com/lifesciences/technical\\_information/techdocs/troubleshootingUltraGAPS\\_ProntoReagents.asp](http://www.corning.com/lifesciences/technical_information/techdocs/troubleshootingUltraGAPS_ProntoReagents.asp)

<http://www.asperbio.com/FAQ.htm>

Microarray Protocols <http://www.zmdb.iastate.edu/zmdb/microarray/protocols.html>

# Data Cleaning and Transformation



14/41

Ensure that the data is of high quality and suitable for analysis.

## Removing Flagged Features

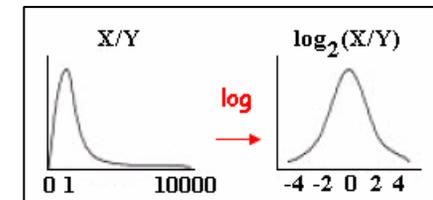
- Remove flagged features from the data set. (Sometimes removing potentially valuable data)
- Refer back to the original image of every flagged feature and to identify the problem that has resulted in flagging. (Require time and resources, not practical)

## Background Subtraction

- The background signal is thought to represent the contribution of **non-specific hybridization** of labeled target to the glass, as well as the **natural fluorescence of the glass** slide itself.
- Deal with negative signal:
  - Remove these features from the analysis (most common approach)  
the high background is taken to represent a local problem with that array, regarded as unreliable.
  - Use the lowest available signal-intensity measurement as the background subtracted intensity (typically value 1).  
idea behind: represents a gene with no or very low expression, and so the lowest value available used.
  - Estimate the true feature intensity (Bayesian).  
Assume: true feature intensity is higher than the background intensity, and so the high background represents some type of experimental error.

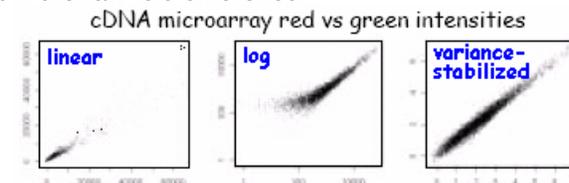
## Taking Logarithm

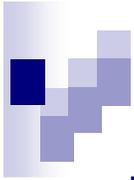
- The objective of log transformation:
  - There should be a reasonably even spread of features across the intensity range.
  - The variability should be constant at all intensity levels.
  - The distribution of experimental errors should be approximately normal.
  - The distribution of intensities should be approximately bell-shaped.
- Other transformation can be applied: aim at ensure the variability is constant at all intensity level.
- Logarithm of base 2
  - The ratio of the raw Cy5 and Cy3 intensities is transformed into the difference between the logs of the intensities of the Cy5 and Cy3 channels.
  - 2-fold up-regulated genes correspond to a log ratio of +1.
  - 2-fold down-regulated genes correspond to a log ratio of -1.
  - Genes are not differentially expressed have a log ratio of 0.
  - The log ratios have a natural symmetry, which reflects that biology and is not present in the raw fold difference.



## Quantification of Expression

$$\begin{aligned} \text{Red intensity (Cy5)} &= R_{fg} - R_{bg} \\ \text{Green intensity (Cy3)} &= G_{fg} - G_{bg} \end{aligned} \quad \rightarrow \quad \begin{aligned} &\log_2 \text{ ratio} \\ &= \text{Log}_2(\text{Cy5/Cy3}) \end{aligned}$$





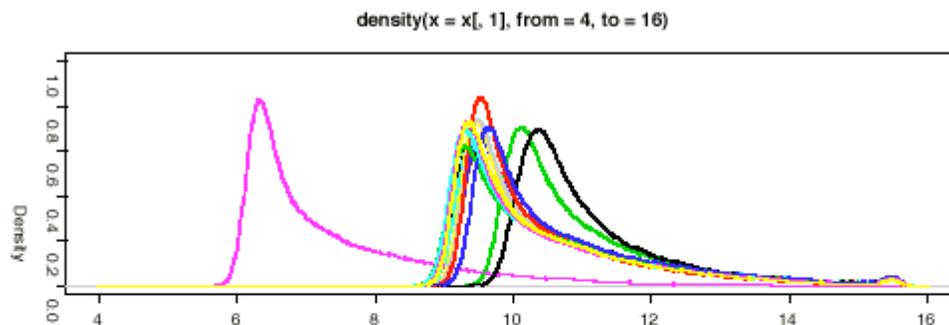
# Diagnostic Plots

## Plots of spot statistics

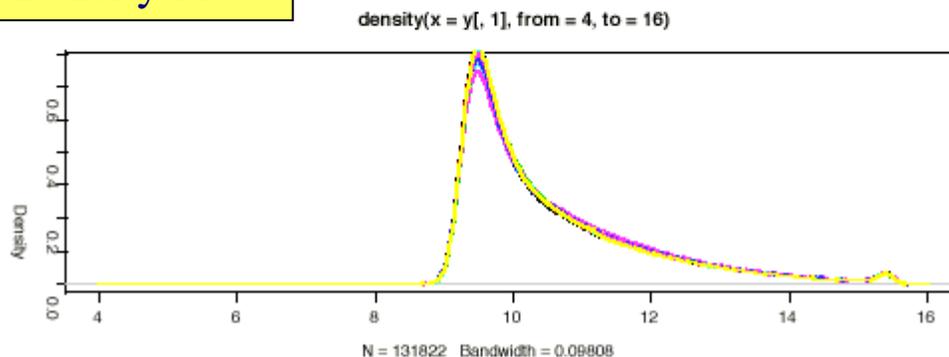
e.g. Red and green log-intensities, Intensity log ratio M, Average log-intensities A,...

## Stratify plots

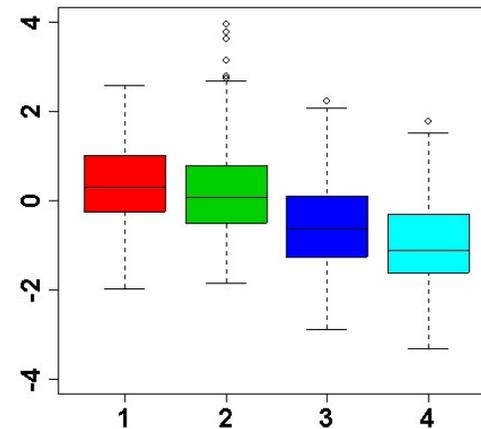
according to layout parameters, e.g. print-tip group, plates,...



Density Plots

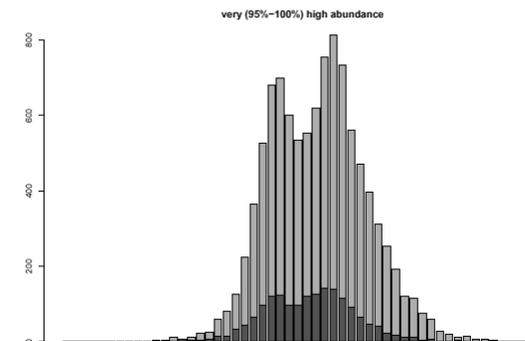


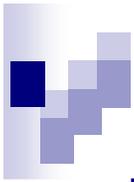
Box Plots



Boxplots within pin-groups

Histogram





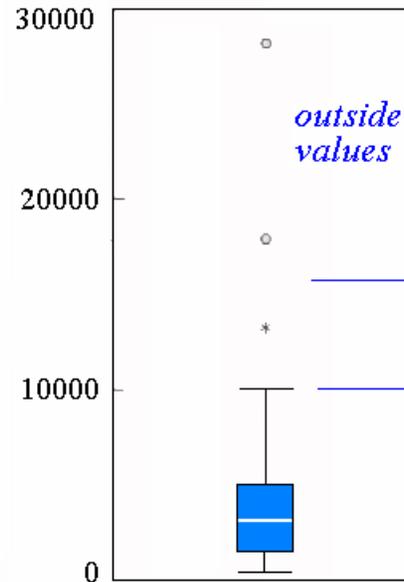
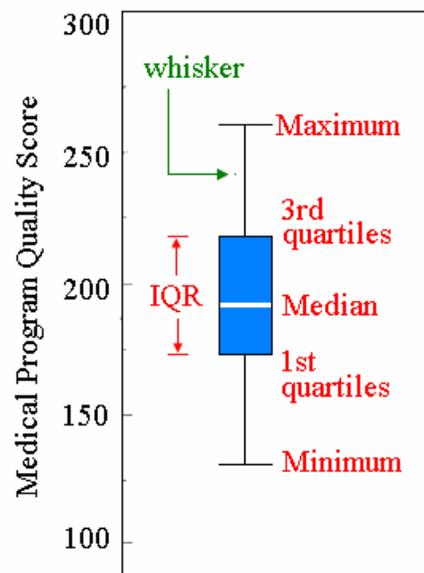
# Box Plots



- Box plots (Tukey 1977, Chambers 1983) are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups

## The box plot can provide answers to the following questions:

- Is a factor significant?
  - Does the location differ between subgroups?
  - Does the variation differ between subgroups?

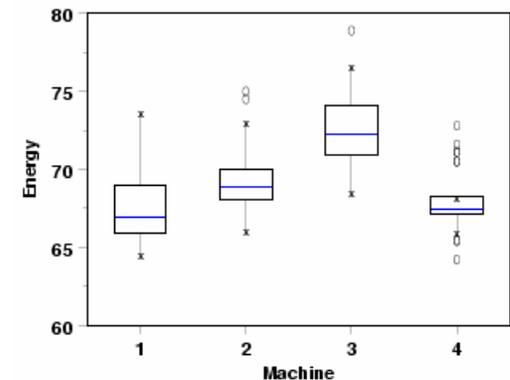


Upper Outer Fence:  
 $x_{0.75} + 3 \text{ IQR}$

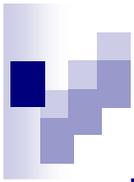
Upper Inner Fence:  
 $x_{0.75} + 1.5 \text{ IQR}$

Lower Inner Fence:  
 $x_{0.25} - 1.5 \text{ IQR}$

Lower Outer Fence:  
 $x_{0.25} - 3 \text{ IQR}$



Further reading: <http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm>



# Histogram

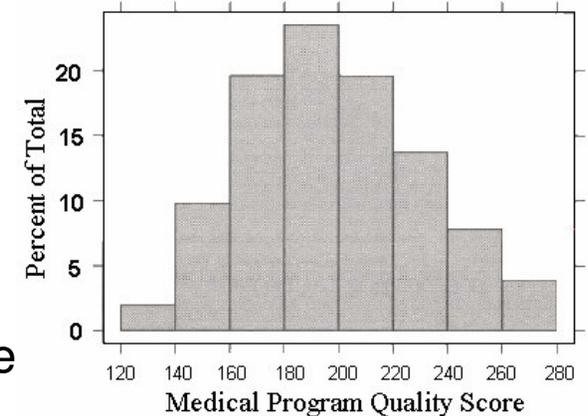


- $1/2h$  adjusts the height of each bar so that the total area enclosed by the entire histogram is 1.
- The area covered by each bar can be interpreted as the probability of an observation falling within that bar

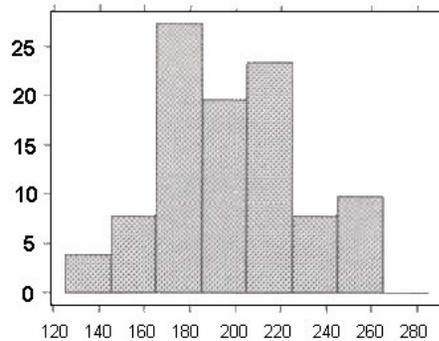
## Disadvantage for displaying a variable's distribution:

- selection of origin of the bins.
- selection of bin widths.
- the very use of the bins is a distortion of information because any data variability within the bins cannot be displayed in the histogram.

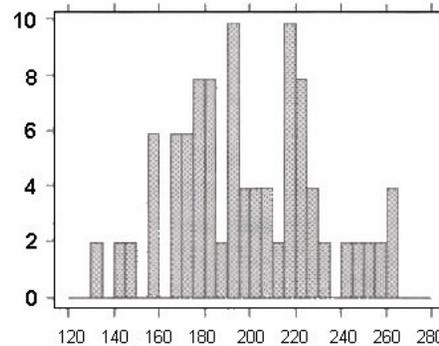
O. Bin origin at 120, bin widths of 20.



A. Bin origin at 125, bin widths of 20.



B. Bin origin at 120, bin widths of 5.



C. Bin origin at 120, bin widths of 10.

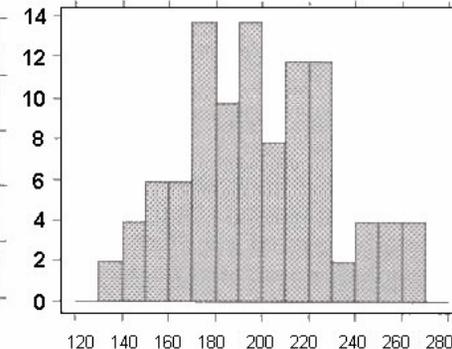
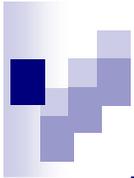
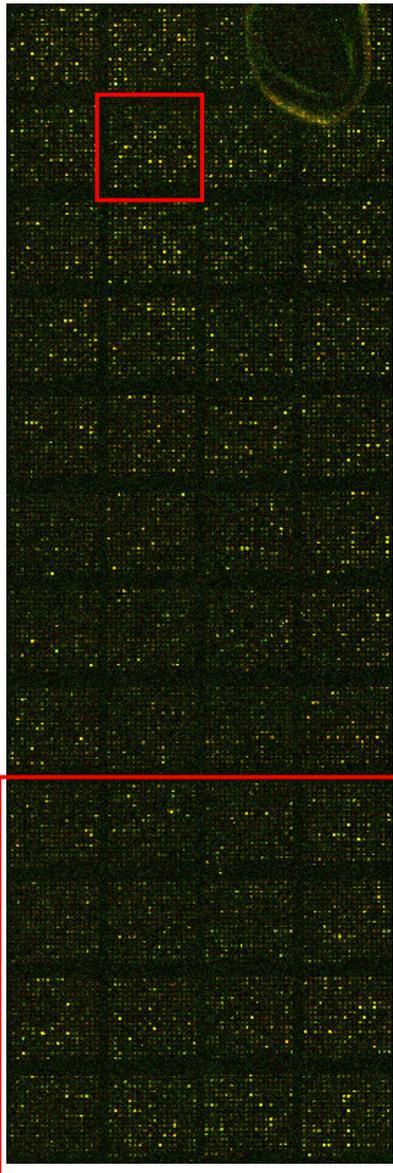


Figure Sources: Jacoby (1997).



# Array Image



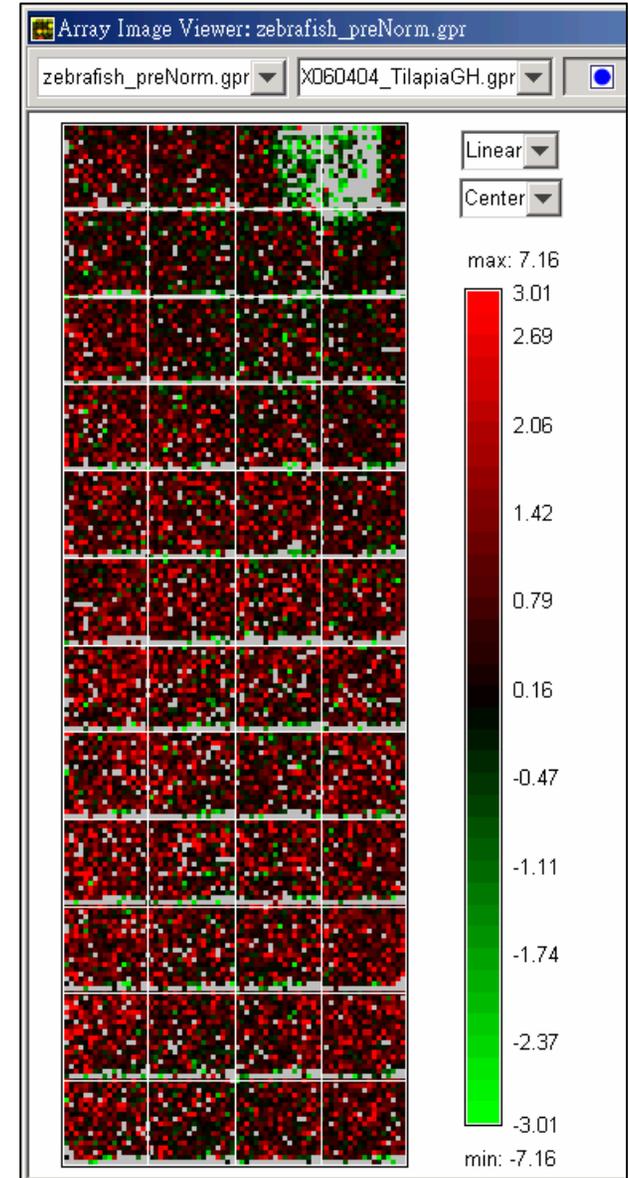
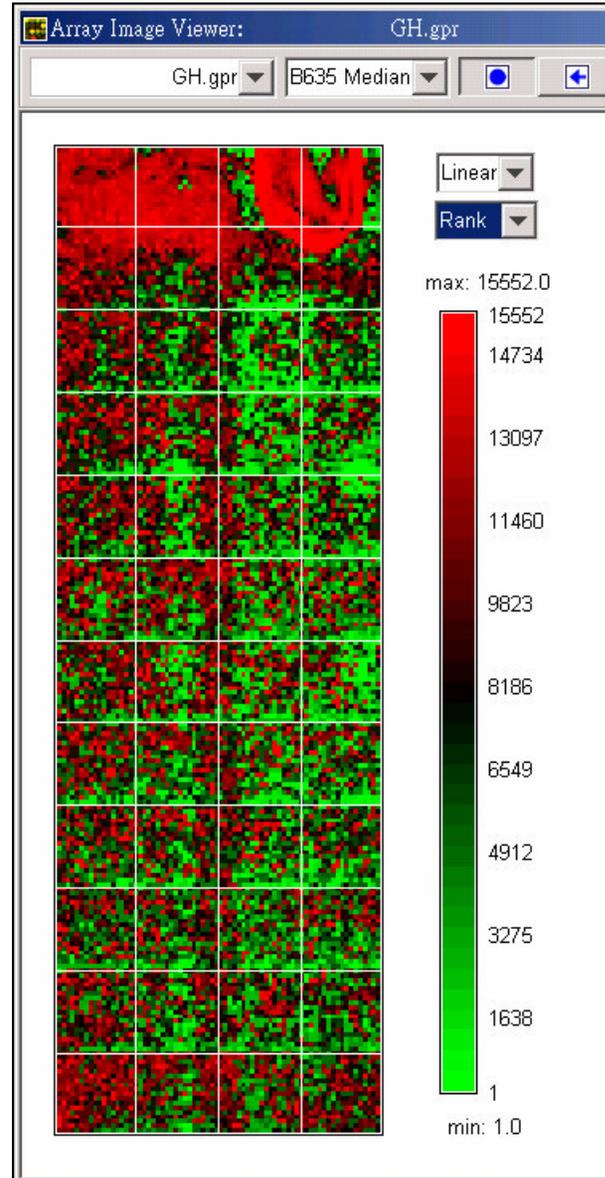
Blocks:  
12 by 4

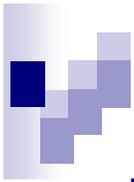
Features:  
18 by 18

Signal  
16-bit  
0~65535

\*.gpr

GAL





# Boxplots and Pin-group Effects

Before  
Normalization

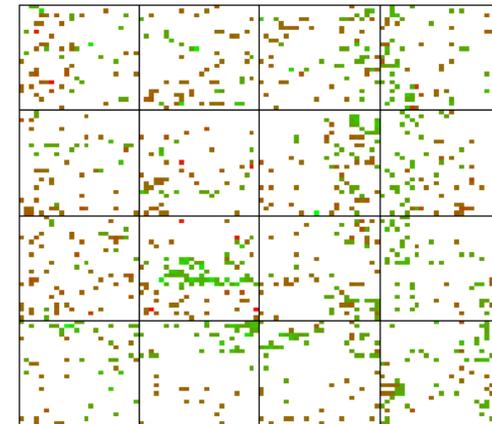
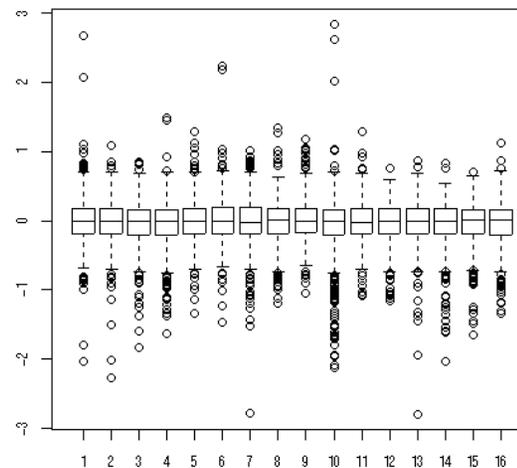
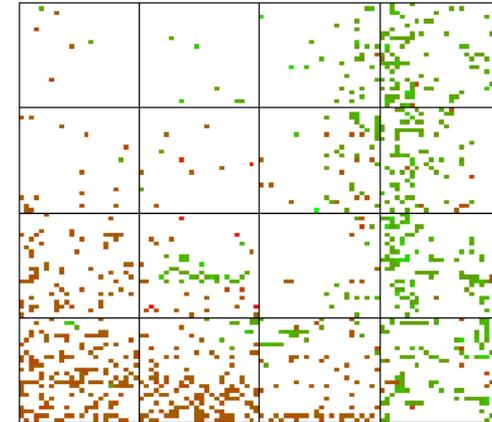
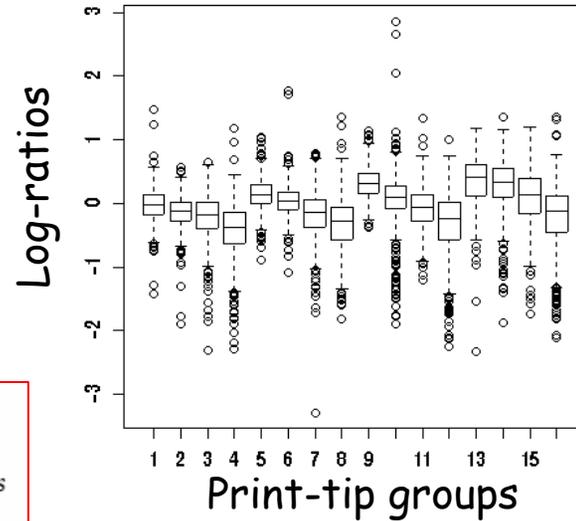
### Spatial Bias

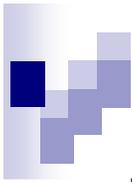
- patterns in the arrays:
- different concentrations of probes
  - plates sorted by biological role
  - uneven hybridization

After  
Normalization

Top 2.5% of ratios red

Bottom 2.5% of ratios green





# Scatterplot and MA plot



20/41

## ■ Features of scatter plot

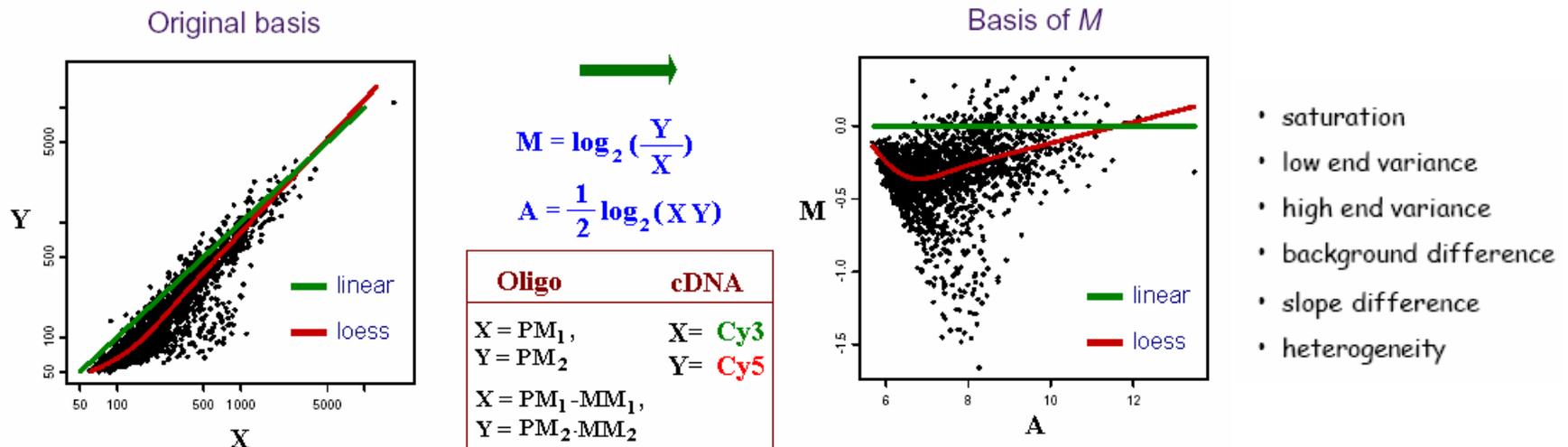
- the substantial correlation between the expression values in the two conditions being compared.
- the preponderance of low-intensity values. (the majority of genes are expressed at only a low level, and relatively few genes are expressed at a high level)

## ■ **Goals:** to identify genes that are differentially regulated between two experimental conditions.

## ■ **Outliers in logarithm scale**

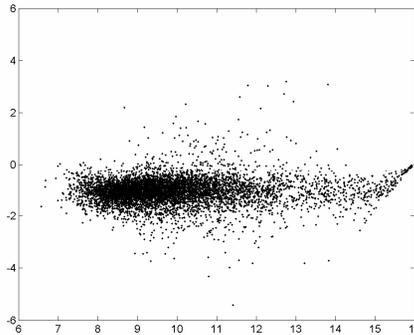
- spreads the data from the lower left corner to a more centered distribution in which the prosperities of the data are easy to analyze.
- easier to describe the fold regulation of genes using a log scale. In log<sub>2</sub> space, the data points are symmetric about 0.

**MA plots (Dudoit *et al.* 2002) can show the intensity-dependant ratio of raw microarray data.**

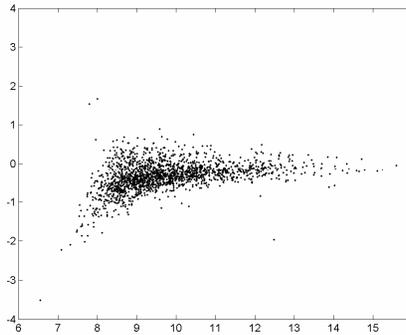


# Common Problems Diagnosed Using MA Plots

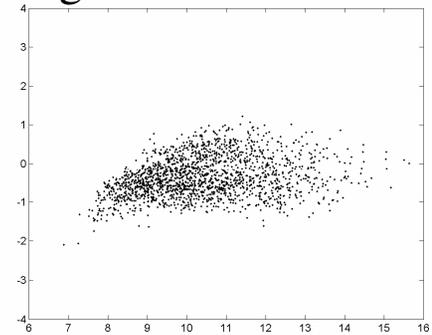
Saturation



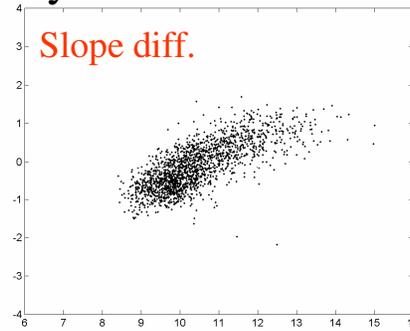
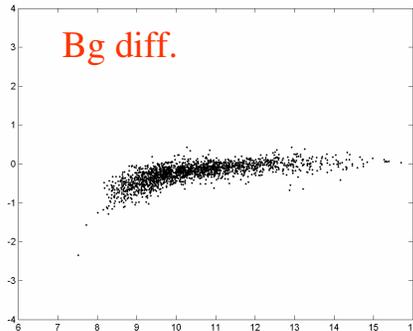
Low end variation



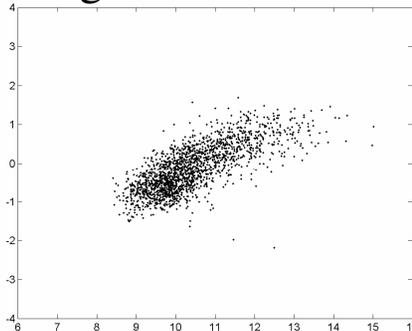
High end variation



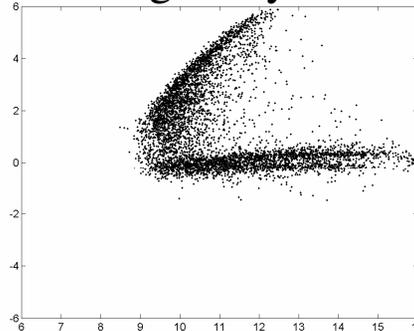
Curvature at Low intensity



Large curvature



Heterogeneity



$$Y_{ik} = a_i + b_i X_{ik} e^{(\eta_k + \zeta_{ik})} + \varepsilon_k + \delta_{ik}$$

$i$ , channel (1,2)

$k$ , gene (1 ... n genes)

$Y_{ik}$  = PMT of channel  $i$  and gene  $k$

$a_i$  = mean background of channel  $i$

$b_i$  = slope of channel  $i$

$X_{ik}$  = RNA abundance at channel  $i$  and gene  $k$

$\eta_k$  = common multiplicative error

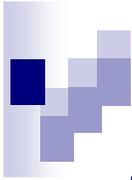
$\zeta_{ik}$  = multiplicative error of channel  $i$

$\varepsilon_k$  = common additive error

$\delta_{ik}$  = additive error of channel  $i$

Roche and Durbin (2001) J. Comp. Bio. 8:557-569

Source: Xiangqin Cui, Jackson Lab



# Normalization



22/41

## Sources of Variation

amount of RNA in the biopsy  
efficiencies of

- RNA extraction
- reverse transcription
- labeling
- photodetection

PCR yield  
DNA quality  
Spotting efficiency, spot size  
cross- or unspecific-hybridization  
stray signal

### Systematic → Normalization

- similar effect on many measurements
- corrections can be estimated from data

### Stochastic → Error Model

- too random to be explicitly accounted for
- noise

## What is normalization?

- Non-biological factor can contribute to the variability of data, in order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized.
- Normalization is a process of reducing unwanted variation across chips. It may use information from multiple chips.

**Dye effect:** Cy5 is usually more bleached than Cy3 (labeling efficiencies).

**Slide effect:** The normalization factor is slide dependent.  
**print-tip, spatial plate effects**

## Why normalization?

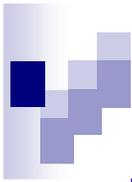
■ Normalization corrects for overall chip brightness and other factors that may influence the numerical value of expression intensity, enabling the user to more confidently compare gene expression estimates between samples. **Ensure that the data is of high quality and suitable for analysis.**

### Main idea

■ Remove the systematic bias in the data as completely possible while preserving the variation in the gene expression that occurs because of biologically relevant changes in transcription.

### Assumption

- The average gene does not change in its expression level in the biological sample being tested.
- Most genes are not differentially expressed or up- and down-regulated genes roughly cancel out the expression effect.



# Self-self Hybridizations

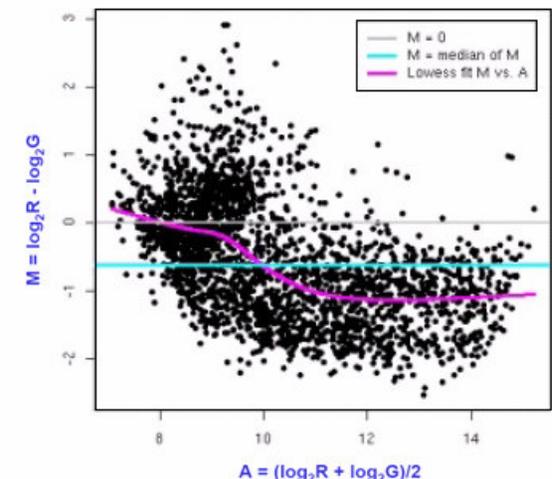
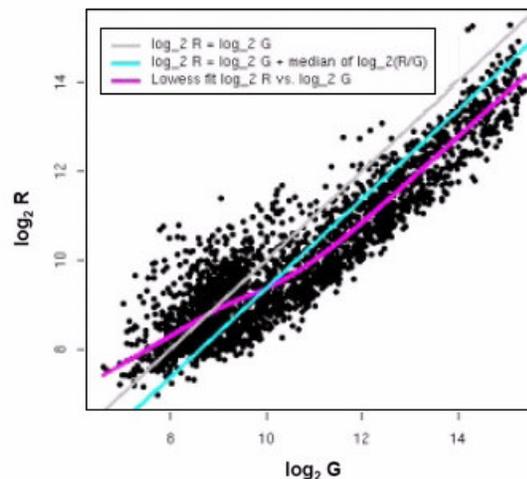


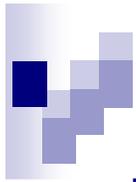
- The need for normalization can be seen most clearly in self-self hybridizations where the same mRNA sample is labeled with the Cy3 and Cy5 dyes.
- The imbalance in the red and green intensities is usually not constant across the spots within and between arrays, and can vary according to overall spot intensity, location, plate origin, etc. These factors should be considered in the normalization.
- By examining self-self hybridizations, where **no true differential expression** is occurring.
- There are dye biases which vary with spot intensity, location on the array, plate origin, pins, scanning parameters, (

## Self-self Hybridizations

### Minimizing Normalization

- dye swap strategies
- replicates
- reference samples
- controls
- error checking and quality control
- beautiful and consistent techniques
- sensible design of arrays and experiments





# Normalization Methods



24/41

## Within-Array Normalization: location

- Correcting for Different Responses of the Cy3 and Cy5 Channels
  - Linear Regression of Cy5 Against Cy3 (**Global Normalization**)
  - Linear Regression of Log Ratio Against Average Intensity
  - Nonlinear Regression of Log Ratio Against Average Intensity (**Lowess Normalization**)
- Correcting for Spatial Effects
  - Two-Dimensional Lowess Regression
  - Block-Block Loess Regression (**Within Print-tip Group Normalization**)

## Within-Array Normalization: scale

### Between-Array Normalization

(to enable comparison of multiple arrays)

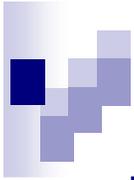
- Centering, Scaling, Distribution Normalization

### Paired-slides Normalization

(dye swap Experiments) (**Self-normalization**)

#### Which Genes To Use

- All genes on the array
- Constantly expressed genes (housekeeping genes)
- Controls (spiked controls or titration series of control sequences)



# Global Normalization

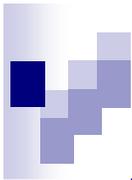


25/41

- The Cy3 and Cy5 are incorporated into cDNA with different efficiencies: without normalization, it would not be possible to accurately assess the relative expression of sample that are labeled with those dyes; genes that are actually expressed at comparable levels would have a ratio different than 1.
- **Constant normalization factor:**
  - **Cy5** and **Cy3** intensities are related by a constant factor,
    - $R = kG$ , and
    - the center of the distribution of log ratios is shifted to zero. (i.e. the average ratio for gene expression is 1. )

$$\log_2 R/G \Rightarrow \log_2 R/G - c = \log_2 R/(kG)$$

- $c = \mathbf{median}$  or  $\mathbf{mean}$  of log ratios for a particular gene set (e.g. housekeeping genes).
- $k = \sum R_i / \sum G_i$  **total intensity**.

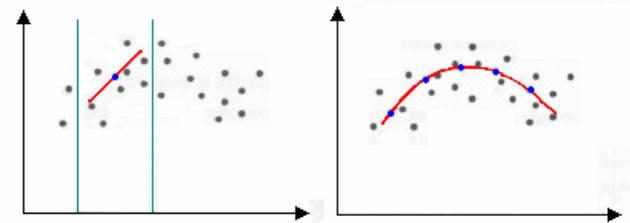


# Loess Normalization



- Loess normalization (Bolstad *et al.*, 2003) is based on **MA plots**. Two arrays are normalized by using a loess smoother.
- Skewing** reflects experimental artifacts such as the
  - contamination of one RNA source with genomic DNA or rRNA,
  - the use of unequal amounts of radioactive or fluorescent probes on the microarray.
- Skewing can be corrected with local normalization: fitting a local regression curve to the data.

Loess regression  
(locally weighted polynomial regression)



1. For any two arrays  $i, j$  with probe intensities  $x_{ki}$  and  $x_{kj}$  where  $k = 1, \dots, p$  represents the probe

2. we calculate

$$M_k = \log_2(x_{ki}/x_{kj}) \text{ and } A_k = \frac{1}{2} \log_2(x_{ki}x_{kj}).$$

3. A normalization curve is fitted to this  $M$  versus  $A$  plot using loess.

Loess is a method of local regression (see Cleveland and Devlin (1988) for details).

4. The fits based on the normalization curve are  $\hat{M}_k$

5. the normalization adjustment is  $M'_k = M_k - \hat{M}_k$ .

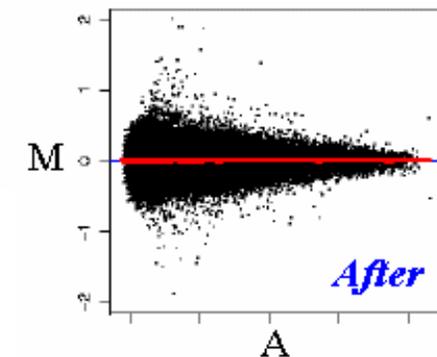
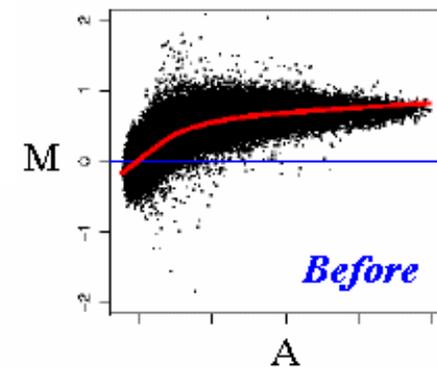
6. Adjusted probe intensities

$$\text{are given by } x'_{ki} = 2^{A_k + \frac{M'_k}{2}} \text{ and } x'_{kj} = 2^{A_k - \frac{M'_k}{2}}.$$

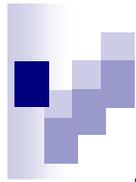
$$M = \log_2\left(\frac{Y}{X}\right)$$

$$A = \frac{1}{2} \log_2(XY)$$

Oligo	cDNA
X = PM <sub>1</sub> ,	X = Cy3
Y = PM <sub>2</sub>	Y = Cy5
X = PM <sub>1</sub> · MM <sub>1</sub> ,	
Y = PM <sub>2</sub> · MM <sub>2</sub>	



$$\log_2 R/G \Rightarrow \log_2 R/G - c(A) = \log_2 R/[k(A)G].$$



# Rank-invariant Method



27/41

**Housekeeping Genes:** beta-actin, GAPDH.

■ Each gene expression value in a single array experiment is divided by the mean expression values of these housekeeping genes.

■ **Assumption:** such genes do not change in their expression values between two conditions.

□ Human Gene Expression (HuGE) database: 7000 gene in 19 tissues: 451 housekeeping genes are commonly expressed across all these tissues.

**Rank-invariant method** (Schadt et al. 2001, Tseng et al. 2001):

- If a particular gene is up- or down- regulated, then its Cy5 rank among whole genome will significantly different from Cy3 rank.
- Iterative selection helps to select a more conserved invariant set when number of genes is large.

■ When the number of genes is small such as the 125 gene

$$S_g = \{g: \text{abs}[\text{rank}(\text{Cy5}_g) - \text{rank}(\text{Cy3}_g)] < 5\}$$

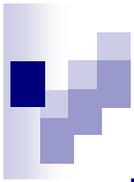
$$S = \{g: |\text{rank}(\text{Cy5}_g) - \text{rank}(\text{Cy3}_g)| < d \ \& \ l < \text{rank}[(\text{Cy5}_g + \text{Cy3}_g)/2] < G - l\}$$

■ If the number of genes is large

$$S_0 = \{g: |\text{rank}(\text{Cy5}_g) - \text{rank}(\text{Cy3}_g)| < p \times G \ \& \ l < \text{rank}[(\text{Cy5}_g + \text{Cy3}_g)/2] < G - l\}$$

$$S_i = \{g: g \in S_{i-1} \ \& \ |\text{rank}_{g \in S_{i-1}}(\text{Cy5}_g) - \text{rank}_{g \in S_{i-1}}(\text{Cy3}_g)| < p \times |S_{i-1}|\}$$

where  $|S_i|$  is the number of genes in set  $S_i$ .



# Two-Dimensional Lowess Regression

It fits a two-dimensional polynomial surface to the data.

1. Calculate the log ratio for each feature on the array.
2. Produce a pseudo image of the log ratios of the features as a function of the x and y coordinates of the features on the array.
3. Perform a two-dimensional Loess fit of the log ratios as a function of the x and y coordinates of the features.
4. For each feature, calculate the normalized log ratio by subtracting the fitted value on the Loess surface from the raw log ratio.

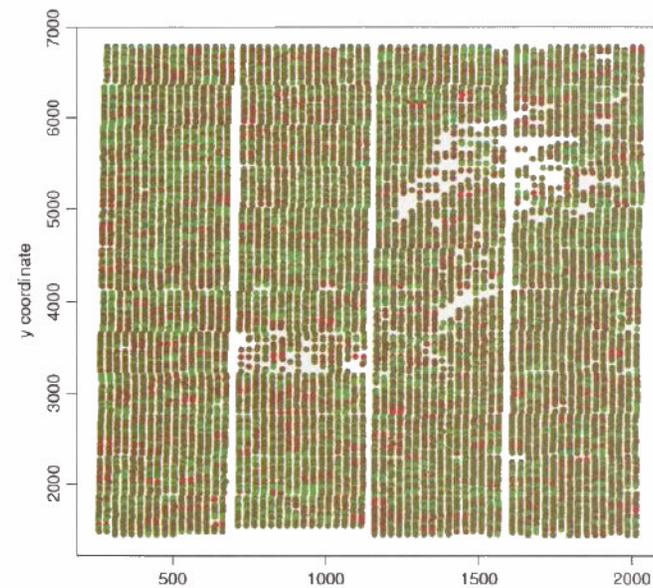
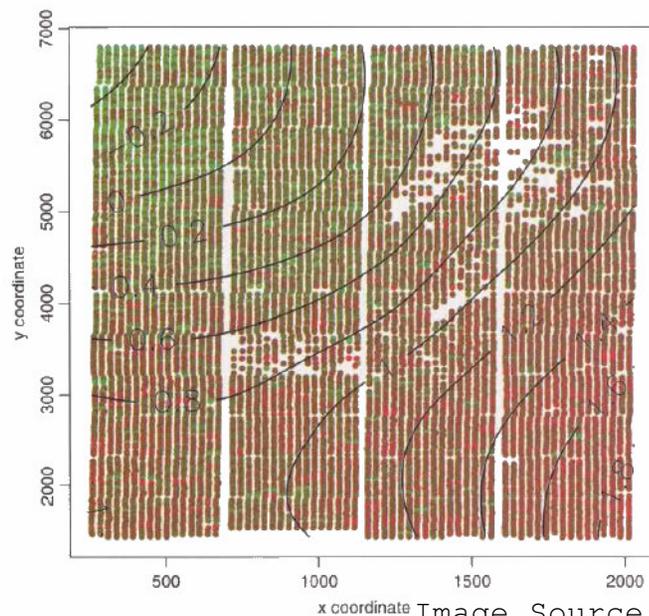


Image Source: Stekel, D. (2003). Microarray bioinformatics.

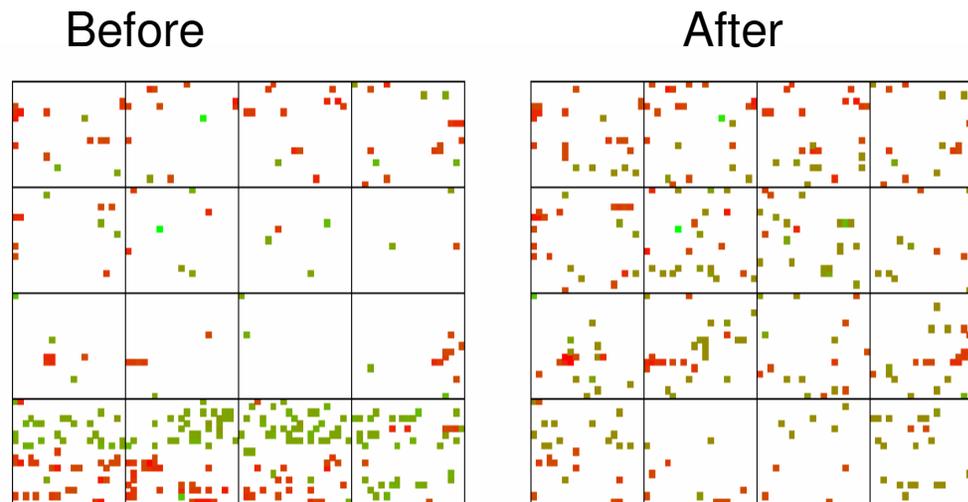
# Within Print-tip Group Normalization

## (print tip + A)-dependent normalization

- Most normalization methods do not correct for spatial effects produced by hybridization artifacts or print-tip or plate effects during the construction of the microarrays. It is possible to correct for both print-tip and intensity-dependent bias by performing LOWESS fits to the data within print-tip groups.

$$\log_2 R/G \Rightarrow \log_2 R/G - c_i(A) = \log_2 R/[k_i(A)G].$$

where  $c_i(A)$  is the LOWESS fit to the MA-plot for the  $i$ th grid only.



# Within-Array Normalization: Scale

- After within-print-tip-group normalization, all the normalized log-ratios from the different print-tip groups will be centered around zero.
- it is possible that the log-ratios from the various print-tip groups have different spread and some scale adjustment is required.

## Assumption

- a relatively small proportion of the genes will vary significantly in expression between the two mRNA samples.
- the spread of the distribution of the log-ratios should be roughly the same for all print-tip groups.

$M_{ij}$  denotes the  $j$ th log-ratio in the  $i$ th print-tip group,  $j = 1, \dots, n_i$ .

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

where  $\sigma^2$  is the variance of the true log-ratios

$a_i^2$  is the scale factor for the  $i$ th print-tip group.

$$\sum_{i=1}^I \log a_i^2 = 0$$

Estimation 1:

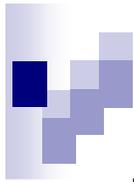
$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[ I ]{\prod_{k=1}^I \sum_{j=1}^{n_k} M_{kj}^2}}$$

Estimation 2:

$$\hat{a}_i = \frac{MAD_i}{\sqrt[ I ]{\prod_{i=1}^I MAD_i}}$$

median absolute deviation  $MAD$

$$MAD_i = \text{median}_j \{ | M_{ij} - \text{median}_j(M_{ij}) | \}.$$



# Between-Array Normalization



31/41

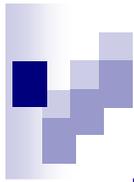
- In order to be able to compare the samples hybridized to different arrays on an equal footing, it is necessary to correct for the variability introduced by using multiple arrays.
- **Assumption:** the variations in the distribution between arrays are a result of experimental conditions and do not represent biological variability.

## Scaling

- Data is scaled to ensure that the means of all the distributions are equal.
- Subtract the mean log ratio (or log intensity) of all of the data on the array from each log ratio (or log intensity) measurement on the array.
- The mean of the measurements on each array will be zero after normalization.
- Use median: more robust measure of the average intensity on an array in situations where there are outliers or the intensities are not normally distributed.

## Centering

- Data is centered to ensure that the means and the standard deviations of all of the distributions are equal.
- Following centering, the mean of the measurements on each array will be zero, and the standard deviation will be 1.
- Particularly useful when calculating the Pearson correlation coefficient of a large number of data sets prior to cluster analysis, because it ensures that the correlation coefficient can define a distance metric on the data.
- **Use median and median absolute deviation from the median (MAD).**
- **This has the advantage of being more robust to outliers than using the mean and standard deviation, but has the disadvantage of not producing a distance metric when using Pearson correlation.**



# Distribution Normalization



32/41

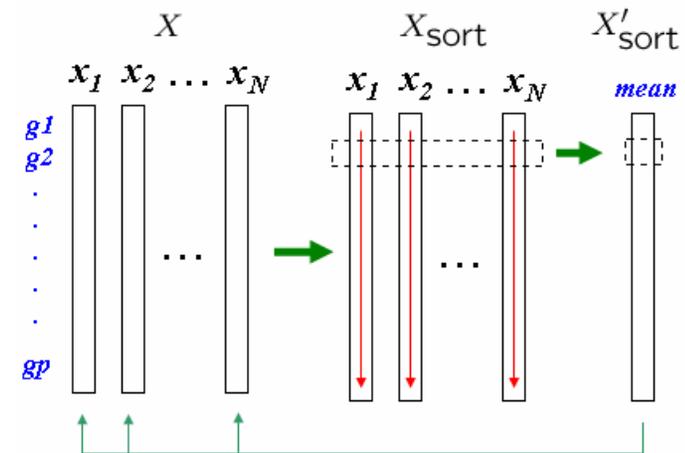
Data is distribution normalized to ensure that the distribution of the data on each of the arrays are identical.

- Center the data.
- For each array, order the centered measurements from lowest to highest.
- Compute a new distribution whose lowest value is the **average** of the value of the lowest expressed gene on each of the array; whose second-lowest value is the **average** of the second-lowest values from each of the array; and so on. until the highest value is the average value of the highest values from each of the arrays.
- Replace each measurement on each array with the corresponding average in the new distribution.  
(If a particular measurement is the 100th largest value on the array, replace it with 100th largest value in the new distribution. )
- Following distribution normalization, the measurements of each array will have mean 0, standard deviation 1, and identical distribution to all other arrays.
- It is useful where the different arrays have different distributions of values.

# Quantiles Normalization

- Quantiles Normalization (Bolstad *et al*, 2003) is a method to make the distribution of probe intensities the same for every chip. That is each chip is really the transformation of an underlying common distribution.
- The two distribution functions are effectively estimated by the sample quantiles.
- The normalization distribution is chosen by averaging each quantile across chips.

1. Given  $N$  datasets of length  $p$  form  $X$  of dimension  $p \times N$  where each dataset is a column
2. Set  $d = \left( \frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}} \right)$
3. Sort each column of  $X$  to give  $X_{\text{sort}}$
4. Project each row of  $X_{\text{sort}}$  onto  $d$  to get  $X'_{\text{sort}}$
5. Get  $X_{\text{norm}}$  by rearranging each column of  $X'_{\text{sort}}$  to have the same ordering as original  $X$

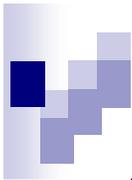


$X_{\text{norm}}$   
rearranging each column of  $X'_{\text{sort}}$   
to have the same ordering as original  $X$

1. If  $q_i = (q_{i1}, \dots, q_{iN})$  is a row in  $X_{\text{sort}}$  then the corresponding row in  $X'_{\text{sort}}$  is given by  $q'_i = \text{proj}_d q_i$
2. The projection is equivalent to taking the average of the quantile in a particular row and substituting this value for each of the individual elements in that row

$$\text{proj}_d q_i = \frac{q_i \cdot d}{d \cdot d} d = \frac{1}{\sqrt{N}} \sum_{j=1}^N q_{ij} d = \left( \frac{1}{N} \sum_{j=1}^N q_{ij}, \dots, \frac{1}{N} \sum_{j=1}^N q_{ij} \right)$$

The  $q$ th quantile of a data set is defined as that value where a  $q$  fraction of the data is below that value and  $(1-q)$  fraction of the data is above that value. For example, the 0.5 quantile is the median.



# Slef-Normalization



Paired-slides normalization applies to dye-swap experiments:  
two hybridizations for two mRNA samples, with dye assignment reversed in the second hybridization.

- first slide by  $\log_2 R/G - c$   
second slide by  $\log_2 R'/G' - c'$
- The normalized log-ratios on the two slides are of equal magnitude and opposite sign.

$$\log_2 R/G - c \approx -(\log_2 R'/G' - c').$$

$c$  and  $c'$  denote the normalization functions for the two slides;

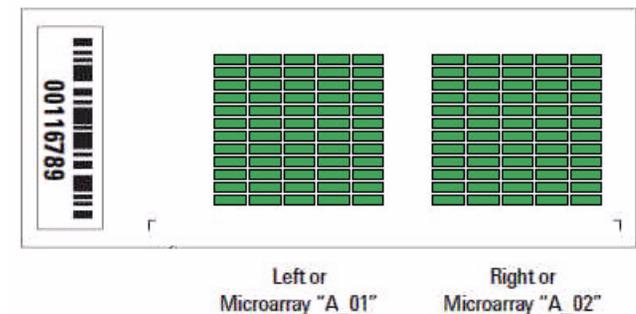
- If  $c \approx c'$ , then  $c \approx \frac{1}{2} \left[ \log_2 R/G + \log_2 R'/G' \right] = \frac{1}{2}(M + M')$ .

- The normalized  $\log_2$ -ratios will then be

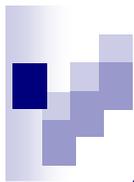
$$\frac{1}{2} \left[ \log_2 R/G - c - (\log_2 R'/G' - c') \right]$$

$$\approx \frac{1}{2} \left[ \log_2 R/G + \log_2 G'/R' \right] = \frac{1}{2} \log_2 \frac{RG'}{GR'} = \frac{1}{2}(M - M').$$

## Agilent Microarray Layout



- In practice,  $c = c(A)$  is estimated by the **lowess** fit to the plot of  $\frac{1}{2}(M + M')$  vs.  $\frac{1}{2}(A + A')$



# Normalization Methods



## Within-array Normalization

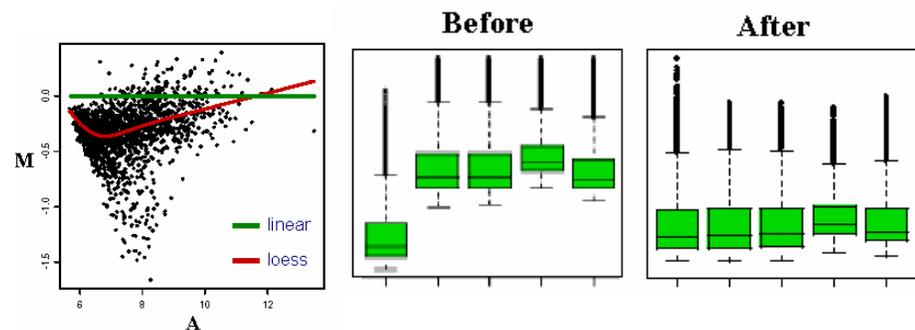
Subject be used for estimating normalization curve				
<b>Location Normalization</b>				
Method	allGenes	Print-tip $i$	2D Location $(x, y)$	SelectedGenes (Controls, Housekeeping, MSP, Invariant set)
constant	global normalization $N = M - c$ $c$ : mean, median	print-tip normalization $N = M - c_i$		
loess (Robust scatterplot smoother: loess, spline,...)	global loess normalization $N = M - c(A)$ $c$ : loess curve	print-tip loess normalization $N = M - c_i(A)$	2D loess normalization $N = M - c(x, y) - c(A)$	Smyth and Speed (2003) $N = M - p_A c_{MSP}(A) - (1 - p_A) c_i(A)$
<b>Scale Normalization</b>				
MAD	global scale normalization $N = s \times M$ $s = 1/mad(A)$	print-tip scale normalization $N = s_i \times M$ $s_i = 1/mad_i(A)$		
STD	standardization $N = M - ave(M)/std(A)$			

## Between-array Normalization

Scale-normalization: scaling of the M-values from a series of arrays so that each array has the same

$$MAD = \text{median}|M - \text{median}(M)|$$

## Paired-array Normalization (Dye-swap)



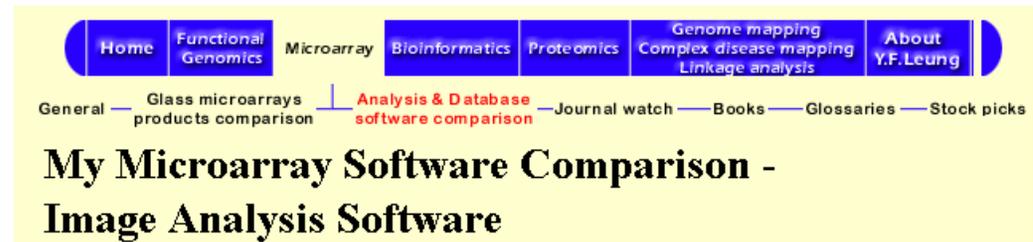
## Image Analysis/Normalization

### Freeware/Shareware

- ScanAlyze**
- Bioconductor: marrayNorm, arrayQuality
- Dapple

### Commercial

- GenePix**
- Spot



[http://ihome.cuhk.edu.hk/~b400559/arraysoft\\_image.html](http://ihome.cuhk.edu.hk/~b400559/arraysoft_image.html)

**BOOK:** Kamberova, G. and Shah, S. 2002. DNA Array Image Analysis: Nuts and Bolts. DNA press.

# ScanAlyze v2.50

Updated November 27, 2002



Eisen Lab

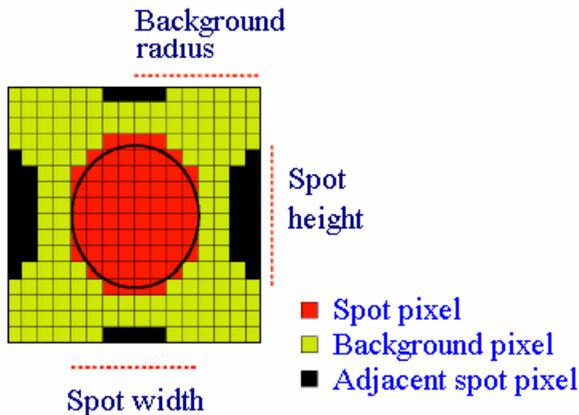
<http://rana.lbl.gov/EisenSoftware.htm>

Eisen MB, Spellman PT, Brown PO and Botstein D. (1998).  
**Cluster Analysis and Display of Genome-Wide Expression Patterns.**  
*Proc Natl Acad Sci U S A* 95, 14863-8.

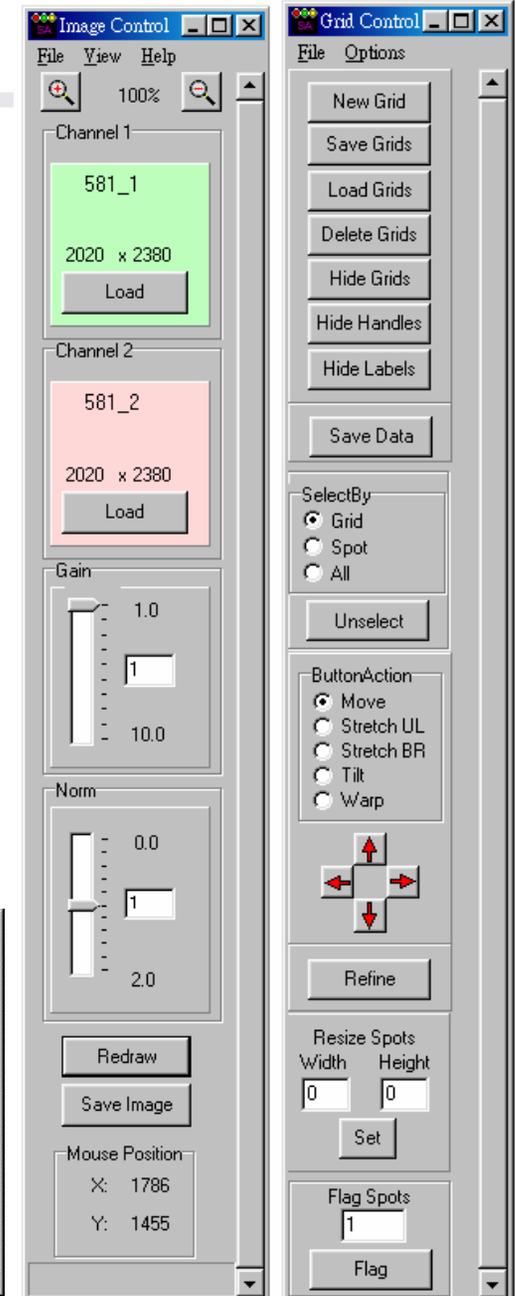
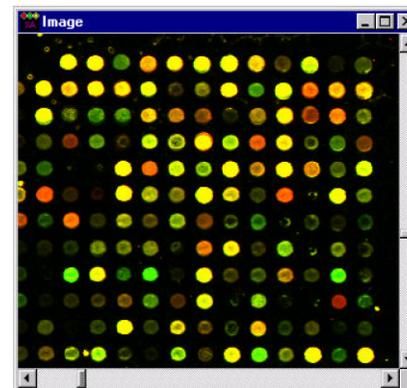
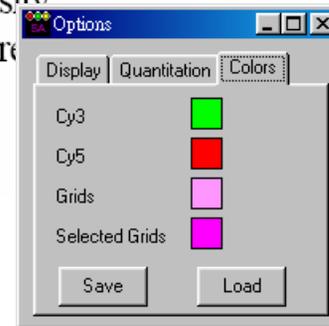
ScanAlyze supports two image formats:

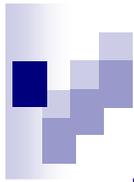
- 1) A raw image file (.SCN) used at Stanford University
- 2) Standard 8- and 16-bit TIFFs (all images are stored internally as 16-bit)

Separate files are required for each fluor analyzed



Lawrence Berkeley National Lab (LBNL)  
 University of California at Berkeley (UCB).





# Dapple v0.88



<http://www.cse.wustl.edu/~jbuhler/research/dapple/>

Jeremy Buhler  
Department of Computer Science and  
Engineering at Washington University  
in St. Louis

- Dapple is a program for quantitating spots on a two-color DNA microarray image.
- Given a pair of images from a comparative hybridization, Dapple finds the individual spots on the image, evaluates their qualities, and quantifies their total fluorescent intensities.

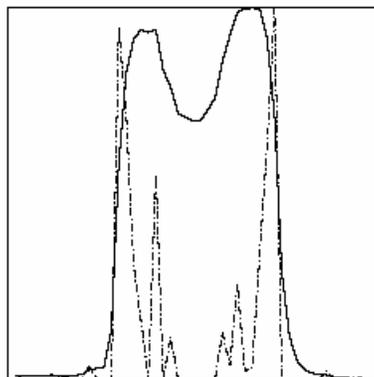
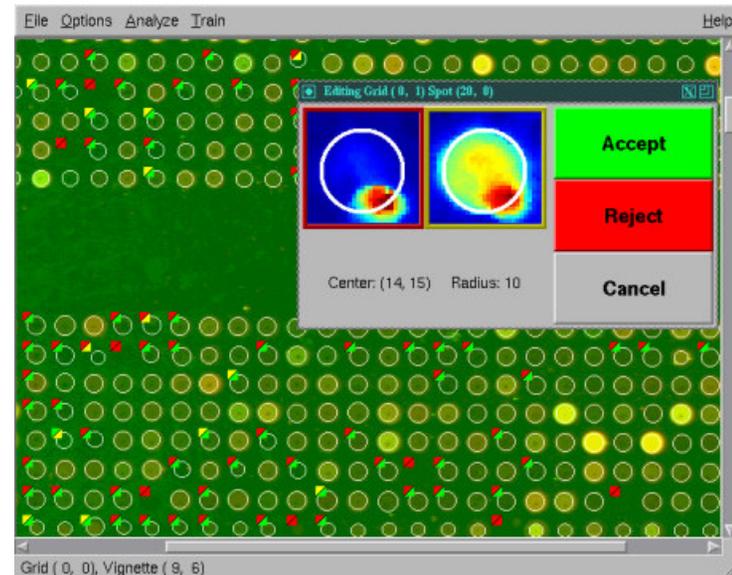


Figure 2

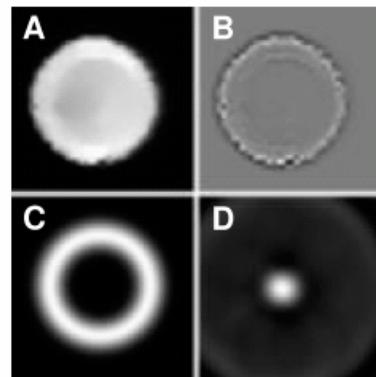
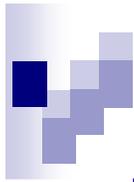


Figure 3

Figure 2: an intensity cross-section of a typical spot (solid line) overlaid with its negative second derivative (dashed line; only regions greater than zero are shown). The negative second derivative is maximized at the upper edge of the spot.

Figure 3: the spot finding process. (A) a vignette image containing a spot; (B) the negative Laplacian  $L$  of the image, which highlights the spot's edge; (C) a circular filter  $M_r$  ( $r = 11$  pixels) which matches the spot's radius; (D) the 2D correlation of  $M_r$  at all offsets against  $L$ , which is maximized at the spot's center.



# Spot v1.2



The R Project for  
Statistical Computing



39/41

<http://spot.cmis.csiro.au/spot/index.php>

- CSIRO provides a free evaluation version of **Spot** image analysis software on a limited 30 day licence.
- Features
  - Automatic grid location.
  - Flexible spot segmentation.
  - Morphological background estimation.
- Segmentation: seeded region growing technique (Adams and Bischof, 1994).
- Batch automatic addressing.
- Segmentation. Seeded region growing (Adams & Bischof 1994): adaptive segmentation method, no restriction on the size or shape of the spots.
- Information extraction
  - Foreground. Mean of pixel intensities within a spot.
  - Background. Morphological opening: non-linear filter which generates an image of the estimated background intensity for the entire slide.
- Spot quality measures.

The screenshot shows a web browser window titled "Spot Home Page - Microsoft Internet Explorer". The address bar contains the URL "http://spot.cmis.csiro.au/spot/index.php". The page content includes the CSIRO logo, the title "Spot: Software for Analysis of Microarray Images", and the text "CSIRO Mathematical and Information Sciences, Image Analysis Group". There are two images showing microarray spots: one with a grid overlay and one with individual spots highlighted in various colors. The page lists several links: "Spot Home", "Purchase Spot", "Spot Price List", "Download Evaluation", "Installation Instructions", "Spot User Manual", and "FAQ". A "Summary of Features:" section lists: "Automatic grid location.", "Flexible spot segmentation.", and "Morphological background estimation." The browser's status bar at the bottom shows "完成" (Done) and "網際網路" (Internet).

# The Bioconductor v1.7



<http://www.bioconductor.org/>

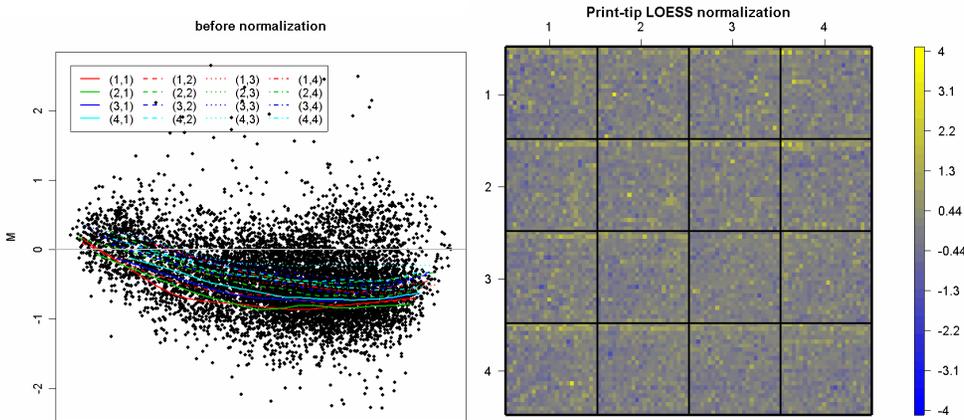


**library(arrayQuality)**

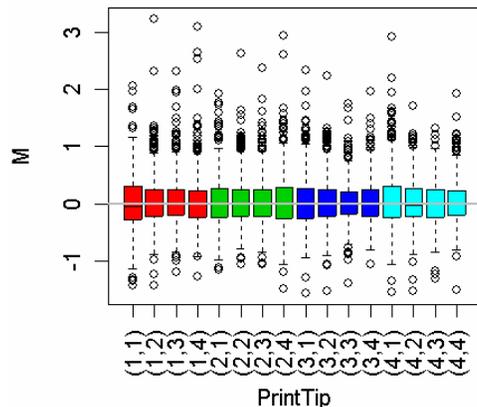
<http://www.sandler-fgcf.ucsf.edu/software/>

**library(arrayMagic)**

two-colour cDNA array quality control and preprocessing



after print-tip loess normalization



**library(marrayInput)**

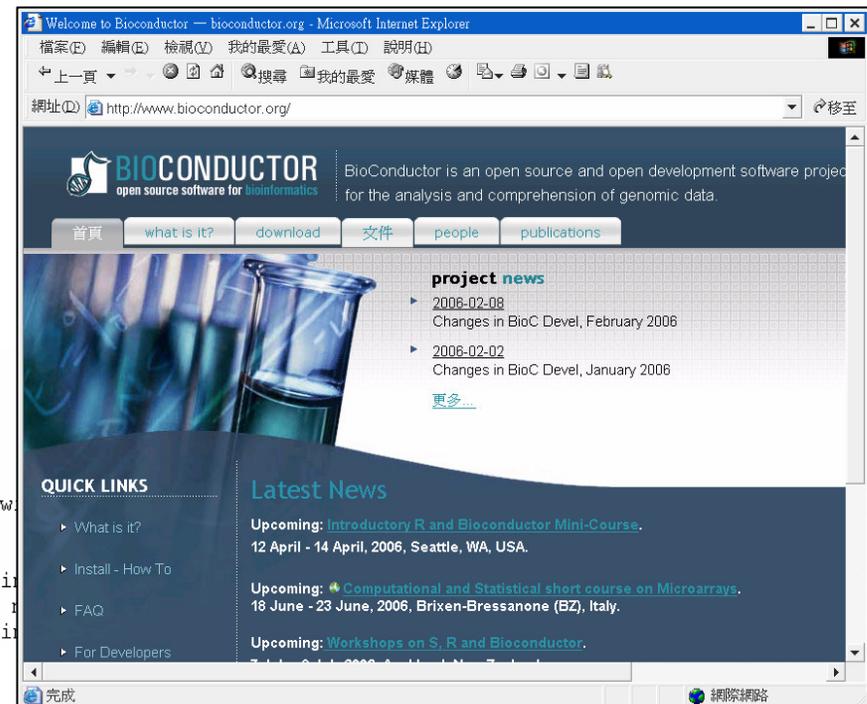
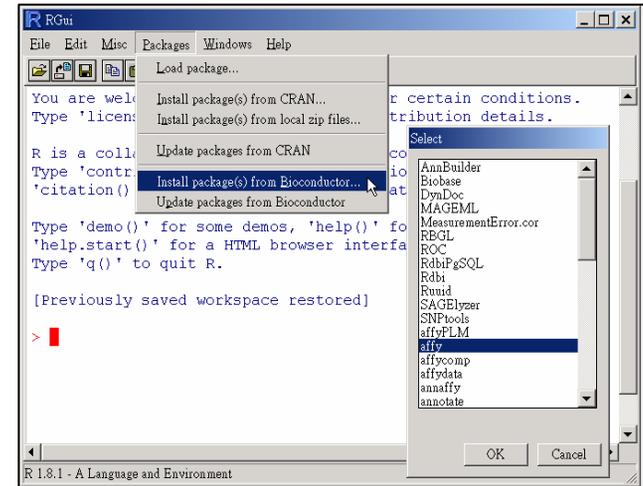
```
data(swirl)
swirl.3 <- swirl[,3]
```

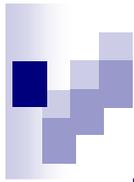
**library(marrayNorm)**

```
swirl.norm.printTipLoess <- maNorm(swirl.3)
```

**library(marrayPlots)**

```
maImage(swirl.norm.printTipLoess, main="array")
maBoxplot(swirl.norm.printTipLoess, main="array")
maPlot(swirl.norm.printTipLoess, main="array")
```





# GenePix Pro v6.0



[http://www.moleculardevices.com/pages/software/gn\\_genepix\\_pro.html](http://www.moleculardevices.com/pages/software/gn_genepix_pro.html)



Microarray Scanners  
Bioinformatics Software

- GenePix Pro is the complete standalone image analysis software for microarrays, tissue arrays and cell arrays

- IMAGING TOOLS
  - Display
  - Feature-Finding
  - Measuring Tools
  - Image Alignment

- ANALYSIS TOOLS
  - Background Subtraction
  - The Feature Viewer
  - The Feature Pixel Plot
  - Advanced Quality Control Flagging
  - Extracting Data from Images
  - Normalization
  - Web Integration

- AUTOMATION
- VISUALIZATIONS
- REPORTS AND SCRIPTING

